Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Relative Performance of Methods Based on Model-Assisted Survey Regression Estimation

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

Release date: October 29, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society’s Emerging Needs

Relative Performance of Methods Based on Model-Assisted Survey Regression Estimation

Erin R. Lundy [1]and J.N.K. Rao[2]

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.

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.

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, many 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.

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 $U = \{1, \dots, N\}$ 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 $p(\cdot)$, 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

[1] Erin R. Lundy, Statistical Integration Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6
[2] J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6

variables, $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \dots, x_{id})^T$ are known for each $i \in U.$ A standard approach is to use the Horvitz-Thompson estimator

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

where $d_i = \pi_i^{-1}$ denotes design weights.

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 $\boldsymbol{x}$ and use this model to predict y values. Specifying a linear regression working model leads to the generalized regression (GREG) estimator (Cassel, Sarndal and Wretman, 1976). Here, we consider the GREG estimator under a linear regression working model

$$y_i = \boldsymbol{x_i^T \beta} + \epsilon_i \qquad (2.1)$$

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

$$\hat{t}_{y,GREG} = \sum_{i\epsilon s} \frac{y_i - \boldsymbol{x_i^T \widehat{\beta}_s}}{\pi_i} + \sum_{i\epsilon U} \boldsymbol{x_i^T \widehat{\beta}_s} \; (2.2)$$

where $\boldsymbol{\widehat{\beta}_s}$ is a vector of estimated regression coefficients.

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 :

$$\hat{t}_{y,GREG} = \sum_{i\in s} \left[1 + (\boldsymbol{t_x} - \boldsymbol{\hat{t}_{x,HT}})^T \left(\sum_{k\in s} \boldsymbol{x_k x_k^T} d_k\right)^{-1} \boldsymbol{x_i}\right] d_i y_i = \sum_{i\in s} w_i \, y_i, \; (2.3)$$

where $\boldsymbol{t_x}$ is the known population total vector of the covariates $\boldsymbol{x}$ and $\boldsymbol{\hat{t}_{x,HT}}$ is the Horvitz-Thompson estimator vector of the covariate population totals $\boldsymbol{t_x} = \sum_{i\,\in U} \boldsymbol{x_i}$.

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 GREG weights W_i , and greatly inflate the variance of the

GREG estimator. Also, 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 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

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{s},\boldsymbol{L}} = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{Y_s} - \boldsymbol{X_s\beta})^{\boldsymbol{T}}\boldsymbol{\Pi}_{\boldsymbol{s}}^{-\boldsymbol{1}}(\boldsymbol{Y_s} - \boldsymbol{X_s\beta}) \; + \; \lambda \textstyle\sum_{j=1}^{p}\left|\beta_j\right|,$$

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

$$\hat{t}_{y,LASSO} = \textstyle\sum_{i\epsilon s}\frac{y_i - \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{s},\boldsymbol{L}}}{\pi_i} + \textstyle\sum_{i\epsilon U}\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{s},\boldsymbol{L}}.$$

2.3 Lasso Calibration Estimators

The 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. 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 $\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{s},\boldsymbol{L}}$. The lasso calibration estimator can be written in the same form as (2.3), where $\boldsymbol{x_i}$ is replaced by $\boldsymbol{x}_{\boldsymbol{i}}^{*} = (1, \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{s},\boldsymbol{L}})^{\boldsymbol{T}}$:

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

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.

2.4 Regression Tree Estimator

The GREG estimator can also be expressed as

$$\hat{t}_{y,r} = \sum_{i\epsilon s} \frac{y_i - \hat{h}_n(x_i)}{\pi_i} + \sum_{i\epsilon U} \hat{h}_n(\boldsymbol{x_i}) \text{ , (2.4)}$$

where $\hat{h}_n(\boldsymbol{x_i})$ is an estimator of the mean function of Y_i given $\boldsymbol{X_i} = \boldsymbol{x_i}$, $h(\boldsymbol{x_i}) = E(Y_i|\boldsymbol{X_i} = \boldsymbol{x_i})$, based on the sample data $(y_i, \boldsymbol{x_i}), i \in s$. As an alternative to a linear regression model, McConville and Toth (2019) proposed estimating $h(\boldsymbol{x})$ with a regression tree model. 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.

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

$$\tilde{h}_n(\boldsymbol{x_i}) = \widetilde{\#}(B_{nk})^{-1} \sum_{i \in s} \pi_i^{-1} y_i \, I(\boldsymbol{x_i} \in B_{nk}) = \tilde{\mu}_{nk}, \qquad (2.5)$$

Where

$$\widetilde{\#}(B_{nk}) = \sum_{i \in s} \pi_i^{-1} \, I(\boldsymbol{x_i} \in B_{nk})$$

is the HT estimator of the population size in box B_{nk}.The regression tree estimator $\hat{t}_{y,TREE}$ is obtained by inserting equation (2.5) into the generalized regression estimator, given in equation (2.4), leading to the post stratified estimator

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

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

Since $\tilde{h}_n(\boldsymbol{x_i})$ 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.

3. Simulation Study using Financing and Growth of Small and Medium Enterprises Survey Data.

Next, 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. The Survey of Financing and Growth of Small and Medium Enterprises (SFGSME) is a periodic survey of enterprises which collects information on the types of financing businesses use.

3.1 Simulation Methodology

We considered sample sizes of $n = \{200, 500, 1000\}$ from the 9115 respondents in the SFGSME dataset. This dataset was treated as the target population and repeated samples were drawn using stratified simple random sampling. 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.

For each sample, models using four categorical *x*-variables, industry (10 categories), employment size (4 categories), region (6 categories) and revenue (8 categories) were used to estimate total amount of trade credit requested and results were compared to the true total. For each of the three different sample sizes, we drew 5000 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 3.1-1 below:

Table 3.1-1
Summary of model assisted estimators considered in simulation study

Estimator	**Auxiliary Data**	**Regression Weights**	**Calibration Totals**
GREG	Marginal totals	Independent of *y*	All auxiliary variables
GREG with forward variable selection (FSTEP)	Individual values	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		
Calibrated lasso (CLASSO)	Individual values	Dependent on *y*	Population size and lasso-fitted mean function

3.2 Performance of Estimators in Terms of Design MSE

We computed design mean square error (MSE) from the 5000 total estimates by sample size and the results are displayed in Table 3.2-1 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. As the sample size increases, the gains in efficiency of the model-assisted survey regression estimators, relative to the HT estimator, are essentially equal. For larger sample sizes, 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.

Table 3.2-1
Ratio of MSE of Each Estimator to MSE of HT Estimator

	n=200	n=500	n=1000
GREG	1.084	0.959	0.954
FSTEP	1.040	0.945	0.958
TREE	0.983	0.963	0.949
LASSO(1-way)	1.009	0.946	0.947
CLASSO(1-way)	1.042	0.952	0.949
LASSO(2-way)	0.981	0.935	0.936
CLASSO(2-way)	1.045	0.959	0.950

The potential gains in efficiency for model-assisted estimators depend on the predictive power of the working model. 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 models:

- Lasso model with main effects only
- Lasso model with main effects and 2-way interactions
- Regression tree model
- Linear regression model with no main effects and some 2-way interactions
- Linear regression model with no main effects and a single 3-way interaction

Table 3.2-2 displays the ratio the design MSE of each estimator to that of the HT estimator under the lasso and regression tree models generating the survey variable of interest for a sample size of *n=1000.* 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. In all cases, significant efficiency gains, relative to the design-based HT estimator, are achieved

Table 3.2-2
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

Table 3.2-3 shows the ratio the design MSE of the estimators to that of the HT estimator, where the survey variable is generated from models with no main effects for sample sizes of *n=200* and *n=1000*. Here, 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 for larger sample sizes. 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 3.2-2 to the 2-way interaction column in Table 3.2-3. 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 3.2-3
Ratio of MSE for Each Estimator to MSE of HT under Models with No Main Effects

	n=200		n=1000	
	2-way interactions	Single 3-way interaction	2-way interactions	Single 3-way interaction
GREG	1.045	1.044	0.935	0.911
FSTEP	1.042	1.013	0.935	0.910
TREE	1.015	0.975	0.824	0.796
LASSO(1-way)	0.982	0.959	0.930	0.899
CLASSO(1-way)	1.031	0.985	0.936	0.902
LASSO(2-way)	0.912	0.957	0.783	0.815
CLASSO(2-way)	0.990	1.010	0.795	0.818

4. Conclusions

We have evaluated the performance of several model-assisted survey regression estimators, in the context of probability sampling, through a simulation study. 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. 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. We therefore generated study variables such that there was stronger relationship between the study variable and the available categorical auxiliary variables. As expected, model-assisted estimators led to significant efficiency gains over the HT estimator in all cases, as reported in Tables 3.2-2 and 3.3-3 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. 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 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.

Ongoing work includes further identification of scenarios where machine learning methods for calibration perform significantly better than the traditional methods. We are also studying model-assisted estimation under non-probability sampling.

References

Cassel CM, Sarndal CE and Wretman JH. (1976), “Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Population”, *Biometrika,*, 63, pp. 615-620.

McConville KS, Breidt FJ, Lee TCM and Moisen GG. (2017), “Model-assisted Survey Regression Estimation with the LASSO”, *Journal of Survey Statistics and Methodology*, 5, pp. 131-158.

McConville KS and Toth D. (2019), “Automated Selection of Post-strata using a Model-assisted Regression Tree Estimator”, *Scandinavian Journal of Statistics*, 46, pp. 389-413.

Tibshirani R. (1996), “Regression Shrinkage and Selection via the LASSO”, *Journal of the Royal Statistical Society, Series B,* 58, pp. 267-288.