Statistica Sinica **30** (2020), 1135-1154
doi:https://doi.org/10.5705/ss.202017.0384

GENERALIZED REGRESSION ESTIMATORS WITH HIGH-DIMENSIONAL COVARIATES

Tram Ta[1], Jun Shao[1,2], Quefeng Li[3] and Lei Wang[4]

[1]*University of Wisconsin, Madison* [2]*East China Normal University*
[3]*University of North Carolina, Chapel Hill and* [4]*Nankai University*

Abstract: Data from a large number of covariates with known population totals are frequently observed in survey studies. These auxiliary variables contain valuable information that can be incorporated into an estimation of the population total of a survey variable in order to improve the estimation precision. We consider a generalized regression estimator formulated under a model-assisted framework, in which a regression model is used for the available covariates, and the estimator retains the basic design-based properties. The generalized regression estimator is shown to improve the efficiency of the design-based Horvitz–Thompson estimator when the number of covariates is fixed. We investigate the performance of the generalized regression estimator when the number of covariates p is allowed to diverge as the sample size n increases. We examine two approaches. First, the model parameter is estimated using the weighted least squares method when $p < n$. Second, the Lasso method is employed when the model parameter is sparse. We show that under an assisted model and certain conditions on the joint distribution of the covariates, as well as the divergence rates of n and p, the generalized regression estimator is asymptotically more efficient than the Horvitz–Thompson estimator, and is robust against a model misspecification. We also study the consistency of the variance estimation for the generalized regression estimator. Our theoretical results are corroborated by simulation studies and an example.

Key words and phrases: Asymptotic efficiency, auxiliary information, high dimension, Lasso, model-assisted, survey sampling.

1. Introduction

In many survey studies, in addition to the observed data from a study variable and related covariates, auxiliary information—for instance from administrative records or results from previous surveys—is available in the form of covariate population totals. This information can be used under the model-assisted framework to improve the precision of the Horvitz–Thompson estimator, a well-known design-based estimator of the total or mean of the survey variable (Cassel, Särndal and Wretman (1977); Särndal, Swensson, and Wretman (2003)).

In this framework, a model is adopted to reduce the estimation variability by utilizing the auxiliary information from the covariates related to the main study variable. Because the model's role is only to assist in the estimation process, the constructed estimator is protected against a model misspecification in the sense that it is still asymptotically design-unbiased and normally distributed when the model is incorrect.

The generalized regression (GREG) estimator, first discussed in Cassel, Särndal and Wretman (1976), and studied extensively in Cassel, Särndal and Wretman (1977); Särndal (1980a,b); Särndal, Swensson, and Wretman (2003), is a popular estimator under the model-assisted framework. It includes a wide range of estimators, notably the ratio estimator and the classical regression estimator (Särndal (1980b)), and is constructed for many survey designs that allow arbitrary inclusion probabilities (Särndal, Swensson, and Wretman (2003)). A closely related estimator is the calibration estimator, which is asymptotically equivalent to the GREG estimator under certain assumptions (Deville and Särndal (1992)). Here, we estimate the population total or mean using the GREG estimator.

In traditional applications that consider a small or moderate number of covariates, the properties of the GREG estimator have been well studied; see, for example, Cassel, Särndal and Wretman (1977) and Särndal, Swensson, and Wretman (2003) for a good overview. A well-known characteristic of the GREG estimator is that when there is a linear regression model between the study variable and the covariates, the estimator is asymptotically more efficient than the Horvitz–Thompson estimator, which is based only on data from the study variable. Moreover, the gain in efficiency is not affected by the fact that the weighted least squares estimator (WLSE) instead of the true regression parameter is used in the GREG estimator.

However, with technological advances, it is now possible to collect data on a large number of covariates, which could even exceed the sample size. See Nascimento Silva and Skinner (1997) for examples of survey data with large numbers of covariates. For example, in the 1990 U.S. Census on law enforcement (`http://archive.ics.uci.edu/ml/datasets/communities+and+crime+unnormalized`), 101 covariates are recorded in a population of 2,195 communities. These covariates include the population of a community, mean people per household, percentage of population by race, median household income, number of people in each age, percentage of a household with a salary, farm, or self-employment income, etc. A complete list of these 101 covariates can be found in the Supplementary Material. Population totals of these covariates can be obtained from the census or administrative records. A more recent example is the electronic health record (Jha et al. (2009)), in which a large number of covariates are recorded for each patient, including demographic information, biometric information, medical records, historical medical test results and so on. In addition, population totals for many covariates are maintained by social and governmental organizations. Another example is that of considering covariate interactions and/or polynomial effects in a regression. In this case, even though the original number of covariates with a known population total is moderate, the number of covariates after adding interactions and/or polynomial terms could be very large (McConville et al. (2017)).

With high-dimensional covariate and auxiliary population information, we wish to know whether the GREG estimator based on the WLSE still improves the efficiency, and whether using a regularized regression estimator leads to a better GREG estimator. To answer these questions, we study the GREG estimator in a setting in which both the number of covariates p and the sample size n are allowed to diverge to infinity. Our first result concerns the GREG estimator based on the WLSE. We show that under a correct regression model and certain assumptions on the joint distribution of the covariates, the GREG estimator using the WLSE is asymptotically equivalent to the GREG estimator using the true regression parameter. Hence, it outperforms the Horvitz–Thompson estimator, as long as $p/n \to 0$. On the other hand, when p/n does not converge to zero, the GREG estimator using the WLSE may not be asymptotically more efficient than the Horvitz–Thompson estimator.

If there are only s of p covariates that are actually related to the study variable, where s diverges slower than the sample size n, although p may be comparable to, or even larger than n, a GREG estimator using a regularized regression estimator can be constructed. Dimension reduction has been studied in Cardot, Goga and Shehzad (2017), in which the authors considered a principal component analysis to reduce the covariate dimension prior to performing a calibration. This calibration approach can also be adopted for the GREG estimator. However, the asymptotic results of their GREG estimator are established under the condition $p^3/n \to 0$, which is much stronger than the condition $p/n \to 0$ for the GREG estimator based on the WLSE (our Theorem 1). As a result, when $p^3/n \to 0$, there is no strong motivation to consider the principal component regression estimator.

The WLSE is unavailable when $p > n$. This may occur in some small-area survey estimations and in economic and biological studies. The principal component calibration approach does not perform well in this high-dimensional problem either. We adopt the Lasso (Tibshirani (1996)) as a regularization method. The use of the Lasso for the GREG was proposed in McConville (2011) and McConville et al. (2017), but they studied the empirical and theoretical properties for fixed p only. Under some conditions on the divergence rates of s and p, we show that the GREG estimator constructed using the Lasso is asymptotically equivalent to the GREG using the true regression parameter when the regression model is correct. In addition, this GREG estimator still possesses asymptotically design-based properties when the assumed model is misspecified. We also study variance estimation for the GREG with the Lasso.

We present simulation results to study how much the Horvitz–Thompson estimator can be improved by the GREG estimators, observe the effect of p on the efficiency gain, and compare the relative performance between the GREG estimators using the WLSE and the Lasso estimator. All technical proofs are available in the Supplementary Material.

2. The Generalized Regression Estimator

Consider a finite population U that consists of N units, for $i = 1, \ldots, N$. For unit i, let y_i be the value of the study variable, and x_i be the p-dimensional vector of covariates. We estimate the finite population total $Y = \sum_{i \in U} y_i$ using data from a sample S of size n selected from U, following a probability plan called sampling design. The value of (y_i, x_i) is observed for unit i in sample S. To estimate Y, Horvitz and Thompson (1952) introduced the following estimator:

$$\hat{Y}_{\text{ht}} = \sum_{i \in S} \frac{y_i}{\pi_i}, \tag{2.1}$$

where $\pi_i > 0$ is the inclusion probability for unit i, which can be calculated from the sampling design, and may depend on some components of x_i. The population mean Y/N can be estimated using $\hat{Y}_{\text{ht}}/N$ or $\hat{Y}_{\text{ht}}/\sum_{i \in S} \pi_i^{-1}$. Under the noninformative sampling assumption, that is, π_i is a known function of x_i, but does not depend on y_i, the Horvitz–Thompson estimator in (2.1) is design–unbiased with respect to the random selection of S from U. Throughout this paper, we assume noninformative sampling and that $\hat{Y}_{\text{ht}} - Y$ is asymptotically normal as $n \to \infty$, under the given sampling design with some conditions; see, for example, Krewski and Rao (1981); Bickel and Freedman (1984); Fuller (2009). When considering asymptotic properties, the finite population is viewed as a member of a sequence of finite populations, with sizes increasing to infinity. Then, the sample is a member of a sequence of samples, with sample sizes increasing to infinity. To abbreviate, we simply write $n \to \infty$.

In addition to the observed x_i, for all $i \in S$, the finite-population total vector $X = \sum_{i \in U} x_i$ is often known in many studies. To use the information provided by the covariates, we consider $\{(x_i, y_i) : i \in U\}$ as realizations from a super-population model. In some applications, it may not be practical to impose an assumption on the entire population U. It may be more realistic to assume that U can be divided into sub-populations, such that an assumption can be made for units within each sub-population. These sub-populations, such as strata or post-strata (Valliant (1993)), are constructed so that (x_i, y_i) in each sub-population is assumed to be unconditionally independent and identically distributed (i.i.d.). Because an estimator of the sub-population total will be constructed using data within each sub-population, and the estimator of the overall population total is the sum of the sub-population total estimators, in what follows, we ignore the sub-populations for notation simplicity; that is, for all $i \in U$, we assume that

$$y_i = \mu + \beta^T x_i + \epsilon_i, \tag{2.2}$$

where μ and β are unknown parameters, a^T is the transpose of a vector a, x_i is an i.i.d. random vector of covariates with an unknown positive-definite covariance matrix Σ, ϵ_i is an independent random variable with mean zero and unknown variance σ_ϵ^2, and x_i is independent of ϵ_i. After the sample S is selected from U, $\{(x_i, y_i), i \in S\}$ are observed.

To take advantage of the available covariate information under model (2.2), Cassel, Särndal and Wretman (1976, 1977) proposed the following GREG estimator of the total Y:

$$\hat{Y}_{\mathrm{gr}} = \hat{Y}_{\mathrm{ht}} + \hat{\beta}^T (X - \hat{X}_{\mathrm{ht}}), \tag{2.3}$$

where $X = \sum_{i \in U} x_i$ is the known finite population total of x_i, $\hat{X}_{\mathrm{ht}}$ is the Horvitz–Thompson estimator of X, defined as in (2.1); that is, $\hat{X}_{\mathrm{ht}} = \sum_{i \in S} x_i / \pi_i$, and $\hat{\beta}$ is an estimator of β in (2.2) based on (y_i, x_i), for $i \in S$. The GREG estimator in (2.3) is the sum of the Horvitz–Thompson estimator $\hat{Y}_{\mathrm{ht}}$ and an adjustment $\hat{\beta}^T (X - \hat{X}_{\mathrm{ht}})$ that is used to increase the efficiency.

To study the properties of the GREG estimator, we first consider an artificial situation where β in (2.2) is known, such that $\hat{\beta} = \beta$ and the estimator in (2.3) is denoted as

$$\hat{Y}_{\mathrm{gr}}^* = \hat{Y}_{\mathrm{ht}} + \beta^T (X - \hat{X}_{\mathrm{ht}}). \tag{2.4}$$

Because $\hat{X}_{\rm ht}$ is the Horvitz–Thompson estimator of X, $\hat{Y}^*_{\rm gr}$ is a design-unbiased estimator of Y, even if model (2.2) is wrong or β is a wrong value. If model (2.2) is correct, then regardless of how large the dimension p is, the variance of $\hat{Y}^*_{\rm gr}$ is smaller than the variance of $\hat{Y}_{\rm ht}$, unless $\beta = 0$, where the variance is calculated with respect to the sampling and model. For this reason, the GREG estimator is referred to as a model-assisted estimator.

In practice, β is unknown; therefore, the GREG estimator, which involves $\hat{\beta}$, is asymptotically design-unbiased and normally distributed, as long as $\hat{\beta}$ does not diverge to infinity. In a traditional setting, the covariate dimension is fixed, in the sense that p does not change as $n \to \infty$. Then, under model (2.2), the GREG estimator is asymptotically more efficient than the Horvitz–Thompson estimator, as long as $\hat{\beta}$ is consistent, because

$$\begin{aligned}\hat{Y}_{\rm gr} - Y &= \hat{Y}_{\rm ht} - Y + \hat{\beta}^T(X - \hat{X}_{\rm ht})\\ &= \hat{Y}_{\rm ht} - Y + \beta^T(X - \hat{X}_{\rm ht}) + (\hat{\beta} - \beta)^T(X - \hat{X}_{\rm ht})\\ &= \hat{Y}^*_{\rm gr} - Y + o_p(1)(X - \hat{X}_{\rm ht})\\ &= \hat{Y}^*_{\rm gr} - Y + o_p(1)(\hat{Y}^*_{\rm gr} - Y),\end{aligned}$$

where $o_p(1)$ denotes a quantity converging to zero in probability. This implies that in a low-dimensional setting, $\hat{Y}_{\rm gr}$ and $\hat{Y}^*_{\rm gr}$ in (2.4) are asymptotically equivalent under model (2.2). Note that we do not need to worry about the efficiency of $\hat{\beta}$.

When p is fixed, $\hat{\beta}$ is typically the following WLSE of β under model (2.2),

$$\hat{\beta}_{\rm wls} = \left\{\sum_{i\in S} \frac{1}{\pi_i}\left(x_i - \frac{\hat{X}_{\rm ht}}{\hat{N}}\right)\left(x_i - \frac{\hat{X}_{\rm ht}}{\hat{N}}\right)^T\right\}^{-1} \sum_{i\in S} \frac{(x_i - \hat{x}_S)y_i}{\pi_i}, \qquad (2.5)$$

where $\hat{N} = \sum_{i\in S} \pi_i{}^{-1}$. The GREG estimator constructed using $\hat{\beta}_{\rm wls}$ is denoted by $\hat{Y}_{\rm gr\_wls}$. If model (2.2) is correct, $n^{1/2}(\hat{\beta}_{\rm wls} - \beta)$ is asymptotically normal with mean zero; thus,

$$\hat{Y}_{\rm gr\_wls} - Y = \hat{Y}^*_{\rm gr} - Y + O_p(n^{-1/2})(\hat{Y}^*_{\rm gr} - Y); \qquad (2.6)$$

that is, $\hat{Y}_{\rm gr\_wls}$ is asymptotically equivalent to $\hat{Y}^*_{\rm gr}$ up to an order of $n^{-1/2}$, where $O_p(a_n)$ denotes a sequence that is bounded in probability by $|a_n|$.

As discussed in the introduction, modern data are often high dimensional. When p is unbounded as $n \to \infty$, we examine whether $\hat{Y}_{\rm gr\_wls}$ is still asymptotically equivalent to $\hat{Y}^*_{\text{gr}}$ such that it improves $\hat{Y}_{\text{ht}}$. The answer is given in the following result.

Theorem 1. *Assume model* (2.2) *with $p < n$ and the following assumptions:*

(A1) $\max_{i\in U} \pi_i^{-1} = O(N/n)$.

(A2) $\sum_{i\in U}(\pi_i^{-1} - 1) \geq c(N^2/n)$ *for a constant $c > 0$, not depending on n and p.*

(A3) *The components of $\Sigma^{-1/2}x_i$ are i.i.d. and have finite fourth-order moments.*

Then, we have the following conclusions:

(a) *If $p/n \to 0$ as $n \to \infty$, then*

$$\hat{Y}_{\text{gr\_wls}} - Y = \hat{Y}^*_{\text{gr}} - Y + O_p\left\{\left(\frac{p}{n}\right)^{1/2}\right\}(\hat{Y}^*_{\text{gr}} - Y) \tag{2.7}$$

and, hence, $\hat{Y}_{\text{gr\_wls}}$ is asymptotically equivalent to $\hat{Y}^*_{\text{gr}}$.

(b) *If $p/n \to \gamma > 0$ as $n \to \infty$, then, in general, $\hat{Y}_{\text{gr\_wls}}$ is not asymptotically equivalent to $\hat{Y}^*_{\text{gr}}$.*

Assumptions (A1) and (A2) involve bounds on the inclusion probabilities. Assumption (A3) is used to obtain the limiting spectral distribution of the functionals of the design matrix. Using the arguments in Bai and Zhou (2008) and Xie (2013), the results in Theorem 1 can also be established if (A3) is replaced by

(A3′) $p^3/n \to \infty$ and $E(x_i^T\Sigma^{-1/2}B\Sigma^{-1/2}x_i - \text{tr}B)^2 = o(p^3/n)$, for any $p \times p$ deterministic matrix B with a bounded spectral norm.

Note that result (2.7) includes result (2.6) for the case of fixed p as a special case. Theorem 1 indicates that under model (2.2), if $p/n \to 0$, then $\hat{Y}_{\text{gr\_wls}}$ is asymptotically more efficient than $\hat{Y}_{\text{ht}}$, and is asymptotically equivalent to $\hat{Y}^*_{\text{gr}}$, which is based on the true β. The difference between $\hat{Y}_{\text{gr\_wls}}$ and $\hat{Y}^*_{\text{gr}}$ depends on the rate of convergence of p/n, as result (2.7) indicates. Thus, it is expected that the efficiency gain by the GREG estimation deteriorates as the rate of p/n increases, although there is no rigorous proof.

When $p/n \to \gamma > 0$, Theorem 1 shows that $\hat{Y}_{\text{gr\_wls}}$ may not be asymptotically equivalent to $\hat{Y}^*_{\text{gr}}$. Consequently, if p diverges at a rate the same as or close to n, then the performance of $\hat{Y}_{\mathrm{gr\_wls}}$ can be even worse than $\hat{Y}_{\mathrm{ht}}$, even if model (2.2) is correct. In the next section, we consider an improvement of $\hat{Y}_{\mathrm{gr\_wls}}$ when the true regression coefficient β is sparse, in the sense that many of its components are zero, although p can still be large.

3. The Lasso Generalized Regression Estimator

Although data today contain many covariates, it is often true that only a few of these are actually related to the study variable. In model (2.2), among p covariates, only s have nonzero regression coefficients (i.e., β-components) and s is fixed or diverges much slower than p. We require a sparse estimator of β when β is sparse because retaining the extraneous variables serves no purpose, but it does increase the variability and model complexity. The WLSE $\hat{\beta}_{\mathrm{wls}}$, however, is not sparse, regardless of whether or not β is sparse. Therefore, we consider the Lasso estimator, denoted by $\hat{\beta}_{\ell_1}$. The GREG estimator in (2.3) using $\hat{\beta} = \hat{\beta}_{\ell_1}$, denoted as $\hat{Y}_{\mathrm{gr\_}\ell_1}$, is well defined, even when $p > n$. In this section, we study the asymptotic properties of $\hat{Y}_{\mathrm{gr\_}\ell_1}$, and show that it improves $\hat{Y}_{\mathrm{gr\_wls}}$ and the Horvitz–Thompson estimator $\hat{Y}_{\mathrm{ht}}$. Furthermore, it is asymptotically equivalent to $\hat{Y}^*_{\mathrm{gr}}$, under some conditions on the sparsity and the diverging rate of p that allows $p/n \to \infty$. It is also design-based robust against a model misspecification.

We use the notation from Section 2. The Lasso estimator $\hat{\beta}_{\ell_1}$ is a solution to the ℓ_1-penalized weighted least squares minimization problem:

$$\min_{b \in R^p} \left[\frac{1}{2n} \sum_{i \in S} \frac{\{y_i - b^T(x_i - \hat{X}_{\mathrm{ht}}/\hat{N})\}^2}{\pi_i} + \lambda \|b\|_1 \right], \tag{3.1}$$

where $\|b\|_1$ is the usual ℓ_1-norm of a vector $b \in R^p$, and $\lambda \geq 0$ is a penalty parameter that may depend on n. The ℓ_1-norm penalty is applied to shrink the estimated coefficients and to select variables, simultaneously. Note that the WLSE $\hat{\beta}_{\mathrm{wls}}$ is a special case of $\hat{\beta}_{\ell_1}$, defined as the solution to (3.1) with $\lambda = 0$.

There is a considerable body of literature devoted to studying the conditions on the covariates x_i in order to guarantee certain good oracle properties of $\hat{\beta}_{\ell_1}$ in terms of prediction or estimation accuracy, and in terms of variable selection consistency. Here, well-known conditions include the restricted null space property (Donoho and Huo (2001)), restricted isometry property (Candes and Tao (2005, 2007)), restricted eigenvalue condition (Bickel, Ritov and Tsybakov (2009)), and irrepresentable condition (Zhao and Yu (2006)). The last condition is quite strong, and is required only if model-selection consistency is of interest.

The restricted null space property has been shown to successfully recover the signal in a noiseless setting; that is, $\epsilon_i = 0$ for all i in (2.2). When ϵ_i in (2.2) is not degenerated, the restricted isometry property is proved to be sufficient for bounding the estimation error.

A relatively weaker condition is the restricted eigenvalue (RE) condition introduced in Bickel, Ritov and Tsybakov (2009), which holds for an $n \times p$ matrix A if

$$\frac{1}{K_{(l,k,A)}} = \min_{\substack{J \subset \{1,\ldots,p\} \\ |J| \le l}} \quad \min_{\substack{v \neq 0 \\ \|v_{-J}\|_1 \le k \|v_J\|_1}} \frac{\|Av\|_2}{\|v_J\|_2} > 0, \tag{3.2}$$

where v_J is a sub-vector of v with components indexed by elements in $J \subset \{1,\ldots,p\}$, v_{-J} is a sub-vector of v with components not in v_J, $|J|$ is the number of elements in J, $\|\cdot\|_2$ is the usual ℓ_2-norm, and l and k are constants. The condition is denoted as $RE(l,k,A)$.

The restricted eigenvalue condition requires that A be positive-definite on a restricted set of vectors in the cone

$$\mathcal{C}_{(l,k)} = \{v \in R^p : \exists J \subset \{1,\ldots,p\}, |J| \le l, \|v_{-J}\|_1 \le k\|v_J\|_1\}, \tag{3.3}$$

hence the name restricted eigenvalue condition. It is shown in the Supplementary Material that the estimation error $\hat{\beta}_{\ell_1} - \beta$ belongs to the cone $\mathcal{C}_{(s,3)}$, that is, $\|(\hat{\beta}_{\ell_1} - \beta)_{-\mathcal{S}}\|_1 \le 3\|(\hat{\beta}_{\ell_1} - \beta)_{\mathcal{S}}\|_1$, where $\mathcal{S}$ contains the indices of all nonzero components of β and $s = |\mathcal{S}|$.

Condition (3.2) was first assumed in Bickel, Ritov and Tsybakov (2009) on a deterministic design matrix to establish a bound on the estimation loss of the signal for the Lasso estimator and the Dantzig selector. Rudelson and Zhou (2013) showed that with high probability and certain conditions, the RE condition holds for a large class of random matrices, including matrices with uniformly bounded entries, and those whose rows follow a sub-Gaussian distribution. In this study, the covariates x_i under the model-assisted framework are random vectors, distributed according to the super-population model. We consider a random design matrix $\mathbf{X}$, whose ith row is x_i, for $i \in S$. If x_i follows a sub-Gaussian distribution and Σ is positive-definite, then under certain assumptions, condition (3.2) holds for $A = \mathbf{X}/n^{1/2}$ with high probability (Rudelson and Zhou (2013)). The performance of $\hat{Y}_{\text{gr\_}\ell_1}$ is stated in the following theorem.

Theorem 2. *Assume* (A1)–(A2) *and the following assumptions:*

(A4) ϵ_i *and* x_i *independently follow sub-Gaussian distributions, with scale fac-*

tors τ and ν, respectively; that is, $E\{\exp(u\epsilon_i)\} \leq \exp(\tau^2 u^2/2)$, *for any real-valued* u, *and* $E\{\exp(t^T x_i)\} \leq \exp(\nu^2 t^T t/2)$ *for any* p*-dimensional vector* t.

(A5) *There exist constants* b_0, b_1, b_2, b_3 *not depending on* n *and* p, *such that* $n \geq b_1 r \log(b_2 p/r)$, *for all* $n \geq b_0$, *where* $r = \min\{s + b_3 s M^2 K^2_{(s,9,\Sigma^{1/2})}, p\}$,

$$M = \max_j \|\Sigma^{1/2} e_j\|_2,$$

and $e_j = (0, \ldots, 1, \ldots, 0)$, *for* $j = 1, \ldots, p$, *form the standard basis of* R^p.

(A6) *The tuning parameter* λ *in* (3.1) *is* $d\tau M(n^{-1}\log p)^{1/2}$ *for a constant* $d \geq 8$.

(i) *If model* (2.2) *holds, then*

$$\|\hat{\beta}_{\ell_1} - \beta\|_1 = O_p\left\{s(n^{-1}\log p)^{1/2}\ MK^2_{(s,3,\Sigma^{1/2})}\right\}, \tag{3.4}$$

and

$$\hat{Y}_{\text{gr\_}\ell_1} - Y = \hat{Y}^*_{\text{gr}} - Y + O_p\Big\{n^{-1/2}\ s\log p\ MK^2_{(s,3,\Sigma^{1/2})}\Big\}(\hat{Y}^*_{\text{gr}} - Y). \tag{3.5}$$

(ii) *If model* (2.2) *is wrong, and* (A4) *holds with* ϵ_i *replaced by* $y_i - x_i^T\beta$, *where* β *is defined as* $\beta = \Sigma^{-1}E(x_1 y_1)$, *then* (3.4) *still holds and*

$$\hat{Y}_{\text{gr\_}\ell_1} - Y = \hat{Y}_{\text{ht}} - Y + \beta^T(X - \hat{X}_{\text{ht}}) + O_p\left\{Nsn^{-1}\log p\ MK^2_{(s,3,\Sigma^{1/2})}\right\}.$$

The result on the estimation loss $\|\hat{\beta}_{\ell_1} - \beta\|_1$ was first established in Bickel, Ritov and Tsybakov (2009) for deterministic x_i, where the RE condition was imposed on the design matrix $\mathbf{X}$. Zhou (2009) showed that an estimation loss with a similar order to that of (3.4) holds when the rows of the random matrix $\mathbf{X}$ follow a sub-Gaussian distribution with a covariance matrix Σ that satisfies the RE condition $RE(s, 3, \Sigma^{1/2})$. The lower bound of the sample size n in Zhou (2009), however, depends on a quantity $\rho(s)$, which is defined as the maximum eigenvalue of Σ, restricted to sparse vectors with at most s nonzero components. We instead make a similar assumption (A5) to that in Rudelson and Zhou (2013), in which the lower bound of n does not depend on $\rho(s)$, but instead a slightly stronger assumption $RE(s, 9, \Sigma^{1/2})$ is used.

Theorem 2 indicates that $\hat{Y}_{\text{gr\_}\ell_1}$ is asymptotically equivalent to $\hat{Y}^*_{\text{gr}}$, even when the working model (2.2) is misspecified, as long as $n^{-1/2}s\log p\ MK^2_{(s,3,\Sigma^{1/2})} \to 0$, which is reasonable because $s\log p$ can be much smaller than n. Hence, $\hat{Y}_{\text{gr\_}\ell_1}$

asymptotically outperforms $\hat{Y}_{\mathrm{ht}}$ if model (2.2) holds. When model (2.2) is misspecified, both $\hat{Y}_{\mathrm{ht}}$ and $\hat{Y}_{\mathrm{gr\_}\ell_1}$ are design-based asymptotically valid, and there is no definite conclusion on the relative performance of $\hat{Y}_{\mathrm{ht}}$ and $\hat{Y}_{\mathrm{gr\_}\ell_1}$, although $\hat{Y}_{\mathrm{gr\_}\ell_1}$ is expected to be better than $\hat{Y}_{\mathrm{ht}}$ if (2.2) is nearly correct. See the simulation results in Section 4.1.

In the study of Cardot, Goga and Shehzad (2017), who used a calibration based on the principal components of the covariates, the number of covariates p was restrictively assumed to satisfy $p^3r^3/n \to 0$ to establish the consistency of the calibration estimator, where r is the number of selected principal components. This condition is much stronger than $p/n \to 0$, under which the GREG estimator with the WLSE is asymptotically equivalent to $\hat{Y}^*_{\mathrm{gr}}$ (Theorem 1). If the covariate x_i is observed for every unit i in the population U, then the assumption $p^3r^3/n \to 0$ can be relaxed to $r^3/n \to 0$. However, such a result has limited application because complete covariate information for the population is not usually available, especially when x_i has a high dimension.

To assess the estimation variability or to make an inference about Y, we need a variance estimator for $\hat{Y}_{\mathrm{gr\_}\ell_1}$. First, consider $\hat{Y}^*_{\mathrm{gr}}$, given by (2.4). If β is treated as known, then a classical variance estimator for $\hat{Y}^*_{\mathrm{gr}}$ is

$$v(\beta) = \sum_{i\in S}\sum_{j\in S} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} \; \frac{y_i - x_i^T\beta}{\pi_i} \; \frac{y_j - x_j^T\beta}{\pi_j}, \tag{3.6}$$

where π_{ij} is the inclusion probability of units i and j in the sample S, for $i \neq j$. When β is unknown, it is substituted by the same estimator $\hat{\beta}$ used in the GREG. In the traditional case where p is fixed, $v(\hat{\beta})$ is defined as (3.6), with β replaced by a consistent $\hat{\beta}$, and is consistent for the variance of $\hat{Y}_{\mathrm{gr}}$ as $n \to \infty$. The next result shows that this remains true when β is estimated using the Lasso method.

Theorem 3. *Assume model* (2.2)*, the conditions of Theorem* 2, $\max_{i,j} |1 - \pi_i\pi_j/\pi_{ij}| = O(n^{-1})$*, and the right-hand side of* (3.4) *converges to zero. Then, the variance estimator* $v(\hat{\beta}_{\ell_1})$ *defined as* (3.6)*, is consistent in the sense that* $v(\hat{\beta}_{\ell_1})/\mathrm{var}(\hat{Y}_{\mathrm{gr\_}\ell_1}) \to 1$ *in probability.*

4. Simulation Studies

4.1. Results based on simple random sampling

In the first simulation study, we consider simple random sampling without replacement (SRSWO). Finite populations of size $N = 10^5$ were generated from three super-population models, described as follows. Covariate vectors x_i were generated from a multivariate normal distribution $N(0, \Sigma)$, with

$$\Sigma = \begin{bmatrix} B & 0 \\ 0 & I_{p/2} \end{bmatrix},$$

where $I_{p/2}$ is the identity matrix of order $p/2$, and B is a $p/2 \times p/2$ symmetric matrix, with diagonal entries equal to one, and every off-diagonal entry equal to zero with probability 0.8, and equal to the value of a random variable following a uniform distribution on (0,1) with probability 0.2. A small positive quantity was added to the diagonal of B to ensure its positive-definiteness.

Different values of p were considered in each model to observe the effect of the number of covariates on the estimators' performance. The following three super-population models were considered:

Model M1: $y_i = \mu + x_i^T\beta + \epsilon_i$, as in (2.2), with $s = p^{1/2}, \beta = (2, \dots, 2, 0, \dots, 0)$, where ϵ_i are i.i.d. $N(0, 1)$, $\mu = \sum_j \beta_j$, and β_j is the jth component of β. In this model, the first $p^{1/2}$ (with rounding) components of β are set to two, and all other entries are zero. The number of relevant variables in this model, therefore, increases as the dimension increases.

Model M2: the same as M1, but the first 10 entries of β are $1, 2, 3, 4, 5,\ 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2$, and all other entries of β are zeros. Therefore, the underlying model has dimension $s = 10$, although p increases. Because the nonzero components of β take different values, the corresponding covariates are correlated with the variable y with different strengths.

Model M3: $y_i = \mu + \beta_1(x_i^{(1)})^2 + \beta_2(x_i^{(2)})^2 + \cdots + \beta_p(x_i^{(p)})^2 + \epsilon_i$, where $x_i^{(j)}$ is the jth component of x_i, $s = 10$, β is the same as that in Model 2, ϵ_i are i.i.d. $N(0, 1)$, and $\mu = \sum_j \beta_j$. The parameter β, however, is still estimated under the assumed model (2.2) in order to investigate the consequences of a model misspecification.

From each finite population generated according to the models, 500 different SRSWO samples of size $n = 500$ were selected. For each sample, $\hat{Y}_{\text{ht}}$, $\hat{Y}_{\text{gr\_wls}}$, $\hat{Y}_{\text{gr\_}\ell_1}$, and the optimal estimator $\hat{Y}_{\text{gr\_opt}}$ proposed in Berger, Tirari and Tille (2003) were computed. A 10-fold cross-validation was used to select the tuning parameter λ in the minimization problem (3.1), where we chose the one with the smallest mean squared error Friedman, Hastie and Tibshirani (2010). Based on the 500 simulations, the standard deviation (SD) of each estimator $\hat{Y}$ and the

Table 1. Standard deviation (SD) and mean squared error (MSE) ratio for $\hat{Y}_{\text{ht}}$, $\hat{Y}_{\text{gr\_wls}}$, $\hat{Y}_{\text{gr\_opt}}$, and $\hat{Y}_{\text{gr\_}\ell_1}$ based on SRSWO.

		SD						
p	s	$\hat{Y}_{\text{ht}}$	$\hat{Y}_{\text{gr\_wls}}$	$\hat{Y}_{\text{gr\_opt}}$	$\hat{Y}_{\text{gr\_}\ell_1}$	$\frac{\text{mse}(\hat{Y}_{\text{ht}})}{\text{mse}(\hat{Y}_{\text{gr\_}\ell_1})}$	$\frac{\text{mse}(\hat{Y}_{\text{gr\_wls}})}{\text{mse}(\hat{Y}_{\text{gr\_}\ell_1})}$	$\frac{\text{mse}(\hat{Y}_{\text{gr\_opt}})}{\text{mse}(\hat{Y}_{\text{gr\_}\ell_1})}$
				Model M1				
10	3	17,099	4,495	4,489	4,477	14.6	1.0	1.0
50	7	29,205	5,559	4,760	4,675	39.0	1.4	1.1
100	10	31,759	7,982	4,988	4,931	41.5	2.6	1.1
200	14	40,717	16,832	5,715	5,019	65.8	11.3	1.6
300	17	44,378	26,973	7,202	5,288	70.4	27.0	2.0
400	20	47,563	38,079	10,121	5,349	79.1	53.2	4.3
				Model M2				
10	10	36,939	4,604	4,521	4,525	63.8	1.0	1.0
50	10	40,344	6,240	4,730	4,689	74.4	1.8	1.0
100	10	33,658	8,320	5,013	4,755	50.1	3.1	1.1
200	10	35,033	14,837	5,849	4,771	53.9	10.2	1.5
300	10	35,740	21,874	7,304	4,803	55.4	21.2	2.3
400	10	33,369	26,788	10,044	4,720	52.5	32.5	4.5
				Model M3				
10	10	88,030	51,215	51,227	51313	2.7	1.1	1.1
50	10	83,894	51,969	51,748	50186	2.7	1.1	1.1
100	10	87,616	54,823	53,962	49398	3.1	1.2	1.2
200	10	86,742	62,090	60,671	49839	3.0	1.5	1.5
300	10	86,010	76,002	67,390	49760	3.0	2.3	1.8
400	10	87,531	106,794	77,554	49498	3.1	4.6	2.4

ratio of mse($\hat{Y}$) for pairs of estimators, where mse($\hat{Y}$) is the mean squared error of $\hat{Y}$, are reported in Table 1 for all three models, M1–M3. All estimators $\hat{Y}_{\text{ht}}$, $\hat{Y}_{\text{gr\_wls}}$, $\hat{Y}_{\text{gr\_opt}}$, and $\hat{Y}_{\text{gr\_}\ell_1}$ have negligible biases of less than 1% of Y and, hence, are not shown in the table.

Based on the SD, the GREG estimators, which incorporate data from the covariates, were more efficient than the Horvitz–Thompson estimator in all but one case, namely, where p is large ($p = 400$), the model is misspecified, and the GREG estimator is based on the WLSE. Under models M1 and M2, the mean squared error of the Horvitz–Thompson estimator was reduced 18 to 100 times as a result of using the auxiliary information. Under model M3, which is an incorrect model, the GREG estimators still outperformed the Horvitz–Thompson estimator in terms of efficiency in most cases, although the improvement was not as large as that observed in models M1–M2, because no auxiliary information was used correctly.

Similarly, based on the SD, not only was $\hat{Y}_{\text{gr\_}\ell_1}$ more efficient than $\hat{Y}_{\text{gr\_wls}}$, but its performance was also more consistent than that of $\hat{Y}_{\text{gr\_wls}}$ when the complexity of the model grows. For instance, under model M2, in which s is fixed, while p increases, $\hat{Y}_{\text{gr\_wls}}$ deteriorates considerably. Table 1 shows that the ratio $\text{mse}(\hat{Y}_{\text{gr\_wls}})/\text{mse}(\hat{Y}_{\text{gr\_}\ell_1})$ is no smaller than one in all cases, and that the difference between the mean squared error ratios becomes more pronounced as p increases. This suggests that when p is large, using $\hat{Y}_{\text{gr\_}\ell_1}$ results in a larger efficiency gain than when using $\hat{Y}_{\text{gr\_wls}}$, even when $p < n$.

Furthermore, $\hat{Y}_{\text{gr\_}\ell_1}$ exhibits comparable performance to that of the optimal estimator $\hat{Y}_{\text{gr\_opt}}$ when the dimension $p \leq 50$, in terms of SD and MSE. However, when p is large, $\hat{Y}_{\text{gr\_}\ell_1}$ outperforms $\hat{Y}_{\text{gr\_opt}}$. This is because $\hat{Y}_{\text{gr\_opt}}$ is not regularized, which means it does not perform well when p is large, although it is still better than the unregularized $\hat{Y}_{\text{gr\_wls}}$.

4.2. Results based on probability proportional to size sampling

In the second simulation study, we considered an unequal probability sampling method, namely, the probability proportional to size without replacement (PPSWO) sampling. The size variable was chosen as five plus the first component of x_i, for $i \in U$, and (x_i, y_i) are generated in the same was as in the first simulation, except that $\mu = 1 + 5\sum_j \beta_j$. More specifically, Tille's algorithm (Tille (1996); Deville and Tille (1998)) was employed to select PPS samples with $\pi_i \propto 5+$ as the first component of x_i.

Finite populations of size $N = 5,000$ were generated, and 500 different samples of size $n = 500$ were selected from each generated finite population. Simulated SD values are given in Table 2, with $\hat{\beta}_{\text{wls}}$ or $\hat{\beta}_{\ell_1}$. Two other quantities are included in Table 2: the estimated SD, that is, the square root of the variance estimator $v(\hat{\beta})$, defined as in (3.6); and the coverage probability (CP) of the 95% confidence interval for Y, based on a normal approximation with the estimated SD.

Overall, the results for SD are similar to those in Table 1 for SRSWO. In addition, the estimated SD is close to the simulated SD, and the CP is close to the nominal value of 95%, except for the case of $\hat{Y}_{\text{gr\_wls}}$, when p is large. The high dimension p has a greater effect on the estimated SD than it does on the estimated β.

Table 2. Standard deviation (SD), estimated SD, and coverage probability (CP) for $\hat{Y}_{\text{ht}}$, $\hat{Y}_{\text{gr\_}\ell_1}$, and $\hat{Y}_{\text{gr\_wls}}$ based on PPSWO.

		SD			estimated SD			CP		
p	s	$\hat{Y}_{\text{ht}}$	$\hat{Y}_{\text{gr\_wls}}$	$\hat{Y}_{\text{gr\_}\ell_1}$	$\hat{Y}_{\text{ht}}$	$\hat{Y}_{\text{gr\_wls}}$	$\hat{Y}_{\text{gr\_}\ell_1}$	$\hat{Y}_{\text{ht}}$	$\hat{Y}_{\text{gr\_wls}}$	$\hat{Y}_{\text{gr\_}\ell_1}$
Model M1										
10	3	1,113	218	219	1,145	219	223	95	96	96
50	7	2,828	223	227	2,913	210	230	93	96	96
100	10	4,391	263	264	4,279	207	243	88	95	92
200	14	6,235	281	266	6,096	190	253	80	95	95
300	17	7,514	431	325	7,369	248	287	73	94	92
400	20	8,456	585	325	8,741	310	289	66	95	94
Model M2										
10	10	3,748	234	244	3,770	224	235	95	94	93
50	10	3,677	244	251	3,748	213	232	95	93	93
100	10	3,793	260	269	3,752	205	245	95	87	92
200	10	3,611	285	249	3,769	188	247	97	81	95
300	10	3,719	378	272	3,748	219	254	95	70	91
400	10	3,885	594	283	3,772	293	262	93	59	92
Model M3										
10	10	23,127	18,038	17,673	22,397	17,775	17,425	94	94	94
50	10	22,304	18,345	17,594	22,372	17,895	17,159	97	95	95
100	10	22,570	16,926	16,119	22,439	17,779	16,941	95	97	97
200	10	22,246	18,485	17,400	22,432	17,633	16,760	95	94	94
300	10	21,042	17,596	15,772	22,367	18,464	16,710	96	96	97
400	10	21,819	16,551	15,359	22,444	17,273	16,468	94	96	96

5. Example

As an example, we consider the 1990 Census on law enforcement (`http://archive.ics.uci.edu/ml/datasets/communities+and+crime+unnormalized`) as a population. The data set consists of data on $N = 2,195$ communities (units) with crime related variables (study variables), such as murders, rapes, robberies, assaults, burglaries, larcenies, auto thefts, and so on. Furthermore, we include 101 covariates, including the population of a community, median household income, per capita income, number of police cars, percentage of officers assigned to drug units, and so on. A list of all 101 covariates is given in the Supplementary Material.

We selected the following six samples from this population:

(a) a simple random sample of size $n = 200$ without replacement;

(b) a simple random sample of size $n = 150$ without replacement;

(c) the first 195 communities, i.e., $S = \{1, \dots, 195\}$;

(d) the last 195 communities, i.e., $S = \{2,001, \dots, 2,195\}$;

(e) a systematic sample of size $n = 220$, i.e., $S = \{1, 11, 21, \dots, 2,191\}$;

(f) a systematic sample of size $n = 439$, i.e., $S = \{1, 6, 11, \dots, 2,191\}$.

After the sure independence screening (Fan and Lv (2008)), we estimate the population totals for murders, rapes, robberies, and assaults (four study variables) using our proposed estimator $\hat{Y}_{\mathrm{gr\_}\ell_1}$, the Horvitz–Thompson estimator $\hat{Y}_{\mathrm{ht}}$, the unregularized $\hat{Y}_{\mathrm{gr\_wls}}$ with a weighted least squares estimator, and the regularized GREG estimator $\hat{Y}_{\mathrm{gr\_sis}}$ with a weighted least squares estimator. The results are summarized in Table 3, which includes the true population totals. Overall, our method gives far more accurate estimates of the total crimes in each category than the competitors do.

6. Discussion

In this study, we established the asymptotic properties of the high-dimensional GREG estimators. We examine two approaches: the GREG estimators are constructed using (a) the WLSE, and (b) the Lasso estimator. When using the weighted least squares method to estimate the regression coefficient, we prove that the number of covariates p should increase at a much slower rate than the sample size n in order for the GREG estimator to perform well. When this condition is not satisfied, the estimator may not be efficient; indeed, its performance deteriorates, as shown in the numerical analysis. Therefore, it is not true that having more variables or auxiliary information will lead to a better Lasso estimator.

The GREG estimator constructed using the Lasso estimator, however, does not suffer from this instability. Because only a small set of variables is retained after the selection, the estimator still performs efficiently, even when p is large, as shown in the numerical study. Our simulation results not only support the theoretical analysis, but also encourage the use of the regularized GREG estimator, owing to its robustness and stability, especially when p is large.

In addition, the eigenvalue behavior of the design matrix plays an important role in the theoretical analysis of both GREG estimators. For the GREG estimator based on the WLSE, a condition is assumed in order to establish the limiting spectral distribution of the design matrix. However, for the GREG estimator based on the Lasso estimator, a restricted eigenvalue condition is assumed.

Table 3. Estimates of the total numbers of murders, rapes, robberies, and assaults in the crime data.

Scenarios	$\hat{Y}_{\text{ht}}$	$\hat{Y}_{\text{gr\_wls}}$	$\hat{Y}_{\text{gr\_}\ell_1}$	$\hat{Y}_{\text{gr\_sis}}$
Total of murders=16,633				
(a)	10,580	11,477	14,600	14,043
(b)	7,521	9,983	12,522	9,995
(c)	52,342	24,455	16,462	20,115
(d)	11,774	12,363	14,594	14,402
(e)	10,885	14,916	15,712	15,451
(f)	15,790	15,818	17,458	16,700
Total of rapes=522,378				
(a)	308,781	330,875	429,309	393,961
(b)	230,036	307,567	477,431	441,959
(c)	1,923,417	828,442	526,686	507,847
(d)	316,170	350,921	430,486	415,684
(e)	271,172	386,957	423,708	433,617
(f)	420,225	440,654	453,555	461,957
Total of robberies=716,317				
(a)	535,361	568,015	643,077	602,668
(b)	404,348	524,835	678,994	592,765
(c)	1,976,468	1,006,262	827,473	716,354
(d)	585,964	596,410	664,838	628,707
(e)	481,852	637,777	698,029	678,851
(f)	593,685	597,974	633,935	652,702
Total of assaults=1,634,471				
(a)	1,398,709	1,502,997	1,681,493	1,675,426
(b)	1,000,144	1,283,311	1,704,307	1,368,095
(c)	3,558,106	2,181,098	1,791,385	1,692,598
(d)	1,516,430	1,525,991	1,667,527	1,580,314
(e)	1,166,383	1,506,206	1,523,222	1,595,661
(f)	1,468,040	1,455,648	1,576,333	1,591,371

If the population total X in (2.3) is not available and is replaced by $\hat{X}$, an estimated total from another survey, then the GREG estimators are still consistent, as long as $\hat{X}$ is consistent. However, their efficiencies depend on the efficiency of $\hat{X}$, even if model (2.2) holds. Another situation in which our result is useful is when y_i has a covariate-dependent nonresponse and x_i is always observed. If, throughout this paper, we replace S with R the set of units with observed y_i, $R \subset S \subset U$, and replace the known X in GREG with $\hat{X} = \sum_{i \in S} x_i/\pi_i$, then $\hat{Y}_{\text{gr\_wls}}$ and $\hat{Y}_{\text{gr\_}\ell_1}$ are the same as the estimators of Y with every missing y_i imputed by $\hat{\beta}_{\text{wls}}^T x_i$ and $\hat{\beta}_{\ell_1}^T x_i$, respectively. Our Theorems 1–2 still apply. That is, WLSE works well when p is small relative to n, and the Lasso works well when β is sparse and p is comparable with n.

Note that similar results may be established if the Lasso estimator is replaced by a sparse estimator of β obtained using other penalized regression or variable selection methods. Our results, together with those of in Cardot, Goga and Shehzad (2017), demonstrate that under certain assumptions, the nice properties of the model-assisted estimators, such as the asymptotic efficiency and consistency, are still preserved in high dimensions.

Supplementary Material

The online Supplementary Material contains all theoretical proofs of Theorems 1–3 and, a complete list of the 101 covariates in the data example.

Acknowledgements

We would like to thank the two referees and an associate editor for their helpful comments. Our research was supported by the National Natural Science Foundation of China (11501208, 11871287, 11831008), the Natural Science Foundation of Tianjin (18JCYBJC41100), the Fundamental Research Funds for the Central Universities, the Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin, the Chinese 111 Project (B14019), the UW-Madison Graduate Research Scholar/Advanced Opportunity Fellowship, and the U.S. National Science Foundation grant (DMS-1612873). Dr. Lei Wang is the corresponding author of this paper.

References

Bai, Z.D. and Zhou, W. (2008). Large sample covariance matrices without independence structures in columns. *Statistica Sinica* **18**, 425–442.

Berger, Y.G., Tirari, M. and Tille, Y. (2003). Towards optimal regression estimation in sample surveys. *Australian and New Zealand Journal of Statistics* **45**, 319–329.

Bickel, P.J. and Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics* **12**, 470–482.

Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* **37**, 1705–1732.

Candes, E. and Tao, T. (2005). Decoding by linear programming. *Information Theory, IEEE Transactions on* **51**, 4203–4215.

Candes, E. and Tao, T. (2007). The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics* **35**, 2313–2351.

Cardot, H., Goga, C. and Shehzad, M.A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica* **27**, 243–260.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63**, 615–620.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1977). *Foundations of Inference in Survey Sampling.* Wiley, New York.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association* **87**, 376–382.

Deville, J.C. and Tille, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika* **85**, 89–101.

Donoho, D.L. and Huo, X. (2001). Uncertainty principles and ideal atomic decomposition. *Information Theory, IEEE Transactions on* **47**, 2845–2862.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B (Statistical Methodological)* **70**, 849–911.

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

Fuller, W.A. (2009). *Sampling Statistics.* John Wiley & Sons, New York.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Jha, A.K., DesRoches, C.M., Campbell, E.G., Donelan, K., Rao, S.R., Ferris, T.G., Shields, A., Rosenbaum, S. and Blumenthal, D. (2009). Use of electronic health records in U.S. hospitals. *New England Journal of Medicine* **360**, 1628–1638.

Krewski, D. and Rao, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics* **9**, 1010–1019.

McConville, K.S. (2011). *Improved Estimation for Complex Surveys Using Modern Regression Techniques.* PhD Dissertation.

McConville, K.S., Breidt, F.J., Lee, T.C.M. and Moisen, G.G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology* **5**, 131–158.

Nascimento Silva, P. and Skinner, C.J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology* **23**, 23–32.

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on* **59**, 3434–3447.

Särndal, C.E. (1980a). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika* **67**, 639–650.

Särndal, C.E. (1980b). A two-way classification of regression estimation strategies in probability sampling. *Canadian Journal of Statistics* **8**, 165–177.

Särndal, C.E., Swensson, B. and Wretman, J. (2003). *Model Assisted Survey Sampling.* Springer

Science & Business Media, New York.

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

Tille, Y. (1996). An elimination procedure for unequal probability sampling without replacement. *Biometrika* **83**, 238–241.

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association* **88**, 89–96.

Xie, J. (2013). Limiting spectral distribution of normalized sample covariance matrices with $p/n \to 0$. *Statistics and Probability Letters* **83**, 543–550.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research* **7**, 2541–2563.

Zhou, S. (2009). Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045.*

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, U.S.A.

E-mail: mytramta@gmail.com

School of Statistics, East China Normal University, Shanghai 200241, China.
Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, U.S.A.

E-mail: shao@stat.wisc.edu

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.

E-mail: quefeng@email.unc.edu

School of Statistics and Data Science, LPMC & KLMDASR, Nankai University, Tianjin 300071, China.

E-mail: lwangstat@nankai.edu.cn

(Received November 2016; accepted July 2018)