Scientia Iranica E (2023) 30(3), 1224–1244

Sharif University of Technology
**Scientia Iranica**
*Transactions E: Industrial Engineering*
http://scientiairanica.sharif.edu

# A novel basis function approach to finite population parameter estimation

**Sh. Ahmed**[*,1] **and J. Shabbir**

*Department of Statistics, Quaid-i-Azam University, Islamabad, 44000, Pakistan.*

Received 6 July 2020; received in revised form 29 December 2020; accepted 8 March 2021

**KEYWORDS**
Superpopulation;
Survey sampling;
Inference;
Basis functions;
Feature matrix;
Non-linear function.

**Abstract.** Modeling nonlinear data is a common practice in data science and Machine Learning (ML). It is aberrant for the outcome of a natural process to vary linearly with the values of input variable(s). A robust and easy methodology is needed for accurately and quickly fitting a sampled dataset with a set of covariates, assuming that the sampled data can be a complicated nonlinear function. A novel approach to the estimation of finite population parameter $\tau$, which is a linear combination of the population values, is considered in this article under superpopulation setting with known Basis Functions Regression (BFR) models. The problems of subsets selection with a single predictor using an automatic matrix approach and ill-conditioned regression models are discussed. Prediction error variance of the proposed estimator is estimated based on widely used feature selection criteria in ML. Finally, the Expected Squared Prediction Error (ESPE) of the proposed estimator and the expectation of estimated error variance under bootstrapping as well as simulation study with different regularizers are obtained to observe the long-run behavior of the proposed estimator.



## 1. Introduction

In survey sampling, researchers mostly prefer random sampling for a valid statistical inference due to its attractive long run properties such as unbiasedness and efficiency in the design-based sense. However, they ignore the importance of underlying model relationship between the survey variable and one or more covariate(s) in the estimation stage. Without exposing an appropriate model relationship between the survey variable and the covariates, researchers in the design-based paradigm have been constructing estimators for estimating unknown population quantities such as population total, mean, variance, etc. relying only on the mechanism of randomization incurred by sampling. They have been utilizing sample estimates and known population parameters of the auxiliary variable(s) in the estimation stage for efficiency improvement. Thousands of estimators used for estimating population parameters have been developed in terms of efficiency improvement and bias reduction using the design-based approach; for instance, some of the related works can be found in [1–7]. On the contrary, advocates of the model-based paradigm emphasize that randomization is a property of error term used in a model; hence, it is neither necessary nor sufficient for a solid statistical inference [8]. In a model-based framework, initially, in [9], a regression model of the response on the covariates was used to predict the non-sampled values and their total, which is assumed to be an unknown and random quantity. Many varieties of model-based

1. *Present address: Department of Mathematics and Statistics, Institute of Business Management, Karachi, Pakistan.*
*. *Corresponding author.*
*E-mail address: shakeel.ahmed@iobm.edu.pk (Sh. Ahmed)*

doi: 10.24200/sci.2021.56353.4682

estimators have been developed for efficiency improvement, bias reduction, and robustness to model failure in the last two decades of the 20th century. In [10,11], researchers worked on estimating a smooth function to be used for predicting the non-sampled values in estimating the total finite population. The asymptotic bias of the regression estimator of population total does not account for the division of the sampling distribution into sampled and non-sampled parts. In [12], a class of estimators was investigated based on local polynomial regression that were weighted linear combinations of the study variables, where the weights were calibrated to control totals that are known. In [13], a model-based approach along with the Local Linear Regression (LLR) was employed to estimate the unknown parameters of the study variable. They particularly derived the properties of the proposed estimator and compared it with NadarayaWatson regression estimator [14,15]. They found that the two estimators were asymptotically equally efficient. In [16], it was found that the calibration estimator based on the columnar model slightly outperformed the Best Linear Unbiased Estimator (BLUE) at a higher band width. In general, the estimator is robust to bandwidth changes and provides exact unbiasedness as well as minimal variance for a specific weighted balanced sample. Calibration-based estimators provide insight into the balance sampling framework, which suggests selecting the samples with the condition of unbiasedness in the presence of reliable auxiliary data. They noticed that the total estimators of population total from a nonparametric regression model provided approximate unbiasedness without imposing restriction on balancing and results close to minimal variance. However, in [17], a more appealing strategy than the kernel regression, e.g., the variable bandwidth LLR approach, was uncovered. In [18], a model-based estimator that worked with penalized spline regression was proposed and the estimator was extended to two-stage sampling [19]. In [12], the Classical Local Polynomial Regression (CLPR) estimator was applied for the regression function to obtain the model-assisted estimator of the total in finite populations. In [20], a method for balancing which equalized the multivariate densities and reduced bias without searching for specifications was developed. The regression function with mixed variable was estimated using a modified form of local constant estimator [21]. The properties of a weighted nonparametric regression estimator were derived using probabilities as weights for complex surveys under combined inference [22]. Several partial solutions for balanced sampling are available in [8,23–25]. In [16], a general method, called the cube method, was proposed which is appropriate for a set of quantitative or qualitative balancing variables and allows unequal probabilities of inclusion. In [26], authors developed a cube method for selecting approximately balanced samples based on equal or unequal inclusion probabilities with a number of auxiliary variables. A balanced sampling strategy was formed in multi-way stratification settings for small area estimation and used to obtain the planned sample size for domains belonging to different partitions of the population (small areas) [27]. The strategy reduces the sampling errors of domain estimates and provides threshold values. In [28], the nonparametric estimation methods were considered for data analysis in complex surveys. Authors in [29] employed the LLR technique to assess the properties of the derived estimator and compared its performance with the existing estimators. The LLR technique can be also used for evaluating entrepreneurial opportunities. Therefore, this application and the following sentence should be added here: Note that the LLR technique is employed for evaluating entrepreneurial opportunities, see [30] for more information about this topic.

The researches documented in the literature consist of a wide variety of model-based estimators constructed under different forms of the relationship between the outcome and the predictors, see [31–38]. Although a wide variety of restricted sampling methods were presented in [8], some of them were based on the Linear Regression Model (LRM), some on polynomial models, and some on proportional and stratified population models. Further, we need a general framework for predicting responses from nonlinear (in variable) functions of auxiliary data. The nonlinear function of the auxiliary variable may be logarithm, power, or exponential form. The problem of concern is the prediction of output variable for non-sampled set based on the relationship between the inputs and outputs in the sampled set and the known values of the input variable(s) in the non-sampled set. There is a lack of data on the non-linear regression model under general prediction theorem. The existence of natural processes whose outcome varies linearly with the values of predictors is aberrant. To fill this gap, this study establishes a general framework using Basis Functions Regression (BFR) model to estimate the values of finite population parameters. A novel approach to the estimation of finite population parameter $\tau$, a linear combination of the population values, is suggested here in the superpopulation setting with known BFR models. The problems of subsets selection with a single predictor under an automatic matrix approach and ill-conditioned regression models are discussed. Prediction error variance of the proposed estimator is estimated based on the widely used feature selection criteria in Machine Learning (ML). Section 2 delineates a model-based estimation developed in the literature with its usual notations. Section 3 describes the proposed basis function approach along with some special cases. Section 4 estimates $\tau$ under regularized

BFR. Section 5 covers variance estimation and comparison of competing variance estimators. Sections 6 and 7 cover model selection and simulation studies, respectively. Section 8 concludes the study with future recommendations.

## 2. Model based estimation

Consider a finite population of size $N$ indexed as $\mathcal{U} = \{1, 2, 3, \cdots, N\}$ with responses $y$ corresponding to a random variable $Y$. In the matrix notation, $\boldsymbol{y} = (y_i, i \in \mathcal{U})$ is the realized stochastic vector of $\boldsymbol{Y} = (\boldsymbol{Y}_i, i \in \mathcal{U})$. Suppose that a sample $s = \{1, 2, 3, \cdots, n\}$ of size $n$ is drawn from the finite population $\mathcal{U}$ using some sampling design SD and $\bar{s} = (1, 2, 3, \cdots, N-n)$ is the set of indices attached to the values of units that are not indexed in $s$. For a given sample $s$, we can rearrange the population vector as $\boldsymbol{y} = (\boldsymbol{y}_s^T, \boldsymbol{y}_{\bar{s}}^T)^T$, where $\boldsymbol{y}_s$ and $\boldsymbol{y}_{\bar{s}}$ are the vectors of $n$ sampled and $N-n$ non-sampled values of the study variable, respectively. The underlying superpopulation model is expressed as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{X}$ is the known and non-stochastic data matrix containing $p$ regressors including intercept, $\boldsymbol{\beta}$ is the corresponding vector of coefficients and $\boldsymbol{\epsilon}$ is the vector of random error terms assumed to be distributed normally with conditional mean vector 0 and variance-covariance matrix $\boldsymbol{\Sigma}$. Further, the data matrix $\boldsymbol{X}$ and covariance matrix $\boldsymbol{\Sigma}$ can be partitioned as follows:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_s \\ \boldsymbol{X}_{\bar{s}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_{\bar{s}\boldsymbol{s}} = \begin{bmatrix} \boldsymbol{\Sigma}_{ss} & \boldsymbol{\Sigma}_{s\bar{s}} \\ \boldsymbol{\Sigma}_{\bar{s}s} & \boldsymbol{\Sigma}_{\bar{s}\bar{s}} \end{bmatrix}.$$

The quantity of interest to be estimated is a linear combination of the population values $\tau(y) = \boldsymbol{\gamma}^T \boldsymbol{y}$, which can be a realization of the random variable $\boldsymbol{\gamma}^T \boldsymbol{Y}$, where $\boldsymbol{\gamma} = (\gamma_i, i \in \mathcal{U})$ is the vector of weights that can be partitioned for sampled and non-sampled values as $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_s^T, \boldsymbol{\gamma}_{\bar{s}}^T)^T$ defined a linear estimator (known as Best Linear Unbiased Predictor (BLUP)) for $\tau(y)$ as $\hat{\tau}(y) = \boldsymbol{g}_s \boldsymbol{Y}_s$, where $\boldsymbol{g}_s = (g_i, i \in s)$ is a vector of constants to be optimized. Under model Eq. (1), ([39]) proposed a general prediction estimator for $\tau(y)$ as follows:

$$\hat{\tau}(y) = \boldsymbol{\gamma}_s^T \boldsymbol{Y}_s + \boldsymbol{\gamma}_{\bar{s}}^T \left[ \boldsymbol{X}_{\bar{s}} \hat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_{\bar{s}s} \boldsymbol{\Sigma}_{ss}^{-1} \left( \boldsymbol{Y}_s - \boldsymbol{X}_s \hat{\boldsymbol{\beta}} \right) \right], \tag{2}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_s^{-1} \boldsymbol{\Sigma}_{ss}^{-1} \boldsymbol{X}_s)^{-1} \boldsymbol{X}_s^T \boldsymbol{\Sigma}_{ss}^{-1} \boldsymbol{Y}_s$ is the Weighted Least Square (WLS) estimator of the vector $\boldsymbol{\beta}$. The variance of $\hat{\tau}(y)$ is given by:

$$\begin{aligned} V_M\left(\hat{\tau}(y) - \tau(y)\right) &= \boldsymbol{\gamma}_{\bar{s}}^T \left( \boldsymbol{\Sigma}_{s\bar{s}} - \boldsymbol{\Sigma}_{\bar{s}s} \boldsymbol{\Sigma}_{ss}^{-1} \boldsymbol{\Sigma}_{s\bar{s}} \right) \boldsymbol{\gamma}_{\bar{s}} \\ &+ \boldsymbol{\gamma}_{\bar{s}}^T \left( \boldsymbol{X}_{\bar{s}} - \boldsymbol{\Sigma}_{\bar{s}s} \boldsymbol{\Sigma}_{ss}^{-1} \boldsymbol{X}_s \right) \\ &\left( \boldsymbol{X}_s^T \boldsymbol{\Sigma}_{ss}^{-1} \boldsymbol{X}_s \right)^{-1} \left( \boldsymbol{X}_{\bar{s}} - \boldsymbol{\Sigma}_{\bar{s}s} \boldsymbol{\Sigma}_{ss}^{-1} \boldsymbol{X}_s \right)^T \boldsymbol{\gamma}_{\bar{s}}. \end{aligned} \tag{3}$$

When the sampled and non-sampled units are uncorrelated, i.e., $\boldsymbol{\Sigma}_{\bar{s}s} = 0$, the BLUP for $\tau(y)$ is reduced to:

$$\hat{\tau}(y) = \boldsymbol{\gamma}_s^T \boldsymbol{Y}_s + \boldsymbol{\gamma}_{\bar{s}}^T \boldsymbol{X}_{\bar{s}} \hat{\boldsymbol{\beta}}, \tag{4}$$

with prediction error variance:

$$\begin{aligned} &V_M\left(\hat{\tau}(y) - \tau(y)\right) \\ &\quad = \boldsymbol{\gamma}_{\bar{s}}^T \left\{ \boldsymbol{\Sigma}_{s\bar{s}} + \boldsymbol{X}_{\bar{s}} \left( \boldsymbol{X}_s^T \boldsymbol{\Sigma}_{ss}^{-1} \boldsymbol{X}_s \right)^{-1} \boldsymbol{X}_{\bar{s}}^T \right\} \boldsymbol{\gamma}_{\bar{s}}. \end{aligned} \tag{5}$$

The assumption of zero correlation does not hold in multistage surveys where the intra-cluster correlation among units within clusters exists. Assuming independent and identically distributed (iid) error term, i.e., $\boldsymbol{\Sigma}_{ss} = \sigma^2 \boldsymbol{I}_n$ and $\boldsymbol{\Sigma}_{s\bar{s}} = \sigma^2 \boldsymbol{I}_{N-n}$, we can write the prediction error variance as follows:

$$\begin{aligned} &V_M\left(\hat{\tau}(y) - \tau(y)\right) \\ &\quad = \sigma^2 \left[ \boldsymbol{\gamma}_{\bar{s}}^T \boldsymbol{\gamma}_{\bar{s}} + \boldsymbol{\gamma}_{\bar{s}}^T \boldsymbol{X}_{\bar{s}} \left( \boldsymbol{X}_s^T \boldsymbol{X}_s \right)^{-1} \boldsymbol{X}_{\bar{s}}^T \boldsymbol{\gamma}_{\bar{s}} \right]. \end{aligned} \tag{6}$$

The general prediction estimator was constructed using a general LRM of $\boldsymbol{Y}$ in a matrix of covariates $\boldsymbol{X}$. It is noteworthy that for generalizing the result from the sample to population, the sampler should make at least one model explicit from the underlying population. That would be possible when the sampler knows the functional form of the underlying population model. Thus, if one is concerned with finite population parameter estimation, it is inevitable to account for the chance of deviation from the model, which is difficult to detect from the data obtained from the sample. In such situations, it is necessary to robustify the sampling mechanism and/or estimator with respect to model failure. One way to robustify is to measure such effects as bias and variance and that how these measures change when the working model deviates from the true model. Royall and Herson [40] emphasized the balancing of a sample to protect the inference against model misspecification. Valliant et al. [8] conducted an extensive study on balance sampling for reducing the effect of bias introduced due to model failure. Apart from balancing, BFR may provide a general framework for estimation of finite population parameters after predicting non-sampled data through BFR models [41].

## 3. Model-based estimation using basis functions

Starting with a single input variable $X$, the corresponding vector basis function is defined as $\boldsymbol{\Phi}(x_i) = (\Phi_0(x_i), \Phi_1(x_i), \cdots \Phi_M(x_i))$ which are attached to the $i$th population unit, where $M$ is the number of basis

functions in the model. The matrix consisting of the basis function is known as a feature matrix in ML terminology and is presented as follows:

$$\boldsymbol{\Phi}=\begin{bmatrix}\Phi_0(x_1) & \Phi_1(x_1) & \Phi_2(x_1) & \cdots & \Phi_{M-1}(x_1)\\ \Phi_0(x_2) & \Phi_1(x_2) & \Phi_2(x_2) & \cdots & \Phi_{M-1}(x_2)\\ \vdots & \vdots & \vdots & & \vdots\\ \Phi_0(x_N) & \Phi_1(x_N) & \Phi_2(x_N) & \cdots & \Phi_{M-1}(x_N)\end{bmatrix}.$$

The population BFR model is defined as follows:

$$\boldsymbol{Y}=\boldsymbol{\Phi\beta}+\boldsymbol{\epsilon}, \tag{7}$$

where $\boldsymbol{\epsilon}$ is the vector of random errors assumed to be distributed normally with mean vector 0 and variance-covariance matrix $\boldsymbol{\Sigma}$. Further, $f(x,\boldsymbol{\beta})=\boldsymbol{\Phi\beta}$ is the population regression function. The basis function $\Phi_j(X)$ is usually found in nonlinear functions in the input variable $x$, allowing the function $E_M(Y|\Phi,\beta)=\Phi\boldsymbol{\beta}$ to be nonlinear in $x$. However, the conditional mean is still linear in parameters $\boldsymbol{\beta}$. For prediction of the non-sampled values of the population parameter $\tau(y)$, the feature matrix $\boldsymbol{\Phi}$ can be partitioned as follows:

$$\boldsymbol{\Phi}=\begin{bmatrix}\boldsymbol{\Phi}_s\\ \boldsymbol{\Phi}_{\bar{s}}\end{bmatrix},$$

where $\boldsymbol{\Phi}_s$ and $\boldsymbol{\Phi}_{\bar{s}}$ are the sub-matrices of features with order $n\times M$ and $(N-n)\times M$, respectively.

**Theorem 1.** The quantity of interest $\tau(y)$ can be estimated using the general linear estimator proposed by Valliant et al. [8] with feature matrix $\boldsymbol{\Phi}$ as:

$$\hat{\tau}(y)=\boldsymbol{\gamma}_s^T\boldsymbol{Y}_s+\boldsymbol{\gamma}_{\bar{s}}^T\left[\boldsymbol{\Phi}_{\bar{s}}\hat{\boldsymbol{\beta}}+\boldsymbol{\Sigma}_{\bar{s}s}\boldsymbol{\Sigma}_{ss}^{-1}\left(\boldsymbol{y}_s-\boldsymbol{\Phi}_s\mathbf{e}\hat{\boldsymbol{t}}\boldsymbol{a}\right)\right], \tag{8}$$

where $\hat{\boldsymbol{\beta}}=(\boldsymbol{\Phi}_s^T\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{\Phi}_s)^{-1}\boldsymbol{\Phi}_s^T\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{y}_s$ is the WLS estimator of $\boldsymbol{\beta}$. The variance of $e(\hat{\tau})$ is given by:

$$\begin{aligned}V\left(e\left(\hat{\tau}\right)\right)&=\boldsymbol{\gamma}_{\bar{s}}^T\left(\boldsymbol{\Sigma}_{s\bar{s}}-\boldsymbol{\Sigma}_{\bar{s}s}\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{\Sigma}_{s\bar{s}}\right)\boldsymbol{\gamma}_{\bar{s}}\\&+\boldsymbol{\gamma}_{\bar{s}}^T\left(\boldsymbol{\Phi}_{\bar{s}}-\boldsymbol{\Sigma}_{\bar{s}s}\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{\Phi}_s\right)\left(\boldsymbol{\Phi}_s^T\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{\Phi}_s\right)^{-1}\\&\left(\boldsymbol{\Phi}_{\bar{s}}-\boldsymbol{\Sigma}_{\bar{s}s}\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{\Phi}_s\right)^T\boldsymbol{\gamma}_{\bar{s}},\end{aligned} \tag{9}$$

where $e(\hat{\tau})=\hat{\tau}(y)-\tau(y)$ is the prediction error.

***Proof.*** Derivation of Eqs. (8) and (9) can be found after replacing the feature matrix $\boldsymbol{\Phi}$ by the data matrix $\boldsymbol{X}$ in general prediction theorem given in [8, Chapter 2]. For simplicity, we assume that noninformative sampling (i.e., the sampled and non-sampled outcomes have the same distribution) conditional on values of the auxiliary variables results in $\boldsymbol{\Sigma}_{\bar{s}s}=0$ and the BLUP for $\tau(y)$ is reduced to:

$$\hat{\tau}(y)=\boldsymbol{\gamma}_s^T\boldsymbol{Y}_s+\boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\Phi}_{\bar{s}}\hat{\boldsymbol{\beta}}, \tag{10}$$

with prediction variance:

$$V(e\left(\hat{\tau}\right))=\boldsymbol{\gamma}_{\bar{s}}^T\left\{\boldsymbol{\Sigma}_{s\bar{s}}+\boldsymbol{\Phi}_{\bar{s}}\left(\boldsymbol{\Phi}_s^T\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{\Phi}_s\right)^{-1}\boldsymbol{\Phi}_{\bar{s}}^T\right\}\boldsymbol{\gamma}_{\bar{s}}. \tag{11}$$

Assuming iid noise in the data, i.e., $\boldsymbol{\Sigma}_{ss}=\sigma^2\boldsymbol{I}_n$ and $\boldsymbol{\Sigma}_{\bar{s}\bar{s}}=\sigma^2\boldsymbol{I}_{N-n}$, the resulting expression for variance of prediction error can be written as follows:

$$V\left(e\left(\hat{\tau}\right)\right)=\sigma^2\left[\boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\gamma}_{\bar{s}}+\boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\Phi}_{\bar{s}}\left(\boldsymbol{\Phi}_s^T\boldsymbol{\Phi}_s\right)^{-1}\boldsymbol{\Phi}_{\bar{s}}^T\boldsymbol{\gamma}_{\bar{s}}\right]. \tag{12}$$

For population total and mean, we set $\gamma_i=1$ and $\gamma_i=\frac{1}{N}$ for all $i\in\mathcal{U}$. We discuss some special cases of the proposed basis function model in the following subsections.

### *3.1. Special cases*

This subsection discusses some members of the BFR model and obtains estimators of total output using the specified models. Model Mean Squared Error (MSE) and bias are studied for the selected cases.

#### *3.1.1. Expansion estimator*

Consider a single constant BFR for estimating the finite population total as follows:

$$y_i=\beta_0+\epsilon_i \qquad \text{for} \quad i=1,2,\cdots,N. \tag{13}$$

The model given in Eq. (13) is known as Homogeneous Population Model (HPM) and obtained by taking $\boldsymbol{\Phi}$ as $N$ dimensional vector of 1's. The expansion estimator for $t_y=\sum_{i\in\mathcal{U}}y_i$ (population total) under HPM is obtained as:

$$\hat{t}_y^E=\sum_{i\in s}y_i+\sum_{i\in\bar{s}}\hat{\beta}_0, \tag{14}$$

where $\hat{\beta}_0=\frac{\sum_{i\in s}y_i}{n}$ is the BLUP for $\beta_0$ obtained based on the Ordinary Least Square (OLS) assumptions. The expansion estimator is unbiased when the underlying model is correct. The prediction error variance of the expansion estimator is given by:

$$V_M\left(\hat{t}_{ys}-t_y\right)=N^2\left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2, \tag{15}$$

which is equivalent to the designed-based variance of the total estimator under Simple Random Sampling Without Replacement (SRSWOR) [see, 1].

#### *3.1.2. Regression estimator*

A single variable linear BFR model with intercept is given by:

$$y_i=\beta_0+\sum_{j=1}^{M-1}\beta_j\Phi_j(x_i) \qquad \text{for} \quad i=1,2,\cdots,N. \tag{16}$$

The total estimator in the linear BFR model is obtained as follows:

$$\hat{t}_{y(reg)}=\sum_{i\in s}y_i+\sum_{i\in\bar{s}}\left\{\hat{\beta}_0+\sum_{j=1}^{M-1}\hat{\beta}_j\Phi_j(x_i)\right\},$$

where $\hat{\beta}_0=\bar{y}-\sum_{j=1}^{M-1}\hat{\beta}_j\bar{\Phi}_{js}$, $\bar{y}=\frac{1}{n}\sum_{i\in s}y_i$, and

$\bar{\Phi}_{js} = \frac{1}{n}\sum_{i\in s}\Phi_j(x_i)$. The total estimator using the BFR model can be simplified to:

$$\hat{t}_{y(reg)} = N\left[\bar{y} + \sum_{j=1}^{M-1}\hat{\beta}_j\left\{\bar{\Phi}_{jU} - \bar{\Phi}_{js}\right\}\right], \tag{17}$$

where $\bar{\Phi}_{jU} = \frac{1}{N}\sum_{i\in\mathcal{U}}\Phi_j(x_i)$ is the population mean corresponding to the $j$th basis function for the whole population. It is easy to prove that $\hat{t}_{y(reg)}$ is unbiased when the working model is a true representation of the underlying population model. On the contrary, if we use an incorrect model, the estimator may suffer misspecification bias. In the model without any basis function, i.e., $M = 1$, the resulting estimator of population total is $t_y = N\bar{y}_s$ with prediction bias $B_M(t_y) = N\sum_{j=1}^{M-1}\beta_j(\bar{\Phi}_{jU} - \bar{\Phi}_{js})$, which is of order $O(n^{-1})$. When the sample size increases, it goes to zero. If the chosen values of $x$'s provide larger mean values of the basis functions, then we get $\bar{\Phi}_{jU} < \bar{\Phi}_{js}$ and the bias $B_M(\hat{t}_{y(reg)})$ becomes negative, and vice versa. The bias is minimized by selecting a sample such that the difference on the right side of bias expression is minimum. In compliance with [8], we call such a sample a balanced sample. Exact balancing is achieved by selecting a sample for which $\bar{\Phi}_{jU} = \bar{\Phi}_{js}$. The prediction error variance for the estimator given in Eq. (17) is given as follows:

$$V_M\left(\hat{t}_{y(reg)} - t_y\right) = N^2\left[\sum_{j=1}^{M-1}\left(\bar{\Phi}_{jU} - \bar{\Phi}_{js}\right)^2 V_M\left(\hat{\beta}_j\right) + \left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2\right]. \tag{18}$$

Considering the case of a single basis function with intercept, i.e., $M = 2$, we have the following variance expression:

$$V_M\left(\hat{t}_{y(reg)} - t_y\right) = N^2\left[\frac{\left(\bar{\Phi}_{1U} - \bar{\Phi}_{1s}\right)^2}{\sum_{i\in s}\left(\Phi_1(x_i) - \bar{\Phi}_{1s}\right)^2} + \left(\frac{1}{n} - \frac{1}{N}\right)\right]\sigma^2. \tag{19}$$

This variance decreases when the mean of the basis function for the sampled and non-sampled units is in agreement and there is a high variation in the sampled values of the basis functions.

*3.1.3. Ratio estimator*

When the variance of the study variable depends on some function $\psi(x)$ of input variable(s), the least square estimator provides higher variance due to the problem of heteroscedasticity. In such situations, WLS method is preferred for estimating superpopulation parameters when the variance structure is known. We consider the following $(M-1)$ degree polynomial model with the basis function containing a single regressor with no intercept as follows:

$$y = f(x,\boldsymbol{\beta}) + \psi(x)\epsilon, \tag{20}$$

where $f(x,\boldsymbol{\beta}) = \sum_{j=1}^{M-1}\beta_j\Phi_j(x)$. The Gamma Population Model (GPM) discussed by Chambers and Clark [42] is obtained by setting $\psi(x) = x^{\gamma^*}$ and the well-known ratio estimator is obtained under GPM with $\gamma^* = \frac{1}{2}$. For $\gamma^* = 0$, we get a linear regression estimator with constant variance. To obtain homoscedastic error term, the WLS method is adopted to estimate the parameters involved in model Eq. (20):

$$y^* = \sum_{j=1}^{M-1}\beta_j\Phi_j^*(x) + \epsilon, \tag{21}$$

where $y^* = \frac{y}{\psi(x)}$ and $\Phi_j^*(x) = \frac{\Phi_j(x)}{\psi(x)}$. For $M = 2$, we have:

$$y^* = \beta_1\Phi_1^*(x) + \epsilon. \tag{22}$$

The BLUE for $\beta_1$ is obtained as $\hat{\beta}_1 = \frac{\sum_{i\in s}\Phi_1^*(x_i)y_i^*}{\sum_{i\in s}\Phi_1^{*2}(x_i)}$ with variance $V_M(\hat{\beta}_1) = \sigma^2\frac{\sum_{i\in s}\Phi_1^2(x_i)}{\left(\sum_{i\in s}\Phi_1^{*2}(x_i)\right)^2}$. The ratio estimator under single basis function is given by:

$$\hat{t}_{y(r)} = \sum_{i\in s}\left[1 + \lambda_i\sum_{i\in\bar{s}}\Phi_1^*(x_i)\right]y_i, \tag{23}$$

where $\lambda_i = \frac{\Phi_1^*(x_i)}{\psi(x_i)\sum_{i\in s}\Phi_1^{*2}(x_i)}$. The prediction error of $\hat{t}_{y(r)}$:

$$\hat{t}_{y(r)} - t_y = \sum_{i\in s}\lambda_i^* y_i - \sum_{i\in\bar{s}} y_i,$$

where $\lambda_i^* = \lambda_i\sum_{i\in\bar{s}}\Phi_1^*(x_i)$. The model bias and the prediction error variance are given by:

$$B_M\left(\hat{t}_{y(r)}\right) = \beta_1\left[\sum_{i\in s}\lambda_i^*\Phi_1(x_i) - \sum_{i\in\bar{s}}\Phi_1(x_i)\right], \tag{24}$$

and:

$$V_M\left(\hat{t}_{y(r)} - \tau(y)\right) = \left[\sum_{i\in s}\lambda_i^{*2}\psi^2(x_i) + \sum_{i\in\bar{s}}\psi^2(x_i)\right]\sigma^2. \tag{25}$$

The model mean squared prediction error is given by:

$$MSE_M\left(\hat{t}_{y(r)}\right) = \left[\sum_{i\in s}\lambda_i^{*2}\psi^2(x_i) + \sum_{i\in\bar{s}}\psi^2(x_i)\right]\sigma^2 + \beta_1^2\left[\sum_{i\in s}\lambda_i^*\Phi_1(x_i) - \sum_{i\in\bar{s}}\Phi_0(x_i)\right]^2. \tag{26}$$

The use of balance sampling with $\sum_{i\in s}\lambda_i^*\Phi_1(x_i)$

$-\sum_{i\in\bar{s}}\Phi_1(x_i)$ leads to unbiasedness, which is consistent with the estimator given in [43].

### *3.2. Some special basis functions*

Previously, we have discussed the prediction estimators using a general BFR model for estimating the values of finite population parameters. The next problem is to choose a reasonable function of the predictor or set of predictors for predicting non-sampled population values of the study variable for obtaining an estimate of $\tau(y)$. The real world is much complicated and we cannot easily adopt a linear model to capture a wide variety of so-called basis functions that we might need in prediction. To capture the complex phenomenon with nonlinear data, scientists have urged using a variety of basis functions that make a more precise prediction [44]. Some commonly used basis functions are briefly discussed in the following subsections.

#### *3.2.1. Polynomial basis functions*

Upon applying polynomial regression for predicting non-sampled values, it is essential to determine the degree of the polynomial before attending to the prediction problem. The problem of determining the degree of polynomial can be solve through visualization of the display of sample data. It is much tougher in case of three or more feature dimensions and it is a complete waste of time if there are interaction feature terms that affect the outcome. For a mutually interacting high-dimensional dataset, one may reach a wrong conclusion if we look at the output with one feature plot at a time. There is no simple way to visualize two or more variables at a time. In this way, we must adopt some ML techniques to fit a high-dimensional dataset, which is an open area for new developments. Consider a single variable basis function $f(x,\beta)=\sum_{l=0}^{M-1}\beta_l x^l$ with the corresponding feature matrix:

$$\mathbf{\Phi}=\begin{bmatrix}1 & x_1 & x_1^2 & \cdots & x_1^{M-1}\\ 1 & x_2 & x_2^2 & \cdots & x_2^{M-1}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_N & x_N^2 & \cdots & x_N^{M-1}\end{bmatrix}.$$

The polynomial used in the feature matrix is of order $M-1$. Determination of the degree of polynomial depends on the nature of relationship between the study variable $y$ and the auxiliary variable $x$. For $M=1$, we get the HPM and $M=2$ LRM with intercept; for $M=3$, we get Quadratic Regression Model (QRM). The polynomial BFRM provides global basis functions that affect the prediction over the whole range of inputs. The number of polynomials increases exponentially with increase in the value of $M$. The local basis functions are considered to be appropriate in prediction problems.

#### *3.2.2. Basis functions with two regressors*

Further, polynomial curve fitting is applicable only for single input variable $x$. It is not easy to generalize it for several input variables. For three input variables associated with the BFR model with $M=2$, we use separate indices for each variable as $J=(j_1,j_2,j_3)$ such that $(j_1+j_2+j_3)\leq(M-1)$ . The corresponding BFRM is:

$$\begin{aligned}y=\sum_{j_1,j_2,j_3}\beta_j\Phi_j(x)+\epsilon=\beta_0+\beta_{100}x_1+\beta_{010}x_2\\ +\beta_{001}x_3+\beta_{110}x_1x_2+\beta_{101}x_1x_3+\beta_{011}x_2x_3\\ +\beta_{200}x_1^2+\beta_{020}x_2^2+\beta_{002}x_3^2+\epsilon,\end{aligned}\tag{27}$$

where $\epsilon$ is random error term. For $p$ covariates, the number of quadratic terms is $[1+p+p(p-1)]/(2+p)$ in the above example and $p=3$; hence, the number of terms is 10. For $p$ input, a general case is $\Phi(x)=\{\prod_{k=1}^{p}x_k^{m_k}:\sum_{k=1}^{p}m_k\leq p\}$. The BFRM is appropriate for two or more variables when we know that there is an interaction effect of the two or more regressors on the output.

#### *3.2.3. Radial Basis Functions (RBF)*

RBF represent another type of real-valued basis functions whose values depend only on the distance from the origin, i.e., $\Phi(x)=\Phi(||x||)$. Alternatively, it may be based on the distance from another point called a center so that $\Phi(x,c)=\Phi(||x-c||)$. The concept of radial basis was initially introduced by Broomhead and Lowe [45], which had been inspired by the study in [46]. Lowe and Broomhead [47] discussed the relationship between "learning" in adaptive-layered networks and fitting of the data of high-dimensional surfaces. RBF is used as a kernel in classification of support vector [48]. Buhmann [49] provided theory and implementation of RBF. Later, Biancolini [50] extended its application to different fields of engineering and physics. RBFs are typically preferred for estimating population parameters when the auxiliary data consist of latitudes and longitudes in spatial studies. In general, we choose a family of basis functions in order to get a good fit to our training data with a small basis set, which consequently provides a moderate number of weights (coefficients) to be estimated. As is known, 'linearity' in the LRM indicates that the model is linear in coefficients rather than in features (or independent variables). Features can be of any degree or have transcendental functions like logarithmic, exponential and sinusoidal, etc. As a result, a surprisingly large number of natural phenomena can be modeled (through approximation) using the linear model with these transformations. Estimators for $\tau(y)$ can be obtained using alternate basis functions in Eq. (8).

## 4. Estimation under regularized regression

In regression analysis, overfitting shows that the de -

pendent variable corresponds exactly (or very close) to a particular dataset and is not able to fit additional data points. Such a condition is termed ill-conditioning in regression analysis. Initially, Tikhonov and Arsenin [51] worked on the mathematical aspect of the ill-posed problems and discussed the matter in their book. In addition to [51,52] suggested a ridge regression method for solving the ill-conditioned linear regression problem. Here, ill-conditioning refers to numerical difficulty in obtaining the inverse of the matrix, which is necessary for obtaining variance of estimators of the superpopulation parameters. The method presented in [52] is, in fact, a crude form of the ridge regression known as zero order regularization [53]. When Neural Network (NN) became popular in the 1980s, the weight decay was invented to deal with prune network connections, which are considered unimportant. Weight decay is soon recognized as an alternate of ridge regression in NN, given that it involves adding penalties to the cost function (sum-squared error). A variety of regularization methods are available in the literature and most of them were cited in [54]. In this section, we confined our discussion to the simple regularization method introduced by Press et al. [52], although our prediction problem can be handled using more advanced regularization methods, e.g., Least Absolute Shrinkage and Selection Operator (LASSO) [55], elastic net regression [56], and their extensions. Selection of a certain regularizer depends on bias and variance trade-off. Regularization reduces variance and increases bias simultaneously, leading to the MSE adjustment. If $E_M(\hat{\boldsymbol{\beta}}_{ridge}) = \boldsymbol{\beta}$ for all $\boldsymbol{\Phi}$'s, then the $\hat{\tau}_{ridge}(y)$ estimator will be unbiased. However, an unbiased estimator may still have larger MSE if the variance of the estimators of superpopulation parameters is higher. Such cases often occur when the regression function is highly sensitive to the choice of sample selection and noise of each training set. The sensitivity causes ill-conditioned regression estimates, as indicated in [51]. To significantly reduce high variation, [52] introduced a small amount of bias so that the net effect results in MSE reduction. Through regularization, one can reach the following cost function (sum-squared error):

$$C = (\boldsymbol{y}_s - \boldsymbol{\Phi}_s\boldsymbol{\beta})^T (\boldsymbol{y}_s - \boldsymbol{\Phi}_s\boldsymbol{\beta}) + v\boldsymbol{\beta}^T\boldsymbol{\beta}, \tag{28}$$

where the positive constant $v$ is called regularizer, which creates bias in the estimate of $\boldsymbol{\beta}$ and reduces variance. By optimizing the cost function given in Eq. (28), we reach the following ridge regression estimator for the coefficient vector $\boldsymbol{\beta}$ as follows:

$$\hat{\boldsymbol{\beta}}_{ridge} = \boldsymbol{Q}_s^{-1}\boldsymbol{\Phi}_s^T\boldsymbol{y}_s, \tag{29}$$

where the matrix $\boldsymbol{Q}_s = \boldsymbol{\Phi}_s^T\boldsymbol{\Phi}_s^T + v\boldsymbol{I}_n$ is symmetric, i.e., $\boldsymbol{Q}_s^T = \boldsymbol{Q}_s$. An estimator of population parameter $\tau(y)$ using the ridge regression model is given by:

$$\hat{\tau}_{ridge}(y) = \boldsymbol{\gamma}_s^T\boldsymbol{y}_s + \boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\Phi}_{\bar{s}}\hat{\boldsymbol{\beta}}_{ridge}, \tag{30}$$

which has model bias:

$$E_M\left(e\left(\hat{\tau}_{ridge}\right)\right) = \boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\Phi}_{\bar{s}}\left[E_M\left(\hat{\boldsymbol{\beta}}_{ridge}\right) - \boldsymbol{\beta}\right],$$

where $e(\hat{\tau}_{ridge}) = \hat{\tau}_{ridge}(y) - \tau(y)$. After some simplification (see, Eq. (54) in the Appendix), the bias is reduced to:

$$B_M\left(\hat{\tau}_{ridge}(y)\right) = -v\boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\Phi}_{\bar{s}}\boldsymbol{Q}_s^{-1}\boldsymbol{\beta}. \tag{31}$$

The bias given in Eq. (31) depends on the regularizer $v$. It can be concluded that the bias tends to decline as $v \to 0$ depending on entries in $\boldsymbol{Q}_s^{-1}$ (also depending on $v$). The variance expression is given by:

$$\begin{aligned} &V_M\left(e\left(\hat{\tau}_{ridge}\right)\right) \\ &\quad= \sigma^2\left[\boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\gamma}_{\bar{s}} + \boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\Phi}_{\bar{s}}\left(\boldsymbol{Q}_s^{-1} - \boldsymbol{Q}_s^{-2}\right)\boldsymbol{\Phi}_{\bar{s}}^T\boldsymbol{\gamma}_{\bar{s}}\right] \\ &\quad= V_M\left(e\left(\hat{\tau}\right)\right) - \sigma^2\boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\Phi}_{\bar{s}}\boldsymbol{Q}_s^{-2}\boldsymbol{\Phi}_{\bar{s}}^T\boldsymbol{\gamma}_{\bar{s}}. \end{aligned} \tag{32}$$

This shows that regularization reduces variance by an amount of $\sigma^2\boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\Phi}_{\bar{s}}\boldsymbol{Q}_s^{-2}\boldsymbol{\Phi}_{\bar{s}}^T\boldsymbol{\gamma}_{\bar{s}}$. This amount increases by increasing the parameter $v$, which ultimately raises the efficiency with a significant bias. The MSE of $\hat{\tau}_B(y)_{ridge}$ is then obtained using bias and variance relation and given by:

$$\begin{aligned} MSE_M\left\{\hat{\tau}_{ridge}(y)\right\} =& \sigma^2\left(\boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\gamma}_{\bar{s}}\right) + \boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\Phi}_{\bar{s}}\left[\sigma^2\left(\boldsymbol{Q}_s^{-1} - \boldsymbol{Q}_s^{-2}\right)\right. \\ &\left. + v^2\boldsymbol{Q}_s^{-1}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{Q}_s^{-1}\right]\boldsymbol{\Phi}_{\bar{s}}^T\boldsymbol{\gamma}_{\bar{s}}. \end{aligned} \tag{33}$$

The amount $\boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\Phi}_{\bar{s}}[v^2\boldsymbol{Q}_s^{-1}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{Q}_s^{-1} - \sigma^2\boldsymbol{Q}_s^{-2}]\boldsymbol{\Phi}_{\bar{s}}^T\boldsymbol{\gamma}_{\bar{s}}$ is the net effect on MSE. The regularization parameter $v$ provides a trade-off between over-fitting (which causes higher variance) and avoiding penalty (which causes increase in bias). Since the first derivative of the variance expression is nonlinear in $v$, it is not straightforward to obtain an explicit expression for $v$ which minimizes Eq. (33). Alternatively, one can adopt model selection criteria to obtain the optimum choice of $v$. Since all the criteria for model selection are nonlinear in $v$, we need some nonlinear optimization problems here. We can use any standard method for this purpose, such as the Newton method. The derivation of the optimum choice of $v$ is left for future study.

## 5. Variance estimation and comparison

After obtaining the prediction error, including bias and variance of the error, the next step is to search for an estimate of the error variance for further statistical analysis, e.g., testing statistical hypothesis about $\tau(y)$ and constructing confidence interval. Unlike the variance estimation methods in design-based paradigm such as Jackknife technique [57], in the model-based

approach, model selection criteria that indirectly provide an estimate of error variance $\sigma^2$ in the model-based approach are utilized. It can be seen that the variance of error term given in Eq. (12) depends on error variance $\sigma^2$ and the auxiliary data from the whole population. When the sub-matrix of the basis function for the non-sampled part is known, we need an estimate of $\sigma^2$ only for estimating the prediction error variance of $\tau(y)$. A sample estimate for the prediction error variances given in Eqs. (12) and (31) can be expressed as:

$$\hat{V}\left(e\left(\hat{\tau}\right)\right)=\hat{\sigma}^2\left\{\boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\gamma}_{\bar{s}}+\boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\Phi}_{\bar{s}}\left(\boldsymbol{\Phi}_s^T\boldsymbol{\Phi}_s\right)^{-1}\boldsymbol{\Phi}_{\bar{s}}^T\boldsymbol{\gamma}_{\bar{s}}\right\}, \quad (34)$$

and:

$$\hat{V}\left(e\left(\hat{\tau}_{ridge}\right)\right)=\hat{V}\left(e\left(\hat{\tau}_{ML}\right)\right)-\hat{\sigma}^2\boldsymbol{\gamma}_{\bar{s}}^T\boldsymbol{\Phi}_{\bar{s}}\boldsymbol{Q}_s^{-2}\boldsymbol{\Phi}_{\bar{s}}^T\boldsymbol{\gamma}_{\bar{s}}. \quad (35)$$

Estimate for $\sigma^2$ based on residuals is a routine practice in survey sampling. The estimate taken from the sampled observations or a part of observations (training set) provides a good measure for average noise in the study variable. We extend different methods for estimating the error variance in estimating finite population parameter $\tau(y)$. The projection matrix, say $\boldsymbol{P}$, plays a key role in obtaining the estimate for $\sigma^2$ using the mentioned methods. The projection matrix defined in Eq. (50) (see the Appendix) is symmetric and idempotent ($\boldsymbol{P^2}=\boldsymbol{P}$) when no regularization is applied. For obtaining estimates for $\sigma^2$, we use the following model selection criteria:

1. Residual method,
2. Cross Validation (CV),
3. Generalized Cross Validation (GCV),
4. Final Prediction Error (FPE) or Akaike's Information Criterion (AIC) 5-Bayesian Information Criterion (BIC).

### 5.1. Residual method or unbiased estimate of variance

Estimators for $\hat{\tau}(y)$ and $\hat{\tau}_{ridge}(y)$ of prediction error variance of the estimator $\hat{\tau}(y)$ are obtained using residual method after inserting $\hat{\sigma}^2_{res}$ in Eqs. (34) and (35), respectively, where:

$$\hat{\sigma}^2_{res}=\frac{1}{n-M}\boldsymbol{y}_s^T\boldsymbol{P}^2\boldsymbol{y}_s, \quad (36)$$

where $\boldsymbol{P}=I_M-\boldsymbol{\Phi}_s\boldsymbol{Q}_s^{-1}\boldsymbol{\Phi}_s^T$. Another widely used model selection criterion is the Unbiased Estimate of Variance (UEV), which is similar to residual variance and is obtained by replacing the total number of parameters by the number of effective parameters in the denominator. The UEV estimator of $\sigma^2$ is given by:

$$\hat{\sigma}^2_{UEV}=\frac{1}{n-M^*}\boldsymbol{y}_s^T\boldsymbol{P}^2\boldsymbol{y}_s, \quad (37)$$

where $M^*=n-trace(P)$ is the effective number of parameters in the model. However, the residual method is not considered as an appropriate measure for predictive power of the model [58]. The predictive power of the model here refers to how well the sampled data will perform in predicting unknown values of the output for the non-sampled part of the population.

### 5.2. Estimation via CV

A new variant of CV involves randomly splitting the data into a training set and a test set $k$ at distinct times. The benefit of doing so is that one can independently choose the size of each test set and number of trials for averaging. Leave-One-Out CV (LOOCV) is a special case of $k$-fold CV with its logical extreme, i.e., taking $k=n$ as the total number of data points. This means that the model is trained $n$ times including all the data except one point and predicting the outcome for that single point. The average prediction error is computed and applied to evaluate the model as an estimated noise. The prediction error variance of the estimator $\hat{\tau}(y)$ under LOO is obtained by replacing $\hat{\sigma}^2$ by $\hat{\sigma}^2_{LOO}$ in Eqs. (34) and (35), where:

$$\hat{\sigma}^2_{LOO}=\frac{1}{n}\boldsymbol{y}_s^T\boldsymbol{P}\{diag\left(\boldsymbol{P}\right)\}^{-2}\boldsymbol{P}\boldsymbol{y}_s. \quad (38)$$

The estimated variance obtained through LOO CV is a good estimate of model variance; however, at first glance, it seems too expensive and tiresome to compute. Luckily, locally weighted regression makes it easy to make regular predictions. It is implied that computing the LOO-XVE consumes no longer time than the residual error, which is the reason why it is preferred as the model selection criterion.

### 5.3. Estimation under GCV

The diagonal matrix $diag(\boldsymbol{P})$ makes LOO mathematically inappropriate. GCV, as its alternate, introduced by Golub et al. [59], is more convenient and is obtained by replacing the matrix $diag(\boldsymbol{P})$ by the average of the diagonal elements multiplied by the identity matrix of order $n$, i.e., $trace\left(\boldsymbol{P}/n\right)\boldsymbol{I}_n$. An estimator for the prediction error variance of $\hat{\tau}(y)$ under GCV is obtained by replacing $\hat{\sigma}^2$ by $\hat{\sigma}^2_{GCV}$ in Eqs. (34) and (35), where $\hat{\sigma}^2_{GCV}$ is defined by:

$$\hat{\sigma}^2_{GCV}=\frac{n\boldsymbol{y}_s^T\boldsymbol{P}^2\boldsymbol{y}_s}{\{trace\left(\boldsymbol{P}\right)\}^2}. \quad (39)$$

GCV is among one of the model selection criteria that includes an adjustment to the average of mean squared prediction error over the training set. It is equivalent to standard residual method given in Eq. (38), $\frac{n}{\{trace\left(\boldsymbol{P}\right)\}^2}=\frac{1}{n-M^*}$, where $M^*=n-sum(diag(P))$ is the effective number of parameters in the model. GCV can be expressed in terms of the effective number of parameters $M^*$ instead of $trace\left(\boldsymbol{P}\right)$ as follows:

$$\hat{\sigma}^2_{GCV} = \frac{n\boldsymbol{y}_s^T\boldsymbol{P}^2\boldsymbol{y}_s}{(n-M^*)^2}. \quad (40)$$

### 5.4. Estimation under FPE

Mallow's $C_p$ [60], named after Colin Lingwood Mallows, is a statistic that assesses the fit of a regression model which is estimated via OLS. This statistic is used in the context of model selection when a number of predictors are available for predicting the outcome, aiming to find the best subset of available predictors. In the case of the Gaussian LRM, Mallows's $C_p$ is equivalent to AIC, a most widely used model evaluation criterion [61], and is used as an alternate of AIC. An estimator for prediction error variance of $\hat{\tau}(y)$ using FPE method is obtained by replacing $\hat{\sigma}^2$ by $\hat{\sigma}^2_{FPE}$ in Eq. (34), where $\hat{\sigma}^2_{FPE}$ is the alternative version of Mallows's $C_p$ [62] and is given by:

$$\hat{\sigma}^2_{FPE} = \frac{1}{n}\left(\boldsymbol{y}_s^T\boldsymbol{P}^2\boldsymbol{y}_s + 2M^*\hat{\sigma}^2_{res}\right) = \frac{n+M^*}{n-M^*}\frac{\boldsymbol{y}_s^T\boldsymbol{P}^2\boldsymbol{y}_s}{n}, \quad (41)$$

where $M^*$ is the effective number of parameters. $\hat{\sigma}^2_{FPE}$ is subject to two limitations:

1. The approximation is valid only for a sufficiently large sample size;
2. It cannot deal with a complex set of models as in the variable selection (feature selection in ML) problems [63].

### 5.5. Estimation under BIC

The BIC developed by Schwarz et al. [64] is a Bayesian argument on the maximum data likelihood. It is related to the AIC and later, Akaike developed his own Bayesian formalism in inspiration from the motive of Schwarz, now mostly referred to as "Akaike's Bayesian Information Criterion" (ABIC) instead of "Bayesian Information Criterion" (BIC) [65]. An estimator of variance of prediction error for $\hat{\tau}(y)$ based on BIC is found after inserting $\hat{\sigma}^2_{BIC}$ by $\hat{\sigma}^2$ in Eqs. (34) and (35), where $\hat{\sigma}^2_{BIC}$ is:

$$\begin{aligned}\hat{\sigma}^2_{BIC} &= \frac{1}{n}\left(\boldsymbol{y}_s^T\boldsymbol{P}^2\boldsymbol{y}_s + \ln(n)M^*\hat{\sigma}^2_{res}\right)\\ &= \frac{n+M^*(\ln(n)-1)}{n-M^*}\frac{\boldsymbol{y}_s^T\boldsymbol{P}^2\boldsymbol{y}_s}{n}, \end{aligned} \quad (42)$$

where $\ln(n)$ is the natural logarithm of $n$. Here, $\hat{\sigma}^2_{BIC}$ measures the unexplained variation in the output variable and the increased number of explanatory variables.

All the mentioned estimators of the prediction error variance can be used for statistical analysis about the finite population parameter $\tau(y)$. To compare the above discussed rival estimators of prediction variance, we write all the variance estimators in the form of $\sigma^2_{abc} = \Gamma_{abc}\boldsymbol{y}_s^T\boldsymbol{P}^2\boldsymbol{y}_s/n$ and have the following natural ordering as follows:

$$\Gamma_{UEV} \le \Gamma_{FPE} \le \Gamma_{GCV} \le \Gamma_{BIC}. \quad (43)$$

The factors $\Gamma$'s are approximated using Taylor's series as:

$$\Gamma_{res} = \frac{n}{n-M^*} = 1 + \frac{M^*}{n} + \frac{M^{*2}}{n} + \frac{M^{*3}}{n} + \cdots$$

$$\Gamma_{FPE} = \frac{n+M^*}{n-M^*} = 1 + \frac{2M^*}{n} + \frac{2M^{*2}}{n^2} + \frac{2M^{*3}}{n^3} + \cdots$$

$$\Gamma_{GCV} = \frac{M^{*2}}{(n-M^*)^2} = 1 + \frac{2M^*}{n} + \frac{3M^{*2}}{n^2} + \frac{4M^{*3}}{n^3} + \cdots$$

$$\begin{aligned}\Gamma_{BIC} &= \frac{n+(\ln(n)-1)M^*}{n-M^*}\\ &= 1 + \ln(n)\left(\frac{M^*}{n} + \frac{M^{*2}}{n} + \frac{M^{*3}}{n} + \cdots\right),\end{aligned}$$

where the subscript "res" denotes the error variance obtained by the residual method. Hence, the estimators of $\hat{\sigma}^2$ obtained by different model selection criteria can be ranked according to the factor $\Gamma$. Hence, variance estimators can be ranked as follows:

$$\begin{aligned}\hat{V}\left(e\left(\hat{\tau}\right)\right)_{UEV} &\le \hat{V}\left(e\left(\hat{\tau}\right)\right)_{FPE} \le \hat{V}\left(e\left(\hat{\tau}\right)\right)_{GCV}\\ &\le \hat{V}\left(e\left(\hat{\tau}\right)\right)_{BIC}, \end{aligned} \quad (44)$$

where the subscripts attached to the estimated variances show the model selection criteria used for estimating $\sigma^2$.

## 6. Model selection

We previously discussed ridge regression (Section 4) as a tool for controlling the trade-off between the bias and variance (Section 5) of the estimators of superpopulation parameters such as $\sigma^2$. Alternatively, one can compare models with different subsets of basis functions selected from a fixed set of candidate models, known as "subset selection" [66]. It is difficult to find the best set among the $2^M - 1$ alternative subsets (each of size $M$) for the purpose of response prediction. To search for an interesting small fraction of all subsets, we need heuristics. Forward selection and backward selection methods are two widely used heuristics for model selection. Although backward selection is a commonly method used for factor screening in multiple regression analysis, it does not seem logical to start with a set including higher order basis functions and, then, come to an effective smaller subset in finite population parameter estimation. On the other end, the forward selection method begins with a null subset and goes by adding one basis function at a time. The process of forward selection stops at the subset which provides minimum sum of squared prediction error.

Although forward selection is a nonlinear algorithm, it still has the following plus points:

(i) The number of hidden units does not need to be fixed in advance;

(ii) It has a tractable criterion for model selection;

(iii) It needs relatively low computational effort.

In forward selection, the model grows at each step by one basis function. To investigate the effect of increasing a new basis function, we introduce some incremental operators (see, the Appendix). We see the effect of adding a new basis function on the bias and variance of the estimator $\hat{\tau}(y)$ in the following subsections.

### 6.1. Model selection under OLS

Reduction in variance on using additional basis function can be computed as follows:

$$IE = V_M\left(\hat{\tau}\right)_m - V_M\left(\hat{\tau}\right)_{m+1} = \sigma^2 \boldsymbol{\gamma}_{\bar{s}}^T \left[\boldsymbol{\Phi}_{\bar{s}m} \boldsymbol{A}_{sm}^{-1} \boldsymbol{\Phi}_{\bar{s}m}^T - \boldsymbol{\Phi}_{\bar{s}(m+1)} \boldsymbol{A}_{s(m+1)}^{-1} \boldsymbol{\Phi}_{\bar{s}(m+1)}^T\right] \boldsymbol{\gamma}_{\bar{s}}.$$

The subscripts $m$ and $(m+1)$ indicate that the quantities are obtained with $M$ and $(M+1)$ basis functions, respectively. The subscripts $s$ and $\bar{s}$ are used to denote the quantities corresponding to the sampled and non-sampled parts, respectively. In the Appendix, using Eq. (53), we get:

$$\begin{aligned} IE =& V_M\left(\hat{\tau}\right)_m - V_M\left(\hat{\tau}\right)_{m+1} \\ =& \sigma^2 \boldsymbol{\gamma}_{\bar{s}}^T \left[\boldsymbol{\Phi}_{\bar{s}m} \boldsymbol{A}_{sm}^{-1} \boldsymbol{\Phi}_{\bar{s}m}^T - \boldsymbol{\Phi}_{\bar{s}(m+1)} \boldsymbol{A}_{s(m+1)}^{-1} \boldsymbol{\Phi}_{\bar{s}(m+1)}^T\right] \boldsymbol{\gamma}_{\bar{s}}, \end{aligned} \tag{45}$$

$$\begin{aligned} IE =& \frac{1}{\triangle} \sigma^2 \boldsymbol{\gamma}_{\bar{s}}^T \left[\boldsymbol{\phi}_{\bar{s}(m+1)} \boldsymbol{\phi}_{s(m+1)}^T \boldsymbol{\Phi}_{sm} \boldsymbol{A}_{sm}^{-1} \boldsymbol{\Phi}_{\bar{s}m}^T \right. \\ &+ \boldsymbol{\Phi}_{\bar{s}m}^T \boldsymbol{A}_{sm}^{-1} \boldsymbol{\Phi}_{sm}^T \boldsymbol{\phi}_{s(m+1)} \boldsymbol{\phi}_{\bar{s}(m+1)}^T \\ &- \boldsymbol{\Phi}_{\bar{s}m} \boldsymbol{A}_{sm}^{-1} \boldsymbol{\Phi}_{sm}^T \boldsymbol{\phi}_{s(m+1)} \boldsymbol{\phi}_{s(m+1)}^T \boldsymbol{\Phi}_{sm} \boldsymbol{A}_{sm}^{-1} \boldsymbol{\Phi}_{\bar{s}m}^T \\ &\left. - \boldsymbol{\phi}_{\bar{s}(m+1)} \boldsymbol{\phi}_{\bar{s}(m+1)}^T \right] \boldsymbol{\gamma}_{\bar{s}}, \end{aligned} \tag{46}$$

where the vector $\boldsymbol{\phi}_{\bar{s}(m+1)}$ shows the $(M+1)$th column of the basis function matrix $\boldsymbol{\Phi}_{s(m+1)}$. The positive increase in efficiency, i.e., $IE > 0$, indicates that using an additional basis function reduces the variance of prediction error. This can be converted into a ratio by Eq. (47) as shown in Box I. The index $IE_R$ measures the relative increase in efficiency of using an additional predictor to our model. The $IE$ can only be seen when we know the variance of the response in advance. In many real applications, we do not have a known value of the variance of the study variable in advance. Then, different estimates obtained in Section 5 are used. Since the estimates involve the basis function matrix, the use of additional basis function affects the estimated variances. One option is to obtain an estimate of variance of $e(\hat{\tau})$ through re-estimation of the regression, which is a difficult task in model selection. Second, we can jointly compute the estimated prediction variance of $\hat{\tau}(y)$ instead of its population counterpart. The third option is to use Eq. (47) and separately obtain the estimates of $\sigma^2$ through incremental operators given in [67]. Although the third option provides a variance expression for the model with $(M+1)$ predictors without recomputing the regression, it does not make a comparison of the two models, namely the model with $M$ basis functions and the one with $(M+1)$ basis functions.

### 6.2. Model selection under regularized regression

When regularization is used for estimating superpopulation parameter $(\beta)$ and then, the finite population parameters are estimated with bias. We first see the change in bias in using the additional basis function in the model as follows:

$$\begin{aligned} \left|B_M\left(\hat{\tau}_{ridge}(y)\right)_m - B_M\left(\hat{\tau}_{ridge}(y)\right)_{m+1}\right| =& v\boldsymbol{\gamma}_{\bar{s}}^T \\ &\left[\frac{1}{\triangle} \boldsymbol{\Phi}_{\bar{s}m} \boldsymbol{Q}_{sm}^{-1} \boldsymbol{\Phi}_{sm} \boldsymbol{\phi}_{s(m+1)} \boldsymbol{\phi}_{s(m+1)} \boldsymbol{\Phi}_{sm} \boldsymbol{Q}_{sm}^{-1} \boldsymbol{\beta}_m \right. \\ &+ \boldsymbol{\phi}_{\bar{s}(m+1)} \boldsymbol{Q}_{21}^{-1} \boldsymbol{\beta}_m + \boldsymbol{\Phi}_{\bar{s}m} \boldsymbol{Q}_{12}^{-1} \beta_{m+1} \\ &\left. + \boldsymbol{\phi}_{\bar{s}(m+1)} \boldsymbol{Q}_{22}^{-1} \beta_{m+1}\right], \end{aligned} \tag{48}$$

where $\beta_{m+1}$ is the $(M+1)$th component of the vector $\boldsymbol{\beta}_{m+1}$, i.e., the effect of additional basis function on the response. A smaller amount of increase in bias implies that the additional variable does not affect the bias of

$$IE_R = \frac{\boldsymbol{\gamma}_{\bar{s}}^T \left[\boldsymbol{\phi}_{\bar{s}(m+1)} \boldsymbol{\phi}_{s(m+1)}^T \boldsymbol{\Phi}_{sm} \boldsymbol{A}_{sm}^{-1} \boldsymbol{\Phi}_{\bar{s}m}^T + \boldsymbol{\Phi}_{\bar{s}m}^T \boldsymbol{A}_{sm}^{-1} \boldsymbol{\Phi}_{sm}^T \boldsymbol{\phi}_{s(m+1)} \boldsymbol{\phi}_{\bar{s}(m+1)}^T\right] \boldsymbol{\gamma}_{\bar{s}}}{\boldsymbol{\gamma}_{\bar{s}}^T \left[\boldsymbol{\Phi}_{\bar{s}m} \boldsymbol{A}_{sm}^{-1} \boldsymbol{\Phi}_{sm}^T \boldsymbol{\phi}_{s(m+1)} \boldsymbol{\phi}_{s(m+1)}^T \boldsymbol{\Phi}_{sm} \boldsymbol{A}_{sm}^{-1} \boldsymbol{\Phi}_{\bar{s}m}^T + \boldsymbol{\phi}_{\bar{s}(m+1)} \boldsymbol{\phi}_{\bar{s}(m+1)}^T\right] \boldsymbol{\gamma}_{\bar{s}}}. \tag{47}$$

Box I

the estimator for a particular value of the ridge parameter. When different cases of regularizations are used for each superpopulation parameter, the amount of increase cannot be computed with this formula. Now, the increase in the efficiency of the ridge regression estimator in using additional basis function is expressed as follows:

$$\begin{aligned} IE_{ridge} =& V_M\left(e\left(\hat{\tau}_{ridge}\right)\right)_m - V_M\left(e\left(\hat{\tau}_{ridge}\right)\right)_{m+1} \\ =& IE - \sigma^2 \boldsymbol{\gamma}_{\bar{s}}^T \left[\boldsymbol{\Phi}_{\bar{s}m} \boldsymbol{Q}_{sm}^{-2} \boldsymbol{\Phi}_{\bar{s}m}^T \right. \\ & \left. - \boldsymbol{\Phi}_{\bar{s}(m+1)} \boldsymbol{Q}_{s(m+1)}^{-2} \boldsymbol{\Phi}_{\bar{s}(m+1)}^T \right] \boldsymbol{\gamma}_{\bar{s}}, \end{aligned} \tag{49}$$

where:

$$\begin{aligned} & \sigma^2 \boldsymbol{\gamma}_{\bar{s}}^T \left[\boldsymbol{\Phi}_{\bar{s}m} \boldsymbol{Q}_{sm}^{-2} \boldsymbol{\Phi}_{\bar{s}m}^T - \boldsymbol{\Phi}_{\bar{s}(m+1)} \boldsymbol{Q}_{s(m+1)}^{-2} \boldsymbol{\Phi}_{\bar{s}(m+1)}^T \right] \boldsymbol{\gamma}_{\bar{s}} \\ & = -\Big( 2\boldsymbol{\Phi}_{\bar{s}m} \boldsymbol{Q}_{sm}^{-1} \triangle^* \boldsymbol{\Phi}_{\bar{s}m}^T + \boldsymbol{\Phi}_{\bar{s}m} \triangle^* \triangle^{*T} \boldsymbol{\Phi}_{\bar{s}m}^T \\ & \quad + \boldsymbol{\phi}_{\bar{s}(m+1)} \boldsymbol{Q}_{21}^{-2} \boldsymbol{\Phi}_{\bar{s}m}^T + \boldsymbol{\Phi}_{\bar{s}m} \boldsymbol{Q}_{12}^{-1} \boldsymbol{\phi}_{\bar{s}(m+1)}^T \\ & \quad + \boldsymbol{\phi}_{\bar{s}(m+1)} \boldsymbol{Q}_{22}^{-2} \boldsymbol{\phi}_{\bar{s}(m+1)}^T \Big), \end{aligned}$$

and:

$$\triangle^* = \triangle^{-1} \boldsymbol{Q}_{sm}^{-1} \boldsymbol{\Phi}_{sm}^T \boldsymbol{\phi}_{s(m+1)} \boldsymbol{\phi}_{s(m+1)}^T \boldsymbol{\Phi}_{sm} \boldsymbol{Q}_{sm}^{-1}.$$

The sub-matrices $\boldsymbol{Q}_{21}^{-2}$, $\boldsymbol{Q}_{12}^{-2}$, and $\boldsymbol{Q}_{22}^{-2}$ are the elements of the matrix $\boldsymbol{Q}_{s(m+1)}^{-2} = \boldsymbol{Q}_{s(m+1)}^{-1} \boldsymbol{Q}_{s(m+1)}^{-1}$ and are defined in the Appendix. Computation of $IE_{ridge}$ is not straightforward, and some algebraic treatment on matrices can still provide a compact form that can be solved numerically. The positive value of the index $IE_{ridge}$ provides evidence of efficiency improvement by adding an additional basis function to the superpopulation model.

## 7. Simulations

Two simulation studies, namely one simulated and one bootstrapped, are conducted to evaluate the error variance of the proposed estimator of $\tau(y)$ ($\gamma_i = 1$ for all $i \in \mathcal{U}$) to find design expected values of the estimated error variance of $\hat{\tau}(y)$. For this purpose, we provide a simulation study using artificially generated population and fitting basis functions (we limit our discussion to polynomial basis function to avoid complexity). The bootstrap study includes a real data set to perform repeated sampling to obtain design-based properties of the estimator and estimate the error variance of the estimator ($\hat{\tau}(y)$). Both Monte Carlo (MC) simulation and bootstrapping are performed in the widely used statistical software R (version 4.0.1). The simulation steps are described below:

(i) To constitute a population exhibiting nonlinear behavior, draw two independent vectors $u^*$ and $v^*$ of length $N = 1000$, each with uniform (0,1). The variables $x$ and $e$ are obtained as the quantile points corresponding to the cumulative probabilities $u^*$ and $v^*$ with normal $N(10, 10)$ and $N(0, 10)$, respectively. We generate the vector of the study variable $y$ as $y = \sin(2\pi x) + e$. Note that for obtaining design-based properties, we generate these variables only once and consider them as a fixed finite population (after observing population characteristics such as mean and variance), while for model-based properties, we need to generate the data repeatedly. We focus on design-bias and design-expected prediction error to see the behavior of the proposed estimator $\hat{\tau}(y)$;

(ii) For fixed $n$, we split data *df* ($y$,$x$,$x^2$,$x^3$,$\cdots$,$x^{M-1}$) (where $M$ is the number of basis functions and *df* denotes data frame) into sampled and non-sampled parts with sizes $n$ and $N - n$ randomly. From sampled data, we estimate superpopulation parameters ($\boldsymbol{\beta}$ and $\sigma^2$). The estimated values of $\sigma^2$ are obtained using different formulas discussed in Section 5;

(iii) Further, we evaluate the proposed estimator of $\tau(y)$ (with $\gamma_i = 1$ for all $i \in \mathcal{U}$) and the estimated variance of $\hat{\tau}(y)$ using different formulas given in Section 5;

(iv) Repeat Steps (ii) and (iii) 30,000 times to obtain design-expected prediction error (i.e., bias) and design-expected squared prediction error of the proposed estimator of $\tau(y)$ and expected values of the estimated variance of $\hat{\tau}(y)$ for different choices of $n$, $M$, and $v$ (for ridge regression).

For bootstrapping, we consider first $N = 203$ hospitals from the hospital data given in [8, Appendix B, Page 424] as our population. The number of beds ($x$) at each hospital is taken as the predictor for the number of patients discharged ($y$). Repeated sampling, as early mentioned for hypothetical population, is performed to study the properties of total estimators and estimated error variance. Expected Squared Prediction Error (ESPE) is obtained as follows:

$$ESPE = \frac{1}{30,000} \sum_{sim} \left\{e\left(\hat{\tau}_{ridge}\right)\right\}^2,$$

where the symbol $\sum_{sim}$ indicates that summation is taken over all 30,000 simulated samples. The ESPE is determined via regularization upon considering $v = 0, 1, 5, 10$ with $v = 0$, representing the case of no regularization. Further, the design expectation of the estimated variance of $\hat{\tau}$ is obtained by the respective formulae after averaging over all the selected samples. We use polynomial basis functions of different orders with intercept (the linear population model) and without intercept (the proportional population model).

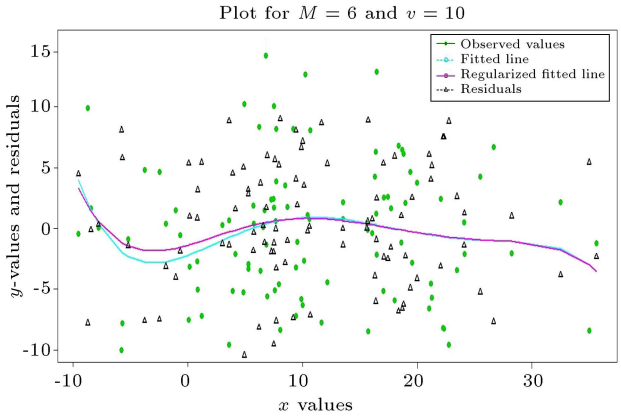


**Figure 1.** Scatter plot for the sample selected from hypothetical population plotting the observed values, fitted values (for simple and penalized regression with polynomial of 5th order, i.e., $M = 6$) and the residuals (for both cases) versus the values of predictor ($x$).

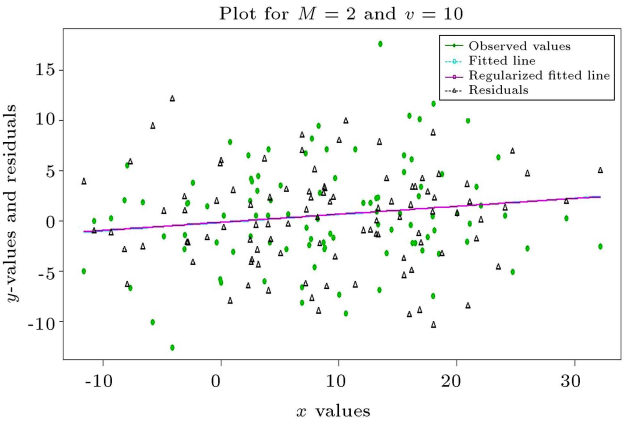


**Figure 2.** Scatter plot for the sample selected from hypothetical population plotting the observed values, fitted values (for simple and penalized regression with $M = 2$), and the residuals (for both cases) versus the values of predictor ($x$).

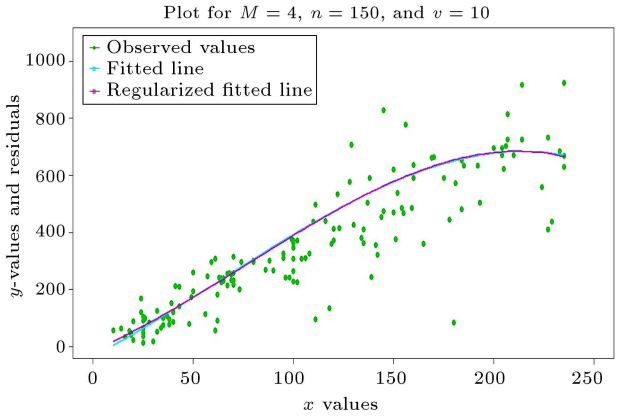


**Figure 3.** Scatter plot for the sample selected from the hospital data plotting the observed values, fitted values (for simple and penalized regression with $M = 4$), and the residuals (for both cases) versus the values of predictor ($x$).

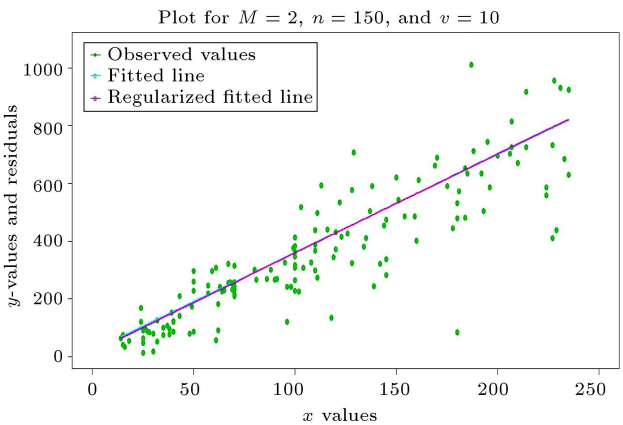


**Figure 4.** Scatter plot for the sample selected from the hospital data plotting the observed values, fitted values (for simple and penalized regression with $M = 2$), and the residuals (for both cases) versus the values of predictor ($x$).

To observe the behavior of fitted models (under both simple and ridge regression), scatter plots with fitted lines and residuals are constructed for both data sets used in this study. The residual values are plotted to evaluate the pattern in the study variable. For both data sets, no significant pattern of heteroscedasticity is observed. Scatter plots between $x$ with observed values of $y$ and fitted values $y$ are shown in Figures 1–4. The scatter plots provide a quick picture about the relationships between the outcome and predictor, which is necessary in choosing an appropriate model.

Tables 1 and 2 provide the design-based behavior of the prediction error of $\hat{\tau}(y)$ for the hypothetically generated population for the models of certain orders with and without intercept, respectively. For simulated data, the results are obtained for HPM ($M = 1$), linear population model ($M = 2$), the quadratic model ($M = 3$), and the higher order polynomial model ($M = 6$). The values of ESPE and expected estimated variances in Tables 1 and 2 are presented after dividing by $10^3$. While Tables 3 and 4 provide the design-based behavior of the prediction error of $\hat{\tau}(y)$ for the real population (hospitals data) for the models of certain orders with and without intercept. The values of ESPE and expected estimated variances in Tables 3 and 4 are reported after dividing by $10^5$ for simplicity. For real data, results are obtained for HPM ($M = 1$), LPM ($M = 2$), quadratic model ($M = 3$), and higher order polynomial (cubic) model ($M = 4$). The estimated error variances of $\hat{\tau}(y)$ are obtained using Eq. (35) based on different estimators of $\sigma^2$. Note that all the results given in Tables 1–4 are provided for a ridge regression estimator with certain choices of $v$ as the variance estimator given in $n$ Eq. (34) is a special case of variance estimator in Eq. (35) with $v = 0$. The ESPE for different combinations of $M$, $v$, and $n$ is enlisted in the third column of Tables 1 and 2. For simulated data in Table 1, the smallest ESPE is observed at $M = 6$

**Table 1.** Simulated ESPE and expected estimated variances of the proposed estimator under linear BFR model.

| | | | Expected estimated variances | | | | |
|---|---|---|---|---|---|---|---|
| $n$ | $v$ | ESPE | Residual | UEV | FPE | GCV | BIC |
| | | | | $M = 1$ | | | |
| 100 | 0 | 353.488 | 231.447 | 233.785 | 236.123 | 236.146 | 242.213 |
| | 1 | 347.545 | 227.343 | 229.616 | 231.890 | 231.912 | 237.812 |
| | 5 | 325.365 | 212.085 | 214.125 | 216.164 | 216.184 | 221.477 |
| | 10 | 300.810 | 195.310 | 197.102 | 198.894 | 198.910 | 203.562 |
| 200 | 0 | 53.696 | 103.360 | 103.879 | 104.399 | 104.401 | 106.112 |
| | 1 | 53.327 | 102.539 | 103.052 | 103.565 | 103.567 | 105.256 |
| | 5 | 51.899 | 99.376 | 99.863 | 100.350 | 100.353 | 101.957 |
| | 10 | 50.215 | 95.673 | 96.131 | 96.589 | 96.591 | 98.099 |
| | | | | $M = 2$ | | | |
| 100 | 0 | 491.880 | 231.843 | 236.574 | 241.306 | 241.402 | 253.632 |
| | 1 | 493.513 | 227.703 | 232.301 | 236.898 | 236.991 | 248.876 |
| | 5 | 499.478 | 213.512 | 217.656 | 221.800 | 221.880 | 232.595 |
| | 10 | 505.863 | 199.860 | 203.575 | 207.290 | 207.359 | 216.968 |
| 200 | 0 | 59.773 | 103.477 | 104.522 | 105.568 | 105.578 | 109.015 |
| | 1 | 59.830 | 102.653 | 103.684 | 104.715 | 104.726 | 108.117 |
| | 5 | 60.045 | 99.600 | 100.580 | 101.561 | 101.571 | 104.795 |
| | 10 | 60.291 | 96.267 | 97.193 | 98.119 | 98.128 | 101.172 |
| | | | | $M = 3$ | | | |
| 100 | 0 | 408.216 | 232.991 | 240.197 | 247.403 | 247.626 | 266.176 |
| | 1 | 408.907 | 228.750 | 235.772 | 242.793 | 243.009 | 261.086 |
| | 5 | 411.410 | 214.308 | 220.708 | 227.108 | 227.299 | 243.781 |
| | 10 | 414.043 | 200.553 | 206.368 | 212.183 | 212.352 | 227.332 |
| 200 | 0 | 50.033 | 103.639 | 105.217 | 106.795 | 106.820 | 112.001 |
| | 1 | 50.009 | 102.806 | 104.366 | 105.926 | 105.950 | 111.071 |
| | 5 | 49.917 | 99.728 | 101.220 | 102.712 | 102.735 | 107.634 |
| | 10 | 49.809 | 96.375 | 97.794 | 99.213 | 99.234 | 103.894 |
| | | | | $M = 6$ | | | |
| 100 | 0 | 324.940 | 290.500 | 309.043 | 327.586 | 328.769 | 375.892 |
| | 1 | 327.414 | 287.212 | 305.406 | 323.600 | 324.753 | 370.999 |
| | 5 | 335.748 | 277.340 | 294.468 | 311.596 | 312.653 | 356.217 |
| | 10 | 343.560 | 269.538 | 285.789 | 302.039 | 303.019 | 344.375 |
| 200 | 0 | 90.969 | 107.627 | 110.955 | 114.284 | 114.387 | 125.263 |
| | 1 | 103.270 | 101.953 | 105.017 | 108.080 | 108.172 | 118.184 |
| | 5 | 97.617 | 104.423 | 107.603 | 110.783 | 110.880 | 121.271 |
| | 10 | 103.270 | 101.953 | 105.017 | 108.080 | 108.172 | 118.184 |

**Table 2.** Simulated ESPE and expected estimated variances of the proposed estimator under proportional BFR model.

| $n$ | $v$ | ESPE | Expected estimated variances | | | | |
|---|---|---|---|---|---|---|---|
| | | | Residual | UEV | FPE | GCV | BIC |
| | | | | $M = 2$ | | | |
| 100 | 0 | 269.879 | 125.303 | 126.568 | 127.834 | 127.847 | 131.131 |
| | 1 | 269.855 | 125.292 | 126.558 | 127.823 | 127.836 | 131.120 |
| | 5 | 269.757 | 125.250 | 126.515 | 127.780 | 127.793 | 131.075 |
| | 10 | 269.635 | 125.198 | 126.462 | 127.726 | 127.738 | 131.019 |
| 200 | 0 | 9.988 | 61.071 | 61.378 | 61.685 | 61.687 | 62.697 |
| | 1 | 9.987 | 61.069 | 61.376 | 61.683 | 61.684 | 62.695 |
| | 5 | 9.985 | 61.061 | 61.368 | 61.675 | 61.676 | 62.687 |
| | 10 | 9.982 | 61.051 | 61.358 | 61.664 | 61.666 | 62.676 |
| | | | | $M = 4$ | | | |
| 100 | 0 | 268.060 | 127.606 | 130.210 | 132.814 | 132.868 | 139.599 |
| | 1 | 268.021 | 127.597 | 130.201 | 132.805 | 132.858 | 139.588 |
| | 5 | 267.866 | 127.562 | 130.164 | 132.766 | 132.819 | 139.544 |
| | 10 | 267.673 | 127.519 | 130.118 | 132.718 | 132.771 | 139.489 |
| 200 | 0 | 8.180 | 61.481 | 62.102 | 62.723 | 62.729 | 64.772 |
| | 1 | 8.179 | 61.479 | 62.100 | 62.721 | 62.728 | 64.770 |
| | 5 | 8.177 | 61.473 | 62.093 | 62.714 | 62.720 | 64.762 |
| | 10 | 8.175 | 61.464 | 62.085 | 62.705 | 62.711 | 64.752 |
| | | | | $M = 6$ | | | |
| 100 | 0 | 403.569 | 161.764 | 166.767 | 171.770 | 171.925 | 184.804 |
| | 1 | 403.568 | 161.757 | 166.759 | 171.762 | 171.916 | 184.793 |
| | 5 | 403.565 | 161.729 | 166.728 | 171.727 | 171.882 | 184.751 |
| | 10 | 403.561 | 161.693 | 166.689 | 171.684 | 171.839 | 184.699 |
| 200 | 0 | 29.688 | 74.447 | 75.581 | 76.714 | 76.732 | 80.454 |
| | 1 | 29.687 | 74.446 | 75.579 | 76.713 | 76.730 | 80.452 |
| | 5 | 29.685 | 74.440 | 75.574 | 76.707 | 76.724 | 80.445 |
| | 10 | 29.681 | 74.434 | 75.567 | 76.700 | 76.717 | 80.437 |

when the sample size is taken 100. It is the smallest at $M = 3$ when sample size is set to 200. ESPE for the simulated data also tends to decrease with increase in $v$. For example, in the case of $n = 200$ and $M = 1$, the ESPE for $v = 0$ is 53.696 while it is 50.215 for $v = 10$. Similarly, for the simulated data in Table 2 (i.e., for the models without intercept), the smallest ESPE is observed at $M = 6$ when the sample size is taken to be 100. It is the smallest at $M = 4$ for both choices of sample size (i.e., $n = 100, 200$). Further, ESPE for the simulated data tends to decrease with increase in $v$. At $n = 200$ and $M = 2$, the ESPE for $v = 0$ is 9.988, while it is 9.982 for $v = 10$. In the real data, the ESPE values are increasing with increase in $v$ for some choices of $M$ while it is decreasing with increase in $v$ for other choices. This is because $v$ on the one side decreases variance and increases bias on the other side. When increase in the bias is dominated, the ESPE tends to increase with increase in $v$, and vice versa. From all the tables, it can be observed that the ESPE declines as $n$ increases.

The estimated variance of prediction error of $\hat{\tau}(y)$ is obtained in Columns 4-8 using residual, UEV, FPE, GCV, and BIC in ascending order (according to their values) from left to right of each table. For numerical study with $M = 4$, $n = 50$, and $v =$

**Table 3.** ESPE and expected estimated variances of the proposed estimator under linear BFR model for hospital population.

| | | | Expected estimated variances | | | | |
|---|---|---|---|---|---|---|---|
| $n$ | $v$ | ESPE | Residual | UEV | FPE | GCV | BIC |
| | | | | $M = 1$ | | | |
| 50 | 0 | 6.6755 | 1115.0647 | 1137.8211 | 1160.5775 | 1161.0419 | 1204.0883 |
| | 1 | 38.3376 | 1082.7802 | 1104.4359 | 1126.0915 | 1126.5246 | 1167.4975 |
| | 5 | 372.9478 | 975.9444 | 994.0174 | 1012.0905 | 1012.4252 | 1046.6466 |
| | 10 | 1105.6943 | 878.3235 | 893.2103 | 908.0971 | 908.3495 | 936.5611 |
| 100 | 0 | 265.7557 | 374.5753 | 378.3589 | 382.1425 | 382.1807 | 391.9994 |
| | 1 | 304.8374 | 370.8607 | 374.5693 | 378.2779 | 378.3150 | 387.9395 |
| | 5 | 478.2710 | 357.5779 | 361.0162 | 364.4544 | 364.4875 | 373.4117 |
| | 10 | 725.2613 | 343.9352 | 347.0905 | 350.2459 | 350.2749 | 358.4662 |
| | | | | $M = 2$ | | | |
| 50 | 0 | 7.6058 | 829.4559 | 864.0166 | 898.5772 | 900.0173 | 964.6580 |
| | 1 | 7.0821 | 804.7302 | 836.9398 | 869.1494 | 870.4387 | 930.7350 |
| | 5 | 5.7066 | 746.2935 | 772.7304 | 799.1672 | 800.1040 | 849.7151 |
| | 10 | 4.7894 | 713.4974 | 736.3547 | 759.2121 | 759.9447 | 802.9160 |
| 100 | 0 | 42.9819 | 279.5041 | 285.2083 | 290.9124 | 291.0289 | 305.7728 |
| | 1 | 42.0453 | 276.6408 | 282.1757 | 287.7106 | 287.8213 | 302.1299 |
| | 5 | 38.9960 | 267.9593 | 272.9638 | 277.9684 | 278.0618 | 291.0060 |
| | 10 | 36.2653 | 261.0667 | 265.6210 | 270.1754 | 270.2548 | 282.0402 |
| | | | | $M = 3$ | | | |
| 50 | 0 | 9.1234 | 832.4586 | 885.5943 | 938.7299 | 942.1215 | 1040.3265 |
| | 1 | 8.5644 | 802.2694 | 850.1327 | 897.9959 | 900.8518 | 989.5114 |
| | 5 | 7.5052 | 767.7585 | 807.7119 | 847.6652 | 849.7452 | 924.0570 |
| | 10 | 7.0038 | 759.5841 | 796.4001 | 833.2161 | 835.0011 | 903.6091 |
| 100 | 0 | 34.8717 | 279.6398 | 288.2884 | 296.9371 | 297.2046 | 319.4683 |
| | 1 | 35.0648 | 276.3260 | 284.5826 | 292.8392 | 293.0859 | 314.3490 |
| | 5 | 35.5802 | 269.7184 | 277.0528 | 284.3872 | 284.5867 | 303.4946 |
| | 10 | 35.9369 | 266.7718 | 273.5537 | 280.3356 | 280.5081 | 298.0036 |
| | | | | $M = 4$ | | | |
| 50 | 0 | 13.3928 | 854.9230 | 929.2641 | 1003.6053 | 1010.0697 | 1145.7472 |
| | 1 | 13.9070 | 827.1330 | 891.7295 | 956.3260 | 961.3725 | 1079.8360 |
| | 5 | 14.4792 | 816.0691 | 872.8424 | 929.6156 | 933.5662 | 1038.1674 |
| | 10 | 14.6302 | 814.8074 | 869.4893 | 924.1712 | 927.8413 | 1028.7242 |
| 100 | 0 | 43.6831 | 282.4436 | 294.2121 | 305.9805 | 306.4709 | 336.6394 |
| | 1 | 44.0972 | 279.0332 | 289.9848 | 300.9363 | 301.3661 | 329.4668 |
| | 5 | 44.7919 | 275.7103 | 285.4503 | 295.1903 | 295.5344 | 320.5647 |
| | 10 | 45.0972 | 275.0052 | 284.2723 | 293.5394 | 293.8517 | 317.6817 |

**Table 4.** ESPE and expected estimated variances of the proposed estimator under proportional BFR model for hospital population.

| $n$ | $v$ | ESPE | Expected estimated variances | | | | |
|---|---|---|---|---|---|---|---|
| | | | Residual | UEV | FPE | GCV | BIC |
| | | | | | $M = 2$ | | |
| 50 | 0 | 38865.250 | 271.521 | 277.062 | 282.603 | 282.716 | 293.198 |
| | 1 | 38865.250 | 271.609 | 277.042 | 282.474 | 282.582 | 292.860 |
| | 5 | 38865.250 | 273.426 | 278.489 | 283.553 | 283.647 | 293.234 |
| | 10 | 38865.250 | 277.924 | 282.635 | 287.346 | 287.425 | 296.352 |
| 100 | 0 | 17746.841 | 184.520 | 186.384 | 188.248 | 188.266 | 193.103 |
| | 1 | 17746.841 | 184.535 | 186.380 | 188.226 | 188.244 | 193.033 |
| | 5 | 17746.841 | 184.867 | 186.645 | 188.423 | 188.440 | 193.053 |
| | 10 | 17746.841 | 185.786 | 187.491 | 189.195 | 189.211 | 193.636 |
| | | | | | $M = 3$ | | |
| 50 | 0 | 1989.648 | 915.590 | 953.740 | 991.889 | 993.479 | 1064.832 |
| | 1 | 1845.946 | 887.065 | 923.259 | 959.452 | 960.929 | 1028.655 |
| | 5 | 1405.980 | 803.258 | 833.744 | 864.230 | 865.387 | 922.520 |
| | 10 | 1051.156 | 739.901 | 766.057 | 792.214 | 793.139 | 842.226 |
| 100 | 0 | 2260.617 | 298.282 | 304.369 | 310.456 | 310.581 | 326.315 |
| | 1 | 2211.905 | 295.136 | 301.097 | 307.059 | 307.179 | 322.590 |
| | 5 | 2039.615 | 284.476 | 290.007 | 295.538 | 295.646 | 309.948 |
| | 10 | 1864.813 | 274.447 | 279.564 | 284.680 | 284.776 | 298.011 |
| | | | | | $M = 4$ | | |
| 50 | 0 | 1011.154 | 1934.051 | 2057.501 | 2180.951 | 2188.831 | 2416.991 |
| | 1 | 905.878 | 1749.106 | 1857.814 | 1966.522 | 1973.278 | 2174.374 |
| | 5 | 629.460 | 1305.618 | 1380.619 | 1455.619 | 1459.928 | 1599.022 |
| | 10 | 448.607 | 1050.837 | 1107.537 | 1164.237 | 1167.297 | 1272.648 |
| 100 | 0 | 962.666 | 531.844 | 548.293 | 564.742 | 565.251 | 607.594 |
| | 1 | 924.836 | 510.483 | 526.069 | 541.655 | 542.130 | 582.258 |
| | 5 | 802.672 | 445.061 | 458.060 | 471.060 | 471.439 | 504.926 |
| | 10 | 694.960 | 392.290 | 403.265 | 414.239 | 414.546 | 442.830 |

0, the estimated variances are 1934.051, 2057.501, 2188.831, and 2416.991, which satisfy the inequality given in Eq. (44). Tables 1–4 provide the evidence that estimated variances are reduced upon raising regularization. A similar statement can be made for the relation between estimated variances and sample size. Among the alternative variance estimators, one must choose the one that is closer to the true variance in prediction error. Hence, the choice of variance estimator depends on the statistical properties like unbiasedness and consistency of the variance estimator. A separate study can be conducted on the choice of variance estimators after conducting a detailed simulation study.

The unbiasedness and consistency of the variance estimators are good measures in this regard. However, these properties are not discussed in this study since our goal was basically the construction of estimator for $\tau$ and discussion of the problems associated with its estimation and inference about the finite population in a model-based setting.

## 8. Conclusion

A general framework of the model-based approach for estimation of finite population parameter $\tau$ (a linear combination of population values), assuming superpopulation setting, was discussed. Some special cases of the proposed general framework were deducted to observe its applicability. Expressions for prediction error variance and model-bias of the proposed estimator were derived. For statistical inference about $\tau$,

estimation of prediction error variance under residual, Generalized Cross Validation (GCV), Unbiased Estimate of Variance (UEV), Final Prediction Error (FPE), and Bayesian Information Criterion (BIC) methods (the widely used feature selection criteria in ML) were considered. The variance introduced under UEV provided minimum variance estimates compared to all other competing estimators with maximum value at BIC. Model selection for finite population parameter in the proposed general framework was discussed using incremental operators by the matrix approach. The model selection was based on a measure, named as increment in efficiency $IE$ which provides guideline on selecting a model with an appropriate number of basis functions. Positive value of $IE$ exhibited an increase in efficiency while adding additional basis functions to the feature matrix. Further ill-conditioning of the regression estimation coped with typical regularization method, which introduced a slight bias in estimates of $\beta$'s, but provided a smaller estimate of the variance of the error term and, consequently, smaller estimated variance of prediction error of $\hat{\tau}$. The current study can be used for estimation of any linear combination of population values; hence, many finite population parameters can be estimated in this general framework. The proposed model-based framework is extended to other sampling designs, multi-level models, and small area estimation. Working with a mixed model for estimating the value of the finite population parameter with basis functions is also recommended.

## Acknowledgments

The authors are very grateful to the reviewers and the editor for providing constructive comments which led to a substantial improvement in the quality of our paper.

## References

1. Cochran, W. "The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce", *The Journal of Agricultural Science*, **30**(2) pp. 262–275 (1940).

2. Murthy, M. "Product method of estimation", *Sankhya: The Indian Journal of Statistics, Series A*, **26**,(1), pp. 69–74 (1964).

3. Upadhyaya, L.N. and Singh, H.P. "Use of transformed auxiliary variable in estimating the finite population mean", *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, **41**(5), pp. 627–636 (1999).

4. Gupta, S. and Shabbir, J. "On improvement in estimating the population mean in simple random sampling", *Journal of Applied Statistics*, **35**(5), pp. 559–566 (2008).

5. Diana, G., Giordan, M., and Perri, P.F. "An improved class of estimators for the population mean", *Statistical Methods & Applications*, **20**(2), pp. 123–140 (2011).

6. Mahdizadeh, M. and Zamanzade, E. "Estimation of a symmetric distribution function in multistage ranked set sampling", *Statistical Papers*, **61**(2), pp. 851–867 (2020).

7. Zamanzade, E. and Mahdizadeh, M. "Using ranked set sampling with extreme ranks in estimating the population proportion", *Statistical Methods in Medical Research*, **29**(1), pp. 165–177 (2020)

8. Valliant, R., Dorfman, A.H., and Royall, R.M., *Finite Population Sampling and Inference: A Prediction Approach*, Number 04; QA276. 6, V3. John Wiley, New York (2000).

9. Godambe, V. "A unified theory of sampling from finite populations", *Journal of the Royal Statistical Society, Series B (Methodological)*, pp. 269–278 (1955).

10. Dorfman, A.H., Hall, P., et al. "Estimators of the finite population distribution function using nonparametric regression", *The Annals of Statistics*, **21**(3), pp. 1452–1475 (1993).

11. Chambers, R.L., Dorfman, A.H., and Wehrly, T.E. "Bias robust estimation in finite populations using nonparametric calibration", *Journal of the American Statistical Association*, **88**(421), pp. 268–277 (1993).

12. Breidt, F.J. and Opsomer, J.D. "Local polynomial regression estimators in survey sampling", *Annals of Statistics*, **16**, pp. 1026–1053 (2000).

13. Kikechi, C.B., Simwa, R.O., and Pokhariyal, G.P. "On local linear regression estimation in sampling surveys", *Far East Journal of Theoretical Statistics*, **53**(5), pp. 291–311 (2017).

14. Nadaraya, E.A. "On estimating regression", *Theory of Probability & Its Applications*, **9**(1), pp. 141–142 (1964).

15. Watson, G.S. "Smooth regression analysis", *Sankhya: The Indian Journal of Statistics, Series A*, **26**(4), pp. 359–372 (1964).

16. Chambers, R., Dorfman, A., and Sverchkov, M.Y. "Nonparametric regression with complex survey data", *Analysis of Survey Data*, pp. 151–174 (2003).

17. Fan, G., *Local Polynomial Modeling and Its Applications*, London (1996).

18. Zheng, H. and Little, R.J. "Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples", *Journal of Official Statistics*, **19**(2), p. 99 (2003).

19. Zheng, H. and Little, R. "Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples" (2004).

20. Hazlett, C. "A balancing method to equalize multivariate densities and reduce bias without a specification search", *Working Draft*, **17** (2013).

21. Sanchez-Borrego, I., Opsomer, J.D., Rueda, M., et al. "Nonparametric estimation with mixed data types in survey sampling", *Revista Matematica Complutense*, **27**(2), pp. 685–700 (2014).

22. Luc, C. "Nonparametric kernel regression using complex survey data", Job Market Paper (2016).

23. Ardilly, P. "Echantillonnage representatif optimuma probabilityes inegales", *Annales d Economie et de Statistique*, pp. 91–113 (1991).

24. Deville, J. "Constrained samples, conditional inference, weighting: Three aspects of the utilization of auxiliary information", In *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*, pp. 5–7 (1992).

25. Hedayat, A. and Majumdar, D. "Generating desirable sampling plans by the technique of tradeoff in experimental design", *Journal of Statistical Planning and Inference*, **44**(2), pp. 237–247 (1995).

26. Deville, J.-C. and Tille, Y. "Efficient balanced sampling: the cube method", *Biometrika*, **91**(4), pp. 893–912 (2004).

27. Falorsi, P.D. and Righi, P. "A unified approach for defining optimal multivariate and multidomains sampling designs", *Topics in Theoretical and Applied Statistics*, pp. 145–152, Springer (2016).

28. Clair, L. "Nonparametric Kernel estimation methods using complex survey data", PhD Thesis (2017).

29. Kikechi, C.B., Simwa, R.O., and Pokhariyal, G.P. "On local linear regression estimation of finite population totals in model based surveys", *American Journal of Theoretical and Applied Statistics*, **7**(3), pp. 92–101 (2018).

30. Rastkhiz, S.E.A., Dehkordi, A.M., Farsi, J.Y., et al. "A new approach to evaluating entrepreneurial opportunities", *Journal of Small Business and Enterprise Development*, **26**(1), pp. 67–84 (2019).

31. Kumar, S., Sisodia, B., Singh, D., et al. "Calibration approach based estimation of finite population total in survey sampling under super population model when study variable and auxiliary variable are inversely related", *Journal of Reliability and Statistical Studies*, **10**(2), pp. 83–93 (2017).

32. Chauhan, S. and Sisodia, B. "Model based prediction of finite population total under super population model", *Journal of Reliability and Statistical Studies*, **11**(2), pp. 57–68 (2018).

33. Kawakubo, Y. and Kobayashi, G. "Small area estimation of general finite-population parameters based on grouped data", arXiv preprint arXiv:1903.07239 (2019).

34. Ahmed, S. and Shabbir, J. "Model based estimation of population total in presence of nonignorable nonresponse", *PloS One*, **14**(10), p. e0222701 (2019a).

35. Ahmed, S. and Shabbir, J. "On use of ranked set sampling for estimating super-population total: Gamma population model", *Scientia Iranica*, **28**(1), (2019b). DOI: 10.24200/SCI.2019.50976.1946

36. Liu, C., Li, H.-C., Fu, K., et al. "Bayesian estimation of generalized gamma mixture model based on variational em algorithm", *Pattern Recognition*, **87**, pp. 269–284 (2019).

37. Molina, I. and Ghosh, M. "Accounting for dependent informative sampling in model-based finite population inference", *TEST*, **30**(1), pp. 1–19 (2020).

38. Jafaraghaie, R. "Prediction of finite population parameters using parametric model under some loss functions", *Communications in Statistics-Theory and Methods*, **51**(4), pp. 863–882 (2020).

39. Royall, R.M. "The linear least-squares prediction approach to two-stage sampling", *Journal of the American Statistical Association*, **71**(355), pp. 657–664 (1976).

40. Royall, R.M. and Herson, J. "Robust estimation in finite populations I", *Journal of the American Statistical Association*, **68**(344), pp. 880–889 (1973).

41. Jekabsons, G. and Zhang, Y. "Adaptive basis function construction: an approach for adaptive building of sparse polynomial regression models", *Machine Learning*, **1**(10), pp. 127–155 (2010).

42. Chambers, R. and Clark, R. "An introduction to model-based survey sampling with applications", *OUP Oxford*, **37** (2012).

43. Deville, J.-C. and Sarndal, C.-E. "Calibration estimators in survey sampling", *Journal of the American Statistical Association*, **87**(418), pp. 376–382 (1992).

44. Krishnaiah, P.R. and Alpaydin, E., *Introduction to Machine Learning*, MIT Press (2009).

45. Broomhead, D.S. and Lowe, D. "Radial basis functions, multi-variable functional interpolation and adaptive networks", Technical Report, Royal Signals and Radar Establishment Malvern (United Kingdom) (1988).

46. Powell, M.J.D. "Restart procedures for the conjugate gradient method", *Mathematical Programming*, **12**(1), pp. 241–254 (1977).

47. Lowe, D. and Broomhead, D. "Multivariable functional interpolation and adaptive networks", *Complex Systems*, **2**(3), pp. 321–355 (1988).

48. Scholkopf, B., Sung, K.-K., Burges, C.J., et al. "Comparing support vector machines with gaussian kernels to radial basis function classifiers", *IEEE Transactions on Signal Processing*, **45**(11), pp. 2758–2765 (1997).

49. Buhmann, M.D., *Radial Basis Functions: Theory and Implementations*, **12**, Cambridge University Press (2003).

50. Biancolini, M.E., *Fast Radial Basis Functions for Engineering Applications*, Springer (2017).

51. Tikhonov, A.N. and Arsenin, V.I., *Solutions of Ill-Posed Problems*, **14**, *Vh Winston* (1977). *Technometrics*, **12**(1), pp. 55–67 (1970).

52. Hoerl, A.E. and Kennard, R.W., *Ridge Regression: Biased Estimation for Non-Orthogonal Problems* (1970).

53. Teukolsky, P. and Teukolsky, S. "Vetterling, and flannery", *Numerical Recipes in C*, **18**, pp. 656–680 (1992).

54. Cartis, C., Gould, N.I., and Toint, P.L. "Universal regularization methods: varying the power, the smoothness and the accuracy", *SIAM Journal on Optimization*, **29**(1), pp. 595–615 (2019).

55. Tibshirani, R. "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), pp. 267–288 (1996).

56. Zou, H. and Hastie, T. "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), pp. 301–320 (2005).

57. Shao, J. and Wu, C.J. "A general theory for jackknife variance estimation", *The Annals of Statistics*, **17**(3), pp. 1176–1197 (1989).

58. Zheng, B. and Agresti, A. "Summarizing the predictive power of a generalized linear model", *Statistics in Medicine*, **19**(13), pp. 1771–1781 (2000).

59. Golub, G.H., Heath, M., and Wahba, G. "Generalized cross-validation as a method for choosing a good ridge parameter", *Technometrics*, **21**(2), pp. 215–223 (1979).

60. Mallows, C.L. "Some comments on $C_p$", *Technometrics*, **15**(4), pp. 661–675 (1973).

61. Boisbunon, A., Canu, S., Fourdrinier, D., et al. "AIC, $C_p$ and estimators of loss for elliptically symmetric distributions", arXiv preprint arXiv:1308.2766 (2013)

62. James, G., Witten, D., Hastie, T., et al., *An Introduction to Statistical Learning*, **112**, Springer (2013).

63. Giraud, C., *Introduction to High-Dimensional Statistics*, Chapman and Hall/CRC (2014).

64. Schwarz, G. "Estimating the dimension of a model", *The Annals of Statistics*, **6**(2), pp. 461–464 (1978).

65. Akaike, H. "On Entropy Maximization Principle, Applications of Statistics", In *Proceedings of the Symposium Held at Wright State University, PR Krishnaiah*, Ed. North-Holland Publishing Company (1977).

66. Rawlings, J.O., Pantula, S.G., and Dickey, D.A., *Applied Regression Analysis: A Research Tool*, Springer Science & Business Media (2001).

67. Orr, M.J.L., *Introduction to Radial Basis Function Networks*, Center for Cognitive Science, Edinburgh University, Scotland, UK. (1996) http://anc.ed.ac.uk/rbf.

68. Lutkepohl, H., *Handbook of Matrices*, **1**, Wiley (1996).

69. Horn, R.A. and Johnson, C.R., *Matrix Analysis Cambridge University Press* (2012).

## Appendix

The projection matrix $\boldsymbol{P}$ based on $M$ covariates is defined by:

$$\boldsymbol{P} = \boldsymbol{I}_M - \boldsymbol{\Phi}_s \boldsymbol{Q}_{sm}^{-1} \boldsymbol{\Phi}_s^T, \tag{50}$$

where $\boldsymbol{Q}_{sm}^{-1} = \boldsymbol{\Phi}_s^T \boldsymbol{\Phi}_s + v\boldsymbol{I}_M$ is the Hessian Matrix [68] based on $\boldsymbol{\Phi}_s$ with $M$ basis functions. Before applying the incremental operator to $\boldsymbol{P}_m$, we find $\boldsymbol{A}_{s(m+1)}^{-1}$ using $\boldsymbol{A}_{sm}^{-1}$. We use the two following useful lemmas from [69] for the purpose of matrix inversion.

**Lemma 1.** For any partitioned square matrix $\boldsymbol{B}$ defined by:

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix},$$

$$\boldsymbol{B}^{-1} = \begin{bmatrix} \boldsymbol{B}_{11}^{-1} + \boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\triangle^{-1}\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1} & -\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\triangle^{-1} \\ -\triangle^{-1}\boldsymbol{A}_{21}\boldsymbol{B}_{11}^{-1} & \triangle^{-1} \end{bmatrix},$$

where $\triangle = \boldsymbol{B}_{22} - \boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}$.

**Lemma 2.** Let the inverse of matrix $\boldsymbol{B}_0^{-1} \in \mathcal{R}^{m\times m}$, $\boldsymbol{X}$, $\boldsymbol{Y}^T \in \mathcal{R}^{m\times r}$, and $\boldsymbol{R} \in \mathcal{R}^{r\times r}$ be known. For computing the inverse of a new matrix $\boldsymbol{B}_1$, we have:

$$\boldsymbol{B}_1 = \boldsymbol{B}_0 + \boldsymbol{XRY}.$$

To compute the inverse of new matrix $\boldsymbol{B}_1$, we have the following relation:

$$\boldsymbol{B}_1^{-1} = \boldsymbol{B}_0^{-1} - \boldsymbol{B}_0^{-1}\boldsymbol{X}(\boldsymbol{Y}\boldsymbol{B}_0^{-1}\boldsymbol{X} + \boldsymbol{R}^{-1})^{-1}\boldsymbol{Y}\boldsymbol{B}_0^{-1}.$$

This relation between the inverse of the original matrix and the appended matrix saves much computation time.

***Decrement in variance***

The Hessian matrix after adding an additional basis function:

$$\boldsymbol{Q}_{s(m+1)}^{-1} = \begin{bmatrix} \boldsymbol{Q}_{sm}^{-1} & \boldsymbol{0} \\ \boldsymbol{0}^T & 0 \end{bmatrix} + \triangle^{-1} \begin{bmatrix} \boldsymbol{Q}_{sm}^{-1}\boldsymbol{\Phi}_{sm}^T\boldsymbol{\phi}_{s(m+1)}\boldsymbol{\phi}_{s(m+1)}^T\boldsymbol{\Phi}_{sm}\boldsymbol{Q}_{sm}^{-1} & -\boldsymbol{Q}_{sm}^{-1}\boldsymbol{\Phi}_{sm}^T\boldsymbol{\phi}_{s(m+1)} \\ -\boldsymbol{\phi}_{s(m+1)}^T\boldsymbol{\Phi}_{sm}\boldsymbol{Q}_{sm}^{-1} & 1 \end{bmatrix}. \quad (52)$$

Box II

$$\begin{aligned} \boldsymbol{\phi}_{r(m+1)}\boldsymbol{Q}_{r(m+1)}^{-1}\boldsymbol{\phi}_{s(m+1)}^T &= \begin{bmatrix} \boldsymbol{\Phi}_{rm} & \boldsymbol{\phi}_{r(m+1)} \end{bmatrix} \times \begin{bmatrix} \boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{12}^{-1} \\ \boldsymbol{Q}_{21}^{-1} & \boldsymbol{Q}_{22}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_{rm}^T \\ \boldsymbol{\phi}_{r(m+1)}^T \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{\Phi}_{rm}\boldsymbol{Q}_{11}^{-1} + \boldsymbol{\phi}_{r(m+1)}\boldsymbol{Q}_{21}^{-1} & \boldsymbol{\Phi}_{rm}\boldsymbol{Q}_{12}^{-1} + \boldsymbol{\phi}_{r(m+1)}\boldsymbol{Q}_{22}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_{rm}^T \\ \boldsymbol{\phi}_{r(m+1)}^T \end{bmatrix} \\ &= \boldsymbol{\Phi}_{rm}\boldsymbol{Q}_{11}^{-1}\boldsymbol{\Phi}_{rm}^T + \boldsymbol{\phi}_{r(m+1)}\boldsymbol{Q}_{21}^{-1}\boldsymbol{\Phi}_{rm}^T + \boldsymbol{\Phi}_{rm}\boldsymbol{Q}_{12}^{-1}\boldsymbol{\phi}_{r(m+1)}^T + \boldsymbol{\phi}_{r(m+1)}\boldsymbol{Q}_{22}^{-1}\boldsymbol{\phi}_{r(m+1)}^T. \end{aligned} \quad (53)$$

Box III

$$\boldsymbol{Q}_{m+1} = \boldsymbol{\Phi}_{s(m+1)}^T\boldsymbol{\Phi}s(m+1) + v\boldsymbol{I}_{m+1} = \begin{bmatrix} \boldsymbol{Q}_{sm} & \boldsymbol{\Phi}_{sm}^T\boldsymbol{\phi}_{s(m+1)} \\ \boldsymbol{\phi}_{s(m+1)}^T\boldsymbol{\Phi}_{sm} & \boldsymbol{\phi}_{s(m+1)}^T\boldsymbol{\phi}_{s(m+1)} + v_{m+1} \end{bmatrix}, \quad (51)$$

where:

$$\boldsymbol{\Phi}_{s(m+1)}^T = \begin{bmatrix} \boldsymbol{\Phi}_{sm} & \boldsymbol{\phi}_{s(m+1)} \end{bmatrix},$$

and:

$$\boldsymbol{Q}_{sm} = \boldsymbol{\Phi}_{sm}^T\boldsymbol{\Phi}_{sm} + v\boldsymbol{I}_m.$$

To obtain the inverse of the Hessian matrix given in Eq. (51), we use Lemma 1; Eq. (52) is shown in Box II, where $\boldsymbol{0}$ is an $m \times 1$ null vector. Further, we have Eq. (53) as shown in Box III, where:

$$\boldsymbol{Q}_{11}^{-1} = \boldsymbol{Q}_{sm}^{-1} + \triangle^{-1}\boldsymbol{Q}_{sm}^{-1}\boldsymbol{\Phi}_{sm}^T\boldsymbol{\phi}_{s(m+1)}\boldsymbol{\phi}_{s(m+1)}^T\boldsymbol{\Phi}_{sm}\boldsymbol{Q}_{sm}^{-1},$$

$$\boldsymbol{Q}_{12}^{-1} = -\triangle^{-1}\boldsymbol{Q}_{sm}^{-1}\boldsymbol{\Phi}_{sm}^T\boldsymbol{\phi}_{s(m+1)},$$

$$\boldsymbol{Q}_{21}^{-1} = -\triangle^{-1}\boldsymbol{\phi}_{s(m+1)}^T\boldsymbol{\Phi}_{sm}\boldsymbol{Q}_{sm}^{-1},$$

$$\boldsymbol{Q}_{22}^{-1} = \triangle^{-1}.$$

We first see the effect on the variance $V_M(\hat{\tau}(y) - \tau(y)) = V_M(\hat{\tau})$ (say) when there is no regularization on parameters, i.e., $v=1$ and $\boldsymbol{Q}_{sm}^{-1} = \boldsymbol{A}_{sm}^{-1}$ for $m=1,2,\cdots,M$. For prediction models with no regularization, we have the variance of $V_M(\hat{\tau})$ with $M$ regressors:

$$V_M(\hat{\tau})_m = (N-n)\sigma^2 + \sigma^2\left[\boldsymbol{\gamma}_r^T\boldsymbol{\Phi}_{rm}\boldsymbol{A}_{sm}^{-1}\boldsymbol{\Phi}_{rm}^T\boldsymbol{\gamma}_r\right],$$

and the variance of $V_M(\hat{\tau})$ with $M+1$ regressors:

$$V_M(\hat{\tau})_{m+1} = (N-n)\sigma^2 + \sigma^2\left[\boldsymbol{\gamma}_r^T\boldsymbol{\Phi}_{r(m+1)}\boldsymbol{A}_{s(m+1)}^{-1}\boldsymbol{\Phi}_{r(m+1)}^T\boldsymbol{\gamma}_r\right].$$

The bias of ridge regression estimator with $M$ basis functions is given by:

$$B_M(\hat{\tau}_{ridge}(y)) = -v\boldsymbol{\gamma}_r^T\boldsymbol{\Phi}_{rm}\boldsymbol{Q}_{sm}^{-1}\boldsymbol{\beta}_m, \quad (54)$$

and for $M+1$ basis function after some matrix multiplication, the bias becomes:

$$\begin{aligned} B_{M+1}(\hat{\tau}_{ridge}(y)) = -v\boldsymbol{\gamma}_r^T\Bigg[&\boldsymbol{\Phi}_{rm}\boldsymbol{Q}_{sm}^{-1}\boldsymbol{\beta}_m \\ &+ \frac{1}{\triangle}\boldsymbol{\Phi}_{rm}\boldsymbol{Q}_{sm}^{-1}\boldsymbol{\Phi}_{sm}\boldsymbol{\phi}_{s(m+1)}\boldsymbol{\phi}_{s(m+1)}\boldsymbol{\Phi}_{sm}\boldsymbol{Q}_{sm}^{-1}\boldsymbol{\beta}_m \\ &+ \boldsymbol{\phi}_{r(m+1)}\boldsymbol{Q}_{21}^{-1}\boldsymbol{\beta}_m + \boldsymbol{\Phi}_{rm}\boldsymbol{Q}_{12}^{-1}\beta_{m+1} \\ &+ \boldsymbol{\phi}_{r(m+1)}\boldsymbol{Q}_{22}^{-1}\beta_{m+1}\Bigg]. \end{aligned} \quad (55)$$

## Biographies

**Shakeel Ahmed** recently, completed his PhD in Statistics from the Department of Statistics Quaid-i-Azam University Islamabad, Pakistan and won Vice Chancellor Gold Medal from the University in 2015. He has published about 15 research papers in internationally reputed journals in the field of survey methodologies and estimation of parameters under new data collection mechanism, especially in the ranked set sampling scheme and the model-based estimation approach. He is now working at the College of Computer Science and Information System, Institute

of Business Management, Karachi, Pakistan as an Assistant Professor. He is currently working on finite population parameter estimation using the emerging prediction approach under model-based setting. The author is also working as an Associate Editor for a reputed Journal "BMC-Public Health Bio-statistics and methods section".

**Javid Shabbir** is working as a Tenured Professor at Statistics at Department of Statistics Quaid-i-Azam University, Islamabad, Pakistan. He completed his PhD in Statistics from Kent University at Canterbury, UK in 1997. He had Post-Doctoral positions at University of Southern Maine, USA in 2003 and University of North Carolina at Greensboro USA in 2005. He has published about 300 article papers in different internationally reputed journals. His area of research includes survey sampling and randomized response techniques. He has supervised many MPhil and PhD students at the department.