Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Relative performance of methods based on model-assisted survey regression estimation: A simulation study

SURVEY
METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

by Erin R. Lundy and J.N.K. Rao

Release date: June 21, 2022

Statistics Canada
Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Relative performance of methods based on model-assisted survey regression estimation: A simulation study

Erin R. Lundy and J.N.K. Rao[1]

Abstract

Use of auxiliary data to improve the efficiency of estimators of totals and means through model-assisted survey regression estimation has received considerable attention in recent years. Generalized regression (GREG) estimators, based on a working linear regression model, are currently used in establishment surveys at Statistics Canada and several other statistical agencies. GREG estimators use common survey weights for all study variables and calibrate to known population totals of auxiliary variables. Increasingly, many auxiliary variables are available, some of which may be extraneous. This leads to unstable GREG weights when all the available auxiliary variables, including interactions among categorical variables, are used in the working linear regression model. On the other hand, new machine learning methods, such as regression trees and lasso, automatically select significant auxiliary variables and lead to stable nonnegative weights and possible efficiency gains over GREG. In this paper, a simulation study, based on a real business survey sample data set treated as the target population, is conducted to study the relative performance of GREG, regression trees and lasso in terms of efficiency of the estimators and properties of associated regression weights. Both probability sampling and non-probability sampling scenarios are studied.

Key Words: Model assisted inference, Calibration estimation; Model selection; Generalized regression estimator.

1. Introduction

At Statistics Canada and several other statistical agencies, there is a growing interest in leveraging auxiliary data, possibly from administrative sources, to improve the efficiency of estimators. Machine learning techniques have become a popular tool in various disciplines for utilizing such auxiliary information. These methods often do not require the distributional assumptions of more traditional methods and are able to adapt to complex non-linear and non-additive relationships between the outcomes and auxiliary variables. Machine learning methods have been applied to survey data in a variety of contexts such as response/adaptive designs, data processing, nonresponse adjustment and weighting (Buskirk, Kirchner, Eck and Signorino, 2018; Kern, Klausch and Kreuter, 2019).

Recently, the use of machine learning techniques to improve the efficiency of estimators of totals and means through model-assisted survey regression estimation under probability sampling has been considered. Model-assisted survey regression estimators of finite population totals may reduce variability and lead to significant gains in efficiency if the available auxiliary variables are strongly associated with the survey variable of interest. Increasingly, a large number of auxiliary variables are available, some of which may be extraneous. In this case, variable selection followed by regression estimation based on the selected model may improve efficiency of the survey regression estimators of finite population totals. We consider finite population estimation using the generalized regression (GREG) estimator with various linear working models (Särndal, Swensson and Wretman, 1992). Model-assisted estimators, using lasso

[1] Erin R. Lundy, Statistical Integration Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6. E-mail: erin.lundy@statcan.gc.ca; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6.

and adaptive lasso methods (McConville, Breidt, Lee and Moisen, 2017) and regression trees (McConville and Toth, 2019), have been applied to survey data. Other nonlinear models, such as penalized splines and neural networks, have been explored for model-assisted estimation; see Breidt and Opsomer (2017) for a survey of these techniques.

Another field of research where the use of model-assisted estimators has been proposed is estimation from non-probability samples. Increasing costs and declining response rates are leading to an expanding interest in the use of non-probability samples. However, the process generating a non-probability sample is unknown and such samples are subject to selection bias. Two commonly used approaches to estimation from non-probability samples are quasi-randomization and superpopulation modeling. In the first, the sample is treated as if it was obtained from probability sampling but with unknown selection probabilities. The pseudo-inclusion probabilities are estimated via a propensity model that uses the sample data in combination with some external data set that covers the targeted population. Machine learning techniques have been employed in the estimation of pseudo-inclusion probabilities or, equivalently, in the construction of pseudo-weights. Kern, Li and Wang (2020) investigated several machine learning techniques to construct pseudo-weights using a propensity score-based kernel weighting for non-probability samples. Rafei, Flannagan and Elliott (2020) developed a pseudo-weighting approach using Bayesian Additive Regression Trees.

In the superpopulation approach, observed values of the variables of interest are assumed to be generated by some model. The model is estimated from the data and, along with external population control data, is used to project the sample to the population. Under this framework, calibration to known population totals of auxiliary variables provides a means of potentially reducing the effect of sample selection bias. Chen, Valliant and Elliott (2018) discussed the implementation of model calibration using adaptive lasso for data based on non-probability sampling. In scenarios where the population totals are estimated, Chen, Valliant and Elliott (2019), incorporated the sampling uncertainty of the benchmarked data, obtained from a probability sample survey, into the variance component of a model-assisted calibration estimator using adaptive lasso regression. Therefore, unlike in the probability sampling context where the use of model-assisted estimation seeks to improve the efficiency of estimators, the use of these techniques in a non-probability sampling context aims to diminish the impact of selection bias.

We consider several lasso-based estimators as well as a regression tree estimator and evaluate their performance in both a probability sampling context and a non-probability sampling set up. In Section 2, the model-assisted estimators considered are discussed. The set up for a simulation study under probability sampling is described in Section 3. The results of the simulation study on the root mean square error of the estimators, relative bias of variance estimators and properties of survey weights are presented in Section 4. Except for the GREG estimator, all the model-assisted estimators considered here involve variable selection and yield, if applicable, regression weights that depend on the survey variable of interest, y. The impact of using a single set of regression weights for multiple related study variables is also investigated in this section. The results of the simulation study using a non-probability sampling scenario are detailed in Section 5. We conclude with a summary of the findings in Section 6.

2. Model-assisted estimation under probability sampling

2.1 GREG estimators

Consider the estimation of a finite population total $t_y = \sum_{i \in U} y_i,$ where $U = \{1, \ldots, N\}$ is the set of units of the finite population and y_i is the value of the survey variable of interest for the unit $i \in U.$ Let $s \subset U$ be a sample selected according to a sampling design $p(.),$ where $p(s)$ is the probability of selecting $s.$ For $i \in U,$ let $\pi_i = \Pr[i \in s]$ denote the first-order inclusion probabilities of the design. We assume $\pi_i > 0$ for all $i \in U.$ Additionally, assume d auxiliary variables, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$ are known for each $i \in U.$ A standard approach is to use the Horvitz-Thompson estimator

$$\hat{t}_{y,\mathrm{HT}} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} d_i y_i$$

where $d_i = \pi_i^{-1}$ denotes design weights. Under this strictly design-based framework, the auxiliary data do not impact the form of the estimator but can impact the design weights, $d_i,$ through the specification of the sampling design.

One strategy to use auxiliary data in estimation is to employ a model-assisted estimator of t_y by specifying a working model for the mean of y given $\mathbf{x}$ and use this model to predict y values. Specifying a linear regression working model leads to the generalized regression (GREG) estimator (Cassel, Särndal and Wretman, 1976). The GREG estimator typically has smaller variance than the Horvitz-Thompson estimator if the working model has some predictor power for $y.$ Here, we consider the GREG estimator under a linear regression working model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i \tag{2.1}$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$, ε_i independent and identically distributed with mean zero and variance σ^2 and $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^T$. The GREG estimator is given by

$$\hat{t}_{y,\mathrm{GREG}} = \sum_{i \in s} \frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_s}{\pi_i} + \sum_{i \in U} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_s \tag{2.2}$$

with the regression coefficients $\boldsymbol{\beta}$ estimated as

$$\hat{\boldsymbol{\beta}}_s = \underset{\beta}{\operatorname{argmin}} \left(\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta}\right)^T \boldsymbol{\Pi}_s^{-1} \left(\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta}\right) = \left(\mathbf{X}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{Y}_s, \tag{2.3}$$

where $\mathbf{X}_s$ is a $n \times (p+1)$ matrix, Y_s is a n-vector and Π_s is an $n \times n$ diagonal matrix of first-order inclusion probabilities for the sampled units.

The GREG estimator can also be written as a weighted sum of the variable of interest, $y,$ yielding regression weights that are independent of y and, therefore, can be applied to any study variable, y:

$$\hat{t}_{y,\mathrm{GREG}} = \sum_{i \in s} \left[1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x,\mathrm{HT}}\right)^T \left(\sum_{k \in s} \mathbf{x}_k \mathbf{x}_k^T d_k \right)^{-1} \mathbf{x}_i \right] d_i y_i = \sum_{i \in s} w_i y_i \tag{2.4}$$

where $\mathbf{t}_x$ is the known population total vector of the covariates $\mathbf{x}$ and $\hat{\mathbf{t}}_{x,\mathrm{HT}}$ is the Horvitz-Thompson estimator vector of the covariate population totals $\mathbf{t}_x = \sum_{i \in U} \mathbf{x}_i$. The regression weights, w_i, are termed calibration weights because they satisfy the calibration constraint $\sum_{i \in s} w_i \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i$. The calibration weight w_i does not depend on the study variable y_i. Note that the GREG estimator (2.4) can alternatively be expressed as

$$\hat{t}_{y,\mathrm{GREG}} = \hat{t}_{y,\mathrm{HT}} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x,\mathrm{HT}}\right)^T \hat{\boldsymbol{\beta}}_{\boldsymbol{s}}$$

which only requires known population totals $\mathbf{t}_x$. For the GREG estimator, the individual population values $\mathbf{x}_i, i \in U$ are not needed.

If a variable selection procedure, such as a forward stepwise procedure, is implemented prior to fitting the linear regression model, then the calibration weights will depend on y as the selected models may vary across study variables. This type of stepwise survey regression estimator is calibrated to the auxiliary variables selected by the variable selection procedure for a specific variable of interest, y.

Using a working linear regression model with many auxiliary variables, including interactions of categorical auxiliary variables, can produce substantially variable weights, and greatly increase the variance of the GREG estimator. Furthermore, some of the regression weights, $w_i, i \in s$, may be negative, thus losing the interpretation of a weight as the number of population units represented by the sampled unit.

2.2 Survey regression estimator with Lasso

If the linear regression model in (2.1) is sparse, i.e., p is large, and, say, only p_0 of the p regression coefficients are nonzero, then the estimation of the zero coefficients in (2.3) leads to extra variation in the GREG estimator (2.2). In this case, model selection to remove extraneous variables could reduce the overall design variance of the GREG estimator, leading to more efficient estimates of finite population totals. The least absolute shrinkage and selection operator (lasso) method, developed by Tibshirani (1996), simultaneously performs model selection and coefficient estimation by shrinking some regression coefficients to zero. The lasso approach estimates coefficients by minimizing the sum of squared residuals subject to a penalty constraint on the sum of the absolute value of the regression coefficients.

McConville et al. (2017) proposed using survey-weight lasso estimated regression coefficients given by

$$\hat{\boldsymbol{\beta}}_{s,L} = \underset{\beta}{\operatorname{argmin}} \left(\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta}\right)^T \boldsymbol{\Pi}_s^{-1} \left(\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{p} \left|\beta_j\right|,$$

where $\lambda \geq 0$. The lasso survey regression estimator for the total t_y is then given by

$$\hat{t}_{y,\,\mathrm{LASSO}} = \sum_{i \in s} \frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,L}}{\pi_i} + \sum_{i \in U} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,L}.$$

The value of the penalty parameter λ must be selected prior to obtaining the estimated coefficients. In general, this process of specifying hyperparameters prior to fitting the final model is called hyperparameter tuning. There are several potential selection criteria that can used to select the value of hyperparameters including Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) or cross-validation. We used a version of cross-validation which incorporates the design weights in our simulation study; see McConville (2011) for discussion of the selection of the penalty parameter for survey-weighted lasso coefficient estimates.

2.3 Survey regression estimator with adaptive Lasso

An issue with the use of the lasso criterion is that by shrinking the regression coefficients towards zero it yields biased estimates for regression coefficients that are far from zero. Under the adaptive lasso criterion (Zou, 2006), the coefficients in the l_1 penalty are weighted by the inverse of a root-n consistent estimator of $\boldsymbol{\beta}$. Therefore, the bias for large coefficients tends to be smaller.

McConville et al. (2017) considered an adaptive lasso survey regression estimator

$$\hat{t}_{y,\,\mathrm{ALASSO}} = \sum_{i \in s} \frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,\,\mathrm{AL}}}{\pi_i} + \sum_{i \in U} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,\,\mathrm{AL}},$$

where

$$\hat{\boldsymbol{\beta}}_{s,\,\mathrm{AL}} = \underset{\beta}{\operatorname{argmin}} \left(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta}\right)^T \boldsymbol{\Pi}_s^{-1} \left(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{p} \frac{\left|\beta_j\right|}{\left|\hat{\beta}_{sj}\right|}$$

and $\hat{\boldsymbol{\beta}}_s$ is given by (2.3). The reliance of the adaptive lasso method on the standard weighted linear regression coefficient estimates, $\hat{\boldsymbol{\beta}}_s$, leads to a loss of efficiency in settings when p is large because the estimates $\hat{\boldsymbol{\beta}}_s$ tend to be very unstable.

2.4 Lasso calibration estimators

The lasso and adaptive lasso methods do not produce regression weights directly, as the estimators cannot be expressed as weighted combinations of the y-values. McConville et al. (2017) developed lasso survey regression weights using a model calibration approach and a ridge regression approximation. These lasso regression weights depend on the variable of interest, y.

The lasso calibration estimator is calculated by regressing the variable of interest, y_i, on an intercept and the lasso-fitted mean function $\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,\,L}$. The lasso calibration estimator can be written in the same form as (2.4), where $\mathbf{x}_i$ is replaced by $\mathbf{x}_i^* = \left(1, \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,\,L}\right)^T$:

$$\hat{t}_{y,\,\mathrm{CLASSO}} = \sum_{i \in s} \left[1 + \left(\mathbf{t}_{x^*} - \hat{\mathbf{t}}_{x^*,\,\mathrm{HT}}\right)^T \left(\sum_{k \in s} \mathbf{x}_k^* \mathbf{x}_k^{*T} d_k\right)^{-1} \mathbf{x}_i^* \right] d_i y_i. \tag{2.5}$$

Similarly, the adaptive lasso calibration estimator is given by

$$\hat{t}_{y,\,\mathrm{CALASSO}} = \sum_{i \in s} \left[1 + \left(\mathbf{t}_{x^{**}} - \hat{\mathbf{t}}_{x^{**},\,\mathrm{HT}} \right)^T \left(\sum_{k \in s} \mathbf{x}_k^{**} \mathbf{x}_k^{**T} d_k \right)^{-1} \mathbf{x}_i^{**} \right] d_i y_i,$$

where the lasso-fitted mean for $\mathbf{x}_i^*$ in (2.5) is replaced by the adaptive lasso fit, $\mathbf{x}_i^{**} = \left(1, \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,\,\mathrm{AL}}\right)^T$. The weights for the lasso calibration estimators are calibrated to the population size N and to the population total of the lasso-fitted mean functions.

2.5 Regression tree estimator

The GREG estimator can also be expressed as

$$\hat{t}_{y,r} = \sum_{i \in s} \frac{y_i - \hat{h}_n(\mathbf{x}_i)}{\pi_i} + \sum_{i \in U} \hat{h}_n(\mathbf{x}_i), \tag{2.6}$$

where $\hat{h}_n(\mathbf{x}_i)$ is an estimator of the mean function of Y_i given $\mathbf{X}_i = \mathbf{x}_i$, $h(\mathbf{x}_i) = E(Y_i \mid \mathbf{X}_i = \mathbf{x}_i)$, based on the sample data $(y_i, \mathbf{x}_i)$, $i \in s$. As an alternative to a linear regression model, McConville and Toth (2019) proposed estimating $h(\mathbf{x})$ with a regression tree model using the following algorithm:

1. Let $k(n)$ be the minimum box size and α be a specified significance level.
2. If the dataset contains at least $2k(n)$ observations then continue to step 3; otherwise, stop.
3. Among the auxiliary variables x_l, $l = 1, \ldots, d$, choose a variable to split the data. The chosen x_l is the variable that shows the largest significance difference after testing the null-hypothesis of homogeneous $E[y \mid x_l]$. If no variable leads to a significant difference, then stop.
4. Split the data into two sets S_L and S_R by splitting based on the value of the selected variable x_l that results in the largest decrease in the estimated mean square error, while satisfying the requirement that each subset contains at least $k(n)$ units.
5. For each of the resulting subsets of the data, return to step 1.

The resulting regression tree model groups the categories of an auxiliary variable based on their relationship to the variable of interest and only includes auxiliary variables and interactions associated with this variable. Importantly, including a categorical variable does not require a split for each category, potentially reducing the model size substantially while still capturing important interactions.

After fitting a regression tree model, we obtain a set of boxes $Q_n = \{B_{n1}, B_{n2}, \ldots, B_{nq}\}$ which partition the data. Let $I(\mathbf{x}_i \in B_{nk}) = 1$ if $\mathbf{x}_i \in B_{nk}$ and 0 otherwise, for $k = 1, \ldots, q$. This means that $I(\mathbf{x}_i \in B_{nk}) = 1$ for exactly one box $B_{nk} \in Q_n$ for every $i \in s$. For every $\mathbf{x}_i \in B_{nk}$, the estimator of $h(\mathbf{x}_i)$ is given by

$$\tilde{h}_n(\mathbf{x}_i) = \tilde{\#}(B_{nk})^{-1} \sum_{i \in s} \pi_i^{-1} y_i I(\mathbf{x}_i \in B_{nk}) = \tilde{\mu}_{nk}, \tag{2.7}$$

where

$$\tilde{\#}\left(B_{nk}\right)=\sum_{i\in s}\pi_i^{-1}I\left(\mathbf{x}_i\in B_{nk}\right)$$

is the HT estimator of the population size in box B_{nk}. The regression tree estimator $\hat{t}_{y,\,\text{TREE}}$ is obtained by inserting equation (2.7) into the generalized regression estimator, given in equation (2.6), leading to the post stratified estimator

$$\hat{t}_{y,\,\text{TREE}}=\sum_k N_k\tilde{\mu}_{nk},$$

where N_k is the number of units in U that belong to box k.

Since $\tilde{h}_n\left(\mathbf{x}_i\right)$ can be written as a linear regression estimator with q indicator function covariates, the regression tree estimator is also a post-stratified estimator, where each box B_{nk} represents a post-stratum. This implies that this estimator is calibrated to the population total of each box, providing a data-driven mechanism, dependent on y, for selecting post-strata that ensures that none of them are empty. As a result, the regression weights are guaranteed to be non-negative. The weights produced by this estimation procedure depend on the variable of interest, y. Therefore, unlike the GREG approach, a single set of generic weights to apply to all study variables is not available. Instead, a set of weights for each survey variable of interest is produced.

2.6 Variance estimation under stratified simple random sampling

Under stratified simple random sampling, a variance estimator of the model-assisted survey regression estimators described above is obtained by the Taylor linearization method and given by

$$\hat{V}\left(\hat{t}_y\right)=\sum_h\frac{N_h\left(N_h-n_h\right)}{n_h}\frac{1}{n_h-1}\sum_{i\in s_h}\left(e_{hi}-\bar{e}_h\right)^2, \tag{2.8}$$

where h indexes the strata, N_h is the number of population units in stratum h, n_h is the number of sampled units s_h in stratum h, $e_{hi}=y_{hi}-\hat{h}_n\left(\mathbf{x}_{hi}\right)$ is the residual of sample unit i in stratum h under the regression model and $\bar{e}_h$ is the average residual in stratum h.

The variance estimators readily extend to more complex sampling designs, but for simplicity we have given the expression only for stratified simple random sampling which is used in the simulation study of Section 3.

3. Simulation study using Financing and Growth of Small and Medium Enterprises Survey data

In this section, we describe a simulation study used to compare the performance of model-assisted survey regression estimators relative to the purely design-based HT estimator. Using the Survey of Financing and Growth of Small and Medium Enterprises data as the population, we compare the estimators in repeated samples of the data to produce estimates of the total amount requested for trade credit which is a particular type of financing.

3.1 Simulation population

The Survey of Financing and Growth of Small and Medium Enterprises (SFGSME) is a periodic survey of enterprises which occurs approximately every three years and collects information on the types of financing businesses use. The sample is stratified by size, defined by the number of employees, the age of the business, industry at the 2-digit North American Industry Classification System (NAICS) and geography. A sample of approximately 17,000 enterprises was selected for the 2017 iteration of the survey.

The Business Register (BR) is the primary source of auxiliary information for business surveys at Statistics Canada. The frame used by the SFGSME was constructed by selecting from Statistic's Canada BR all enterprises with between 1 and 499 employees and a minimum gross revenue of $30,000. Non-profit enterprises as well as enterprises belonging to certain industry subgroups were excluded from the target population. The BR contains information on the location, number of employees, industry as well as revenue for each enterprise in the population.

3.2 Simulation methodology

We conducted a simulation study to compare the relative performance of several model-assisted survey regression estimators, using three and four categorical auxiliary variables. We considered sample sizes of $n = \{200; 500; 1{,}000\}$ from the 9,115 respondents in the SFGSME dataset. This dataset was treated as the target population and repeated samples were drawn using stratified simple random sampling as this is the design commonly used by statistical agencies for business surveys. We assumed there are two strata, where stratum A consists of units with revenue of less than $2.5 million and stratum B consists of units with revenue greater than $2.5 million. We assumed equal sample sizes in each stratum but most of the units in the population, approximately 70%, belong to stratum A. Under this sampling design, larger revenue units are over-represented, resulting in an unequal probability sampling design. Preliminary simulations using a simple random sample design were also considered and yielded similar results. The minimum sample size considered was $n = 200$ because for smaller sample sizes and 28 categories of x-variables, there were often categories without a sampled unit. In this case, it is not possible to calibrate the GREG estimator to all the pre-specified marginal totals.

For each sample, models using three x-variables, industry (10 categories), employment size (4 categories) and region (6 categories) were used to estimate total amount of trade credit requested and results were compared to the true total. We also considered a fourth variable, revenue, with 8 categories. For each combination of the three different sample sizes, and the two sets of auxiliary variables, with 20 and 28 main effects categories, we drew 5,000 repeated stratified random samples from the target population. For each sample, we implemented the HT estimator and several model-assisted survey estimators as summarized in Table 3.1 below:

Table 3.1
Summary of model assisted estimators considered in simulation study

Estimator	**Auxiliary Data**	**Regression Weights**	**Calibration Totals**
GREG	Marginal totals Considered main effects only	Independent of y	All auxiliary variables
GREG with forward variable selection (FSTEP)	Individual values Considered main effects only	Dependent on y	Selected auxiliary variables
Regression Tree (TREE)	Individual values	Dependent on $y,$ strictly positive	Population size of each box
Lasso (LASSO)	Individual values Considered main effects (1-way) and two-way interactions (2-way)		
Calibrated lasso (CLASSO)	Individual values Considered main effects (1-way) and two-way interactions (2-way)	Dependent on y	Population size and lasso-fitted mean function
Adaptive lasso (ALASSO)	Individual values Considered main effects only		
Calibrated adaptive lasso (CALASSO)	Individual values Considered main effects only	Dependent on y	Population size and lasso-fitted mean function

We initially also considered adaptive lasso and adaptive lasso calibration estimators using all main effects and 2-way interactions, but estimates of the coefficients under the GREG linear model, $\hat{\boldsymbol{\beta}}_s,$ were highly unstable leading to singularity issues.

All computations were completed in R (Version 3.4.0, 2017). The HT, GREG, regression tree and lasso estimators were calculated using the package **mase** (McConville, Tang, Zhu, Li, Cheung and Toth, 2018) and the adaptive lasso coefficients were computed using the package **glmnet** (Friedman, Hastie, Simon, Qian and Tibshirani, 2017). The function `cv.glmnet` was used to select the value of the penalty parameter for the lasso estimators. We used a 10-fold cross validation procedure which allows for the inclusion of design weights. For the regression tree estimator, the minimum box size $k(n)$ was specified as 25 and the level of significance α was 0.05. We also considered a minimum box size of 10 units. For small sample sizes, there was a small gain in efficiency relative to a minimum box size of 25. For sample sizes of $n = 1{,}000,$ different choices for the minimum box size yielded similar results in term of mean square error. Forward stepwise selection for the FSTEP estimator was based on minimizing the Akaike Information Criteria (AIC) and was performed using the function `stepAIC` in the MASS package (Ripley, Venables, Bates, Hornik, Gebhardt and Firth, 2017).

In regressing the amount of trade credit requested for the entire finite population on the 28 marginal categories, the adjusted coefficient of determination was approximately $R^2 = 0.22$ when both main effects and two-way interaction effects were considered. For the population model with main effects only the number of significant effects was 15 and for the population model with main effects and two-way interactions, there were 2 significant main effects and 29 significant interaction effects. These population-level results indicate that useful predictive models should be sparse and that there may be important two-way interactions.

Fitting regression tree models to the amount of trade credit requested resulted in 25 splits. The first split was based on revenue, indicating that this is the auxiliary data that is most strongly related to the amount of trade credit requested. There were splits based on all four of the auxiliary variables considered: revenue, industry, employment size and geography. This is consistent with the conclusions that useful predictive models should be sparse but allow for higher order interactions.

4. Results of the simulation study

4.1 Performance of estimators in terms of design MSE

We computed design bias and design mean square error (MSE) from the 5,000 total estimates by sample size and number of marginal categories. The percentage absolute relative design bias was less than 2 percent for all the estimators for all scenarios. As expected, for all estimators, the bias decreases as the sample size increases.

Figure 4.1 displays the MSE of the HT, GREG, GREG with forward variable selection, regression tree and calibrated lasso estimators by sample size, based on the 5,000 simulated samples. The MSE values are similar for the adaptive and non-calibrated versions of the lasso estimators. For all the estimators, the decrease in MSE is much more pronounced from $n = 200$ to $n = 500$ than from $n = 500$ to $n = 1{,}000.$ This is likely due to the small sample size, relative to the number of categories for the auxiliary variables. It may not be possible to explore all the potential effects, particularly higher order effects, with only 200 sampled units.

Table 4.1 displays the ratio the design MSE of each estimator to the MSE of the HT estimator for the total amount of trade credit requested. For $n = 200,$ the regression tree estimator and the lasso (2-way) estimator with two factor interaction effects are the only model-assisted estimators that provide any efficiency gains, relative to the HT estimator, when the number of categories of auxiliary variables used is large. As the sample size increases, the gains in efficiency of the model-assisted survey regression estimators, relative to the HT estimator, are essentially equal. Using any of the model-assisted estimators when $n = 1{,}000$ results in a slight gain in efficiency, relative the HT estimator. There is little efficiency advantage for model-assisted estimators over the HT estimator, indicating that the auxiliary variables are not strongly related to the variable of interest.

Figure 4.1 Comparison of mean square error for HT, GREG, FSTEP, regression tree and calibrated lasso estimators (1-way and 2-way) for the total amount of trade credit requested.

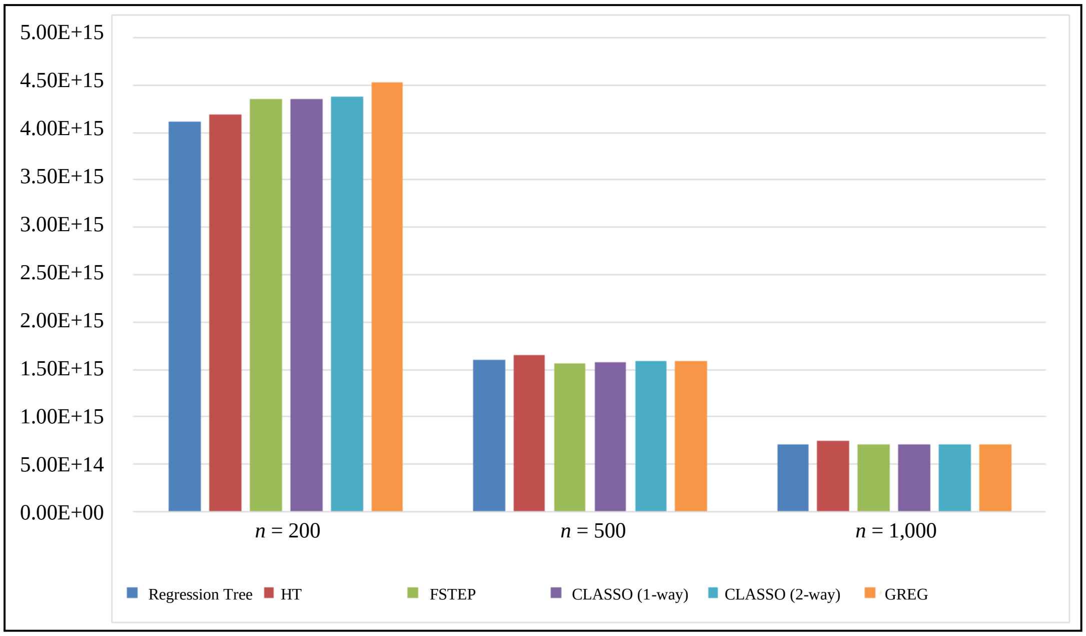

Table 4.1
Ratio of MSE of each estimator to MSE of HT estimator with 20 and 28 marginal categories

	20 categories			28 categories		
	$n=200$	$n=500$	$n=1{,}000$	$n=200$	$n=500$	$n=1{,}000$
GREG	1.067	1.011	0.994	1.084	0.959	0.954
FSTEP	1.036	1.009	0.994	1.040	0.945	0.958
TREE	1.023	1.007	0.977	0.983	0.963	0.949
LASSO (1-way)	1.020	0.995	0.986	1.009	0.946	0.947
CLASSO (1-way)	1.047	1.004	0.990	1.042	0.952	0.949
LASSO (2-way)	0.999	0.995	0.952	0.981	0.935	0.936
CLASSO (2-way)	1.061	1.029	0.966	1.045	0.959	0.950
ALASSO	1.024	0.999	0.986	1.021	0.948	0.948
CALASSO	1.040	1.005	0.989	1.037	0.951	0.949

The potential gains in efficiency for model-assisted estimators depend on the predictive power of the working model. In our simulation population, the strength of the relationship between the variable of interest and the available auxiliary variables is weak, leading to only slight efficiency gains relative to the purely design-based HT estimator. Therefore, to further explore the differences between the various model-assisted survey estimators, we ran additional simulations using different survey variables of interest, generated according to the following procedure:

1. Assuming a lasso model with main effects only, we obtained the lasso coefficient estimates for the amount of trade credit requested, y_i, using the population values for the auxiliary variables $\mathbf{x}_i$, including revenue.

2. We used the coefficient estimates $\hat{\beta}_L$ obtained in step 1 and the population values for $\mathbf{x}_i$ to generate a new survey variable of interest

$$y_i^* = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_L + \mathbf{u}_i,$$

where $\mathbf{u}_i$ is a normally distributed random variable with mean 0 and standard deviation σ chosen such that the adjusted coefficient of determination is approximately $R^2 = 0.5.$
3. We drew 5,000 repeated samples from the target population and calculated the mean square error of each estimator of the total t_{y^*}.
4. Steps 1-3 were repeated by fitting a lasso regression model with main effects and 2-way interactions and a regression tree model using the algorithm detailed in Section 2.5.

Table 4.2 displays the ratio the design MSE of each estimator to that of the HT under the three different models generating the survey variable of interest for a sample size of $n = 1{,}000.$ As expected, the estimator based on the correctly specified working model is the most efficient. In the case where the true generating model contains only main effects, assuming a working model with higher order interactions results in a slight loss in efficiency. If two-way or higher order interactions are present, the regression tree and lasso-based estimators fitted with two-way interactions are more efficient than the model-assisted estimators based on working models with only main effects. When the generating model is a regression tree, the regression tree estimator yields modest efficiency gains over the 2-way lasso-based estimators. This can be explained by the fact that the regression tree model groups the categories of an auxiliary variable based on their relationship to the variable of interest and, therefore, reduces the model size. In all cases, significant efficiency gains, relative to the design-based HT estimator, are achieved.

Table 4.2
Ratio of MSE for each estimator to MSE of HT under different models generating survey variable of interest

	LASSO (1-way)	LASSO (2-way)	Regression Tree
GREG	0.749	0.855	0.878
FSTEP	0.749	0.855	0.876
TREE	0.803	0.821	0.778
LASSO (1-way)	0.747	0.850	0.871
CLASSO (1-way)	0.747	0.851	0.873
LASSO (2-way)	0.763	0.761	0.826
CLASSO (2-way)	0.763	0.765	0.833
ALASSO	0.750	0.849	0.872
CALASSO	0.750	0.851	0.873

4.2 Performance under other scenarios

We also examined the performance of the lasso-based and regression tree estimators under scenarios where there are no main effects, only 2-way interactions. We generated a fourth survey variable of interest using the lasso regression model with main effects and 2-way interactions as described in the procedure above. However, in step 2, we set all coefficients estimates corresponding to main effects equal to 0.

The first column of Table 4.3 (called no multicollinearity) shows the ratio the design MSE of the estimators to that of the HT estimator, where the survey variable is generated from a model with no main effects for sample sizes of $n = 1{,}000$. Under this scenario, the lasso estimators with 2-way interactions and the regression tree estimator are significantly more efficient than model-assisted estimators based on main effects only models. Relative to the commonly used GREG estimator, the efficiency gains for the lasso estimators with 2-way interactions and the regression tree estimator are significantly greater when there are no main effects. This is evident by comparing LASSO 2-way column in Table 4.2 to first column in Table 4.3. The relative MSE is very similar for the 2-way lasso and regression tree estimators but closer to 1 for GREG and 1-way lasso estimators.

Table 4.3
Ratio of MSE for each estimator to MSE of HT under generating model with no main effects and in the absence/presence of multicollinearity

	No Multicollinearity	Duplicated Variable	Collapsed Categories
GREG	0.935	-	-
TREE	0.824	0.850	0.842
LASSO (1-way)	0.930	0.945	0.942
CLASSO (1-way)	0.936	0.953	0.951
LASSO (2-way)	0.783	0.795	0.773
CLASSO (2-way)	0.795	0.809	0.781

For administrative data with many variables, it is not uncommon for some variables to be colinear or nearly colinear. For example, information on both the total number of employees and the number of full-time equivalent employees is often available. The GREG estimator, and by extension the FSTEP estimator and adaptive lasso estimators, fail in the presence of collinearity as the design matrix is singular. We investigated the performance for regression tree and lasso estimators in the presence of multicollinearity. We considered two types of multicollinearity:

- Duplicate of existing categorical variable. We created three new indicator variables corresponding to employment size.
- Collapsed categories of existing auxiliary variable: We created a new indicator variable corresponding to the three highest categories of revenue.

The MSE, relative to the HT estimator, for $n = 1{,}000$ is shown in columns 2 and 3 of Table 4.3. These results are very similar to those in the first column of Table 4.3 without the presence of multicollinearity. The regression tree and lasso estimators provide an automatic way of removing colinear auxiliary variables without impacting the potential efficiency gains. It should be noted that other methods, such as principal component analysis, can be used to eliminate collinearity but require some expertise.

4.3 Performance of variance estimators in terms of relative bias

Variance estimators based on (2.8) were constructed for each estimator. Table 4.4 displays the percentage relative bias of each estimator for the total amount of trade credit requested. For comparison purposes, the theoretically unbiased variance estimator of the HT estimator is included in this table. This variance estimator is equivalent to the expression provided in (2.8) where $e_i = y_i - \bar{y}_s$. The variance estimators for the model-assisted survey regression estimators have substantial negative bias which increases as the number of auxiliary variables, p, increases. The magnitude of negative bias is largest for the lasso-based estimators fitted using 2-way interactions. For small sample sizes, the negative bias is smallest for the regression tree estimator. As well, for small sample sizes, there is a substantial difference in bias between the GREG and FSTEP estimators. Performing variable selection prior to calculating the standard GREG calibration estimator appears to reduce the bias of the variance estimator in this case. The bias reduces for all model-assisted survey regression estimators as the sample size increases.

Table 4.4
Percent relative bias of variance estimators

	20 categories			28 categories		
	$n = 200$	$n = 500$	$n = 1{,}000$	$n = 200$	$n = 500$	$n = 1{,}000$
GREG	-12.44	-4.16	-1.60	-22.23	-10.86	-6.99
FSTEP	-7.05	-3.60	-1.62	-14.07	-7.71	-6.73
TREE	-5.79	-5.53	-2.81	-8.45	-12.93	-10.83
LASSO (1-way)	-7.79	-2.96	-1.14	-12.42	-9.49	-6.44
CLASSO (1-way)	-10.08	-3.74	-1.61	-16.01	-9.84	-6.52
LASSO (2-way)	-11.94	-11.57	-7.62	-16.12	-15.14	-13.08
CLASSO (2-way)	-19.99	-15.09	-9.06	-25.87	-19.04	-15.14
ALASSO	-8.69	-3.61	-1.41	-14.52	-9.43	-6.38
CALASSO	-9.40	-3.78	-1.48	-15.80	-9.64	-6.46
HT	5.19	5.72	5.82	4.90	-0.11	1.66

Given the bias of the variance estimators seen here, particularly for small sample sizes, a possible concern is the quality of the first-order Taylor expansion approximation. For a large number of categorical auxiliary variables, the remainder term in the Taylor expansion may no longer be negligible for small sample sizes. An alternative variance estimator for the lasso estimators was considered by McConville et al. (2017) but yielded only slight improvements in terms of bias reduction. An additional concern is properly accounting for the inherently data driven procedure used to estimate the regression tree and lasso models. The regression tree model has splits while the lasso models have a penalty parameter both depending on the sample.

4.4 Properties of the survey weights

Regression weights are directly available for the GREG, FSTEP, regression tree, lasso calibration (1-way and 2-way) and adaptive lasso calibration estimators. We investigated the properties of the weights for these estimators in our simulations.

Large variation in the values of weights is undesirable as they allow some units to be much more influential than others. Positive weights are preferred by national statistical organizations as a negative weight no longer holds the interpretation of the number of population units represented by the sampled unit.

First, we computed the average, over repeated samples, of the empirical within-sample variance of the weights:

$$\overline{\text{var}}\left(\mathbf{w}\right) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n-1}\sum_{j\in s^{(r)}}\left(w_j^{(r)} - \bar{w}^{(r)}\right)^2,$$

where $s^{(r)}$ is the r^{th} simulated sample, $\bar{w}^{(r)} = \frac{1}{n}\sum_{j\in s^{(r)}} w_j^{(r)}$ and $w_j^{(r)}$ is the weight of the j^{th} unit in the r^{th} simulated sample.We also computed the average coefficient of variation (CV) of the weights:

$$\overline{\text{CV}}\left(\mathbf{w}\right) = \frac{1}{R}\sum_{r=1}^{R}\frac{\sqrt{\frac{1}{n-1}\sum_{j\in s^{(r)}}\left(w_j^{(r)} - \bar{w}^{(r)}\right)^2}}{\bar{w}^{(r)}}$$

Table 4.5 displays the average variance and average CV for the weights across samples when revenue was included as an auxiliary variable. The weights for the GREG estimator and, to a lesser extent the FSTEP estimator, are much more variable than the weights for the regression tree and lasso-based estimators, particularly for small sample sizes. The variability of the weights for the three lasso-based approaches is very similar and is always slightly lower than the variability of the weights for the regression tree estimator.

Table 4.5
Average variance (CV) for weights across samples

	***n* = 200**	***n* = 500**	***n* = 1,000**
GREG	728.18 (0.59)	77.14 (0.48)	16.41 (0.44)
FSTEP	462.81 (0.47)	67.45 (0.45)	15.90 (0.44)
TREE	374.43 (0.42)	59.35 (0.42)	14.70 (0.42)
CLASSO (1-way)	354.57 (0.41)	56.21 (0.41)	14.03 (0.41)
CLASSO (2-way)	361.83 (0.42)	56.60 (0.41)	14.06 (0.41)
CALASSO	354.29 (0.41)	56.28 (0.41)	14.03 (0.41)

We also computed the proportion of simulated samples where the regression weights contained negative values. As mentioned in Section 2.5, by construction, the weights for the regression tree estimator are guaranteed to be strictly positive. When the sample size was 200, the GREG estimator calibrated to 20 marginal categories yielded negative weights for approximately 3% of the repeated samples. There were no negative weights when the sample size was 500 or 1,000. For the GREG estimator calibrated to 28 marginal categories, approximately 27% of the repeated samples of size 200 contained negative weights and less than 0.5% of the repeated samples of size 500 contained negative weights. The GREG weights are unstable when the sample size is small, especially if the GREG estimator is calibrated to auxiliary variables with many categories. Using forward stepwise variable selection with the GREG estimator resulted in a substantial decrease in the number of simulated samples with negative weights for small sample sizes. The FSTEP estimator applied to the 28 marginal categories yielded negative weights in approximately 0.5% of the repeated samples of size 200. There were no negative weights observed for the lasso calibration estimator with only main effects or adaptive lasso calibration estimator. Using the lasso calibration estimator with 2-way interactions resulted in negative weights in less than 0.05% of the simulated samples.

4.5 Estimation based on a single set of weights

A major drawback in the implementation of the regression tree and the calibrated lasso-based approaches is that the estimation procedures yield variable-specific weights. We conducted additional simulations in which a single set of variable-specific weights was applied to other related survey variables of interest. In the context our business survey data, we considered four survey variables of interest, the amount of trade credit requested as well as the amount requested for three additional types of financing: line of credit, business credit card and leasing financing. We examined the impact on bias and loss of efficiency in using a single set of weights, determined by a primary variable of interest, to estimate the total amount requested for the remaining three survey variables of interest. Specifically, we calculated the percentage absolute relative design bias for the estimators of the total amount requested and the variance estimators. We also calculated the ratio of the MSE for the regression tree and three calibrated lasso-based approaches using the set of weights corresponding to a primary variable of interest to the MSE for the estimators using variable-specific weights. For brevity, we considered only settings with 28 marginal categories.

The percentage absolute relative design bias was less than 2 percent for all of the estimators for all scenarios. For all estimators and primary variable of interest, the bias decreases as the sample size increases.

Unlike the bias of the variance estimators based on variable-specific weights, the bias of the variance estimators based on a single set of weights for a primary variable of interest does not necessarily decrease as the sample size increases. As well, the bias is not strictly in one direction and may be positive or negative. For the regression tree and calibrated lasso-based approaches, the bias of the variance estimators is substantially larger for the primary variable of interest used to calculate the single set of weights than for the other study variables. The data driven nature of these estimators means that the estimated variance for the primary variable of interest is underestimated, as shown in Table 4.4.

Table 4.6 displays the ratio of the design MSE of each estimator with weights determined by a primary variable of interest to that of the estimator with variable-specific weights, calculated separately for each of the four study variables for n equal to 200 and 500. Using a single set of weights determined by a primary variable of interest results in a similar or slightly higher MSE than using variable-specific weights. Here, the loss in efficiency is modest, less than 8% in all settings considered. Similar results were obtained for the case $n = 1{,}000$. There is no clear pattern in terms of loss of efficiency and sample size.

Table 4.6
Ratio of MSE for each estimator with weights determined by primary variable of interest to MSE for estimator with variable-specific weights

		Trade Credit		Line of Credit		Business Credit Card		Lease Financing	
	n	**200**	**500**	**200**	**500**	**200**	**500**	**200**	**500**
Primary variable: Trade Credit	TREE	-	-	1.01	0.97	0.99	1.00	0.99	1.00
	CLASSO (1-way)	-	-	0.99	0.99	1.01	0.99	1.00	1.01
	CLASSO (2-way)	-	-	0.93	0.94	0.92	0.98	0.92	0.97
	CALASSO	-	-	0.97	0.99	1.01	0.99	0.96	1.00
Primary variable: Line of Credit	TREE	1.06	0.97	-	-	0.98	1.00	0.98	0.97
	CLASSO (1-way)	0.96	0.98	-	-	0.99	1.01	0.99	0.99
	CLASSO (2-way)	0.95	0.96	-	-	0.92	0.98	0.93	0.96
	CALASSO	0.97	0.98	-	-	0.99	1.00	0.96	0.98
Primary variable: Business Credit Card	TREE	1.06	1.01	1.06	0.97	-	-	0.99	1.02
	CLASSO (1-way)	0.99	1.02	0.98	0.97	-	-	0.99	1.02
	CLASSO (2-way)	0.98	1.00	0.95	0.93	-	-	0.92	0.99
	CALASSO	1.00	1.02	0.97	0.97	-	-	1.00	1.01
Primary variable: Lease Financing	TREE	1.07	1.03	1.06	1.05	0.99	1.02	-	-
	CLASSO (1-way)	0.99	1.05	0.98	1.04	0.99	1.02	-	-
	CLASSO (2-way)	0.97	1.02	0.96	1.01	0.92	0.99	-	-
	CALASSO	1.00	1.05	0.98	1.05	1.00	1.01	-	-

5. Estimation under non-probability sampling

In this section, we study the effect of selection bias on the survey regression estimators under non-probability sampling. For this purpose, we studied two types of selection bias possibly present in non-probability samples. In particular, we considered a scenario in which the probability of selection depends only on the auxiliary data available for all units in the population, and a scenario in which the probability of selection depends on the survey variable of interest. In both scenarios, we evaluated the absolute relative bias (ARB), $\left|\hat{t}_y - t_y\right| \big/ t_y$, for each estimator of the total. Following Chen, Valliant and Elliott (2018), we treat the non-probability sample as a simple random sample and set the design weights equal to $d_i = N/n$ for the estimation of total t_y as the selection process for non-probability samples is unknown in practice.

5.1 Selection probabilities depend on auxiliary data

We drew repeated samples using the same stratified SRS design as in Section 4. Table 5.1 displays the ARB of each estimator of the total amount of trade credit requested assuming $d_i = N/n$, when the sample is in fact selected using disproportionate stratified random sampling.

As expected, the wholly designed-based HT estimator has the largest bias, and this bias does not decrease as the sample size increases. The ARB of model-assisted estimators decreases as the sample size

n increases. The GREG estimator has the smallest bias, particularly for small sample sizes. Furthermore, the GREG estimator is approximately unbiased if revenue is included as one of the auxiliary variables for calibration. However, if stepwise variable selection is used, the GREG estimator is no longer unbiased for small sample sizes. On the other hand, if revenue is not included as a calibration variable, the GREG estimator is slightly biased. The lasso-based and, to a smaller extent, the regression tree estimators suffer from small sample bias for $n = 200$ when revenue is correctly included as an auxiliary variable. This is most apparent for the standard lasso estimators that do not include calibration to known population totals. For n equal to 500 or 1,000, including revenue as an auxiliary variable, substantially decreases the bias for the regression tree and calibrated lasso estimators but only slightly decreases the bias for the lasso estimators without calibration. This indicates that the additional calibration step is important for diminishing the effect of selection bias, especially if the sample size is small.

Table 5.1
Percent ARB of each estimator under stratified sampling with revenue and without revenue included as an auxiliary variable

	Revenue			**Without Revenue**		
	$n = 200$	$n = 500$	$n = 1{,}000$	$n = 200$	$n = 500$	$n = 1{,}000$
GREG	0.31	0.06	0.06	4.84	5.12	4.71
FSTEP	2.67	0.44	0.06	9.20	5.18	4.92
TREE	4.15	1.04	0.50	17.40	10.20	8.94
LASSO (1-way)	17.42	5.10	2.32	16.32	8.88	6.49
CLASSO (1-way)	7.99	0.83	0.20	9.04	5.22	4.59
LASSO (2-way)	25.36	14.28	8.40	26.31	15.16	9.89
CLASSO (2-way)	10.72	1.44	1.02	14.19	5.56	3.84
ALASSO	14.95	5.63	3.00	14.35	8.64	6.51
CALASSO	9.63	2.54	1.25	9.27	5.77	4.92
HT	49.45	48.84	48.81	49.08	49.29	48.60

These results indicate that when the selection probability depends on a known auxiliary variable, including it in the working model for the GREG estimator effectively diminishes the effect of selection bias. This was not the case for the model-assisted estimators that involved variable selection. Performing variable selection may increase bias as auxiliary variables that are predictive in terms of selection probability may not be selected and properly accounted for. The lasso estimators can be constructed such that user-specified variables are always included in the working regression model. These user-specified variables can be added to x_i^* in equation (2.5) to force calibration to corresponding population totals. Unfortunately, the underlying selection mechanism is unknown in practice and, therefore, correctly identifying variables which impact selection probability is challenging.

5.2 Selection probabilities depend on the study variable

Next, we drew repeated samples using Poisson sampling where the sampling probabilities depends on the survey variable of interest. We assume the Poisson sampling probabilities are given by:

$$\operatorname{logit}\left(p_i\right) = \beta_0 + \beta_1 y_i$$

where y_i is the amount of trade credit requested in millions of dollars, $\beta_1 = 0.5$ and $\beta_0 = -3.80, -2.85, -2.10.$ The intercept values, $\beta_0,$ were chosen such that we obtained sample sizes of approximately 200, 500 and 1,000 units, averaged over the simulated samples. Under this sampling design, units with larger amounts requested for trade credit have a higher probability of being sampled and, therefore, are over-represented. Table 5.2 displays the ARB of each estimator of the total amount of trade credit requested assuming $d_i = N/n,$ when the sample is selected using the above informative Poisson sampling. Here, all the estimators are heavily biased because the population model does not hold due to informative sampling. The magnitude of the bias is very similar across estimators and does not substantially decrease as the sample size increases. The inclusion or exclusion of revenue as an auxiliary variable does not impact the bias.

Table 5.2
Percent ARB of each estimator under Poisson sampling with revenue and without revenue included as an auxiliary variable

	Revenue			**Without Revenue**		
	β_0 = -3.8	**β_0 = -2.85**	**β_0 = -2.1**	**β_0 = -3.8**	**β_0 = -2.85**	**β_0 = -2.1**
GREG	23.53	22.27	20.45	24.74	22.91	21.21
FSTEP	24.54	22.55	20.58	25.16	23.24	21.15
TREE	24.07	22.73	20.15	24.93	22.47	20.55
LASSO (1-way)	24.29	22.73	20.65	25.45	23.29	21.38
CLASSO (1-way)	23.02	22.30	20.47	24.74	22.99	21.23
LASSO (2-way)	23.15	22.06	20.17	24.66	22.73	20.62
CLASSO (2-way)	20.11	20.18	19.01	22.62	21.63	19.98
ALASSO	24.44	22.72	20.66	25.50	23.21	21.36
CALASSO	23.91	22.46	20.53	25.10	23.01	21.25
HT	29.12	27.95	25.57	29.36	27.53	25.45

6. Conclusions

We have evaluated the performance of several model-assisted survey regression estimators, in the context of both probability and non-probability sampling, through a simulation study. First, we discuss the overall conclusions from our simulation study using probability samples with a stratified SRS design. In the context of our business survey data with all categorical auxiliary variables, the regression tree estimator and the lasso (2-way) estimator with two factor interaction effects are the only model-assisted estimators that provide any efficiency gains, relative to the HT estimator, when the sample size is small and the number of categories of auxiliary variables used is large. As well, the variance estimator for the regression tree estimator is the least biased in this scenario. As the sample size increases, the difference in efficiency between the model-assisted survey regression estimators becomes negligible and all are slightly more efficient than the HT estimator. In general, the potential gains in efficiency for model-assisted estimators over the HT estimator depend on the predictive power of the model. In our simulation population, the strength of the relationship between the study variable and the available categorical auxiliary variables is somewhat weak as judged by the adjusted coefficient of determination R^2 around 0.20. We therefore generated study variables leading to larger R^2 values around 0.50 by making the model error variance smaller. As expected, model-assisted estimators led to significant efficiency gains over the HT estimator in all cases, as reported in Table 4.2 which shows that the regression tree estimator and the lasso estimator with interaction effects yield improved efficiency over the commonly used GREG estimator if two-factor interactions are present. Moreover, the regression weights for the tree estimator and the calibration weights for the lasso calibration estimators are much less variable, particularly for small sample sizes, than the weights for the GREG. We also examined the performance of the lasso-based and regression trees estimators under a scenario with no main effects and only two-factor interactions are present and another scenario where multi-collinearity among the auxiliary variables is present. In the latter scenario, GREG is not applicable, and we show that the regression tree and lasso estimators provide an automatic way of removing colinear auxiliary variables without impacting the potential efficiency gains. Overall, we recommend using either lasso (2-way) or regression tree estimators in terms of efficiency when two factor interactions are likely to be present among the categorical auxiliary variables. Even in the case of models with only main effects, both methods perform well relative to GREG in terms of MSE because the lasso (2-way) estimator automatically shrinks regression coefficients associated with the interactions to zero while the regression tree estimator does not require specification of the mean function. In other contexts where there is evidence of complex non-linear and non-additive relationships between the survey variable of interest and auxiliary variables, the use of other tree-based machine learning methods, such as xgboost and random forests, should be studied.

In Section 4.3, we studied the performance of variance estimators in terms of relative bias and showed that all the variance estimators exhibit significant underestimation for sample size $n = 200$ and 28 x-categories. Relative bias of the regression tree variance estimator did not decrease as the sample size increased, unlike in the other cases, and it could be due to overfitting. In the context of random forests method, Dagdoug, Goga and Haziza (2021) examined a procedure based on cross-validation which led to small relative biases and good coverage rates. It would be worthwhile to study a similar procedure for variance estimation of the regression tree estimator.

A major drawback of the regression tree and lasso-based approaches is that the estimation procedures do not yield a set of generic weights that can be applied to all study variables, $y.$ A possible alternative approach is to derive regression weights based on a primary variable of interest and apply that set of weights to related study variables. In the survey context considered here, using a single set of weights for a group of related variables resulted in little loss of efficiency, relative to the use of variable-specific weights. As well, the bias of the estimators remained negligible. Under this approach, the desirable properties of the regression weights, low variability and, in the case of the regression tree estimator, strictly positive weights are maintained. However, the asymptotic properties of the lasso and regression tree survey estimators have not been derived for a single set of weights, applied to multiple study variables.

We also considered the use of model-assisted survey regression estimators for data from mis-specified probability sampling, treated as a non-probability sample. When the probability of selection depends on an observed auxiliary variable, the bias of the model-assisted estimators decreases as the sample size increases. Including the appropriate auxiliary variable in the working model for the GREG estimator effectively removes the selection bias. Achieving this in practice is difficult as the selection process is unknown. Performing variable selection can increase the bias for model-assisted survey regression estimators as the auxiliary variables related to the selection probability may not be included in the regression model. In fact, in our simulations, correctly including revenue as a potential auxiliary variable did not necessarily decrease the bias of the lasso estimators.

When the probability of selection depends on the survey variable of interest, all the estimators are heavily biased. The magnitude of the bias is similar across estimators and does not greatly decrease as the sample size increases. In our simulation population, the auxiliary variables are not highly predictive for the survey variables of interest. Examining the impact of the strength of the relationship between the auxiliary variables and the variable of interest when informative selection is present warrants more investigation.

Sample selection bias may not be reduced by using a non-probability sample alone, as demonstrated in our simulation study. Methods based on integrating a non-probability sample observing the study variables and associated auxiliary variables with a probability sample observing only the same auxiliary variables have the potential of reducing selection bias through modeling the participation probabilities (Chen, Li and Wu, 2020). Dual frame screening methods are also available when the study variable is observed in both samples and the units in the probability sample belonging to the non-probability sample can be identified without linkage errors without the need to model the participation probabilities (Kim and Tam, 2020; Rao, 2021 and Beaumont, 2020). However, the dual frame method is effective only when the sampling fraction for the non-probability sample is large. We are studying the above methods in the context of business surveys, for example integrating survey data with incomplete administrative data treated as a non-probability sample.

Acknowledgements

We thank Dr. Wesley Yung for initiating this work and for his constructive comments and suggestions. We also thank the reviewers, the editor and the associate editor for constructive comments and suggestions.

References

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46, 1, 1-28. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020001/article/00001-eng.pdf.

Breidt, F.J., and Opsomer, J.D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2), 190-205.

Buskirk, T.D., Kirchner, A., Eck, A. and Signorino, C.S. (2018). An introduction to machine learning methods for survey researchers. *Survey Practice*, 11(1), 1-10.

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

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(523), 2011-2021.

Chen, J.K.T., Valliant, R.L. and Elliott, M.R. (2018). Model-assisted calibration of non-probability sample survey data using adaptive LASSO. *Survey Methodology*, 44, 1, 117-144. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2018001/article/54963-eng.pdf.

Chen, J.K.T., Valliant, R.L. and Elliott, M.R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3), 657-681.

Dagdoug, M., Goga, C. and Haziza, D. (2021). Model-assisted estimation through random forests infinite population sampling. *Journal of the American Statistical Association* (to appear).

Friedman, J., Hastie, T., Simon, N., Qian, J. and Tibshirani, R. (2017). *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models.* R package version 2.0-13.

Kern, C., Klausch, T. and Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods*, 13(1), 73-93.

Kern, C., Li, Y. and Wang, L. (2020). Boosted kernel weighting-using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*. https://doi.org/10.1093/jssam/smaa028.

Kim, J.K., and Tam, S.M. (2020). Data integration combing big data and survey sample data for finite population inference. *International Statistical Review*, 89(2), 382-401.

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 Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46(2), 389-413.

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

McConville, K.S., Tang, B., Zhu, G., Li, S., Cheung, S. and Toth, D. (2018). *mase: Model-Assisted Survey Estimators*. R package version 0.1.1.

Rafei, A., Flannagan, C.A. and Elliott, M.R. (2020). Big data for finite population inference: Applying quasi-random approaches to naturalistic driving data using Bayesian additive regression trees. *Journal of Survey Statistics and Methodology*, 8(1), 148-180.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83(1), 242-272 (published online April 2020).

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. and Firth, D. (2017). *MASS: Modern Applied Statistics with S.* R package version 7. 3-47.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag Publishing.

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

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association,* 101(476), 1418-1429.