<table>
<tr><th colspan="2">Statistica Sinica Preprint No: SS-2017-0384</th></tr>
<tr><td><b>Title</b></td><td>Generalized Regression Estimators with
High-Dimensional Covariates</td></tr>
<tr><td><b>Manuscript ID</b></td><td>SS-2017-0384</td></tr>
<tr><td><b>URL</b></td><td>http://www.stat.sinica.edu.tw/statistica/</td></tr>
<tr><td><b>DOI</b></td><td>10.5705/ss.202017.0384</td></tr>
<tr><td><b>Complete List of Authors</b></td><td>Tram Ta
Jun Shao
Quefeng Li and
Lei Wang</td></tr>
<tr><td><b>Corresponding Author</b></td><td>Jun Shao</td></tr>
<tr><td><b>E-mail</b></td><td>shao@stat.wisc.edu</td></tr>
<tr><td colspan="2">Notice: Accepted version subject to English editing.</td></tr>
</table>

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 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 estimation of the population total of a survey variable to improve the estimation precision. We consider the generalized regression estimator formulated under the model-assisted framework in which a regression model is utilized to make use of the available covariates while the estimator still has basic design-based properties. The generalized regression estimator has been shown to improve the efficiency of the design-based Horvitz-Thompson estimator when the number of covariates is fixed. In this study, 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 where the model parameter is estimated using the weighted least squares method when $p < n$ and the LASSO method 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 model misspecification. We also study the consistency of 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—which comes from, for instance, 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 et al. (1977), Särndal et al. (2003)). In this framework, a model is adopted to reduce the estimation variability by utilizing the auxiliary information from covariates related to the main study variable. Since the model's role is only to assist in the estimation process, the constructed estimator is protected against 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 et al. (1976) and studied extensively in Cassel et al. (1977), Särndal (1980a), Särndal (1980b), and Särndal et al. (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 et al., 2003). A closely related estimator is the calibration estimator, which is asymptotically equivalent to the GREG estimator under certain assumptions (Deville and Särndal, 1992). In this paper, we focus on the estimation of the population total or mean using the GREG estimator.

In traditional applications where a small or moderate number of covariates are considered, properties of the GREG estimator have been well-studied; see, for example, Cassel et al. (1977) and Särndal et al. (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 GREG 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 the technological advances, it is now possible to collect data on a large number of covariates, which could even exceed the sample size. Dated back to Nascimento Silva and Skinner (1997), the authors gave many 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, such as population for community, mean people per household, percentages of population in race groups, median household income, number-s of people in age groups, percentage of households with 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 is recorded for each patient, such as patient's demographic information, biometric information, medical records, historical medical test results, etc., and population totals for many covariates are maintained in social and governmental organizations. Another example is when we consider covariate interactions and/or polynomial effects in regression, in which case even though the original number of covariates with know 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, it is of interest 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, the present paper studies the GREG estimator in a setting that 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, and 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 0, 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, the GREG estimator using a regularized regression estimator can be constructed. Dimension reduction has been studied in Cardot et al. (2014) in which the authors considered a principal component analysis to reduce the covariate dimension prior to performing calibration. This calibration approach can also be adopted to form the GREG estimator. However, asymptotic results of their GREG estimator was established under the condition $p^3/n \to 0$, which is much stronger than $p/n \to 0$ 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 $p > n$, a problem that can occur in some small area survey estimation and in many economic and biological studies. The principal component calibration approach also does not perform well in this high-dimensional problem. We adopt the LASSO (Tibshirani, 1996) as a regularization method. The use of LASSO in GREG was proposed in McConville (2011) and McConville et al. (2017), but they only studied empirical and theoretical properties in the case of a fixed p. Under some conditions on the divergence rates of s and p, we show that the GREG estimator constructed using 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 GREG with LASSO.

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

2. The Generalized Regression Estimator

Consider a finite population U that consists of N units labeled $i = 1, 2, ..., N$. Associated with unit i, let y_i be the value of the study variable and x_i be the p-dimensional vector of covariates. We consider the estimation of the finite population total $Y = \sum_{i \in U} y_i$, using data from a sample S of size n selected from U according to some probability plan called sampling design. The value of (y_i, x_i) is observed for unit i in the sample S. To estimate the total Y, Horvitz and Thompson (1952) introduced the following estimator which was also named after the authors:

$$\hat{Y}_{\text{ht}} = \sum_{i \in S} y_i / \pi_i \tag{1}$$

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 $\hat{Y}_{\rm ht}/N$ or $\hat{Y}_{\rm ht}/\sum_{i\in S}\pi_i^{-1}$. Under the noninformative sampling assumption, i.e., π_i is a known function of x_i but does not depend on y_i, the Horvitz-Thompson estimator in (1) is design-unbiased with respect to the random selection of S from U. Throughout this paper, we assume noninformative sampling and that $\hat{Y}_{\rm ht} - Y$ is asymptotically normal as $n \to \infty$ under the given sampling design with some conditions; see, for example, Bickel and Freedman (1984), Krewski and Rao (1981) and 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, and the sample is then 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 make use of 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 is 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 poststrata (Valliant, 1993), are constructed so that (x_i, y_i)'s in each sub-population can be assumed to be unconditionally independent and identically distributed. Since 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, i.e., we assume that for all $i \in U$,

$$y_i = \mu + \beta^T x_i + \epsilon_i \tag{2}$$

where μ and β are unknown parameters, a^T is the transpose of a vector a, x_i's are independent and identically distributed random vectors of covariates with an unknown positive-definite covariance matrix Σ, ϵ_i's are independent random variables with mean 0 and unknown variance σ_ϵ^2, and x_i's are independent of ϵ_i's. 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), Cassel et al. (1976, 1977) proposed the following GREG estimator of the total Y:

$$\hat{Y}_{\text{gr}} = \hat{Y}_{\text{ht}} + \hat{\beta}^T (X - \hat{X}_{\text{ht}}), \tag{3}$$

where $X = \sum_{i \in U} x_i$ is the known finite population total of x_i's, $\hat{X}_{\text{ht}}$ is the Horvitz-Thompson estimator of X defined as (1), i.e. $\hat{X}_{\text{ht}} = \sum_{i \in S} x_i / \pi_i$, and $\hat{\beta}$ is an estimator of β in (2) based on (y_i, x_i), $i \in S$. The GREG estimator in (3) is the sum of the Horvitz-Thompson estimator $\hat{Y}_{\mathrm{ht}}$ and an adjustment $\hat{\beta}^T(X - \hat{X}_{\mathrm{ht}})$ which is used to increase the efficiency.

To study the property of the GREG estimator, we first consider an artificial situation where β in (2) is known so that $\hat{\beta} = \beta$ and the estimator in (3) is denoted as

$$\hat{Y}^*_{\mathrm{gr}} = \hat{Y}_{\mathrm{ht}} + \beta^T(X - \hat{X}_{\mathrm{ht}}) \tag{4}$$

Since $\hat{X}_{\mathrm{ht}}$ is the Horvitz-Thompson estimator of X, $\hat{Y}^*_{\mathrm{gr}}$ is a design-unbiased estimator of Y even if model (2) is wrong or β is a wrong value. If model (2) is correct, then regardless of how large the dimension p is, the variance of $\hat{Y}^*_{\mathrm{gr}}$ is smaller than the variance of $\hat{Y}_{\mathrm{ht}}$ unless $\beta = 0$, where the variance is with respect to both 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 not exactly, but asymptotically design-unbiased and normally distributed as long as $\hat{\beta}$ does not diverge to infinity. In traditional setting, the covariate dimension is fixed in the sense that p does not change as $n \to \infty$. Then, under model (2), the GREG estimator is asymptotically more efficient than the Horvitz-

Thompson estimator as long as $\hat{\beta}$ is consistent, since

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

where $o_p(1)$ denotes a quantity converging to 0 in probability. This implies that in low-dimensional setting, $\hat{Y}_{\rm gr}$ and $\hat{Y}^*_{\rm gr}$ in (4) are asymptotically equivalent under model (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),

$$\hat{\beta}_{\rm wls} = \left\{\sum_{i\in S} \frac{1}{\pi_i}\left(x_i - \hat{X}_{\rm ht}/\hat{N}\right)\left(x_i - \hat{X}_{\rm ht}/\hat{N}\right)^T\right\}^{-1} \sum_{i\in S} \frac{(x_i - \hat{x}_S)y_i}{\pi_i} \qquad (5)$$

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) is correct, $n^{1/2}(\hat{\beta}_{\rm wls} - \beta)$ is asymptotically normal with mean 0, and thus

$$\hat{Y}_{\rm gr\_wls} - Y = \hat{Y}^*_{\rm gr} - Y + O_p(n^{-1/2})(\hat{Y}^*_{\rm gr} - Y) \qquad (6)$$

i.e., $\hat{Y}_{\rm gr\_wls}$ 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 section, 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 $\hat{Y}_{\rm gr}^*$ so that it improves $\hat{Y}_{\rm ht}$. The answer is given in the following result.

Theorem 1. *Assume model (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 independent and identically distributed and have finite 4th order moments.*

Then we have the following conclusions.

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

$$\hat{Y}_{\rm gr\_wls} - Y = \hat{Y}_{\rm gr}^* - Y + O_p\{(p/n)^{1/2}\}(\hat{Y}_{\rm gr}^* - Y) \tag{7}$$

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

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

Assumptions (A1) and (A2) involve bounds on the inclusion probabilities. Assumption (A3) is used to obtain the limiting spectral distribution of 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 $E(x_i^T \Sigma^{-1/2} B \Sigma^{-1/2} x_i - \mathrm{tr} B)^2 = o(p^3/n)$ for any $p \times p$ deterministic matrix B with bounded spectral norm.

Note that result (7) includes result (6) for the case of fixed p as a special case. Theorem 1 indicates that under model (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 (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 $\hat{Y}_{\text{gr\_wls}}$ can be even worse than $\hat{Y}_{\text{ht}}$, even if model (2) is correct. In the next section we consider an improvement of $\hat{Y}_{\text{gr\_wls}}$ 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 nowadays contain many covariates, it is often true that only a few of these available covariates are actually related to the study variable. In model (2), this amounts to that, among p covariates, only s of them have non-zero regression coefficients, i.e., β-components, and s is fixed or diverges much slower than p. It is desired to have a sparse estimator of β when β is sparse since retaining the extraneous variables serves no purpose but increases the variability and model complexity. The WLSE $\hat{\beta}_{\text{wls}}$, however, is not sparse regardless of whether β is sparse or not. Therefore, we consider the LASSO estimator, denoted by $\hat{\beta}_{\ell_1}$. The GREG estimator in (3) using $\hat{\beta} = \hat{\beta}_{\ell_1}$, denoted as $\hat{Y}_{\text{gr\_}\ell_1}$, is well defined even when $p > n$. In this section, we study the asymptotic properties of $\hat{Y}_{\text{gr\_}\ell_1}$ and show that it improves $\hat{Y}_{\text{gr\_wls}}$ as well as the Horvitz-Thompson estimator $\hat{Y}_{\text{ht}}$ and is asymptotically equivalent to $\hat{Y}^*_{\text{gr}}$, under some conditions on sparsity and diverging rate of p which allows $p/n \to \infty$. It is also design-based robust against 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:

$$\min_{b \in R^p} \left[\frac{1}{2n} \sum_{i \in S} \frac{\{y_i - b^T(x_i - \hat{X}_{\text{ht}}/\hat{N})\}^2}{\pi_i} + \lambda \|b\|_1 \right] \tag{8}$$

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 select variables simultaneously. Note that the WLSE $\hat{\beta}_{\text{wls}}$ is the special case of $\hat{\beta}_{\ell_1}$ defined as a solution to (8) with $\lambda = 0$.

There is a considerable literature devoted to studying conditions on the covariates x_i's in order to guarantee certain good oracle properties of $\hat{\beta}_{\ell_1}$ in terms of prediction or estimation accuracy and variable selection consistency. Some of the most well-known conditions are the restricted null space property (Donoho and Huo, 2001), the restricted isometry property (Candes and Tao, 2005, 2007), the restricted eigenvalue condition (Bickel et al., 2009), and the 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 the noiseless setting, i.e., $\epsilon_i = 0$ for all i in (2). When ϵ_i's in (2) are not degenerated, the restricted isometry property was proved to be sufficient for bounding the estimation error.

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

$$\frac{1}{K_{(l,k,A)}} = \min_{\substack{J \subset \{1,\ldots,p\} \\ |J| \le l}} \min_{\substack{v \neq 0 \\ \|v_{-J}\|_1 \le k \|v_J\|_1}} \frac{\|Av\|_2}{\|v_J\|_2} > 0 \tag{9}$$

where v_J is the sub-vector of v with components indexed by elements in $J \subset \{1, \ldots, p\}$, v_{-J} is the 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, l and k are constants. The condition is denoted as $RE(l, k, A)$.

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

$$\mathcal{C}_{(l,k)} = \{v \in R^p : \exists J \subset \{1, \ldots, p\}, |J| \le l, \|v_{-J}\|_1 \le k \|v_J\|_1\}, \tag{10}$$

and 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 $\mathcal{C}_{(s,3)}$, i.e., $\|(\hat{\beta}_{\ell_1} - \beta)_{-\mathcal{S}}\|_1 \leq 3\|(\hat{\beta}_{\ell_1} - \beta)_{\mathcal{S}}\|_1$, where $\mathcal{S}$ contains indices of all non-zero components of β and $s = |\mathcal{S}|$.

Condition (9) was first assumed in Bickel et al. (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 matrices whose rows follow a sub-Gaussian distribution. In this study, covariates x_i's under the model-assisted framework are random vectors distributed according to the super-population model. We consider a random design matrix **X** whose ith row is x_i, $i \in S$. If x_i's follow a sub-Gaussian distribution and Σ is positive definite, then under certain assumptions, condition (9) holds for $A = \mathbf{X}/n^{1/2}$ with high probability (Rudelson and Zhou, 2013) . 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 factor τ and ν, i.e., $E\{\exp(u\epsilon_i)\} \leq \exp(\tau^2 u^2/2)$ 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 $e_j = (0, .., 1, .., 0)$, $j = 1, ..., p$*, form the standard basis of* R^p*.*

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

(i) If model (2) holds, then

$$\|\hat{\beta}_{\ell_1} - \beta\|_1 = O_p\left\{s(n^{-1}\log p)^{1/2}\, MK^2_{(s,3,\Sigma^{1/2})}\right\} \tag{11}$$

and

$$\hat{Y}_{\mathrm{gr}\_\ell_1} - Y = \hat{Y}^*_{\mathrm{gr}} - Y + O_p\left\{n^{-1/2}\, s\log p\, MK^2_{(s,3,\Sigma^{1/2})}\right\}(\hat{Y}^*_{\mathrm{gr}} - Y). \tag{12}$$

(ii) If model (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 (11) still holds and*

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

The result on estimation loss $\|\hat{\beta}_{\ell_1} - \beta\|_1$ was first established in Bickel et al. (2009) for deterministic x_i's, where the RE condition was imposed on the design matrix **X**. Zhou (2009) also showed that an estimation loss with a similar order as (11) holds when the rows of the random matrix **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 s-parse vectors with at most s nonzero components. We instead make a similar assumption (A5) as in Rudelson and Zhou (2013) in which the lower bound of n does not depend on $\rho(s)$, but a slightly stronger $RE(s, 9, \Sigma^{1/2})$ assumption is used.

Theorem 2 indicates that $\hat{Y}_{\mathrm{gr\_}\ell_1}$ is asymptotically equivalent to $\hat{Y}^*_{\mathrm{gr}}$, even when working model (2) is misspecified, as long as $n^{-1/2} s \log p \, MK^2_{(s,3,\Sigma^{1/2})} \to 0$, which is reasonable since $s \log p$ can be much smaller than n. Hence, $\hat{Y}_{\mathrm{gr\_}\ell_1}$ asymptotically outperforms $\hat{Y}_{\mathrm{ht}}$ if model (2) holds. When model (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) is nearly correct. See the simulation results in Section 4.1.

In the study of Cardot et al. (2014) where the authors used calibration based on the principal components of the covariates, the number of covariates p was restrictively assumed to satisfy $p^3 r^3 / n \to 0$ 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 using 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 since complete covariate information in the entire population is in general not available, especially when x_i has a high dimension.

To assess the estimation variability or make inference about Y, we need a variance estimator for $\hat{Y}_{\mathrm{gr\_}\ell_1}$. First, consider $\hat{Y}^*_{\mathrm{gr}}$ given by (4). If β is treated as known, then a classical variance estimator for $\hat{Y}^*_{\mathrm{gr}}$ is

$$v(\beta) = \sum_{i\in S}\sum_{j\in S} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} \, \frac{y_i - x_i^T\beta}{\pi_i} \, \frac{y_j - x_j^T\beta}{\pi_j}, \tag{13}$$

where π_{ij} is the inclusion probability of units i and j in the sample S, $i \neq j$. When β is unknown, it is substituted by the same estimator $\hat{\beta}$ used in GREG. In the traditional case where p is fixed, $v(\hat{\beta})$ defined as (13) with β replaced by a consistent $\hat{\beta}$ is consistent for the variance of $\hat{Y}_{\mathrm{gr}}$ as $n \to \infty$. The next result shows that this is still true when β is estimated by LASSO.

Theorem 3. *Assume model (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 (11) converges to 0. Then, the variance estimator* $v(\hat{\beta}_{\ell_1})$ *defined as (13) 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 considered 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's 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 whose diagonal entries are equal to 1 and every off-diagonal entry is 0 with probability 0.8 and equal to the value of a random variable having the 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) with $s = p^{1/2}, \beta = (2, \ldots, 2, 0, \ldots, 0)$, where ϵ_i's 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 were set to 2 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 ten entries of β were $1, 2, 3, 4, 5$, $0.2, 0.2,\ 0.2, 0.2, 0.2$, and all other entries of β were zeros. Therefore, the underlying model has a dimension $s = 10$ although p increases. Since non-zero components of β took different values, the corresponding covariates were correlated with the variable y with different strength.

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's are i.i.d. $N(0, 1)$, and $\mu = \sum_j \beta_j$. The parameter β was, however, still estimated under the assumed model (2) in order to investigate the consequences of 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 et al. (2003) were computed. A 10-fold cross-validation was used to select the tuning parameter λ in the minimization problem (8) and the one with the smallest mean squared error was chosen (Friedman et al., 2010). Based on the 500 simulations, the standard deviation (SD) of each estimator $\hat{Y}$ and ratio of mse$(\hat{Y})$ for pairs of estimators, where $\mathrm{mse}(\hat{Y})$ is the mean squared error of $\hat{Y}$, were reported in Table 1 for all three models M1-M3. All estimators $\hat{Y}_{\mathrm{ht}}$, $\hat{Y}_{\mathrm{gr\_wls}}$, $\hat{Y}_{\mathrm{gr\_opt}}$ and $\hat{Y}_{\mathrm{gr\_}\ell_1}$ have negligible biases less than 1% of Y and hence were 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, 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 folds by utilizing the auxiliary information. Under model M3, which is a wrong 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 what was observed in models M1-M2, since the auxiliary information was not utilized in a correct way.

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

It also can be seen that $\hat{Y}_{\text{gr\_}\ell_1}$ has comparable performance to the optimal estimator $\hat{Y}_{\text{gr\_opt}}$ when 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 so it does not perform well when p is large, although it is 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, the probability proportional to size without replacement (PPSWO) sampling. The size variable was chosen to be 5 plus the first component of x_i, $i \in U$, and (x_i, y_i)'s are generated the same as those in the first simulation, except that $\mu = 1 + 5\sum_j \beta_j$. More specifically, Tille's algorithm (Tillé, 1996; Deville and Tille, 1998) was employed to select PPS samples with $\pi_i \propto 5+$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 squared root of the variance estimator $v(\hat{\beta})$ defined as (13), and the coverage probability (CP) of the 95%

confidence interval for Y by normal approximation with estimated SD.

Overall, the results on 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 95%, except for the case of $\hat{Y}_{\text{gr\_wls}}$ when p is large. The high dimension p has more effects on the estimated SD than the estimated β.

5. Example

As an example, we considered the 1990 Census on law enforcement (http://archive.ics.uci.edu/ml/datasets/communities+and+crime+unnormalized) as a population, which consists of $N = 2,195$ communities (units) with crime related variables (study variables) such as murders, rapes, robberies, assaults, burglaries, larcenies, auto thefts, etc., and 101 covariates including population for community, median household income, per capita income, number of police cars, percent of officers assigned to drug units, etc. A list of all 101 covariates is Lgiven 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., $S = \{2001, \dots, 2195\}$;

(e) a systematic sample of size $n = 220$, i.e., $S = \{1, 11, 21, \ldots, 2191\}$;

(f) a systematic sample of size $n = 439$, i.e., $S = \{1, 6, 11, \ldots, 2191\}$.

We estimate the population totals of 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 weighted least squares estimator, and the regularized GREG estimator $\hat{Y}_{\mathrm{gr\_sis}}$ with weighted least squares estimator after sure independence screening (Fan and Lv, 2008). The results are summarized in Table 3, which includes the true population totals. Overall, our method gives much more accurate estimate of the total crimes in each category than the competitors.

6. Discussion

In this study, asymptotic properties of the high-dimensional GREG estimators are established. We examine two approaches where the GREG estimators are constructed using the WLSE and the LASSO estimator. When using the weighted least squares method to estimate the regression coefficient, we prove that the number of covariate 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 the more variables or auxiliary information we use, the better the estimator is.

The GREG estimator constructed using the LASSO estimator, however, does not suffer from this instability. Since only a small set of variables is retained after the selection, the estimator is still able to perform 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 since it is more robust and stable, especially when p is large.

It can also be observed that 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 to establish the limiting spectral distribution of the design matrix, while for the GREG estimator based on the LASSO estimator, the restricted eigenvalue condition is assumed.

If the population total X in (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, but their efficiencies depend on the efficiency of $\hat{X}$ even if model (2) holds. Another situation in which our result is useful is when y_i has covariate-dependent nonresponse and x_i is always observed. If we replace S in the retire paper by R, the set of units with observed y_i's, $R \subset S \subset U$, and replace the known X in GREG by $\hat{X} = \sum_{i \in S} x_i/\pi_i$, then $\hat{Y}_{\text{gr\_wls}}$ and $\hat{Y}_{\text{gr\_}\ell_1}$ are the same as 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, i.e., 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.

It should be noted that similar results may be established if the LASSO estimator is replaced by a sparse estimator of β obtained by using other penalized regression or variable selection methods. The results in this paper, together with those obtained in Cardot et al. (2014), demonstrate that under certain assumptions, nice properties of the model-assisted estimators such as the asymptotic efficiency and consistency are still preserved in high dimension.

Supplementary Material

The supplementary material contains all theoretical proofs of Theorems 1-3 and a complete list of 101 covariates in data example.

Acknowledgement

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

References

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

Berger, Y. G., M. Tirari, and Y. Tille (2003). Towards optimal regression estimation in sample surveys. *Australian and Newzeland Journal of Statistics 45*(3), 319–329.

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

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

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

Candes, E. J. and T. Tao (2005). Decoding by linear programming. *Information Theory, IEEE Transactions on 51*(12), 4203–4215.

Cardot, H., C. Goga, and M.-A. Shehzad (2014). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *arXiv preprint arXiv:1406.7686*.

Cassel, C. M., C. E. Särndal, and J. H. Wretman (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika 63*(3), 615–620.

Cassel, C.-M., C. E. Särndal, and J. H. Wretman (1977). *Foundations of inference in survey sampling*. Wiley.

Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association 87*(418), 376–382.

Deville, J.-C. and Y. Tille (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika 85*(1), 89–101.

Donoho, D. L. and X. Huo (2001). Uncertainty principles and ideal atomic decomposition. *Information Theory, IEEE Transactions on 47*(7), 2845–2862.

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

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

Fuller, W. A. (2009). *Sampling statistics*. John Wiley & Sons.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685.

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

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

McConville, K. S. (2011). Improved estimation for complex surveys using modern regression techniques. *PhD Dissertation*.

McConville, K. S., F. J. Breidt, T. C. M. Lee, and G. G. Moisen (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology 5*(2), 131–158.

Nascimento Silva, P. and C. J. Skinner (1997). Variable selection for regression estimation in finite populations. *Survey Methodology 23*(1), 23–32.

Rudelson, M. and S. Zhou (2013). Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on 59*(6), 3434–3447.

Särndal, C. E. (1980a). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika 67*(3), 639–650.

Särndal, C.-E. (1980b). A two-way classification of regression estimation strategies in probability sampling. *Canadian Journal of Statistics 8*(2), 165–177.

Särndal, C.-E., B. Swensson, and J. Wretman (2003). *Model assisted survey sampling*. Springer Science & Business Media.

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

Tillé, Y. (1996). An elimination procedure for unequal probability sampling without replacement. *Biometrika 83*(1), 238–241.

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association 88*(421), 89–96.

Xie, J. (2013). Limiting spectral distribution of normalized sample covariance matrices with p/n→ 0. *Statistics & Probability Letters 83*(2), 543–550.

Zhao, P. and B. Yu (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, Nankai University, Tianjin 300071, China.

E-mail: leiwang.stat@gmail.com

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

		SD						
p	s	$\hat{Y}_{\mathrm{ht}}$	$\hat{Y}_{\mathrm{gr\_wls}}$	$\hat{Y}_{\mathrm{gr\_opt}}$	$\hat{Y}_{\mathrm{gr\_}\ell_1}$	$\frac{\mathrm{mse}(\hat{Y}_{\mathrm{ht}})}{\mathrm{mse}(\hat{Y}_{\mathrm{gr\_}\ell_1})}$	$\frac{\mathrm{mse}(\hat{Y}_{\mathrm{gr\_wls}})}{\mathrm{mse}(\hat{Y}_{\mathrm{gr\_}\ell_1})}$	$\frac{\mathrm{mse}(\hat{Y}_{\mathrm{gr\_opt}})}{\mathrm{mse}(\hat{Y}_{\mathrm{gr\_}\ell_1})}$
				Model M1				
10	3	17099	4495	4489	4477	14.6	1.0	1.0
50	7	29205	5559	4760	4675	39.0	1.4	1.1
100	10	31759	7982	4988	4931	41.5	2.6	1.1
200	14	40717	16832	5715	5019	65.8	11.3	1.6
300	17	44378	26973	7202	5288	70.4	27.0	2.0
400	20	47563	38079	10121	5349	79.1	53.2	4.3
				Model M2				
10	10	36939	4604	4521	4525	63.8	1.0	1.0
50	10	40344	6240	4730	4689	74.4	1.8	1.0
100	10	33658	8320	5013	4755	50.1	3.1	1.1
200	10	35033	14837	5849	4771	53.9	10.2	1.5
300	10	35740	21874	7304	4803	55.4	21.2	2.3
400	10	33369	26788	10044	4720	52.5	32.5	4.5
				Model M3				
10	10	88030	51215	51227	51313	2.7	1.1	1.1
50	10	83894	51969	51748	50186	2.7	1.1	1.1
100	10	87616	54823	53962	49398	3.1	1.2	1.2
200	10	86742	62090	60671	49839	3.0	1.5	1.5
300	10	86010	76002	67390	49760	3.0	2.3	1.8
400	10	87531	106794	77554	49498	3.1	4.6	2.4

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

p	s	SD			estimated SD			CP		
		$\hat{Y}_{\rm ht}$	$\hat{Y}_{\rm gr\_wls}$	$\hat{Y}_{{\rm gr\_}\ell_1}$	$\hat{Y}_{\rm ht}$	$\hat{Y}_{\rm gr\_wls}$	$\hat{Y}_{{\rm gr\_}\ell_1}$	$\hat{Y}_{\rm ht}$	$\hat{Y}_{\rm gr\_wls}$	$\hat{Y}_{{\rm gr\_}\ell_1}$
Model M1										
10	3	1113	218	219	1145	219	223	95	96	96
50	7	2828	223	227	2913	210	230	93	96	96
100	10	4391	263	264	4279	207	243	88	95	92
200	14	6235	281	266	6096	190	253	80	95	95
300	17	7514	431	325	7369	248	287	73	94	92
400	20	8456	585	325	8741	310	289	66	95	94
Model M2										
10	10	3748	234	244	3770	224	235	95	94	93
50	10	3677	244	251	3748	213	232	95	93	93
100	10	3793	260	269	3752	205	245	95	87	92
200	10	3611	285	249	3769	188	247	97	81	95
300	10	3719	378	272	3748	219	254	95	70	91
400	10	3885	594	283	3772	293	262	93	59	92
Model M3										
10	10	23127	18038	17673	22397	17775	17425	94	94	94
50	10	22304	18345	17594	22372	17895	17159	97	95	95
100	10	22570	16926	16119	22439	17779	16941	95	97	97
200	10	22246	18485	17400	22432	17633	16760	95	94	94
300	10	21042	17596	15772	22367	18464	16710	96	96	97
400	10	21819	16551	15359	22444	17273	16468	94	96	96

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=16633				
(a)	10580	11477	14600	14043
(b)	7521	9983	12522	9995
(c)	52342	24455	16462	20115
(d)	11774	12363	14594	14402
(e)	10885	14916	15712	15451
(f)	15790	15818	17458	16700
Total of rapes=522378				
(a)	308781	330875	429309	393961
(b)	230036	307567	477431	441959
(c)	1923417	828442	526686	507847
(d)	316170	350921	430486	415684
(e)	271172	386957	423708	433617
(f)	420225	440654	453555	461957
Total of robberies=716317				
(a)	535361	568015	643077	602668
(b)	404348	524835	678994	592765
(c)	1976468	1006262	827473	716354
(d)	585964	596410	664838	628707
(e)	481852	637777	698029	678851
(f)	593685	597974	633935	652702
Total of assaults=1634471				
(a)	1398709	1502997	1681493	1675426
(b)	1000144	1283311	1704307	1368095
(c)	3558106	2181098	1791385	1692598
(d)	1516430	1525991	1667527	1580314
(e)	1166383	1506206	1523222	1595661
(f)	1468040	1455648	1576333	1591371