Optimal Subsampling Methods in Bike Sharing Data Analysis

Xiaofeng Zhao (✉ zxfstats@ncwu.edu.cn)

North China University of Water Resources and Electric Power <https://orcid.org/0000-0001-6245-8031>

Research Article

Keywords: Bike Sharing data analysis, Subsampling, Optimal subsampling algorithm

Posted Date: April 28th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-458508/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Optimal Subsampling Methods in Bike Sharing Data Analysis

Xiaofeng Zhao

Received: date / Accepted: date

Abstract Bike sharing system are popular around the world. Traditional bike sharing system require the bikes to be returned to fixed stations, while modern system allows users to leave bikes wherever they like, ready for the next user to pick them up. Smartphone use GPS signal to keep track of its bikes and monitor where most bikes are used and where to place them. Smartphone simultaneously collect many other information such as weather condition, temperature and so on, these features have influence on the delivering amount of bikes. Due to the extensive number of smartphone users, big data technique is required to handle this situation. We apply subsample method to this smartphone collected big data. In this paper, we derive non-uniform sampling distributions and propose optimal subsampling algorithm. We apply the proposed optimal subsampling algorithm to analyze the smartphone collected bike sharing data set, perform extensive computer experiments to evaluate the numerical performance of the proposed sampling algorithm. Our results indicated that the proposed optimal algorithm outperformed the uniform method and have faster running time than using the whole data set.

Keywords Bike Sharing data analysis · Subsampling · Optimal subsampling algorithm

Xiaofeng Zhao
School of Mathematics and Statistics, North China University
of Water Resources and Electrical Power, Zhengzhou, Henan,
China, 450046
Tel.: +8615239687109
Fax: +86-0371-69310032
E-mail: zxfstats@ncwu.edu.cn

1 Introduction

With the universal spread of the smartphone, it is more convenient to collect users' data. Big data analysis is inevitable required to deal with the smartphone collected data. A typical example of smartphone collected data is bike sharing data, people use mobile app to get access to the bike sharing system and consequently generate huge amount of data to process. Our goal is to use the smartphone collected information building model to predict the number of delibered sharing bikes for different locations. Due to the nonnegative integer characteristic of the response variable, we referred this kind of data as count data. Count data are observations of the number of occurrences of a behavior in a fixed period of time, for example, hospital visits, blog comments, car/bike renters, and questionnaire respondents. Analysis of count data is an important task in social sciences and economics. Since linear regression does not take into account the restricted number of count response values, it is not an appropriate technique for count data. GLM model is the best suitable choice for count data. Meanwhile, these smartphone collected data satisfy the definition of big data. Big data are on a massive scale with regard to volume, velocity, variety, and veracity that exceed both the capacity of the conventional software tools and operating systems and the physical spaces of computers, see e.g. Fan, et al. [1]. These smartphone collected data shares two inherent characteristics of the big data: (1) the data is too large for a desktop to store, and (2) the computing task takes too long time to accomplish. Divide-and-conquer approach easily solve the first problem, see e.g. Bhlmann, et al. [2]. The subsampling approach simultaneously solve both of them. In this paper, our focus is to choose a subsample as a surrogate for the full data,

use GLM model completing the data analysis. Uniform sampling is often used to take subsample, since it is simple in mathematic and easy in computation. Monte Carlo and bootstrap method are the two representative cases, see e.g. Freedman, et al. [3], Barbeet al. [4]. Although uniform subsampling provide a solution for intensive computing, it can not detect important observation and is not effective in extracting information. We consider non-uniform sampling to overcome this drawbacks. Mathematicians, computer scitists and statisticians have already made important progress in this area. Drineas, et al. [5,6] constructed fast Monte Carlo algorithms to approximate matrix multiplication and developed randomized algorithms for faster least squares approximation. Drineas, et al. [7] presented a sampling algorithm for the least squares fit problem and studied its algorithmic properties. A key feature of the above algorithms is the non-uniform sampling. Drineas, et al. [8] also proposed the non-uniform distribution to develop fast algorithms to approximate the product of two matrices, the idea is to minimize the expected squared Frobenius distance of the product and its approximate. Ma and Sun et al. [9] used the leverage scores as non-uniform importance sampling distributions for big data linear regression. Zhu, et al. [10] obtained optimal subsampling distributions for large sample linear regression and proposed the OPT and PL subsampling method in linear regression model, they discussed the sampling probability by minimizing the trace of the intermittent part of the variance-covariance matrix of the subsampling estimator, derived asymptotic normality and performed simulations and real data analysis. Wang, et al. [11] (2017) proposed non-uniform subsampling probabilities that minimize the asymptotic mean squared error of subsampling estimator in logistic regression, established consistency and asymptotic normality of the estimator. Xu, et al. [12] studied subsampled newton methods with non-uniform sampling. Wang, et al. [13] developed information-based subdata selection for large linear regression. Peng and Tan [14,15] investigated A-optimal subsampling for Big Data linear regression and constructed fast algorithms. Avron, et al. [16] used random-sampling and random-mixing techniques to describe a fast LS solver for dense highly overdetermined systems. Ma, et al. [17] conduct the leverage score based non-uniform subsampling method, this method used the estimate from a subsample taken randomly from the full sample to approximate the full sample ordinary least square estimate, they proposed BLEV, SLEV, and LEVUNW method to perform subsampling. Their work is mainly in linear regression and logistic regression area. In this paper, we focus on smartphone collected bike sharing data, we extend the sub-

sample method beyond linear regression and logistic regression. We summarize our main contributions as follows:

- We generalize the subsampling method into more complex data structure, expanded this method to analyse smartphone collected bike sharing data, demonstrate the feasibility to other count data situation.
- We develop the non-uniform distributions by minimizing the trace of certain variance-covariance matrices, which results in our less variance estimator, we show the consistency and asymptotic normality properties of our estimator, we pursue statistical inference of our non-uniform distributions.
- we introduce the two stage algorithm to reduce the computing time, we conduct extensive computer experiments to illustrate our results, we seek the algorithmic properties.

The rest of the paper is organized as follows. In Section 2, We introduce the smartphone data collection and integrate subsampling method into the framework of count data analysis. We also apply subsampling method to GLM model and derive the corresponding optimal ditributions. In Section 3, We introduce the two stage algorithm. In Section 4, we show our computer experiment result, we compare our proposed sampling methods to the uniform sampling, we compare the running time of our methods to the full data. We present the performance evaluation results of our proposed methods. Finally, we conclude our work in section 5.

2 Smartphone Data Collection System and Optimal Subsampling Method

2.1 Smartphone Data Collection System

Bike sharing systems allow people to rent a bike at one of the rental stations scattered around the city, use them for a short traveling. The rental station includes one terminal and several bike stands. When people use a bike, they use their smartphone scan the QR code on the bike, the smartphone APP records the real time transaction and communicates with the central control facility, if the transaction was approved, the central control facility send signal to open the locker, during this time, people can use the bike. After people finish the travel, they lock the bike, the bike send signal to the central control facility, the central control facility records the location of the capable bike. Operators of bike sharing systems make this information available to the users online, other people could get to know where the bike are capable of using. A important factor in the success of a bike sharing system is its ability to meet the

demand for bikes at each station. Our goal is to predict the amount of bikes delivered at each station. The delivery amount of sharing bike have close relationship with several key features, The information of the key feature are getting from the smartphone. When the user begin to use the smartphone APP to rent the bike, the smartphone starts to collect the data, it records the data at each time point, the intervals between each time point are preseted. When the user returns the bike. we take averaged values as our variable value. The temperature, humidity and some weather condition related variables are optimal choices as the predict variables.

2.2 Optimal Subsampling Method

The dependent variable of the smartphone bike sharing data is the amount of delivered bikes, since its inborn characteristics of non-negativity and integer attribute, we refer this kind of data as count data. In a count data analysis model, the mean of a count response y_i and covariate vector $\mathbf{x}_i$ satisfy

$$E(y_i) = Var(y_i) = h(\mathbf{x}_i^T \beta), \quad i = 1, \dots, n \quad (1)$$

where $\beta \in \mathbf{R}^p$ is a regression parameter and h is an inverse link function. Typically, $h(t) = \exp(t)$ (the inverse log link). The estimate $\hat{\beta}$ is obtained by maximizing the likelihood function, which is equivalent to solve the following equation,

$$\sum_{i=1}^n (y_i - \exp(\mathbf{x}_i^T \beta)) \mathbf{x}_i = 0 \quad (2)$$

For ordinary sample size, apply Neton's method to iteratively obtain the numerical solution,

$$\hat{\beta}^{k+1} = \hat{\beta}^k - \mathbf{H}_0^{-1} \mathbf{S}_0 \quad (3)$$

Where

$$\mathbf{H}_0 = \sum_{i=1}^n \exp(\mathbf{x}_i^T \hat{\beta}^k) \mathbf{x}_i \mathbf{x}_i^T,$$

$$\mathbf{S}_0 = \sum_{i=1}^n (y_i - \exp(\mathbf{x}_i^T \hat{\beta}^k)) \mathbf{x}_i$$

$\hat{\beta}^k$ is the parameter estimate at the k th iteration, stop the process until $\|\hat{\beta}^{k+1} - \hat{\beta}^k\| < \epsilon$, ϵ is the preseted value to control the convergence.

We then consider the case that the sample size n is very large and the estimate $\hat{\beta}$ is not available. In this situation, We take a subsample as a surrogate of the whole sample, use the subsample estimate to approximate the whole sample. Let $\pi_i, i = 1, \dots, n$ be a sampling distribution on the n data points. A special case is $\pi_i = \frac{1}{n}$

for the uniform distribution. We use the sampling distribution to take a subsample without replacement. The subsample is represented as $(\mathbf{x}_i^{*T}, y_i^*), i = 1, \dots, r$, with the subsample size $r < n$ and $\pi_i^*, i = 1, \dots, r$ are the corresponding sampling probabilities. We now approximate the maximum likelihood estimate $\hat{\beta}$ by the subsample estimate which solves the equation

$$\sum_{i=1}^r \frac{1}{r\pi_i^*} (y_i^* - \exp(\mathbf{x}_i^{*T} \beta)) \mathbf{x}_i^* = 0. \quad (4)$$

Similarly, the subsample parameter estimate $\hat{\beta}_r$ can be caculated from the following equation.

$$\hat{\beta}_r^{k+1} = \hat{\beta}_r^k - \mathbf{H}_1^{-1} \mathbf{S}_1 \quad (5)$$

where

$$\mathbf{H}_1 = \sum_{i=1}^r \frac{\exp(\mathbf{x}_i^{*T} \hat{\beta}_r^k) \mathbf{x}_i^* \mathbf{x}_i^{*T}}{\pi_i^*},$$

$$\mathbf{S}_1 = \sum_{i=1}^r \frac{(y_i^* - \exp(\mathbf{x}_i^{*T} \hat{\beta}_r^k)) \mathbf{x}_i^*}{\pi_i^*}$$

Under certain condition, subsample estimate $\hat{\beta}_r$ is consistent with full sample estimate $\hat{\beta}$, moreover, as $n \rightarrow \infty$ and $r \rightarrow \infty$

$$\hat{\beta}_r - \hat{\beta} \rightarrow N(0, \mathbf{V}) \quad (6)$$

where

$$\mathbf{V} = \mathbf{H}^{-1} \mathbf{V}_m \mathbf{H}^{-1},$$

$$\mathbf{H} = \frac{1}{n} \sum_{i=1}^n \exp(\mathbf{x}_i^T \hat{\beta}) \mathbf{x}_i \mathbf{x}_i^T,$$

$$\mathbf{V}_m = \frac{1}{rn^2} \sum_{i=1}^n \frac{(y_i - \exp(\mathbf{x}_i^T \hat{\beta}))^2 \mathbf{x}_i \mathbf{x}_i^T}{\pi_i}$$

The subsample estimator $\hat{\beta}_r$ have the two good properties, consistency property shows that $\hat{\beta}_r$ is asymptotically unbiased, normality ensures the subsample estimator to be capable to make statistical inference.

Uniform sampling distribution is not a good choice, the next step is to find optimal distributions. The purpose is to find distributions to minimize the varince matrix $\mathbf{V}$ and $\mathbf{V}_m$. We choose trace criterion to minimize a matrix. Through minimizing the trace of $\mathbf{V}$ and $\mathbf{V}_m$, several types of distribution were proposed:

$$\hat{\pi}_i^{(2)} = \frac{\|(\mathbf{X}^T \mathbf{W}(\hat{\beta}) \mathbf{X})^{-1} \mathbf{x}_i\| |\hat{e}_i|}{\sum_{i=1}^n \|(\mathbf{X}^T \mathbf{W}(\hat{\beta}) \mathbf{X})^{-1} \mathbf{x}_i\| |\hat{e}_i|}, \quad (7)$$

$$\hat{\pi}_i^{(1)} = \frac{\|(\mathbf{X}^T \mathbf{W}(\hat{\beta}) \mathbf{X})^{-\frac{1}{2}} \mathbf{x}_i\| |\hat{e}_i|}{\sum_{i=1}^n \|(\mathbf{X}^T \mathbf{W}(\hat{\beta}) \mathbf{X})^{-\frac{1}{2}} \mathbf{x}_i\| |\hat{e}_i|}, \quad (8)$$

$$\hat{\pi}_i^{(0)} = \frac{\|\mathbf{x}_i\| |\hat{e}_i|}{\sum_{i=1}^n \|\mathbf{x}_i\| |\hat{e}_i|}, \quad (9)$$

$$\bar{\pi}_i^{(2)} = \frac{\|(\mathbf{X}^\top \mathbf{W}(\hat{\beta}) \mathbf{X})^{-1} \mathbf{x}_i\| \hat{g}_i}{\sum_{i=1}^n \|(\mathbf{X}^\top \mathbf{W}(\hat{\beta}) \mathbf{X})^{-1} \mathbf{x}_i\| \hat{g}_i}, \quad (10)$$

$$\bar{\pi}_i^{(1)} = \frac{\|(\mathbf{X}^\top \mathbf{W}(\hat{\beta}) \mathbf{X})^{-\frac{1}{2}} \mathbf{x}_i\| \hat{g}_i}{\sum_{i=1}^n \|(\mathbf{X}^\top \mathbf{W}(\hat{\beta}) \mathbf{X})^{-\frac{1}{2}} \mathbf{x}_i\| \hat{g}_i}, \quad (11)$$

$$\bar{\pi}_i^{(0)} = \frac{\|\mathbf{x}_i\| \hat{g}_i}{\sum_{i=1}^n \|\mathbf{x}_i\| \hat{g}_i}, \quad i = 1, \dots, n \quad (12)$$

where

$$\mathbf{X} = \begin{Bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{Bmatrix}$$

$$\mathbf{W}(\hat{\beta}) = \text{Diag}(\hat{\mu}_i), \quad \hat{\mu}_i = \exp(\mathbf{x}_i^T \hat{\beta}),$$

$$\hat{e}_i = y_i - \hat{\mu}_i, \quad \hat{g}_i = \sqrt{\hat{\mu}_i} \quad i = 1, \dots, n$$

3 Two Stage algorithm

By looking at (7)-(12), the optimal sample distributions are all depend on the full sample estimator $\hat{\beta}$. Since in the frame work of big data analysis, the full sample estimate $\hat{\beta}$ is usually not available. We propose a two stage algorithm to approximate optimal sample distributions. In stage one, we take a uniform subsample to get a pioneer estimate of $\hat{\beta}$, which can be used to approximate the optimal sample distributions, then we use these approximated optimal sample distributions to draw the second stage subsample, caculate our subsample estimate $\hat{\beta}_r$ based on the second stage subsample.

Two Stage Algorithm:

- Stage 1 Take subsample with replacement of size r_0 using uniform $\pi_i = \frac{1}{n}$ and obtain the inial estimate $\hat{\beta}_{r_0}$. Substitute $\hat{\beta}$ with $\hat{\beta}_{r_0}$ in (7)-(12) to caculate the approximate sampling probabilities.
- Stage 2 Take subsample with replacement of size r using the sampling probabilities obtained in Stage 1. Caculate the subsample estimate $\hat{\beta}_r$ according to (5).

The two stage algorithm reduces the computation time, totally we only use $r_0 + r$ data points. Our caculation process involve iteration, so we only take a portion of the whole data set greatly speed up the computation time.

4 Smartphone Bike Sharing Data Results

This data set is available from the UCI Machine Learning Repository website. Bike sharing systems can be considered new generation of old-fashioned bike rentals. The use of this system is not restricted to rentals and returns at the same docking station, bikes can be returned to any docking station after usage. Predicting the hourly bike request will help in planing, expanding and maintaining adequate number of bikes. In United States, the bike sharing system has been proved to be very successful in major cities, including Washington, DC, New York, Chicago, Los Angles, where bikes sharing has become a popular transportation option. Our goal in this example is to build statistical models to predict hourly request of bikes in Washington DC area. There are totally 17,389 observations in this data set, the response variables are the casual bike rentals. The predictor variables are season, workingday, daytime, weathersit, temp, hum and windspeed. All these features are available from the users'smartphones. The season variables include spring indicator, summer indicator and fall indicator, winter is the reference level. Variable workingday indicates whether a day is a working day, the reference level is weekend of holiday. Variable daytime indicates if the time is between 7am to 22pm, with the referencing time range from 0am to 5am representing the reference level night time. The season, workingday, daytime variables can be directly obtained from the users' smartphones. Weathersit variable has 4 categories, the first category represents clear, few clouds, and partly cloudy; the second category represents mist plus cloudy, mist plus broken clouds, and mist; the third category represents light snow, light rain; and the fourth category represents heavy rain, and snow plus frog. Since the fourth category only has three observations, we combine the third and fourth categories. We choose the first category as the reference. Temp is normalized temperature in Celsius. Hum is normalized humidity. Windspeed is normalized wind speed. The weathersit, temp, hum and windspeed data is collected in this way. There are 11 regression coefficients for the predictor variables including the intercept.

The computer experiments are carried out in the follwing way, we take a uniform subsample of size $r_0 = 200$ to get an initial estimates $\hat{\beta}_{r_0}$; then plug in to the formulas(7-12) to approximate the optimal subsampling distributions; take subsample of size r according to the approximate distributions to get $\hat{\beta}_r$; We calculate the empirical mean square errors for different subsample sizes r for each of $B = 1000$ subsamples using the

formula:

$$MSE = \frac{1}{B} \sum_{b=1}^B \|\hat{\beta}_{r,b} - \hat{\beta}\|^2,$$

where $\hat{\beta}_{r,b}$ is the estimate from the b^{th} subsample with subsample size r .

Table 1 Averaged estimates, theoretical standard errors(Tse), empirical standard errors(Ese), and P-values based on 1000 subsamples, the pre-subsample size $r_0 = 200$, $r = 400$

	$\hat{\pi}^{(2)}$				$\hat{\pi}^{(1)}$				$\hat{\pi}^{(0)}$			
	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value
Intercept	1.2297	0.1735	0.1748	< 0.001	1.2280	0.1859	0.1863	< 0.001	1.2294	0.2219	0.2217	< 0.001
Summer	0.5862	0.1091	0.1098	< 0.001	0.5850	0.1209	0.1195	< 0.001	0.5853	0.1417	0.1399	< 0.001
Fall	0.3495	0.1292	0.1286	0.0069	0.3503	0.1379	0.1385	0.0111	0.3502	0.1560	0.1551	0.0248
Winter	0.6195	0.1032	0.1012	< 0.001	0.6192	0.1117	0.1130	< 0.001	0.6197	0.1323	0.1323	< 0.001
Workingdays	-0.9169	0.0575	0.0592	< 0.001	-0.9185	0.0501	0.0486	< 0.001	-0.9166	0.0486	0.0481	< 0.001
Daytime	1.6246	0.1025	0.1034	< 0.001	1.6252	0.1107	0.1100	< 0.001	1.6245	0.1315	0.1329	< 0.001
W2_cloudy	0.0190	0.0672	0.0662	0.7777	0.0189	0.0608	0.0606	0.7554	0.0183	0.0617	0.0606	0.7672
W3_rain	-0.3821	0.1169	0.1169	0.0011	-0.3839	0.1244	0.1232	0.0020	-0.3837	0.1501	0.1516	0.0105
Temp	0.5273	0.0432	0.0413	< 0.001	0.5258	0.0417	0.0430	< 0.001	0.5262	0.0448	0.0450	< 0.001
Hum	-0.2028	0.0317	0.0302	< 0.001	-0.2014	0.0297	0.0310	< 0.001	-0.2011	0.0308	0.0311	< 0.001
Windspeed	0.0036	0.0273	0.0284	0.8946	0.0040	0.0243	0.0234	0.8690	0.0038	0.0251	0.0267	0.8795

Fig. 1 Percentages of the 95% confidence intervals which caught the full sample MLE $\hat{\beta}_2$

Figure 2 shows that the coverage probabilities of proposed methods achieve the nominal 95% in both small and large subsample sizes.

Table 2 The length ratios of the 95% confidence intervals of proposed subsampling to uniform subsampling, subsample size $r = 400$

	$\hat{\pi}^{(2)}$	$\hat{\pi}^{(1)}$	$\hat{\pi}^{(0)}$	$\bar{\pi}^{(2)}$	$\bar{\pi}^{(1)}$	$\bar{\pi}^{(0)}$
Intercept	0.6349	0.6804	0.8121	0.8452	0.8814	0.9984
Summer	0.5312	0.5883	0.6898	0.7669	0.8457	0.9614
Fall	0.5435	0.5798	0.6562	0.7885	0.8384	0.9234
Winter	0.5724	0.6197	0.7339	0.8089	0.8648	0.9914
Workingdays	0.7005	0.6102	0.5929	0.9840	0.8387	0.7888
Daytime	0.7974	0.8609	0.9127	0.9068	0.9564	0.9237
W2_cloudy	0.7076	0.6400	0.6495	0.9489	0.8407	0.8219
W3_rain	0.5209	0.5543	0.6685	0.7203	0.7516	0.8746
Temp	0.5981	0.5775	0.6209	0.8541	0.8007	0.8203
Hum	0.6374	0.5972	0.6183	0.8907	0.8139	0.8066
Windspeed	0.6426	0.5712	0.5913	0.8863	0.7804	0.7898

Table 1 continued

	$\hat{\pi}^{(2)}$				$\hat{\pi}^{(1)}$				UNIF			
	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value
Intercept	1.2297	0.2310	0.2308	< 0.001	1.2278	0.2408	0.2421	< 0.001	1.2295	0.2728	0.2735	< 0.001
Summer	0.5860	0.1575	0.1570	0.0002	0.5855	0.1757	0.1744	0.0008	0.5852	0.1975	0.1986	0.0030
Fall	0.3490	0.1875	0.1864	0.0627	0.3504	0.1994	0.1994	0.0788	0.3504	0.2196	0.2180	0.1106
Winter	0.6204	0.1458	0.1465	< 0.001	0.6196	0.1559	0.1571	0.0001	0.6203	0.1787	0.1778	0.0005
Workingdays	-0.9176	0.0807	0.0824	< 0.001	-0.9174	0.0688	0.0680	< 0.001	-0.9182	0.0647	0.0661	< 0.001
Daytime	1.6242	0.1294	0.1295	< 0.001	1.6243	0.1358	0.1343	< 0.001	1.6256	0.1535	0.1535	< 0.001
W2_cloudy	0.0172	0.0901	0.0887	0.8484	0.0189	0.0798	0.0779	0.8128	0.0187	0.0781	0.0764	0.8108
W3_rain	-0.3828	0.1617	0.1624	0.0179	-0.3838	0.1687	0.1681	0.0229	-0.3831	0.1963	0.1960	0.0510
Temp	0.5277	0.0616	0.0632	< 0.001	0.5277	0.0578	0.0593	< 0.001	0.5280	0.0592	0.0603	< 0.001
Hum	-0.2020	0.0443	0.0440	< 0.001	-0.2020	0.0405	0.0410	< 0.001	-0.2029	0.0401	0.0385	< 0.001
Windspeed	0.0031	0.0377	0.0380	0.9335	0.0034	0.0332	0.0320	0.9176	0.0036	0.0336	0.0338	0.9154

The P-values in Table 1 are based on theoretical standard errors. The averaged parameter estimates are very close among all the 7 methods, but our proposed methods have the smallest standard errors.

Table 2 shows the confidence interval length ratios of the proposed methods versus uniform method. First of all, all of the values in Table 2 is less than 1, indicating that the 95% confidence intervals constructed by our proposed methods have shorter length compared to uniform subsampling. Second, Table 6.4 shows that the coverage probabilities of proposed methods achieve the nominal 95% in both small and large subsample sizes.

That is, our methods are more accurate while maintaining the nominal coverage probability.

Fig. 2 MSE plot for different subsample sizes

Figure 2 is the MSE plots, as subsample size r increases, we found that the largest MSE's over different subsample sizes are given by the uniform methods, the smallest MSE's over different subsample sizes are given by the $\hat{\pi}^{(2)}$ sampling. All the proposed methods are under the uniform method line, which means our methods have smaller MSE.

Table 3 MSE ratios of the proposed subsampling to uniform subsampling, the pre-subsample size $r_0 = 200$

r	160	400	800	1600	4000	8000
$\hat{\pi}^{(2)}$	0.3556	0.3567	0.3578	0.3827	0.4173	0.3739
$\hat{\pi}^{(1)}$	0.3961	0.4033	0.4020	0.4298	0.3955	0.4144
$\hat{\pi}^{(0)}$	0.5393	0.5409	0.5550	0.5712	0.6105	0.6278
$\bar{\pi}^{(2)}$	0.6717	0.6711	0.6771	0.6727	0.7308	0.7945
$\bar{\pi}^{(1)}$	0.7252	0.7268	0.7333	0.7389	0.7330	0.8532
$\bar{\pi}^{(0)}$	0.9156	0.9160	0.9287	0.9416	0.9254	0.9062

Table 3 contains the MSE ratio of the proposed sampling methods to uniform sampling for different subsample sizes r . In Table 3, all the values are less than 1, indicating our proposed sampling distributions produce smaller MSE than the uniform sampling. The highest reduction is about 65%. $\hat{\pi}^{(k)}$ produce smaller MSEs than $\bar{\pi}^{(k)}$, $\hat{\pi}^{(2)}$ is the best among $\hat{\pi}^{(k)}$, and $\bar{\pi}^{(2)}$ is the best among $\bar{\pi}^{(k)}$, $k = 0, 1, 2$.

Table 4 The CPU times in seconds for $r_0 = 200$

r	200	500	1000	1500	2000	3000
$\hat{\pi}^{(2)}$	4.191	4.205	4.226	4.241	4.567	4.632
$\bar{\pi}^{(2)}$	2.313	2.334	2.356	2.395	3.025	3.564
Full data CPU seconds 5.872						

In order to evaluate the computation efficiency, we report the running times for computing $\hat{\beta}_r$ by using $\hat{\pi}^{(2)}$. The experiment is carried out using R programming language. Those values were computed on a desktop with Intel i5 processor and 8GB memory. We recorded the CPU times for 1000 repetitions, then average the time to make the comparison fair. The results are in Table 4. We observe that the $\hat{\pi}^{(2)}$ require more time than $\bar{\pi}^{(2)}$ method. All the proposed methods have significant less computing times than the full data

Table 5 Averaged iterations using Newton's method. The iterations for full data set are 8.4

r	$\hat{\pi}^{(2)}$		$\bar{\pi}^{(2)}$		Uniform
	Step1	Step2	Step1	Step2	
500	8.89	8.77	8.67	8.49	8.40
1000	8.75	8.56	8.56	8.23	8.80
1500	8.56	8.32	8.59	8.39	8.54
2000	8.55	8.01	8.58	8.53	8.34
2500	8.60	8.91	8.62	8.85	8.27

In Table 5, we can see all the proposed methods have similar number of iterations, indicating smaller subsample sizes do not necessarily increase the iterations for Newton's method.

5 Conclusions

In this paper, subsampling method was introduced, we proposed optimal subsampling algorithm and establish the consistency and asymptotic normality theorem. we apply optimal subsample algorithm to the smartphone collected bike sharing data. Experimental results shows our proposed methods have better performance than the uniform sampling method in terms of less variance and less computation time compared to full data. There are important issues that we will investigate in the future.

1. The smartphone collected bike sharing data used in this paper is not big enough, we will apply our method to the data set which contains millions observations, also we would like to further investigate other types of smartphone collected data.
2. We have developed the optimal sampling method to analyse the smartphone collected bikesharing data.

This data have a feature that the data matrix have huge amount of rows. Other types of smartphone collected data matrix may also have huge amount of columns. How to handle this problems worth of future study.

6 Ethics declarations

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of interest

The authors declare that they have no conflict of interest.

Funding details

None.

Informed consent

Informed consent was obtained from all individual participants included in the study.

7 Authorship contributions

Xiaofeng Zhao developed the idea of the study, contributed to the acquisition and interpretation of data. Xiaofeng Zhao provided review and revised the manuscript. Xiaofeng Zhao read and approved the final manuscript.

References

1. FAN, J., HAN, F. AND LIU, H. (2013). Challenges of big data analysis. *arXiv:1308.1479*.
2. BHLMANN, P., KANE M., DRINEAS P., MARK VAN DER LAAN (EIDOTRS) (2016). *Handbook of Big Data*. Chapman and Hall/CRC.
3. FREEDMAN, D.A. (1981). Bootstrapping regression models. *Ann. Statist.* **9**(6): 1218–1228.
4. BARBE, P. AND BERTAIL, P. (1995). *Weighted bootstrap*. Lecture Notes in Statistik. Vol. 98, Springer, New York.
5. DRINEAS P., MAGDON-ISMAIL, M., MAHONEY M.W. and WOODRUFF, D.P. (2012d). Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, **13**: 3475–3506.
6. DRINEAS P., KANNAN R. and MAHONEY M.W. (2006a). Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, **36**: 132–157.
7. DRINEAS, P., MAHONEY, M.W., MUTHUKRISHNAN, S. and SARLÓS, T. (2010). Faster least squares approximation. *Numerische Mathematik*, **117**(2): 219–249.
8. DRINEAS P., MAHONEY M.W. and MUTHUKRISHNAN S. (2006b). Sampling algorithms for ℓ_2 regression and applications. *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1127–1136.

9. MA, P. AND SUN, X. (2014). Leveraging for big data regression. *Computational Statistics*. textbf7 (1): 70-76.
10. ZHU, R., MA, P., MAHONEY, M.W. AND YU, B. (2015). Optimal subsampling Approaches for Large Sample Linear Regression. *arXiv:1509.0511.v1* [stat.ME].
11. WANG, H., ZHU, R., AND MA, P. (2015). Optimal subsampling for large sample logistic regression. *JASA DOI: 10.1080/01621459.2017.1292914*.
12. XU, P., YANG J., MAHONEY, M.W. (2015). Subsampled Newton Methods with Non-uniform Sampling *arXiv:1607.00559* [math.OC].
13. WANG, H., MA, Y., STUFKEN, J (2017). Information-Based Optimal Subdata Selection for Big Data Linear Regression *arXiv:1710.10382* [stat.ME].
14. PENG, H. AND TAN, F. (2018a). A Fast Algorithm For Computing The A-optimal Sampling Distributions In Big Data Linear Regression. *Preprint*.
15. PENG, H. AND TAN, F. (2018b). Big Data Linear Regression Via A-optimal Subsampling. Submitted to *Ann. Statist.*
16. AVRON, H., MAYMOUNKOV, P. and TOLEDO, S. (2010). Blendenpik: Supercharging LAPACK's least-squares solver. *SIAM Journal on Scientific Computing*, **32**: 1217–1236.
17. MA, P., MAHONEY, M.W. AND YU, B. (2015). A statistical perspective on algorithmic leveraging *Journal of Machine Learning Research*. **16** (April): 861–911.

Figures

Figure 1

Percentages of the 95% condence intervals which caught the full sample MLE B^2

Figure 2

MSE plot for different subsample sizes