Statistica Sinica Preprint No: SS-2023-0434	
Title	Balanced Subsampling for Big Data with Categorical Predictors
Manuscript ID	SS-2023-0434
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0434
Complete List of Authors	Lin Wang
Corresponding Authors	Lin Wang
E-mails	linwang@purdue.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

BALANCED SUBSAMPLING FOR BIG DATA WITH CATEGORICAL PREDICTORS

Lin Wang

Department of Statistics, Purdue University

Abstract: Supervised learning under measurement constraints is a common challenge in statistical and machine learning. In many applications, despite extensive design points, acquiring responses for all points is often impractical due to resource limitations. Subsampling algorithms offer a solution by selecting a subset from the design points for observing the response. Existing subsampling methods primarily assume numerical predictors, neglecting the prevalent occurrence of big data with categorical predictors across various disciplines. This paper proposes a novel balanced subsampling approach tailored for data with categorical predictors. A balanced subsample significantly reduces the cost of observing the response and possesses three desired merits. First, it is nonsingular and, therefore, allows linear regression with all dummy variables encoded from categorical predictors. Second, it offers optimal parameter estimation by minimizing the generalized variance of the estimated parameters. Third, it allows robust prediction in the sense of minimizing the worst-case prediction error. We demonstrate the superiority of balanced subsampling over existing methods through extensive simulation studies and a real-world application.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

Key words and phrases: Data labeling, D-optimality, Experimental design, Orthogonal array, Robust prediction.

1. Introduction

Supervised learning under measurement constraints is a common challenge in statistical and machine learning (Wang et al., 2017; Meng et al., 2021). In many applications, despite the availability of extensive predictor observations (design points), acquiring the observations of the response variable for all design points is frequently impractical due to resource limitations. For example, consider a scenario in healthcare where researchers aim to develop a predictive model for patient outcomes based on a diverse set of health-related predictors. A large set of predictor observations, such as patient demographics, medical history, and genetic information, is readily available. However, obtaining the corresponding response variable, such as a medical condition or treatment outcome, may involve invasive procedures or expensive diagnostic tests. Given the constraints of limited resources, observing the response of every individual in the dataset becomes practically impossible. Consequently, selecting an informative subsample from the set of design points to observe becomes crucial and challenging.

In recent years, there has been a growing interest in developing design-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

based optimal subsampling methods. Most existing methods focus on numerical predictors in various learning models, such as linear regression (Wang et al., 2019; Ma et al., 2015; Wang et al., 2021), generalized linear models (Wang et al., 2018; Ai et al., 2021; Cheng et al., 2020), linear mixed models (Zhu et al., 2024), quantile regression (Wang and Ma, 2021; Ai et al., 2021), nonparametric regression (Meng et al., 2020; Zhang et al., 2024), Gaussian process modeling (He and Hung, 2022), and model-free scenarios (Mak and Joseph, 2018; Shi and Tang, 2021; Song et al., 2022).

Big data with categorical predictors are frequently encountered in many scientific research areas (Huang et al., 2014; Zuccolotto et al., 2018; Johnson et al., 2018). Numerical predictors may also be binned into categorical ones for better modeling and interpretation (Kanda, 2013; Yu et al., 2022). Despite numerous studies on subsampling methods, they do not apply to data with categorical predictors, so researchers have no choice but to use simple random subsamples, for example, see Maronna and Yohai (2000) and Yu et al. (2022). However, simple random subsamples may bring significant issues, especially for data with categorical predictors. To illustrate, consider that categorical predictors are commonly encoded using dummy variables in regression models. In a dataset with p categorical predictors, each having q_j levels, $j = 1, \ldots, p$, these predictors are coded to $\sum_{j=1}^{p}(q_j - 1)$ dummy

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

variables, which substantially increases the dimensionality of the regression task. Thus, as pointed out by Maronna and Yohai (2000) and Koller and Stahel (2017), singular subsamples (or more accurately, subsamples with singular information matrices) are frequently encountered when dealing with categorical predictors due to the high dimensionality of the dummy variables. Consequently, a simple random subsample cannot facilitate the estimation of the effect for every dummy variable, even though the full data allows for such estimation. This deficiency arises because the subsample lacks crucial information in the full data, and avoiding such substantial information loss is paramount. Furthermore, even among nonsingular subsamples, there can be significant variations in the accuracy of parameter estimation. Identifying the subsample that enables optimal parameter estimation is also important.

This paper proposes a novel method named "balanced subsampling" designed specifically for subsampling data with categorical predictors. The selected subsample achieves a combinatorial balance between values (levels) of the predictors and, therefore, enjoys three desired merits. First, a balanced subsample is generally nonsingular and thus allows the estimation of all parameters in linear (ANOVA) regression. Second, a balanced subsample provides the optimal parameter estimation in the sense of mini-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

mizing the generalized variance of the estimated parameters. Third, when the established model is used for prediction, the model trained on a balanced subsample provides robust predictions in the sense of minimizing the worst-case prediction error. For practical use, we develop an algorithm that sequentially selects data points from the full data to obtain a balanced subsample.

The remainder of the paper is organized as follows. Section 2 presents the issues of simple random subsampling for data with categorical predictors, which motivates us to develop a new subsampling method. Section 3 proposes the balanced subsampling method and develops an efficient algorithm for sequentially subsampling from big data. Section 4 examines the performance of balanced subsampling through extensive simulations, and Section 5 demonstrates the utility of using balanced subsamples in a real-world application. Section 6 offers concluding remarks. Supplementary Materials provide proof of technical results and discuss the computational complexity of the proposed algorithm.

2. Motivations

Let $X = (x_1, \ldots, x_N)^T$ denote the design matrix of the full data, where $x_i = (x_{i1}, \ldots, x_{ip})^T$ consists of the values of p categorical predictors, each

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

with q_j levels for $j = 1, \ldots, p$ and is coded to $q_j - 1$ binary dummy variables. Let $z_{i,jk}$ be the value of the kth dummy variable for x_{ij} and Z the matrix formed by $z_{i,jk}$. Linear regression on the dummy variables is given by:

$$y_i = \beta_0 + \sum_{k=1}^{q_1-1} \beta_{1k} z_{i,1k} + \cdots + \sum_{k=1}^{q_p-1} \beta_{pk} z_{i,pk} + \varepsilon_i, \tag{2.1}$$

where β_{jk} are parameters to be estimated and ε_i is the independent random error with mean 0 and variance σ^2. It is intuitive to assume that $M = Z^T Z$ is nonsingular, enabling linear regression on the full data if all responses can be observed.

We consider taking a subsample of size n from the full data X, denoted as X_s, and observe its corresponding response vector y_s. The OLS estimator for $\beta = (\beta_0, \beta_{11}, \ldots, \beta_{p(q_p-1)})^T$ based on the subsample is given by

$$\hat{\beta}_s = (Z_s^T Z_s)^{-1} (Z_s^T y_s), \tag{2.2}$$

where $Z_s = (z_1^*, \ldots, z_n^*)^T$ are the rows in Z corresponding to points in X_s. We have three concerns regarding the subsample X_s.

2.1 Nonsingularity

The subsample should allow the estimation of all parameters in β, which is possible only if the information matrix, $M_s = Z_s^T Z_s$, is nonsingular. However, singular subsamples (subsamples with a singular information matrix)

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

are frequently encountered when dealing with categorical predictors, which can be illustrated by the following two toy examples.

Example 1. Assume the full data contain a single categorical predictor with 2 repetitions of 5 levels, that is, $X = (1, 1, 2, 2, 3, 3, 4, 4, 5, 5)^T$. Use dummy variables, then

$$Z = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}^T.$$

Consider choosing a subsample X_s with 5 points, then M_s is nonsingular only if X_s contains at least one observation of each level. Out of the $\binom{10}{5} = 252$ possible subsamples, only 2^5 of them are good in this way, and the probability of obtaining such a subsample with simple random sampling is $2^5/252 = 12.7\%$.

Example 2. Suppose the full data have $N = 1000$ points and $p = 2$ predictors. We generate data from an independent bivariate normal distribution with mean 0 and variance 1, divide the range of either predictor into 5 equal-sized intervals, and code the values according to which interval they fall. Then each predictor includes 5 levels, and the two predictors have

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

25 possible pairs of levels. We select a subsample of size $n = 25$. There are $\binom{1000}{25} \approx 10^{49}$ possible subsamples from simple random sampling. An exhaustive examination of all those subsamples is infeasible. Therefore, we randomly investigate 10^5 of them, and only 4.81% have nonsingular information matrices. It is not easy to obtain a nonsingular subsample from simple random sampling.

2.2 Optimal estimation

Even among the nonsingular subsamples, the accuracy of parameter estimation varies greatly across different subsamples.

Example 3. We continue Example 2. For all the nonsingular subsamples with $n = 25$ (out of the 10^5 investigated random subsamples), we generate the response variable y through the model

$$y_i = 1 + z_{i,11} + \cdots + z_{i,14} + z_{i,21} + \cdots + z_{i,24} + \varepsilon_i, \tag{2.3}$$

where $\varepsilon_i \sim N(0, 1)$, and train the model in (2.1) on each of the nonsingular subsamples. We repeat this process $T = 1000$ times and examine the empirical mean squared error (MSE):

$$\text{MSE} = T^{-1} \sum_{t=1}^{T} \|\hat{\beta}^{(t)} - \beta\|^2, \tag{2.4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

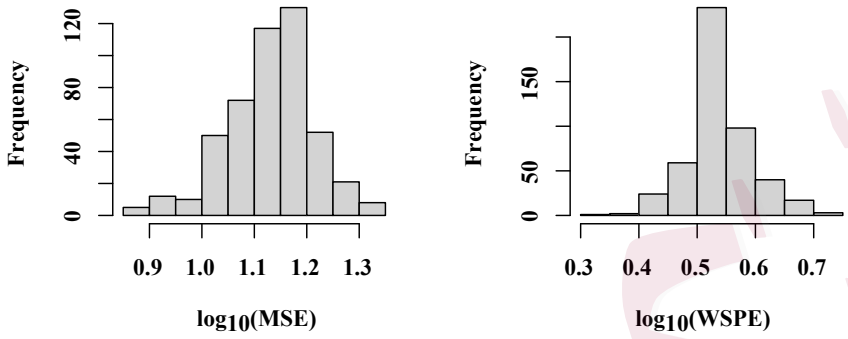

Figure 1: Histograms of $\log_{10}(\text{MSE})$ and $\log_{10}(\text{WSPE})$ of the trained model on each subsample with a nonsingular information.

where $\hat{\beta}^{(t)}$ is the OLS estimate of β via a subsample in the tth repetition, $t = 1, \ldots, 1000$. Figure 1 (left) shows the histogram of $\log_{10}(\text{MSE})$ for all nonsingular subsamples. The MSE varies dramatically, with the minimum as low as $10^{0.85} = 7.1$ achieved by only a couple of subsamples. Recall that we examined 10^5 random subsamples, and only a couple of them allows the "optimal" estimation. It is very hard to obtain such "optimal" subsamples from simple random sampling.

2.3 Robust prediction

We hope the trained model on a subsample provides "robust" prediction, where the terminology "robust" can be understood in the sense of performing

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

well in the worst-case scenario.

Example 4. We continue Example 3 and examine the empirical worst-case squared prediction error (WSPE) for all nonsingular subsamples:

$$\text{WSPE} = \max_{x \in \mathcal{X}} \left\{ T^{-1} \sum_{t=1}^{T} (y^{(t)} - z^T \hat{\beta}^{(t)})^2 \right\} \tag{2.5}$$

where $\mathcal{X}$ includes all the 25 possible pairs of levels for the two predictors, z is the vector of dummy variables for any $x \in \mathcal{X}$, $y^{(t)}$ is the response in the tth repetition, and $\hat{\beta}^{(t)}$ is the OLS estimate of β via a subsample, for $t = 1, \ldots, 1000$. Figure 1 (right) shows the histogram of $\log_{10}(\text{WSPE})$ for all the nonsingular subsamples. The minimum of WSPE is $10^{0.31} = 2.0$ achieved by a single subsample, which is, again, almost impossible to obtain from simple random sampling.

3. Balanced Subsampling

In this section, we propose the balanced subsampling method and develop a computationally efficient algorithm to implement it. The proposed method targets the above three concerns: providing a nonsingular subsample, enabling optimal parameter estimation, and ensuring robust prediction.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

3.1 The method

We first consider the nonsingularity of a subsample and provide the following important result.

Theorem 1. *Let $\lambda_{\min}(M_s)$ be the smallest eigenvalue of M_s. For a subsample X_s,*

$$\lambda_{\min}(M_s) \geq n\nu(1 - f(X_s))$$

where n is the subsample size of X_s, ν is a positive constant independent of X_s,

$$f(X_s) = \sqrt{\sum_{j=1}^{p}\sum_{u=1}^{q_j} q_j^2 \left[\frac{1}{q_j} - \frac{n_j(u)}{n}\right]^2 + \sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p}\sum_{u=1}^{q_j}\sum_{v=1}^{q_k} q_j q_k \left[\frac{1}{q_j q_k} - \frac{n_{jk}(u,v)}{n}\right]^2}, \tag{3.1}$$

$n_j(u)$ is the number of times that the uth level of the jth predictor is observed in X_s, and $n_{jk}(u,v)$ is the number of times that the pair of levels (u,v) is observed for the jth and kth predictors in X_s. Therefore, M_s is nonsingular if $f(X_s) < 1$.

Theorem 1 indicates that we can search for the subsample that minimizes $f(X_s)$ to ensure that the M_s is nonsingular. By (3.1), $f(X_s)$ has two critical components: (a) $f_j = \sum_{u=1}^{q_j}[1/q_j - n_j(u)/n]^2$ that measures the balance of levels for the jth predictor, and (b) $f_{jk} = \sum_{u=1}^{q_j}\sum_{v=1}^{q_k}[1/(q_j q_k) - n_{jk}(u,v)/n]^2$ that measures the balance of level combinations for the jth

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

and kth predictors. Clearly, if $f_j = 0$, all levels of the jth predictor are observed the same number of times in X_s so that they achieve the perfect balance; if $f_{jk} = 0$, all pairs of levels for the jth and kth predictors are observed the same number of times in X_s. Such balance is called combinatorial orthogonality, and a matrix possessing combinatorial orthogonality is called an orthogonal array.

Generally, an orthogonal array of strength t is a matrix where entries of each column of the matrix come from a fixed finite set of q_j levels for $j = 1, \ldots, p$, arranged in such a way that all ordered t-tuples of levels appear equally often in every selection of t columns of the matrix. The t is called the strength of the orthogonal array. Readers are referred to Hedayat et al. (1999) for a comprehensive introduction to orthogonal arrays. Here is an example of an orthogonal array with $p = 3$ predictors, each having 3 levels, and strength $t = 2$:

$$\begin{pmatrix} 1 & 1 & 1 & 2 & 2 & 2 & 3 & 3 & 3 \\ 1 & 2 & 3 & 1 & 2 & 3 & 1 & 2 & 3 \\ 1 & 2 & 3 & 2 & 3 & 1 & 3 & 1 & 2 \end{pmatrix}^T .$$

Each pair of levels in any two columns of the orthogonal array appears once. Clearly, we have the following lemma.

Lemma 1. *A subsample X_s forms an orthogonal array of strength two if*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

and only if $f(X_s) = 0$.

Now we show that the subsample minimizing $f(X_s)$ also allows the optimal estimation of parameters. To see this, note that $\hat{\beta}_s$ in (2.2) is an unbiased estimator of β with

$$\mathrm{Var}(\hat{\beta}_s|X_s) = \sigma^2 M_s^{-1} = \sigma^2 (Z_s^T Z_s)^{-1}. \tag{3.2}$$

The $\mathrm{Var}(\hat{\beta}_s|X_s)$ is a function of X_s (in the form of Z_s), which indicates again that the subsampling strategy is critical in reducing the variance of $\hat{\beta}_s$. To minimize $\mathrm{Var}(\hat{\beta}_s|X_s)$, we seek the X_s which, in some sense, minimizes M_s^{-1}. This is typically done, in experimental design strategy, by minimizing an optimality function $\psi(M_s^{-1})$ of the matrix M_s^{-1} (Kiefer, 1959; Atkinson et al., 2007). A common choice for ψ is the determinant, which is akin to the criterion of D-optimality for the selection of optimal experimental designs.

Theorem 2. *A subsample X_s is D-optimal for the model in* (2.1) *if* $f(X_s) = 0$.

Cheng (1980) showed that an orthogonal array of strength two is universally optimal, i.e., optimal under a wide variety of criteria by minimizing the sum of a convex function of coefficient matrices for the reduced normal equations. However, Cheng's result does not apply to the dummy coding

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

system, so his result does not include Theorem 2. To the best of our knowledge, Theorem 2 originally shows the optimality of orthogonal arrays for the commonly used dummy coding for categorical predictors.

Next, we show that minimizing $f(X_s)$ also allows robust prediction. Let $\mathcal{X}$ denote the set of all possible level combinations of predictors, that is, $\mathcal{X} = \{x = (x_1, \ldots, x_p) : x_j = 1, \ldots, q_j, j = 1, \ldots, p\}$, then $\#\mathcal{X} = \prod_{j=1}^{p} q_j$. For any $x \in \mathcal{X}$, let z be the coded vector of x and Y the random variable of its response with $E(Y) = z\beta$ and $\mathrm{Var}(Y) = \sigma^2$. The WSPE is given by $\max_{x\in\mathcal{X}} E[(Y - z^T\hat{\beta}_s)^2|X_s]$, where

$$E[(Y-z^T\hat{\beta}_s)^2|X_s] = E[(Y-z^T\beta)^2]+E[(z^T\beta-z^T\hat{\beta}_s)^2|X_s] = \sigma^2(1+z^TM_s^{-1}z). \tag{3.3}$$

The WSPE is a function of X_s (in the form of M_s), which indicates again that the subsampling strategy is critical in reducing WSPE. The following theorem shows that the WSPE is minimized when $f(X_s) = 0$.

Theorem 3. *Let $Q = 1 + \sum_{j=1}^{p}(q_j - 1)$. For a subsample X_s of size n,*

$$\max_{x\in\mathcal{X}} E[(Y - z^T\hat{\beta}_s)^2|X_s] \geq \sigma^2(1 + Q/n). \tag{3.4}$$

The equality in (3.4) *holds if $f(X_s) = 0$.*

Theorems 1–3 indicate that $f(X_s) = 0$ ensures model estimability as well as optimal estimation and robust prediction. Considering that a full

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

dataset may generally do not contain a subset with exact zero $f(X_s)$, the objective of the balanced subsampling is to achieve an approximate balance via the optimization problem:

$$\begin{aligned} X_s^* &= \arg\min_{X_s \subseteq X} f(X_s) \\ &s.t. \quad X_s \text{ contains } n \text{ points.} \end{aligned} \tag{3.5}$$

The optimization problem in (3.5) is not easy to solve. The computation of $f(X_s)$ requires the examination of balance for every single predictor and every pair of predictors in X_s, so it requires $O(np^2)$ operations to compute $f(X_s)$ for any X_s. In addition, an exhaustive search for all possible X_s requires $O(N^n)$ operations, making it infeasible for even moderate sizes of the full data. There are many types of algorithms for finding optimal designs and among them, exchange algorithms are among the most popular. For the reasons argued in Wang et al. (2021), these algorithms are cumbersome in solving the subsampling problem in (3.5). We will propose a sequential selection algorithm to efficiently select subsample points.

3.2 A sequential algorithm

The following result is critical in developing the algorithm.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

Theorem 4. *For a subsample $X_s = (x^*_{ij})$, $i = 1, \ldots, n$ and $j = 1, \ldots, p$,*

$$f^2(X_s) = 2n^{-2} \sum_{1 \le i < l \le n} [\delta(x^*_i, x^*_l)]^2 + C, \tag{3.6}$$

where

$$\delta(x^*_i, x^*_l) = \sum_{j=1}^{p} q_j \delta_1(x^*_{ij}, x^*_{lj}), \tag{3.7}$$

$\delta_1(x^*_{ij}, x^*_{lj})$ is 1 if $x^*_{ij} = x^*_{lj}$ and 0 otherwise, and $C = n^{-1}(\sum_{j=1}^{p} q_j)^2 + p - \sum_{j=1}^{p} q_j - p^2$.

By Theorem 4, the optimization in (3.5) can be achieved by minimizing $\sum_{1 \le i < l \le n} [\delta(x^*_i, x^*_l)]^2$. To select an n-point subsample, we start with a random point x^*_1 and select $x^*_2, \ldots, x^*_n$ sequentially. Suppose we have already selected m points, then the $(m+1)$th point is selected by

$$x^*_{m+1} = \arg\min_x \left\{ \sum_{i=1}^{m-1} \sum_{l=i+1}^{m} [\delta(x^*_i, x^*_l)]^2 + \sum_{i=1}^{m} [\delta(x^*_i, x)]^2 \right\} = \arg\min_x \Delta(x)$$

where

$$\Delta(x) = \sum_{i=1}^{m} [\delta(x^*_i, x)]^2, \tag{3.8}$$

and the minimization is over $x \in X \backslash \{x^*_1, \ldots, x^*_m\}$. Since

$$\Delta(x) = \sum_{i=1}^{m-1} [\delta(x^*_i, x)]^2 + [\delta(x^*_m, x)]^2,$$

we only need to compute $\delta(x^*_m, x)$ to update $\Delta(x)$ in the $(m+1)$th iteration. Each iteration has a complexity of $O(Np)$, and the overall complexity of

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

Algorithm 1 Balanced Subsampling (Sequential Selection)

Input: a sample (dataset) X, a required subsample size n

Output: a subsample X_s

Set $m = 1$ and randomly select x_1^* from X

for each $x \in X \backslash \{x_1^*\}$ **do**

Compute $\Delta(x)$ via (3.8)

end for

while $m < n$ **do**

Find $x_{m+1}^* = \arg\min_x \Delta(x)$ and include x_{m+1}^* in X_s

for each $x \in X \backslash \{x_1^*, \ldots, x_{m+1}^*\}$ **do**

Update $\Delta(x) \leftarrow \Delta(x) + [\delta(x_{m+1}^*, x)]^2$

end for

$m \leftarrow m + 1$

end while

the algorithm is $O(Npn)$. Algorithm 1 outlines this sequential selection algorithm.

Example 5. We continue Examples 2–4 to demonstrate the effectiveness of Algorithm 1 by comparing it with existing popular subsampling methods, including simple random sampling (denoted as UNI for consistency with the literature, as it assigns a uniform weight to all observations), information-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

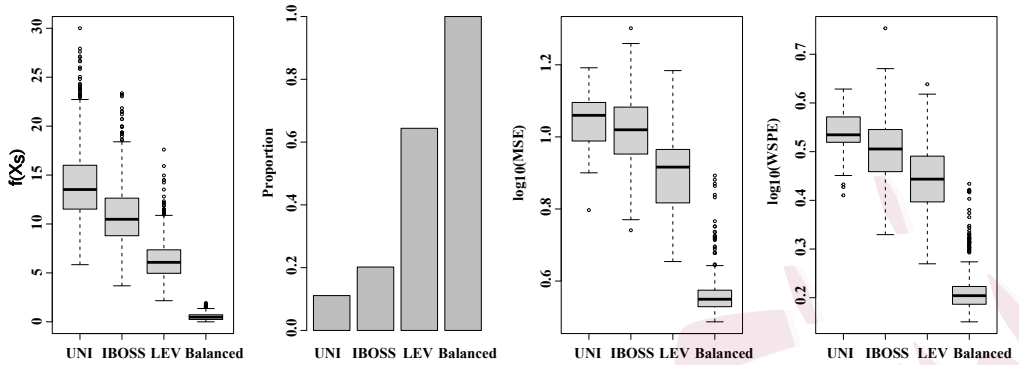

Figure 2: The values of $f(X_s)$, proportion of nonsingular subsamples, MSE, and WSPE for subsamples generated using various methods.

based optimal subdata selection (IBOSS, Wang et al. (2019)), and leveraging subsampling (LEV, Ma and Sun (2015)). We generate the full data following the procedure outlined in Example 2 and select subsamples of size $n = 25$ using different methods. We repeat this process 1000 times. The left two panels of Figure 2 plot the values of $f(X_s)$ and proportions of nonsingular subsamples for each method. UNI often misses many pairs of levels, resulting in high $f(X_s)$ values and a low proportion of nonsingular subsamples. IBOSS and LEV, applied to the dummy variables, offer better balance than UNI, resulting in higher proportions of nonsingular subsamples. Balanced subsamples obtained from Algorithm 1 consistently exhibit smaller $f(X_s)$ values, indicating a higher degree of balance compared to other subsamples, and are consistently nonsingular. For each subsam-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

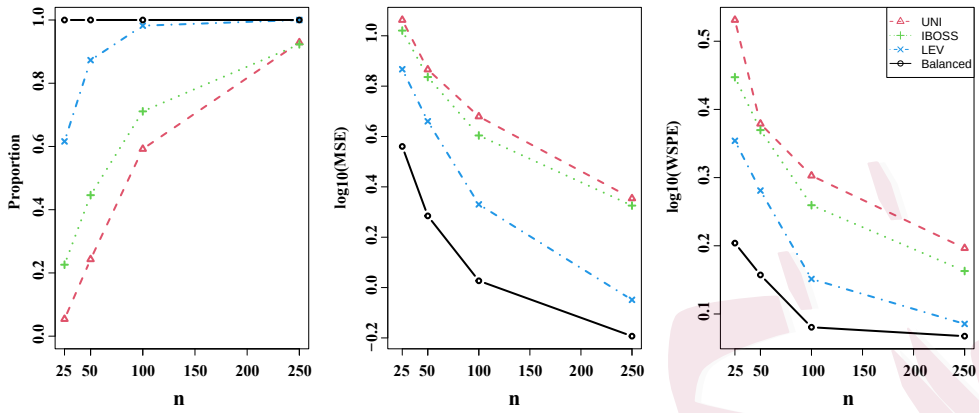

Figure 3: The proportions of nonsingular subsamples (left), MSEs of estimated parameters (middle), and WSPEs (right) for different subsampling methods in Example 5.

ple, we repeatedly generate the response for $T = 1000$ times through the model in (2.3) and examine the empirical MSE in (2.4) and WSPE in (2.5), displayed in the right two panels of Figure 2. The balanced subsamples demonstrate significantly smaller MSEs and WSPEs, clearly outperforming other subsampling methods.

We further explore the performance of subsampling methods across different subsample sizes $n = 25, 50, 100, 250$. For each subsample size, we repeatedly generate the full data and the response for $T = 1000$ times and select a subsample of size n using different methods. Figure 3 plots the proportions of nonsingular subsamples, the empirical MSE (2.4), and

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

the empirical WSPE (2.5). The balanced subsamples always significantly outperform other approaches for any subsample size due to their balance on levels of predictors. Specifically, a balanced subsample is consistently nonsingular. Increasing the subsample size may enhance the nonsingularity of other subsamples, but they still provide much worse parameter estimation and response prediction than a balanced subsample.

4. Simulation studies

We conduct simulation studies to assess the merits of the balanced subsampling method relative to existing subsampling schemes. Consider $p = 20$ predictors each with $q_j = j + 1$ levels for $j = 1, \ldots, p$. The simulation is replicated $T = 1000$ times. In each replication, we generate values of the predictors under three structures:

Case 1. Covariates are independent, and each follows a discrete uniform distribution with q_j levels.

Case 2. Covariates are independent, and for each predictor, the q_j levels have probabilities proportional to $1, \ldots, q_j$.

Case 3. Generate each point x_i from multivariate normal distribution: $x_i \sim$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

$N(0, \Sigma)$ with

$$\Sigma = \left(0.5^{\xi(j,k)}\right), \tag{4.1}$$

where $\xi(j,k)$ is equal to 0 if $j = k$ and 1 otherwise. Discretize $[-3, 3]$ to q_j intervals of equal length, and let $x_{ij} = u$ if x_{ij} falls into the uth interval. Let $x_{ij} = 1$ if $x_{ij} < -3$ and $x_{ij} = q_j$ if $x_{ij} > 3$.

The response data are generated from the linear model in (2.1) with the true value of β being a vector of unity and $\sigma = 1$. We investigate four settings of the full data size $N = 5 \times 10^3, 10^4, 5 \times 10^4$, and 10^5, and two settings of the subsample size $n = 500$ and 2000. Four subsampling approaches, UNI, IBOSS, LEV, and the balanced subsampling, are evaluated by comparing the proportions of nonsingular subsamples, MSEs (2.4), and WSPEs (2.5). To accelerate LEV, we use a fast Singular Value Decomposition method implemented in the R package "corpcor". Since the set of all level combinations $\mathcal{X}$ contains $\prod_{j=1}^{p} q_j = 5 \times 10^{19}$ points, it is infeasible to evaluate predictive performance across the entire set $\mathcal{X}$. Instead, we randomly sample 10^6 points in $\mathcal{X}$ to compute WSPE. Note that the comparisons of subsampling approaches on MSE and WSPE are independent of the settings of the true parameters β or σ, as can be seen from (3.2) and (3.3).

Figure 4 compares the subsampling methods for the full data gener-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

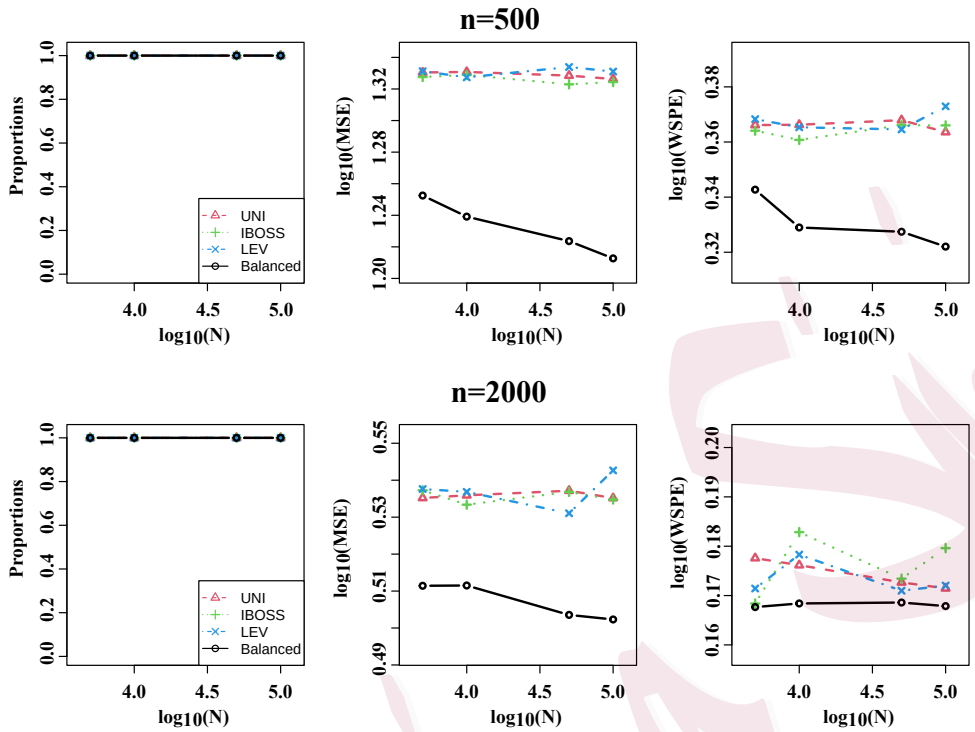

Figure 4: The proportions of nonsingular subsamples (left), MSEs of estimated parameters (middle), and WSPEs (right) for different subsampling methods for predictors in Case 1.

ated in Case 1. In this case, predictor levels in the full data are highly balanced, so all subsamples closely resemble balance and are nonsingular. Even so, the balanced subsamples consistently provide more accurate parameter estimation and slightly better prediction than other methods. Note that when the full data are highly balanced, the increase in the full data size does not have a substantial contribution to the balance of subsamples.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

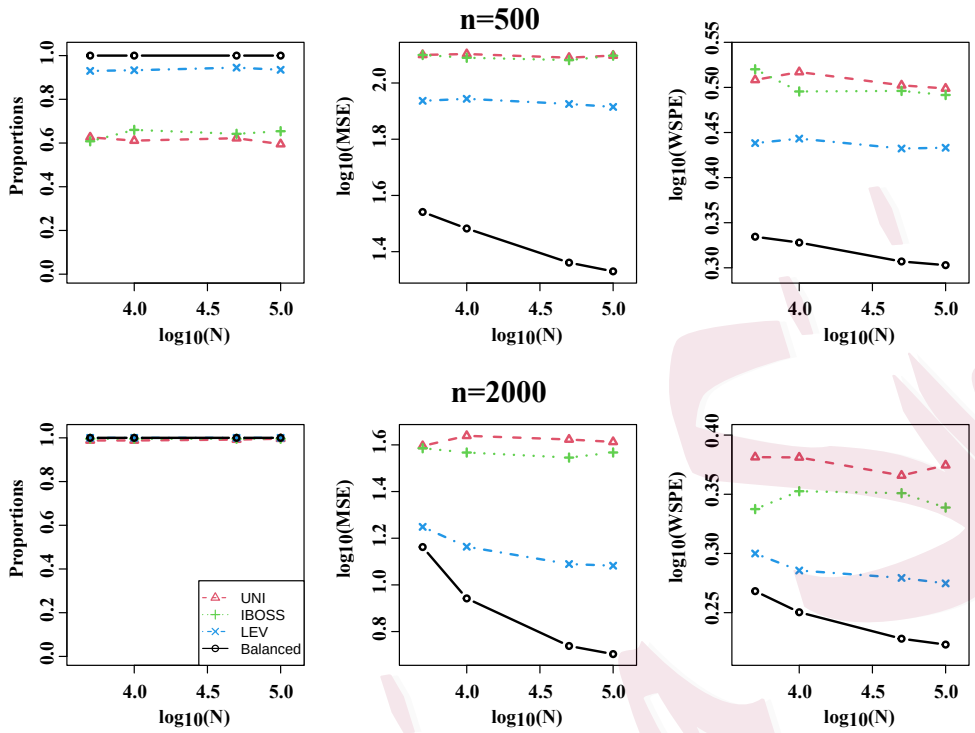

Figure 5: The proportions of nonsingular subsamples (left), MSEs of estimated parameters (middle), and WSPEs (right) for different subsampling methods for predictors in Case 2.

Therefore, balanced subsampling exhibits only a slight improvement in parameter estimation and relatively flat WSPE as the full data size increases. Considering that we can only examine a subset of $\mathcal{X}$, the WSPE may even slightly fluctuate as the full data size increases.

Figure 5 plots results for Case 2, where levels of predictors are unbalanced, which is typically the case in real practice. We observe that, when

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

the subsample size is $n = 500$, only around 60% of UNI and IBOSS subsamples are nonsingular, 90% of LEV subsamples are nonsingular, whereas all balanced subsamples are nonsingular. With just 500 observations, it becomes possible to estimate a model that includes all dummy variables, indicating that a balanced subsample enables significant cost savings in observing the response. Increasing the subsample size to $n = 2000$ may improve the proportions of nonsingularity for other methods, but the estimation and prediction obtained from those subsamples are still much worse than the balanced subsamples. Notably, a balanced subsample with $n = 500$ exhibits better accuracy in parameter estimation when compared with UNI and IBOSS subsamples with $n = 2000$ (around 1.6 on $\log_{10}(\text{MSE})$). This observation once again underscores the significant savings achieved by utilizing a balanced subsample. More importantly, the MSEs and WSPEs from the balanced subsamples decrease fast as the full data size N increases, even though the subsample size is fixed at $n = 500$ or 2000. This trend demonstrates that the balanced subsampling extracts more information from the full data with a fixed subsample size when the full data are more informative. IBOSS has this nice property for continuous predictors (Wang et al., 2019), but not for categorical predictors because of the high association between dummy variables.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

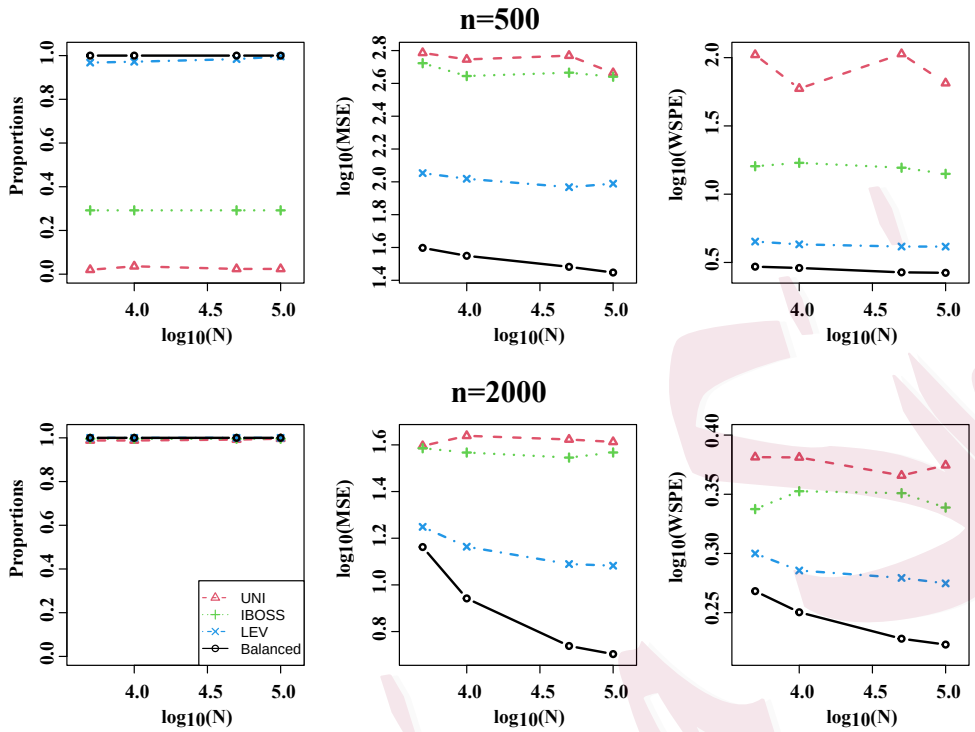

Figure 6: The proportions of nonsingular subsamples (left), MSEs of estimated parameters (middle), and WSPEs (right) for different subsampling methods for predictors in Case 3.

Figure 6 examines Case 3, where predictors are correlated in the full data. For $n = 500$, almost all UNI subsamples are singular, and only less than 40% of IBOSS subsamples are nonsingular. LEV performs well in terms of nonsingularity but has worse MSE and WSPE compared to balanced subsampling. For either setting of the subsample size, we observe a more significant superiority of the balanced subsamples and a decreasing

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

trend of MSEs and WSPEs as the full data size increases. This is because the balanced subsamples have reduced correlation and enhanced combinatorial orthogonality between predictors, which helps reduce the collinearity between predictors and therefore allows a more accurate estimate of parameters. Notably, a balanced subsample with $n = 500$ exhibits comparable accuracy in parameter estimation and worst-case prediction when compared with UNI and IBOSS subsamples with $n = 2000$, which again demonstrates the significant savings achieved by a balanced subsample in observing the response.

5. Real data application

We consider the application to an online store offering clothing for pregnant women. The data are from five months of 2008 and include, among others, product category (4 levels), product code (217 levels), color (14 levels), model photography (2 levels), location of the photo on the page (6 levels), page number (5 levels), country of origin of the IP address of customers clicking the page (47 levels), month (5 levels), and product price in US dollars (continuous). The data contain more predictors to study the behavior patterns of customers. We are only using the above predictors to predict the product price and demonstrate the superiority of balanced sub-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

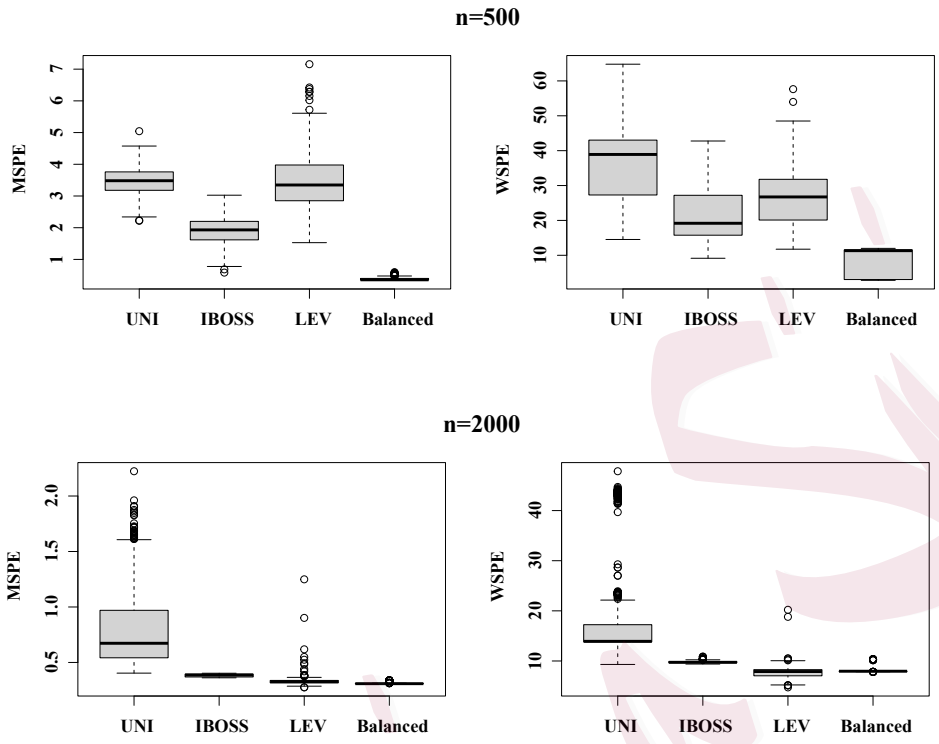

Figure 7: The MSPEs and WSPEs over the full data for different subsampling methods with $n = 500$ (top) and 2000 (bottom).

samples. Further information on the dataset can be found in Łapczyński and Białowąs (2013).

The full dataset has $n = 165,474$ data points. All predictors are categorical and are coded to dummy variables, which results in 293 dummy variables in total (intercept included). The response variable, product price, is accessible for the full data, facilitating the observation of responses for each subsample and allowing for comparison of their predictive performance.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

We consider subsample sizes $n = 500$ and 2000 and select subsamples from the data with different methods. Considering that some of the dummy variables may not be significant in real-world data, we use LASSO regression (Tibshirani, 1996) to select important ones and train a predictive model. On each subsample, a LASSO regression model is trained via the R package "glmnet" (Friedman et al., 2010) with the parameter λ selected by cross-validation. The trained model is then used to predict the product price over the full data. We expect that a balanced subsample should outperform other subsamples in a penalized regression due to its linear nature. To investigate this, we do 500 repetitions of this process and plot the MSPE (mean squared prediction error) and WSPE of the trained models over the full data in Figure 7. Balanced subsamples consistently yield superior predictions compared to other subsamples. In particular, UNI subsamples exhibit the poorest predictive performance of the full dataset. Their selection of markedly different points and subsequent divergent predictions result in unstable performance. In contrast, IBOSS and LEV subsamples demonstrate greater stability and improved performance due to their more balanced predictors. Overall, balanced subsamples consistently outperform others across both settings of subsample size.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

6. Discussion

In this paper, we propose the balanced subsampling method for big data with categorical predictors. The selected subsample achieves a balance among the predictor levels, maximizing the overall information provided by the subsample. A balanced subsample is typically nonsingular and allows more accurate parameter estimation and prediction. Simulations and a real-world application confirm the improved performance of the subsample selected by the balanced subsampling over other available subsampling methods. A balanced subsample offers substantial cost savings when observing the response. Although this paper assumes binary dummy variables for coding the categorical predictors, all theoretical results work if any nonsingular coding system (for example, an orthogonal coding framework such as the orthogonal polynomial) is used for coding the predictors. The superiority of a balanced subsample does not depend on the coding system.

We adopt Algorithm 1 to minimize $f(X_s)$ (alternatively, $\sum_{1\le i<l\le n}[\delta(x_i^*, x_l^*)]^2$) due to its straightforward implementation and efficiency. A primary drawback of Algorithm 1 is its lack of guaranteed optimization of $f(X_s)$. If an optimal solution is imperative, various optimization algorithms, such as the simulated annealing algorithm (Morris and Mitchell, 1995), can be employed. Convergence is ensured for these algorithms with sufficient iter-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

ations. Nonetheless, when dealing with extensive datasets, computational challenges may impede the efficacy of such algorithms. Although the current Algorithm 1 does not guarantee the minimization of $f(X_s)$, our extensive numerical results are strong evidence that the algorithm tends to efficiently produce a dramatically improved subsample relative to those found by other methods.

Balanced subsampling can be combined with robust regression, such as the S estimator (Rousseeuw and Yohai, 1984), to improve the robustness of the trained model to possible outliers. Training the estimator involves repeatedly selecting small and nonsingular subsamples from the full data, which, as discussed in this paper and in (Koller and Stahel, 2017), is infeasible via simple random sampling. The randomness and nonsingularity of balanced subsamples make them applicable to training such estimators, although their performance for this purpose requires further study.

Balanced subsampling can be extended to accommodate numerical or mixed-type predictors. To do so, we may first discretize the numerical predictors and then apply Algorithm 1. The selected subsample will cover the region of the full data evenly and uniformly, thereby fostering a fair study and enabling robust predictions.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

Supplementary Material

The online supplementary material provides proofs of the theoretical results and discusses the computational complexity of the proposed algorithm.

Acknowledgments

Wang is supported by the U.S. National Science Foundation (DMS-2413741) and the Central Indiana Corporate Partnership AnalytiXIN Initiative.

References

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

Ai, M., J. Yu, H. Zhang, and H. Wang (2021). Optimal subsampling algorithms for big data regressions. *Statistica Sinica 31*(1), 749–772.

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

Cheng, C.-S. (1980). Orthogonal arrays with variable numbers of symbols. *The Annals of Statistics 8*(2), 447–453.

Cheng, Q., H. Wang, and M. Yang (2020). Information-based optimal

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

subdata selection for big data logistic regression. *Journal of Statistical Planning and Inference 209*, 112–122.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software 33*(1), 1–22.

He, L. and Y. Hung (2022). Gaussian process prediction using design-based subsampling. *Statistica Sinica 32*(2), 1165–1186.

Hedayat, A., N. Sloane, and J. Stufken (1999). *Orthogonal arrays: theory and applications*. Springer, New York.

Huang, D., R. Li, and H. Wang (2014). Feature screening for ultrahigh dimensional categorical data with applications. *Journal of Business & Economic Statistics 32*(2), 237–244.

Johnson, A. C., C. G. Ethun, Y. Liu, A. G. Lopez-Aguiar, T. B. Tran, G. Poultsides, V. Grignol, J. H. Howard, M. Bedi, T. C. Gamblin, et al. (2018). Studying a rare disease using multi-institutional research collaborations vs big data: Where lies the truth? *Journal of the American College of Surgeons 227*(3), 357–366.

Kanda, Y. (2013). Investigation of the freely available easy-to-use software

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

'EZR' for medical statistics. *Bone marrow transplantation 48*(3), 452–458.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society, Series B 21*(2), 272–304.

Koller, M. and W. A. Stahel (2017). Nonsingular subsampling for regression s estimators with categorical predictors. *Computational Statistics 32*(2), 631–646.

Łapczyński, M. and S. Białowąs (2013). Discovering patterns of users' behaviour in an e-shop-comparison of consumer buying behaviours in poland and other european countries. *Studia Ekonomiczne 151*, 144–153.

Ma, P., M. W. Mahoney, and B. Yu (2015). A statistical perspective on algorithmic leveraging. *The Journal of Machine Learning Research 16*(1), 861–911.

Ma, P. and X. Sun (2015). Leveraging for big data regression. *Wiley Interdisciplinary Reviews: Computational Statistics 7*(1), 70–76.

Mak, S. and V. R. Joseph (2018). Support points. *The Annals of Statistics 46*(6A), 2562–2592.

Maronna, R. A. and V. J. Yohai (2000). Robust regression with both

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

continuous and categorical predictors. *Journal of Statistical Planning and Inference 89*(1-2), 197–214.

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

Meng, C., X. Zhang, J. Zhang, W. Zhong, and P. Ma (2020). More efficient approximation of smoothing splines via space-filling basis selection. *Biometrika 107*(3), 723–735.

Morris, M. D. and T. J. Mitchell (1995). Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference 43*(3), 381–402.

Rousseeuw, P. and V. Yohai (1984). Robust regression by means of s-estimators. In *Robust and Nonlinear Time Series Analysis: Proceedings of a Workshop Organized by the Sonderforschungsbereich 123 "Stochastische Mathematische Modelle", Heidelberg 1983*, pp. 256–272. Springer.

Shi, C. and B. Tang (2021). Model-robust subdata selection for big data. *Journal of Statistical Theory and Practice 15*(4), 1–17.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

Song, D., N. M. Xi, J. J. Li, and L. Wang (2022). scsampler: fast diversity-preserving subsampling of large-scale single-cell transcriptomic data. *Bioinformatics 38*(11), 3126–3127.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*(1), 267–288.

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

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

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

Wang, L., J. Elmstedt, W. K. Wong, and H. Xu (2021). Orthogonal subsampling for big data linear regression. *Annals of Applied Statistics 15*(3), 1273–1290.

Wang, Y., A. W. Yu, and A. Singh (2017). On computationally tractable

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

selection of experiments in measurement-constrained regression models. *The Journal of Machine Learning Research 18*(1), 5238–5278.

Yu, Y., S.-K. Chao, and G. Cheng (2022). Distributed bootstrap for simultaneous inference under high dimensionality. *Journal of Machine Learning Research 23*(195), 1–77.

Zhang, Y., L. Wang, X. Zhang, and H. Wang (2024). Independence-encouraging subsampling for nonparametric additive models. *Journal of Computational and Graphical Statistics 33*(4), 1424–1433.

Zhu, J., L. Wang, and F. Sun (2024). Group-orthogonal subsampling for hierarchical data based on linear mixed models. *Journal of Computational and Graphical Statistics 33*(3), 1037–1046.

Zuccolotto, P., M. Manisera, and M. Sandri (2018). Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions. *International journal of sports science & coaching 13*(4), 569–589.

Lin Wang

Department of Statistics

Purdue University, West Lafayette, IN 47907, USA

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0434

E-mail: linwang@purdue.edu