Journal of Survey Statistics and Methodology (2017) **5**, 131–158

MODEL-ASSISTED SURVEY REGRESSION ESTIMATION WITH THE LASSO

KELLY S. MCCONVILLE*
F. JAY BREIDT
THOMAS C. M. LEE
GRETCHEN G. MOISEN

In the U.S. Forest Service's Forest Inventory and Analysis (FIA) program, as in other natural resource surveys, many auxiliary variables are available for use in model-assisted inference about finite population parameters. Some of this auxiliary information may be extraneous, and therefore model selection is appropriate to improve the efficiency of the survey regression estimators of finite population totals. A model-assisted survey regression estimator using the lasso is presented and extended to the adaptive lasso. For a sequence of finite populations and probability sampling designs, asymptotic properties of the lasso survey regression estimator are derived, including design consistency and central limit theory for the estimator and design consistency of a variance estimator. To estimate multiple finite population quantities with the method, lasso survey regression weights are developed, using both a model calibration approach and a ridge regression approximation. The gains in efficiency of the lasso estimator over the full regression estimator are demonstrated through a simulation study estimating tree canopy cover for a region in Utah.

Kelly S. McConville is Assistant Professor, Department of Mathematics and Statistics, Swarthmore College, Swarthmore. PA. F. Jay Breidt is Professor, Department of Statistics, Colorado State University, Fort Collins, CO. Thomas C. M. Lee is Professor, Department of Statistics, University of California, Davis, CA. Gretchen G. Moisen is Research Forester, U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Ogden, UT.

This work was partially supported by the U.S. National Science Foundation (SES 0922143 to KSM and FJB; 1007520, 1209226, and 1209232 to TCML). The authors would like to acknowledge the USFS Forest Inventory and Analysis Program as well as the USFS Remote Sensing Applications Center for the data.

*Address correspondence to Kelly S. McConville, Department of Mathematics and Statistics, Swarthmore College, 500 College Ave, Swarthmore, PA, 19091, USA; E-mail: kmcconv1@swarthmore.edu.

doi: 10.1093/jssam/smw041 Advance access publication 26 April 2017

KEYWORDS: Adaptive lasso; Calibration estimation; Complex surveys; Generalized regression estimation; Model-assisted inference; Model selection.

1. INTRODUCTION

1.1 Model-Assisted Estimation

Consider estimation of a finite population total $t_y = \sum_{j \in U} y_j$, where $U = \{1, \ldots, N\}$ is the set of elements of the finite population and y_j is the value of a response variable for the jth element. Let $s \subset U$ be selected according to a sampling design $\mathrm{p}(\cdot)$, where $\mathrm{p}(s)$ is the probability of selecting s. For $j, k \in U$, let $\pi_j = \Pr[j \in s] = \sum_{s \subset U: j \in s} \mathrm{p}(s)$ denote the first-order inclusion probabilities of the design and $\pi_{jk} = \Pr[j, k \in s] = \sum_{s \subset U: j,k \in s} \mathrm{p}(s)$ the second-order inclusion probabilities. Assuming $\pi_j > 0$ for all $j \in U$, the design is a probability sampling design and the Horvitz and Thompson (1952) estimator

$$\hat{t}_{y,HT} = \sum_{j \in s} \frac{y_j}{\pi_j} = \sum_{j \in U} y_j \frac{I_j}{\pi_j}, \tag{1}$$

where $I_j = 1$ if $j \in s$ and $I_j = 0$ otherwise, is design unbiased for t_y in the sense that

$$\mathrm{E_p}(\hat{t}_{y,HT}) = \sum_{s \subset U} \hat{t}_{y,HT}(s)\mathrm{p}(s) = t_y.$$

No model or other structure has been assumed for y.

In the United States Forest Service's Forest Inventory and Analysis Program (FIA), numerous response variables y are collected in the field or through manual interpretation of aerial photography using a systematic or random sampling design. Data collection can be extremely expensive, so there is strong interest in using ancillary data from remote sensing and other spatial data sources to improve efficiency in inventory estimates and reduce inventory costs. For $j \in U$, let $\boldsymbol{x}_j = (1, x_{j1}, \ldots, x_{jp})^T$ denote the vector of ancillary data. The key assumption of our method is that $t_{\boldsymbol{x}} = \sum_{j \in U} \boldsymbol{x}_j$ is known and $\{\boldsymbol{x}_j\}_{j \in s}$ is observed for the sample. We do not require the stronger condition that $\{\boldsymbol{x}_j\}_{j \in U}$ is observed for the entire finite population. But if such element-level data are in fact available, then our key assumption is met not only for the original variables, but also for any of their transformations, singly or in combination with others (e.g., interaction terms).

In the FIA, a wealth of relevant ancillary data are available across the United States, including spectral values and indices provided through satellite programs such as Landsat, topographic and bioclimatic variables derived from digital elevation models, and predicted vegetation surfaces compiling remotely sensed data layers into information most relevant to natural resource land managers. As is common in such applications, these data $\{\boldsymbol{x}_j\}_{j \in U}$ are known for the entire finite population, so that we do in fact have a very large vector of potential covariates from the original variables and their transformations.

By contrast, in household surveys it would be unusual to have this kind of rich ancillary data at the element level of person within household. If rich ancillary data were instead available at the household level, then we conjecture that the techniques we describe could be adapted by developing model-assisted methods at the cluster level instead of the element level, as is commonly done in related contexts (see, for example, Särndal, Swensson, and Wretman 1992, section 8.4). We do not pursue cluster-level ancillary data further in this paper.

One way to use these ancillary data in estimation is to compute a model-assisted estimator of t_y by specifying a working model for the mean of y given $\boldsymbol{x}$ and using this model to predict y values. Linear working models lead to classical survey poststratification, ratio, and regression estimators (Cochran 1977, chapters 5–7), all of which are special cases of the generalized regression estimator (GREG) (Cassel, Särndal, and Wretman 1976; Särndal, Swensson, and Wretman 1992, chapter 6). Other specifications of the working model lead to model-assisted estimators based on local polynomial regression (Breidt and Opsomer 2000), penalized splines (Breidt, Claeskens, and Opsomer 2005; McConville and Breidt 2013), neural networks (Montanari and Ranalli 2005b), regression splines (Goga 2005), and additive and generalized additive models (Opsomer, Breidt, Moisen, and Kauermann 2007; Wang and Wang 2011), among many others. See Särndal (2010) for some general review of model-assisted estimation and Montanari and Ranalli (2005a) and Breidt and Opsomer (2009) for nonparametric and semiparametric model-assisted methods. We caution that many of these model-assisted estimators require the stronger condition that $\{\boldsymbol{x}_j\}_{j \in U}$ is known for the entire finite population.

In this paper, we consider the GREG under a linear working model,

$$y_j = \boldsymbol{x}_j^T \boldsymbol{\beta} + \epsilon_j \tag{2}$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{\mathrm{p}})^{\mathrm{T}}$ and ϵ_j independent and identically distributed with mean zero and variance σ^2. We emphasize that the working model is a device used to motivate estimators and is not assumed to hold, particularly in the design-based asymptotic results described below. The GREG is then

$$\hat{t}_{y,greg} = \frac{\sum_{j \in s} y_j - \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s}{\pi_j} + \sum_{j \in U} \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s \tag{3}$$

(Cassel, Särndal, and Wretman 1976; Särndal, Swensson, and Wretman 1992, section 6.3), with regression parameters estimated via

$$\hat{\boldsymbol{\beta}}_s = \underset{\beta}{\operatorname{argmin}} \, (\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta})^T \Pi_s^{-1} (\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}) = (\boldsymbol{X}_s^T \Pi_s^{-1} \boldsymbol{X}_s)^{-1} \boldsymbol{X}_s^T \Pi_s^{-1} \boldsymbol{Y}_s, \quad (4)$$

where $\boldsymbol{X}_s = [\boldsymbol{x}_j^T]_{j \in s}$ is an $n \times (p+1)$ matrix, $\boldsymbol{Y}_s = [y_j]_{j \in s}$ is an n-vector, and $\Pi_s = \operatorname{diag}(\pi_j)_{j \in s}$ is an $n \times n$ diagonal matrix of the first-order inclusion probabilities for the sampled elements.

The GREG is asymptotically design unbiased and consistent for t_y under mild design conditions, even if the working model is not correct. The GREG typically has smaller variance than the Horvitz-Thompson estimator if the working model has some predictive power for y; see Sarndal, Swensson and Wretman (1992, Section 6.7), Fuller (2009, Section 2.2).

1.2 Model Selection and the Lasso

In the FIA, the numerous layers of ancillary data $\{\boldsymbol{x}_j\}_{j \in U}$ are frequently correlated and potentially do not have significant relationships with the variables of interest. In this setting, model selection to remove extraneous variables could reduce the variance of the GREG, making it a more efficient estimate of the finite population total.

Two widely used methods of model selection are best subsets selection and stepwise selection. If the number of potential covariates is large, then best subsets selection becomes computationally time consuming. While stepwise selection is more computationally efficient, the discrete solution path can result in a model that is locally, but not globally, the best model. Silva and Skinner (1997) considered the use of best subsets and forward stepwise regression to estimate a finite population quantity under simple random sampling without replacement.

The "least absolute shrinkage and selection operator" (lasso) method proposed by Tibshirani (1996) simultaneously performs model selection and coefficient estimation by shrinking coefficients to zero. The lasso method finds coefficients that minimize the sum of the squared residuals subject to a constraint on the sum of the absolute value of the coefficients. More specifically, the coefficient estimates for lasso are given by:

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \, (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|, \quad (5)$$

where the estimate of the intercept β_o is not penalized and $\lambda \geq 0$ (Tibshirani 1996). The lasso model selection method is computationally efficient since the solution path is piece-wise linear (Efron, Hastie, Johnstone, and Tibshirani 2004) and it selects the global solution since the lasso criterion is convex. Therefore, the lasso method is often superior to the best subsets method and the stepwise method.

The literature on the lasso and related methods is vast and growing. Asymptotic results include Knight and Fu (2000), who derive limiting distributions of lasso-like estimators when (2) is assumed to hold. In the high-dimension/low sample size setting of more covariates than observations, various asymptotic conditions on the design matrix in (2) have been proposed to establish oracle inequalities and variable selection properties of the lasso. Among many examples, see Zou (2006), Bunea, Tsybakov, and Wegkamp (2007), Van De Geer and Bühlmann (2009), Raskutti, Wainwright, and Yu (2011), and the references therein. These results are not directly applicable in our context of sampling from a finite population and applying design-based methods because (2) is only a working model and not an inferential target. Nonetheless, our method's development below will proceed as if (2) holds with a fixed number of covariates, p.

1.3 Survey Regression Estimation with the Lasso

If the model in (2) is sparse, meaning only p_0 of the p coefficients are nonzero, then estimation of the zero coefficients in (4) leads to extra variation in (3). A reduced model could reduce the overall design variance of the GREG; hence we propose using the following survey-weighted lasso coefficient estimates in the GREG:

$$\hat{\boldsymbol{\beta}}_s^{(L)} = \underset{\beta}{\operatorname{argmin}}\,(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta})^T \Pi_s^{-1}(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}) + \lambda \sum_{i=1}^{p} |\beta_i|, \tag{6}$$

where $\lambda \geq 0$. In computing the coefficient estimates, one can leave a subset of the coefficients unpenalized by excluding those coefficients from the penalty term. In section 3, we discuss how to ensure the estimator is calibrated to the population totals of the unpenalized predictors, a problem of considerable interest in survey practice (e.g., Deville and Särndal 1992; Särndal 2010). The survey-weighted lasso coefficient estimates can be found using one of the various algorithms constructed to find (5) since we can rewrite (6) as

$$\hat{\boldsymbol{\beta}}_s^{(L)} = \underset{\beta}{\operatorname{argmin}}\,(\boldsymbol{Y}_s^* - \boldsymbol{X}_s^*\boldsymbol{\beta})^T (\boldsymbol{Y}_s^* - \boldsymbol{X}_s^*\boldsymbol{\beta}) + \lambda \sum_{i=1}^{p} |\beta_i|,$$

where $\boldsymbol{Y}_s^* = \Pi_s^{-1/2}\boldsymbol{Y}_s$, $\boldsymbol{X}_s^* = \Pi_s^{-1/2}\boldsymbol{X}_s$ and $\Pi_s^{-1/2} = \operatorname{diag}(\pi_j^{-1/2})_{j\in s}$. The lasso survey regression estimator for t_y is then

$$\hat{t}_{y,\text{lasso}} = \sum_{j\in s} \frac{y_j - \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(L)}}{\pi_j} + \sum_{j\in U} \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_{\text{s}}^{(L)}. \tag{7}$$

From (6) and (7), it is evident that only $\{\boldsymbol{x}_j\}_{j\in s}$ and $\sum_{j\in U}\boldsymbol{x}_j$ are required for computation of the lasso survey regression estimator.

For selecting the penalty parameter λ, one can use a survey-weighted version of the Akaike (1973) Information Criterion (AIC), the Schwarz (1978) Bayesian Information Criterion (BIC), or cross-validation. We used cross-validation to select the penalty parameter in the simulation study of section 4.

1.4 Survey Regression Estimation with the Adaptive Lasso

A shortcoming of the lasso criterion is that by shrinking it produces biased estimates for coefficients that are far from zero. In the adaptive lasso criterion function (Zou 2006), the coefficients in the l_1 penalty are weighted by the inverse of a root-n consistent estimator, and therefore large coefficients tend to have less bias.

To estimate the finite population total, t_y, we consider an adaptive lasso survey regression estimator,

$$\hat{t}_{y,\text{alasso}} = \sum_{j\in s}\frac{y_j - \boldsymbol{x}_j^T\hat{\boldsymbol{\beta}}_{\mathrm{s}}^{(AL)}}{\pi_j} + \sum_{j\in U}\boldsymbol{x}_j^T\hat{\boldsymbol{\beta}}_{\mathrm{s}}^{(AL)}, \tag{8}$$

where the estimated coefficient vector based on the sample is

$$\hat{\boldsymbol{\beta}}_s^{(AL)} = \underset{\beta}{\operatorname{argmin}}\,(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta})^T\Pi_s^{-}1(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}) + \lambda\sum_{i=1}^{p}\frac{|\beta_i|}{|\hat{\beta}_{si}|}, \tag{9}$$

and the equation for $\hat{\boldsymbol{\beta}}_s$ is found in (4). Only $\{\boldsymbol{x}_j\}_{j\in s}$ and $\sum_{j\in U}\boldsymbol{x}_j$ are required for computation of the adaptive lasso survey regression estimator.

To compute the survey-weighted adaptive lasso coefficient values, we can transform the criterion in (9) to look like the criterion in (5):

$$\left(\Pi_s^{-1/2}\boldsymbol{Y}_s - \Pi_s^{-1/2}\boldsymbol{X}_s\boldsymbol{V}^{-1}\mathbf{V}\boldsymbol{\beta}\right)^T\left(\Pi_s^{-}1/2\boldsymbol{Y}_s - \Pi_s^{-1/2}\boldsymbol{X}_s\boldsymbol{V}^{-1}\mathbf{V}\boldsymbol{\beta}\right) + \lambda\sum_{i=1}^{p}\frac{|\beta_i|}{|\hat{\beta}_{si}|}$$

$$(\boldsymbol{Y}_s^* - \boldsymbol{X}_s^*\boldsymbol{\beta}^*)^T(\boldsymbol{Y}_s^* - \boldsymbol{X}_s^*\boldsymbol{\beta}^*) + \lambda\sum_{i=1}^{p}|\beta_i^*|,$$

where $\boldsymbol{V}$ is the $(p+1)\times(p+1)$ diagonal matrix of the penalty vector $(1, |\hat{\beta}_{s1}|^{-1}, \ldots, |\hat{\beta}_{sp}|^{-1})$. Then we proceed with the original algorithm using the transformed covariate matrix $\boldsymbol{X}_s^* = \Pi_s^{-1/2}\boldsymbol{X}_s\boldsymbol{V}^{-1}$ and the transformed study variable vector $\boldsymbol{Y}_s^* = \Pi_s^{-1/2}\boldsymbol{Y}_s$ to obtain $\hat{\boldsymbol{\beta}}_s^{(AL)}$. The survey-weighted adaptive lasso coefficient values are found by back transforming, $\hat{\boldsymbol{\beta}}_s^{(AL)} = \boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}_s^{(AL)*}$, and are plugged into (8).

2. ASYMPTOTIC PROPERTIES OF THE LASSO SURVEY REGRESSION ESTIMATOR

We study the asymptotic properties of the lasso survey regression estimator in a design-based setting in which both sample size and population size go to infinity. The dominant error in survey regression estimation is the sampling error, as can be seen by writing (for any parameter estimates $\hat{\boldsymbol{\beta}}_N$ and corresponding finite population target $\boldsymbol{\beta}_N$)

$$|\hat{t}_y - t_y| = \left| \sum_{i \in U} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_N)\left(\frac{I_i}{\pi_i} - 1\right) + \left\{\sum_{i \in U} \boldsymbol{x}_i^T \left(\frac{I_i}{\pi_i} - 1\right)\right\}(\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N)\right|$$

$$\leq O_{\mathrm{P}}\left(\frac{N}{\sqrt{n}}\right) + \sum_{j=1}^{p} O_{\mathrm{P}}\left(\frac{N}{\sqrt{n}}\right) O_{\mathrm{P}}(|\beta_{Nj} - \hat{\beta}_{Nj}|),$$

where the $\sum_{i \in U} x_{ij}(I_i \pi_i^{-1} - 1)$ terms are uniformly $O_{\mathrm{P}}(N/\sqrt{n})$ for $j = 1, 2, \ldots, p$ under very mild design conditions. The usual $O_{\mathrm{P}}(N/\sqrt{n})$ of survey regression estimation, dominated by residual variation from $\{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_N\}$, will then continue to hold as long as the number of covariates grows slower than our ability to estimate their coefficients, $\max_{1 \leq j \leq p} O_{\mathrm{P}}(|\beta_{Nj} - \hat{\beta}_{Nj}|) = o_{\mathrm{P}}(p^{-1})$ uniformly for $j = 1, 2, \ldots, p$.

In this paper, we assume that the number of covariates is fixed and show in theorem 2.1 that the coefficient estimation error is $O_{\mathrm{P}}(N^{-1/2})$. The argument above suggests that our results will hold much more generally. Assumptions are specified in section 2.1 and are used in section 2.2 to establish design consistency of the lasso survey regression estimators and a design-based central limit theorem, showing that the lasso estimator has the same asymptotic properties as the GREG (result 5 of Deville and Särndal [1992] is a similar result for general calibration estimators). This means that the effect of shrinkage and selection is asymptotically negligible in the estimation of finite population totals. In finite samples, however, shrinkage and selection via the lasso can reduce the large weight adjustments often seen in GREG estimators and can lead to substantial efficiency gains, as shown via simulation in section 4. We also propose a variance estimator and establish its design consistency.

Throughout this section, we denote the survey-weighted lasso coefficients of (6) by $\hat{\boldsymbol{\beta}}_N$ to simplify the notation and to emphasize the dependence on N.

2.1 Design Assumptions

Consider the classical survey asymptotic framework of nested populations, $U_1 \subset U_2 \subset \cdots \subset U_N \subset \cdots$. Let $s_N \subset U_N$ be selected according to a sampling design $\mathrm{p}_N(\cdot)$, where $|s_N| = n_N = n$ is the size of the Nth sample. We suppress the subscript N in n as well as in π_j and π_{jk} for simplicity of notation. Let $\Delta_{jk} = \pi_{jk} - \pi_j\pi_k$. Assume the following conditions as $N \to \infty$ with p fixed:

D1. The penalty parameter satisfies $\lambda_N = o(\sqrt{N})$.

D2. The sampling rate $nN^{-1} \to \pi \in (0, 1)$.

D3.

- The matrices

$$\hat{\mathbf{C}}_N = \frac{1}{N}\sum_{i\in U_N} \mathbf{x}_i\mathbf{x}_i^T \frac{I_i}{\pi_i} \text{ and } \mathbf{C}_N = \frac{1}{N}\sum_{i\in U_N} \mathbf{x}_i\mathbf{x}_i^T$$

are positive definite and $\hat{\boldsymbol{C}}_N - \boldsymbol{C}_N = o_{\mathrm{p}}(1)$ elementwise.

- There exists $\boldsymbol{C}$ positive definite such that $\boldsymbol{C}_N - \boldsymbol{C} = o(1)$ elementwise.
- There exists $\boldsymbol{D} \in \mathbb{R}^{p+1}$ such that

$$\boldsymbol{D}_N - \boldsymbol{D} = \frac{1}{N}\sum_{i\in U_N} \boldsymbol{x}_i y_i - \boldsymbol{D} = o(1)$$

elementwise.

D3. The matrix $\boldsymbol{\Sigma} = \lim\limits_{\mathrm{N}\to\infty} \boldsymbol{\Sigma}_{\mathrm{N}}$ exists and is positive definite, where $\boldsymbol{\Sigma}_{\mathrm{N}}$ is the design covariance matrix

$$\boldsymbol{\Sigma}_{\mathrm{N}} = \begin{bmatrix} \Sigma_N^{(xyxy)} & \Sigma_N^{(xyxx_o)} & \cdots & \Sigma_N^{(xyxx_p)} \\ \Sigma_N^{(xx_oxy)} & \Sigma_N^{(xx_oxx_o)} & \vdots & \\ \vdots & \vdots & & \\ \Sigma_N^{(xx_pxy)} & \Sigma_N^{(xx_pxx_o)} & \cdots & \Sigma_N^{(xx_pxx_p)} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{n}{N^2}\sum\sum_{i,j\in U_N} \frac{\Delta_{ij}}{\pi_i\pi_j}\boldsymbol{x}_i y_i \boldsymbol{x}_j^T y_j & \cdots & \frac{n}{N^2}\sum\sum_{i,j\in U_N} \frac{\Delta_{ij}}{\pi_i\pi_j}\boldsymbol{x}_i y_i \boldsymbol{x}_j^T x_{jp} \\ \vdots & & \vdots \\ \frac{n}{N^2}\sum\sum_{i,j\in U_N} \frac{\Delta_{ij}}{\pi_i\pi_j}\boldsymbol{x}_i x_{ip} \boldsymbol{x}_j^T y_j & \cdots & \frac{n}{N^2}\sum\sum_{i,j\in U_N} \frac{\Delta_{ij}}{\pi_i\pi_j}\boldsymbol{x}_i x_{ip} \boldsymbol{x}_j^T x_{jp} \end{bmatrix}$$

of the following $(p + 2)(p + 1)$ vector of centered, standardized Horvitz-Thompson estimators:

$$z_N = \begin{bmatrix} \dfrac{\sqrt{n}}{N}\displaystyle\sum_{i\in U_N} \boldsymbol{x}_i y_i \left(\dfrac{I_i}{\pi_i}-1\right) \\ \left[\dfrac{\sqrt{n}}{N}\displaystyle\sum_{i\in U_N}\boldsymbol{x}_i x_{ik}\left(\dfrac{I_i}{\pi_i}-1\right)\right]_{k=0}^{p}\end{bmatrix}, \qquad (10)$$

where $x_{i0} \equiv 1$.

D4. The normalized, centered, Horvitz-Thompson estimators defined in (10) satisfy a central limit theorem: $z_N \rightarrow^{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

D5. The subvector $z_N^* = (z_{N,1}, z_{N,p+2}, z_{N,p+3}, \ldots, z_{N,2p+3})$ of the vector z_N defined in (10) has a design-consistent covariance matrix estimator

$$\begin{aligned}\hat{\Sigma}_N^* &= \begin{bmatrix} \dfrac{n}{N^2}\displaystyle\sum\sum_{i,j\in U_N}\dfrac{\Delta_{ij}}{\pi_i\pi_j}\dfrac{I_iI_j}{\pi_{ij}}y_iy_j & \dfrac{n}{N^2}\displaystyle\sum\sum_{i,j\in U_N}\dfrac{\Delta_{ij}}{\pi_i\pi_j}\dfrac{I_iI_j}{\pi_{ij}}y_i\boldsymbol{x}_j^T \\ \dfrac{n}{N^2}\displaystyle\sum\sum_{i,j\in U_N}\dfrac{\Delta_{ij}}{\pi_i\pi_j}\dfrac{I_iI_j}{\pi_{ij}}\boldsymbol{x}_iy_j & \dfrac{n}{N^2}\displaystyle\sum\sum_{i,j\in U_N}\dfrac{\Delta_{ij}}{\pi_i\pi_j}\dfrac{I_iI_j}{\pi_{ij}}\boldsymbol{x}_i\boldsymbol{x}_j^T \end{bmatrix} \\ &= \begin{bmatrix} \hat{\Sigma}_N^{(yy)} & \hat{\Sigma}_N^{(yx)} \\ \hat{\Sigma}_N^{(xy)} & \hat{\Sigma}_N^{(xx)} \end{bmatrix},\end{aligned}$$

in the sense that $\hat{\Sigma}_N^* - \Sigma_N^* = o_{\mathrm{p}}(1)$ elementwise where Σ_N^* is the covariance matrix of z_N^*.

Remark 1. Our asymptotic formulation of fixed p and penalty growing as in (D1) is chosen for the finite population regression estimation context in which the dominant errors are sampling errors, not the errors in estimation of finite population regression coefficients. In our section 4 application to an environmental resource survey, N grows like the number of pixels in an image of a landscape, while p grows like the number of information layers (e.g., imagery types) for that landscape, which is necessarily very small relative to N. Hence we have not considered formulations in which p grows with N, though such an asymptotic structure might usefully extend the scope of LASSO application in surveys. Such formulations are beyond the scope of this paper.

Remark 2. Our main interest is in asymptotic comparison of the GREG to the lasso survey regression estimator, so our assumptions are sufficient to ensure design consistency and asymptotic normality of the GREG and design consistency of its conventional variance estimator. Weaker assumptions, like those in Isaki and Fuller (1982), Robinson and Särndal (1983), or Breidt and Opsomer (2000), could be used to establish results of the type assumed here.

Remark 3. The assumed design-based central limit theorem in (D5) is fundamental in survey practice, in which weighted point estimates are computed, variances are estimated accounting for the complexity of the design, and corresponding normal-theory confidence intervals are constructed. Examples of designs under which such central limit theory holds include simple random sampling with and without replacement (Hájek 1960), stratified unequal probability samples with replacement (Krewski and Rao 1981), and stratified simple random sampling without replacement (Bickel and Freedman 1984), among many others. See Fuller (2009, section 1.3) for review of some of the relevant literature.

Remark 4. Assumption (D3) ensures that the finite population coefficient vector defined by

$$\boldsymbol{\beta}_N = (\boldsymbol{X}_U^T \boldsymbol{X}_U)^{-1} \boldsymbol{X}_U^T \boldsymbol{Y}_U,$$

where $\boldsymbol{X}_U = [\boldsymbol{x}_j^T]_{j \in U}$ is an $N \times (p+1)$ matrix and $\boldsymbol{Y}_U = [y_j]_{j \in U}$ is an N-vector, converges as $N \to \infty$ to the vector $\boldsymbol{\beta}^* = \boldsymbol{C}^{-1}\boldsymbol{D} \in \mathbb{R}^{p+1}$. The assumption does not imply that $\boldsymbol{\beta}_N$ converges to $\boldsymbol{\beta}$, the coefficient vector in the working model.

Remark 5. Since $\boldsymbol{x}_i$ contains an intercept term, (D4) covers $\lim\limits_{N \to \infty} \Sigma_N^{(x_o y x_o y)} = \Sigma^{(x_o y x_o y)}$ where

$$\Sigma_N^{(x_o y x_o y)} = \Sigma_N^{(yy)} = \frac{n}{N^2} \sum\sum_{i,j \in U_N} \frac{\Delta_{ij}}{\pi_i \pi_j} y_i y_j,$$

and similarly $\lim\limits_{N \to \infty} \Sigma_N^{(x x_o x x_o)} = \Sigma^{(x x_o x x_o)}$ where

$$\Sigma_N^{(x x_o x x_o)} = \Sigma_N^{(xx)} = \frac{n}{N^2} \sum\sum_{i,j \in U_N} \frac{\Delta_{ij}}{\pi_i \pi_j} \boldsymbol{x}_i \boldsymbol{x}_j^T.$$

2.2 Asymptotic Results

Proofs of the results in this section are omitted but are detailed in McConville (2011). We first establish a central limit theorem for the survey-weighted lasso coefficients as estimates of the finite population coefficients.

Theorem 2.1 Under assumptions (D1)–(D5), the survey-weighted lasso coefficients $\hat{\boldsymbol{\beta}}_N$ satisfy

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_N) \xrightarrow{D} \mathcal{N}(0, \pi^{-1} \boldsymbol{C}^{-1} \boldsymbol{V} \boldsymbol{C}^{-1})$$

as $N \to \infty$, where the matrix $\mathbf{V}$ is defined by

$$\boldsymbol{V} = \Sigma^{(xyxy)} - 2 \sum_{k=0}^{p} \beta_k^* \Sigma^{(x x_k x y)} + \sum_{k=0}^{p} \sum_{l=0}^{p} \beta_k^* \beta_l^* \Sigma^{(x x_k x x_l)}$$

and $[\beta_k^*]_{k=0}^p = \boldsymbol{C}^{-1}\boldsymbol{D}$.

Unlike theorem 2 of Knight and Fu (2000), theorem 2.1 is derived in the finite population setting without independent and identically distributed errors and without a true parameter vector $\boldsymbol{\beta}$. The result implies that the effect of shrinkage and selection is asymptotically negligible since the survey-weighted lasso coefficient estimators consistently target the finite population regression parameters, just like the usual weighted least squares estimators in (4). It follows immediately that the model-assisted estimator with survey-weighted lasso coefficients will be asymptotically equivalent to the difference estimator, with estimated coefficients replaced by finite population coefficients, a result stated in the next theorem. The GREG shares this equivalence.

Theorem 2.2 Under assumptions (D1)–(D5), the estimator $\hat{t}_{\boldsymbol{y},}$lasso is asymptotically equivalent to the difference estimator,

$$\hat{t}_{y,\text{diff}} = \sum_{j \in s} \frac{y_j - \boldsymbol{x}_j^T \boldsymbol{\beta}_N}{\pi_j} + \sum_{j \in U_N} \boldsymbol{x}_j^T \boldsymbol{\beta}_N,$$

in the sense that

$$\frac{\sqrt{n}}{N} \left(\hat{t}_{y,\text{lasso}} - \hat{t}_{y,\text{diff}} \right) = o_{\text{p}}(1),$$

so that

$$\left\{ \text{Var}_{\text{p}} \left(\hat{t}_{y,\text{diff}} \right) \right\}^{-1/2} \left(\hat{t}_{y,\text{lasso}} - t_y \right) \xrightarrow{D} N(0,1), \tag{11}$$

where

$$\text{Var}_{\text{p}} \left(\hat{t}_{y,\text{diff}} \right) = \sum\sum_{i,j \in U_N} \Delta_{ij} \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_N}{\pi_i} \frac{y_j - \boldsymbol{x}_j^T \boldsymbol{\beta}_N}{\pi_j}.$$

A standard variance estimator with exactly the same form as that of the GREG is design consistent for the variance of the difference estimator and hence can be plugged into the central limit theorem of (11) and used to generate asymptotically valid confidence intervals. This is the content of the following theorem and corollary.

Theorem 2.3 Under assumptions (D1)–(D6),

$$\hat{V}\left(\hat{t}_{y,\text{lasso}} \right) = \sum\sum_{i,j \in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_N \right)}{\pi_i} \frac{\left(y_j - \boldsymbol{x}_j^T \boldsymbol{\beta}_N \right)}{\pi_j} \tag{12}$$

$$= \mathrm{Var}_{\mathrm{p}}\left(\hat{t}_{y,\mathrm{diff}}\right) + o_{\mathrm{p}}\left(\frac{N^2}{n}\right).$$

Corollary 1 Under assumptions (D1)–(D6),

$$\{\hat{V}(\hat{t}_{y,\mathrm{lasso}})\}^{-1/2}(\hat{t}_{y,\mathrm{lasso}} - t_y) \xrightarrow{D} N(0,1).$$

3. WEIGHTS FOR THE LASSO SURVEY REGRESSION ESTIMATOR

In practice, it is often the case that many, possibly hundreds of, finite population quantities need to be estimated from the same survey data set. The GREG has the interesting property that

$$\hat{t}_{y,\mathrm{greg}} = \sum_{j \in s}\left[1 + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x,HT})^T\left(\sum_{k \in s}\frac{\boldsymbol{x}_k\boldsymbol{x}_k^T}{\pi_k}\right)^{-1}\boldsymbol{x}_j\right]\frac{1}{\pi_j}y_j = \sum_{j \in s}\omega_j(s)y_j, \quad (13)$$

where $\boldsymbol{t}_x$ is the population total vector of the covariates and $\hat{\boldsymbol{t}}_{x,HT}$ is the corresponding Horvitz-Thompson estimator vector of the covariate totals (Särndal, Swensson, and Wretman 1992; section 6.5). The regression weights $\{\omega_j(s)\}_{j \in s}$ do not depend on y and so can be applied to any response variable. As long as the study variables relate even weakly with the covariates, the GREG weights produce a more efficient estimator than the Horvitz-Thompson weights $\{\pi_j^{-1}\}_{j \in s}$.

A drawback of the lasso survey regression estimator is the lack of regression weights since the lasso coefficients are not linear combinations of the y-values. We consider two approaches to address this drawback and generate regression weights for the lasso survey regression estimator: a model calibration approach in section 3.1 and a ridge regression approximation in section 3.2.

3.1 Weights via Model Calibration

Since the lasso method does not produce an estimator that is linear in y, the lasso survey regression estimator cannot be written as a linear combination of the y values in the sample. To obtain weights, we employ the model calibration method of Wu and Sitter (2001), used in related contexts by Montanari and Ranalli (2005b) and Opsomer et al. (2007). The resulting calibration estimator can be written as a weighted sum of the sampled study variable as in (13) but with the caveat that the weights now depend on the sampled study variable, y.

The lasso calibration estimator is found by regressing the study variable, y_j, on an intercept and the lasso-fitted mean function, $\boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(L)}$, over the sample. Because this calibration step involves linear regression, the lasso calibration estimator can be written in the same form as (13) where $\boldsymbol{x}_j$ is replaced by $\boldsymbol{x}_j^* = (1, \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(L)})^T$:

$$\hat{t}_{y,\text{cal}} = \sum_{j \in s} \left[1 + (\boldsymbol{t}_{x^*} - \hat{\boldsymbol{t}}_{x^*,HT})^T \left(\sum_{k \in s} \frac{\boldsymbol{x}_k^* \boldsymbol{x}_k^{*T}}{\pi_k} \right)^{-1} \boldsymbol{x}_j^* \right] \frac{1}{\pi_j} y_j. \tag{14}$$

Since $\boldsymbol{x}_j^T \boldsymbol{\beta}_s^{(L)}$ is dependent on $\{\boldsymbol{x}_j, y_j\}_{j \in s}$, the weights in the lasso calibration estimator are dependent on the study variable, y. This dependence implies that the utility of applying these weights to other study variables depends on how correlated the variables are with y.

In section 4, we compare the lasso calibration estimator with various other finite population total estimators and consider an adaptive lasso calibration estimator where the lasso-fitted mean function in $\boldsymbol{x}_j^*$ of (14) is replaced with the adaptive lasso fit, $\boldsymbol{x}_j^{**} = (1, \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(AL)})^T$:

$$\hat{t}_{y,\text{acal}} = \sum_{j \in s} \left[1 + (\boldsymbol{t}_{x^{**}} - \hat{\boldsymbol{t}}_{x^{**},HT})^T \left(\sum_{k \in s} \frac{\boldsymbol{x}_k^{**} \boldsymbol{x}_k^{**T}}{\pi_k} \right)^{-1} \boldsymbol{x}_j^{**} \right] \frac{1}{\pi_j} y_j. \tag{15}$$

The weights in (14) and (15) are calibrated to the population size N and to the population total of the lasso-fitted mean function. If desired, additional auxiliary variables can be added to $\boldsymbol{x}_j^*$ or $\boldsymbol{x}_j^{**}$ to force exact calibration to the corresponding population totals. Added auxiliary variables are included in the working regression model even if they were eliminated in the lasso estimation.

3.2 Weights via Ridge Regression Approximation

In a model-based framework, Bardsley and Chambers (1984) introduced and Chambers (1996) extended a ridge regression estimator for estimating finite population quantities when there are many potential predictors and multicollinearity may be a problem. Similar ridge regression estimators have been developed in the model-assisted, design-based context by Rao and Singh (1997) and Théberge (2000); see Beaumont and Bocci (2008) for a review of this and related methods. The form of the model-assisted ridge regression estimator is

$$\hat{t}_{y,\text{rr}} = \sum_{j \in s} \left[1 + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x,HT})^T \left(\sum_{k \in s} \frac{\boldsymbol{x}_k \boldsymbol{x}_k^T}{\pi_k} + \Lambda \right)^{-1} \boldsymbol{x}_j \right] \frac{1}{\pi_j} y_j \tag{16}$$

where Λ is a diagonal matrix of non-negative cost terms. The ridge regression weights are typically less variable than the GREG weights. In particular, Λ can be chosen so that the resulting weights are non-negative.

Although the lasso coefficients do not have a closed form solution, Tibshirani (1996) approximated the coefficient estimates with a ridge regression formula to derive their standard errors. We use this ridge regression approximation as another way to construct regression weights for the lasso estimator. Tibshirani (1996) rewrote the penalty term as $\sum_{i=1}^{p} \beta_i^2 |\beta_i|^{-1}$, allowing him to obtain the ridge regression coefficient estimates, to which we have added design weights:

$$\hat{\boldsymbol{\beta}}_s^{(\text{ridge})} = (\boldsymbol{X}_s^T \Pi_s^{-1} \boldsymbol{X}_s + \mu \boldsymbol{Q}^{-})^{-1} \boldsymbol{X}_s^T \Pi_s^{-1} \boldsymbol{Y}_s, \tag{17}$$

where $\boldsymbol{Q}$ is the diagonal matrix of the vector $(0, |\hat{\beta}_{s1}^{(L)}| + \eta, \ldots, |\hat{\beta}_{sp}^{(L)}| + \eta)$, η is a small positive number, and $\boldsymbol{Q}^{-}$ is a generalized inverse of $\boldsymbol{Q}$. The ridge regression penalization has a negative correlation with the magnitude of the lasso coefficients. To achieve exact calibration on the population total of a predictor, the corresponding value in the diagonal of $\boldsymbol{Q}$ should be set to zero. The penalty parameter μ is chosen so that $\sum_{i=1}^{p} |\hat{\beta}_{si}^{(\text{ridge})}| = \sum_{i=1}^{p} |\hat{\beta}_{si}^{(L)}|$, where $\hat{\beta}_{si}^{(L)}$ is defined in (6). The lasso ridge regression estimator is

$$\hat{t}_{y,\text{ridge}} = \sum_{j \in s} \left[1 + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x,HT})^T \left(\sum_{k \in s} \frac{\boldsymbol{x}_k \boldsymbol{x}_k^T}{\pi_k} + \mu \boldsymbol{Q}^{-} \right)^{-1} \boldsymbol{x}_j \right] \frac{1}{\pi_j} y_j. \tag{18}$$

This estimator has the same form as (16), with Λ replaced by $\mu^{-1}\boldsymbol{Q}$.

We can also construct an adaptive lasso ridge regression estimator, analogous to the adaptive lasso calibration estimators. The regression coefficient estimates take the form of (17), but $\boldsymbol{Q}$ is now the diagonal matrix of the vector $(0, |\hat{\beta}_{s1}^{(L)}||\hat{\beta}_{s1}| + \eta, \ldots, |\hat{\beta}_{sp}^{(L)}||\hat{\beta}_{sp}| + \eta)$. The weights in (18) are again dependent on the study variable, y, because the weights are a function of the lasso coefficients, $\hat{\boldsymbol{\beta}}_s^{(L)}$. In section 4, we compare all of the lasso-based estimators to the Horvitz-Thompson estimator, the full regression estimator, a regression estimator with forward stepwise selection, and the classic ridge regression estimator.

4. LASSO SURVEY REGRESSION ESTIMATION FOR TREE CANOPY COVER IN UTAH

To compare the lasso-based regression estimators to other estimators, we conducted a simulation study using photo-interpreted tree canopy cover sample data and ancillary layers of processed remote sensing data in Utah. The estimators considered are detailed below.

Abbreviation	Estimator	Equation
LASSO	Lasso survey regression estimator	(7)
ALASSO	Adaptive lasso survey regression estimator	(8)
CLASSO	Lasso calibration estimator	(14)
CALASSO	Adaptive lasso calibration estimator	(15)
RLASSO	Lasso ridge regression estimator	(18)
RALASSO	Adaptive lasso ridge regression estimator	(18) with modified $\boldsymbol{Q}$
RIDGE	Survey ridge regression estimator	(16)
FSTEP	Forward stepwise regression estimator	
GREG	Survey regression estimator	(13)
HT	Horvitz-Thompson estimator	(1)

All computations were completed in R (R Core Team 2015). The LASSO and ALASSO coefficent estimates were computed using the function `glmnet` in the `glmnet` package (Friedman, Hastie, and Tibshirani 2010). The function `cv.glmnet`, which allows for the inclusion of weights, was used to select the penalty parameter. McConville (2011) also considered survey-weighted versions of AIC, BIC, and cross-validation to select the penalty parameter; all yielded similar results in terms of the mean squared error of the resulting survey regression estimators. For the RIDGE estimator, we chose $\Lambda = \gamma \boldsymbol{I}$ with γ selected as the smallest positive value so that the weights are all greater than one. For both RLASSO and RALASSO, we selected μ similarly. The FSTEP estimator was fit using the function `step` in the `leaps` package (Lumley 2009).

McConville (2011) compared these model-assisted estimators to the corresponding model-based estimators and found the model-based estimators to be much less efficient once the sampling design was informative. We will restrict our attention to model-assisted estimators, comparing the various lasso-based estimators to the full regression estimator, the classic ridge regression estimator, and the Horvitz-Thompson estimator. We caution that our asymptotic theory strictly applies only to LASSO, so we evaluate our weighted approximations to LASSO via simulation.

4.1 Utah Tree Canopy Cover Data Set

The quantity of interest is the total amount of tree canopy cover for a region of Utah. Understanding and quantifying tree canopy cover, which is an aerial measure of the amount of ground covered by tree crowns, is relevant to many applications, including forest management, fire modeling, air pollution mitigation, stream and water temperature, and carbon storage. The photo-interpreted data used here arise from a national pilot project conducted by FIA and the U.S. Forest Service Remote Sensing Applications Center as part of the development of the updated 2011 National Land Cover Database (NLCD) tree canopy cover layer. An intensive (approximately 1 km x 1 km) grid of photo plots was established over a pilot area the approximate size of one Landsat scene in southern Utah. Each photo plot consisted of 105 dots distributed in a 90 m × 90 m square area. Each dot was characterized as falling on a tree crown or not. The response variable of percent tree canopy cover was defined as the proportion of tree dots identified on the photo plot. Predictor variables included transformed aspect, slope, topographic positional index, elevation, land cover and tree canopy cover from the 2001 NLCD (Homer, Huang, Yang, Wylie, and Coan 2004) and Landsat-5 reflectance bands. See Coulston, Moisen, Wilson, Finco, Cohen, et al. (2012) for more details on the data used in this study.

Each of the auxiliary variables is available at a finer resolution than the photo-interpreted data. The auxiliary variables were collected on a 30 by 30 meter grid, and therefore there are nine observations of every covariate for each photo-interpreted observation. To collapse the auxiliary information, the mean, standard deviation, minimum, and maximum are taken of the nine observations. All variables are standardized by subtracting the empirical mean and dividing by the empirical standard deviation. The following variables are included in the working model:

Variable	Description
CAN_MEAN	Mean of 2001 National Land Cover Database tree canopy cover estimates
CAN_STD	Standard deviation of 2001 National Land Cover Database tree canopy cover estimates
CTI_MEAN	Mean of the Compound Topographic Index
CTI_STD	Standard deviation of the Compound Topographic Index
DEM_MEAN	Mean of the Digital Elevation Model
DEM_STD	Standard deviation of the Digital Elevation Model
SLOPE_MEAN	Mean of the slope
SLOPE_STD	Standard deviation of the slope
TASPCOS_MIN	Minimum of the cosine transformed aspect
TASPCOS_MAX	Maximum of the cosine transformed aspect
TASPCOS_MEAN	Mean of the cosine transformed aspect
TASPCOS_STD	Standard deviation of cosine transformed aspect

We treated the high intensity grid of $N = 4{,}151$ grid points as the finite population of interest. We emphasize that in our simulation and in other real applications with remote sensing and digital elevation data of these types, all of the auxiliary variables listed above are available at every grid point, as are any transformations of the variables, singly or in combination with others (e.g., interaction terms). That is, we have known $\{\boldsymbol{x}_j\}_{j \in U}$, not just our minimal requirement of known $\{\boldsymbol{x}_j\}_{j \in s}$ and $\sum_{j \in U} \boldsymbol{x}_j$.

In regressing tree canopy cover for the entire finite population on an additive model of the $p = 12$ covariates above, the coefficient of determination was $R^2 = 0.57$ and only half of the regression coefficients were statistically different from zero. The regression on the model with main effects and all two-way interactions ($p = 78$) had $R^2 = 0.60$ and 11 regression coefficients were statistically different from zero. These population-level results suggest that useful predictive models should be sparse, so that model selection is appropriate.

We selected $M = 2000$ replicate samples using both simple random sampling without replacement (SI) and stratified simple random sampling without replacement (STSI). The ten counties in the population served as the strata. Sample sizes for the ten strata were $n \times (0.3, 0.1, 0.04, 0.04, 0.06, 0.06, 0.06, 0.1, 0.2, 0.04)$ for $n = 50$, 100, and 200. This stratified sampling scheme resulted in an informative, unequal probability sampling design where the inclusion probabilities were positively correlated with the study variable.

4.2 Design Bias and Design Mean Squared Error

We computed design bias and design mean squared error (MSE) by averaging across the M replicate samples. The percent relative design bias was less than 2.6 percent in absolute value for all of the estimators under all of the sampling schemes, with three exceptions: RIDGE under STSI with $p = 12$, $n = 50$, –5.0 percent relative bias and GREG under SI, and STSI with $p = 78$, $n = 100$, 7.45 percent relative bias, and 5.85 percent relative bias, respectively. Table 1 displays the ratio of the design MSE of each estimator to that of the LASSO. As the sample size increases, the substantial efficiency advantage of the lasso estimators over the GREG becomes less pronounced, a result consistent with the asymptotic theory. The calibration or ridge approximation used to obtain weights does not seem to decrease the design efficiency.

The adaptive lasso method appears to have higher design MSE than the lasso counterpart, and we conjecture this is because the covariates were standardized. The loss of efficiency of the adaptive method is particularly acute for the larger model; this is in large part due to the reliance of the adaptive method on the initial round of estimates, which are extremely variable (as shown by the poor performance of GREG). McConville (2011) conducted simulation studies similar to Example 4 in Tibshirani (1996), where the working model was large but the true model was sparse, and found the adaptive lasso method

Table 1. Ratio of Design MSE for Each Estimator to Design MSE of LASSO under Simple Random Sampling without Replacement and Stratified Simple Random Sampling without Replacement; 2 Working Models Considered: Additive Model (p = 12) and Two-Way Interaction Model (p = 78)

	$p = 12$				$p = 78$			
	SI		STSI		SI		STSI	
	$n = 50$	$n = 100$	$n = 50$	$n = 100$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
ALASSO	1.08	1.03	1.21	1.01	1.35	1.09	3.12	1.30
CLASSO	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00
CALASSO	1.10	1.03	1.24	1.02	1.40	1.11	3.46	1.34
RLASSO	1.06	1.05	0.95	1.00	1.06	1.04	0.97	0.97
RALASSO	1.09	1.05	0.99	1.02	1.10	1.06	1.20	1.05
RIDGE	1.11	1.07	1.01	1.05	1.32	1.21	1.32	1.35
FSTEP	1.17	1.06	1.46	1.11	2.22	1.19	17.02	2.20
GREG	1.50	1.14	1.91	1.24	209.49	3.57	887.05	20.55
HT	1.98	2.12	1.85	1.95	2.10	2.12	1.98	2.09

MSE, mean squared error; SI, simple random sampling without replacement; STSI, stratified random sampling without replacement.

did produce coefficient estimates closer to the true model coefficients and therefore was more design efficient.

In the Utah simulation study with $p = 12$ predictors, the lasso selected models with 4.49 variables on average, the adaptive lasso selected models with 4.05 variables, and forward stepwise selected models with 3.35 variables. When there were $p = 78$ potential predictors, the lasso, adaptive lasso, and forward stepwise selected, on average, 10.32 variables, 13.55 variables, and 13.22 variables, respectively.

4.3 Variance Estimation and Confidence Interval Coverage

Variance estimators based on (12) were constructed for each estimator. We also constructed the alternate variance estimator presented by Särndal, Swensson, and Wretman (1989) where the π_i^{-1}, π_j^{-1} in (12) are replaced by the corresponding weights. Since LASSO and ALASSO cannot be rewritten as a weighted sum of the response variable, they do not have an alternate variance estimator.

As seen in table 2, the alternative variance estimators for the lasso estimators have substantial negative bias, but less negative bias than those of the RIDGE, FSTEP, or the GREG. The standard variance estimator performed slightly worse than the alternative variance estimator in nearly all cases and is not displayed here. The performance of the theoretically unbiased variance estimator for HT is included for comparison. The bias reduces in all cases at the larger

Table 2. Percent Relative Bias of Variance Estimators, with the Standard Variance Estimator (12) Used for the Nonlinear Estimators LASSO and ALASSO, and the Alternate Variance Estimator (See Text) Used for All Other Cases

	$p = 12$				$p = 78$			
	SI		STSI		SI		STSI	
	$n = 50$	$n = 100$	$n = 50$	$n = 100$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
LASSO	−23	−15	−38	−26	−26	−17	−37	−30
ALASSO	−30	−18	−51	−29	−44	−27	−78	−48
CLASSO	−22	−14	−39	−26	−27	−18	−38	−31
CALASSO	−29	−17	−52	−28	−47	−28	−80	−50
RLASSO	−24	−15	−39	−30	−35	−23	−43	−34
RALASSO	−27	−15	−42	−32	−34	−25	−43	−33
RIDGE	−28	−15	−40	−33	−28	−16	−38	−28
FSTEP	−29	−16	−54	−34	−65	−32	−87	−69
GREG	−28	−17	−46	−35	−78	−45	−79	−61
HT	−1	−2	4	−3	−6	2	−3	−4

SI, simple random sampling without replacement; STSI, stratified random sampling without replacement.

Table 3. Coverage of Nominal 95% Confidence Intervals, with the Standard Variance Estimator (12) Used for the Nonlinear Estimators LASSO and ALASSO, and the Alternate Variance Estimator (See Text) Used for All Other Cases

	$p = 12$				$p = 78$			
	SI		STSI		SI		STSI	
	$n = 50$	$n = 100$	$n = 50$	$n = 100$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
LASSO	90	93	82	89	89	92	86	88
ALASSO	89	92	78	89	86	90	79	86
CLASSO	90	93	81	89	89	92	85	88
CALASSO	89	93	79	89	86	90	77	86
RLASSO	89	92	82	89	88	91	84	88
RALASSO	89	92	81	88	88	90	85	89
RIDGE	88	93	80	87	90	92	85	89
FSTEP	89	93	79	88	82	89	64	77
GREG	88	92	80	86	65	83	61	76
HT	94	94	91	92	93	95	92	93

SI, simple random sampling without replacement; STSI, stratified random sampling without replacement.

sample size. The bias in the variance estimators leads to undercoverage of nominal 95 percent confidence intervals, as shown in table 3. The coverage improves slightly in all cases using the alternative variance estimator, and improves as sample size increases.

Given the coverage levels seen here, a natural concern is the quality of the normal approximations in theorem 2.1 and corollary 1. In these simulations, the distributions of the lasso estimators of individual regression coefficients may or may not be well approximated by normal distributions: if the coefficient is small and likely to be shrunk to zero, then of course the normal approximation is poor. Such effects are, however, washed out in the lasso survey regression estimators of the total. Conventional diagnostics indicate that the distributions of the lasso survey regression estimators are well approximated by normal distributions, even at sample size 50. The GREG estimator has somewhat heavier tails and a worse approximation to normality, as expected given the large weight variation in GREG.

4.4 Properties of the Survey Weights

As discussed in section 3.1, a single set of weights is often applied to several study variables, with estimators taking the form of a linear combination of the sampled study variable. The *j*th weight can be heuristically interpreted as the number of similar elements in the population that the *j*th element in the sample represents. Large differences in values of weights are undesirable, as they allow some elements to be much more influential than others. Positive weights are essential because a negative weight no longer carries the described interpretation and can lead to nonsensical estimates. There is an extensive survey literature on weight properties, particularly in the construction of calibration estimators (Deville and Särndal 1992; Särndal 2010) and in the construction of model-assisted estimators with many calibration constraints (Rao and Singh 1997; Théberge 2000; Beaumont and Bocci 2008). All of the model-assisted estimators presented here that can be written as linear combinations of y variables have weights of the form $\pi_j^{-1} + w_j^*$, where the first component is the Horvitz-Thompson weight and the second component is the model adjustment. We now study the properties of the weights for the estimators in our simulation experiment.

In calibration estimation (Deville and Särndal 1992; Särndal 2010), weights are constructed that reproduce known population-level information, while remaining as close as possible (under some metric) to the original Horvitz-Thompson weights. One way to assess this calibration property of the weights is to compute

$$\frac{1}{M}\sum_{j=1}^{M}\frac{|w_j - \pi_j^{-1}|}{\pi_j^{-1}} \times 100\%,$$

the percent relative average absolute distance between the Horvitz-Thompson weights and the weights for the various methods. As shown in table 4, the calibration weights moved the least since the model adjustment is only calibrating

Table 4. Percent Relative Average Absolute Distance between the Survey Weights and the Horvitz-Thompson Weights

	$p = 12$				$p = 78$			
	SI		STSI		SI		STSI	
	$n = 50$	$n = 100$	$n = 50$	$n = 100$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
CLASSO	9.79	6.92	14.63	10.32	6.88	4.70	10.07	7.47
CALASSO	10.12	6.94	15.61	10.56	7.23	4.79	12.16	7.90
RLASSO	29.99	23.98	41.14	30.25	25.40	19.79	34.36	23.70
RALASSO	30.23	23.33	41.66	30.50	24.91	19.38	32.63	22.60
RIDGE	34.15	26.51	41.96	33.14	24.13	19.56	33.32	22.39
FSTEP	18.69	12.55	43.76	24.93	29.64	17.14	162.83	57.20
GREG	59.74	31.66	140.00	62.86	741.62	126.22	5,238.29	566.33

SI, simple random sampling without replacement; STSI, stratified random sampling without replacement.

the estimator on the population size and the fitted values. As the estimators are calibrated on more quantities, the model adjustment term becomes more variable and the weights show greater movement, with the GREG weights moving the most.

To understand better how the weights vary within a sample, we computed the average over replicate samples of the empirical within-sample variance of the weights:

$$\overline{var}(\boldsymbol{w}) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{var}\left(\left\{w_j^{(m)}\right\}_{j\in s^{(m)}}\right) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n-1}\sum_{j\in s^{(m)}}\left(w_j^{(m)} - \bar{w}^{(m)}\right)^2,$$

where $s^{(m)}$ is the mth replicate sample, $\bar{w}^{(m)} = n^{-1}\sum_{j\in s^{(m)}} w_j^{(m)}$ and $w_j^{(m)}$ is the jth weight in the mth replicate sample.

We are also interested in how much the weight for element $j \in U$ varies from sample to sample when element j is in the sample. We computed the empirical variance of the weight for element j across all M_j replicate samples in which j appeared. We then averaged these empirical variances across all j in the universe:

$$\begin{aligned}\overline{var}(w_j|j\in s) &= \frac{1}{N}\sum_{j\in U}\mathrm{var}(w_j|j\in s)\\ &= \frac{1}{N}\sum_{j\in U}\frac{1}{M_j-1}\sum_{m=1}^{M}\left(w_j^{(m)} - \bar{w}_j\right)^2 I_{\{j\in s^{(m)}\}},\end{aligned}$$

where $\bar{w}_j = M_j^{-1}\sum_{m=1}^{M} w_j^{(m)} I_{\{j\in s^{(m)}\}}$ and $M_j = \sum_{m=1}^{M} I_{\{j\in s^{(m)}\}}$.

Table 5. Average Variances for Weights within and across Samples

	Weight variances for $p = 12$							
	$\overline{var}(\boldsymbol{w})$				$\overline{var}(w_j \mid j \in s)$			
	SI		STSI		SI		STSI	
Estimators	$n = 50$	$n = 100$	$n = 50$	$n = 100$	$n = 50$	$n = 100$	$n = 50$	$n = 100$
CLASSO	147	18	7,630	1,869	144	18	1,134	154
CALASSO	162	18	7,673	1,870	158	18	1,285	158
RLASSO	1104	186	7,934	1,973	1,080	184	4,145	848
RALASSO	1147	179	8,076	1,985	1,121	177	4,494	885
RIDGE	1404	228	6,990	1,923	1,380	226	4,093	990
FSTEP	666	67	11,388	2,109	599	65	8,351	877
GREG	5252	352	26,201	3,026	3,761	315	24,685	2,658
HT	0	0	7,490	1,854	0	0	0	0
	Weight variances for $p = 78$							
	$\overline{var}(\boldsymbol{w})$				$\overline{var}(w_j \mid j \in s)$			
	SI		**STSI**		**SI**		**STSI**	
Estimators	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
CLASSO	18	2	1,862	555	18	2	148	26
CALASSO	22	2	1,946	556	21	2	310	32
RLASSO	229	37	1,871	555	227	37	757	180
RALASSO	227	36	1,914	557	225	36	837	183
RIDGE	228	40	1,760	537	227	39	698	173
FSTEP	819	30	38,515	1,159	516	28	40,373	1,083
GREG	477,982	1,497	3,579,436	24,977	325,841	869	3,268,319	31,495
HT	0	0	1,854	554	0	0	0	0

SI, simple random sampling without replacement; STSI, stratified random sampling without replacement.

Table 5 displays both of these variance statistics for the weights within and across the samples. We only see variability in the HT weights under the unequal probability sampling and when measuring the variance within the sample. Since the calibration estimator weights are highly correlated with the HT weights, the variance measures are only slightly higher for the calibration estimators than for the HT. The GREG weights have the highest variability because of the need to calibrate on many covariates. The FSTEP weights are less variable than GREG, but substantially more variable than those from the lasso methods when $p = 78$ and $n = 100$ and under the unequal probability sampling designs. Although the RIDGE, RLASSO, and RALASSO also are based on a full model, the variability in the weights is controlled by the penalization.

A concern with highly variable weights is the possibility of negative weights. The average percentage of negative weights was computed for each sampling design. The calibration weights were negative in only 0.027 percent of all cases and the ridge regression weights were never negative by construction, while the FSTEP weights varied from 0.33 percent to 13 percent negative weights and the GREG weights varied from 1 percent to 46 percent negative weights, on average, depending on the sampling design.

Although the small variability in the weights of the calibration estimators is desirable, the weights still depend on the study variable, y. Both the GREG and HT weights are independent of y and only depend on the sample, s. Therefore, it is important to assess how well the y-dependent weights perform, in comparison with the y-independent weights, when applied to other study variables of interest. McConville (2011) conducted extensive simulation studies in which the lasso weights were applied to other variables. We conducted some additional simulations specifically for the Utah data set by creating noisy versions of canopy cover and applying the y-dependent weights to these new variables. Our results (not shown) are entirely consistent with McConville (2011): the design MSE of the estimators produced using lasso weights was smaller than the design MSE of the estimators with HT weights if the additional study variable was correlated with y, and the design MSE was roughly equal to the HT design MSE if the additional study variable was uncorrelated with y.

4.5 Domain Estimation

Other study variables that are of particular importance in applications are domain-restricted y values, $y_j I_{\{j \in A\}}$, where A is any subset of U. In the FIA context, A could be a county or other geographic domain, or could be defined by the values of other covariates, such as the set of all grid points j with aspect ranging from north-northwest to north-northeast.

Highly variable weights are increasingly important as domain sizes decrease and the presence or absence of individual elements becomes more critical. To

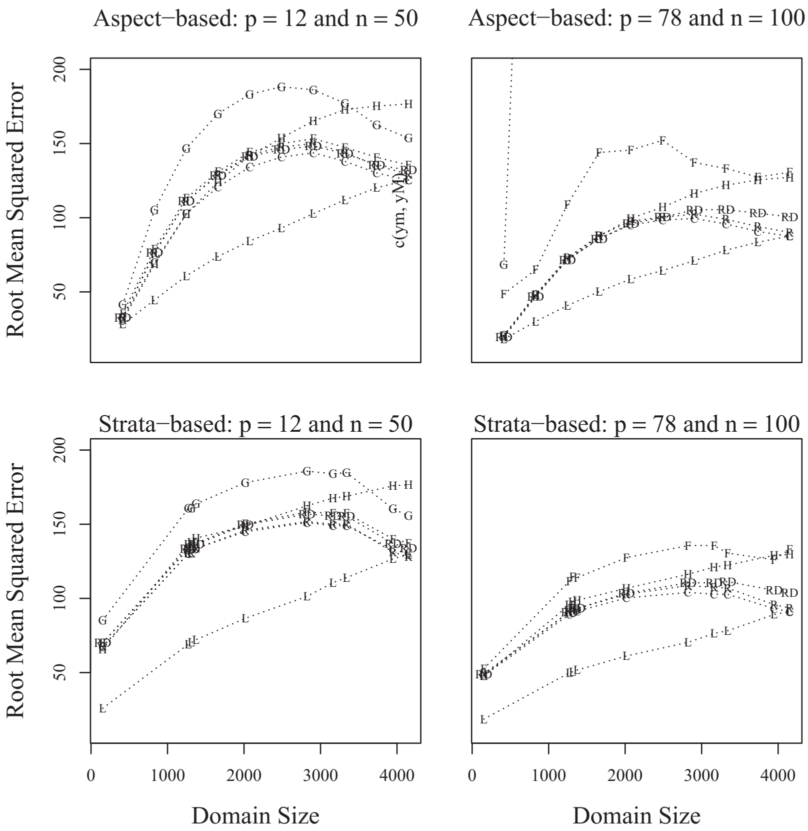

Figure 1. Root-Mean Squared Error of Estimators, Excluding the Adaptive Versions, for Aspect-Based Domains (Top) and Strata-Based Domains (Bottom), under Simple Random Sampling with $p = 12$, $n = 50$ (left), and $p = 78$, $n = 100$ (Right). Plotting Symbols are L = LASSO, C = CLASSO, R = RLASSO, RD = RIDGE, F = FSTEP, G = GREG, H = HT. GREG is not shown in the bottom right panel because its root mean squared error exceeds 200.

assess the impact of weight variation in the FIA example, we constructed two sets of ten nested "strata-based" domains of increasing size. The first set was obtained by setting domain 1 equal to stratum 1, domain 2 equal to stratum 1 combined with stratum 2, etc.: $U_1, U_1 \cup U_2, \ldots, U_1 \cup U_2 \cup \cdots \cup U_{10} = U$. The second set was formed by using the quantiles of transformed aspect to define 10 groups, then combining sequentially as above to form ten nested “aspect-based” domains.

For SI with $p = 12$ and $n = 50$, results are qualitatively very similar for estimation of the domain total on both sets of domains, as shown in the left column of figure 1. The adaptive versions of the lasso estimators are omitted from this figure for clarity, but in every case they track their nonadaptive versions closely. The GREG is the worst for estimation at all domain sizes except the largest two or three, with root MSE values larger even than HT. For the smallest domains, all of the estimators except GREG and nonlinear LASSO are very comparable, with GREG being worse and LASSO much better. For moderate to large domains, nonlinear LASSO maintains its dominance, with the remaining estimators ordered (from best to worst) as weighted lasso, RIDGE, FSTEP, HT, then GREG. The loss in efficiency from using weighted approximations to nonlinear LASSO is quite striking. Finally, for the whole population, the lasso estimators are all comparable, beating RIDGE narrowly and beating GREG by a good margin. HT is the worst for this largest domain.

For $p = 78$ and $n = 100$, results as shown in the right column of figure 1 are for the most part similar to those for the smaller model. The exceptions are that FSTEP behaves much worse than it does for the smaller model, and GREG's RMSE is so large that it cannot be displayed on the same scale. For $p = 78$ and $n = 200$, results (not shown) are qualitatively similar to those for $p = 12$ and $n = 50$.

5. SUMMARY

Based on the FIA simulation study, LASSO or ALASSO would be recommended in the situation (unusual in our experience) where survey regression estimates are needed, but survey weights are not needed. Of the two, the adaptive version ALASSO would seem to be preferable since it attempts to correct for some deficiencies of LASSO. In our simulations, however, ALASSO did not perform as well as LASSO, in part because of standardization of our covariates and in part due to poor performance (particularly in the large model) of the least squares coefficient estimators used in determining adaptive weights. For the more standard situation in which survey weights are desired, our calibrated CLASSO and ridged RLASSO performed very well and had similar computational demands. CLASSO was slightly better in many respects than RLASSO. Calibrated adaptive CALASSO and ridged adaptive RALASSO suffered from some of the same computational problems as ALASSO in our study, with RALASSO having some advantages over CALASSO. Our lasso-based methods tended to have less variable weights than GREG, ridged GREG (RIDGE), or GREG with forward stepwise selection (FSTEP). Less weight variation means much lower chance of negative weights and much better domain estimation properties. Both CLASSO and RLASSO yielded similar RMSE properties in estimation for domains, dominating GREG, RIDGE, and FSTEP. Like GREG, RIDGE suffers from the logistical limitation that it requires processing and maintaining all of the auxiliary data layers, while the lasso methods offer the possibility of dropping some covariates entirely. This is desirable in a production environment, potentially reducing costs and increasing processing speed.

This paper was motivated by survey regression estimation for a major natural resources survey, with large amounts of auxiliary data obtained from remote sensing. Practitioners naturally want to apply modern regression techniques like the LASSO in this context. Application of these methods in other survey contexts may be limited by the availability of auxiliary information. In some countries, population registries contain rich auxiliary information and the LASSO methods may have applicability. In the United States, government agencies are increasingly interested in improving survey data products through the use of "big data" available from various sources, including social media, internet data and sensor networks. Methods developed in this paper may have future application with such auxiliary information.

References

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, eds. B.N. Petrov and F. Caski, pp. 267–281, Budapest: Akademiai Kiado.

Bardsley, P., and R. L. Chambers (1984), "Multipurpose Estimation from Unbalanced Samples," *Applied Statistics*, 33, 290–299.

Beaumont, J. F., and C. Bocci (2008), "Another Look at Ridge Calibration," *Metron*, 66, 5–20.

Bickel, P. J., and D. A. Freedman (1984), "Asymptotic Normality and the Bootstrap in Stratified Sampling," *The Annals of Statistics*, 12, 470–482.

Breidt, F. J., G. Claeskens, and J. D. Opsomer (2005), "Model-Assisted Estimation for Complex Surveys Using Penalised Splines," *Biometrika*, 92, 831–846.

Breidt, F. J., and J. D. Opsomer (2000), "Local Polynomial Regression Estimators in Survey Sampling," *Annals of Statistics*, 28, 1026–1053.

——— (2009), "Nonparametric and Semiparametric Estimation in Complex Surveys," *Sample Surveys: Theory, Methods and Inference, Handbook of Statistics*, 29, 103–119.

Bunea, F., A. Tsybakov, and M. Wegkamp (2007), "Sparsity Oracle Inequalities for the Lasso," *Electronic Journal of Statistics*, 1, 169–194.

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, 615–620.

Chambers, R. L. (1996), "Robust Case-Weighting for Multipurpose Establishment Surveys," *Journal of Official Statistics*, 12, 3–32.

Cochran, W. G. (1977), *Sampling Techniques (3rd ed.)*, New York: John Wiley & Sons.

Coulston, J. W., G. G. Moisen, B. T. Wilson, M. V. Finco, W. B. Cohen, and C. K. Brewer (2012), "Modeling Percent Tree Canopy Cover: A Pilot Study," *Photogrammetric Engineering and Remote Sensing*, 78, 715–727.

Deville, J. C., and C. E. Särndal (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, 87, 376–382.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004), "Least Angle Regression," *Annals of Statistics*, 32, 407–499.

Friedman, J., T. Hastie, and R. Tibshirani (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22.

Fuller, W. A. (2009), *Sampling Statistics*, New Jersey: Wiley.

Goga, C. (2005), "Réduction de la Variance dans les Sondages en Présence D'information Auxiliarie: Une Approache Non Paramétrique Par Splines de Régression," *Canadian Journal of Statistics*, 33, 163–180.

Hájek, J. (1960), "Limiting Distributions in Simple Random Sampling from a Finite Population," *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 5, 361–374.

Homer, C., C. Huang, L. Yang, B. Wylie, and M. Coan (2004), "Development of a 2001 National Land-Cover Database for the United States," *Photogrammetric Engineering and Remote Sensing*, 70, 829–840.

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, 663–685.

Isaki, C. T., and W. A. Fuller (1982), "Survey Design under the Regression Superpopulation Model," *Journal of the American Statistical Association*, 77, 89–96.

Knight, K., and W. Fu (2000), "Asymptotics for Lasso-Type Estimators," *Annals of Statistics*, 28, 1356–1378.

Krewski, D., and J. Rao (1981), "Inference from Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods," *The Annals of Statistics*, 9, 1010–1019.

Lumley, T. (2009), Leaps: Regression Subset Selection. R package version 2.9. Available at http://CRAN.R-project.org/package=leaps.

McConville, K. S. (2011), "Department of Statistics Improved Estimation For Complex Surveys Using Modern Regression Techniques," unpublished Ph.D. thesis, Colorado State University.

McConville, K. S., and F. J. Breidt (2013), "Survey Design Asymptotics for the Model-Assisted Penalised Spline Regression Estimator," *Journal of Nonparametric Statistics*, 25, 745–763.

Montanari, G. E., and M. G. Ranalli (2005a), "Nonparametric Methods in Survey Sampling," *New Developments in Classification and Data Analysis*, 100, 203–210.

——— (2005b), "Nonparametric Model Calibration Estimation in Survey Sampling," *Journal of the American Statistical Association*, 100, 1429–1442.

Opsomer, J. D., F. J. Breidt, G. G. Moisen, and G. Kauermann (2007), "Model-Assisted Estimation of Forest Resources with Generalized Additive Models (with Discussion)," *Journal of the American Statistical Association*, 102, 400–416.

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at http://www.R-project.org/.

Rao, J. N. K., and A. C. Singh (1997), "A Ridge-Shrinkage Method for Range-Restricted Weight Calibration in Survey Sampling," *ASA Proceedings of the Section on Survey Research Methods*, pp. 57–65.

Raskutti, G., M. J. Wainwright, and B. Yu (2011). "Minimax Rates of Estimation for High-Dimensional Linear Regression Over-Balls," *Information Theory, IEEE Transactions On*, 57, 6976–6994.

Robinson, P. M., and C. E. Särndal (1983), "Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling," *Sankhyā: The Indian Journal of Statistics, Series B*, 45, 240–248.

Särndal, C. E. (2010), "The Calibration Approach in Survey Theory and Practice," *Survey Methodology*, 33, 99–119.

Särndal, C. E., B. Swensson, and J. Wretman (1989), "The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total," *Biometrika*, 76, 527–537.

——— (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

Silva, N., and C. J. Skinner (1997), "Variable Selection for Regression Estimation in Finite Populations," *Survey Methodology*, 23, 23–32.

Théberge, A. (2000), "Calibration and Restricted Weights," *Survey Methodology*, 26, 99–108.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Van De Geer, S. A., and P. Bühlmann (2009), “On the Conditions Used to Prove Oracle Results for the Lasso,” *Electronic Journal of Statistics*, 3, 1360–1392.

Wang, L., and S. Wang (2011), “Nonparametric Additive Model-Assisted Estimation for Survey Data,” *Journal of Multivariate Analysis*, 102, 1126–1140.

Wu, C., and R. R. Sitter (2001), “A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data,” *Journal of the American Statistical Association*, 96, 185–193.

Zou, H. (2006), “The Adaptive Lasso and its Oracle Properties,” *Journal of the American Statistical Association*, 101, 1418–1429.