A Novel Basis Function Approach to Finite Population Parameter Estimation

Shakeel Ahmed and Javid Shabbir
Department of Statistics Quaid-i-Azam University, Islamabad, 44000, Pakistan

March 3, 2021

ABSTRACT

Modeling non-linear data is a common practice in data science and machine learning (ML). It is aberrant to get a natural process whose outcome varies linearly with the values of input variable(s). A robust and easy methodology is needed for accurately and quickly fitting a sampled data set with a set of covariates assuming that the sampled data could be a complicated non-linear function. A novel approach for estimation of finite population parameter $\tau$, 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 single predictor under an automatic matrix approach, and ill-conditioned regression models are discussed. Prediction error variance of the proposed estimator is estimated under 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.

Keyword: Superpopulation, Survey sampling, Inference, Basis functions, Feature matrix, Non-linear function.

*Corresponding author address: Department of Mathematics and Statistics, Institute of Business Management, Karachi, Pakistan. E-mail: shakeel.ahmed@iobm.edu.pk, Mobile:+92-323-9710282
1 Introduction

Researchers, in survey sampling, mostly favored random sampling for a valid statistical inference due to its attractive long run properties such as unbiasedness and efficiency in designed based sense. However, they ignored the importance of underlying model relationship between the survey variable and one or more covariate(s) at estimation stage. Without exposing an appropriate model relationship between the survey variable and the covariates, researchers in design-based paradigm have been constructing estimators for the unknown population quantities such as population total, mean, variance etc, relying only on randomization mechanism incurred by sampling. They have been utilizing sample estimates and known population parameters of the auxiliary variable(s) in estimation stage for efficiency improvement. Thousands of estimators for estimating population parameters have been developed in regard with efficiency improvement and bias reduction under design-based approach such works can be found in [1], [2], [3], [4], [5], [6] and [7]. On contrary, believers in model-based paradigm emphasis that the randomization is a property of error term used in model hence it is not necessary nor sufficient condition for a solid statistical inference [8]. In model-based framework, initially, [9] used a regression model of the response on the covariates to predict the non-sampled values and their total which is assumed as unknown and random quantity. Many varieties of model-based estimators have been developed for efficiency improvement, bias reduction and maintaining robustness to model failure in last two decades of 20th century. [10] and [11] have worked on estimating a smooth function and used for predicting the non-sampled values in estimation of finite population total. The asymptotic bias of the regression estimator of population total does not account the division of the sampling distribution into sampled and non-sampled parts. [12] worked on a class of estimators based on local polynomial regression which are weighted linear combinations of the study variables, where the weights are calibrated to control totals which are known. [13] employed model based approach, using the local linear regression (LLR), to estimate the unknown parameters of the study variable. They particularly derived the properties of the proposed estimator and compared with Nadaraya-Watson regression estimator [14, 15] and shown that the two estimators are asymptotically equally efficient. [16] observed that the calibration estimator based on the columnar model perform slightly better than the best linear unbiased estimator (BLUE) at 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. The calibration based estimators provide a guideline to the balance sampling frame work which suggest to select the samples with the condition of unbiasedness in presence of reliable auxiliary data. They noticed that the total estimators of population total from a nonparametric regression model provide approximate unbiasedness without imposing restriction on balancing and results near minimal variance. However, [17] uncovered a more appealing strategy than the kernel regression, e.g, the variable bandwidth LLR approach. [18] proposed a model-based estimator that works with penalized spline regression, and extended the estimator to two-stage sampling [19]. [12] used the classical local polynomial regression (CLPR) estimator for the regression function for obtaining the model assisted estimator of the total in finite populations. [20] developed a method for balancing that equalize the multivariate densities and reduce bias without searching specifications. [21] estimated the regression function with mixed variable using a modified form of local constant estimator . [22] derived properties of weighted nonparametric regression estimator, using probabilities as weight, for complex surveys under combined inference. Several partial solutions for balanced sampling are available in [23], [24], [25] and [8]. [16] proposed a general method, called the cube method, which is appropriate for a set of quantitative or qualitative balancing variables and allows unequal probabilities of inclusion. [26] developed the cube method for the selection of approximately balanced samples based on equal or unequal inclusion probabilities with a number of auxiliary variables. [27] developed a balanced sampling strategy in multi-way stratification settings for small area estimation and used it to obtain planned sample size for domains belonging to different partitions of the population (small areas). The strategy lowers the sampling errors of domain estimates and provided threshold values. [28] considered the nonparametric estimation methods for data analysis in complex surveys. [29] used the LLR technique to asses the properties of the derived estimator and compare its performance with the existing estimators. The LLR technique can be also used for evaluating entrepreneurial opportunities. Therefore, add this application and the following sentence here: Note that the LLR technique
can also be used for evaluating entrepreneurial opportunities, see [30] for more information about this topic.

The researches documented in literature consist a wide variety of model-based estimators constructed under different forms of the relationship between the outcome and the predictors, see [31], [32], [33], [34], [35], [36], [37] and [38]. Although a wide variety of restricted sampling methods have been in [8] some of them are based on linear regression model, some on polynomial models and some on proportional and stratified population models. Further, we need a general framework for predicting responses from non-linear (in variable) functions of auxiliary data. The non-linear function of the auxiliary variable may be logarithm, some power or exponential form. The problem of concern is then the prediction of output variable for non-sampled set based on the relationship between the inputs and outputs in sampled set and the known values of the input variable(s) in non-sampled set. No study has been detected discussing the non-linear regression model under general prediction theorem. It is aberrant to exist natural processes whose outcome varies linearly with the values of predictors. To fill this gap, in this study, we establish a general framework under BFR model for estimation of finite population parameters. A novel approach for estimation of finite population parameter \( \tau \), 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 single predictor under an automatic matrix approach, and ill-conditioned regression models are discussed. Prediction error variance of the proposed estimator is estimated under widely used feature selection criteria in ML. Section 2 delineates model-based estimation developed in literature with its usual notations. Our proposed basis function approach with some special cases is described in Section 3. Estimation of \( \tau \) under regularized BFR is considered in Section 4. Section 5 covers variance estimation and comparison of competing variance estimators. Model selection and simulation studies are covered in Sections 6 and 7 respectively. Section 8 concludes the study with some future recommendations.

2 Model Based Estimation

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

\[
Y = X\beta + \epsilon, \tag{1}
\]

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

\[
X = \begin{bmatrix} X_s \\ X_{\bar{s}} \end{bmatrix} \quad \text{and} \quad \Sigma_{ss} = \begin{bmatrix} \Sigma_{ss} & \Sigma_{s\bar{s}} \\ \Sigma_{s\bar{s}} & \Sigma_{\bar{s}\bar{s}} \end{bmatrix}.
\]

The quantity of interest, to be estimated, is the linear combination of the population values \( \tau(y) = \gamma^T y \) which is a realization of the random variable \( \gamma^T Y \), where \( \gamma = (\gamma_i, i \in \mathcal{U}) \) is the vector of weights which can also be partitioned for sampled and non-sampled values as \( \gamma = (\gamma_s^T, \gamma_{\bar{s}}^T)^T \). [8] defined a linear estimator (known as best linear unbiased predictor (BLUP)) for \( \tau(y) \) as \( \hat{\tau}(y) = g_s Y_s \), where \( g_s = (g_i, i \in s) \) is a vector of constants to be optimized. Under Model Equation (1), [39] proposed the general prediction estimator for \( \tau(y) \) as

\[
\hat{\tau}(y) = \gamma_s^T Y_s + \gamma_{\bar{s}}^T \left[ X_s \hat{\beta} + \Sigma_{ss} \Sigma_{ss}^{-1} (Y_s - X_s \hat{\beta}) \right], \tag{2}
\]
where $\hat{\beta} = (X_s^{-1}\Sigma_{ss}^{-1}X_s)^{-1}X_s^T\Sigma_{ss}^{-1}Y_s$ is the weighted least square (WLS) estimator of the vector $\beta$. The variance of $\hat{\tau}(y)$, is given by

$$V_M(\hat{\tau}(y) - \tau(y)) = \gamma_s^T\left(\Sigma_{ss} - \Sigma_{ss}\Sigma_{ss}^{-1}\Sigma_{ss}\right)\gamma_s + \gamma_s^T\left(X_s - \Sigma_{ss}\Sigma_{ss}^{-1}X_s\right)\left(X_s^T\Sigma_{ss}^{-1}X_s\right)^{-1}\left(X_s - \Sigma_{ss}\Sigma_{ss}^{-1}X_s\right)^T\gamma_s.$$  

(3)

When sampled and non-sampled units are uncorrelated i.e. $\Sigma_{ss} = 0$, the BLUP for $\tau(y)$ reduces to

$$\hat{\tau}(y) = \gamma_s^TY_s + \gamma_s^TX_s\hat{\beta}$$

with prediction error variance

$$V_M(\hat{\tau}(y) - \tau(y)) = \gamma_s^T\left(\Sigma_{ss} + X_s\left(X_s^T\Sigma_{ss}^{-1}X_s\right)^{-1}X_s^T\right)\gamma_s.$$  

(5)

The general prediction estimator was constructed using a general linear regression model of $Y$ on a matrix of covariates $X$. It is noteworthy that for generalizing the result from 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 through sample. In such situations, it is necessary to robustify the sampling mechanism and/or estimator from model failure. One way to robustify is to measure the effects such as bias and variance that show how these measures changes when the working model deviates from the true model. [40] emphasized on the balancing of a sample to protect the inference against model misspecification. [8] done an extensive work 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 following [41].

### 3 Model based Estimation Using Basis Functions

Starting with a single input variable $X$, the corresponding vector basis function is defined as $\Phi(x_i) = (\Phi_0(x_i), \Phi_1(x_i), \ldots, \Phi_M(x_i))$ attached to the $i$th population unit, where $M$ is the number of basis functions in the model. The matrix comprising the basis function is known as feature matrix in ML terminology and presented as

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

The population BFR model is defined by

$$Y = \Phi\beta + \epsilon,$$  

(7)
where \( e \) is the vector of random errors assumed to be distributed normally with mean vector 0 and variance-covariance matrix \( \Sigma \). Further, \( f(x, \beta) = \Phi \beta \) is the population regression function. The basis function \( \Phi_j(X) \) usually found in non-linear functions in the input variable \( x \) which allows the function \( E_M(Y|\Phi, \beta) = \Phi \beta \) as a non-linear in \( x \). But the conditional mean is still linear in parameters \( \beta \). For prediction of the non-sampled values of the population parameter \( \tau(y) \) the feature matrix \( \Phi \) can be partitioned as

\[
\Phi = \begin{bmatrix} \Phi_s & \Phi_s \end{bmatrix},
\]

where \( \Phi_s \) and \( \Phi_s \) 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 general linear estimator proposed by [8] with feature matrix \( \Phi \) as:

\[
\hat{\tau}(y) = \gamma_s^T Y_s + \gamma_s^T \left[ \Phi_s \hat{\beta} + \Sigma_{ss}^{-1} \sum (y_s - \Phi_s \hat{\beta}) \right],
\]

where \( \hat{\beta} = (\Phi_s^T \Sigma_{ss}^{-1} \Phi_s)^{-1} \Phi_s^T \Sigma_{ss}^{-1} y_s \) is the weighted least square (WLS) estimator of \( \beta \). The variance of \( e(\tau) \) is given by

\[
V(e(\tau)) = \gamma_s^T \left( \Sigma_{ss} - \Sigma_{ss} \Sigma_{ss}^{-1} \Sigma_{ss} \right) \gamma_s + \gamma_s^T \left( \Phi_s - \Sigma_{ss} \Sigma_{ss}^{-1} \Phi_s \right) \left( \Phi_s^T \Sigma_{ss}^{-1} \Phi_s \right)^{-1} \gamma_s,
\]

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

**Proof:** Derivation of Equations (8) and (9) can be found after replacing the feature matrix \( \Phi \) by the data matrix \( X \) in general prediction theorem given in [8, Chapter 2]. For simplicity, we assume non-informative sampling (i.e. the sampled and non-sampled outcomes have same distribution) conditional on values of the auxiliary variables resulting \( \Sigma_{ss} = 0 \), the BLUP for \( \tau(y) \) reduced to

\[
\hat{\tau}(y) = \gamma_s^T Y_s + \gamma_s^T \Phi_s \hat{\beta}
\]

with prediction variance

\[
V(e(\tau)) = \gamma_s^T \left( \Sigma_{ss} \Phi_s \right) \left( \Phi_s^T \Sigma_{ss}^{-1} \Phi_s \right)^{-1} \gamma_s.
\]

Assuming iid noise in the data i.e. \( \Sigma_{ss} = \sigma^2 I_n \) and \( \Sigma_{ss} = \sigma^2 I_{N-n} \), the resulting expression for variance of prediction error can be written as

\[
V(e(\tau)) = \sigma^2 \left[ \gamma_s^T \gamma_s + \gamma_s^T \Phi_s \left( \Phi_s^T \Phi_s \right)^{-1} \Phi_s^T \gamma_s \right].
\]

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

### 3.1 Special Cases

In this subsection, we discuss some members of BFR model and obtain estimators of total output under 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 finite population total as follow

\[ y_i = \beta_0 + \epsilon_i \quad \text{for } i = 1, 2, ..., N. \]  

(13)

The model given in Equation (13) is known as homogeneous population model (HPM) and obtained by taking \( \Phi \) as \( N \) dimensional vector of 1’s. The expansion estimator for \( t_y = \sum_{i \in U} y_i \) (population total) under HPM is obtained as:

\[ \hat{t}_y^E = \sum_{i \in s} y_i + \sum_{i \in s} \hat{\beta}_0, \]

(14)

where \( \hat{\beta}_0 = \frac{\sum_{i \in s} y_i}{n} \) is the best linear unbiased predictor (BLUP) for \( \beta_0 \) obtained under the ordinary least square (OLS) assumptions. The expansion estimator is unbiased when underlying model is correct. The prediction error variance of the expansion estimator, is given by

\[ V_M(\hat{t}_{ys} - t_y) = N^2(\frac{1}{n} - \frac{1}{N})\sigma^2 \]

(15)

which is equivalent to the designed-based variance of 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

\[ y_i = \beta_0 + \sum_{j=1}^{M-1} \beta_j \Phi_j(x_i) \quad \text{for } i = 1, 2, ..., N. \]  

(16)

The total estimator under linear BFR model is obtained as

\[ \hat{t}_y(\text{reg}) = \sum_{i \in s} y_i + \sum_{i \in s} \left\{ \hat{\beta}_0 + \sum_{j=1}^{M-1} \hat{\beta}_j \Phi_j(x_i) \right\}, \]

where \( \hat{\beta}_0 = \bar{y} - \sum_{j=1}^{M-1} \hat{\beta}_j \bar{\Phi}_j, \bar{y} = \frac{1}{n} \sum_{i \in s} y_i \) and \( \bar{\Phi}_j = \frac{1}{n} \sum_{i \in s} \Phi_j(x_i) \). The total estimator under BFR model can be simplified to

\[ \hat{t}_y(\text{reg}) = N \left[ \bar{y} + \sum_{j=1}^{M-1} \hat{\beta}_j \{ \Phi_{jU} - \bar{\Phi}_j \} \right], \]

(17)

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

\[ V_M(\hat{t}_y(\text{reg}) - t_y) = N^2 \left[ \sum_{j=1}^{M-1} (\Phi_{jU} - \bar{\Phi}_j)^2 V_M(\hat{\beta}_j) + (\frac{1}{n} - \frac{1}{N})\sigma^2 \right]. \]

(18)
Consider the case of single basis function with intercept i.e. $M = 2$, we have following variance expression

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

This variance decreases when mean of the basis function for sampled and non-sampled units coincide and there is a high variation in 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 estimation of superpopulation parameters when the variance structure is known. We consider following $(M - 1)$ degree polynomial model with basis function contained single regressor with no intercept as:

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

where $f(x, \beta) = \sum_{j=1}^{M-1} \beta_j \Phi_j(x)$. The gamma population model (GPM) discussed by [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 linear regression estimator with constant variance. To obtain homoscedastic error term, we adopt WLS method to estimate the parameters involved in model Equation (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 best linear unbiased estimator (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 $V_M(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i \in s} \Phi_1^2(x_i)}$. The ratio estimator under single basis function, is given by

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

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

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

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

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

and

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

The model mean squared prediction error is given by

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

Use of balance sampling with $\sum_{i \in s} \lambda_i^* \Phi_1(x_i) - \sum_{i \in s} \Phi_1(x_i)$ results in unbiasedness which coincides with estimator given in [43].
3.2 Some Special Basis Functions

Previously, we discussed the prediction estimators under general BFR model for estimating 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 estimator of \( \tau(y) \). The world we are living is much complicated and we can’t easily adopt a linear model to capture the wide variety of so called basis functions that we might need in prediction. To capture the complex phenomenon with data having non-linear behavior data scientists urged on use of a variety of basis functions that make more precise prediction [44]. Some commonly used basis functions are briefly discussed in the following subsections.

3.2.1 Polynomial Basis Functions

While applying polynomial regression for predicting non-sampled values it is essential to decide the degree of the polynomial before going toward the prediction problem. The question of "how many degree of the polynomial can be answered through visual display of sample data (when feature dimension is one or two)". It is much tougher in case of three or higher dimensions of the feature and it is complete wastage of time if there exists interaction terms of the features which influence the outcome. For mutually-interacting high-dimensional data set, we can reach to 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 move toward some machine learning technique 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 corresponding feature matrix

\[
\Phi = \begin{bmatrix}
1 & x_1^1 & x_1^2 & \ldots & x_1^{M-1} \\
1 & x_2^1 & x_2^2 & \ldots & x_2^{M-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_N^1 & x_N^2 & \ldots & x_N^{M-1}
\end{bmatrix}
\]

The polynomial used in the feature matrix is of order \( M-1 \). The determination of 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 homogeneous population model (HPM), \( M = 2 \) linear regression model (LRM) with intercept and for \( M = 3 \), we get quadratic regression model (QRM). The polynomial BFRM provides global basis functions which effect the prediction over the whole input range. The number of polynomials increases exponentially with increase in \( M \). While the local basis functions are considered as appropriate in prediction problems.

3.2.2 Basis Functions with two Regressors

Further, the polynomial curve fitting is applicable only for single input variable \( x \). It is not easy to generalized it for several input variables. For three input variables 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

\[
y = \sum_{j_1,j_2,j_3} \beta_j \Phi_j(x) + \epsilon
\]

\[
= \beta_0 + \beta_{100} x_1 + \beta_{010} x_2 + \beta_{001} x_3 + \beta_{110} x_1 x_2 + \beta_{101} x_1 x_3 + \beta_{011} x_2 x_3 + \beta_{200} x_1^2 + \beta_{020} x_2^2 + \beta_{002} x_3^2 + \epsilon
\]  

(27)

where \( \epsilon \) is random error term. For \( p \) covariates the number of quadratic terms is \( \frac{1+p+p(p-1)}{2+p} \). In above example, we have \( p = 3 \), hence the number of terms is 10. For \( p \) input a general case of

\[
\Phi_j(x) = \left\{ \prod_{k=1}^{p} x_k^{m_k} : \sum_{k=1}^{p} m_k \leq p \right\}
\]

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)

Radial basis functions are 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 some other point, called a center, so that $\Phi(x, c) = \Phi(||x - c||)$. The concept of radial basis was initially introduced by [45] and stemmed from [46]. [47] discussed the relationship between “learning” in adaptive-layered networks and fitting of data in high-dimensional surfaces. RBFs are used as a kernel in classification of support vector [48]. [49] provided theory and implementation of RBF. Later [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 we know that the ‘linearity’ in linear regression refers to the model being linear in coefficients, and not necessarily in features (or independent variables). Features can be of any degree or 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 after using alternate basis functions in Equation (8).

4 Estimation Under Regularized Regression

In regression analysis, over-fitting refers to the dependent variable corresponding exactly (or very close) to a particular data set and not further able to fit additional data points. Such situations are termed as ill-conditioning in regression analysis. Initially, [51] worked on the mathematical aspect of the ill-posed problems and discussed the problem in their book. Aside from [51], [57] suggested ridge regression method for solving ill-conditioned linear regression problem. Here ill-conditioning refers to numerical difficulties in obtaining the inverse of the matrix which is necessary in obtaining variance of estimators of the superpopulation parameters. [57] method was actually a crude form of the ridge regression now known as zero order regularization [52]. When neural network (NN) became famous, in 1980’s, the weight decay is invented to deal with prune network connections that are considered to be unimportant. Weighted decay is soon recognized as alternate of ridge regression in NN as it involves adding penalties to the cost function (sum-squared error). A variety of regularization methods are available in literature and most of them are cited in [53]. In this section, we confined our discussion to the simple regularization method introduced by [57] although our prediction problem can be handled by using more advanced regularization methods e.g. Least Absolute Shrinkage and Selection Operator (LASSO) [54], elastic net regression [55] and their extensions. The selection of certain regularizer depends on bias and variance trade-off. As regularization reduces variance on one side by increasing bias on the other side resulting an adjustment in the MSE. If $E_M(\beta_{\text{ridge}}) = \beta$ for all $\Phi$, then the estimator $\hat{\beta}_{\text{ridge}}(y)$ will be unbiased. However an unbiased estimator may still have larger MSE if the variance of the estimators of superpopulation parameters are 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 in the sense of [51]. To reduce the high variation significantly [57] suggested to introduce small amount of bias so that the net effect results in reduction in MSE. Under regularization, we have following cost function (sum-squared error)

$$C = (y_s - \Phi_s \beta)^T (y_s - \Phi_s \beta) + v \beta^T \beta,$$

where the positive constant $v$ is called regularizer which creates bias in the estimate of $\beta$ and reduces variance on the other side. Optimizing the cost function given in Equation (28), we have following ridge regression estimator for the coefficient vector $\beta$ as:

$$\hat{\beta}_{\text{ridge}} = Q_s^{-1} \Phi_s^T y_s,$$
where the matrix \( Q_\tau = \Phi_\tau^T \Phi_\tau^T + vI_n \) is symmetric i.e. \( Q_\tau^T = Q_\tau \). An estimator of population parameter \( \tau(y) \) using ridge regression model, is given by

\[
\hat{\tau}_{\text{ridge}}(y) = \gamma_\tau^T y_s + \gamma_\tau^T \Phi_s \hat{\beta}_{\text{ridge}}
\]

which has model bias

\[
E_M(e(\hat{\tau}_{\text{ridge}})) = \gamma_\tau^T \Phi_s [E_M(\hat{\beta}_{\text{ridge}}) - \beta],
\]

where \( e(\hat{\tau}_{\text{ridge}}) = \hat{\tau}_{\text{ridge}}(y) - \tau(y) \). After some simplification (see Equation (54) Appendix), the bias reduces to

\[
B_M(\hat{\tau}_{\text{ridge}}(y)) = -v \gamma_\tau^T \Phi_s Q_s^{-1} \beta.
\]

The bias given in Equation (31) depends on the regularizer \( v \). It can be concluded that the bias tends to decline as \( v \rightarrow 0 \) depending on entries in \( Q_s^{-1} \) (which also depend on \( v \)). The variance expression is given by

\[
V_M(e(\hat{\tau}_{\text{ridge}})) = \sigma^2 \left[ \gamma_\tau^T \gamma_s + \gamma_\tau^T \Phi_s \left( Q_s^{-1} - Q_s^{-2} \right) \Phi_s^T \gamma_s \right]
= V_M(e(\hat{\tau})) - \sigma^2 \gamma_\tau^T \Phi_s Q_s^{-2} \Phi_s^T \gamma_s.
\]

This shows that regularization reduces variance by an amount of \( \sigma^2 \gamma_\tau^T \Phi_s Q_s^{-2} \Phi_s^T \gamma_s \). This amount increases by increasing the parameter \( v \) which ultimately increases the efficiency with larger amount of bias. The MSE of \( \hat{\tau}_g(y)_{\text{ridge}} \) is then obtained using bias and variance relation and given by

\[
MSE_M(\hat{\tau}_{\text{ridge}}(y)) = \sigma^2 (\gamma_\tau^T \gamma_s) + \gamma_\tau^T \Phi_s \left[ \sigma^2 (Q_s^{-1} - Q_s^{-2}) + v^2 Q_s^{-1} \beta \Phi_s^T Q_s^{-1} \right] \Phi_s^T \gamma_s
\]

The amount \( \gamma_\tau^T \Phi_s [v^2 Q_s^{-1} \beta \Phi_s^T Q_s^{-1} - \sigma^2 Q_s^{-2}] \Phi_s^T \gamma_s \) 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 non-linear in \( v \). So it is not straightforward to obtain an explicit expression for \( v \) which minimizes Equation (33). Alternatively, one can adopt model selection criteria to obtain an optimum choice of \( v \). Since all the criteria for model selection are also non-linear in \( v \), we need some non-linear optimization problem here. We can use any standard method for this purpose, such as the Newton method. We leave the derivation of optimum choice of \( v \) for future study.

5 Variance Estimation and Comparison

After obtaining the prediction error, 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 to variance estimation methods in design-based paradigm such as Jackknife technique [56], in model based approach, we utilize model selection criteria which indirectly provide estimate of error variance \( \sigma^2 \) in model-based approach. It can be seen that the variance of error term given in Equation (12) depends on error variance \( \sigma^2 \) and the auxiliary data from the whole population. When sub-matrix of the basis function for the non-sampled part is known, we need estimate for \( \sigma^2 \) only for estimating the prediction error variance of \( \tau(y) \). A sample estimate for the prediction error variances given in Equations (12) and (31) can be expressed as

\[
\hat{V}(e(\hat{\tau})) = \hat{\sigma}^2 \left\{ \gamma_\tau^T \gamma_s + \gamma_\tau^T \Phi_s \left( \Phi_s^T \Phi_s \right)^{-1} \Phi_s^T \gamma_s \right\},
\]

and

\[
\hat{V}(e(\hat{\tau}_{\text{ridge}})) = \hat{V}(e(\hat{\tau}_{\text{ridge}})) - \hat{\sigma}^2 \gamma_\tau^T \Phi_s Q_s^{-2} \Phi_s^T \gamma_s.
\]
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 estimation of finite population parameter $\tau(y)$. The projection matrix, say $P$, plays key role in obtaining the estimate for $\sigma^2$ using below mentioned methods. The projection matrix defined in Equation (50) (see Appendix) is symmetric and idempotent ($P^2 = P$) when no regularization is applied. For obtaining estimates for $\sigma^2$, we use following model selection criteria.

1- Residual Method
2- Cross Validation (CV)
3- Generalized Cross Validation (GCV)
4- Final Prediction Error 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 $\hat{\tau}_{\text{ridge}}(y)$ of prediction error variance of the estimator $\hat{\tau}(y)$ are obtained under residual method after inserting $\hat{\sigma}^2_{\text{res}}$ in Equations (34) and (35) respectively, where

$$\hat{\sigma}^2_{\text{res}} = \frac{1}{n-M}y_s^TP^2y_s,$$

(36)

where $P = I_M - \Phi_sQ_s^{-1}\Phi_s^T$. Another most widely used model selection criterion is unbiased estimate of variance (UEV) which is similar to residual variance and is obtained by replacing the number of total parameter by the number of effective parameter in the denominator. The UEV estimator of $\sigma^2$, is given by

$$\hat{\sigma}^2_{\text{UEV}} = \frac{1}{n-M^*}y_s^TP^2y_s,$$

(37)

where $M^* = n - \text{trace}(P)$ 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 non-sampled part of the population.

5.2 Estimation under by Cross Validation (CV)

A new variant of CV involve randomly split the data into a training set and a test set $k$ 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 $k = n$, the total number of data points. It means that the model is trained $n$ times including all the data except of 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 $\hat{\sigma}^2_{\text{LOO}}$ in Equations (34) and (35), where

$$\hat{\sigma}^2_{\text{LOO}} = \frac{1}{n}y_s^TP\{\text{diag}(P)\}^{-2}Py_s.$$

(38)

The estimated variance obtained under LOO cross validation is a good estimate of model variance, but at first glance it seems more expensive and tiresome to compute. Luckily, locally weighted regression makes it easy as they make regular predictions. It means that computing the LOO-XVE consumes no more time than the residual error that’s why it is preferred as model selection criteria.

5.3 Estimation under Generalized Cross Validation

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

$$
\hat{\sigma}_{GCV}^2 = \frac{ny_s^TP^2y_s}{\text{trace}(P)}.
$$

Equation (39)

GCV is among one of the model selection criteria which includes an adjustment to the average of mean squared prediction error over the training set. It is equivalent to standard residual method given in Equation (38), when $\frac{n}{\text{trace}(P)} = \frac{1}{n-M^*}$, where $M^* = n - \text{sum(diag}(P))$ is the effective number of parameters in the model. GCV can also be expressed in term of the effective number of parameters $M^*$ instead of $\text{trace}(P)$ as

$$
\hat{\sigma}_{GCV}^2 = \frac{ny_s^TP^2y_s}{(n-M^*)^2}.
$$

Equation (40)

5.4 Estimation under Final Prediction Error (FPE)

Mallow’s $C_p$ [60], named after Colin Lingwood Mallows, is a statistic which assesses the fit of a regression model that is estimated via ordinary least squares (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 the available predictors. Under Gaussian linear regression model Mallows’s $C_p$ is equivalent to Akaike’s Information Criterion (AIC), a most widely used model evaluation criterion [61], and is used as an alternate of AIC. An estimator for prediction error variance of $\hat{\tau}(y)$ under final prediction error (FPE) method is obtained by replacing $\hat{\sigma}^2$ by $\hat{\sigma}_{FPE}^2$ in Equation Equation (34), where $\hat{\sigma}_{FPE}^2$, an alternative version of Mallows’s $C_p$ [62], and is given by

$$
\hat{\sigma}_{FPE}^2 = \frac{1}{n} (y_s^TP^2y_s + 2M^*\hat{\sigma}_{res}^2) = \frac{n + M^*}{n - M^*} \frac{y_s^TP^2y_s}{n},
$$

Equation (41)

where $M^*$ is the effective number of parameters. The $\hat{\sigma}_{FPE}^2$ suffers from two limitations: 1. the approximation is valid only for large enough sample size and 2. it can’t deal complex set of models as in the variable selection (feature selection in machine learning) problems [63].

5.5 Estimation under Bayesian Information Criterion (BIC)

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

$$
\hat{\sigma}_{BIC}^2 = \frac{1}{n} (y_s^TP^2y_s + \ln(n)M^*\hat{\sigma}_{res}^2) = \frac{n + M^*}{n - M^*} \frac{y_s^TP^2y_s}{n}
$$

Equation (42)

where $\ln(n)$ is the natural logarithm of $n$. Here $\hat{\sigma}_{BIC}^2$ 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 above discussed competing estimators of prediction variance, we write all the variance estimators in the form of $\hat{\sigma}_{abc}^2 = \Gamma_{abc} y_s^TP^2y_s/n$ and have following natural ordering as

$$
\Gamma_{UEV} \leq \Gamma_{FPE} \leq \Gamma_{GCV} \leq \Gamma_{BIC}
$$

Equation (43)
The factors $\Gamma$s are approximated by using Taylor’s series as:

$$\Gamma_{Res} = \frac{n}{n - M^*} = 1 + \frac{M^*}{n} + \frac{M^{s2}}{n} + \frac{M^{s3}}{n^2} + ...,$$

$$\Gamma_{FPE} = \frac{n + M^*}{n - M^*} = 1 + \frac{2M^*}{n} + \frac{2M^{s2}}{n^2} + \frac{2M^{s3}}{n^3} + ...,$$

$$\Gamma_{GCV} = \frac{M^{s2}}{(n - M^*)^2} = 1 + \frac{2M^*}{n} + \frac{3M^{s2}}{n^2} + \frac{4M^{s3}}{n^3} + ...,$$

$$\Gamma_{BIC} = \frac{n + (\ln(n) - 1)M^*}{n - M^*} = 1 + \ln(n)\left(\frac{M^*}{n} + \frac{M^{s2}}{n^2} + \frac{M^{s3}}{n^3} + ...ight),$$

where the subscript ”Res” is used to denote the error variance obtained through residual method. Hence the estimators of $\hat{\sigma}^2$ obtained through different model selection criteria can be ranked according to the factor $\Gamma$. Hence variance estimators can also be ranked as

$$\hat{V}(e(\hat{\tau}))_{UEV} \leq \hat{V}(e(\hat{\tau}))_{FPE} \leq \hat{V}(e(\hat{\tau}))_{GCV} \leq \hat{V}(e(\hat{\tau}))_{BIC}$$

(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 an interesting small fraction of all subset, we need heuristics. Forward selection and backward selection methods are two widely used heuristics for model selection. Although backward selection is a common 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 hand, the forward selection method starts 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 an algorithm of non-linear type, still it has the following plus points.

(i). The number of hidden units is not need to be fixed in advance.
(ii). It has a tractable criteria for model selection.
(iii). It needs relatively low computational effort.

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

6.1 Model Selection Under OLS

The reduction in variance on using additional basis function can be computed as

$$IE = V_M(\hat{\tau})_m - V_M(\hat{\tau})_{m+1}$$

$$= \sigma^2 \gamma_s^T \left[ \Phi_{sm} A_{sm}^{-1} \Phi_{sm}^T - \Phi_{s(m+1)} A_{s(m+1)}^{-1} \Phi_{s(m+1)}^T \right] \gamma_{\bar{s}}.$$

The subscripts $m$ and $(m+1)$, are used to denote 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. From Appendix using Equation (53), we get

$$IE = V_M(\hat{\tau})_m - V_M(\hat{\tau})_{m+1}$$

$$= \sigma^2 \gamma_s^T \left[ \Phi_{sm} A_{sm}^{-1} \Phi_{sm}^T - \Phi_{s(m+1)} A_{s(m+1)}^{-1} \Phi_{s(m+1)}^T \right] \gamma_{\bar{s}}$$

(45)
\[ IE = \frac{1}{\bar{\sigma}^2} \gamma s^T \left[ \phi_{s(m+1)}^T \Phi_{sm}^{-1} \Phi_{sm}^T + \Phi_{sm}^{-1} \Phi_{sm} \phi_{s(m+1)} \phi_{s(m+1)}^T - \phi_{s(m+1)}^T \phi_{s(m+1)} \right] \gamma s, \]

where the vector \( \phi_{s(m+1)} \) shows the \((M+1)\)th column of the basis function matrix \( \Phi_{sm} \). The positive increase in efficiency i.e. \( IE > 0 \) indicates that using an additional basis function decrease the variance of prediction error. This can also be converted to a ratio as

\[ IE_R = \frac{\gamma s^T \left[ \phi_{s(m+1)}^T \phi_{s(m+1)} + \Phi_{sm}^{-1} \Phi_{sm} \phi_{s(m+1)} \phi_{s(m+1)}^T - \phi_{s(m+1)}^T \phi_{s(m+1)} \right] \gamma s}{\gamma s^T \left[ \phi_{sm}^{-1} \phi_{sm} + \Phi_{sm}^{-1} \Phi_{sm} \phi_{s(m+1)} \phi_{s(m+1)}^T - \phi_{s(m+1)}^T \phi_{s(m+1)} \right] \gamma s} \]

The index \( IE_R \) measures the relative increase in efficiency on using 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 don’t have a known value of the variance of the study variable in advance. Then different estimates obtained in Section 5 section are used. Since the estimates involve the basis function matrix, the use of additional basis function also effects 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. Secondly, we can jointly compute the estimated prediction variance of \( \hat{\tau}(y) \) instead of its population counterpart. Third option is to use Equation (47) and separately obtain the estimates of \( \sigma^2 \) through incremental operators given in [67]. The third option although provides a variance expression for the model with \((M+1)\) predictors without recomputing the regression, it does not provide a comparison among two models (i.e. model with \( M \) basis functions and the model with \((M+1)\) basis functions).

\subsection{6.2 Model Selection Under Regularized Regression}

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

\[ |B_M(\hat{\tau}_{ridge}(y)) - B_M(\hat{\tau}_{ridge}(y))|_{m+1} = V_M(e(\hat{\tau}_{ridge}))|_{m+1} = \gamma s^T \left[ \phi_{sm}^{-1} \phi_{sm} \phi_{s(m+1)} \phi_{s(m+1)}^T - \phi_{s(m+1)}^T \phi_{s(m+1)} \right] \gamma s, \]

where \( \beta_{m+1} \) is the \((M+1)\)th component of the vector \( \beta_{m+1} \) i.e. the effect of additional basis function on the response. A smaller amount of increase in bias means that the additional variable is not effecting the bias of the estimator for a particular value of the ridge parameter. When different amount of regularizations are used for each superpopulation parameter then the amount of increase can not be computed with this formula. Now the increase in efficiency of the ridge regression estimator on using additional basis function can be expressed as follow

\[ I_{E_{ridge}} = V_M(e(\hat{\tau}_{ridge}))|_{m+1} - V_M(e(\hat{\tau}_{ridge}))|_{m+1} = IE - \sigma^2 \gamma s^T \left[ \phi_{sm}^{-1} \phi_{sm} \phi_{s(m+1)} \phi_{s(m+1)}^T - \phi_{s(m+1)}^T \phi_{s(m+1)} \right] \gamma s, \]

where

\[ \sigma^2 \gamma s^T \left[ \phi_{sm}^{-1} \phi_{sm} \phi_{s(m+1)} \phi_{s(m+1)}^T - \phi_{s(m+1)}^T \phi_{s(m+1)} \right] \gamma s = \left( \phi_{s(m+1)}^T Q_{21}^{-2} \phi_{s(m+1)} + \phi_{s(m+1)}^T Q_{12}^{-2} \phi_{s(m+1)} + \phi_{s(m+1)}^T Q_{22}^{-2} \phi_{s(m+1)} \right) \gamma s, \]

and \( \Delta^* = \left( \phi_{s(m+1)}^T Q_{21}^{-2} \phi_{s(m+1)} + \phi_{s(m+1)}^T Q_{12}^{-2} \phi_{s(m+1)} + \phi_{s(m+1)}^T Q_{22}^{-2} \phi_{s(m+1)} \right) \gamma s \) is the elements of the matrix \( Q_{21}^{-2} \), \( Q_{12}^{-2} \) and \( Q_{22}^{-2} \) are the elements of the matrix \( Q_{21}^{-2} \) and are defined in Appendix A. Computation of \( I_{E_{ridge}} \) is not straightforward still some algebraic treatment on matrices can provide a compact form that can be solved numerically. A positive value of the index \( I_{E_{ridge}} \) provides evidence of efficiency improvement by adding additional basis function to the superpopulation model.
7 Simulations

Two simulation studies (one simulated and one bootstrapped) are conducted to evaluate the error variance of the proposed estimator of $\tau(y)$ ($\gamma_i = 1$ for all $i \in U$) 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). While the bootstrap study includes a real data set to perform repeated sampling to obtain design-based properties of the estimator and estimated error variance of the estimator ($\hat{\tau}(y)$).

Both MC simulation and bootstrapping are performed in the widely used statistical software R (version 4.0.1). The simulation steps are described as below:

(i). To constitute a population showing non-linear behavior, draw two independent vectors $u^*$ and $v^*$ of length $N = 1000$ each with uniform $(0,1)$. The variable $x$ 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 deal 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 $df(y,x,x^2,x^3,\ldots,x^{M-1})$ (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 ($\beta$ and $\sigma^2$). The estimated values of $\sigma^2$ are obtained using different formula discussed in Section 5.

(iii). Further, we evaluate the proposed estimator of $\tau(y)$ (with $\gamma_i = 1$ for all $i \in U$) and the estimated variance of $\hat{\tau}(y)$ under different formula 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 hospital data given in [8, Appendix B, Page 424] as our population. The number of beds ($x$) in 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 estimator and estimated error variance. Expected squared prediction error (ESPE) are obtained as:

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

where the symbol $\sum_{sim}$ is used to denote that summation is taken over all 30,000 simulated samples. The ESPE are obtained under regularization taking $v = 0, 1, 5, 10$ with $v = 0$ representing the case of no regularization. Further, the design expectation of estimated variance of $\hat{\tau}$ are obtained through respective formulae after averaging over all selected samples. We use polynomial basis functions of different orders with intercept (the linear population model) and without intercept (the proportional population model). 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 also plotted to assess the pattern in 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 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 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 $10^3$ for simplicity. For real data, results are obtained for HPM ($M = 1$), LPM ($M = 2$), the quadratic model ($M = 3$) and the higher order polynomial (cubic) model ($M = 4$). The estimated error variances of $\hat{\tau}(y)$ are obtained from Equation (35) based on different estimators of $\sigma^2$. Note that all results given in Tables 1–4 are provided for ridge regression estimator with certain choices of $v$ as the variance estimator given in $n$ Equation (34) is a special case of variance estimator in Equation (35) with $v = 0$. The ESPE for different combinations of $M$, $v$ and $n$ is enlisted in third column of Tables 1–2. For simulated data in Table 1, smallest ESPE is observed at $M = 6$ when sample size is taken 100. It turns smallest at $M = 3$ when sample size is set to 200. ESPE for simulated data also tends to decrease with increase in $v$. For example, for $n = 200$ and $M = 1$, the ESPE for $v = 0$ is 53.696 while it is 50.215 for $v = 10$. Similarly, for simulated data in Table 2 (i.e. for the models without intercept), smallest ESPE is observed at $M = 6$ when sample size is taken 100. It turns smallest at $M = 4$ for both choices of sample size (i.e. $n = 100, 200$). Further, ESPE for simulated data also 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 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 of the fact that $v$ at one side decreases variance and increases bias on the other side. When increase in bias is dominated the ESPE tends to increase with increase in $v$ and Vice versa. From all tables one can distill that the ESPE goes down as $n$ increases.

The estimated variance of prediction error of $\hat{\tau}(y)$ is obtained in columns 4-8 under 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 $v = 0$, the estimated variances are 1934.051, 2057.501, 2188.831 and 2416.991 which satisfy the inequality given in Equation (44). Tables 1–4 also provide the evidence that estimated variances decrease by increasing amount of regularization. Similar statement can be made for the relation between estimated variances and sample size. Among alternative variance estimators one must choose the one which is nearer to the true variance in prediction error. Hence, the choice of variance estimator depends on the statistical properties like unbiased 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 as our goal was basically the construction of estimator for $\tau$ and discussing problems associated with its estimation and inference about the finite population in model based setting.

8 Conclusion

A general framework of model-based approach for estimation of finite population parameter $\tau$ (a linear combination of population values), assuming superpopulation setting, is discussed. Some special cases of the proposed general framework are deducted to observe its applicability. Expressions for prediction error variance and model-bias of the proposed estimator are derived. For statistical inference about $\tau$, estimation of prediction error variance under residual, GCV, UEV, FPE and BIC methods (the widely used feature selection criteria in ML) are also considered. The variance introduced under UEV provides minimum variance estimates than all other competing estimators with maximum at BIC. Model selection for finite population parameter under the proposed general framework is also discussed using incremental operators under matrix approach. The model selection is based on a measure, named as increment in efficiency $IE$ which provide guideline for selecting a model with appropriate number of basis function. Positive value of $IE$ shows increase in efficiency while adding additional basis functions to the feature matrix. Further ill-conditioning of the regression estimation is also coped with typical regularization method which introduce slight bias in estimates of $\beta$’s but provide 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 in estimation of any linear combination of population values, hence many finite population parameters can be estimated using this general framework. The proposed model-based framework can be extended to other sampling designs, multi-level models and small area estimation. Working with mixed model for estimating 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 lead to a substantial improvement in 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, pp. 69–74 (1964).

[3] Upadhyaya, L. N. and Singh, H. P. (1999). “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 multi-stage 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. (2000). ”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. (2017). ”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. (1964). "Smooth regression analysis". *Sankhya: The Indian Journal of Statistics, Series A*, 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), pp. 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., and Arcos, A. (2014). "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 trade-off 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):893–912 (2004).

[27] Falorsi, P. D. and Righi, P. " A unified approach for defining optimal multivariate and multi-domains 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. (2018). "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., and Azar, A. "A new approach to evaluating entrepreneurial opportunities". *Journal of Small Business and Enterprise Development* (2019).

[31] Kumar, S., Sisodia, B., Singh, D., and Basak, P. (2017). "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*, pages 83–93.

[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 non-response". PloS One, 14(10):e0222701 (2019a).

[35] Ahmed, S. and Shabbir, J. (2019b). "On use of ranked set sampling for estimating super-population total: Gamma population model". Scientia Iranica DOI10.24200/SCI.2019.50976.1946.

[36] Liu, C., Li, H.-C., Fu, K., Zhang, F., Datcu, M., and Emery, W. J. "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, 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, pp. 1–20 (2020).

[39] Royall, R. M. (1976). "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", 37. OUP Oxford, (2012).

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

[44] Krishnaiah PR. 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., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V. (1997). "Comparing support vector machines with gaussian kernels to radial basis function classifiers". IEEE transactions on Signal Processing, 45(11), pp. 2758–2765.

[49] Buhmann, M. D. "Radial basis functions: theory and implementations", volume 12. Cambridge university press (2003)

[50] Biancolini, M. E. (2017). "Fast radial basis functions for engineering applications". Springer.

[51] Tikhonov, A. N. and Arsenin, V. I. "Solutions of ill-posed problems", volume 14. Vi Winston, (1977). Technometrics, 12(1), pp. 55–67 (1970).

[52] Press, Teukolsky, and S. Teukolsky. "Vetterling, and Flannery". Numerical Recipes in C, 18, pp. 656–680, (1992).

[53] 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).
Appendix

The projection matrix $P$ based on $M$ covariates is defined by

$$ P = I_M - \Phi_s Q_{sm}^{-1} \Phi_s^T, \quad (50) $$

where $Q_{sm}^{-1} = \Phi_s^T \Phi_s + v I_M$ is the Hessian Matrix \[68\] based on $\Phi_s$ with $M$ basis functions. Before applying the incremental operator to $P_m$, we find $A_{s(m+1)}^{-1}$ using $A_{sm}^{-1}$. We use following two useful lemmas from \[69\] for the purpose of matrix inversion

**Lemma 1** For a any partitioned square matrix $B$ defined by

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

$$ B^{-1} = \begin{bmatrix} B_{11}^{-1} + B_{11}^{-1} B_{12} \Delta^{-1} B_{21} B_{11}^{-1} & -B_{11}^{-1} B_{12} \Delta^{-1} \\ -\Delta^{-1} A_{21} B_{11}^{-1} & A_{22}^{-1} \end{bmatrix}, $$

where $\Delta = B_{22} - B_{21} B_{11}^{-1} B_{12}$.

**Lemma 2**
Let the inverse of matrix $B_0^{-1} \in \mathbb{R}^{m \times m}$, $X, Y^T \in \mathbb{R}^{m \times r}$ and $R \in \mathbb{R}^{r \times r}$ all are known. For computing the inverse of a new matrix $B_1$ such that

$$B_1 = B_0 + XRY.$$

To compute the inverse of new matrix $B_1$, we have following relation

$$B_1^{-1} = B_0^{-1} - B_0^{-1}X(YB_0^{-1}X + R^{-1})^{-1}YB_0^{-1}$$

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

### 8.1 Decrement in Variance

The Hessian matrix after adding an additional basis function

$$Q_{m+1} = \Phi_{s(m+1)}^T \Phi_s(m + 1) + vI_{m+1} = Q_{sm} + \Phi_{s(m+1)}^T \Phi_{sm} + \Phi_{s(m+1)}^T \Phi_{s(m+1)} + v_{m+1}$$

where $\Phi_{s(m+1)} = [ \Phi_{sm} \Phi_{s(m+1)} ]$ and $Q_{sm} = \Phi_{sm} \Phi_{sm} + vI_m$. To obtain inverse of the Hessian matrix given in Equation (51) we use Lemma 1, with

$$Q_{s(m+1)}^{-1} = [ \begin{array}{cc} Q_{sm}^{-1} & 0 \\ 0 & 0 \end{array} ] + \Delta^{-1} [ \begin{array}{ccc} Q_{sm}^{-1} & \Phi_{s(m+1)}^T & -Q_{sm}^{-1} \\ \Phi_{sm} & \Phi_{s(m+1)} & 0 \\ -Q_{sm}^{-1} & 0 & 1 \end{array} ]$$

where $0$ is an $m \times 1$ null vector. Further

$$\phi_{r(m+1)}Q_{r(m+1)}^{-1} \phi_{r(m+1)}^T = [ \begin{array}{c} \phi_{rm} \end{array} ] [ \begin{array}{ccc} Q_{11}^{-1} & Q_{12}^{-1} \\ Q_{21}^{-1} & Q_{22}^{-1} \end{array} ] [ \begin{array}{c} \phi_{rm} \end{array} ]$$

$$= [ \begin{array}{ccc} \phi_{rm}Q_{11}^{-1} + \phi_{r(m+1)}Q_{21}^{-1} & \phi_{rm}Q_{12}^{-1} + \phi_{r(m+1)}Q_{22}^{-1} \\ \phi_{rm} \phi_{r(m+1)} & \phi_{rm} \phi_{r(m+1)} \end{array} ] [ \begin{array}{c} \phi_{rm} \end{array} ]$$

$$= \phi_{rm}Q_{11}^{-1} \phi_{rm}^T + \phi_{r(m+1)}Q_{21}^{-1} \phi_{r(m+1)}^T + \phi_{rm}Q_{12}^{-1} \phi_{r(m+1)}^T + \phi_{r(m+1)}Q_{22}^{-1} \phi_{r(m+1)}^T$$

where $Q_{11}^{-1} = Q_{sm}^{-1} + \Delta^{-1}Q_{sm}^{-1} \Phi_{s(m+1)}^T \Phi_{s(m+1)} \Phi_{sm} Q_{sm}^{-1}$, $Q_{12}^{-1} = -\Delta^{-1}Q_{sm}^{-1} \Phi_{sm} \Phi_{s(m+1)}$, $Q_{21}^{-1} = -\Delta^{-1} \Phi_{s(m+1)}^T \Phi_{sm} Q_{sm}^{-1}$ and $Q_{22}^{-1} = \Delta^{-1}$. 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 $Q_{sm}^{-1} = A_{sm}^{-1}$ for $m = 1, 2, ..., M$. For prediction models with no regularization, we have the variance of $V_M(\hat{\tau})$ with $M$ regressors

$$V_M(\hat{x}) = (N - n)\sigma^2 + \sigma^2 [\gamma_x^T \Phi_{rm} A_{sm}^{-1} \Phi_{rm}^T \gamma_x].$$

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

$$V_M(\hat{\tau})_{m+1} = (N - n)\sigma^2 + \sigma^2 [\gamma_x^T \Phi_{r(m+1)} A_{sm}^{-1} \Phi_{r(m+1)}^T \gamma_x].$$

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

$$B_M(\hat{\tau}_{ridge}(y)) = -v_T \Phi_{rm} Q_{sm}^{-1} \beta_m.$$

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

$$B_{M+1}(\hat{\tau}_{ridge}(y)) = -v_T \Phi_{rm} Q_{sm}^{-1} \beta_m + \frac{1}{\Delta} \Phi_{rm} Q_{sm}^{-1} \Phi_{sm} \Phi_{s(m+1)} \Phi_{sm} Q_{sm}^{-1} \beta_m + \phi_{r(m+1)}Q_{21}^{-1} \beta_m + \phi_{r(m+1)}Q_{22}^{-1} \beta_{m+1}.$$
List of Figures
Figure 1: Scatter plot for 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 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 sample selected from 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 sample selected from 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)

List of Tables
Table 1: Simulated ESPE and expected estimated variances of the proposed estimator under linear BFR model.
Table 2: Simulated ESPE and expected estimated variances of the proposed estimator under proportional BFR model.
Table 3: ESPE and expected estimated variances of the proposed estimator under linear BFR model for Hospital population.
Table 4: ESPE and expected estimated variances of the proposed estimator under proportional BFR model for Hospital population.
Figure 2:

Figure 3:
Figure 4:
| $n$ | $v$ | ESPE | Residual | UEV | FPE | GCV | BIC |
|-----|-----|------|----------|-----|-----|-----|-----|
|     |     |      |          |     |     |     |     |
| $M = 1$ |     |      |          |     |     |     |     |
| 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 |     |
| 100 | 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 |     |
|     |      |      |          |     |     |     |     |
| 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 |     |
| 200 | 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  |     |
|     |      |      |          |     |     |     |     |
| $M = 2$ |     |      |          |     |     |     |     |
| 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 |     |
| 100 | 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 |     |
|     |      |      |          |     |     |     |     |
| 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 |     |
| 200 | 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 |     |
|     |      |      |          |     |     |     |     |
| $M = 3$ |     |      |          |     |     |     |     |
| 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 |     |
| 100 | 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 |     |
|     |      |      |          |     |     |     |     |
| 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 |     |
| 200 | 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 |     |
|     |      |      |          |     |     |     |     |
| $M = 6$ |     |      |          |     |     |     |     |
| 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 |     |
| 100 | 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 |     |
|     |      |      |          |     |     |     |     |
| 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 |     |
| 200 | 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 1: Caption Table 1
| $n$ | $v$ | ESPE | Residual | UEV | FPE | GCV | BIC |
|-----|-----|------|----------|-----|-----|-----|-----|
|     |     |      |          |     |     |     |     |
|     |     |      |          |     |     |     |     |
|     |     |      |          |     |     |     |     |

Table 2: Caption Table 2
| $n$ | $v$ | ESPE  | Residual | UEV    | FPE    | GCV    | BIC    |
|-----|-----|-------|----------|--------|--------|--------|--------|
|     |     |       |          |        |        |        |        |
|     |     | $M = 1$ |          |        |        |        |        |
| 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 |
| 50  | 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 |
|     |     | $M = 2$ |          |        |        |        |        |
| 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 |
| 50  | 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 |
|     |     | $M = 3$ |          |        |        |        |        |
| 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 |
| 50  | 10  | 7.0038 | 759.5841 | 796.4001 | 833.2161 | 835.0011 | 903.6091 |
|     |     | $M = 4$ |          |        |        |        |        |
| 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   | 10  | 14.4792 | 816.0691 | 872.8424 | 929.6156 | 933.5662 | 1038.1674 |
|     |     | $M = 5$ |          |        |        |        |        |
| 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   | 10  | 44.7919 | 275.7103 | 285.4503 | 295.1903 | 295.5344 | 320.5647 |
|     |     | $M = 6$ |          |        |        |        |        |
| 0   |     | 45.0972 | 275.0052 | 284.2723 | 293.5394 | 293.8517 | 317.6817 |

Table 3: Caption Table 3
| $n$ | $v$ | ESPE | Residual | UEV  | FPE  | GCV  | BIC   |
|-----|-----|------|----------|------|------|------|-------|
| 0   | 2   | 38865.250 | 271.521 | 277.062 | 282.603 | 282.716 | 293.198 |
| 1   | 2   | 38865.250 | 271.609 | 277.042 | 282.474 | 282.582 | 292.860 |
| 50  | 5   | 38865.250 | 273.426 | 278.489 | 283.553 | 283.647 | 293.234 |
| 10  | 5   | 38865.250 | 277.924 | 282.635 | 287.346 | 287.425 | 296.352 |
| 0   | 1   | 17746.841 | 184.520 | 186.384 | 188.248 | 188.266 | 193.103 |
| 1   | 1   | 17746.841 | 184.535 | 186.380 | 188.226 | 188.244 | 193.033 |
| 100 | 5   | 17746.841 | 184.867 | 186.645 | 188.423 | 188.440 | 193.053 |
| 10  | 5   | 17746.841 | 187.786 | 187.491 | 189.195 | 189.211 | 193.636 |
| 0   | 2   | 1989.648  | 915.590 | 953.740 | 991.889 | 993.479 | 1064.832 |
| 1   | 2   | 1845.946  | 887.065 | 923.259 | 959.452 | 960.929 | 1028.655 |
| 50  | 5   | 1405.980  | 803.258 | 833.744 | 864.230 | 865.387 | 922.520 |
| 10  | 5   | 1051.156  | 739.901 | 766.057 | 792.214 | 793.139 | 842.226 |
| 0   | 1   | 2260.617  | 298.282 | 304.369 | 310.456 | 310.581 | 326.315 |
| 1   | 1   | 2211.905  | 295.136 | 301.097 | 307.059 | 307.179 | 322.590 |
| 100 | 5   | 2039.615  | 284.476 | 290.007 | 295.538 | 295.646 | 309.948 |
| 10  | 5   | 1864.813  | 274.447 | 279.564 | 284.680 | 284.776 | 298.011 |
| 0   | 2   | 1011.154  | 1934.051 | 2057.501 | 2180.951 | 2188.831 | 2416.991 |
| 1   | 2   | 905.876   | 1749.106 | 1857.814 | 1966.522 | 1973.278 | 2174.374 |
| 50  | 5   | 629.460   | 1305.618 | 1380.619 | 1455.619 | 1459.928 | 1599.022 |
| 10  | 5   | 448.607   | 1050.837 | 1107.537 | 1164.237 | 1167.297 | 1272.648 |
| 0   | 1   | 962.666   | 531.844 | 548.293 | 564.742 | 565.251 | 607.594 |
| 1   | 1   | 924.836   | 510.483 | 526.069 | 541.655 | 542.130 | 582.258 |
| 100 | 5   | 802.672   | 445.061 | 458.060 | 471.060 | 471.439 | 504.926 |
| 10  | 5   | 694.960   | 392.290 | 403.265 | 414.239 | 414.546 | 442.830 |

Table 4: Caption Table 4