Statistics and Its Interface Volume 0 (0000) 1–13

Online Multiple Learning with Working Sufficient Statistics for Generalized Linear Models in Big Data*

Tonglin Zhang[†,‡], and Baijian Yang[§]

The article proposes an online multiple learning approach to generalized linear models (GLMs) in big data. The approach relies on a new concept called working sufficient statistics (WSS), formulated under traditional iteratively reweighted least squares (IRWLS) for maximum likelihood of GLMs. Because the algorithm needs to access the entire data set multiple times, it is impossible to directly apply traditional IRWLS to big data. To overcome the difficulty, a new approach, called one-step IRWLS, is proposed under the framework of the online setting. The work investigates two methods. The first only uses the current data to formulate the objective function. The second also uses the information of the previous data. The simulation studies show that the results given by the second method can be as precise and accurate as those given by the exact maximum likelihood. A nice property is that one-step IRWLS successfully avoids the memory and computational efficiency barriers caused by the volume of big data. As the size of the WSS does not vary with the sample size, the proposed approach can be used even if the size of big data is much higher than the memory size of the computing system.

AMS 2000 subject classifications: Primary 62J12, 62F10; secondary 62E20.

Keywords and phrases: Big data, Generalized linear models, One-step IRWLS, Online multiple learning, Parallel computation, Working sufficient statistics.

1. INTRODUCTION

Recently, rapid advances in science and technology have brought unprecedented opportunities and challenges to tackle much larger and more complicated data in academics and industry. Due to the situation that data are recorded everyday, sizes of big data can be as high as thousands of petabytes, leading to a large volume and a wide variety of data that have to be treated carefully online before integration into a system. Because of volume and variety issues, it is more efficient to study a number of models together rather than a single model when data are accessed. The goal of the article is to develop such an approach.

The development of statistical approaches for big data under the online setting is more challenging than that under the offline setting. In many situations, massive data arrive in streams or large chunks, leading to the need of online approaches. An online approach provides updating results sequentially without storage requirements of previous results. Rather than an offline approach, it cannot simultaneously access the entire data even once. It can only use the current data and a set of summary information for the previous data. The size of the summary information must be lower than the memory size of the computing system. An updated set of summary information is derived after the current data and the previous summary information are combined. Its size cannot be significantly inflated by the combination. Estimates of model parameters and their variance-covariance matrix can be only computed by the set of summary information.

Our work fulfills recent research topics on big data, which has gained remarkable attention in both academic and industry. Because of the memory and computational efficiency barriers, the implementation of traditional approaches is impossible. Many approaches have been developed. Examples include the sampling-based [8, 16], the divide-and-conquer [7], the divide-and-recombine [6], the sequential updating [3], the online updating [12, 15], and the scanning data by rows [18, 19, 20, 22]. In the case when the size of big data is higher than the memory size of a personal computer, the computation is often distributed across multiple processors such that the entire job can be finished in a reasonable amount of time [10]. Following this line, the MapReduce [2] and Spark [17] frameworks have been widely adopted. Their main interest is to approximately compute maximum likelihood estimates (MLEs) of model parameters in a given statistical model. If another model is considered, then the entire approach must be used again. Since the access of a big data set is time-consuming, it is inefficient to study individual models separately, leading to the need of multiple learning approaches.

arXiv: math/0000.0000

[*] Footnote to the title with the 'thankstext' command.

[†] Some comment

[‡] First supporter of the project

[§] Second supporter of the project

The construction of summary information is important in statistical approaches for big data under the online setting. This issue has been previously studied under the framework of linear models for normal data [18, 19]. The basic technique is the usage of a set of sufficient statistics (SS), which is given by the matrix of the cross product of variables involved in the linear model. The MLEs of model parameters and their standard errors can be exactly derived by the matrix of the cross product. Because the size of the matrix does not depend on the sample size, the method can be used even if the sample size is extremely large. The matrix of the cross product can be easily updated if new observations are collected, implying that it can also be used under the online setting. More importantly, the method can be used to study a number of models simultaneously as one can purposely construct the matrix for all the related models.

We extend the idea of SS from linear models to GLMs. An obvious difficulty is that the size of SS in a GLM for nonnormal data is often identical or close to the size of the observed data. We cannot use SS to reduce the memory needed in the computation. To overcome the difficulty, we study traditional iteratively reweighted least squares (IRWLS) for MLEs of GLMs [5]. We develop (online) one-step IRWLS by a new concept called working sufficient statistics (WSS). We note that traditional IRWLS is the standard fitting method for MLEs of GLMs, which has been adopted by many software packages, such as R, SAS, Python, and MatLab. Traditional IRWLS uses the weighted least squares (WLS) method under a working weighted linear model (WWLM) in each iteration. As the WWLM is a linear model for normal data, the IRWLS successfully changes the computation from nonnormal data to normal data. Then, we derive our one-step IRWLS.

We investigate two kinds of implementations. The first kind of implementations is divide-and-conquer. It does not use any information of previous data in the construction of the objective function for the current data. The second kind of implementations is online updating. It uses the information of previous data in the construction. The difference between the two kinds of implementations only affects the construction of the working response vector and the working weights in the WWLM. The rest steps are identical. We find that the second kind of implementations provides more reliable results than the first.

The primary difference between SS and WSS is that the derivation of SS only needs the likelihood function, but the derivation of WSS also needs numerical algorithms. The difficulty caused by the size of big data can be overcome by using the concept of WSS. In the case when the size of data is lower than the memory size of a personal computer, we compare our approach with traditional IRWLS. We apply one-step IRWLS to GLMs for binomial big data with the logistic link and Poisson data with the log link. We find that the results given by one-step IRWLS are close to those given by traditional IRWLS. We also apply our approach to the case when the size of data exceeds the memory size of a personal computer, where traditional IRWLS cannot be applied. Similar to the SS approach for linear models, our approach can also be used to study a number of models simultaneously. Therefore, we classify it as an online multiple learning approach for GLMs in big data.

The article is organized as follows. In Section 2, we provide a brief review of traditional IRWLS for small or moderate data. In Section 3, we present our approach. In Section 4, we specify our approach to three well-known statistical models. In Section 5, we evaluate our proposed approach using Monte Carlo simulations. In Section 6, we apply our approach to a real data example. In Section 7, we provide a discussion.

2. TRADITIONAL IRWLS

Traditional IRWLS is a standard fitting method for GLMs which has been adopted by many software packages. It can provide exact MLEs of GLMs with arbitrary links. Traditional IRWLS is an iterative method. It computes the MLE of a WWLM for normal data in each of its iterations. The weight and response values of the WWLM are updated iteratively. The exact MLE is derived if the algorithm converges.

GLMs are defined on exponential family distributions [1, Chapter 4]. The purpose is to model expected values of a response variable via a number of explanatory variables. Three components are needed to define a GLM. The random component consists of a response vector $\mathbf{y} = (y_1, \ldots, y_n)^\top$, where $y_1, \ldots, y_n$ are independently obtained from an exponential family distribution, and n is the sample size. An exponential family distribution has a probability mass function (PMF) or a probability density function (PDF) as

$$f(y_i) = \exp\left[\frac{y_i\omega_i - b(\omega_i)}{a(\phi)} + c(y_i, \phi)\right], \tag{1}$$

where ω_i is a canonical parameter representing the location, and ϕ is a dispersion parameter representing the scale. The linear component $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top$ is a vector related to explanatory variables by $\eta_i = \mathbf{x}_i^\top\boldsymbol{\beta}$ for all $i \in \{1, \ldots, n\}$, where $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{i(p-1)})^\top$ is the ith observed vector of explanatory variables, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$is a vector of regression coefficients. The link function $g(\cdot)$ connects $\mu_i = \mathrm{E}(y_i) = b'(\omega_i)$ and η_i through

$$\eta_i = g(\mu_i) = g[b'(\omega_i)] = \mathbf{x}_i^\top\boldsymbol{\beta} \tag{2}$$

for all $i \in \{1, \ldots, n\}$. The variance of the response is given by $\mathrm{V}(y_i) = a(\phi)b''(\omega_i)$. The variance function of the model is $v(\mu) = b''\{h^{-1}[g(\mu)]\}$, where $\omega_i = h(\mathbf{x}_i^\top\boldsymbol{\beta})$ is the inverse function obtained by (2). If the canonical link is used, then (2) reduces to $\eta_i = \omega_i = g(\mu_i) = \mathbf{x}_i^\top\boldsymbol{\beta}$, implying that $h(\cdot)$ is the identity function.

The MLEs of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$, must be solved numerically if the distribution is not normal. A popular and well accepted method is traditional IRWLS [5]. It is modified from the Fisher scoring method, and is identical to the Newton-Raphson method under the canonical link.

Let $\boldsymbol{\mu}^{(r)} = (\mu_1^{(r)}, \ldots, \mu_n^{(r)})^\top$ and $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top)^\top$, where $\mu_i^{(r)} = g^{-1}(\eta_i^{(r)})$, $\eta_i^{(r)} = \mathbf{x}_i^\top \boldsymbol{\beta}^{(r)}$, and $\boldsymbol{\beta}^{(r)}$ is the rth iterated value of $\hat{\boldsymbol{\beta}}$. Let

$$w(\eta) = (\partial\mu/\partial\eta)^2 / b''[h(\eta)] \tag{3}$$

and

$$u(\eta) = \eta + (y - \mu)(\partial\eta/\partial\mu) \tag{4}$$

be functions of η, where y, μ, and η are the general notations of the response, the expected value, and the linear component given by (1) and (2), respectively. Then, $\mathbf{W}^{(r)} = \mathrm{diag}(w_1^{(r)}, \ldots, w_n^{(r)})$ with $w_i^{(r)} = w(\eta_i^{(r)})$ is the working weight matrix and $\mathbf{u}^{(r)} = (u_1^{(r)}, \ldots, u_n^{(r)})^\top$ with $u_i^{(r)} = u(\eta_i^{(r)})$ is the working response vector given by traditional IRWLS. The Fisher-scoring method updates the solution of $\hat{\boldsymbol{\beta}}$ by $(\mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{X})\boldsymbol{\beta}^{(r+1)} = \mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{u}^{(r)}$. Equivalently, $\boldsymbol{\beta}^{(r+1)}$ is the MLE of the WWLM given by

$$\mathbf{u}^{(r)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\{\mathbf{W}^{(r)}\}^{-1})$, and $\sigma^2 = a(\phi)$. Traditional IRWLS provides the final solution of $\hat{\boldsymbol{\beta}}$ with initial $\mathbf{u}^{(0)}$ and $\mathbf{W}^{(0)}$. The ith components of $\mathbf{u}^{(0)}$ and $\mathbf{W}^{(0)}$ may be taken as the conventional choices $u_i^{(c)} = u^{(c)}(y_i) = g(y_i)$ and $w_i^{(c)} = w^{(c)}(y_i) = \{[g'(y_i)]^2 b''[h(u_i^{(c)})]\}^{-1}$, respectively, leading to $\mathbf{u}^{(0)} = \mathbf{u}^{(c)}$ and $\mathbf{W}^{(0)} = \mathbf{W}^{(c)}$ in many software packages. The formulations of $u_i^{(c)}$ and $w_i^{(c)}$ are often modified if $g(y_i)$ is not well-defined. In general, $\mathbf{u}^{(0)}$ and $\mathbf{W}^{(0)}$ are not necessary to be identical to $\mathbf{u}^{(c)}$ and $\mathbf{W}^{(c)}$, respectively. The implementation of IRWLS can be flexible. After $\hat{\boldsymbol{\beta}}$ is derived, a straightforward method to estimate ϕ is given by moment estimation [9] as

$$a(\hat{\phi}) = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{b''[h(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}})]}, \tag{6}$$

where $\hat{\mu}_i = b'[h(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}})]$. If ϕ is not present in (1), then (6) is not needed. This occurs in Bernoulli, binomial, and Poisson models.

3. APPROACH

We present our approach in this section. The basic version of (online) one-step IRWLS is introduced in Section 3.1. The extended online updating version is introduced in Section 3.2. A framework for online parallel computation is developed in Section 3.3. The asymptotic properties of our approach are studied in Section 3.4. A multiple learning framework is proposed in Section 3.5. An offline extension is displayed in Section 3.6.

3.1 One-Step IRWLS

The aim of one-step IRWLS is to overcome the difficulty caused by the volume of big data under the online setting. It assumes that an online learning system receives data in a data stream. The system sequentially receives blocks (or trunks) of data denoted by $\mathcal{D}_t = \{(\mathbf{y}_t, \mathbf{X}_t)\}$ for all $t \in \{1, \ldots, T\}$, where T is the current time, t with $t < T$ represents the previous times, $\mathbf{y}_t = (y_{t1}, \ldots, y_{tn_t})^\top$ is the response vector, $\mathbf{X}_t = (\mathbf{x}_{t1}^\top, \ldots, \mathbf{x}_{tn_t}^\top)^\top$ with $\mathbf{x}_{ti} = (1, x_{ti1}, \ldots, x_{ti(p-1)})^\top$ is the design matrix, and n_t is the sample size at time t. The total sample size is $n = \sum_{t=1}^{T} n_t$. Then, $\mathcal{D}_{T+} = \bigcup_{t=1}^{T} \mathcal{D}_t$ is the set of the entire data, and $\mathcal{D}_{(T-1)+}$ is the set of previous data.

We assume that y_{ti} is independently observed from (1) with the relationship between y_{ti} and $\mathbf{x}_{ti}$ given by (2), for all $i \in \{1, \ldots, n_t\}$ and $t \in \{1, \ldots, T\}$. Under the online setting, the system can use $\mathcal{D}_T$ but not $\mathcal{D}_t$ for any $t < T$. Instead, it can use a set of summary information for $\mathcal{D}_{(T-1)+}$. An updated set of summary information is derived by combining the information of $\mathcal{D}_T$ with the set of the summary information for $\mathcal{D}_{(T-1)+}$. The size of the set of summary information cannot be significantly inflated as T increases. Otherwise, the computation will be out-of-memory quickly.

If y_{ti} is normally distributed, then the GLM becomes a linear model. A set of summary information can be straightforwardly constructed by the cross product of the response and explanatory variables, which is equivalent to the set of SS in the model. It can be easily updated by combining the information in $\mathcal{D}_T$ with the set of SS for the previous data. We will revisit this problem in Section 4.1.

If the distribution of y_{ti} is not normal, then the size of SS is often identical to (or slightly lower than) the size of the entire data. It is impossible to use the method of SS to define the set of summary information. We recommend using working sufficient statistics (WSS).

We find that the traditional IRWLS given by (5) can be specified with the working response and weight values. Once they are derived, (5) becomes a weighted linear model (WLM). The computation of the next iterated value of $\hat{\boldsymbol{\beta}}$ only needs the matrix of the cross product in the WLM, which provides the definition of WSS. The usage of WSS can provide the exact answer of the next iterated value of $\hat{\boldsymbol{\beta}}$. As the entire data set is accessed multiple times, this method cannot be used under the online setting.

We propose one-step IRWLS to overcome the difficulty. The approach only needs initial values of the working response and the working weight vectors. The idea is to construct a WLM for the MLE of the GLM, leading to a normal approach to nonnormal data. As the solution provided by the approach is not the exact MLE, it is important to investigate its theoretical properties. This issue will be discussed in Section 3.4.

One-step IRWLS starts with initial $\mathbf{u}_t^{(0)} = (u_{t1}^{(0)}, \ldots, u_{tn_t}^{(0)})^\top$ and $\mathbf{W}_t^{(0)} = \mathrm{diag}(w_{t1}^{(0)}, \ldots, w_{tn_t}^{(0)})$, which depend on $\mathcal{D}_{t+}$ only, for all $t \in \{1, \dots, T\}$. They may not be identical to $\mathbf{u}_t^{(c)}$ and $\mathbf{W}_t^{(t)}$, the conventional choices used by many software packages. The WWLM given by (5) at t becomes

$$\mathbf{u}_t^{(0)} = \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\epsilon}_t, \tag{7}$$

where $\boldsymbol{\epsilon}_t \sim^{ind} \mathcal{N}(\mathbf{0}, \sigma^2\{\mathbf{W}_t^{(0)}\}^{-1})$ for all $t \in \{1, \dots, T\}$. Because (7) is basically a normal model, we can use properties of normal likelihood to study computational issues.

We express the set of WSS as an unstructured array, which is developed under the loglikelihood function of (7) as

$$\ell^{(0)}(\boldsymbol{\beta}, \sigma^2) = \sum_{t=1}^{T} \ell_t^{(0)}(\boldsymbol{\beta}, \sigma^2), \tag{8}$$

where

$$\begin{aligned}\ell_t^{(0)}(\boldsymbol{\beta}, \sigma^2) = &-\frac{n_t}{2}[\log(2\pi) + \log\sigma^2] - \frac{1}{2}\sum_{i=1}^{n_t}\log w_{ti}^{(0)} \\ &- \frac{1}{2\sigma^2}\sum_{i=1}^{n_t} w_{ti}^{(0)}(u_{ti}^{(0)} - \mathbf{x}_{ti}^\top\boldsymbol{\beta})^2\end{aligned} \tag{9}$$

is the loglikelihood function of $\boldsymbol{\beta}$ and σ^2 of (7) in $\mathcal{D}_t$. By the standard method, we can show that $\{s_t^{(0)}, \mathbf{s}_t^{(0)}, \mathbf{S}_t^{(0)}\}$ with $s_t^{(0)} = \sum_{i=1}^{n_t} w_{ti}^{(0)}\{u_{ti}^{(0)}\}^2$, $\mathbf{s}_t^{(0)} = \sum_{i=1}^{n_t} w_{ti}^{(0)} u_{ti}^{(0)}\mathbf{x}_{ti}$, and $\mathbf{S}_t^{(0)} = \sum_{i=1}^{n_t} w_{ti}^{(0)}\mathbf{x}_{ti}\mathbf{x}_{ti}^\top$ is a set of SS of (7) in $\mathcal{D}_t$. We express those by an unstructured array as

$$\mathcal{S}_t^{(0)} = (s_t^{(0)}, \mathbf{s}_t^{(0)}, \mathbf{S}_t^{(0)}), \tag{10}$$

such that

$$\mathcal{S}_{T+}^{(0)} = \sum_{t=1}^{T} \mathcal{S}_t^{(0)} = (s_{T+}^{(0)}, \mathbf{s}_{T+}^{(0)}, \mathbf{S}_{T+}^{(0)}) \tag{11}$$

with $s_{T+}^{(0)} = \sum_{t=1}^{T} s_t^{(0)}$, $\mathbf{s}_{T+}^{(0)} = \sum_{t=1}^{T} \mathbf{s}_t^{(0)}$, and $\mathbf{S}_{T+}^{(0)} = \sum_{t=1}^{T} \mathbf{S}_t^{(0)}$ is an unstructured array of SS of (7) in $\mathcal{D}_{T+}$. Then, $\mathcal{S}_t^{(0)}$ and $\mathcal{S}_{T+}^{(0)}$ are the unstructured arrays of WSS for the GLM given by (1) and (2) in $\mathcal{D}_t$ and $\mathcal{D}_{T+}$, respectively.

In both $\mathcal{S}_t^{(0)}$ and $\mathcal{S}_{T+}^{(0)}$, the first component is a numerical value, the second component is a p-dimensional vector, and the third component is a $p \times p$-dimensional matrix. Their sizes are identical and do not vary with n. Given that $\mathcal{S}_{(T-1)+}^{(0)}$ is derived, we can use it to compute $\mathcal{S}_T^{(0)}$, leading to an updating equation as

$$\mathcal{S}_{T+}^{(0)} = \mathcal{S}_{(T-1)+}^{(0)} + \mathcal{S}_T^{(0)}. \tag{12}$$

Therefore, $\mathcal{S}_{T+}^{(0)}$ can be updated sequentially, indicating that it can be used as the set of summary information under the online setting.

One-step IRWLS uses $\mathcal{S}_{T+}^{(0)}$ to fit (7). The result can be directly obtained by traditional WLS. The one-step IRWLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{one} = \hat{\boldsymbol{\beta}}_{one,T} = \{\mathbf{S}_{T+}^{(0)}\}^{-1}\mathbf{s}_{T+}^{(0)} \tag{13}$$

with the variance-covariance matrix as

$$\hat{\mathrm{V}}(\hat{\boldsymbol{\beta}}_{one}) = \hat{\sigma}_{one}^2\{\mathbf{S}_{T+}^{(0)}\}^{-1}, \tag{14}$$

where

$$\hat{\sigma}_{one}^2 = \frac{1}{n}\left(s_{T+}^{(0)} - \{\mathbf{s}_{T+}^{(0)}\}^\top\{\mathbf{S}_{T+}^{(0)}\}^{-1}\mathbf{s}_{T+}^{(0)}\right) \tag{15}$$

is the MSE of the model. We can derive all of those by $\mathcal{S}_{T+}^{(0)}$ only. We use the Wald statistic $z_j = \hat{\beta}_{one,j}/\hat{\sigma}_{\hat{\beta}_{one,j}}$ to test significance of β_j for all $j \in \{1, \dots, p-1\}$, where $\hat{\beta}_{one,j}$ is the jth component of $\hat{\boldsymbol{\beta}}_{one}$ given by (13), $\hat{\sigma}_{\hat{\beta}_{one,j}}$ is its standard error given by (14), and the p-value of z_j is calculated by the standard normal distribution.

Definition 3.1. *Let $u_{ti}^{(0)}$ and $w_{ti}^{(0)}$ be constructed by $\mathcal{D}_{t+}$ only with $t \in \{1, \dots, T\}$. The loglikelihood functions given by (8) and (9) are called the working loglikelihood functions of (7) in $\mathcal{D}_t$ and $\mathcal{D}_{T+}$, respectively. The arrays of sufficient statistics given by (10) and (11) are called the unstructured arrays of working sufficient statistics (WSS) of (7) in $\mathcal{D}_t$ and $\mathcal{D}_{T+}$, respectively. The (online) one-step IRWLS updates the unstructured array of WSS by (12) as T increases. It uses (13), (14), and (15) to approximately calculate the MLE of $\boldsymbol{\beta}$ and the estimate of its variance-covariance matrix in the GLM defined by (1) and (2), respectively.*

There are two kinds of implementations. The first is divide-and-conquer (D&C), called one-step IRWLS under D&C. It does not use any information of previous data in the construction of the working response and the work weight vectors. We introduce it in the following of this subsection. The second is online updating (UPD), called one-step IRWLS under UPD. It uses the information of previous data. We will introduce it in the next subsection.

We provide two estimators under the framework of D&C. The first estimator is derived by directly following the conventional method, which chooses $\mathbf{u}_t^{(0)} = \mathbf{u}_t^{(c)}$ and $\mathbf{W}_t^{(0)} = \mathbf{W}_t^{(c)}$ in Definition 3.1. The rest steps are exactly identical to (13), (14), and (15). The results are denoted by $\hat{\boldsymbol{\beta}}_{con}$, $\hat{\mathrm{V}}(\hat{\boldsymbol{\beta}}_{con})$, and $\hat{\sigma}_{con}^2$, respectively. We call this the conventional version of one-step IRWLS.

The second estimator is derived by using the individual MLEs of $\boldsymbol{\beta}$ in $\mathcal{D}_t$, denoted by $\hat{\boldsymbol{\beta}}_t$, in the construction of $\boldsymbol{\mu}_t^{(0)}$ and $\mathbf{W}_t^{(0)}$. Following traditional IRWLS, we have $u_{ti}^{(0)} = u(\hat{\eta}_{ti})$ and $w_{ti}^{(0)} = w(\hat{\eta}_{ti})$, where $\hat{\eta}_{ti} = \mathbf{x}_{ti}^\top\hat{\boldsymbol{\beta}}_t$, for $i \in \{1, \dots, n_t\}$. Then, $\mathbf{u}_t^{(0)} = \mathbf{X}_t\hat{\boldsymbol{\beta}}_t + (\mathbf{y}_t - \hat{\boldsymbol{\mu}}_t)(\partial\hat{\boldsymbol{\eta}}_t/\partial\hat{\boldsymbol{\mu}}_t)$ and $\mathbf{W}_t^{(0)} = \mathrm{diag}\{(\partial\hat{\boldsymbol{\mu}}_t/\partial\hat{\boldsymbol{\eta}}_t)^2/b''[h(\hat{\boldsymbol{\eta}}_t)]\}$. The results given by (13), (14), and (15) are denoted by $\hat{\boldsymbol\beta}_{mle}$, $\hat{\mathrm{V}}(\hat{\boldsymbol\beta}_{mle})$, and $\hat\sigma^2_{mle}$, respectively. We call this the MLE version of one-step IRWLS.

We next investigate the relationship between the MLE version of one-step IRWLS and the aggregated estimating equation (AEE) method proposed by [7]. Let

$$\ell_{t,\phi}(\boldsymbol\beta) = \sum_{i=1}^{n_t}\left\{c(y_{ti},\phi) + \frac{y_{ti}h(\mathbf{x}_{ti}^\top\boldsymbol\beta) - b[h(\mathbf{x}_{ti}^\top\boldsymbol\beta)]}{a(\phi)}\right\} \tag{16}$$

be the true loglikelihood function of the model based on $\mathcal{D}_t$. Let $\dot\ell_{t,\phi}(\boldsymbol\beta)$ be its gradient vector and $\ddot\ell_{t,\phi}(\boldsymbol\beta)$ be its Hessian matrix with respect to $\boldsymbol\beta$. The jth component of $\dot\ell_{t,\phi}(\boldsymbol\beta)$ is

$$\frac{\partial\ell_{t,\phi}(\boldsymbol\beta)}{\partial\beta_j} = \frac{1}{a(\phi)}\sum_{i=1}^{n_t}\frac{x_{tij}(y_{ti}-\mu_{ti})}{b''[h(\eta_{ti})]}\frac{\partial\mu_{ti}}{\partial\eta_{ti}}, \tag{17}$$

for all $j \in \{0,\ldots,p-1\}$. The $(j_1, j_2)th$ entry of $\ddot\ell_{t,\phi}(\boldsymbol\beta)$ is

$$\begin{aligned}\frac{\partial^2\ell_{t,\phi}(\boldsymbol\beta)}{\partial\beta_{j_1}\partial\beta_{j_2}} =& \frac{1}{a(\phi)}\sum_{i=1}^{n_t}\{(y_{ti}-\mu_{ti})h''(\eta_{ti}) \\ &- b''[h(\eta_{ti})][h'(\eta_{ti})]^2\}x_{tij_1}x_{tij_2},\end{aligned} \tag{18}$$

for all $j_1, j_2 \in \{0,\ldots,p-1\}$. The estimating function for $\boldsymbol\beta$ in $\mathcal{D}_t$ is $\dot\ell_{t,\phi}(\boldsymbol\beta) = 0$ and the solution is $\hat{\boldsymbol\beta}_t$. Let

$$\mathbf{A}_t = -a(\phi)\ddot\ell_{t,\phi}(\hat{\boldsymbol\beta}_t) \tag{19}$$

for all $t \in \{1,\ldots,T\}$. The Taylor expansion of $-a(\phi)\dot\ell_{t,\phi}(\boldsymbol\beta)$ at $\hat{\boldsymbol\beta}_t$ is $-a(\phi)\dot\ell_{t,\phi}(\boldsymbol\beta) = \mathbf{A}_t(\boldsymbol\beta - \hat{\boldsymbol\beta}_t) + \mathbf{R}_t$, where $\mathbf{R}_t$ is a remainder. The AEE estimator of $\boldsymbol\beta$ for the entire data set is formulated under the D&C framework, which combines individual estimators as

$$\hat{\boldsymbol\beta}_{aee} = \left(\sum_{t=1}^T \mathbf{A}_t\right)^{-1}\sum_{t=1}^T\mathbf{A}_t\hat{\boldsymbol\beta}_t. \tag{20}$$

It is a solution to $\sum_{t=1}^T \mathbf{A}_t(\boldsymbol\beta - \hat{\boldsymbol\beta}_t) = \mathbf{0}$.

Theorem 3.1. *If g is the canonical link in (2), then $\hat{\boldsymbol\beta}_{mle} = \hat{\boldsymbol\beta}_{aee}$.*

Proof: If g is the canonical link, then $h(\cdot)$ is the identical function and $\partial\mu_{ti}/\partial\eta_{ti} = b''(\hat\eta_{ti})$, $u_{ti}^{(0)} = \mathbf{x}_{ti}^\top\hat{\boldsymbol\beta}_t + (y_{ti} - \hat\mu_{ti})/b''(\hat\eta_{ti})$, $w_{ti}^{(0)} = b''(\hat\eta_{ti})$, and $-a(\phi)\partial^2\ell_{t,\phi}(\boldsymbol\beta)/(\partial\beta_{j1}\partial\beta_{j2}) = \sum_{i=1}^{n_t} b''(\eta_{ti})x_{tij_1}x_{tij_2}$. Then, $\mathbf{S}_t^{(0)} = \sum_{i=1}^{n_t} w_{ti}^{(0)}\mathbf{x}_{ti}\mathbf{x}_{ti}^\top = \mathbf{X}^\top\mathbf{W}_t^{(0)}\mathbf{X}_t$. By taking $\eta_{ti} = \hat\eta_{ti}$ and $\mu_{ti} = \hat\mu_{ti}$ in (17), we have $\mathbf{X}_t^\top\mathbf{W}_t^{(0)}\mathbf{u}_t^{(0)} = \mathbf{X}_t^\top\mathbf{W}_t^{(0)}\mathbf{X}_t\hat{\boldsymbol\beta}_t + \dot\ell_{t,\phi}(\hat{\boldsymbol\beta}_t)$. By $\dot\ell_{t,\phi}(\hat{\boldsymbol\beta}_t) = \mathbf{0}$, we have $\mathbf{s}_t^{(0)} = \mathbf{X}_t^\top\mathbf{W}_t^{(0)}\mathbf{X}\hat{\boldsymbol\beta}_t$. Thus, $\mathbf{S}_{T+}^{(0)} = \sum_{t=1}^T\mathbf{A}_t$ and $\mathbf{s}_{T+}^{(0)} = \sum_{t=1}^T\mathbf{A}_t\hat{\boldsymbol\beta}_t$, implying the conclusion. □

Corollary 3.1. *If we modify (19) as $\mathbf{A}_t = \mathbf{A}_t(\hat{\boldsymbol\beta}_t)$ with $\mathbf{A}_t(\boldsymbol\beta) = \mathrm{E}[-a(\phi)\ddot\ell_{t,\phi}(\boldsymbol\beta)]$, then $\hat{\boldsymbol\beta}_{mle} = \hat{\boldsymbol\beta}_{aee}$.*

Proof: If g is the canonical link in (2), then $\ddot\ell_{t,\phi}(\boldsymbol\beta) = \mathrm{E}[\ddot\ell_{t,\phi}(\boldsymbol\beta)]$ and we draw the conclusion by Theorem 3.1. For any other links, using the same method in the proof of Theorem 3.1, we can also show $\mathbf{S}_t^{(0)} = \sum_{t=1}^T\mathbf{X}_t^\top\mathbf{W}_t^{(0)}\mathbf{X}_t$ and $\mathbf{s}_{T+}^{(0)} = \sum_{t=1}^T\mathbf{A}_t\hat{\boldsymbol\beta}_t$. We draw the conclusion. □

Corollary 3.2. *If $w_{ti}^{(0)} = (y_{ti} - \hat\mu_{ti})h''(\hat\eta_{ti}) - (\partial\hat\mu_{ti}/\partial\hat\eta_{ti})^2/b''[h(\hat\eta_{ti})]$ is used in the computation of $\hat{\boldsymbol\beta}_{mle}$, then $\hat{\boldsymbol\beta}_{mle} = \hat{\boldsymbol\beta}_{aee}$.*

Proof: The conclusion can be similarly shown by the method of Corollary 3.1. □

We have proposed the conventional and MLE versions of one-step IRWLS under the framework of D&C. We show that the MLE version is equivalent to AEE if the canonical link is used in (2). With slight modifications, they can also be equivalent to each other under other links. A concern in the implementation of $\hat{\boldsymbol\beta}_{mle}$ and $\hat{\boldsymbol\beta}_{aee}$ is their existence. This can be caused by rank deficiency problems in $\mathbf{X}_t$ for some $t \in \{1,\ldots,T\}$ [12]. This usually does not occur in the conventional version of one-step IRWLS.

3.2 Online Updating

Under the online setting, all $\mathcal{S}_t^{(0)}$ for $t \in \{1,\ldots,T-1\}$ have been derived before the construction of the current $\mathcal{S}_T$, leading to our one-step IRWLS under UPD. In this method, we use $\mathcal{S}_{(T-1)+}$ in the construction $\mathbf{u}_T^{(0)}$ and $\mathbf{W}_T^{(0)}$. We then use (12) to compute $\mathcal{S}_{T+}$. This can be sequentially implemented under the online setting by increasing T.

By the combination of the working loglikelihood function for the previous data and the true loglikelihood function for the current data, we obtain an updating loglikelihood function as

$$\ell_{upd}(\boldsymbol\beta,\sigma^2) = \sum_{t=1}^{T-1}\ell_t^{(0)}(\boldsymbol\beta,\sigma^2) + \ell_{T,\phi}(\boldsymbol\beta), \tag{21}$$

where the first term is given by (9) and the second term is given by (16). We estimate $\boldsymbol\beta$ by

$$\check{\boldsymbol\beta} = \operatorname*{argmax}_{\boldsymbol\beta}\{\ell_{upd}(\boldsymbol\beta,\sigma^2)\}. \tag{22}$$

An iterative algorithm for (22) is obtained by the idea of traditional IRWLS. In particular, let $\check{\boldsymbol\beta}^{(r)}$ be the rth iterated value of $\check{\boldsymbol\beta}$. Then, the rth working response vector is $\check{\mathbf{u}}^{(r)} = (\check u_1^{(r)},\ldots,\check u_{n_T}^{(r)})^\top$ and the rth working weight matrix is $\check{\mathbf{W}}^{(r)} = \mathrm{diag}(\check w_1^{(r)},\ldots,\check w_{n_T}^{(r)})$, where $\check u_i^{(r)} = u(\check\eta_i^{(r)})$, $\check w_i^{(r)} = w(\check\eta_i^{(r)})$, and $\check\eta_i^{(r)} = \mathbf{x}_{Ti}^\top\check{\boldsymbol\beta}^{(r)}$. Let $\check\ell^{(r)}(\boldsymbol\beta,\sigma^2)$ be the loglikelihood function of (5) after $\mathbf{X}$ and $\mathbf{W}^{(r)}$ are replaced by $\mathbf{X}_T$ and $\check{\mathbf{W}}^{(r)}$, respectively. Then, $\check\ell^{(r)}(\boldsymbol\beta,\sigma^2)$ is the rth working loglikelihood function of $\ell_{T,\phi}(\boldsymbol\beta)$. The next iterated value of $\check{\boldsymbol\beta}$ is

$$\check{\boldsymbol\beta}^{(r+1)} = \operatorname*{argmax}_{\boldsymbol\beta}\left\{\sum_{t=1}^{T-1}\ell_t^{(0)}(\boldsymbol\beta,\sigma^2) + \check\ell^{(r)}(\boldsymbol\beta,\sigma^2)\right\}. \tag{23}$$

It provides the final $\check{\boldsymbol{\beta}}$ if the algorithm converges. We generally express $\check{\mathbf{u}} = (\check{u}_1, \ldots, \check{u}_{n_T})^\top$ as the working response vector and $\check{\mathbf{W}} = \mathrm{diag}(\check{w}_1, \ldots, \check{w}_{n_T})$ as the working weighted matrix in the computation of $\mathcal{S}_T^{(0)}$ in one-step IRWLS under UPD, where $\check{u}_i = u(\check{\eta}_i)$, $\check{w}_i = w(\check{\eta}_i)$, and $\check{\eta}_i = \mathbf{x}_{Ti}^\top \check{\boldsymbol{\beta}}$. The result given by (13) is denoted by $\check{\boldsymbol{\beta}}_{upd}$.

We also modify $\check{\boldsymbol{\beta}}_{upd}$. We use $\check{\boldsymbol{\beta}}^{(r)}$ for a selected r but not the final solution given by (23) to approximate $\ell_{T,\phi}(\boldsymbol{\beta})$. Because the result given by (13) depends on $\check{\boldsymbol{\beta}}^{(0)}$, we can propose many modifications. Here, we only introduce two. In the first, we choose $\check{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{con}$ when $T = 1$. We call this the simplified version. We denote the result given by (13) as $\check{\boldsymbol{\beta}}_{sim,t}$ if the method is applied until $\mathcal{D}_t$ or $\check{\boldsymbol{\beta}}_{sim}$ if it is applied to all $t \in \{1, \ldots, T\}$. Thus, we have $\check{\boldsymbol{\beta}}_{sim} = \check{\boldsymbol{\beta}}_{sim,T}$.

In the second, we choose $\check{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}_T$ in (23). If $T = 1$, then $\check{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_1$ because $\hat{\boldsymbol{\beta}}_1$, the MLE of $\boldsymbol{\beta}$ for data in $\mathcal{D}_1$, is a maximizer of the likelihood function; otherwise, we need (23). We denote the result given by (13) as $\check{\boldsymbol{\beta}}_{mle}^{(r)}$ if we treat the rth iterated value given by (23) as the final solution of $\check{\boldsymbol{\beta}}$. Then, $\check{\boldsymbol{\beta}}_{mle}^{(0)} = \hat{\boldsymbol{\beta}}_{aee}$ if g is the canonical link.

We compare our $\check{\boldsymbol{\beta}}_{mle}^{(r)}$ with the estimator given by the cumulatively updated estimating equation (CUEE) approach proposed by [12]. The CUEE estimator of $\boldsymbol{\beta}$ is obtained by the Taylor expansion of $-a(\phi)\dot{\ell}_{T,\phi}(\boldsymbol{\beta})$ at

$$\text{(24)}\quad \tilde{\boldsymbol{\beta}}_{n_T,T} = (\tilde{\mathbf{A}}_{T-1} + \mathbf{A}_T)^{-1} \left(\sum_{t=1}^{T-1} \tilde{\mathbf{A}}_{n_t,t} \tilde{\boldsymbol{\beta}}_{n_t,t} + \mathbf{A}_T \hat{\boldsymbol{\beta}}_T \right),$$

where $\tilde{\mathbf{A}}_0 = \mathbf{0}$, $\tilde{\boldsymbol{\beta}}_0 = \mathbf{0}$, $\tilde{\mathbf{A}}_t = \sum_{i=1}^t \tilde{\mathbf{A}}_{n_i,i}$, and $\tilde{\mathbf{A}}_{n_t,t} = -a(\phi)\ddot{\ell}_{t,\phi}(\tilde{\boldsymbol{\beta}}_{n_t,t})$. Note that (24) provides the current $\tilde{\boldsymbol{\beta}}_{n_T,T}$ based on previous $\tilde{\boldsymbol{\beta}}_{n_t,t}$ for all $t \in \{1, \ldots, T-1\}$. We assume that $\tilde{\boldsymbol{\beta}}_{n_t,T}$ given by (24) is implemented by increasing T from 1 to the current time in a data stream, such that we have all the previous $\tilde{\boldsymbol{\beta}}_{n_t,t}$. Then, (24) provides the current $\tilde{\boldsymbol{\beta}}_{n_T,T}$ value. The CUEE estimator of $\boldsymbol{\beta}$ is

$$\text{(25)}\quad \begin{aligned} \check{\boldsymbol{\beta}}_{cuee} = &\left\{ \sum_{t=1}^{T-1} \tilde{\mathbf{A}}_{n_t,t} + \tilde{\mathbf{A}}_{n_T,T} \right\}^{-1} [\mathbf{a}_{T-1} \\ &+ \tilde{\mathbf{A}}_{n_T,T} \tilde{\boldsymbol{\beta}}_{n_T,T} + \mathbf{b}_{T-1} + a(\phi)\dot{\ell}_{T,\phi}(\tilde{\boldsymbol{\beta}}_{n_T,T})], \end{aligned}$$

where $\mathbf{a}_T = \mathbf{a}_{T-1} + \tilde{\mathbf{A}}_{n_T,T}\tilde{\boldsymbol{\beta}}_{n_T,T}$, $\mathbf{b}_T = \mathbf{b}_{T-1} + a(\phi)\dot{\ell}_{T,\phi}(\tilde{\boldsymbol{\beta}}_{n_T,T})$, $\mathbf{a}_0 = \mathbf{b}_0 = \mathbf{0}_p$, and $\tilde{\mathbf{A}}_0 = \mathbf{0}_{p \times p}$.

Theorem 3.2. *If g is the canonical link, or g is not the canonical link but we use the modification given by Corollary 3.1 or 3.2, then $\check{\boldsymbol{\beta}}_{mle}^{(1)} = \check{\boldsymbol{\beta}}_{cuee}$.*

Proof: The conclusion can be similarly proven by the method in the proofs of Theorem 3.1, and Corollary 3.1 and 3.2. The detail is omitted. □

We summarize the five estimators obtained in Sections (3.6) and (3.2) in Table (1). Among those, two are developed under D&C and three are developed under UPD.

Table 1. Five estimators given by one-step IRWLS under divide-and-conquer (D&C) and online updating (UPD), where $u^{(c)}(y) = g(y)$, $w^{(c)}(y) = \{[g'(y)]^2 b''[h(g(y))]\}^{-1}$, $\hat{\boldsymbol{\beta}}_t$ is the MLE of $\boldsymbol{\beta}$ in $\mathcal{D}_t$, and $\check{\boldsymbol{\beta}}_t^{(r)}$ and $\check{\boldsymbol{\beta}}_t$ are the solutions of $\boldsymbol{\beta}$ given by (23) for a selected r and by (22), respectively, when UPD is applied from $\mathcal{D}_1$ to $\mathcal{D}_t$.

			$t=1$	$t>1$
D&C	$\hat{\boldsymbol{\beta}}_{con}$	$u_{ti}^{(0)}$	$u^{(c)}(y_{ti})$	$u^{(c)}(y_{ti})$
		$w_{ti}^{(0)}$	$w^{(c)}(y_{ti})$	$w^{(c)}(y_{ti})$
	$\hat{\boldsymbol{\beta}}_{aee}$	$u_{ti}^{(0)}$	$u(\mathbf{x}_{ti}^\top \hat{\boldsymbol{\beta}}_t)$	$u(\mathbf{x}_{ti}^\top \hat{\boldsymbol{\beta}}_t)$
		$w_{ti}^{(0)}$	$w(\mathbf{x}_{ti}^\top \hat{\boldsymbol{\beta}}_t)$	$w(\mathbf{x}_{ti}^\top \hat{\boldsymbol{\beta}}_t)$
UPD	$\check{\boldsymbol{\beta}}_{sim}$	$u_{ti}^{(0)}$	$u^{(c)}(y_{ti})$	$u(\mathbf{x}_{ti}^\top \check{\boldsymbol{\beta}}_{sim,t-1})$
		$w_{ti}^{(0)}$	$w^{(c)}(y_{ti})$	$w(\mathbf{x}_{ti}^\top \check{\boldsymbol{\beta}}_{sim,t-1})$
	$\check{\boldsymbol{\beta}}_{cuee}$	$u_{ti}^{(0)}$	$u(\mathbf{x}_{ti}^\top \hat{\boldsymbol{\beta}}_t)$	$u(\mathbf{x}_{ti}^\top \check{\boldsymbol{\beta}}_t^{(1)})$
		$w_{ti}^{(0)}$	$w(\mathbf{x}_{ti}^\top \hat{\boldsymbol{\beta}}_t)$	$w(\mathbf{x}_{ti}^\top \check{\boldsymbol{\beta}}_t^{(1)})$
	$\check{\boldsymbol{\beta}}_{upd}$	$u_{ti}^{(0)}$	$u(\mathbf{x}_{ti}^\top \hat{\boldsymbol{\beta}}_t)$	$u(\mathbf{x}_{ti}^\top \check{\boldsymbol{\beta}}_t)$
		$w_{ti}^{(0)}$	$w(\mathbf{x}_{ti}^\top \hat{\boldsymbol{\beta}}_t)$	$w(\mathbf{x}_{ti}^\top \check{\boldsymbol{\beta}}_t)$

Under D&C, the mathematical formulations for the working response and the working weight vectors in all $\mathcal{D}_t$ are identical, but not under UPD. As previous information is used, we expected that UPD can provide better estimators than D&C. This is confirmed in our simulation studies.

3.3 Online Parallel Computation

The major issue in one-step IRWLS is the computation of $\mathcal{S}_{T+}^{(0)}$ by (12). This can be easily implemented if $\mathcal{D}_T$ is not large in size. An online parallel computation framework is needed only when $\mathcal{D}_T$ is also large in size.

Suppose that $\mathcal{D}_T$ is large and has been partitioned into K subsets. Let $\mathcal{D}_{Tk} = \{(\mathbf{y}_{Tk}, \mathbf{X}_{Tk})\}$ be the kth subset, such that $\mathcal{D}_T$ can be obtained by $\mathbf{y}_T = (\mathbf{y}_{T1}^\top, \ldots, \mathbf{y}_{TK}^\top)^\top$ and $\mathbf{X}_T = (\mathbf{X}_{T1}^\top, \ldots, \mathbf{X}_{TK}^\top)^\top$. The online parallel computation calculates $\mathcal{S}_{Tk}^{(0)} = (s_{Tk}^{(0)}, \mathbf{s}_{Tk}^{(0)}, \mathbf{S}_{Tk}^{(0)})$ for all $k \in \{1, \ldots, K\}$ individually, where $s_{Tk}^{(0)}$, $\mathbf{s}_{Tk}^{(0)}$, and $\mathbf{S}_{Tk}^{(0)}$ are defined similarly as those given by (10). Then, it derives $\mathcal{S}_T^{(0)}$ by $\mathcal{S}_T^{(0)} = \sum_{k=1}^K \mathcal{S}_{Tk}^{(0)}$. Once $\mathcal{S}_T^{(0)}$ is available, the task of online parallel computation is over.

We have two scenarios to implement online parallel computation. In the first, we do not need any previous information in the computation of $\mathcal{S}_T^{(0)}$. It is used in the implementation of $\hat{\boldsymbol{\beta}}_{con}$, $\hat{\boldsymbol{\beta}}_{aee}$(i.e., $\hat{\boldsymbol{\beta}}_{mle}$) by the methods given by Section 3.1. The corresponding results are not affected by the choices of $\mathcal{D}_{Tk}$. In the second, we need to use the previous information. It occurs in the computation of $\check{\boldsymbol{\beta}}_{sim}$, $\check{\boldsymbol{\beta}}_{upd}$, and $\check{\boldsymbol{\beta}}_{cuee}$. We need to modify (21) and (22) for $\mathcal{S}_{Tk}$ for each $k \in \{1, \ldots, K\}$. The corresponding results are affected by the choices of $\mathcal{D}_{Tk}$. The method can be developed under the D&C framework.

The major difference between offline and online parallel computation is that offline parallel computation can use the entire data but online parallel computation cannot. Online parallel computation is not needed if the entire data set is large but the current data set is not. The previous data set cannot be accessed by online parallel computation. It can only be given by a set of summary information.

3.4 Asymptotics

All modifications of $\hat{\boldsymbol{\beta}}_{one}$ and $\hat{\sigma}^2_{one}$ given by (13) and (15) can be treated as M-estimators of $\boldsymbol{\beta}$ and σ^2, respectively. Thus, the asymptotic properties of the one-step IRWLS can be evaluated by the traditional M-estimation approach. We study this issue under a broader context. It includes all of the estimators discussed in Sections 3.1 and 3.2.

We focus on the asymptotic properties of the quadratic form given by the last term of (9), which can be generally expressed as

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n} w_i(u_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2, \tag{26}$$

where $w_i = w(y_i, \mathbf{x}_i)$ and $u_i = u(y_i, \mathbf{x}_i)$, and $w(\cdot,\cdot)$ and $u(\cdot,\cdot)$ are smooth functions. Let

$$\tilde{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}). \tag{27}$$

Here, $\tilde{\boldsymbol{\beta}}$ could be one of $\hat{\boldsymbol{\beta}}_{con}$, $\hat{\boldsymbol{\beta}}_{aee}$ (i.e., $\hat{\boldsymbol{\beta}}_{mle}$), $\check{\boldsymbol{\beta}}_{sim}$, $\check{\boldsymbol{\beta}}_{cuee}$ (i.e., $\check{\boldsymbol{\beta}}^{(1)}_{mle}$), or $\check{\boldsymbol{\beta}}_{upd}$ in our asymptotic studies, depending the choices of $w(\cdot,\cdot)$ and $u(\cdot,\cdot)$ in (26). Let

$$q(\boldsymbol{\beta}) = \mathrm{E}_{\boldsymbol{\beta}_0}[Q(\boldsymbol{\beta})] \tag{28}$$

and

$$\boldsymbol{\beta}^* = \operatorname*{argmin}_{\boldsymbol{\beta}} q(\boldsymbol{\beta}), \tag{29}$$

where $\boldsymbol{\beta}_0 = (1, \beta_{01}, \ldots, \beta_{0(p-1)})^\top$ is the true parameter vector. By the traditional M-estimation approach, we can show that $\tilde{\boldsymbol{\beta}} \overset{P}{\to} \boldsymbol{\beta}^*$ as $n \to \infty$. Moreover, we can also derive the asymptotic normality. Since $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}_0$ may be different, we need to evaluate their relationship.

Based on traditional regularity conditions for consistency and asymptotic normality of the maximum likelihood and M-estimation approaches, we propose our regularity conditions as follows.

Regularity conditions:

(C1) $(y_i, \mathbf{x}_i^\top)^\top$ are iid copies of $(y, \mathbf{x}^\top)^\top$ for all $i \in \{1, \ldots, n\}$, and the distribution of $(\mathbf{y}, \mathbf{x}^\top)^\top$ does not vary with n.
(C2) $w(\cdot,\cdot)$ is always positive.
(C3) The domain of $\boldsymbol{\beta}$ in (26) is compact, and $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}^*$ are the interior points.
(C4) $w(\cdot,\cdot)$ and $u(\cdot,\cdot)$ are third-order continuous, and the third-order partial derivative operators can be passed under the integral sign in the expected value operators.
(C5) There exists a continuous function $\psi(y, \mathbf{x})$ such that $\mathrm{E}_{\boldsymbol{\beta}}[\psi(y,\mathbf{x})] < \infty$ and $w(y,\mathbf{x})[u(y,\mathbf{x}) - \mathbf{x}^\top\boldsymbol{\beta}]^2 < \psi(y,\mathbf{x})$ for all y, $\mathbf{x}$, and $\boldsymbol{\beta}$.
(C6) For any $\boldsymbol{\beta}$ and sufficiently small $\rho > 0$, $\sup_{|\boldsymbol{\beta}' - \boldsymbol{\beta}| < \rho} w(y,\mathbf{x})[u(y,\mathbf{x}) - \mathbf{x}^\top\boldsymbol{\beta}]^2$ is measurable in $\mathbf{y}$ and $\mathbf{x}$.
(C7) $\sup_{\boldsymbol{\beta}} |Q(\boldsymbol{\beta}) - q(\boldsymbol{\beta})| \overset{a.s}{\to} 0$.
(C8) $w(y,\mathbf{x})[u(y,\mathbf{x}) - \mathbf{x}^\top\boldsymbol{\beta}]^2 = w(y,\mathbf{x})[u(y,\mathbf{x}) - \mathbf{x}^\top\boldsymbol{\beta}']^2$ for all y and $\mathbf{x}$ if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}'$.
(C9) There exist continuous functions $\psi_1(y,\mathbf{x})$ and $\psi_2(y,\mathbf{x})$ such that $\mathrm{E}_{\boldsymbol{\beta}_0}[\psi_1(y,\mathbf{x})] < \infty$, $\mathrm{E}_{\boldsymbol{\beta}_0}[\psi_2(y,\mathbf{x})] < \infty$, each component of the gradient vector of $w(y,\mathbf{x})[u(y,\mathbf{x}) - \mathbf{x}^\top\boldsymbol{\beta}]^2$ is bounded in absolute value by $\psi_1(y,\mathbf{x})$, and each component of the Hessian matrix of $w(y,\mathbf{x})[u(y,\mathbf{x}) - \mathbf{x}^\top\boldsymbol{\beta}]^2$ is bounded in absolute value by $\psi_2(y,\mathbf{x})$.
(C10) Neither $\mathbf{B}_{\boldsymbol{\beta}_0} = \mathrm{E}_{\boldsymbol{\beta}_0}[w(y,\mathbf{x})\mathbf{x}\mathbf{x}^\top]$ nor $\mathbf{A}_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}) = \mathrm{E}_{\boldsymbol{\beta}_0}\{w^2(y,\mathbf{x})[u(y,\mathbf{x}) - \mathbf{x}^\top\boldsymbol{\beta}]^2\mathbf{x}\mathbf{x}^\top\}$ is singular.

Condition (C1) indicates that the asymptotic properties are studied under the iid case, such that we can directly use the traditional proofs for consistency of M-estimation. Condition (C2) is required as $Q(\boldsymbol{\beta})$ must be the objective function given by the WLS approach. Conditions (C3)–(C8) are modified from the standard conditions for consistency of maximum likelihood [4, Chapter 16] and M-estimation [13, Chapter 5]. Condition (C9) is proposed for the asymptotic normality, which has been previously used for asymptotic normality of maximum likelihood [4, Chapter 17]. Condition (C10), which has been previously used by [21], ensures that the variance-covariance matrix in the asymptotic normality is not singular.

Lemma 3.1. *If (C1)–(C8) hold, then* $\tilde{\boldsymbol{\beta}} \overset{a.s.}{\to} \boldsymbol{\beta}^*$.

Proof. We draw the conclusion by the standard proof for consistency of M-estimation; see the proof of Theorem 5.7 in [13] or Theorem 17 in [4] for details. □

Lemma 3.2. *If (C1)–(C10) hold, then* $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \rightsquigarrow \mathcal{N}[\mathbf{0}, \mathbf{B}^{-1}_{\boldsymbol{\beta}_0}\mathbf{A}_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}^*)\mathbf{B}^{-1}_{\boldsymbol{\beta}_0}]$, *where* $\rightsquigarrow$ *means converges in distribution.*

Proof. The asymptotic normality can be shown by the standard methods for the asymptotic normality of M-estimation; see the proof of Theorem 5.21 in [13]. The variance-covariance matrix in the asymptotic normality can be directly derived by the standard formulation. □

Lemma 3.3. *Let* $\boldsymbol{\beta}^*_J$ *and* $\mathbf{x}_J$ *be sub-vectors of* $\boldsymbol{\beta}^*$ *and* $\mathbf{x}$ *chosen from lags in* $\boldsymbol{\beta}^*$ *and* $\mathbf{x}$ *indexed by* J *for any* $J \subseteq \{1, \ldots, p-1\}$, *respectively. Suppose that (C1)–(C8) hold. Then,* $\boldsymbol{\beta}^*_J = \mathbf{0}$ *if and only if there exists a function* $C(\mathbf{x})$, *which does not depend on* $\mathbf{x}_J$, *such that* $\mathrm{E}_{\boldsymbol{\beta}_0}[w^{1/2}(y,\mathbf{x})u(y,\mathbf{x})|\mathbf{x}_J] = C(\mathbf{x})\mathrm{E}_{\boldsymbol{\beta}_0}[w^{1/2}(y,\mathbf{x})\mathbf{x}^\top\boldsymbol{\beta}|\mathbf{x}_J]$.

Proof. Based on the properties of $\mathrm{E}[Q(\boldsymbol{\beta}|\mathbf{x}_J]$, we find that $\mathrm{E}_{\boldsymbol{\beta}_0}\{w(y,\mathbf{x})[u(y,\mathbf{x}) - \mathbf{x}^\top\boldsymbol{\beta}]^2|\mathbf{x}_J\}$ and $\mathrm{E}_{\boldsymbol{\beta}_0}\{\mathrm{E}_{\boldsymbol{\beta}_0}[w^{1/2}(y,\mathbf{x})u(y,\mathbf{x})|\mathbf{x}_J] - \mathrm{E}_{\boldsymbol{\beta}_0}[w^{1/2}(y,\mathbf{x})\mathbf{x}^\top\boldsymbol{\beta}|\mathbf{x}_J]\}^2$

are minimized at the same function of $\mathbf{x}_J$, implying the conclusion. □

Lemma 3.4. *Assume that (C1)–(C8) are satisfied. Let $\boldsymbol{\beta}_{0J}$ with the first component for the intercept and the rest components be the vector composed by those in $\boldsymbol{\beta}_0$ with lags indexed by $J \subseteq \{1, \ldots, p-1\}$. Then, (a) $\boldsymbol{\beta}_J^* = \mathbf{0}$ if $\boldsymbol{\beta}_{0J} = \mathbf{0}$, and (b) $\boldsymbol{\beta}_J^* \neq \mathbf{0}$ if $\mathrm{E}_{\boldsymbol{\beta}_0}[w^{1/2}(y,\mathbf{x})u(y,\mathbf{x})|\mathbf{x}_J]/\mathrm{E}_{\boldsymbol{\beta}_0}[w^{1/2}(y,\mathbf{x})\mathbf{x}^\top\boldsymbol{\beta}|\mathbf{x}_J]$ always varies with $\mathbf{x}_J$ when $\boldsymbol{\beta}_{0J} \neq \mathbf{0}$.*

Proof. The conclusion can be implied by Lemma 3.3. □

Theorem 3.3. *Let $w(y,\mathbf{x})$ and $u(y,\mathbf{x})$ be those given by Definition 3.1, respectively. Suppose that (C1)–(C10) hold. Assume that $\mathrm{E}_{\boldsymbol{\beta}_0}[w^{1/2}(y,\mathbf{x})u(y,\mathbf{x})|\mathbf{x}_J]/\mathrm{E}_{\boldsymbol{\beta}_0}[w^{1/2}(y,\mathbf{x})\mathbf{x}^\top\boldsymbol{\beta}|\mathbf{x}_J]$ always varies with $\mathbf{x}_J$ for any $J \subseteq \{1, \ldots, p-1\}$ when $\boldsymbol{\beta}_{0J} \neq 0$. If $\beta_{0j} = 0$, then $z_j \rightsquigarrow \mathcal{N}(0, \tau_j^2)$, where τ_j^2 is a positive constant which is determined by Lemma 3.2 and (14). If $\beta_{0j} \neq 0$, then $\lim_{n\to\infty} P(|z_{0j}| < C) = 0$ for any $C \in \mathbb{R}^+$.*

Proof: By Lemma 3.2, we conclude that $\sqrt{n}(\hat{\beta}_{one,j} - \beta_j^*) \rightsquigarrow \mathcal{N}(0, \tau_j^2)$, where β_j^* is the jth component of $\boldsymbol{\beta}^*$, and τ_j^2 is the jth diagonal element of $\mathbf{B}_{\boldsymbol{\beta}_0}^{-1}\mathbf{A}_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}^*)\mathbf{B}_{\boldsymbol{\beta}_0}^{-1}$ divided by the jth diagonal element of $n\hat{\mathrm{V}}(\hat{\boldsymbol{\beta}}_{one})$. By the Strong Law of Large Number (SLLN), we conclude $n\hat{\mathrm{V}}(\hat{\boldsymbol{\beta}}_{one}) \overset{a.s.}{\to} n^{-1}\mathrm{E}_{\boldsymbol{\beta}_0}(\mathbf{S}_{T+}^{(0)}) = \mathrm{E}_{\boldsymbol{\beta}_0}[w^{1/2}(y,\mathbf{x})\mathbf{x}\mathbf{x}^\top]$. Then $\beta_j^* = 0$ if and only if $\beta_{j0} = 0$ by Lemma 3.4, and the conclusion is drawn. □

Corollary 3.3. *Assume that all conditions of Theorem 3.3 hold. If $w(y,\mathbf{x})[u(y,\mathbf{x}) - \mathbf{x}^\top\boldsymbol{\beta}^*]^2$ and $w(y,\mathbf{x})\mathbf{x}\mathbf{x}^\top$ are uncorrelated, then $\tau_j^2 = 1$.*

Proof: Note that $\hat{\sigma}_{one}^2 \overset{P}{\to} \mathrm{E}_{\boldsymbol{\beta}_0}\{w(y,\mathbf{x})[u(y,\mathbf{x}) - \mathbf{x}^\top\boldsymbol{\beta}^*]^2\}$ and $\mathbf{A}_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}^*) = \mathbf{B}_{\boldsymbol{\beta}_0}\mathrm{E}_{\boldsymbol{\beta}_0}\{w(y,\mathbf{x})[u(y,\mathbf{x}) - \mathbf{x}^\top\boldsymbol{\beta}]^2\}$ under the condition of the corollary. Then, we draw the conclusion by combining this with (14). □

Theorem 3.3 points out that the Wald test based on z_j given by (13) and (14) is consistent. Corollary 3.3 points out that it is appropriate to use the standard normal distribution to calculate the p-value of z_j under the condition. If the condition is violated but the correlation between $w(y,\mathbf{x})[u(y,\mathbf{x}) - \mathbf{x}^\top\boldsymbol{\beta}^*]^2$ and $w(y,\mathbf{x})\mathbf{x}\mathbf{x}^\top$ is small, then we can still approximately use the method to calculate the p-value of z_j. Because $\boldsymbol{\beta}^*$ may be different from $\boldsymbol{\beta}_0$, the estimators given by one-step IRWLS may be biased, but it can still provide significance in the model. This is used in finding the optimal model, indicating that main properties of GLMs can be identified.

3.5 Multiple Learning

Multiple learning has more advantages than individual learning since it can provide results of a number of statistical models simultaneously. As the optimal model is unknown, a number of candidate models must be investigated. The usage of an individual learning procedure is inefficient as it can only provide results of a single statistical model by one access of data. If another model is considered, then the entire procedure must be used again, leading to another access of the data. The framework of individual learning procedures has been extensively adopted by many statistical software packages. Examples include all of packages in R and SAS. To fit a candidate model, the data command in R and SAS must be included to indicate the name of the data set to be accessed. This is not a concern if the size of data is small, but it is a serious problem in big data as the access of the entire data set is time-consuming.

A multiple learning procedure can be straightforwardly proposed by (21). Note that the second term is the true loglikelihood function for the current data. It can vary with respect to our interest. For the first term, when a number of GLMs are involved, we can purposely construct $\mathbf{x}_i$ in (11) such that it includes all of the explanatory variables. Then, the combination of the two terms can be used to study all of the related models. After $\mathcal{S}_{T+}^{(0)}$ is derived, we can remove the redundant elements and obtain the corresponding WSS for each candidate model. It can provide the solutions to the model without an access of the data. A similar idea can be found in criterion-based variable selection procedures for regression models [14]. Our method can extend this idea to GLMs. Based on WSS, we can test significance of parameters and identify the optimal model by traditional multiple testing methods (e.g., by the Bonferroni method). We can also carry out a backward or forward selection procedure.

Since only summations are used in the computation of $\mathcal{S}_{T+}^{(0)}$, the multiple learning procedure can be applied even if rank deficiency is present. We may face a situation that rank deficiency is present in $\mathcal{S}_t^{(0)}$ given by (10) occasionally for some t but never in $\mathcal{S}_{T+}^{(0)}$ given by (11). Thus, we do not need to account for rank deficiency in our approach. On the other hand, even if the rank deficiency problem is always present, it can be overcome by using the generalized inverse, such as the Moore-Penrose pseudoinverse, in matrix computation. Since we only need to provide predicted values for the linear components, the rank deficiency problem does not affect the computation of the WSS.

3.6 Offline Extension

We can extend our approach to the offline setting. We assume that the observed data set is massive, such that we can only load portion of the data to memory. Rather than an online approach, the offline approach can access the entire data multiple times by trunks, leading to the possibility for the exact MLEs.

Following traditional IRWLS, we treat (5) as the WWLM given by the rth iteration, and derive a working loglikelihood function for normal data. Then, we obtain an unstructured array of WSS for the rth iteration. It is used to calculate the next iterated value of the MLE. The memory needed in the entire computation does not depend on n, indicating that the offline extension can be implemented even if the size of the data is extremely large.

Suppose that the entire data set $\mathcal{D} = \{(\mathbf{y}, \mathbf{X})\}$ has been partitioned into K subsets $\mathcal{D}_k = \{(\mathbf{y}_k, \mathbf{X}_k)\}$ for all $k \in \{1, \ldots, K\}$, such that we have $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_K^\top)^\top$ and $\mathbf{X} = (\mathbf{X}_1^\top, \ldots, \mathbf{X}_K^\top)^\top$ for the entire data. Let $\mathbf{u}_k^{(r)}$ and $\mathbf{W}_k^{(r)}$ be the working response vector and the working weight matrix given by $\mathcal{D}_k$. If $\boldsymbol{\beta}^{(r)}$ is available, then $\mathbf{u}_k^{(r)}$ and $\mathbf{W}_k^{(r)}$ can be derived by $\mathcal{D}_k$ only. Let $\mathcal{S}_k^{(r)} = (s_k^{(r)}, \mathbf{s}_k^{(r)}, \mathbf{S}_k^{(r)})$ and $\mathcal{S}^{(r)} = \sum_{k=1}^K \mathcal{S}_k^{(r)}$, where $s_k = \{\mathbf{u}^{(r)}\}^\top \mathbf{W}^{(r)} \mathbf{u}^{(r)}$, $\mathbf{s}_k^{(r)} = \mathbf{X}_k^\top \mathbf{W}^{(r)} \mathbf{u}^{(r)}$, and $\mathbf{S}_k^{(r)} = \mathbf{X}_k^\top \mathbf{W}^{(r)} \mathbf{X}_k$. Then, $\mathcal{S}_k^{(r)}$ and $\mathcal{S}$ are the sets of WSS in $\mathcal{D}_k$ and $\mathcal{D}$ given by the rth iteration, respectively. They are used to compute $\boldsymbol{\beta}^{(r+1)}$, the next iterated value of $\hat{\boldsymbol{\beta}}$, by a method similar to that given by (13), implying that the entire IRWLS can be applied. Meanwhile, we also compute the Fisher information of $\boldsymbol{\beta}^{(r+1)}$ by a method similar to (14). The offline extension can provide exact solutions of $\hat{\boldsymbol{\beta}}$ as well as its variance-covariance matrix. Since it needs to access the entire data multiple times, the method cannot be applied to the online setting.

4. SPECIFICATION

We specify our approach to three typical GLMs. The first is linear models for normal data, where the unstructured array of WSS given by (11) becomes an unstructured array of SS, leading to an exact online approach to big data regression. The second is binomial regression. It includes the logistic, probit, and cloglog models for binomial or Bernoulli data. We focus on the logistic model because of its popularity. The third is Poisson regression. It is carried out by loglinear models for Poisson data. Because the dispersion parameter ϕ is not present in binomial and Poisson models, we propose an adjustment of our one-step IRLWS, such that we always have $\hat{\sigma}^2_{one} = 1$ in our approach. This is consistent with the assumptions of binomial and Poisson models.

4.1 Regression

We specify our approach to the WLM given by

$$y_{ti} = \mathbf{x}_{ti}^\top \boldsymbol{\beta} + \epsilon_{ti}, \tag{30}$$

for all $t \in \{1, \ldots, T\}$ and $i \in \{1, \ldots, n_t\}$, where $\epsilon_{ti} \sim \mathcal{N}(0, \sigma^2/w_{ti})$ independently, and w_{ti} is known. It is derived by assuming that $f(y_{ti})$ given by (1) is a normal density with known variance ratios. Without the usage of any transformation, we choose $u_{ti}^{(0)} = y_{ti}$ and $w_{ti}^{(0)} = w_{ti}$. The WWLM in our approach is identical to the WLM given by (30).

The unstructured array of SS in $\mathcal{D}_t$ is $\mathcal{S}_t = (s_t, \mathbf{s}_t, \mathbf{S}_T)$, where $s_t = \sum_{i=1}^{n_t} w_{ti} y_{ti}^2$, $\mathbf{s}_t = \sum_{i=1}^{n_t} w_{ti} y_{ti} \mathbf{x}_{ti}$, and $\mathbf{S}_t = \sum_{i=1}^{n_t} w_{ti} \mathbf{x}_{ti} \mathbf{x}_{ti}^\top$. The unstructured array of SS in $\mathcal{D}_{T+}$ is $\mathcal{S}_{T+} = \sum_{t=1}^T \mathcal{S}_t$. The exact value of $\hat{\boldsymbol{\beta}}$ and its variance-covarinace matrix can be derived by (13), (14), and (15), respectively.

For asymptotic properties, we assume that w_{ti} are constants, $\mathbf{x}_{ti}$ are identically and independently derived, and the conditional distribution of y_{ti} given $\mathbf{x}_{ti}$ is given by (30). Then, we have $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ in (29). Condition (C10) becomes $\mathbf{A}_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}) = \sum_{t=1}^T \sum_{i=1}^{n_t} \mathrm{E}_{\boldsymbol{\beta}_0}[w_{ti}^2 (y_{ti} - \mathbf{x}_{ti}^\top \boldsymbol{\beta})^2 \mathbf{x}_{ti} \mathbf{x}_{ti}^\top]/n$ and $\mathbf{B}_{\boldsymbol{\beta}_0} = \sum_{t=1}^T \sum_{i=1}^{n_t} \mathrm{E}_{\boldsymbol{\beta}_0}(w_{ti} \mathbf{x}_{ti} \mathbf{x}_{ti}^\top)/n$, respectively. Because $\mathrm{E}_{\boldsymbol{\beta}_0}(w_{ti}^{1/2} y_{ti} | \mathbf{x}_{ti}) = \mathrm{E}_{\boldsymbol{\beta}_0}(w^{1/2} \mathbf{x}_{ti}^\top \boldsymbol{\beta} | \mathbf{x}_{ti}) = w_{ti} \mathbf{x}_{ti}^\top \boldsymbol{\beta}$, the conclusion of Lemma 3.3 holds. Because residuals and estimators are independent, Conditions of Theorem 3.3 and Corollary 3.3 hold. Theorem 3.3 and Corollary 3.3 are eqilvalent to the properties of the Wald statistic. This is because z_j follows the t-distribution, which converges to the standard normal distribution as $n \to \infty$.

We may also use a location-scale transformation on y_{ti} to define the working response, leading to $u_{ti}^{(0)} = (y_{ti} - \delta_1)/\delta_2$ and $w_{ti}^{(0)} = w_{ti}$ for some $\delta_1 \in \mathbb{R}$ and $\delta_2 \in \mathbb{R}^+$. If this is adopted, then $\boldsymbol{\beta}^*$ given by (29) is not identical to $\boldsymbol{\beta}_0$. This is an important issue to be addressed when our approach is applied to binomial and Poisson data.

4.2 Binomial

The logistic linear model for binomial data assumes that $y_{ti} \sim \mathcal{B}(m_{ti}, \pi_{ti})$ independently with

$$\log \frac{\pi_{ti}}{1 - \pi_{ti}} = \mathbf{x}_{ti}^\top \boldsymbol{\beta} \tag{31}$$

for all $t \in \{1, \ldots, T\}$ and $i \in \{1, \ldots, n_t\}$, where $m_{ti} \in \mathbb{N}$ and $\pi_{ti} \in (0, 1)$. By traditional IRWLS, we obtain $u_{ti}^{(c)} = \log[(y_{ti}+0.5)/(m_{ti}-y_{yi}+1)]$ and $w_{ti}^{(c)} = m_{ti}(y_{ti}+0.5)(m_{ti}-y_{yi}+0.5)/(m_{ti}+1)^2$. The WWLM is derived by putting these into (7), leading to the conventional version of one-step IRWLS. Since $\sigma^2 = a(\phi) = 1$, we propose an adjustment of our approach by $\hat{\boldsymbol{\beta}}_{\hat{\sigma}} = \hat{\boldsymbol{\beta}}_{con}/\hat{\sigma}_{con}$ with the variance-covariance matrix given by $\hat{\mathrm{V}}(\hat{\boldsymbol{\beta}}_{\hat{\sigma}}) = \{\mathbf{S}_{T+}^{(0)}\}^{-1}$.

In the MLE version of one-step IRWLS, we choose $w_{ti}^{(0)} = m_{ti} \hat{\pi}_{ti}(1 - \hat{\pi}_{ti})$ and $u_{ti}^{(0)} = \hat{\eta}_{ti} + (y_{ti} - m_{ti}\hat{\pi}_{ti})/w_{ti}^{(0)}$ with $\pi_{ti}^{(0)} = e^{\hat{\eta}_{ti}}/(1 + e^{\hat{\eta}_{ti}})$ and $\eta_{ti} = \mathbf{x}_{ti}^\top \hat{\boldsymbol{\beta}}_t$ for all $t \in \{1, \ldots, T\}$ and $i \in \{1, \ldots, n_t\}$. We still compute the working loglikelihood function $\ell^{(0)}(\boldsymbol{\beta}, \sigma^2)$ by (8). Then, we calculate $\hat{\boldsymbol{\beta}}_{mle}$, which is also $\hat{\boldsymbol{\beta}}_{aee}$, by maximizing $\ell^{(0)}(\boldsymbol{\beta}, \sigma^2)$. To compute $\ell_{upd}(\boldsymbol{\beta}, \sigma^2)$, we replace the second term on the right-hand side of (21) by

$$\ell_T(\boldsymbol{\beta}) = \sum_{i=1}^{n_T} [y_{Ti} \mathbf{x}_{Ti}^\top \boldsymbol{\beta} - m_{Ti} \log(1 + e^{\mathbf{x}_{Ti}^\top \boldsymbol{\beta}})] + C, \tag{32}$$

where $C = \sum_{i=1}^{n_T} \log \binom{m_{Ti}}{y_{Ti}}$. The second term on the right-hand side of (23) is

$$\check{\ell}^{(r)}(\boldsymbol{\beta}) = -\frac{1}{2}(\check{\mathbf{u}}^{(r)} - \mathbf{X}_T \boldsymbol{\beta})^\top \check{\mathbf{W}}^{(r)} (\check{\mathbf{u}}^{(r)} - \mathbf{X}_T \boldsymbol{\beta}) + C, \tag{33}$$

where $C = -(1/2)[n_T \log(2\pi) + \sum_{i=1}^{n_T} \log \check{w}_i]$. Obvious, we have $\check{\boldsymbol{\beta}}_{mle}^{(1)} = \check{\boldsymbol{\beta}}_{cuee}$.

4.3 Poisson

In the loglinear model for Poisson data, we assume that $y_{ti} \sim \mathcal{P}(\mu_{ti})$ independently with

$$(34) \qquad \log \mu_{ti} = \mathbf{x}_{ti}^\top \boldsymbol{\beta}$$

for all $t \in \{1, \ldots, T\}$ and $i \in \{1, \ldots, n_t\}$. We follow traditional IRWLS and choose $u_{ti}^{(c)} = \log(y_{ti} + 0.5)$ and $w_{ti}^{(c)} = y_{ti} + 0.5$. It provides the conventional version of one-step IRWLS. Note that $\sigma^2 = a(\phi) = 1$ holds. We can also define $\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$. In the MLE version of one-step IRWLS, we choose $u_{ti}^{(0)} = w_{ti}^{(0)} = e^{\mathbf{x}_{ti}^\top \hat{\boldsymbol{\beta}}_t}$. By maximizing $\ell^{(0)}(\boldsymbol{\beta}, \sigma^2)$ given by (8), we obtain $\hat{\boldsymbol{\beta}}_{mle}$, which is also $\hat{\boldsymbol{\beta}}_{aee}$. The second term in $\ell_{upd}(\boldsymbol{\beta}, \sigma^2)$ given by (21) is

$$(35) \qquad \ell_T(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{n_T} \log(y_{Ti}!) + \sum_{i=1}^{n_T} (y_{Ti} \mathbf{x}_{ti}^\top \boldsymbol{\beta} - e^{\mathbf{x}_{Ti}^\top \boldsymbol{\beta}}).$$

We can apply (23). Further, we also have $\check{\boldsymbol{\beta}}_{mle}^{(1)} = \check{\boldsymbol{\beta}}_{cuee}$.

5. SIMULATION

We evaluated the advantage of our approach via simulations. All of the simulations were carried out by a third generation Intel core-i7 2.8GHz processor with 16GB DDR 3 memory. We evaluated the performance of our approach based on a single processor. We studied the binomial and Poisson models. We found that conclusions for the Poisson model were similar to those for the binomial model. Therefore, we decided to only display our results for the binomial model.

5.1 Comparison with Traditional IRWLS

We compared our approach with the traditional IRWLS approach in the logistic linear model for binomial data. To ensure that the traditional IRWLS approach could be applied, we assumed that the size of the data was not large, such that the entire data set could always be loaded to memory of a personal computer. We compared the precision of seven estimators. They are $\hat{\boldsymbol{\beta}}_{con}$, $\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$, $\hat{\boldsymbol{\beta}}_{aee}$, $\check{\boldsymbol{\beta}}_{sim}$, $\check{\boldsymbol{\beta}}_{cuee}$, $\check{\boldsymbol{\beta}}_{upd}$, and $\hat{\boldsymbol{\beta}}$. The computation of $\hat{\boldsymbol{\beta}}$ was carried out by glm in R.

We assumed that the entire data set contained $p - 1 = 20$ explanatory variables, such that we had $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{20})^\top$ and $\mathbf{x}_{ti} = (1, x_{ti1}, \ldots, x_{ti20})^\top$ in (31). We chose $m_{ti} = 1$ for all $t \in \{1, \ldots, T\}$ and $i \in \{1, \ldots, n_t\}$, such that the binomial model became a Bernoulli model. We assumed that n_t were all equal to n_0, such that the total number of observations was $n = Tn_0$. We generated all of the explanatory variables independently from $\mathcal{N}(0, 1)$. We only set the coefficients for the first and second explanatory variables nonzero, such that we had $\beta_{j,0} \neq 0$ if and only if $j = 1, 2$, where $\boldsymbol{\beta}_0 = (\beta_{0,0}, \beta_{1,0}, \ldots, \beta_{20,0})^\top$ was the true parameter vector. The true model was

$$(36) \qquad \log \frac{\pi_{ti}}{1 - \pi_{ti}} = x_{ti1}\beta_{1,0} + x_{ti2}\beta_{2,0}.$$

We evaluated the performance of the seven estimators based on 1000 simulation replications from (36). In each replication, we first generated $\beta_{1,0}$ and $\beta_{2,0}$ independently from $\mathcal{U}(a, 2a)$ and then generated y_{ti} independently from $\mathcal{B}(1, \pi_{ti})$ with π_{ti} given by (36). We calculated the mean square errors (MSEs) of the seven estimators in each replication. The MSE was defined by $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ with $\tilde{\boldsymbol{\beta}}$ to be one of the seven estimators. We compared the averages of these MSE values based on these replications (Table 2). Our results indicated that the performance of $\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$ was better than that of $\hat{\boldsymbol{\beta}}_{con}$. We investigated the reason and found that the usage of a location-scale transformation on $\mathbf{u}^{(0)}$ in (7) could significantly affect the MSE values. Note that the dispersion parameter was not present. This problem could be partially overcome by the adjustment of $\hat{\sigma}$, which was given by $\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$. The MSE values of $\hat{\boldsymbol{\beta}}_{aee}$ were lower than those given by $\hat{\boldsymbol{\beta}}_{con}$ and $\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$, indicating that the usage of individual MLE of trunks improved the precision. The MSE values of $\check{\boldsymbol{\beta}}_{sim}$, $\check{\boldsymbol{\beta}}_{cuee}$, $\check{\boldsymbol{\beta}}_{upd}$, and $\hat{\boldsymbol{\beta}}$ were all close, indicating that the results given by online updating could be as precise as those given by the exact MLE. The difference between implementation methods of online updating can be ignored.

We next studied the multiple learning problems that we have presented in Section 3.5. We treated the model with all of the explanatory variables as a full model, and studied a number of reduced models. A reduced models was derived by removing a few explanatory variables from the full model. We investigated significance of explanatory variables in all of the reduced models. We kept all of the significant variables and removed all of the insignificant variables. As multiple explanatory variables were involved, we adjusted the multiple testing problems with the 0.01 significant level by the Bonferroni method. The final model was derived by removing all of the insignificant variables. We compared the final models with the true models given by the seven estimators by calculating the percentage of correctly identified models (Table 3), where we classified the final model as a correct model if it only contained nonzero coefficients of the true model. We found that the final models were almost identical to the true model when the strength of parameters became large. The difference between the seven methods could be ignored.

In summary, we find that parameter estimates given by one-step IRWLS can be significantly different from the true parameter, because they may be affected by the usage of the location-scale transformation on the working response values. This can be overcome by incorporating an estimate of the dispersion parameter in the WWLM. Significance of explanatory variables can be correctly identified by any of $\hat{\boldsymbol{\beta}}_{con}$, $\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$, $\hat{\boldsymbol{\beta}}_{aee}$, $\check{\boldsymbol{\beta}}_{sim}$, $\check{\boldsymbol{\beta}}_{cuee}$, and $\check{\boldsymbol{\beta}}_{upd}$. The usage of previous data information in one-step IRWLS significantly enhances precision of the results, which may be as precise as those given by the exactly MLE given by traditional IRWLS based on the entire data.

Table 2. Root MSE and absolute bias of $\hat{\boldsymbol{\beta}}_{con}$, $\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$, $\hat{\boldsymbol{\beta}}_{aee}$, $\check{\boldsymbol{\beta}}_{sim}$,$\check{\boldsymbol{\beta}}_{cuee}$, $\check{\boldsymbol{\beta}}_{upd}$, and $\hat{\boldsymbol{\beta}}$ for selected $n_0(\times 10^2)$, $T(\times 10)$, and a based on 1000 replications from (36).

				MSE			
(n_0, T, a)	$\hat{\boldsymbol{\beta}}_{con}$	$\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$	$\hat{\boldsymbol{\beta}}_{aee}$	$\check{\boldsymbol{\beta}}_{sim}$	$\check{\boldsymbol{\beta}}_{cuee}$	$\check{\boldsymbol{\beta}}_{upd}$	$\hat{\boldsymbol{\beta}}$
(1, 1, 0.2)	0.258	0.350	0.262	0.296	0.300	0.300	0.309
(1, 1, 0.4)	0.484	0.350	0.317	0.308	0.317	0.317	0.328
(1, 2, 0.2)	0.235	0.246	0.195	0.209	0.211	0.210	0.214
(1, 2, 0.4)	0.471	0.247	0.266	0.220	0.223	0.223	0.227
(1, 5, 0.2)	0.217	0.163	0.143	0.134	0.134	0.134	0.135
(1, 5, 0.4)	0.464	0.163	0.232	0.141	0.142	0.142	0.143
(1, 10, 0.2)	0.211	0.119	0.121	0.093	0.093	0.093	0.094
(1, 10, 0.4)	0.465	0.123	0.222	0.100	0.100	0.100	0.101
(2, 1, 0.2)	0.232	0.246	0.196	0.206	0.210	0.210	0.213
(2, 1, 0.4)	0.469	0.247	0.219	0.220	0.222	0.222	0.226
(2, 2, 0.2)	0.219	0.176	0.140	0.146	0.147	0.147	0.148
(2, 2, 0.4)	0.464	0.179	0.169	0.157	0.157	0.157	0.159
(2, 5, 0.2)	0.212	0.120	0.095	0.093	0.093	0.093	0.094
(2, 5, 0.4)	0.460	0.121	0.131	0.099	0.098	0.098	0.099
(2, 10, 0.2)	0.208	0.093	0.075	0.066	0.066	0.066	0.066
(2, 10, 0.4)	0.460	0.095	0.117	0.070	0.070	0.070	0.070
(5, 1, 0.2)	0.217	0.160	0.128	0.129	0.131	0.131	0.133
(5, 1, 0.4)	0.464	0.163	0.140	0.147	0.141	0.141	0.142
(5, 2, 0.2)	0.211	0.120	0.092	0.093	0.093	0.093	0.094
(5, 2, 0.4)	0.461	0.122	0.103	0.103	0.100	0.100	0.100
(5, 5, 0.2)	0.209	0.088	0.060	0.059	0.060	0.060	0.060
(5, 5, 0.4)	0.460	0.089	0.072	0.064	0.063	0.063	0.063
(5, 10, 0.2)	0.207	0.074	0.044	0.042	0.042	0.042	0.042
(5, 10, 0.4)	0.458	0.076	0.058	0.045	0.044	0.044	0.044
(10, 1, 0.2)	0.211	0.121	0.093	0.094	0.094	0.094	0.095
(10, 1, 0.4)	0.459	0.122	0.099	0.112	0.099	0.099	0.100
(10, 2, 0.2)	0.209	0.094	0.065	0.066	0.066	0.066	0.066
(10, 2, 0.4)	0.462	0.095	0.071	0.076	0.070	0.070	0.070
(10, 5, 0.2)	0.208	0.073	0.042	0.042	0.042	0.042	0.042
(10, 5, 0.4)	0.462	0.076	0.048	0.046	0.044	0.044	0.045
(10, 10, 0.2)	0.205	0.065	0.031	0.030	0.030	0.030	0.030
(10, 10, 0.4)	0.455	0.068	0.037	0.032	0.031	0.031	0.031
				Bias			
(n_0, T, a)	$\hat{\boldsymbol{\beta}}_{con}$	$\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$	$\hat{\boldsymbol{\beta}}_{aee}$	$\check{\boldsymbol{\beta}}_{sim}$	$\check{\boldsymbol{\beta}}_{cuee}$	$\check{\boldsymbol{\beta}}_{upd}$	$\hat{\boldsymbol{\beta}}$
(1, 1, 0.2)	0.200	0.063	0.262	0.017	0.010	0.010	0.015
(1, 1, 0.4)	0.450	0.071	0.317	0.058	0.017	0.018	0.025
(1, 2, 0.2)	0.203	0.056	0.195	0.014	0.010	0.011	0.008
(1, 2, 0.4)	0.450	0.060	0.266	0.037	0.017	0.018	0.010
(1, 5, 0.2)	0.202	0.057	0.143	0.006	0.005	0.005	0.005
(1, 5, 0.4)	0.449	0.060	0.232	0.017	0.009	0.010	0.005
(1, 10, 0.2)	0.201	0.055	0.121	0.005	0.005	0.005	0.003
(1, 10, 0.4)	0.452	0.059	0.222	0.010	0.006	0.006	0.004
(2, 1, 0.2)	0.200	0.059	0.196	0.018	0.007	0.007	0.009
(2, 1, 0.4)	0.447	0.064	0.219	0.060	0.011	0.011	0.012
(2, 2, 0.2)	0.201	0.055	0.140	0.012	0.006	0.006	0.005
(2, 2, 0.4)	0.448	0.061	0.169	0.033	0.008	0.009	0.006
(2, 5, 0.2)	0.202	0.056	0.095	0.005	0.004	0.004	0.004
(2, 5, 0.4)	0.448	0.059	0.131	0.015	0.006	0.006	0.004
(2, 10, 0.2)	0.201	0.055	0.075	0.003	0.002	0.002	0.002
(2, 10, 0.4)	0.449	0.059	0.117	0.009	0.004	0.004	0.002
(5, 1, 0.2)	0.202	0.057	0.128	0.018	0.004	0.004	0.005
(5, 1, 0.4)	0.449	0.061	0.140	0.059	0.004	0.004	0.006
(5, 2, 0.2)	0.201	0.055	0.092	0.010	0.003	0.003	0.003
(5, 2, 0.4)	0.448	0.058	0.103	0.033	0.005	0.005	0.003
(5, 5, 0.2)	0.202	0.055	0.060	0.004	0.002	0.002	0.002
(5, 5, 0.4)	0.449	0.058	0.072	0.014	0.003	0.003	0.002
(5, 10, 0.2)	0.201	0.056	0.044	0.002	0.001	0.001	0.001
(5, 10, 0.4)	0.448	0.059	0.058	0.007	0.001	0.001	0.002
(10, 1, 0.2)	0.201	0.056	0.093	0.018	0.004	0.004	0.004
(10, 1, 0.4)	0.447	0.059	0.099	0.059	0.004	0.004	0.004
(10, 2, 0.2)	0.202	0.056	0.065	0.009	0.002	0.002	0.002
(10, 2, 0.4)	0.451	0.058	0.071	0.032	0.003	0.003	0.002
(10, 5, 0.2)	0.203	0.055	0.042	0.004	0.001	0.001	0.001
(10, 5, 0.4)	0.451	0.058	0.048	0.014	0.001	0.001	0.001
(10, 10, 0.2)	0.200	0.055	0.031	0.002	0.001	0.001	0.001
(10, 10, 0.4)	0.445	0.058	0.037	0.007	0.001	0.001	0.001

Table 3. Percentage of correctly identified models by $\hat{\boldsymbol{\beta}}_{con}$, $\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$, $\hat{\boldsymbol{\beta}}_{aee}$, $\check{\boldsymbol{\beta}}_{sim}$,$\check{\boldsymbol{\beta}}_{cuee}$, $\check{\boldsymbol{\beta}}_{upd}$, and $\hat{\boldsymbol{\beta}}$ for selected $n_0(\times 10^2)$, $T(\times 10)$, and a based on 1000 replications from (36).

(n_0, T, a)	$\hat{\boldsymbol{\beta}}_{con}$	$\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$	$\hat{\boldsymbol{\beta}}_{aee}$	$\check{\boldsymbol{\beta}}_{sim}$	$\check{\boldsymbol{\beta}}_{cuee}$	$\check{\boldsymbol{\beta}}_{upd}$	$\hat{\boldsymbol{\beta}}$
(1, 1, 0.2)	37.4	37.4	92.1	44.9	46.0	46.4	38.1
(1, 1, 0.4)	0.9	0.9	10.2	0.9	0.8	0.9	0.8
(1, 2, 0.2)	5.4	5.4	45.1	6.4	6.4	6.6	5.2
(1, 2, 0.4)	0.7	0.7	0.0	0.7	0.6	0.4	0.6
(1, 5, 0.2)	1.1	1.1	0.9	0.9	0.6	0.6	1.0
(1, 50, 0.4)	0.9	0.9	0.0	0.9	0.9	0.9	1.0
(1, 10, 0.2)	0.9	0.9	0.0	0.8	0.8	0.7	0.9
(1, 10, 0.4)	0.5	0.5	0.0	0.4	0.4	0.4	0.5
(2, 1, 0.2)	5.6	5.6	15.2	6.7	6.7	6.8	5.8
(2, 1, 0.4)	1.9	1.9	0.0	2.3	1.6	1.6	1.8
(2, 2, 0.2)	0.8	0.8	0.5	0.7	0.7	0.7	0.8
(2, 2, 0.4)	0.7	0.7	0.1	1.1	0.8	0.8	0.8
(2, 5, 0.2)	0.6	0.6	0.0	0.5	0.4	0.4	0.6
(2, 5, 0.4)	0.6	0.6	0.0	0.4	0.4	0.4	0.5
(2, 10, 0.2)	0.7	0.7	0.0	0.7	0.7	0.7	0.7
(2, 10, 0.4)	0.6	0.6	0.0	0.6	0.5	0.5	0.5
(5, 1, 0.2)	1.1	1.1	0.3	1.0	0.9	0.9	1.0
(5, 1, 0.4)	1.1	1.1	0.4	1.1	0.9	0.9	1.1
(5, 2, 0.2)	0.8	0.8	0.2	0.7	0.4	0.4	0.6
(5, 2, 0.4)	1.2	1.2	0.2	1.3	0.9	0.9	1.1
(5, 5, 0.2)	1.0	1.0	0.5	0.8	1.0	1.0	1.0
(5, 5, 0.4)	0.6	0.6	0.0	0.6	0.5	0.5	0.6
(5, 10, 0.2)	0.7	0.7	0.4	0.7	0.7	0.7	0.7
(5, 10, 0.4)	1.1	1.1	0.5	1.1	1.1	1.1	1.1
(10, 1, 0.2)	1.1	1.1	0.7	1.3	0.9	0.9	1.1
(10, 1, 0.4)	1.2	1.2	0.8	1.5	1.0	1.0	1.2
(10, 2, 0.2)	1.2	1.2	0.8	1.4	1.1	1.1	1.2
(10, 2, 0.4)	1.0	1.0	0.6	1.1	1.0	1.0	1.0
(10, 5, 0.2)	1.0	1.0	0.6	1.0	1.0	1.0	1.0
(10, 5, 0.4)	1.3	1.3	0.4	1.2	1.2	1.2	1.2
(10, 10, 0.2)	1.6	1.6	0.7	1.5	1.6	1.6	1.6
(10, 10, 0.4)	0.7	0.7	0.3	0.7	0.7	0.7	0.6

5.2 Implementation to Big Data

We implemented our approach to big data. We assumed that the data set contained 1000 explanatory variables such that we had $p = 1001$ in (31). We choose $m_{ti} = 1$ for all $t \in \{1, \ldots, T\}$ and $i \in \{1, \ldots, n_t\}$ with fixed $n_t = n_0 = 10^4$. We chose $T \in \{1, 5, 10, 20, 50, 100, 200\}$ such that we could control the size of the data. For each selected T, we generated T data sets and wrote them to the hard disk. We evaluated properties of $\hat{\boldsymbol{\beta}}_{con}$, $\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$, $\hat{\boldsymbol{\beta}}_{aee}$, $\check{\boldsymbol{\beta}}_{sim}$, $\check{\boldsymbol{\beta}}_{cuee}$, and $\check{\boldsymbol{\beta}}_{upd}$ in one-step IRWLS. We also evaluated properties of the traditional MLE, given by $\hat{\boldsymbol{\beta}}$ for the entire data.

To implement one-step IRWLS, we first loaded individual data sets sequentially, and then calculated the values of the six estimators. As only individual data set was needed, our approach could be used for arbitrarily large T (e.g., a few thousand). We did not find any difficulties in all of the cases that we studied. Our approach always provided the results of the six estimators even when $T = 200$. The size of the entire data was about 37.3GB. The time of the computation was proportional to T. The time taken for $\hat{\boldsymbol{\beta}}_{con}$, $\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$, and $\check{\boldsymbol{\beta}}_{sim}$ was about one minute when $T = 1$ and slightly over three hours when $T = 200$. The time taken for $\hat{\boldsymbol{\beta}}_{aee}$, $\check{\boldsymbol{\beta}}_{cuee}$, and $\check{\boldsymbol{\beta}}_{upd}$ was about nine times longer than that for $\hat{\boldsymbol{\beta}}_{con}$, $\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$, and $\check{\boldsymbol{\beta}}_{sim}$. This was because the computation of the MLEs for individual data sets in $\hat{\boldsymbol{\beta}}_{aee}$, $\check{\boldsymbol{\beta}}_{cuee}$, and $\check{\boldsymbol{\beta}}_{upd}$ was time-consuming.

We must load the entire data set to memory in the computation of $\hat{\boldsymbol{\beta}}$. We successfully derived $\hat{\boldsymbol{\beta}}$ by the glm function of R when $T = 20$ but not when $T = 50$. The size of the entire data was about 1.49GB when $T = 20$ or about 3.73GB when $T = 50$. We checked the memory consumption issue by Windows Task Manager when $T = 20$. We found that the size of memory consumption used by the glm function of R was about 10GB, which was over six times of the size of the data. This amount was close to the maximum memory capacity in size that we could use in our computation. Because the glm function in R needed to open a few matrices with sizes identical to the size of the data, the memory consumption in size was often many times higher than the size of the data. We found that traditional IRWLS could not be used if the size of the data was over 20% of the memory in size of the computing system.

6. APPLICATION

We applied our approach to the airline data set, which has been studied by many authors [14, 12, 22]. The data set can be downloaded from the ASA (American Statistical Association) website. The airline data contained flight delay information from 1987 to 2008 over hundreds of airports in the United States. The airline data were given by data sets for individual years. Except for 1987, all of those had over five million flights. The entire data set contained over 100 million flights. The size of the entire data was over 60GB.

Since some important information was lost in the data sets before 1994, we decided to analyze the data from 1995 to 2008. We used the variable for minutes of arrival flight delay to define a Bernoulli response variable. According to the definition given by FAA (United States Federal Aviation Administration), a flight is considered late if it is delayed at least 15 minutes. Then, we defined the value of the Bernoulli response variable equal to one if the arrival flight was delayed at least 15 minutes or zero otherwise. We studied seven continuous explanatory variables and one factor variable. The seven explanatory variables were *actual elapsed time* (x_1), *CRS elapsed time* (x_2), *air time* (x_3), *departure delay* (x_4), *distance* (x_5), *taxi in* (x_6), and *taxi out* (x_7). The factor variable was days of week (i.e., Monday-Sunday). After cleaning missing variables, the final data contained over 84 million observations (i.e., $n \geq 8.4 \times 10^7$) with size about 40GB. Each individual year data set between 1995 and 2008 had over five million observations.

We assumed that the online analysis was carried out by files for individual years according to the calendar order. To compare, we attempted to use the traditional IRWLS given by the glm function in R to analyze these files. We chose the file for 1995, which had over 5.2×10^6 records. We studied two models. The first only contained the seven continuous explanatory variables. We only studied the main effects of these variables. The second also contained the factor variable, where we considered the interaction effects between the factor and continuous explanatory variables. We studied the second model because we wanted to investigate properties of the multiple learning procedure introduced in Section 3.5.

Table 4. Estimates of coefficients and their standard errors ($\times 10^{-2}$) given by $\hat{\boldsymbol{\beta}}_{aee}$, $\check{\boldsymbol{\beta}}_{cuee}$, and $\check{\boldsymbol{\beta}}_{upd}$ in the model with actual elapsed time (x_1), CRS elapsed time (x_2), and arrival delay (x_4) only.

	$\hat{\boldsymbol{\beta}}_{aee}$		$\check{\boldsymbol{\beta}}_{cuee}$		$\check{\boldsymbol{\beta}}_{upd}$	
	Est	SE	Est	SE	Est	SE
Intercept	−75.78	5.35	−130.52	13.34	−178.32	45.50
x_1	4.88	0.35	8.42	0.86	11.49	2.93
x_2	−4.88	0.35	−8.42	0.86	−11.49	2.93
x_4	4.88	0.34	8.42	0.86	11.49	2.93

Table 5. Estimates of coefficients given by $\check{\boldsymbol{\beta}}_{upd}$ with respect to days of week when only actual elapsed time (x_1), CRS elapsed time (x_2), and arrival delay (x_4) are considered.

Days of Week	Number of Flights	Intercept	x_1	x_2	x_4
Mon	12373903	−178.31	11.50	−11.50	11.50
Tue	12266222	−179.15	11.55	−11.55	11.55
Wed	12315917	−178.76	11.52	−11.52	11.52
Thu	12329282	−178.52	11.51	−11.51	11.51
Fri	12363950	−178.31	11.50	−11.50	11.49
Sat	10782918	−178.06	11.48	−11.48	11.48
Sun	11778432	−177.02	11.41	−11.41	11.41

We loaded individual year data to memory and applied the glm function of R. We checked memory consumption of the two models by Windows Task Manager. The first model used less than 2GB memory in size. The second model used over 14GB in size. We could not fit the second model for two years data. We then applied our proposed approach to the entire data. We loaded files for individual years to memory sequentially and then computed the set of WSS by the online approach. After the set of WSS was derived, we computed all of the six estimators that we considered in the previous section as well as their variance-covariance matrices.

We followed exactly the same procedure that we had used in the previous section. After the WSS matrix for the entire data was derived, we calculated the estimates of parameters and their standard errors. We studied all of the reduced models obtained of the first model, which was obtained by removing a few explanatory variables from it. In all of these, we checked overdispersion issues and found that it was not a concern. Therefore, we decided to ignore this issue. We found that x_1, x_2, and x_4 were important variables, because their absolute z-values were all over 400. We also found that x_3, x_5, x_6, and x_7 were unimportant variables. For example, the p-values for x_3, x_5, x_6, and x_7 given by $\check{\boldsymbol{\beta}}_{upd}$ were 0.85, 0.35, 0.81, and 0.30, respectively. Then, we removed x_3, x_5, x_6, and x_7 and refitted the model by the WSS only. We calculated all of the six estimators. We found that the results of $\hat{\boldsymbol{\beta}}_{con}$, $\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$, and $\check{\boldsymbol{\beta}}_{sim}$ were far away from those of $\hat{\boldsymbol{\beta}}_{aee}$, $\check{\boldsymbol{\beta}}_{cuee}$, and $\check{\boldsymbol{\beta}}_{upd}$. This was because MLEs of individual years were used in the computation of $\hat{\boldsymbol{\beta}}_{aee}$, $\check{\boldsymbol{\beta}}_{cuee}$, and $\check{\boldsymbol{\beta}}_{upd}$. Then, we decided to focus on $\hat{\boldsymbol{\beta}}_{aee}$, $\check{\boldsymbol{\beta}}_{cuee}$, and $\check{\boldsymbol{\beta}}_{upd}$. Their results are displayed in Table 4.

We implemented the multiple learning procedure in fitting the second model when the factor variable for days of week was involved. We defined the full model as the model with all of the main effects of the continuous and factor variables, and the interaction effects between the continuous and the factor variables. The full model contained seven continuous and one factor explanatory variables. The size of $\boldsymbol{\beta}$ was $8 \times 7 = 56$, and the size of the WSS matrix was $(56+1)^2 = 3249$. We calculated the WSS matrix for the full model only. We only used it to analyze the full model and its reduced models, where we treated the first model as one of the reduced models. We also found that x_1, x_2, and x_4 were more important than the other four explanatory variables. We fitted the model with the three important explanatory variables, the factor variable for days of week, and their interaction effects (Table 5). Our results showed that all of the three interaction effects were significant, even though the coefficients were not far away from those given by the model without the factor variable. We used the WSS matrix obtained from the second model to calculate estimates of parameters for the first model. We found the WSS matrix was identical to that we had obtained previously. The results were identical to those given by Table 4.

7. DISCUSSION

In this article, we propose an online multiple learning approach for GLMs in big data. The approach can overcome the memory and computational efficiency barriers caused by the size of the data. The technique of the approach is developed based on a new concept called working sufficient statistics (WSS), which becomes sufficient statistics (SS) in linear models for nonnormal data. Since the size of SS in GLMs for nonnormal data is often identical to the size of data, we cannot use SS to reduce the memory needed in the computation. The usage of WSS successfully solves the problem. Since the size of the WSS does not depend on the number of observations, our method can be used even if the size of data is much higher than the memory size of the computing system.

Similar to SS in regression, we can construct the set of WSS based on a family of GLMs. As far as the set of WSS is available, all of the GLMs in the family can be fitted simultaneously without the need of another access of the data. This induces a multiple learning approach. If an individual learning approach is used, then one must access the entire data set again if another model is considered. Because the access of big data is time consuming, a multiple learning approach is more efficient than an individual learning approach.

The difference between online and offline learning is that an online learning approach can use information of previous data but an offline learning approach cannot. Since GLMs for nonnormal data are fitted by an iterative algorithm, the usage of previous information can significantly enhance the precision of the computation. We have developed at least two methods in the one-step IRWLS. The first method (i.e., the method for $\hat{\boldsymbol{\beta}}_{con}$, $\hat{\boldsymbol{\beta}}_{\hat{\sigma}}$, and $\hat{\boldsymbol{\beta}}_{aee}$) does not use any information of the previous data to construct the objective function for the current data. This method can be easily modified to offline learning. Therefore, the first method can be treated as either an online or an offline learning approach. The second method (i.e., the method for $\check{\boldsymbol{\beta}}_{sim}$, $\check{\boldsymbol{\beta}}_{cuee}$ and $\check{\boldsymbol{\beta}}_{upd}$) uses the information of the previous data. It is completely an online method, which cannot be modified to offline learning. Our research shows that online and offline learning approaches for big data should be developed separately. This is left to future research.

ACKNOWLEDGMENTS

The author appreciates great comments from Editor-in-Chief Ming-Hui Chen and four anonymous referees that significantly improved the quality of the article.

Received 0 12 0000

REFERENCES

[1] Agresti, A. (2002). *Categorical Data Analysis*, Wiley, Hoboken, New Jersey.

[2] Dean, J. and Ghamawat, S. (2004). MapRedue: simplified data processing on large clusters. In *Proceeding of OSDI*, 137-150.

[3] Enea, M. (2009). Fitting linear models and generalized linear models with large data sets in R. *Statistical Methods for the Analysis of Large Datasets: book of short papers*, 411-414.

[4] Ferguson, T.S. (1996). *A Course in Large Sample Theory*, Chapman & Hall/CRC Press, Boca Raton, Florida.

[5] Green, P.J. (1984). Iteratively weighted least squares for maximum likelihood estimation and some robust and resistant alternatives. *Journal of Royal Statistical Society Series B*, **46**, 149-192.

[6] Guha, S., Hafen, R., Rounds, J., Xia, J., Li, J., Xi, B., and Cleveland, W.S. (2012). Large complex data: divide and recombine (D&R) with Rhipe. *Stat*, **1**, 53-67.

[7] Lin, N. and Xi, R. (2011). Aggregated estimating equation estimation. *Statistics and Its Interface*, **4**, 73-83.

[8] Ma, P, and Sun, X. (2015). Leveraging for big data regression. *WIREs Computational Statistics*, **7**, 70-76.

[9] McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics*, **11**, 59-67.

[10] Meeker, W.Q., and Hong, Y. (2014). Reliability meets big data: opportunities and challenges. *Qualify Engineering*, **26**, 102-116.

[11] Reyes-Ortiz, J.L., Oneto, L., and Anguita, D. (2015). Big Data Analytics in the Cloud: Spark on Hadoop vs MPI/OpenMP on Beowulf. *Precedia Computer Science*, **53**, 121-130.

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

[13] van der Vaart, A.W. (1998). *Asymptotic Statistics*, Cambridge University Press, Cambridge, UK.

[14] Wang, C., Chen, M.-H., Schifano, E., Wu., J., and Yan, J. (2016). Statistical methods and computing for big data. *Statistics and Its Interface*, **9**, 399-411.

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

[16] Wang, H., Zhu, R., and Ma, P. (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, to appear.

[17] Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., and Stoica, I. (2010). Spark: cluster computing with working sets. *HotCloud*, **10**, Article 10.

[18] Zhang, T. and Yang, B. (2017a). Box-Cox transformation in big data. *Technometrics*, **59**, 189-201.

[19] Zhang, T. and Yang, B. (2017b). An exact approach to ridge regression for big data. *Computational Statistics*, **32**, 909-928.

[20] Zhang, T. and Yang, B. (2018). Dimension reduction for big data. *Statistics and Its Interface*, **11**, 295-306.

[21] Zhang, T. (2019). General Gaussian estimation. *Journal of Multivariate Analysis*, **169**, 234-247.

[22] Zhang, T. and Yang, B. (2019). Accounting for factor variables in big data regression. *Statistica Sinica*, to appear.

Tonglin Zhang
Department of Statistics
Purdue University
250 North University Street
West Lafayette, IN 47907-2066
USA
E-mail address: tlzhang@purdue.edu

Baijian Yang
Department of Computer and Information Technology
Purdue University
250 North University Street
West Lafayette, IN 47907-2066
USA
E-mail address: byang@purdue.edu