Original Research Article

Evaluating Model-Assisted Estimators: A Comparative Study in High-Dimensional Survey Data

ABSTRACT

Model-assisted estimators have gained significant attention due to their ability to efficiently utilize auxiliary information during the estimation process. These estimators rely on a working model that links the survey variable to the auxiliary variables, which is then fitted to the sample data to generate predictions. These predictions are subsequently integrated into the estimation procedures. In this study, were explores various model-assisted estimators including Generalized Regression (GREG), Ridge regression, Lasso regression, CART (Classification and Regression Tree), Random Forest, Cubist and Principal Components Regression (PCR) estimator. The analysis involved 2,000 samples of size 50 (n/N ≈ 10%) and employed a stepwise variable selection method to determine the most significant auxiliary variables, incrementally adding them to the model. The performance of these estimators was assessed using relative bias (RB), relative root mean square error (RRMSE) and relative efficiency (RE). Our findings reveal that tree-based models like CART and Random Forest and penalized regression estimators such as Ridge and Lasso display robustness with increased number of auxiliary variables. Among all the estimators, Random Forest consistently yielded the lowest RRMSE, particularly with five auxiliary variables, demonstrating superior efficiency. Conversely, the GREG estimator exhibited poor performance as the number of auxiliary variables increased. This study underscores the importance of selecting suitable model-assisted estimation procedures tailored to the data characteristics and the relationship between survey and auxiliary variables within this high-dimensional dataset.

Keywords: *Design consistency, GREG, CART, random forest, Cubist, PCR, RB, RRMSE, RE.*

1. Introduction

"The purpose of a sample survey is to gather information about a population by sample of that population. Since surveying an entire population can be costly, time-consuming, or impractical in some situation, sample surveys allow researchers to make inferences about the population's characteristics. One of the key purposes of sample surveys is to estimate various characteristics or parameters of a population, such as averages, proportions, totals, or variances. Population totals can be estimated unbiasedly using the well-known Horvitz–Thompson estimator" (1). "In the absence of non-sampling errors, the Horvitz–Thompson estimator is unbiased, where the properties of estimators are assessed based on the sampling design, as discussed in" (2). "However, in certain cases, the Horvitz–Thompson estimator can exhibit high variance. Its efficiency can be enhanced by incorporating auxiliary information that leverages the relationship between the survey variable Y and a set of auxiliary variables" ***X*** (3). "This approach, known as model-assisted estimation, utilizes a working model to construct point estimators. When the working model accurately captures the relationship between Y and the auxiliary variables, model-assisted estimators are expected to be more efficient than the Horvitz–Thompson estimator" (4)

"The class of model-assisted estimators encompasses a diverse range of procedures, many of which have been extensively explored both theoretically and empirically in the literature. When the working model is a standard linear regression model, the resulting estimator is the well-known generalized regression estimator (GREG), as discussed by Särndal" (5). Additionally,

model-assisted methods have been developed using other approaches, including, panelise regression estimators such as ridge regression (6) and lasso regression (7). Apart of these, machine learning based estimators are also available in literature such as principal component regression (8), regression trees (9,10), random forests (11) and Cubist (12).

"With the rapid advancements in information technology, now we can access to a diverse array of data sources, many of which encompass a vast number of observations across a wide range of variables. Traditionally, the properties of model-assisted estimators have been well established within the customary asymptotic framework of finite population sampling. This framework assumes that both the population size N and the sample size n increase indefinitely, while the number of auxiliary variables p remains fixed, implying that n must be large relative to p. However, this conventional approach is insufficient when dealing with high-dimensional data sets where p may be comparable to n, or even exceed it, i.e., $p > n$" (13). A more suitable asymptotic framework would account for the scenario where p, in addition to N and n, also tends to infinity. In this context, (8) explored dimension reduction via principal component analysis and demonstrated the design consistency of the resulting calibration estimator and (14) describes feature selection techniques for high-dimensional data with very small sample sizes. More recently, (15) examined the properties of the Generalized Regression (GREG) estimator from a model-based perspective, allowing for p to diverge, while (16) investigated the asymptotic variance of the calibration estimator as the number of calibration variables p approaches infinity. (17) provided an empirical comparison of several model-assisted estimators using data from the Irish Commission for Energy Regulation (CER) Smart Metering Project that was conducted in 2009–2010 (CER, 2011), first they generate study variable then applied different model-assisted estimators and draw the conclusions. To establish a general consistency result for a class of model-assisted estimators when the number of auxiliary variables p is permitted to grow to infinity. This class encompasses not only the Generalized Regression (GREG) estimator but also model-assisted estimators that utilize random forest, CART (classification and regression tree), principal component regression, penalization methods, including ridge regression and lasso. Despite these advancements, there remains a significant research gap. Specifically, there is a lack of comprehensive comparative analyses that assess the performance and consistency of various model-assisted estimators, particularly machine learning-based approaches, under high-dimensional conditions where p can exceed or be comparable to n. Addressing this gap through a thorough comparative study would provide valuable insights into the effectiveness of different estimators in such high-dimensional settings. In this study, we utilize real data that already includes the study variable, eliminating the need to generate it for our analysis. We compare various model-assisted estimators using several evaluation criteria, including Relative Bias, Percent Relative Root Mean Square Error (% RRMSE), and Relative Efficiency with respect to the Horvitz–Thompson estimator.

1. MATERIAL AND METHODS

2.1 Data description

In this study, data were used regarding applicability on real world scenario. The Boston Housing dataset used in this study is publicly available on the UCI Machine Learning Repository. The original dataset was first published by Harrison and Rubinfeld (18), available at: https://archive.ics.uci.edu/ml/datasets/Housing. The following figure shows all variables in this dataset. It consisted of 13 auxiliary variables with one study variable and 506 observations. Description of Boston Housing dataset as given in list 1.

list 1: Description of Boston Housing dataset

Code	Description
CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq. ft.
INDUS	Proportion of non-retail business acres per town
CHAS	Charles river dummy variable (= 1 if tract bounds river, o otherwise
NOX	Nitric oxides concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to1940
DIS	Weighted distance to five Boston employment centres
RAD	Index of accessibility to radial highways
TAX	Full-value property-tax rate per $ 10,000
PTRATIO	Pupil-teacher ratio by town
B	1000 (BK-0.63)^2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MDEV	Median value of owner-occupied homes in $1000's

2.2 Methodology

We consider a finite population U of size N, and s a sample selected from U according to a sampling design $d(s)$. Let $\pi_i = P(k \in s)$ and $\pi_{ij} = P(i, j \in s)$ be the first and second-order inclusion probability for any units $i, j \in U$. We are interested in estimating

$$t_y = \sum_{i \in U} y_i \quad \ldots(1)$$

the population total of the survey variable Y. In the absence of non-sampling errors, the Horvitz– Thompson estimator is design-unbiased for t_y provided that $\pi_i > 0$ for all $i \in U$.

The Horvitz– Thompson estimator ($\hat{t}_\pi$):

$$\hat{t}_\pi = \sum_{i \in s} \frac{y_i}{\pi_i} \quad \ldots(2)$$

Let, $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{ip})'$ be the $\boldsymbol{x}$-vector of auxiliary variables associated with unit i. The $N \times p$ sampling design matrix is denoted by $\boldsymbol{X}_U = (\boldsymbol{x}_i)_{i \in U}$ and its sample of size s denoted by $\boldsymbol{X}_s = (\boldsymbol{x}_i)_{i \in s}$.

In the Model-assisted estimation we assume following underlying model:

$$y_i = f(\boldsymbol{x}_i) + e_i, \qquad i \in U, \qquad \ldots(3)$$

where, $f(\boldsymbol{x}_i)$ is an any unknown function of $\boldsymbol{x}$ and the error e_i are independent noise variables with mean 0 and constant variances σ^2.

The unknown function $f(\boldsymbol{x}_i)$ is estimated by $\hat{f}(\boldsymbol{x}_i)$ from the sample data $(\boldsymbol{x}_i, y_i)$; $i \in s$. Then the model-assisted estimator based on fitted model is:

$$\hat{t}_{ma} = \sum_{i \in U} \hat{f}(\boldsymbol{x}_i) + \sum_{i \in s} \frac{y_i - \hat{f}(\boldsymbol{x}_i)}{\pi_i} \qquad \ldots(4)$$

where, $\hat{f}(\boldsymbol{x}_i)$ denotes the predicted value of $f(\boldsymbol{x}_i)$ under the working model.

2.2.1 The GREG estimator

Suppose that the regression function $f(\boldsymbol{x}_i)$ is approximated by a linear combination of p auxiliary variables. The working model (4) reduces to

$$y_i = \boldsymbol{x}_i \boldsymbol{\beta} + e_i, \qquad i \in U, \qquad \ldots(5)$$

here, $\boldsymbol{\beta}$ is a vector of unknown coefficients. The least square estimate of $\boldsymbol{\beta}$ at the population level is given by:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}_U{}' \boldsymbol{X}_U)^{-1} \boldsymbol{X}_U \boldsymbol{y}_U \qquad \ldots(6)$$

$$\widehat{\boldsymbol{\beta}} = (\sum_{i \in U} \boldsymbol{x}_i \boldsymbol{x}_i{}')^{-1} \sum_{i \in U} \boldsymbol{x}_i y_i \qquad \ldots(7)$$

In sampling scenario, the $\widehat{\boldsymbol{\beta}}$ cannot be computed because the y-values are recorded for sample units only. The estimate of $\widehat{\boldsymbol{\beta}}$ at the sample level can be estimated through following adjustment:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{s}} = \left(\sum_{i \in s} \frac{\boldsymbol{x}_i \boldsymbol{x}_i{}'}{\pi_i} \right)^{-1} \sum_{i \in s} \frac{\boldsymbol{x}_i y_i}{\pi_i} \qquad \ldots(8)$$

Now the GREG estimator:

$$\hat{t}_{greg} = \sum_{i \in U} \boldsymbol{x}_i{}' \widehat{\boldsymbol{\beta}}_{\boldsymbol{s}} + \sum_{i \in s} \frac{y_i - \boldsymbol{x}_i{}' \widehat{\boldsymbol{\beta}}_{\boldsymbol{s}}}{\pi_i} \qquad \ldots(9)$$

2.2.1.1 Consistency of GREG estimator in high dimensional data

Consider a growing series of embedded finite populations $\{U_v\}$ $v \in N$, with sizes $\{N_v\}$. For each finite population U_v, a sample of size n_v is selected based on a sampling design, which involves first-order inclusion probabilities π_i and second-order inclusion probabilities π_{ij}. In this framework, as v increases, the finite population sizes $\{N_v\}$, the sample size and the number of auxiliary variables approach infinity.

The following assumptions are required to establish the consistency of the GREG estimators in a high-dimensional setting.

1) There exists a positive constant C such that $N_v^{-1} \sum_{i \in U_v} y_i{}^2 < C$.
2) We assume that $lim_{v \to \infty} \frac{n_v}{N_v} = \pi \in (0, 1)$.

3) There exists a positive constant k such that $min_{i\in U_v}\ \pi_i \geq k > 0$; also, we assume that $lim\ sup_{v\to\infty}\ n_v\ max_{\ i\neq j\in U_v}\ |\pi_{ij} - \pi_i\pi_j| < \infty$.
4) There exists a positive constant D such that, for all $i \in U_v, ||x_i|| \leq D\ p_v$, where $||\ \cdot\ ||$ denotes the usual Euclidean norm.
5) We assume that $||\hat{\beta}||_1 = O_P(p_v)$, where $\hat{\beta}$ is the least square estimator and $||\ .\ ||_1$ denotes the L 1 norm.

The first three assumption used by (19) to establish the consistency of GREG estimator in a fixed dimensional setting. The additional fourth and fifth assumption introduced for consistency of GREG estimator when p_v grow to infinity. Under the above five assumption we can establish:

$$\frac{1}{N}(\widehat{t}_{greg} - t_y) = O_P\left(\sqrt{\frac{p_v^3}{n_v}}\right) \quad \ldots(10)$$

This result highlights the fact that the rate of convergence decreases as the number of auxiliary variables increases. The result assures that the GREG estimator can remain consistent even if the number of auxiliary variables p_v grows, as long as the growth condition $\frac{p_v^3}{n_v} = O(1)$ is met. This result also reveals that if number of auxiliary variables p_v increase the rate of consistency deceases (5).

2.2.2 Penalized least square estimators

When multicollinearity is present in the data, the least squares estimates can become unstable and misleading, as the predictor variables are highly correlated, making it difficult to determine their individual contributions to the outcome. In a classical linear regression setting, penalization techniques like ridge regression and lasso can be employed to mitigate these issues. These methods introduce a penalty to the least squares criterion, which helps to shrink the coefficients and reduce variance, leading to more reliable estimates. Specifically, these estimators are obtained by minimizing a penalized least squares criterion at the population level, which balances the trade-off between fitting the data well and maintaining model simplicity. Let β_{pen} be an estimator of obtained through the penalized least square criterion at the sample level:

$$\widehat{\boldsymbol{\beta}}_{pen} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i\in s} \frac{(y_i - \boldsymbol{x_i}'\boldsymbol{\beta})^2}{\pi_i} + \lambda P(\beta) \quad \ldots(11)$$

where, y_i is the observed value of the response variable for the ith observation. λ is the regularization parameter that controls the strength of the penalty and P(β) is the penalty function applied to the coefficients β.

Ridge regression (6) adds a penalty proportional to the square of the coefficients, which is effective in handling multicollinearity by keeping all predictors but shrinking their influence. The formula for the Ridge regression estimator $\widehat{\boldsymbol{\beta}}_{ridge}$ is:

$$\widehat{\boldsymbol{\beta}}_{ridge} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i\in s} \frac{(y_i - \boldsymbol{x_i}'\boldsymbol{\beta})^2}{\pi_i} + \lambda \sum_{j\in p} \beta_j{}^2 \quad \ldots(12)$$

Lasso (Least Absolute Shrinkage and Selection Operator) regression (7) also modifies the OLS estimation, but it uses a penalty proportional to the absolute value of the coefficients. The formula for the Lasso estimator $\widehat{\boldsymbol{\beta}}_{lasso}$ is:

$$\widehat{\boldsymbol{\beta}}_{lasso} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i\in s} \frac{(y_i - \boldsymbol{x_i}'\boldsymbol{\beta})^2}{\pi_i} + \lambda \sum_{j\in p} |\beta_j| \quad \ldots(13)$$

2.2.3 Classification and Regression Trees (CART)

A decision tree is a tree-like model that is used for making decisions. It consists of nodes that represent decision points, and branches that represent the outcomes of those decisions. The decision points are based on the values of the input variables, and the outcomes are the possible classifications or predictions. A decision tree is constructed by recursively partitioning the input data into subsets based on the values of the input variables. Each partition corresponds to a node in the tree, and the partitions are chosen so as to minimize the impurity of the resulting subsets.

The CART algorithm is a decision tree-based algorithm that can be used for both classification and regression problems in machine learning. It works by recursively partitioning the training data into smaller subsets using binary splits. The tree starts at the root node, which contains all the training data, and recursively splits the data into smaller subsets until a stopping criterion is met. At each node of the tree, the algorithm selects a feature and a threshold that best separates the training data into two groups, based on the values of that feature. This is done by choosing the feature and threshold that maximizes the information gain or the Gini impurity, which are measures of how well a split separates the data. Once the tree is built, it can be used to make predictions by traversing the tree from the root node to a leaf node that corresponds to the input data. For regression problems, the prediction is the average of the target values in the leaf node.

Let we have small dataset with only 2 explanatory variable X_1 and X_2. The tree begins at the root node, where it evaluates the condition $X_1 \leq t_1$. If this condition is true, the tree moves to the left subtree; if false, it moves to the right. In the left subtree, another decision is made based on $X_2 \leq t_2$, leading to either region R_1 if true or region R_2 if false. On the right side of the root node, the tree makes a further split at $X_1 \leq t_3$. If this condition holds, the path leads to region R_3 ; otherwise, the tree continues to a subsequent decision node where $X_2 \leq t_4$ is evaluated. Depending on the result, the input is directed to either region R_4 or R_5. Each leaf node, representing regions R_1 to R_5, corresponds to a final prediction: in regression trees, the prediction value is the average of the target values of all the data points that fall into particular region while in classification trees, it would be the majority class of the data points (20).

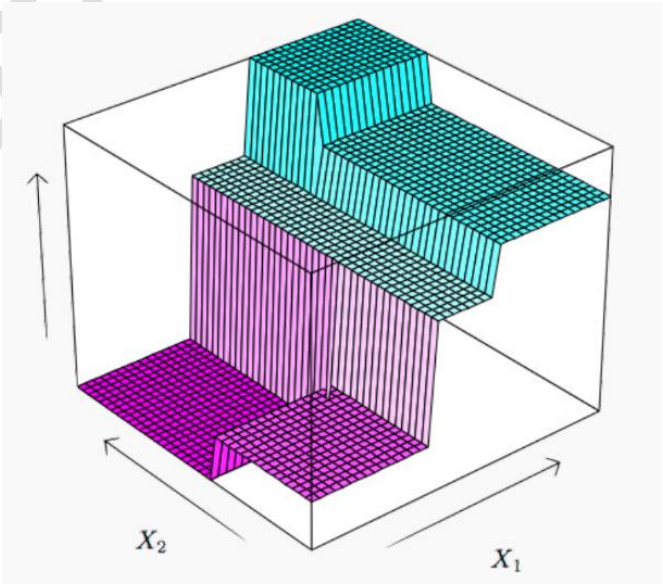

Fig. 1: Visualization of small dataset, in that Y is a vertical axis.

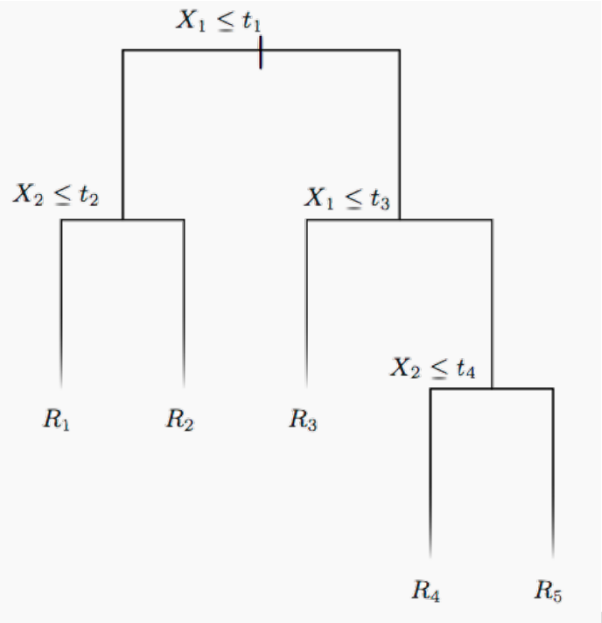

Fig. 2: Decision tree of small dataset.

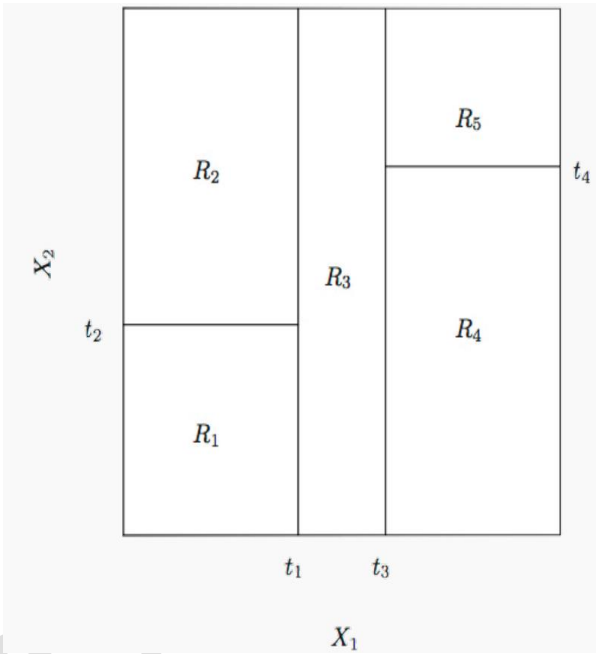

Fig. 3: Region (R_i) for prediction.

2.2.4 Random Forest

Random forest is a supervised learning algorithm, meaning that the data on which it operates contains labels or outcomes. It works by creating many decision trees, each built on randomly chosen subsets of the data. The model then aggregates the outputs of all of these decision trees to make an overall prediction for unseen data points. In this way, it can process larger datasets and capture more complex associations than individual decision trees.

Random forests start by creating multiple decision trees. Each tree is built using a subset of the training data, selected through a process called bootstrapping. Bootstrapping involves randomly selecting N observations (rows) from the training dataset with replacement. This means some observations may be selected multiple times, while others may not be selected at all.

Ensemble learning is the process of using multiple models, trained over the same data, averaging the results of each model ultimately finding a more powerful predictive/classification result. Random forests combine multiple trees to make a prediction (21).

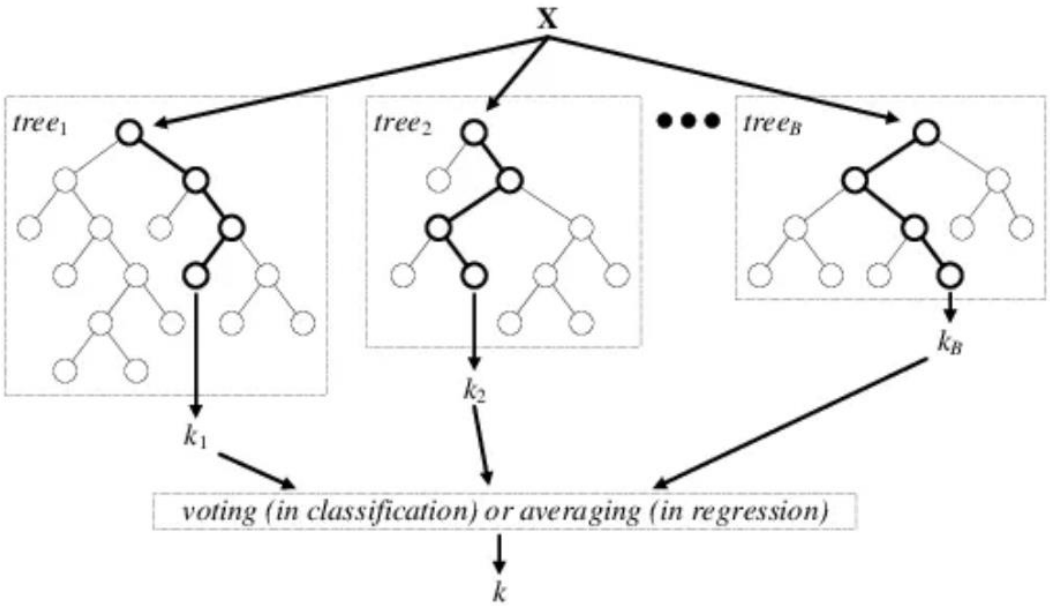

Fig. 4: B number of decision trees by Random Forest machine learning algorithm

2.2.5 Cubist

Cubist is a rule-based machine learning model. Introduced by Quinlan (22). It is an algorithm that combines regression trees with linear models. These elements work together to improve prediction accuracy.

A basic model tree is created using fit the linear regression model corresponding for each node. The data is recursively split into subsets based on the values of the predictor variables, forming a decision tree. Each split is chosen to minimize the prediction error for the target variable. For each terminal node (leaf) in the tree, a rule is created. This rule represents the path from the root of the tree to that particular leaf. The rule is composed of the conditions on the predictor variables that define the splits along the path. In each node of the tree, a linear model is fitted between the target variable Y and the auxiliary variables are used to split the tree.

Let consider the jth terminal node A_j. There is a path leading from the root node to this terminal node A_j, which involves a subset of the total auxiliary variables. For example, suppose the tree in Figure 5 partitions the data into five segments. The linear model at node A_1 would be constructed using the variables X_1, X_4 and X_6 which form the red path in the figure, making the number of auxiliary variables is 3. On the other hand, the linear model at node A_4 would be based solely on X_1 (the green path). The coefficients $\beta_j \in \mathbb{R}^p$ (p is the number of auxiliary variables) for the linear model at node A_j are estimated using standard weighted least squares (22).

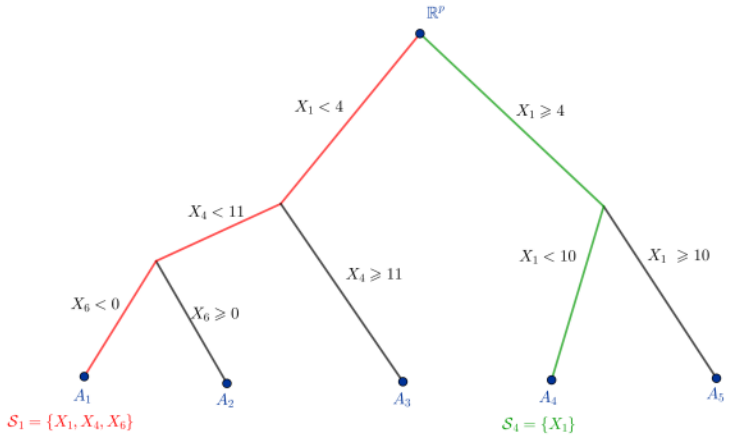

Fig. 5: Decision tree structure by Cubist machine learning algorithm

2.2.6 Principal Components regression

Principal Component Regression (PCR) is a regression technique that combines Principal Component Analysis (PCA) with linear regression. It is used primarily when dealing with multicollinearity in a dataset, where predictor variables are highly correlated with each other. The PCA transforms the original correlated variables into principal components, which are orthogonal to each other and have variances in decreasing order.

Let $\boldsymbol{X}$ be the matrix of p original auxiliary variables then jth principal component, denoted by $\boldsymbol{Z_j}$, is defined as follows

$$\boldsymbol{Z_j} = \boldsymbol{X v_j}$$

Where, $\boldsymbol{v_j}$ is a jth eigen vector, $j = 1, 2, \ldots, p$.

The important property of principal component regression is that, instead of using all the principal components, PCR typically selects the first few components that capture most of the variance in the predictors. The number of components to retain is usually determined by examining a scree plot or by cross-validation. The selected principal components are then used as predictors in a linear regression model to predict the response variable. Since the principal components are uncorrelated, the multicollinearity problem is effectively removed.

Let, consider the only first r (with $r < p$) principal components, $\boldsymbol{Z_1}, \boldsymbol{Z_2}, \ldots, \boldsymbol{Z_j}$, which correspond to the r largest eigenvalues.

The least squares estimation, at the population level, of $\widehat{\boldsymbol{\beta}}$, is

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i \in s} \frac{\boldsymbol{z_{ir}}\, \boldsymbol{z_{ir}}'}{\pi_i}\right)^{-1} \sum_{i \in s} \frac{\boldsymbol{z_{ir}}\, y_i}{\pi_i} \qquad \ldots (14)$$

Where, $\boldsymbol{z_{ir}} = (z_{i1}, \ldots, z_{i2})$ is the vector containing the values of the first r PCs computed for the i^{th} individual. $\widehat{\boldsymbol{\beta}}$ cannot be computed because the y-values are recorded for sample units only.

2.3 Criteria for Evaluating Estimators

The measure of bias of the model-assisted estimators $\hat{t}_{ma}$, computed the Monte Carlo percent Relative Bias defined as:

$$\%RB \;=\; 100 \times \frac{1}{R}\sum_{r=1}^{R}\frac{(\hat{t}_{ma}^{(r)} - t_y)}{t_y} \qquad \ldots(15)$$

where, $\hat{t}_{ma}^{(r)}$ and t_y denotes the estimator $\hat{t}_{ma}$ at the r^{th} iteration, $r = 1,2,..,R$ and population total respectively.

The Relative Root Mean Square Error (%RRMSE) is defined as:

$$\%RRMSE \;=\; 100 \times \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{t}_{ma}^{(r)} - t_y}{t_y}\right)^2} \qquad \ldots(16)$$

The %RRMSE measures how much the model's predictions differ from the actual true values, on average. This comparison is done multiple times (over several replications), and the result is expressed as a percentage. The lower the %RRMSE, indicating better accuracy.

The measure of efficiency, they computed the percentage Relative Efficiency (%RE), using the Horvitz–Thompson estimator given by $\hat{t}_{\pi}$. That is,

$$\%RE \;=\; 100 \times \frac{MSE(\hat{t}_{ma})}{MSE(\hat{t}_{\pi})} \qquad \ldots(17)$$

where,

$$MSE(\hat{t}_{ma}) = \tfrac{1}{R}\textstyle\sum_{r=1}^{R}\left(\hat{t}_{ma}^{(r)} - t_y\right)^2 \qquad \ldots(18)$$

$$MSE(\hat{t}_{\pi}) = \tfrac{1}{R}\textstyle\sum_{r=1}^{R}\left(\hat{t}_{\pi}^{(r)} - t_y\right)^2 \qquad \ldots(19)$$

2. Result and Discussion

This study conducts a comprehensive comparison of model-assisted estimators, focusing on their efficiency and bias in estimating population totals using Simple Random Sampling Without Replacement (SRSWOR) under varying conditions of auxiliary variables. We selected R = 2000 samples, of size 50 which corresponds to a sampling fraction n/N of about 10% with simple random sampling without replacement.

In the analysis, we employed a stepwise variable selection method to identify the most significant auxiliary variables, arranging them in descending order of significance. This approach enabled us to systematically assess the impact of progressively incorporating additional auxiliary variables into the model. Our objective was to observe the evolutionary changes in the performance of estimators as the number of auxiliary variables used in model prediction increased.

To achieve this, we initiated the analysis with a base model including only three auxiliary variables and incrementally increased the number by two, up to a maximum of eleven variables. This incremental approach allowed us to carefully evaluate and compare the performance of various model-assisted estimators under different conditions, thereby providing insights into the relationship between the number of auxiliary variables and estimator efficiency.

In each sample, we computed model-assisted estimators of the form

$$\hat{t}_{ma} = \sum_{i \in U} \hat{f}(\boldsymbol{x}_i) + \sum_{i \in s} \frac{y_i - \hat{f}(\boldsymbol{x}_i)}{\pi_i} \quad \ldots(20)$$

where the predictors $\hat{f}(\boldsymbol{x}_i)$, were obtained using the different model procedures such as GREG, Ridge regression, Lasso regression, CART, Random Forest with 500 trees, Cubist and Principal components regression based on the first few components kept which exhibit more than 90% of total variability.

The performance metrics, including relative bias (RB), relative root mean square error (RRMSE), and relative efficiency (RE) for various regression models, are summarized in Tables 1, 2, and 3, respectively. Figure 7 presents a comparative study of the performance of different estimators, illustrating the variations in relative efficiency across the models.

Table 1: Relative bias (%) of model-assisted estimators for the estimation of the population total of with SRSWOR (n = 50) and increasing number of auxiliary variables.

Estimators	**RB (%)**				
No. of X's	**3**	**5**	**7**	**9**	**11**
HT	0.18	0.18	0.18	0.18	0.18
GREG	-0.39	-0.52	-0.76	-0.67	-0.69
Ridge	-0.32	-0.38	-0.45	-0.36	-0.34
Lasso	-0.39	-0.48	-0.63	-0.52	-0.51
CART	-0.11	-0.2	-0.23	-0.21	-0.16
Random Forest	-1.17	-1.16	-1.24	-0.87	-0.6
Cubist	-2.06	-1.79	-2.12	-1.96	-1.81
PCR	-0.23	-0.22	-0.16	0.04	0.05

Table 2: Percentage relative root mean square error of model-assisted estimators for the estimation of the population total of with SRSWOR (n = 50) and increasing number of auxiliary variables.

Estimators	**%RRMSE**				
No. of X's	**3**	**5**	**7**	**9**	**11**
HT	5.50	5.50	5.50	5.50	5.50
GREG	3.48	3.30	3.39	3.47	3.76
Ridge	3.48	3.26	3.28	3.31	3.44
Lasso	3.49	3.33	3.39	3.43	3.55
CART	3.52	3.59	3.60	3.62	3.61
Random Forest	2.98	2.86	3.02	2.90	2.93
Cubist	5.09	4.22	4.35	4.16	4.05
PCR	3.79	3.66	3.81	4.92	4.92

Table 3: Relative efficiency (%) of model-assisted estimators for the estimation of the population total of with SRSWOR (n = 50) and increasing number of auxiliary variables.

Estimators	**RE (%)**				
No. of X's	**3**	**5**	**7**	**9**	**11**
GREG	39.88	35.98	37.99	39.79	46.77
Ridge	40.07	35.02	35.44	36.2	39.06
Lasso	40.12	36.64	37.83	38.73	41.56

CART	40.85	42.54	42.89	43.3	42.97
Random Forest	29.32	27.03	30.14	27.82	28.44
Cubist	85.68	58.89	62.36	57.19	54.05
PCR	47.37	44.25	47.8	79.89	79.85

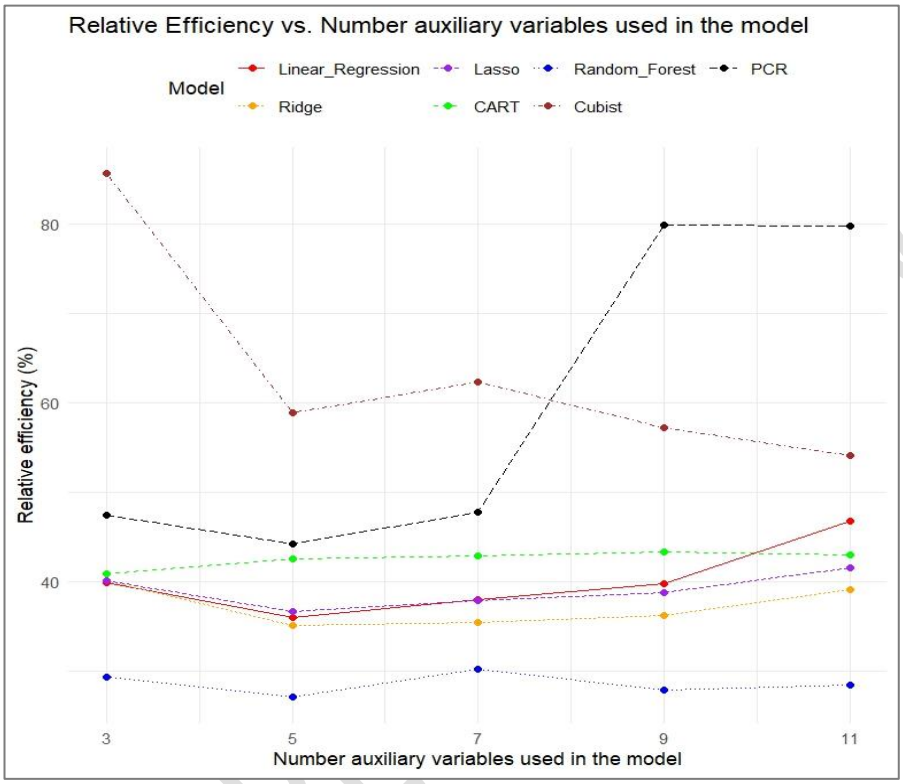

Fig. 6: Relative efficiency of model-assisted estimators for the estimation of the population total of with SRSWOR (n = 50) and increasing number of auxiliary variables

The table 1 presents the elative bias percentages of various model-assisted estimators used for estimating population totals in simple random sampling without replacement (SRSWOR) while varying the number of auxiliary variables. The Generalized Regression (GREG) estimator shows increasing negative bias from -0.39 to -0.69 as auxiliary variables increase, suggesting potential overfitting. Ridge and Lasso estimators exhibit slight negative bias but remain less effective with more variables. Tree-based estimators like CART and Random Forest present decreased bias with added variables, indicating robustness. In contrast, Cubist shows the highest negative bias values, hinting at sensitivity to complexity. Table 2 shows that the Random Forest estimator achieves the lowest percentage relative root mean square error (%RRMSE), peaking at 2.86 with five auxiliary variables, indicating superior efficiency compared to others. Conversely, the Generalized Regression (GREG) estimators demonstrate an increase in error as more variables are added, highlighting its diminished efficacy. The accompanying figure depict the relative efficiencies with respect to HT estimator of the different modal assisted estimators, where Random Forest stands out, significantly improving as the number of auxiliary variables rises, while estimators like Lasso and CART show stability

but lower overall efficiency. Overall, these results underscore the importance of estimator selection based on both bias and error metrics to optimize estimation accuracy.

3. Conclusions

Our examination of model-assisted estimation procedures in high-dimensional settings reveals significant insights into the performance of various methodologies. Analysis based on Boston household pricing data reveal the relationship between the survey variable and auxiliary information is well-captured by penalized estimators such as ridge and lasso demonstrate, CART and Random Forest based estimators gave robust performance, exhibiting high efficiency. Conversely, the model-assisted estimators based on random forests give best estimator with low value of %RRMSE. All the model-assisted estimators perform better than HT estimator. However, the Generalized Regression Estimator (GREG) demonstrates limitations in scenarios with numerous auxiliary variables, suffering from poor performance. These findings underscore the importance of selecting appropriate estimation methods based on the characteristics of the data and the nature of the relationships involved.

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

Option 2:

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc have been used during writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology

Details of the AI usage are given below:

1.

2.

3.

References

1. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc. 1952;47(260):663–85.

2. Deville JC, Särndal CE. Calibration estimators in survey sampling. J Am Stat Assoc. 1992;87(418):376–82.

3. Merkouris T. On the Most Effective Use of Continuous Auxiliary Variables in Regression Estimation in Survey Sampling. International Statistical Review. 2024;

4. Stefan M, Hidiroglou MA. Jackknife bias-corrected generalized regression estimator in survey sampling. J Surv Stat Methodol. 2024;12(1):211–31.

5. Särndal CE. On π-inverse weighting versus best linear unbiased weighting in probability sampling. Biometrika. 1980;67(3):639–50.

6. Goga C, Shehzad MA. Overview of ridge regression estimators in survey sampling. Université de Bourgogne: Dijon, France. 2010;

7. McConville KS, Breidt FJ, Lee TCM, Moisen GG. Model-assisted survey regression estimation with the lasso. J Surv Stat Methodol. 2017;5(2):131–58.

8. Cardot H, Goga C, Shehzad MA. Calibration and partial calibration on principal components when the number of auxiliary variables is large. Stat Sin. 2017;243–60.

9. McConville KS, Toth D. Automated selection of post-strata using a model-assisted regression tree estimator. Scandinavian Journal of Statistics. 2019;46(2):389–413.

10. Toth D, Eltinge JL. Building consistent regression trees from complex sample data. J Am Stat Assoc. 2011;106(496):1626–36.

11. Dagdoug M, Goga C, Haziza D. Model-assisted estimation through random forests in finite population sampling. J Am Stat Assoc. 2023;118(542):1234–51.

12. Dagdoug M, Goga C, Haziza D. Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison. J Surv Stat Methodol. 2023;11(1):141–88.

13. Buchner J. Nested sampling methods. Statistic Surveys. 2023;17:169–215.

14. Kuncheva LI, Matthews CE, Arnaiz-González A, Rodríguez JJ. Feature selection from high-dimensional data with very low sample size: A cautionary tale. arXiv preprint arXiv:200812025. 2020;

15. Ta T, Shao J, Li Q, Wang L. Generalized regression estimators with high-dimensional covariates. Stat Sin. 2020;30(3):1135.

16. Chauvet G, Goga C. Asymptotic efficiency of the calibration estimator in a high-dimensional data setting. J Stat Plan Inference. 2022;217:177–87.

17. Dagdoug M, Goga C, Haziza D. Model-assisted estimation in high-dimensional settings for survey data. J Appl Stat. 2023;50(3):761–85.

18. Harrison Jr D, Rubinfeld DL. Hedonic housing prices and the demand for clean air. J Environ Econ Manage. 1978;5(1):81–102.

19. Robinson PM, Särndal CE. Asymptotic properties of the generalized regression estimator in probability sampling. Sankhyā: The Indian Journal of Statistics, Series B. 1983;240–8.

20. Breiman L. Classification and regression trees. Routledge; 2017.

21. Breiman L. Random forests. Mach Learn. 2001;45:5–32.

22. Quinlan JR. Learning with continuous classes. In: 5th Australian joint conference on artificial intelligence. World Scientific; 1992. p. 343–8.