Model-Based Estimation of Population Parameters with Applications

DOCTOR OF PHILOSOPHY

IN

STATISTICS

BY

SHAKEEL AHMED

DEPARTMENT OF STATISTICS

FACULTY OF NATURAL SCIENCES

QUAID-I-AZAM UNIVERSITY, ISLAMABAD

2020

In The Name of Allah The Most Beneficent The Most Merciful

Model-Based Estimation of Population Parameters with Applications

By

Shakeel Ahmed

A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN STATISTICS

Supervised By

Prof. Dr. Javid Shabbir

Department of Statistics
Faculty of Natural Sciences
Quaid-i-Azam University, Islamabad
2020

CERTIFICATE

Model-Based Estimation of Population Parameters with Applications

by

SHAKEEL AHMED

(Reg.No.03221613001)

A thesis submitted in the partial fulfillment of the requirements for the degree of the Doctor of Philosophy

We accept this dissertation as conforming to the required standard

Dr. Muhammad Zakria
(External Examiner)

Dr. Muhammad Hanif
(External Examiner)

Prof. Dr. Javid Shabbir
(Supervisor)

3.04/20

Dr. Ijaz Hussain
(Chairman)

DEPARTMENT OF STATISTICS
QUAID-I-AZAM UNIVERSITY
ISLAMABAD, PAKISTAN
2020

QUAID-I-AZAM UNIVERSITY
DEPARTMENT OF STATISTICS

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in this thesis, entitled "**Model-Based Estimation of Population Parameters with Applications**" was conducted by Shakeel Ahmed under supervision of **Prof. Dr. Javid Shabbir** No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to QUAID-I-AZAM UNIVERSITY ISLAMABAD in partial fulfillment of the requirements for the degree of Doctor of Philosophy in field of STATISTICS in the DEPARTMENT OF STATISTICS, QUAID-I-AZAM UNIVERSITY ISLAMABAD.

Name & Sign of Student **Shakeel Ahmed**

1. **Name & address of External Examiner**

Dr. Muhammad Zakria
Associate Professor
Department of Statistics,
Allama Iqbal Open University, Islamabad.

2. **Name & address of External Examiner**

Dr. Muhammad Hanif
Associate Professor and Chairman
Department of Mathematics and Statistics,
PMAS-Arid Agriculture University, Rawalpindi

3. **Name & address of Supervisor**

Prof. Dr. Javid Shabbir 30/4/2020
Department of Statistics,
Quaid-i-Azam University, Islamabad.

Prof. Dr. Muhammad Saddiq
Dean, Faculty of Natural Sciences

Dr. Ijaz Hussain 30/04/20
Chairman

AUTHOR'S DECLARATION

I **SHAKEEL AHMED** hereby state that my PhD thesis titled **Model-based estimation of population parameter with applications** is my own work and has not been submitted previously by me for taking any degree from **Quaid-I-Azam University Islamabad** or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my Graduate the university has the right to withdraw my PhD degree.

SHAKEEL AHMED
Reg. No. 03221613001
Date: 30-04-2020

Plagiarism Undertaking

I solemnly declare that the research work presented in the thesis titled **"Model-Based Estimation of Population Parameters with Applications "** is solely my research work with no significant contribution from any other person. Small contributions/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Quaid-I-Azam University Islamabad toward plagiarism. Therefore I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of PhD degree, the University reserves the rights to withdraw/revoke my PhD degree and that HEC and the University has the right to publish my name on the HEC/University Website on which names of students are placed who submitted plagiarized thesis.

Sh

Shakeel Ahmed

Reg. No. 03221613001

Date: 30-04-2020

DEDICATED TO

My father, who has always been a source of
Inspiration, Zeal and Strength for me.
My mother, whose prayers brought me to
this stage
All members of my family.

ACKNOWLEDGMENT

First and foremost I praise and pay gratitude to Allah Almighty, the Lord and creator of the Universe. All respect and gratitude goes to the Holy Prophet Muhammad(PBUH) and his family (Aal-e-Muhammad) whos enlighten our hearts with the light of Islam and whose way of life has been a continuous guidance for us.

I deem it my utmost pleasure to avail this opportunity to convey my gratitude and deep sense of obligation to my respected supervisor Prof. Dr. Javid Shabbir for his guidance, scholarly criticism, kind behavior and moral support throughout this research. I also wish to express my sincere thanks to Dr. Ijaz Hussain, Chairman, Department of Statistics and other respected teachers for their help and guidance during my course and research work at the department. Further, I want to pay my sincere gratitude to Prof. Sat Gupta for his scholarly guidance in my research works.

My sincere thanks goes to my parents for their love and support throughout my life without their support, prayers and encouragement it would have been impossible for me to complete this work. Special credit goes to my beloved father Mr. Hassan whose continuous effort and support brought me to this stage. Special thanks to my brother Mr. Sajid Ali and my sisters for their encouragement and prayers. I would like to thank all my seniors for their cooperation especially Muhammad Aslam, Akbar Ali, Farooq Shah and colleagues Umair Sohail, Saima khan and Fakhra Batool. Finally, I would say thanks to all who have been supportive and cooperative to me during my research work.

SHAKEEL AHMED

LIST OF PUBLICATIONS

Ahmed, S., & Shabbir, J. (2019). On use of Ranked Set Sampling for estimating Super-population Total: Gamma Population Model. *Scientia Iranica*, DOI: 10.24200/sci.2019.50976.1946.

Ahmed, S., & Shabbir, J. (2019). Model based estimation of population total in presence of non-ignorable non-response. , *PloS one*, 14(10):e0222701.

Ahmed, S., & Shabbir, J. (2019).Model-based estimation of finite population parameter: A general basis function approach (Submitted).

Ahmed, S., & Shabbir, J. (2019).Finite population parameter estimation
under Bayesian basis function regression (Submitted).

Ahmed, S., & Shabbir, J. (2019).Domain estimation under ranked set sampling without replacement (Submitted).

Ahmed, S., & Shabbir, J. (2019).Model-based Estimation of Fertility Rates: Analysis on PDHS 2017-18 Data. (Submitted).

ABSTRACT

The problem of finite population parameter estimation, in superpopulation settings, is receiving considerable attention in the field of survey sampling. In this dissertation, we develop a general framework of model-based approach for estimation of finite population parameter τ (a linear combination of population values), assuming superpopulation setting under basis function regression model. Bayesian version of the proposed general framework is also studied by assuming the Gaussian distribution for the error term, for incorporating prior information about the superpopulation parameters. 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 $\hat{\tau}$ are derived. For statistical inference about τ, estimation of prediction error variance under different model selection criteria residual, generalized cross validation (GCV), unbiased estimated variance (UEV), final prediction error (FPE) and Bayesian information criteria (BIC) methods, is also considered. An index for increase in efficiency on using additional basis functions, named as increment in efficiency (IE), is also devised under, simple, ridge and Bayesian regression. The index provides a logical guideline for selecting a model with appropriate number of basis functions or covariates. Non-response problem in the study variable is dealt based on sub-sampling technique, known as Hansen and Hurwitz technique, under model-based approach with the assumption that the responding and non-responding population have different models and the occurrence of non-response is observable just like a stratification variable in stratified sampling. Design-based efficiency comparisons are made based on real and simulated data sets. Under linear population model (linear in parameter as well as in variables), the total estimator with sub-sampling is model-unbiased and has smaller model-variance as compared to predictive estimator based on sampled respondents only. We presents new version of ranked set sampling for obtaining more dispersed units with title, the ranked set sampling without replacement (RSSWOR), based on the assumption that the finite population is coming from an infinite superpopulation via some stochastic process with finite mean and variance. Both mathematical expressions and Monte-Carlo experiment support the superiority of the total estimator under RSSWOR over the competitors under simple random sampling without replacement (SRSWOR) for a special model, so called, gamma population model (GPM). Estimation of sub-population total under a new version of ranked set sampling for obtaining a without replacement sample with GPM (general form of proportional population model) is also provided. Further, the model relationship between the study variable and the auxiliary variable for whole population is used to predict the non-sampled values to establish a domain specific estimator for total. The superiority of the domain specific total estimator under RSSWOR over the total estimator under SRSWOR for specific cases are also shown mathematically as well as through Monte-Carlo experiment.

Finally, we analyze the birth history data from Pakistan Demographic Health Survey

(PDHS) 2017-18 using three separate models taking 1-year period births for first, 3-years period birth for second and 5-years period births for third models as the responses and 24 regressors instead of basis function of single regressor. Mainly, the Poisson regression model with log-link function is used for modeling purpose. To deal with responses having large variance and many 0's as well as a few very large values, we use negative binomial (NB), zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB) models as extensions of Poisson model. We also conduct the estimation of regression models under Bayesian paradigm assuming normal priors for each coefficients including intercept. The posterior means are obtained using rjags (R Just Another Gibbs Sampler) package in R Statistical software. The posterior means for each coefficients are observed closer to classical estimates for Poisson models. Some model diagnostics are applied to check the validity of estimation procedure. The model-based fertility rates i.e. age specific fertility rate (ASFR), total fertility rate (TFR), general fertility rate (GFR) and gross reproduction rate (GRR) are obtained using predicted response under the estimated models. We provide an illustration of predictive approach through bootstrap sampling from the PDHS 2017-18 individual recode data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Review of Literature

1.1 Background of the Study

The basic goal of survey sampling is to obtain the accurate estimates of totals, means, ratios and proportions etc. for the characteristics of interest in a finite population. In classical or design-based inference, the finite population is considered as fixed with known values corresponding to each unit. For statistical inference, randomization is achieved from random sampling mechanism used for collecting data. While the model-based approach assumes that the finite population values are the realization of some stochastic process with fixed but unknown parameters. The mechanism of randomness arise from these stochastic variables and the variable follows some models known as superpopulation. In modern era, we can relate the model-based approach with supervised learning for predicting the values of non-sampled observation. The current research is a part of such effort which explicitly introduced a generalize frame work for obtaining model-based for finite population parameters with minimum prediction error. Before going in depth literature, we introduce the basic terminologies being used in the study in upcoming section.

1.2 Definitions of Terminologies

This section gives definitions to some useful terminologies which are being used in the dissertation.

1.2.1 Design-Based Estimation

Design-based method of estimation uses the randomization incurred during selection of units from the population where the values for the units in the population are considered to be fixed. It is assumed that the variation in the estimates arises from the fact that the estimates are based on a random sample (probability sample) drawn rather than non-random sample (non-probability sample) from a finite population. Design-based estimation include construction of estimators without considering the underlying model relation ship between the variables. The properties of estimators are than studied via sampling distribution resulted from repeated sampling of same size under similar conditions. The estimate are usually constructed for population quantities, such as totals, means, proportions, or ratios etc. of the survey variable (usually known as study variable).

1.2.2 Model-Based Estimation

As name suggest, the model-based approach assumes values on population units are the realization of stochastic variables with specified lower order moments (mean which is typically unknown and variance assumed to be known). These stochastic variables depend on the auxiliary data and random error terms. Model-based supporters believe that variation in estimates is resulted by the noise in the model (error term). Hence properties of estimates in model-based approach are studied using the distribution of error term rather than sampling distribution. A model is typically constructed by expressing some dependent variables as a function of some covariates (independent variables). This functional relationship is then utilized for estimation of finite population quantities such as mean, total, proportion and ratio etc. The model relationship is considered for estimation of missing values in the response variable variables at unit level as well as at aggregated or cluster levels. More details about model-based estimation and inference can be found from upcoming chapters.

1.2.3 Basis Functions

In many cases, we are left with a little information about the underlying nature of the phenomenon being modeled. Then, we typically jump to a few models in machine-learning that are broadly-used and much effective for many problems. These include basis function regression (including polynomial, Radial, and Gaussian basis functions), Artificial Neural

Networks (ANN), and k-Nearest Neighbors (KNN). In basis function regression, we need to model our responses (say y) that depends on some underlying function (say $\phi(x)$) of independent variables (x) such that for some non-sampled values of the predictors we'll be able to accurately predict the future values of the responses. The function $\phi(x)$ is called the basis function and the resulting model is known as basis function model. There are several methods for constructing basis function models, all of which are based on particular assumptions. Basis functions allow us to opt for an elastic model including different functions of a single covariate or two covariates including their interactions. For example, in modeling births we may use age of mother as covariate and basis functions can be generated as age and squared age etc. The polynomial basis function based on covariate age is then used to model the birth data and birth specific rates are computed using the developed model. Other basis function such as Guassian basis function are widely used in Bayesian regression and the radial basis function are used in collection of geographical samples. The basis functions models allow us to use non-linear basis functions of predictors to obtain linear predicted responses. In practice there are many other possible choices for basis functions, including sinusoidal functions, polynomials, basis functions from different families, such as monomials and radial basis functions (RBFs) etc. In general one ideally wants to choose a family of basis functions in order to get a good fit to the responses with a small basis set. So that the number of weights to be estimated is not too large. Polynomial and Gaussian basis functions and their use in estimation theory will be demonstrated in Chapter 2.

1.2.4 Superpopulation Models

As we know, a finite population consists of units whose values are considered to be known and fixed. Hence, the population quantities for finite population are fixed and known if a census is conducted. However, there may exists another type of statistical object associated to the values that constitute a less well defined finite population. This is the statistical model for these values, often known as a superpopulation model. In this study, a statistical model for the population is defined as a specification of the statistical properties of the population values of the variables of interest. In some situations the model may be exactly specified, in the sense that it identifies a stochastic process explicitly that generated the population values. In general, such models are usually rather weakly specified, i.e. it specify only first and second order moments of the population distribution of the survey variables and higher order moments are often unspecified. In all cases there will be some parameters associated with the model specification (known as superpopulation parameters) whose values are unknown. For example, let the values $y_1, y_2, ..., y_N$ be the N independent and identically distributed (IID) realizations of a random variable with mean μ and variance σ^2. In this case, the mean and variance are hypothetical constructs that could never be obtained exactly even if a census of the same population was carried out. Hence μ and σ^2 are the parameters of this superpopulation model. In model-based approach, one should first estimate superpopulation parameters for estimating the model before going to estimate the quantity of interest which is the finite population parameters, where the finite population is considered as a single realization of superpopulation.

1.2.5 Non-Informative Sample

The basic assumption of model-based estimation that allows generalization of the estimated statistical model from sampled population for prediction of non-sampled values of the response variable is the non-informative sampling. More generally, a method of sampling is said to be non-informative for inference about a superpopulation parameters for a variable if the same superpopulation model also satisfies for the values of this variable in the sample, i.e. we can make valid statistical inferences about the superpopulation parameters after fitting the superpopulation model to the sample data. This is equivalent to say that the conditional distributions for the study variable y given some covariates whose values are assumed to be known for all population units are same for sample and non-sampled parts of the population.

1.2.6 Ranked Set Sampling

Ranked set sampling is a data collection mechanism based on random sampling from an infinite or finite population after ranking observations via some inexpensive ranking mechanism. The method works by selecting initial samples of relatively smaller size and rank them within themselves according to certain ranking mechanism. Then smallest ranked unit is observed from 1st set, second smallest from second set and so on. The process stops after observing largest unit from last set. The whole process can be repeated a specified number of cycles to obtain the required number of observations. The ranked set sampling is preferred only when: (i) ranking small groups of unit are easy and economical and (ii) taking measurement from larger samples are expensive and time consuming. Despite of these limitations, ranked set sampling is preferred over simple random sampling (when it is applicable) due to attractive design-based properties. In this study, we cover application of ranked set sampling to model-based approach as well as small area estimation. The details about proposed modification and their application can be found in respective chapters.

1.2.7 Small Area Estimation

The term "small area", generally, refers to a small geographical area or domain such as a county or state also called “small domain", i.e. a particular sub-part of an area. If a survey is carried out for the whole population (for example, a state-wide or nation-wide survey), the sample size within any particular area may be too small to provide accurate estimates from the data at hand. Small area estimation is one of the statistical techniques involving the estimation of parameters for such small sub-populations. The technique is used when the sub-population of interest is included in a larger survey. Statistical models are widely used to obtain small area level estimates. The detail discussion of small domain estimation under different population models for single regressor will be provided in Chapter 6.

1.2.8 Bayesian Prediction

The controversy of using Bayesian and Frequentist frameworks in statistical analysis is one of the most important academic discussion that Statisticians engaged in. Rather than blindly jumping into one side, one should to learn both methods analysis and apply them where seem appropriate. In this way, recently, Bayesian method of estimation and inference have been extensively used. One of the areas to focus in applied Bayesian inference is Bayesian linear modeling. The most important aspect of the Bayesian learning process is explaining a relationship and generalizing it to others, and this study consists of an attempt to use the Bayesian Linear Regression (BLR) for predicting the outcome for non-sampled set. What we result from the frequentist linear regression is an estimate of the model parameters from only the training data set (the sampled data set in our problem). Our model is informed completely by the sampled data: in this way, everything that we need to recognize our model is available in the sampled data. However, if the sample size is small, one might like to express the estimate as a distribution of possible values of the parameter given the sample information. This is the situation where Bayesian Linear Regression is needed.

1.2.9 Age-specific Fertility Rates (ASFR)

The ASFR is the number of births occurred during a given reference period per 1,000 women expose to the risk of fertility, in single year, three-years or five-years age groups. ASFRs are typically calculated by dividing the number of births in a period of three years preceding the survey on the women-years of exposure to fertility in same reference period. It is usually calculated for seven age groups of five years each (i.e. 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, and 45–49). For any age group g, B_g denotes the number of births to women in age group g during the reference period, and E_g denotes the number of women-years of exposure in age group g during the same reference period ($g = 1, 2, ..., 7$ for 5-years grouping and $g = 15, 16, 2, ...49$ for single year grouping). The ASFR in age group g can be expressed as follows:

$$ASFR_g = \frac{B_g}{E_g} \times 1000 \tag{1.2.1}$$

where 1000 is multiplied to show the rate per 1000 women-years of exposure. The information about the exact date of birth (DOB) of child from DHS data are utilized for directly calculating the numerator B_g. For calculating the denominator E_g, the exact date of birth (DOB) of each woman is used for summing up the number of women-years of exposure in age group g, by considering the fact that a woman can participate into two or more age groups in a reference period. For examples and further illustrations about the calculation of the women-years of exposure, readers are referred to Masset (2016).

1.2.10 Total Fertility Rate (TFR)

The TFR measures the women's fertility in a hypothetical way. It can be described as the total number of children who would be born per 1000 woman if they were to pass through their reproductive age according to a given schedule of ASFR subjected to no mortality. In DHS surveys, the TFR is computed on the basis of $ASFR_g$ for $g = 1, 2, ..., 7$ as follows:

$$TFR_g = 5 \times \sum_{g=1}^{7} ASFR_g \tag{1.2.2}$$

1.2.11 General Fertility Rate (GFR)

The GFR is the mean number of children that a woman gives during her whole reproductive period, and obtained by dividing the total number of births during a specified period on the total number of women expose to risk of fertility, during the same specific period. In DHS.rates Package, the GFR is obtained by using following formula:

$$GFR_g = \frac{\sum_{g=1}^{7} B_g}{\sum_{g=7}^{6} E_g} \tag{1.2.3}$$

1.2.12 Gross Reproduction Rate (GRR)

It is the number of daughters a woman expected to have if she lived all of her childbearing age, which is round about to the age of 49. It is computed on the basis of the ASFR and sex ratio at birth during that period. Like TFR, GRR assumes that the hypothetical group (cohort) of women pass from birth over their reproductive age with no mortality hence, not studied here. This assumption is valid when one is interested in comparing levels of fertility over time. Although Net Reproduction Rate (NRR) is a more realistic measure of women's reproduction it need data on women mortality.

$$TFR_g = 5 \times \sum_{g=1}^{7} ASFR_g \times P_g \tag{1.2.4}$$

where P_g is the proportion of female births to the women in age group g

1.3 Literature Review

Researchers, in survey sampling, always favored random sampling for a valid statistical inference due to its attractive long run properties such as unbiasedness and efficiency in design-based sense. They have been ignoring the importance of underlying model relationship between the survey variable and one or more covariate(s) (auxiliary variable(s)) at estimation stage. Without exposing an appropriate model relationship between the survey variable and the available covariates researchers in design-based paradigm have been constructing estimators for the unknown population quantities such as total, mean, variance etc, relying only on randomization mechanism proponed by sampling mechanism. They have been utilizing sample estimates and known population parameters of the auxiliary variable(s) at 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 Cochran (1940), Murthy (1964), Upadhyaya and Singh (1999), Gupta and Shabbir (2008) and Diana et al. (2011).

On contrary, supporters of model-based paradigm emphasis that randomization is a property of error term used in the model (superpopulation model already defined in introduction) hence it is not necessary nor sufficient condition for a rigorous statistical inference (Valliant, 2000). In the model-based framework, initially, Godambe (1955) used a simple 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. Dorfman et al. (1993) and Chambers et al. (1993) have worked on estimating a smooth function and used for predicting the non-sampled values in estimating finite population total. The asymptotic bias of this form of regression estimator of population total does not account the division on the sampling density.

In similar direction, Breidt and Opsomer (2000) worked on a class of estimators based on local polynomial regression models which are weighted linear combinations of the study variables, where the weights were calibrated to control totals which are known. In general, the estimator is robust to bandwidth changes, and provides exact unbiasedness as well as minimal variance for a specific weighted balanced sample. They noticed that the total estimators for population total from a nonparametric regression model provide approximate unbiasedness without imposing restriction on balancing and result near minimal variance. However, Fan (1996) uncovered a more appealing strategy than the kernel regression,e.g, the variable bandwidth LLR approach. Chambers et al. (2003) observed that the calibration estimator based on the columnar model performs slightly better than the best linear unbiased estimator (BLUE) at higher bandwidth. Zheng and Little (2003) proposed a model-based estimator that works with penalized spline regression, and extended the estimator to two-stage sampling (Zheng and Little, 2004). Breidt and Opsomer (2000) used the classical local polynomial regression (CLPR) estimator for estimating the regression function to obtain the model assisted estimator of the total in finite populations. Several partial solutions for balanced sampling are available in Ardilly (1991), Deville (1992), Hedayat and Majumdar (1995) and

Valliant (2000). Chambers et al. (2003) 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. Deville and Tillé (2004) developed the cube method for the selection of approximately balanced samples based on equal or unequal inclusion probabilities with a number of auxiliary variables. Hazlett (2013) developed a method for balancing that equalize the multivariate densities and reduce bias without searching specifications. Sánchez-Borrego et al. (2014) estimated the regression function with mixed variable using a modified form of local constant estimator. Luc (2016) derived properties of weighted nonparametric regression estimator, using probabilities as weight, for complex surveys under combined inference. Falorsi and Righi (2016) 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 than given threshold values. Kikechi et al. (2017) 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 (Nadaraya (1964), Watson (1964)) and shown that the two estimators are asymptotically equally efficient. Clair (2017) considered the nonparametric estimation methods for data analysis in complex surveys. Kikechi et al. (2018) used the LLR technique to asses the properties of the derived estimator and compare its performance with the existing estimators.

Aside from use of the auxiliary information i.e. previous knowledge about the parameters involved in density one can also be incorporated while predicting future values. This strategy is known as Bayesian Prediction (BP). It has many applications in quality control, reliability engineering and biological sciences. A wide range of literature is available regarding predictive inference for future observations. Some of related works are cited as: Aitchison and Dunsmore (1975), Särndal et al. (1978), Sinha (1990), Raqab and Madi (2002), and A Alamm et al. (2007) etc. Fushiki (2011) worked on estimation of prediction error using K-fold cross validation under Bayesian framework. Further, Jochems et al. (2016) developed a predictive model using multiple hospitals data and provide a real life proof of the concept. For a detailed survey on Bayesian predictive estimation readers are referred to Vehtari and Ojanen (2012). Further, Wu et al. (2012) proposed a prediction model for cyber attacks that rely on Bayesian network. Cai et al. (2014) constructed an effective algorithm to build the failure prediction Bayesian network (FPBN) model under data mining technology. They used the conception of FPBN to describe the state of components and system and their cause-effect relationships. Similarly, Nazerfard and Cook (2015) proposed an activity prediction model using Bayesian networks together with a two-step inference process to predict both the next activity features and labels. Minka (2000) derived the posterior and predictive density for linear multivariate regression under Gaussian noise with zero-mean. He contributed to the discussion by giving careful attention on demonstrating evidence based feature selection and illustrating the predictive distribution. Basis function regression, which is the main concern of our research, is also discussed in the note. Recently, Pettit (1986) considered the importance of model diagnostic in choice of Bayesian model and conclude that diagnostics can be justified by approximating the probability of a simple model as the true one. Smith (1986) worked on Bayesian thoughts on modeling and choice of model in detail. More related literature on Bayesian model fitting and their diagnostics using MCMC are provided in Chapter 7.

In statistical investigations, once data collection is completed, one has to bear some, perhaps a considerable amount of non–response. Although a significant resource are employed to improve data collection process to avoid the non-response still about 20% non–response rate is commonly accepted. Non-response has major two categories the item non-response and the unit non-response. Item non-response occurs when one or more questions in the questionnaire are left unanswered during the survey. While a unit non-response occurs when one or more unit(s) do not response at all or are missing. In either cases non–response in sample surveys always leads to non-sampling error in estimation of the population parameters and yields biased estimates which ultimately spoils inference about the population of interest. Such type of bias can not be vanished even for large sample sizes. When non-response occurs completely at random then the best way to deal with is to impute the projected values of the outcome variable corresponding to non-respondents (Yuan, 2000). On contrary, when non-response factor (e.g, age, sex or/and income status etc.) is correlated with the outcome variable then the usual imputation methods fail to cope with the situation. In such situations, the population parameters and the behavior of the population (superpopulation model) may differ among the responding population (respondents) and the responding populations (non-respondent).

There are several approaches for checking whether there is a difference between the behavior of populations of respondents and non-respondents and evaluating potential bias due to non–response: (i) specific follow-up of non-respondents and (ii) analysis of the characteristics of respondents and non-respondents which are known prior to survey. Barton et al. (1980) used demographic information (education, age, employment status, state of residence, field of employment etc.) to compare the respondents and the non-respondents. Information regarding non-respondents may come from previous surveys of same population (in the case of longitudinal surveys or with rotation groups) or by using some external data sources (e.g. administrative data etc.). Wood et al. (2006) suggested a method for adjustment of non-ignorable non-response in studies involving one or more additional attempts to contact initial non-responders. Peytchev et al. (2009) worked on changing in survey estimates as a function of additional calls under the similar protocol as well as under a different protocol. Biemer et al. (2013) considered the use of level-of-effort para-data to model the mechanism of non–response in surveys and for adjusting non–response bias, specially bias that is not missing at random (NMAR) or non-ignorable. The approach was based on unconditional maximum likelihood estimation model that adapted and extended the prior work to cope with the complexities encountered in large-scale surveys.

For similar situation, Knudsen et al. (2010) examined whether non-participation in a census-based health study was related with poorer health status, using the Hordaland Health Study conducted in western Norway in 1997-1999. They aimed to determine whether health problems were over–represented in nonparticipants and to explore the consequences of participation bias on relation between outcomes and exposures. Statistical techniques for dealing with non–ignorable non–response based on a propensity–to–respond score has been developed by Copas and Farewell (1998) assuming both item as well as unit non–response.

Moreover, Moore (2014) proposed an approach of increasing blood supply by collecting blood more frequently from the selected donors for studying the relationship between aging the population and blood transfusion. The primary aim of their proposed "interval trial" was to observe whether donation intervals can be acceptably and safely decreased to optimize blood supply while maintaining the health status of donors. The health status of a cohort of 1991 Gulf War veterans was periodically assessed by Huang and Tai-kang (2009). They compared various health outcomes of veterans with those of their peers in military who were not posted to the Gulf. Another example in which one can make utilization of sub-sampling method can be found in White et al. (2011), where missing data and incomplete randomized interventions were common. These problems complicate the analysis as well as interpretation of controlled randomized trials (CRT), and are rarely handled well in practice. Guan et al. (2018) modeled the non-response probabilities as logistic functions of the survey variable and related covariates in the survey with callback. They proposed maximum likelihood semi-parametric estimators of the parameters in the response probabilities. They further suggested, an efficient estimator for the mean of the study variable using the estimated response probabilities. The method was employed to data taken from the Singapore Life Panel Survey, a survey of health spending utilizing a census-based sample of individuals 50-70 years old, assuming that non-response was related to the health status.

In real surveys, as discussed in above cited works, non-response occurrence is not missing at random (NMAR) or, in other word, it is non-ignorable. When the occurrence of non-response in sample survey is related to the outcome of the survey, a valid statistical inference about the target population is quiet difficult. Among the two possible solutions one prefers to opt the sub-sampling method instead of call back due to resource and time constrain. In this regard, Hansen and Hurwitz (1946) introduced a well known procedure for sub-sampling (follow-up) the non-respondents. The method includes sub-sampling a portion of non-respondents from the first sample with the assumption that some stronger mode of interview is applied for the purpose of sub-sampling non-respondents, consequently, all persons give full response on second call. On the basis of sub-sampling procedure introduced by Hansen and Hurwitz (1946), many authors including Khare and Srivastava (1993, 1995), Khare and Sinha (2009) and Singh and Kumar (2008) worked on mean estimation under designed-based approach ignoring model relationship between the study variable and the known covariates. Ericson (1967) suggested Hansen and Hurwitz Hansen and Hurwitz (1946) type estimator under Bayesian paradigm using squared error loss function (SELF). Later on Smouse (1982) considered Bayesian approach of estimation under a general model using Hansen and Hurwitz (1946) technique. As we already discussed earlier that theory of survey sampling usually focuses on design-based approach, which often based on the probability mechanism that is used for selecting samples. In many occasions a design-based approach does not perform well or at all. For example; (i)in administrative data from incomplete registers or in internet surveys, we can't use probability mechanism for selecting samples as there no appropriate sampling frame exist, (ii) in situations, where the sample size is too small to obtain the reliable estimates. This is particularly the case if the level of detail for which figures must be produced is high, such that the sample size is small in the various sub-populations.

As we already discussed, there are several approaches for handling the problem of non-response in sample literature. A suitable approach may be chosen according to the type of non-response (full or partial), the accessibility of the auxiliary variable(s) and the validity of the underlying response model for handling the problem. In general, re weighting is used to deal with full (non-availability of units) non-response. Imputation is preferably applied for dealing with partial non-response although it can be applied for full non-response if appropriate auxiliary information is available. Re-weighting eliminates or at least reduces total non-response bias (Särndal, 2007; Holt and Elliot, 1991). While the sub-sampling method introduced by Hansen and Hurwitz (1946) provides a good adjustment for non-response bias and yield unbiased estimator for the population mean when the non-response variable R is significantly correlated with the survey outcome. In this study, we develop a model-based estimator for population parameter τ by adjusting non-response using sub-sampling procedure. The detailed discussion on how model-based approach works in finite population parameter estimation in presence of non-ignorable non-response is included in Chapter 4.

In same era, many survey sampling practitioners have worked on improved methods of data collection. Among them Ranked Set Sampling (RSS) technique is a good alternative, in terms of relative efficiency, to Simple Random Sampling (SRS) for obtaining experimental data that are considered as true representative of the population under investigation. This is true across all of the sciences including agricultural, biological, environmental, engineering, physical, medical, and social sciences. Because in RSS, measurements are likely to be more regularly spaced as compared to SRS. The RSS procedure creates stratification of the entire population at the sampling stage i.e. we are randomly select samples from the subpopulations of small, medium and large units without constructing the subpopulations (strata) in advance. Ranked set sampling method, proposed originally by McIntyre (1952) to estimate mean pasture yields, has recently been modified by many researchers to estimate the population parameters with improved efficiency and applicability. Dell and Clutter (1972) showed that the sample mean is an unbiased estimator of the population mean under RSS for both perfect as well as imperfect ranking. To take advantage from the negative correlation between the observations, Patil et al. (1995) extended the idea of ranked set sampling for finite population assuming sampling without replacement. Later on, Muttlak (2003) suggested median ranked set sampling (MRSS) for the estimation of finite population mean. Al-Saleh and Al-Omari (2002) used multistage ranked set sampling (MSRSS) to improve the efficiency of an estimator of the population mean for certain values of the sample size. Although MSRSS results improved estimators of the population parameter than RSS does, this sampling scheme requires a huge number of population units to be ranked before actual quantifications which questions on its applicability. Mahdizadeh and Zamanzade (2019) developed a new variation on MRSS called multistage paired ranked set sampling (MSPRR) to reduce ranking burden in MRSS and use it for estimation of body fat. Many other authors have worked on estimation of parameters in RSS (see Samawi and Muttlak (1996), Bouza (2002), Ohyama et al. (2008) and Al-Omari and Jaber (2008) among others). Ranked set sampling has been applied, after modifications, for estimation of different population parameters such as mean, median, distribution function etc. Moreover, Haq et al. (2014) proposed a mixture of simple random sampling and ranked set sampling for estimation of population mean and median. Salehi and Ahmadi (2015) worked on estimation of stress-strength reliability with the help of record values obtained through ranked set sampling. Ahmed and Shabbir (2019a) suggested extreme-cum-median ranked set sampling for estimation of population mean by sub-sampling non-respondents. Similarly, Priya and Thomas (2016) developed a method for estimation of common location and scale parameters using suitable ranked set sampling schemes. Barreto and Barnett (1999) considered a form ranked set sampling for obtaining best linear unbiased estimator for regression coefficients with replicated observations. The efficiency of estimators are compared with traditional estimator obtained under SRSWR. Chen and Wang (2004) developed sampling strategies with reduced cost and increase efficiency of the regression analysis for a lung cancer study using the concept of RSS. In this study we are concerned with estimation of finite population parameter τ (specially total) under a newly suggested ranking mechanism, called, ranked set sampling without replacement (RSSWOR) by utilizing model relationship. A single covariate model, known as gamma population model (GPM) in model-based literature Chambers et al. (2003), between the study variable and the auxiliary variable is assumed. We opt the single predictor case as the ranking mechanism is possible for two correlated variables (outcome and regressor) only. The introduction to ranked set sampling without replacement and related works are given in Chapter 5.

Another major concern in survey sampling lies in estimation of parameters for certain sub-populations, known as “domains". In many fields of research, the sampling frame for domains or sub-populations are out of date so it is not possible to classify units prior to sampling and domain membership can be observed only after survey is conducted. Example of such situations can be found in clinical studies, where patients are classified according to severity level. In public health surveys, researchers have interest in estimating well-being status of children for certain racial or ethical groups separately. In agriculture surveys, farmers are classified according to the size of forms harvested. Similarly, in industrial research firms are classified according to their sizes i.e. large, medium and small. In all of these studies, we can obtain separate estimate of the characteristics of interest for each domain after observing the domain membership. Data gathered from sample surveys can be utilized to get reliable direct estimates (based on data obtained only from the sample units in the area of interest) for larger areas or domains, but in small areas sample sizes are rarely large to get direct estimates with adequate precision for small areas. To overcome this deficiency, data from related areas are utilized to find indirect estimators that increase sample size and increase the precision in estimation of small area characteristics. Demographers have been using a wide variety of methods for small area estimation of population characteristics of interests in post-censal years. Fay and Herriot (1979) and Purcell and Kish (1980) suggested post-censal estimates for local small domains. Drew et al. (1982) evaluated techniques for small area for Canadian Labour Force Survey (CLFS). Pfeffermann et al. (1996) considered labour force trend estimation for small areas. Later on, Brown et al. (2001) provided an evaluation of small area estimation methods with application to the unemployment estimates from the UK labour force survey. In same year, Ambler et al. (2001) obtained estimate of the International Labour Organization (ILO) unemployment for small areas by combining unemployment benefits data and LFS data. Works related to such methods can be found in (Rao, 1994, 2003) and You (2008).

Currently, Whitworth et al. (2017) considered estimation of uncertainty in spatial micro-simulation approaches for small area estimation. Similarly, indirect estimates for small area mean or total incorporating information about membership are found in Chambers and Clark (2012). They have also suggested model-based estimators for small area parameters. The model-based small area estimators are obtained using some implicit or explicit models which links related small areas via supplementary information such information might be previous values of the variable of interest or some covariates highly related with the study variable. The utilization of the auxiliary information for efficiency improvement and weighting adjustment in post-stratified sample has also been observed in many articles (Casady and Valliant, 1993; Zhang, 2000; Breidt et al., 2008; Dever and Valliant, 2010, and reference there in).

Apart from using model relationship for enhancing efficiency of the small area estimators, one can also look for some efficient sampling scheme. Ranked set sampling scheme, the most accepted sampling scheme in term of efficiency, can be practiced for obtaining domain specific estimates of finite population parameter (specially total). We consider the estimation of domain specific total under single-covariate model using RSSWOR. Readers are referred to Chapter 6 for a detailed explanation of the proposed technique for estimation of domain specific parameters.

Finally, we turn the direction of our discussion to the application of the proposed theoretical frameworks to the demographic health survey data sets. Demographic and Health Surveys (DHS) are nationally representative household surveys which have been conducting since 1984 in more than 85 countries. The DHS were basically designed to explore demographic, family planning and fertility data collected in the Contraceptive Prevalence Surveys (CPS) Chamratrithirong et al. (1986) and World Fertility Surveys (WFS) Lightbourne et al. (1982), and to provide a necessary resources for the monitoring and evaluation of vital statistics and health indicators in developing countries. The DHS collect data on a wide range of objectives with a focus on fertility indicators, maternal and child health, reproductive health, nutrition, mortality and health behavior in adults. The main advantages of the DHS are high response rates, employment of qualified and trained interviewers, national coverage, worldwide standardized data collection procedures and consistent material over time and comparable across populations cross-sectionally as well as over time.

In last 35 years, the DHS Program has regulated more than 300 surveys in more than 90 countries in Asia, Africa and South America. These surveys were based on representative samples at national level that allow for national and sub-national estimates. After 2012-13, the DHS program conducted another demographic health survey in Pakistan for updating detailed information on demographic characteristics of the population during the year 2017-18. Following standard rules of DHS program PDHS 2017-18 is conducted on the basis of stratified two-stage sampling design, where Enumeration Areas (EAs) obtained from Census 2017 are selected on first stage as Primary Sampling Units (PSUs). On the second stage, a sample of 28 households was selected from each selected clusters or PSU. From the selected households, ever-married women of reproductive age (15–49) who stayed in the household the last night before the survey, from the households selected, were considered eligible to fill a questionnaire designed for women. Additionally, in a sub-sample of households, all men of reproductive age are considered as eligible to complete the man-questionnaire. The core questions asked to women in DHS surveys included questions about their birth history, fertility preferences and use of family planning methods etc. To calculate key fertility indicators, such as the general fertility rate (GFR), total fertility rate (TFR) and age-specific fertility rates (ASFR), produced by the DHS surveys the data on birth history (total number of live births) and age of woman at the time of the survey were utilized.

Measuring fertility indicators based on household surveys such as the DHS and Multiple Indicator Cluster Surveys (MICS) is challenging especially in low- and middle-income countries like Pakistan, where functional vital registration systems (VRS) are poor or do not exist at all. Such indicators are needed to assess the progress toward the United Nations (UN) Sustainable Development Goals (SDGs), which is especially to "*Ensure healthy lives and promote well-being for all at all ages*" calls for improving the maternal, newborn, and child health IGME (2018); Abel et al. (2016). To produce the fertility indicators, and other DHS tables, the DHS Program mostly uses the Census and Survey Processing System (CSPro)(CSPro is a public domain software package utilized by hundreds of organizations and millions of individuals for entering, editing, tabulating, and analyzing census and survey data). To replicate the rates produced by the DHS Program, common statistical packages such as SAS, SPSS, STATA or R are widely used Schoumaker (2013); Masset (2016). Recently, Elkasabi (2019) introduced the DHS.rates R package to calculate demographic indicators from DHS datasets, such as fertility and childhood mortality indicators. Motivating from the literature, in Chapter 7, we fit some non-linear regression models to birth data from Pakistan Demographic Health Survey 2017-18. Model based fertility rates i.e. ASFR, TFR, GFR and GRR are devised using predicted response obtained from the regression models after partitioning data into sampled and non-sampled parts. Detail about model-based fertility estimation is found in Chapter 7.

1.4 Objectives of the Study

The aim of this research is to investigate different aspects of model–based estimation of parameters. The basic purpose is to suggest methods which capture more information from the underlying phenomenon and to predict behavior for non-sampled population and consequently for overall population. The detailed objectives are:

(i). To provide a general framework for model-based estimation of finite population parameters under basis functions regression for a more accurate modeling of response and prediction of non-sampled part of the responses.

(ii). To extend the general framework to Bayesian paradigm for ensuring utilization of prior knowledge about the model behavior.

(iii). To cope with problem of non-response for a special case of the suggested frame.

(iv). To study the model-based frame work under sampling without replacement and certain models especially gamma population model a generalized model with single covariate.

(vi). To obtain spate estimates for small domain under a special case of proposed framework under ranked set sampling without replacement.

(vii). To provide model-based rates for fertility in Pakistan based on Pakistan Demographic Health Survey (PDHS) data collected during year 2017-18.

1.5 Outline of the Study

The formation of upcoming chapters are described in this section. Chapter 2 provides a general framework for model-based estimation of finite population parameters under different basis functions. The estimator of prediction error variance is also established using different model selection criteria theoretically as well as through simulation studies. Chapter 3 extends the suggested general frame work to Bayesian approach and studies the properties of estimators and estimated error variance via theoretical and simulated methods. Chapter 4 utilizes the model relationship between the study variable and some covariate(s) for handling non-ignorable non-response and obtaining an unbiased estimator for the population total under the sub-sampling technique. A model unbiased linear predictor for the population total in presence of non-ignorable non-response is proposed assuming unit non-response and design-based properties of the estimator are studied. In Chapter 5, we study the model-based framework under sampling without replacement and certain models especially gamma population model (GPM) a generalized model with single covariate. The proposed framework in Chapter 5 is used to obtain separate estimates for sup-population under ranked set sampling in Chapter 6. Chapter 7 provides model-based fertility rate for Pakistan based on PDHS data collected during year 2017-18. Chapter 8 covers an overall conclusion of the thesis along with some future recommendations.

Chapter 2

Model-Based Estimation of Population Parameters-A General Framework

2.1 Outline

Modeling non-linear data is a common task in the fields of data science and machine learning. It is very rare to obtain a natural process whose outcome varies linearly with the values of input variable(s). Therefore a robust and easy methodology is needed for accurately and quickly fitting a sample data set with a set of covariates assuming that the sample data could be a complicated non-linear function. In this chapter, the model-based approach for estimation of finite population parameter τ (a linear combination of the population values), assuming superpopulation setting, under basis functions regression models is discussed. Apart from the estimation of finite population parameter, we attempt to answer the question of: how one decides the order of polynomial under single predictor for modeling? How variable selection effects the finite population parameter estimation under multivariate regression model? Is there any easy method to automate the process for estimation of finite population parameters under basis function regression? How do we cope with ill-conditioning for prediction problem? Estimation of prediction error variance, under widely used feature selection criteria in machine learning (ML), are also considered. 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. Section 2.2 delineates model-based estimation developed in literature with its usual notations. Our proposed basis function approach with some special cases is described in Section 2.3. Estimation of τ under regularized basis function regression is given in Section

2.4. Section 2.5 covers variance estimation and comparison of competting variance estimators. Model selection and simulation studies are covered in Sections 2.6 and 2.7. Section 2.8 concludes the study with some future recommendations.

2.2 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 $y = (y_i, i \in U)$ 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 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:

$$Y = X\beta + \varepsilon, \tag{2.2.1}$$

where X is the known and non-stochastic data matrix containing p regressors including intercept, β is the corresponding vector of coefficients and ε be the vector of random error terms assumed to be distributed normally with mean vector 0 and variance-covariance matrix Σ. Further the data matrix X and covariance matrix Σ can be partitioned as

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

The quantity of interest, to be estimated, is a 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 $\gamma = (\gamma_s^T, \gamma_{\bar{s}}^T)^T$. Valliant (2000) defined a linear estimator (known as the 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 (2.2.1), Royall (1976) has given the general prediction estimator for $\tau(y)$ as

$$\hat{\tau}(y) = \gamma_s^T Y_s + \gamma_{\bar{s}}^T \left[X_{\bar{s}}\hat{\beta} + \Sigma_{\bar{s}s}\Sigma_{ss}^{-1}\left(Y_s - X_{\bar{s}}\hat{\beta}\right)\right], \tag{2.2.2}$$

where $\hat{\beta} = \left(X_s^{-1}\Sigma_{ss}^{-1}X_s\right)^{-1}X_s^T\Sigma_{ss}^{-1}Y_s$ is the weighted least square (WLS) estimator of the vector β. The variance of $\hat{\tau}(y)$, is given by

$$\begin{aligned} V_M(\hat{\tau}(y) - \tau(y)) =& \gamma_{\bar{s}}^T\left(\Sigma_{s\bar{s}} - \Sigma_{\bar{s}s}\Sigma_{ss}^{-1}\Sigma_{s\bar{s}}\right)\gamma_{\bar{s}} + \gamma_{\bar{s}}^T\left(X_{\bar{s}} - \Sigma_{\bar{s}s}\Sigma_{ss}^{-1}X_s\right)\left(X_s^T\Sigma_{ss}^{-1}X_s\right)^{-1} \\ & \left(X_{\bar{s}} - \Sigma_{\bar{s}s}\Sigma_{ss}^{-1}X_s\right)^T\gamma_{\bar{s}}. \end{aligned} \tag{2.2.3}$$

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

$$\hat{\tau}(y) = \gamma_s^T Y_s + \gamma_{\bar{s}}^T X_{\bar{s}}\hat{\beta} \tag{2.2.4}$$

with prediction error variance

$$V_M(\hat{\tau}(y) - \tau(y)) = \gamma_{\bar{s}}^T\left\{\Sigma_{s\bar{s}} + X_{\bar{s}}\left(X_s^T\Sigma_{ss}^{-1}X_s\right)^{-1}X_{\bar{s}}^T\right\}\gamma_{\bar{s}}. \tag{2.2.5}$$

This assumption violates for multistage surveys where intra-cluster correlation exists among units within clusters. Assuming independent and identically distributed (iid) error term i.e. $\Sigma_{ss} = \sigma^2 I_n$ and $\Sigma_{s\bar{s}} = \sigma^2 I_{N-n}$, we can write the prediction error variance as follow

$$V_M(\hat{\tau}(y) - \tau(y)) = \sigma^2\left[\gamma_{\bar{s}}^T\gamma_{\bar{s}} + \gamma_{\bar{s}}^T X_{\bar{s}}\left(X_s^T X_s\right)^{-1}X_{\bar{s}}^T\gamma_{\bar{s}}\right]. \tag{2.2.6}$$

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 underlying population model. Thus if one is concerned with superpopulation sampling, 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 how these measures changes when the working model is different from the underlying model. Royall and Herson (1973) emphasized on the balancing of a sample to protect the inference against model misspecification. In early stage of modern sampling, Valliant (2000) done an extensive work on balance sampling for reducing the effect of bias introduced due to model failure. Instead of dealing these specific problems we define a general estimation approach after predicting non-sampled data through basis function regression models (Jekabsons and Zhang, 2010).

2.3 Model based Estimation Using Basis Functions

The general prediction approach does not provide any general guideline about sample selection and use of appropriate model. Although a wide variety of restricted sampling are available in Valliant (2000) where some of them are based on linear regression model, some on polynomial models and some on proportional population model. Although the literature covers a wide range of functions of the auxiliary variable, a general framework for predicting responses from non-linear (in variable) functions of auxiliary data may provide a general guideline for selecting a sample from the population. The non-linear function of the auxiliary variable may be logarithm, some power, exponential of the auxiliary variable. In current section, we use machine learning (ML) terminologies and techniques to assist prediction of the values of the outcome corresponding to the non-sampled units for estimation of the finite population total. In ML the regression approach is considered as supervised learning, the study variable and the auxiliary variable(s) are named as output and input variables respectively. The main aim is to divide sampled data into training and test sets to check the predictive performance of the model. 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.

We start with a basic example of linear regression. For a single input variable X, the corresponding vector basis function is defined as $\Phi(x_i) = \big(\Phi_0(x_i), \Phi_1(x_i), \ldots \Phi_M(x_i)\big)$ attached to the ith population unit, where M is the number of basis in regression function $g(x, \beta)$. The matrix comprising the basis function is known as feature matrix in ML terminology, which is presented as

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

The population regression model can be expressed as

$$Y = \Phi\beta + \varepsilon \tag{2.3.1}$$

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

$$\Phi = \begin{bmatrix} \Phi_s \\ \Phi_{\bar{s}} \end{bmatrix}$$

where Φ_s and $\Phi_{\bar{s}}$ 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 general linear estimator proposed by Valliant (2000) with feature matrix Φ as:

$$\hat{\tau}(y) = \gamma_s^T Y_s + \gamma_{\bar{s}}^T \left[\Phi_{\bar{s}}\hat{\beta} + \Sigma_{\bar{s}s}\Sigma_{ss}^{-1}\left(y_s - \Phi_{\bar{s}}\hat{\beta}\right)\right], \tag{2.3.2}$$

where $\hat{\beta} = \left(\Phi_s^T \Sigma_{ss}^{-1} \Phi_s\right)^{-1} \Phi_s^T \Sigma_{ss}^{-1} y_s$ is the WLS estimator of β using basis functions rather than linear regressors.

The variance of prediction error of the proposed estimator ($e(\hat{\tau})$) is given by

$$\begin{aligned} V\left(e(\hat{\tau})\right) =& \gamma_{\bar{s}}^{T}\left(\Sigma_{s\bar{s}}-\Sigma_{\bar{s}s}\Sigma_{ss}^{-1}\Sigma_{s\bar{s}}\right)\gamma_{\bar{s}}+\gamma_{\bar{s}}^{T}\left(\Phi_{\bar{s}}-\Sigma_{\bar{s}s}\Sigma_{ss}^{-1}\Phi_{s}\right)\left(\Phi_{s}^{T}\Sigma_{ss}^{-1}\Phi_{s}\right)^{-1} \\ & \left(\Phi_{\bar{s}}-\Sigma_{\bar{s}s}\Sigma_{ss}^{-1}\Phi_{s}\right)^{T}\gamma_{\bar{s}}, \end{aligned} \tag{2.3.3}$$

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

Proof: Derivation of Equations (2.3.2) and (2.3.3) can be found after replacing the feature matrix Φ by the data matrix X in general prediction theorem given in (Valliant, 2000, Chapter 2). For simplicity, we assume non-informative sampling conditional on values of the auxiliary variables resulting $\Sigma_{\bar{s}s} = 0$, the BLUP for $\tau(y)$ reduced to

$$\hat{\tau}(y) = \gamma_{s}^{T}Y_{s} + \gamma_{\bar{s}}^{T}\Phi_{\bar{s}}\hat{\beta} \tag{2.3.4}$$

with prediction variance

$$V\left(e(\hat{\tau})\right) = \gamma_{\bar{s}}^{T}\left\{\Sigma_{s\bar{s}} + \Phi_{\bar{s}}\left(\Phi_{s}^{T}\Sigma_{ss}^{-1}\Phi_{s}\right)^{-1}\Phi_{\bar{s}}^{T}\right\}\gamma_{\bar{s}}. \tag{2.3.5}$$

Assuming iid noise in the data, we have $\Sigma_{ss} = \sigma^2 I_n$ and $\Sigma_{\bar{s}\bar{s}} = \sigma^2 I_{N-n}$. The resulting expression for variance of prediction error is

$$V\left(e(\hat{\tau})\right) = \sigma^2\left[\gamma_{\bar{s}}^{T}\gamma_{\bar{s}} + \gamma_{\bar{s}}^{T}\Phi_{\bar{s}}\left(\Phi_{s}^{T}\Phi_{s}\right)^{-1}\Phi_{\bar{s}}^{T}\gamma_{\bar{s}}\right]. \tag{2.3.6}$$

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

2.3.1 Special Cases

In this subsection, we discuss some members of basis function model and obtain estimators of total output under specified models. Model mean squared error and bias are studied for the selected cases.

2.3.1.1 Expansion Estimator

Consider a single constant basis function for estimating finite population total i.e. taking Φ as N dimensional vector of 1's.

$$y = \beta_0 + \varepsilon. \tag{2.3.7}$$

The expansion estimator for $t_y = \sum_{i \in \mathcal{U}} y_i$ (population total) under homogeneous population is obtained as follows:

$$\hat{t}_y^E = \sum_{i \in s} y_i + \sum_{i \in \bar{s}} \hat{\beta}_0, \tag{2.3.8}$$

where $\hat{\beta}_0 = \frac{\sum_{i \in s} y_i}{n}$ is the best linear unbiased predictor (BLUP) for β_0 obtained under the ordinary least square assumptions. The expansion estimator is unbiased when underlying model is correct. The prediction error variance of the expansion estimator, is given by

$$V_M(\hat{t}_{ys} - t_y) = N^2 \big(\frac{1}{n} - \frac{1}{N}\big)\sigma^2 \tag{2.3.9}$$

which is equivalent to the designed-based variance of total estimator under simple random sampling without replacement (SRSWOR) (see Cochran, 1940).

2.3.1.2 Regression Estimator

Assuming single variable linear regression model with basis functions including intercept, we have

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

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 $\bar{\Phi}_{js} = \frac{1}{n}\sum_{i \in s} \Phi_j(x_i)$. After some simplification, we get

$$\hat{t}_{y(reg)} = N\big[\bar{y} + \sum_{j=1}^{M-1} \hat{\beta}_j \{\bar{\Phi}_{jU} - \bar{\Phi}_{js}\}\big], \tag{2.3.10}$$

where $\bar{\Phi}_{jU} = \frac{1}{N}\sum_{i\in\mathcal{U}}\Phi_j(x_i)$. It is easy to show that $\hat{t}_{y(reg)}$ 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 some bias. Consider 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 $B_M(t_y) = N\sum_{j=1}^{M-1}\beta_j\left(\bar{\Phi}_{jU} - \bar{\Phi}_{js}\right)$ 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 $\bar{\Phi}_{jU} < \bar{\Phi}_{js}$ and the bias $B_M(\hat{t}_{y(reg)})$ becomes negative. The bias can be minimized by selecting a sample such that the difference on right side of bias expression is minimum. Following Valliant (2000), we call such a sample as balanced sample. Exact balanced sample is achieved by selecting a sample for which $\bar{\Phi}_{jU} = \bar{\Phi}_{js}$. The prediction error variance for the estimator given in (2.3.10), is given by

$$V_M(\hat{t}_{y(reg)} - t_y) = N^2\left[\sum_{j=1}^{M-1}\left(\bar{\Phi}_{jU} - \bar{\Phi}_{js}\right)^2 V_M(\hat{\beta}_j) + \left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2\right]. \tag{2.3.11}$$

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

$$V_M(\hat{t}_{y(reg)} - t_y) = N^2\left[\frac{\left(\bar{\Phi}_{1U} - \bar{\Phi}_{1s}\right)^2}{\sum_{i\in s}\left(\Phi_1(x_i) - \bar{\Phi}_{1s}\right)^2} + \left(\frac{1}{n} - \frac{1}{N}\right)\right]\sigma^2. \tag{2.3.12}$$

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

2.3.1.3 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 heteroscedasticity. In such situations, weighted least square 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:

$$y = f(x,\beta) + \psi(x)\varepsilon, \tag{2.3.13}$$

where $f(x,\beta) = \sum_{j=1}^{M-1}\beta_j\Phi_j(x)$. The gamma population model discussed by Chambers and Clark (2012) is obtained by setting $\psi(x) = x^{\gamma^*}$ and the well known ratio estimator is obtained by setting $\gamma^* = \frac{1}{2}$. For $\gamma^* = 0$, we get linear regression estimator with constant variance. To obtain homoscedastic error term, we adopt following weighted least square method to estimate (2.3.13).

$$y^* = \sum_{j=1}^{M-1} \beta_j \Phi_j^*(x) + \varepsilon, \tag{2.3.14}$$

where $y^* = \frac{y}{\psi(x)}$ and $\Phi_j^*(x) = \frac{\Phi_j(x)}{\psi(x)}$ for $M = 2$

$$y^* = \beta_1 \Phi_1^*(x) + \varepsilon. \tag{2.3.15}$$

The best linear unbiased estimator (BLUE) for β_1 is then 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

$$\begin{aligned} \hat{t}_{y(r)} =& \sum_{i \in s} y_i + \sum_{i \in r} \Phi_1^*(x_i) \frac{\sum_{i \in s} \Phi_1^*(x_i) y_i^*}{\sum_{i \in s} \Phi_1^{*2}(x_i)} \\ =& \sum_{i \in s} \left[1 + \lambda_i \sum_{i \in r} \Phi_1^*(x_i)\right] y_i, \end{aligned} \tag{2.3.16}$$

where $\lambda_i = \frac{\Phi_1^*(x_i)}{\psi(x_i) \sum_{i \in s} \Phi_1^{*2}(x_i)}$.

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

where $\lambda_i^* = \lambda_i \sum_{i \in r} \Phi_1^*(x_i)$.
The model bias and variance, are given by

$$B_M(\hat{t}_{y(r)}) = \beta_1 \left[\sum_{i \in s} \lambda_i^* \Phi_1(x_i) - \sum_{i \in r} \Phi_1(x_i)\right] \tag{2.3.17}$$

and

$$V_M(\hat{t}_{y(r)} - \tau(y)) = \left[\sum_{i \in s} \lambda_i^{*2} \psi^2(x_i) + \sum_{i \in s} \psi^2(x_i)\right] \sigma^2. \tag{2.3.18}$$

The model mean squared is obtained as:

$$MSE_M(\hat{t}_{y(r)}) = \left[\sum_{i\in s}\lambda_i^{*2}\psi^2(x_i) + \sum_{i\in r}\psi^2(x_i)\right]\sigma^2 + \beta_1^2\left[\sum_{i\in s}\lambda_i^{*}\Phi_1(x_i) - \sum_{i\in r}\Phi_0(x_i)\right]^2. \quad (2.3.19)$$

A balance sampling with $\sum_{i\in s}\lambda_i^{*}\Phi_1(x_i) - \sum_{i\in r}\Phi_1(x_i)$ results $\hat{t}_{y(r)}$ in unbiasedness. This is same as the calibration estimator with single predictor given in Deville and Särndal (1992).

2.3.2 Some Special Basis Functions

The next problem is to choose a reasonable function of the predictor or set of predictors for prediction. The world we are living is much complicated, and we can't easily adopt a linear models to capture the wide variety of so called basis functions that we might need in prediction. To capture complex phenomenon with non-linear behavior data, scientists urged on use of a wide variety of basis functions that make more precise prediction (Alpaydin, 2009). Some widely used basis functions that can be employed to parameter estimation in finite population settings are as follow.

2.3.2.1 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 between 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. Before going to the complex non-linear functions for predicting non-sampled values, one should opt linear regression to look up. As we know that the 'linearity' in linear regression model refers that the model is 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.

Consider polynomial basis function $f(x,\beta) = \sum_{l=0}^{M-1} \beta_l x^l$, here the feature matrix is

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

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, $M=2$ linear regression model with intercept and for $M=3$, we get quadratic regression model. The polynomial basis models provide 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.

2.3.2.2 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. The prediction problem for three input variables for the case of $M=2$ is considered here. We use separate index for each variable as $J=(j_1,j_2,j_3)$ such that $(j_1+j_2+j_3) \leq (M-1)$.

$$\begin{aligned} y &= \sum_{j_1,j_2,j_3} \beta_j \Phi_j(x) + \varepsilon \\ &= \beta_{000} + \beta_{100}x_1 + \beta_{010}x_2 + \beta_{001}x_3 + \beta_{110}x_1x_2 + \beta_{101}x_1x_3 \\ &+ \beta_{011}x_2x_3 + \beta_{200}x_1^2 + \beta_{020}x_2^2 + \beta_{002}x_3^2 + \varepsilon \end{aligned} \tag{2.3.20}$$

where ε is random error term and the subscripts of the coefficients with values 1 show that the corresponding regressor is present and 0 represent its absence. For p covariates the number of quadratic terms is $\left[1+p+p(p-1)\right]/(2+p)$ in above example, we have $p=3$, hence the number of terms is 10. For p inputs, a general case is the regression model with basis functions $\Phi(x) = \left\{ \prod_{k=1}^{p} x_k^{m_k} : \sum_{k=1}^{p} m_k \leq p \right\}$.

2.3.2.3 Radial Basis Functions (RBF)

Radial basis functions are the 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 based on the distance from some other point, called a center, so that $\Phi(x,c) = \Phi(||x-c||)$. In general, a function $\Phi(x)$ is said to be radial basis function if it can be expressed as $\Phi(x) = \Phi(||x||)$. The concept of radial basis was initially introduced by Broomhead and Lowe (1988) stemmed from Powell (1977).

Lowe and Broomhead (1988) 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 (Scholkopf et al., 1997). Buhmann (2003) provided theory and implementation of RBF. Later Biancolini (2017) extended its application in different fields of engineering and Physics. Radial basis functions are typically preferred for estimating population parameters when the auxiliary data consist of latitudes and longitudes. 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.

2.4 Estimation Under Regularized Regression

In regression analysis, over-fitting means the outcome of an analysis that corresponds exactly or very close to a particular data set, and therefore failure to fit additional data points. Such situations are termed as ill-conditioning in regression analysis. Initially, Tikhonov and Arsenin (1977) worked on mathematical aspect of the ill-posed problems and discussed the problem in their book. Aside from Tikhonov and Arsenin (1977), Hoerl and Kennard (1970) 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. Hoerl and Kennard (1970) method was actually a crude form of the ridge regression now known as zero order regularization (Press and Flannery, 1992). 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 is available in literature and most of them are cited in Cartis et al. (2019). In this section, we confined our discussion to the simple regularization method introduced by Hoerl and Kennard (1970) although our prediction problem can be handled by using more advanced regularization methods e.g Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), elastic net regression (Zou and Hastie, 2005) 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 mean squared error. If $E_M\left(\hat{\beta}_{ridge}\right) = \beta$ for all Φ, then the total estimator will be unbiased. However an unbiased estimator may still have larger mean squared error 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 Tikhonov and Arsenin (1977). To reduce the high variation significantly Hoerl and Kennard (1970) suggested to introduce small amount of bias so that net effect, shown in (2.4.5), is a reduction in mean squared error. Under regularization, we have following cost function (sum-squared error)

$$C = \left(y_s - \Phi_s\beta\right)^T \left(y_s - \Phi_s\beta\right) + v\beta^T\beta \tag{2.4.1}$$

where the positive constant v is called regularizer which creates bias in the estimate of β and reduces variance on the other side. Optimizing the cost function given in (2.4.1), we have following ridge regression estimator for the coefficient vector

$$\hat{\beta}_{ridge} = Q_s^{-1}\Phi_s^T y_s, \tag{2.4.2}$$

where $Q_s = \Phi_s^T\Phi_s^T + vI_n$. The matrix Q_s is symmetric i.e. $Q_s^T = Q_s$. An estimator of population parameter $\tau(y)$ using ridge regression for model estimation is given by

$$\hat{\tau}_{Ridg}(y) = \gamma_s^T y_s + \gamma_{\bar{s}}^T \Phi_{\bar{s}}\hat{\beta}_{ridge} \tag{2.4.3}$$

which has model bias

$$E_M\big(e(\hat{\tau}_{ridge})\big) = \gamma_{\bar{s}}^T \Phi_{\bar{s}} \big[E_M\big(\hat{\beta}_{ridge}\big) - \beta\big],$$

where $e(\hat{\tau}_{ridge}) = \hat{\tau}_{ridge}(y) - \tau(y)$. After some simplification (see Appendix A.5), we have

$$B_M\big(\hat{\tau}_{ridge}(y)\big) = -v\gamma_{\bar{s}}^T \Phi_{\bar{s}} Q_s^{-1}\beta. \tag{2.4.4}$$

This amount of bias depends on the regularizer v and we can infer that the bias tends to reduces as $v \to 0$ depending on entries in Q_s^{-1} (which also depend on v). Now we observe the effect on the error variance of $\hat{\tau}_{ridge}(y)$. The variance expression is given by

$$\begin{aligned} V_M\big(e(\hat{\tau}_{ridge})\big) &= \sigma^2 \left[\gamma_{\bar{s}}^T \gamma_{\bar{s}} + \gamma_{\bar{s}}^T \Phi_{\bar{s}} \left(Q_s^{-1} - Q_s^{-2}\right) \Phi_{\bar{s}}^T \gamma_{\bar{s}}\right] \\ &= V_M\big(e(\hat{\tau})\big) - \sigma^2 \gamma_{\bar{s}}^T \Phi_{\bar{s}} Q_s^{-2} \Phi_{\bar{s}}^T \gamma_{\bar{s}}. \end{aligned} \tag{2.4.5}$$

This shows that regularization reduces variance by an amount of $\sigma^2 \gamma_{\bar{s}}^T \Phi_{\bar{s}} Q_s^{-2} \Phi_{\bar{s}}^T \gamma_{\bar{s}}$. This amount increases by increasing the parameter v which ultimately increases the efficiency with larger amount of bias. The mean squared error of $\hat{\tau}_B(y)_{\bar{s}idge}$ is then obtained using bias and variance relation as follow

$$\begin{aligned} MSE_M\{\hat{\tau}_{ridge}(y)\} &= \sigma^2 \left[\gamma_{\bar{s}}^T \gamma_{\bar{s}} + \gamma_{\bar{s}}^T \Phi_{\bar{s}} \left(Q_s^{-1} - Q_s^{-2}\right) \Phi_{\bar{s}}^T \gamma_{\bar{s}}\right] + v^2 \gamma_{\bar{s}}^T \Phi_{\bar{s}} Q_s^{-1} \beta\beta^T Q_s^{-1} \Phi_{\bar{s}}^T \gamma_{\bar{s}} \\ &= \sigma^2 \left(\gamma_{\bar{s}}^T \gamma_{\bar{s}}\right) + \gamma_{\bar{s}}^T \Phi_{\bar{s}} \left[\sigma^2 \left(Q_s^{-1} - Q_s^{-2}\right) + v^2 Q_s^{-1} \beta\beta^T Q_s^{-1}\right] \Phi_{\bar{s}}^T \gamma_{\bar{s}} \end{aligned} \tag{2.4.6}$$

The amount $\gamma_{\bar{s}}^T \Phi_{\bar{s}} [v^2 Q_s^{-1} \beta\beta^T Q_s^{-1} - \sigma^2 Q_s^{-2}] \Phi_{\bar{s}}^T \gamma_{\bar{s}}$ is the net effect in reduction of mean squared error. 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 optimization of (2.4.5) with respect to v is not straightforward. 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 have 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.

2.5 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 such as Jackknife technique (Shao et al., 1989), in model-based approach, we utilize model selection criteria which indirectly provide estimate of error variance σ^2 for obtaining estimate for prediction error variance of $\hat{\tau}(y)$. It can be seen that the variance of error term given in (2.3.6) depends on error variance σ^2 and the auxiliary data (data on basis functions) from the whole population. When we have known sub-matrix of the basis function for the non-sampled part as well, we need estimate for σ^2 only for estimating prediction variance of $\tau(y)$. Hence a sample estimate for the prediction error variances given in (2.3.6) and (2.4.4) can be expressed as

$$\hat{V}\left(e(\hat{\tau})\right) = \hat{\sigma}^2\left\{\gamma_{\bar{s}}^T\gamma_{\bar{s}} + \gamma_{\bar{s}}^T\Phi_{\bar{s}}\left(\Phi_s^T\Phi_s\right)^{-1}\Phi_{\bar{s}}^T\gamma_{\bar{s}}\right\}, \tag{2.5.1}$$

and

$$\hat{V}\left(e(\hat{\tau}_{ridge})\right) = \hat{V}\left(e(\hat{\tau}_{ML})\right) - \hat{\sigma}^2\gamma_{\bar{s}}^T\Phi_{\bar{s}}Q_s^{-2}\Phi_{\bar{s}}^T\gamma_{\bar{s}}. \tag{2.5.2}$$

Estimation of σ^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. An estimator of prediction error variance of the estimator $\hat{\tau}(y)$ is obtained under residual method by replacing $\hat{\sigma}^2_{res}$ by $\hat{\sigma}^2$ in (2.5.1) and (2.5.2), where

$$\hat{\sigma}^2_{res} = \frac{1}{n-M}y_s^T P^2 y_s, \tag{2.5.3}$$

where $P = I_M - \Phi_s Q_s^{-1}\Phi_s^T$ is the projection matrix which is symmetric and idempotent. The projection matrix defined in A.1 (see Appendix A) is idempotent i.e. $P^2 = P$ if no regularization is used. Another most widely used model selection criteria is unbiased estimate of variance (UEV) which is similar to residual variance obtained by replacing the number of total parameter by the number of effective parameter in the denominator. The UEV estimator of σ^2, is given by

$$\hat{\sigma}^2_{UEV} = \frac{1}{n - M^*} y_s^T P^2 y_s, \tag{2.5.4}$$

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 (Zheng and Agresti, 2000). 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. A model is considered to be best whose estimated prediction error is minimum among all other alternative models. In following subsections, we provide some alternative variance estimation methods. We extend these 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 σ^2 using above mentioned methods. For obtaining estimates for σ^2, we use following model selection criteria.

1- Cross Validation (CV)

2- Generalized Cross Validation (GCV)

3-Final Prediction Error or Akaike's Information Criterion (AIC)

4-Bayesian Information Criterion (BIC).

We developed different estimators for the error variance of $\hat{\tau}(y)$ using estimates for σ^2 obtained under different model selection criteria in the following subsections.

2.5.1 Variance Estimation under Cross Validation (CV)

The simplest kind of cross validation is the holdout method in which the data set is divided into two halves, i.e. the training set and the testing set. An appropriate method of estimation is used to estimate the parameters of function (in ML language to trained the model) using the training set. Then the estimated model is used to predict the outputs for the test data (which is holdout during training). The prediction errors it makes are averaged to give the mean absolute prediction error, which is considered as an alternate tool of model evaluation. This method takes no longer time in computation. However, the estimate of variance may have higher variance as it depends heavily on the separation mechanism of the data into test and training. k-fold CV is an improved form of the holdout CV method in which the data set is splitted into k subsets. The method of holdout is re-runed k times by considering one of the k subsets as the test set and training the model with the help of remaining $(k-1)$ subsets each time. Then the errors from all k trials are accumulated to compute the average error. This k-fold CV method has smaller error than holdout method but consumes relatively more time in computation. The variance of the estimate of prediction error variance is reduced as k increases. A new variant of this method is to randomly split the data into a training and 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 (LOO) cross validation 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 $\hat{\sigma}^2_{LOO}$ in (2.5.1) and (2.5.2), where

$$\hat{\sigma}^2_{LOO} = \frac{1}{n} y_s^T P \big\{ \text{diag}(P) \big\}^{-2} P y_s \tag{2.5.5}$$

The estimated variance obtained under LOO cross validation error is a good estimate of model variance, but at first glance it seems more expensive and tiresome to compute. Luckily, locally weighted regressions make 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.

2.5.2 Variance Estimation under Generalized Cross validation

The diagonal matrix diag(P) makes LOO mathematically inappropriate. Its alternate, GCV introduced by Golub et al. (1979), is more convenient and is obtained by replacing the matrix diag(P) by the average of the diagonal elements multiplied by the identity matrix of order n i.e. 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 $\hat{\sigma}^2_{GCV}$ in (2.5.1) and (2.5.2), where $\hat{\sigma}^2_{GCV}$ is

$$\hat{\sigma}^2_{GCV} = \frac{n y_s^T P^2 y_s}{\big\{ \text{trace}(P) \big\}^2}. \tag{2.5.6}$$

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 (2.5.5), when $\frac{n}{\left\{\text{trace}(P)\right\}^2} = \frac{1}{n-M^*}$, where $M^* = n - sum(diag(P))$ 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 trace(P) as

$$\hat{\sigma}^2_{GCV} = \frac{ny_s^T P^2 y_s}{\left(n - M^*\right)^2}. \tag{2.5.7}$$

2.5.3 Variance Estimation under Final Prediction Error (FPE)

Mallows's C_p (Mallows, 1973), named after Colin Lingwood Mallows, is a statistic which assess 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. A smaller value of C_p indicating relatively precise fit and vice versa. Under Gaussian linear regression model Mallows's C_p is equivalent to Aikake's Information Criterion(AIC), a most widely used model evaluation criterion (Boisbunon et al., 2013) and can be 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 $\hat{\sigma}^2_{FPE}$ in Equation (2.5.1), where $\hat{\sigma}^2_{FPE}$, an alternative version of Mallows's C_p (James et al., 2013), and is given by

$$\hat{\sigma}^2_{FPE} = \frac{1}{n}\left(y_s^T P^2 y_s + 2M^* \hat{\sigma}^2_{res}\right) = \frac{n+M^*}{n-M^*}\frac{y_s^T P^2 y_s}{n}, \tag{2.5.8}$$

where M^* is the effective number of parameters. The $\hat{\sigma}^2_{FPE}$ suffers from two limitations: 1. the approximation is valid only for large enough sample size and 2. it can't deal with complex set of models as in the variable selection (feature selection in machine learning) problems (Giraud, 2014).

2.5.4 Variance Estimation under Bayesian Information Criterion (BIC)

The BIC developed by (Schwarz et al., 1978), 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 "Akaike's Bayesian Information Criterion" instead of BIC "a Bayesian Information

Criterion" (Akaike, 1977). An estimator of variance of prediction error for $\hat{\tau}(y)$ based on BIC is found by inserting $\hat{\sigma}^2_{BIC}$ by $\hat{\sigma}^2$ in Equations (2.5.1) and (2.5.2), where $\hat{\sigma}^2_{BIC}$ is

$$\begin{aligned}\hat{\sigma}^2_{BIC} &= \frac{1}{n}\left(y_s^T P^2 y_s + \ln(n) M^* \hat{\sigma}^2_{res}\right)\\ &= \frac{n + M^*\left(\ln(n) - 1\right)}{n - M^*}\frac{y_s^T P^2 y_s}{n},\end{aligned} \tag{2.5.9}$$

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 variance of prediction error 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 $\sigma^2_{abc} = \Gamma_{abc} y_s^T P^2 y_s / n$ and have following natural ordering as

$$\Gamma_{UEV} \leq \Gamma_{FPE} \leq \Gamma_{GCV} \leq \Gamma_{BIC} \tag{2.5.10}$$

The factors Γs are approximated by using Taylor's series as:

$$\begin{aligned}\Gamma_{\bar{s}es} &= \frac{n}{n - M^*} = 1 + \frac{M^*}{n} + \frac{M^{*2}}{n} + \frac{M^{*3}}{n} + \ldots.\\ \Gamma_{FPE} &= \frac{n + M^*}{n - M^*} = 1 + \frac{2M^*}{n} + \frac{2M^{*2}}{n^2} + \frac{2M^{*3}}{n^3} + \ldots.\\ \Gamma_{GCV} &= \frac{M^{*2}}{\left(n - M^*\right)^2} = 1 + \frac{2M^*}{n} + \frac{3M^{*2}}{n^2} + \frac{4M^{*3}}{n^3} + \ldots.\\ \Gamma_{BIC} &= \frac{n + \left(\ln(n) - 1\right)M^*}{n - M^*} = 1 + \ln(n)\left(\frac{M^*}{n} + \frac{M^{*2}}{n} + \frac{M^{*3}}{n} + \ldots.\right)\end{aligned}$$

Hence the estimators of $\hat{\sigma}^2$ obtained through different model selection criteria can be ranked according to the factor Γ. Hence variance estimators can also be ranked as

$$\hat{V}\left(e(\hat{\tau})\right)_{UEV} \leq \hat{V}\left(e(\hat{\tau})\right)_{FPE} \leq \hat{V}\left(e(\hat{\tau})\right)_{GCV} \leq \hat{V}\left(e(\hat{\tau})\right)_{BIC} \tag{2.5.11}$$

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

2.6 Model Selection

We previously discussed ridge regression (Section 2.4) as a tool for controlling the trade-off between the bias and variance (Section 2.5) of the estimators of superpopulation parameters such as σ^2. Alternatively, one can compare models with different subsets of basis functions selected from a fixed set of candidate models, known as "subset selection" (Rawlings et al., 2001). 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 also widely used for factor screening in multiple regression analysis our aim is to estimate the finite population parameter(s) as a prediction problem. So it does not seem logical to start with a larger group of covariates or with a higher order polynomials (in case of single covariate) and then come to an effective smaller subset. On the other end, 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 in Appendix (see Appendix A). We see the effect of adding a new predictor on the bias and variance of the total estimator in following subsections.

2.6.1 Model Selection Under Ordinary Least Square

The reduction in variance (increase in efficiency) on using additional basis function can be computed as

$$\begin{aligned} IE =& V_M(\hat{\tau})_m - V_M(\hat{\tau})_{m+1} \\ =& \sigma^2 \gamma_{\bar{s}}^T \left[\Phi_{\bar{s}m} A_{sm}^{-1} \Phi_{\bar{s}m}^T - \Phi_{\bar{s}(m+1)} A_{s(m+1)}^{-1} \Phi_{\bar{s}(m+1)}^T \right] \gamma_{\bar{s}}. \end{aligned} \tag{2.6.1}$$

The subscripts m and $(m+1)$ are used to denote that the quantities are obtained with M basis functions, $(M+1)$ basis functions, with s and $\bar{s}$ for sampled and non-sampled populations respectively. From (A.3) (see Appendix A) and (2.6.1), we get

$$\begin{aligned} IE =& V_M(\hat{\tau})_m - V_M(\hat{\tau})_{m+1} \\ =& \sigma^2 \gamma_{\bar{s}}^T \left[\Phi_{\bar{s}m} A_{sm}^{-1} \Phi_{\bar{s}m}^T - \Phi_{\bar{s}(m+1)} A_{s(m+1)}^{-1} \Phi_{\bar{s}(m+1)}^T \right] \gamma_{\bar{s}} \end{aligned} \tag{2.6.2}$$

$$\begin{aligned} IE =& \frac{1}{\triangle} \sigma^2 \gamma_{\bar{s}}^T \Big[\phi_{\bar{s}(m+1)} \phi_{s(m+1)}^T \Phi_{sm} A_{sm}^{-1} \Phi_{\bar{s}m}^T + \Phi_{\bar{s}m}^T A_{sm}^{-1} \Phi_{sm}^T \phi_{s(m+1)} \phi_{\bar{s}(m+1)}^T \\ & - \Phi_{\bar{s}m} A_{sm}^{-1} \Phi_{sm}^T \phi_{s(m+1)} \phi_{s(m+1)}^T \Phi_{sm} A_{sm}^{-1} \Phi_{\bar{s}m}^T - \phi_{\bar{s}(m+1)} \phi_{\bar{s}(m+1)}^T \Big] \gamma_{\bar{s}}, \end{aligned} \tag{2.6.3}$$

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

$$IE_R = \frac{\gamma_{\bar{s}}^T \left[\phi_{\bar{s}(m+1)} \phi_{s(m+1)}^T \Phi_{sm} A_{sm}^{-1} \Phi_{\bar{s}m}^T + \Phi_{\bar{s}m}^T A_{sm}^{-1} \Phi_{sm}^T \phi_{s(m+1)} \phi_{\bar{s}(m+1)}^T \right] \gamma_{\bar{s}}}{\gamma_{\bar{s}}^T \left[\Phi_{\bar{s}m} A_{sm}^{-1} \Phi_{sm}^T \phi_{s(m+1)} \phi_{s(m+1)}^T \Phi_{sm} A_{sm}^{-1} \Phi_{\bar{s}m}^T + \phi_{\bar{s}(m+1)} \phi_{\bar{s}(m+1)}^T \right] \gamma_{\bar{s}}} \tag{2.6.4}$$

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 response in advance. Then different estimates obtained in previous section are applied. 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 (2.6.4) and separately obtain the estimates of σ^2 through incremental operators given in Orr et al. (1996). 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).

2.6.2 Model Selection Under Regularized Regression

Under regularization, the absolute change in bias can be expressed as

$$|B_M\big(\hat{\tau}_{ridge}(y)\big)_m - B_M\big(\hat{\tau}_{ridge}(y)\big)_{m+1}| = v\gamma_{\bar{s}}^T\,\big[\frac{1}{\triangle}\Phi_{\bar{s}m}Q_{sm}^{-1}\Phi_{sm}\phi_{s(m+1)}\phi_{s(m+1)}\Phi_{sm}Q_{sm}^{-1}\beta_m + \phi_{\bar{s}(m+1)}Q_{21}^{-1}\beta_m + \Phi_{\bar{s}m}Q_{12}^{-1}\beta_{m+1} + \phi_{\bar{s}(m+1)}Q_{22}^{-1}\beta_{m+1}\big], \tag{2.6.5}$$

where β_{m+1} is the $(M+1)$th component of the vector β_{m+1} i.e. the effect of additional basis function on the response. Q_{12}^{-1}, Q_{21}^{-1} and Q_{22}^{-1} are defined in Appendix A. 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

$$\begin{aligned} IE_{ridge} &= V_M\big(e(\hat{\tau}_{ridge})\big) - V_{M+1}\big(e(\hat{\tau}_{ridge})\big) \\ &= IE - \sigma^2\gamma_{\bar{s}}^T\,\big[\Phi_{\bar{s}m}Q_{sm}^{-2}\Phi_{\bar{s}m}^T - \Phi_{\bar{s}(m+1)}Q_{s(m+1)}^{-2}\Phi_{\bar{s}(m+1)}^T\big]\gamma_{\bar{s}}, \end{aligned} \tag{2.6.6}$$

where
$\sigma^2\gamma_{\bar{s}}^T\,\big[\Phi_{\bar{s}m}Q_{sm}^{-2}\Phi_{\bar{s}m}^T - \Phi_{\bar{s}(m+1)}Q_{s(m+1)}^{-2}\Phi_{\bar{s}(m+1)}^T\big]\gamma_{\bar{s}} = -\big(2\Phi_{\bar{s}m}Q_{sm}^{-1}\triangle^*\Phi_{\bar{s}m}^T + \Phi_{\bar{s}m}\triangle^*\triangle^{*T}\Phi_{\bar{s}m}^T + \phi_{\bar{s}(m+1)}Q_{21}^{-2}\Phi_{\bar{s}m}^T + \Phi_{\bar{s}m}Q_{12}^{-1}\phi_{\bar{s}(m+1)}^T + \phi_{\bar{s}(m+1)}Q_{22}^{-2}\phi_{\bar{s}(m+1)}^T\big)$
and $\triangle^* = \triangle^{-1}Q_{sm}^{-1}\Phi_{sm}^T\phi_{s(m+1)}\phi_{s(m+1)}^T\Phi_{sm}Q_{sm}^{-1}$. Q_{21}^{-2}, Q_{12}^{-2} and Q_{22}^{-2} are the elements of the matrix $Q_{s(m+1)}^{-2} = Q_{s(m+1)}^{-1}Q_{s(m+1)}^{-1}$. Computation of IE_{ridge} is not straight forward still some algebraical treatment on matrices can provide a compact form that can be solved numerically. A positive value of the index IE_{ridge} provides evidence of efficiency improvement by adding additional basis function to the superpopulation model which is the main concern of our study.

2.7 Simulations

Two simulation studies (one simulated and one bootstrapped) are conducted to evaluate the error variance of the proposed estimator of $\hat{\tau}(y)$ ($\gamma_i = 1$ for all $i \in \mathcal{U}$) and to find expected values of the estimated error variance of $\hat{\tau}(y)$. For this purpose, first we provide a simulation study using artificially generated population and fitting basis functions (we limit our discussion to polynomial basis function to avoid complexity). Secondly, we take a real data set and perform repeated sampling to obtain design-based properties of the estimator and estimated error variance of the estimator $e(\hat{\tau})$. 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 (10,10) and (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 $df(y,x,x^2,x^3,...x^{M-1})$ (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 (β and σ^2). The estimated values of σ^2 are obtained using different formula discussed in Section 2.5.

(iii). Further we evaluate the the proposed estimator of $\tau(y)$ (with $\gamma_i = 1$ for all $i \in \mathcal{U}$) and the estimated variance of $\hat{\tau}(y)$ under different formula given in 2.5.

(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 203 hospitals from hospital data given in (Valliant, 2000, 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:

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

where the $\sum_{sim}$ is used to denote that summation is taken over all 30000 simulated samples. The ESPE are obtained under regularization taking $v = 0, 1, 5, 10$ with $v = 0$ representing no regularizer. Further the design expectation of estimated variance of $\hat{\tau}$ are obtained through respective formula after averaging over all selected samples. We use polynomial basis functions of different orders with intercept and without intercept. Scatter plots between x with observed values of y and fitted values y are shown in Figures 2.1–2.4. The scatter plots give a quick picture about the relationships between the outcome and predictor which is necessary in choosing appropriate model.

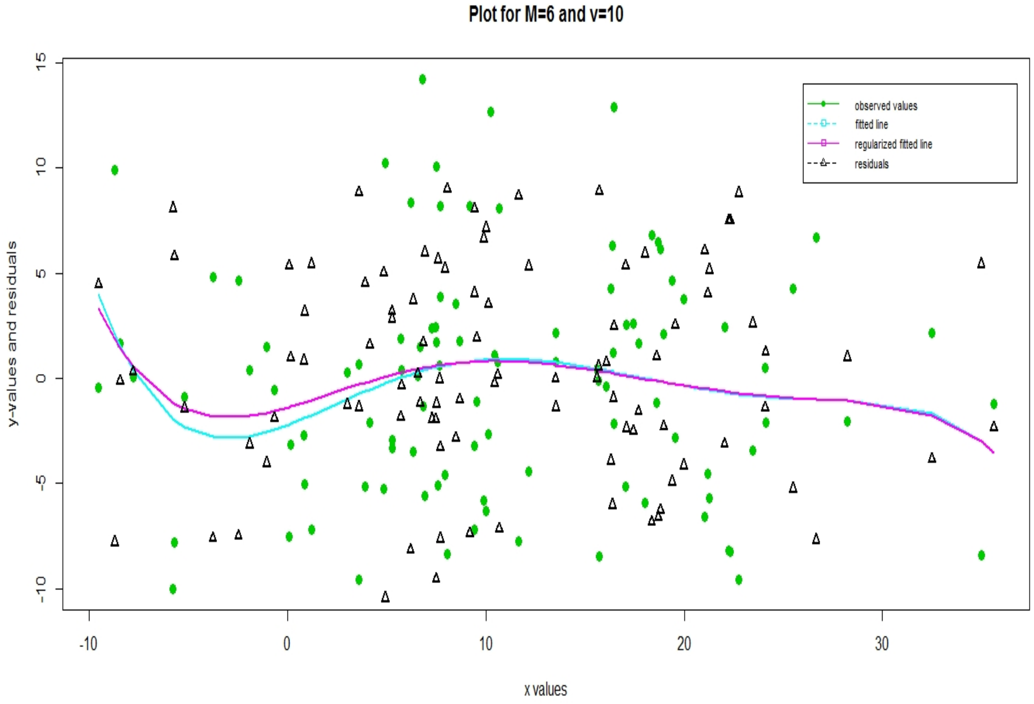

Figure 2.1: Scatter plot between x and y for sample selected from hypothetical population with $M = 6$

Scatter plot between x and y for sample selected from hypothetical population, fitted line for simple and penalized regression with polynomial of 5th order i.e $M = 6$ are displayed in Figure 2.1. The residual values are displayed on same plot shown via triangles.

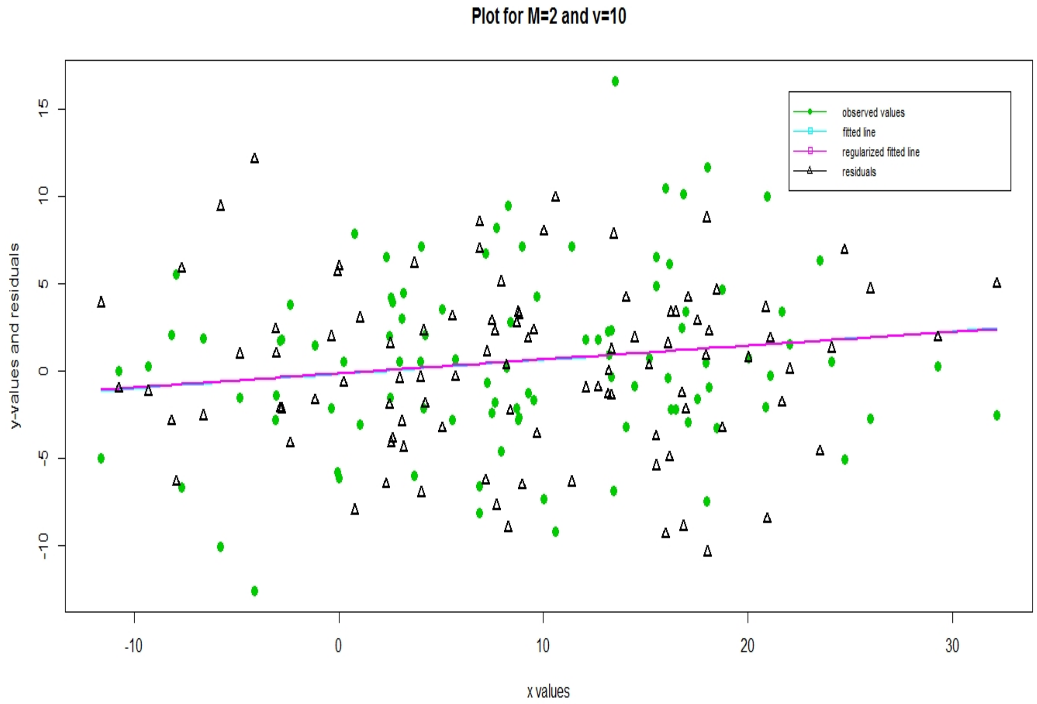

Figure 2.2: Scatter plot between x and y for sample selected from hypothetical population with $M = 2$

Scatter plot between x and y for sample selected from hypothetical population, fitted line for simple and penalized regression with polynomial of 1st order i.e $M = 2$ are displayed in Figure 2.2. The residual values are displayed on same plot shown via triangles

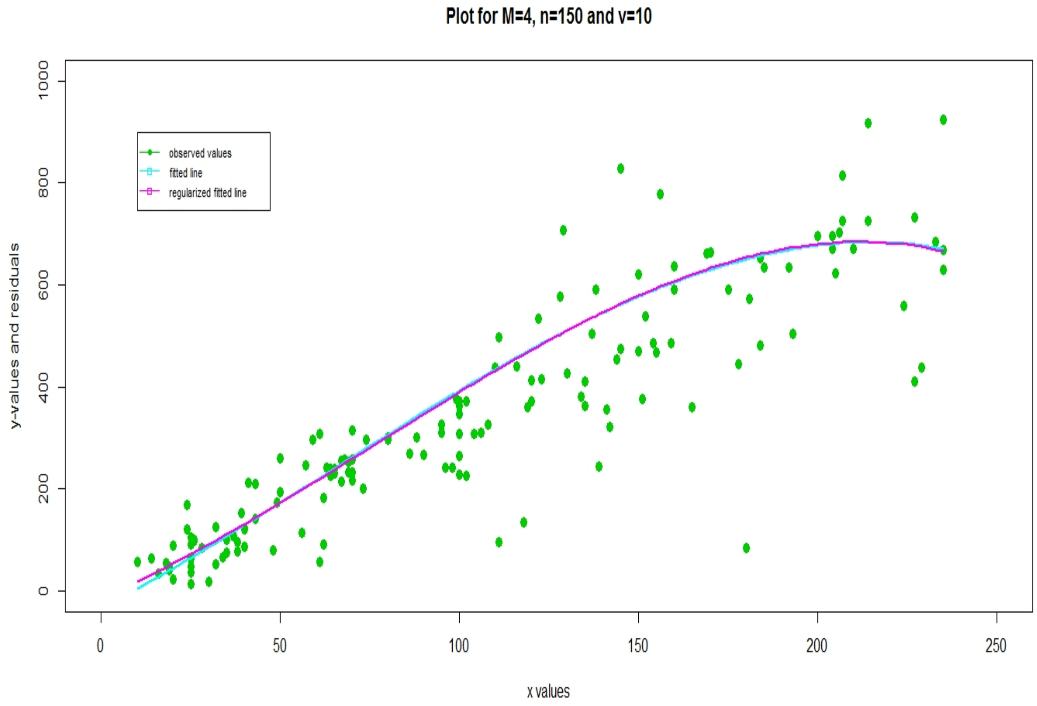

Figure 2.3: Scatter plot between x and y for sample selected from population Hospitals with $M = 4$

Figure 2.3 gives scatter plot between x and y for sample selected from population Hospitals selected from Valliant (2000), fitted line for simple and penalized regression with polynomial of 3rd order i.e $M = 4$ are displayed.

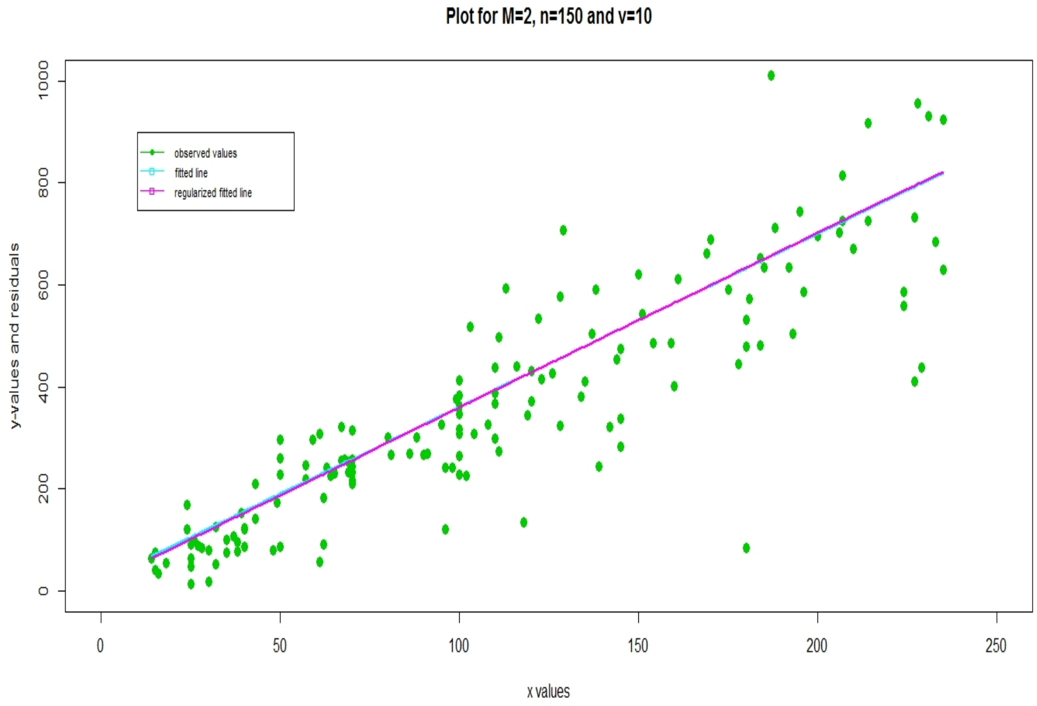

Figure 2.4: Scatter plot between x and y for sample selected from population Hospitals with $M = 2$

Figure 2.4 provides scatter plot between x and y for sample selected from population Hospitals selected from Valliant (2000), fitted line for simple and penalized regression with polynomial of 1st order i.e $M = 2$ are displayed.

Results computed from hypothetical population are given in Tables 2.1. and 2.2. for the regression models with and without intercepts respectively. Tables 2.1. and 2.2. 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 under homogeneous population model ($M = 1$), 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 2.1. and 2.2. are presented after dividing on 10^3. While Tables 2.3. and 2.4. 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.

Table 2.1.: Simulated results for linear basis function model

			Expected estimated variances				
n	v	ESPE	RES	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 2.2.: Simulated results for proportional basis function model

			Expected estimated variances				
n	v	ESPE	RES	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		29.688	74.447	75.581	76.714	76.732	80.454
	0	29.687	74.446	75.579	76.713	76.730	80.452
	1	29.685	74.440	75.574	76.707	76.724	80.445
	5	29.681	74.434	75.567	76.700	76.717	80.437

Table 2.3.: Bootstrapped results for proportional basis function model

			Expected estimated variances				
n	v	ESPE	RES	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 2.4.: Bootstrapped results for proportional basis function model

			Expected estimated variances				
n	v	ESPE	RES	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

The values of ESPE and expected estimated variances in Tables 2.3. and 2.4. are reported after dividing 10^5 for the sake of space. For real data, results are obtained for homogeneous population model ($M = 1$), linear population model ($M = 2$), 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 (2.5.2) after incorporating different estimators of σ^2. Note that all results given in Tables 2.1.–2.4. are provided for ridge regression estimator with certain choices of v as the variance estimator given (2.5.1) is a special case of variance estimator in

(2.5.2) with $v = 0$. The ESPE for different combinations of M, v and n is enlisted in third column of Tables 2.1.–2.2.. For simulated data in Table 2.1., smallest ESPE is observed at $M = 6$ when sample size is taken 100. It turns smallest at $M = 3$ when sample size is fixed at 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 2.2. (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 (2.5.11). Additionally, one can infer the following relation from our empirical evidences $\hat{V}\left(e(\hat{\tau})\right)_{RES} \leq \hat{V}\left(e(\hat{\tau})\right)_{UEV} \leq \hat{V}\left(e(\hat{\tau})\right)_{FPE} \leq \hat{V}\left(e(\hat{\tau})\right)_{GCV} \leq \hat{V}\left(e(\hat{\tau})\right)_{BIC}$. Tables 2.1.–2.4. 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 we must choose the one which is nearer to the true variance in prediction error. 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 τ and discussing problems associated with its estimation and inference about the finite population in model-based setting.

2.8 Conclusion

A general framework of model-based approach for estimation of finite population parameter τ (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 τ, 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 competting estimators with maximum at BIC which can also be observed from simulation study. 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 β'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 multi-level models and small area estimation.

Chapter 3

Model Based Estimation of Parameters under Bayesian Approach

3.1 Outline

The controversy of using Bayesian and Frequentist frameworks in statistical analysis is one of the most important academic discussion that statisticians engaged in. Rather than blindly jumping into one side, one should learn both methods of analysis and apply them where seem appropriate. In this way, recently, Bayesian method of estimation and inference have been extensively used. As we already discussed in previous chapters that the utilization of the superpopulation models for estimation of population parameters is an advantageous practice, when it is easy to recognize the relationship between the study variable and one or more auxiliary variable(s). In certain situations, one may also have prior information about the distribution of the parameters involved in the parent density during estimation of parameters such as mean, variance etc. In such situations, non-sampled values of the population units can easily be predicted using Bayesian regression approach. In this chapter, Bayesian basis function regression model is employed for predicting the values of the non-sampled units for developing Bayesian predictive estimator for the finite population parameter $\tau(y)$(a linear combination of the population values), assuming superpopulation setting. 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 in classical sense. Section 3.2 delineates the proposed basis function regression model and estimator of $\tau(y)$ under Bayesian framework. Section 3.3

covers variance estimation and comparison of competting variance estimators. Simulation studies are covered in Sections 3.4. Section 3.5 concludes the study with some future recommendations.

3.2 Model Based Estimation Under Bayesian Framework

The general prediction approach given in Chapter 2 was constructed using a general linear regression model of Y on a matrix of basis functions ϕ rather than just on regressors. Although the basis function method is flexible and generalizable to different special cases it does not utilize the prior information about the superpopulation parameter β while estimating finite population parameter $\tau(y)$. The Bayesian practitioners think that without including prior knowledge about the parameter estimation just based on sample information is not a fruitful process. One of the areas to focus in applied Bayesian inference is Bayesian linear modeling. The most important aspect of the Bayesian learning process is explaining a relationship and generalizing it to others, and this study is our attempt to use the Bayesian Linear Regression (BLR) for predicting the outcome for non-sampled set. What we result from the frequentist linear regression is an estimate of the model parameters from only the training data set(the sampled data set in our problem). Our model is informed completely by the sampled data: in this way, everything that we need to recognize our model is available in the sampled data. However, if the sample size is small, one might like to express the estimate as a distribution of possible values of the parameter given the sample information. This is the situation where Bayesian Linear Regression is needed. Again consider the notation used in Chapter 2 the population basis function regression model can be written as

$$y = g(x, \beta) + \varepsilon, \tag{3.2.1}$$

where $g(x,\beta) = \sum_{j=0}^{M-1} \beta_j \Phi_j(x) = \Phi\beta$. The basis function matrix Φ and the vector of parameter β are already defined in Section 2.2. In Bayesian paradigm, we introduce linear regression with the help of probability distributions rather than just finding estimates based on sampled data with un-specified distribution. The output variable, y, is assumed to be the outcome of random sample drawn from a probability distribution conditioning on the known covariates. Let the conditional distribution of the sampled data y with the conditional mean

$g(x,\beta)$ and the constant variance σ^2 is

$$P(y|g(x,\beta),\sigma^2) = N\big(y|g(x,\beta),\sigma^2\big). \tag{3.2.2}$$

Following Baye's Rule" the posterior distribution of the parameter β given the data vector y_s is

$$P(\beta|y_s) = \frac{P(y_s|g(x,\beta),\sigma^2) \times P(\beta)}{P(y_s)}, \tag{3.2.3}$$

where $P(\beta)$ and $P(y_s)$ are the prior distribution of the parameter β and the marginal distribution of the sampled data y_s respectively. Before going to our prediction problem, we look for a Gaussian prior for parameter β with mean vector μ_0 and variance covariance matrix Σ_0. The likelihood function $P(y_s|g(x,\beta),\sigma^2)$ is a product of Gaussian noise model. It is easy to derive the posterior distribution of β for given data using (3.2.3) which is also Gaussian, i.e.

$$P(\beta|y_s) = N\big(\beta|\mu_n,\Sigma_n\big), \tag{3.2.4}$$

where $\mu_n = \sigma^2\Sigma_n\big(\sigma^2\Sigma_0\mu_0 + \Phi^T y_s\big)$ and $\Sigma_n = \sigma^2\big(\Sigma_0^{-1} + \Phi^T\Phi\big)^{-1} = \sigma^2 Q_{sm}^{*-1}$. It is important to note that the Baye's estimator of the parameter β is the posterior mean under squared error loss function (SELF) (Zellner, 1986). Which is equivalent to the value of β for which posterior probability is maximum i.e. mode of the posterior distribution. With smaller values of the variance and covariance matrix of prior Σ_0 the Baye's estimate β reduce to maximum likelihood estimate. On the other side, for zero observation (i.e.$n = 0$) posterior mean reduces to prior mean. Further, if the sample observations arrive sequentially then posterior distribution at a point works as prior distribution for subsequent observations. Before generalizing the Gaussian prior, we derive predictive distribution to predict the outcome at non-sample data points. Following notations from Section 2.2, we can write the posterior predictive distribution for non-sampled values of the output $y_{\bar{s}}$ given sampled output y_s as

$$P(y_{\bar{s}}|y_s,\Phi) = \int P(y_{\bar{s}}|\beta,\Phi_{\bar{s}}) \times P(\beta|y_s,\Phi_s)d\beta$$

with explicit dependence on prior parameter Σ_0, noise parameter σ^2 and target variable in sampled data y_s.

$$P(y_{\bar{s}}|y_s,\Phi,\Sigma_0,\sigma^2) = \int P(y_{\bar{s}}|\beta,\Phi_{\bar{s}},\sigma^2) \times P(\beta|y_s,\Phi_s)d\beta, \tag{3.2.5}$$

where $P(y_{\bar{s}}|\beta,\Phi_{\bar{s}},\sigma^2) = N(y_{\bar{s}}|\Phi_{\bar{s}}\beta,\sigma^2 I_{\bar{s}})$ and $I_{\bar{s}}$ is an identity matrix of order $r = N-n$. Using convolution theorem on two Gaussian distributions of the right hand side of (3.2.5), we get Gaussian predictive distribution.

$$P(y_{\bar{s}}|y_s,\Phi_{\bar{s}},\Sigma_0,\sigma^2) = N\big(y_{\bar{s}}|\mu_n\Phi_{\bar{s}},\Sigma_n(x)\big), \tag{3.2.6}$$

where $\Sigma_n(x) = \sigma^2 + \Phi_s^T\Sigma_n\Phi_s$. The first term in $\Sigma_n(x)$ is noise in data and the second term is uncertainty attached with parameter β. It can be seen that as sample points increases $\Sigma_n(x)$ become smaller, mathematically $\Sigma_{n+1}(x) \leq \Sigma_n(x)$. In other word, as $n \to 0$ variance goes to zero and the variance of the posterior predictive distribution depends only on the noise term σ^2. The posterior predictive estimator for $y_{\bar{s}}$ under SELF is the mean of posterior predictive distribution i.e. $E_M(y_{\bar{s}}|y_s,\Phi_{\bar{s}},\Sigma_0,\sigma^2) = \Phi_{\bar{s}}\mu_n$ (see Ahmad et al. (2007)), where $\mu_n = \Sigma_n\big(\Sigma_0^{-1}\mu_0 + \frac{\Phi_s^T y_s}{\sigma^2}\big)$. It is interesting to note that at point $j \in \bar{s}$ the mean of predictive distribution can be expressed as linear combination of observed output i.e. $E_M(y_{\bar{s}}|y_s,\Phi_{\bar{s}},\Sigma_0,\sigma^2) = \Phi_{\bar{s}}\Sigma_n\Sigma_0^{-1}\mu_0 + \Phi_{\bar{s}}\Sigma_n\Phi_s y_s$. This property gives birth to Kernel regression, a non-parametric regression approach (Bierens, 1988) which is not included in this study.

Consider the estimation of a linear function of output variables from a finite population chosen from a superpopulation as $\tau(y) = \gamma^T y = \gamma_s^T y_s + \gamma_{\bar{s}}^T y_{\bar{s}}$. The problem is to estimate the non-sampled part of $\tau(y)$ using the Bayesian learning method i.e. $\hat{y}_{\bar{s}} = E_M(y_{\bar{s}}|y_s,\Phi_{\bar{s}},\Sigma_0,\sigma^2) = \Phi_{\bar{s}}\mu_n$. Consequently, we get following Bayesian predictive estimator for total output of the finite population

$$\hat{\tau}_B(y) = \gamma_s^T y_s + \gamma_{\bar{s}m}^T\Phi_{\bar{s}}Q_{sm}^{*-1}\big(\Phi_{sm}^T y_s + \mu_0^*\big) \tag{3.2.7}$$

with prediction error $\hat{\tau}_B(y) - \tau(y) = \gamma_{\bar{s}m}^T\Phi_{\bar{s}}Q_{sm}^{*-1}\big(\Phi_{sm}^T y_s + \mu_0^*\big) - \gamma_{\bar{s}}^T y_{\bar{s}}$. The model bias is

$$E_M\big(e(\hat{\tau}_B)\big) = \gamma_{\bar{s}}^T\Phi_{\bar{s}}\big\{\big(Q_{sm}^{*-1}A_{sm} - I_M\big)\beta + Q_{sm}^{*-1}\mu_0^*\big\} \tag{3.2.8}$$

The model variance for prediction error is given by

$$V_M\left(\hat{\tau}_B(y)-\tau(y)\right)=\sigma^2\left[\gamma_{\bar{s}}^T\gamma_{\bar{s}}+\gamma_{\bar{s}}^T\Phi_{\bar{s}}Q_{sm}^{*-1}A_{sm}Q_{sm}^{*-1}\Phi_{\bar{s}}^T\gamma_{\bar{s}}\right], \tag{3.2.9}$$

where $Q_{sm}^{*-1}=\Phi_s^T\Phi_s+\sigma_0^2I_M$ is the Hessian matrix (Böhning, 1992) based on Φ_s with M basis functions. The model MSE of $\hat{\tau}_B(y)$ can be obtain by using the relation $MSE_M\left(\hat{\tau}_B(y)\right)=MSE_M\left(\hat{\tau}_B(y)\right)+\left\{B_M\left(\hat{\tau}_B(y)\right)\right\}^2$. Assuming same prior variance for all superpopulation model parameters as $\Sigma_0=\sigma_0^2I_M$, where I_M is an identity matirx of order M. We can see that $\sigma_0^2\to 0$ results $B_M\left(\hat{\tau}_B(y)\right)\to 0$ and $V_M\left(\hat{\tau}_B(y)\right)\to\sigma^2\left[N-n+\gamma_{\bar{s}}^T\Phi_{\bar{s}}\left(\Phi_s^T\Phi_s\right)^{-1}\Phi_{\bar{s}}^T\gamma_{\bar{s}}\right]$. We generalize the results obtain in Equations (3.2.7)-(3.2.9) by writing the

$$\hat{\tau}_B(y)=\gamma_s^Ty_s+\gamma_{\bar{s}}^T\Phi_{\bar{s}}\left[(I_M-\Lambda_{sm})\hat{\beta}_{ml}+\Lambda_{sm}\mu_0\right] \tag{3.2.10}$$

with prediction error $e(\hat{\tau}_B)=\hat{\tau}_B(y)-\tau(y)=\gamma_{\bar{s}}^T\left[\Phi_{\bar{s}}\left\{(I_M-\Lambda_{sm})\hat{\beta}_{ml}+\Lambda_{sm}\mu_0\right\}-y_{\bar{s}}\right]$, where $\Lambda_{sm}=\sigma^2Q_{sm}^{*-1}\Sigma_0^{-1}$ is the matrix of weights which depends on noise in data, prior variance and available auxiliary data. The model bias for the general form $\hat{\tau}_B(y)$ can be expressed as

$$E_M\left(e(\hat{\tau}_B)\right)=\gamma_{\bar{s}}^T\Phi_{\bar{s}}\Lambda_{sm}\left[\mu_0-\beta\right] \tag{3.2.11}$$

and the model variance of prediction error, is given by

$$V_M\left(e(\hat{\tau}_B)\right)=\sigma^2\left[\gamma_{\bar{s}}^T\gamma_{\bar{s}}+\gamma_{\bar{s}}^T\Phi_{\bar{s}}(I_M-\Lambda_{sm})A_{sm}^{-1}\left(I_M-\Lambda_{sm}\right)^T\Phi_{\bar{s}}^T\gamma_{\bar{s}}\right]. \tag{3.2.12}$$

For $\Sigma_0=\sigma_0^2I_M$ it can be observed that $\sigma_0^2\to 0$ and $\mu_0\to\beta$ result $B_0\left(\hat{\tau}_B(y)\right)\to 0$ and $V_M\left(\hat{\tau}_B(y)-\tau(y)\right)\to\sigma^2\left[\gamma_{\bar{s}}^T\gamma_{\bar{s}}+\gamma_{\bar{s}}^T\Phi_{\bar{s}}\left(\Phi_s^T\Phi_s\right)^{-1}\Phi_{\bar{s}}^T\gamma_{\bar{s}}\right]$. In applications, however, we are forced to take a decision about how to act, i.e. we require an optimal point-like prediction. For this, we need a loss function, $l(\hat{y},y)$, which measures the loss incurred by predicting the value $\hat{y}$ when the actual value is y. To this end, decision theorists have defined many loss functions see (Zellner, 1994). The loss function may be equal to the absolute deviation between the predicted and the true value, called absolute loss function. Symmetric loss functions such as squared error and quadratic loss function (James and Stein, 1992) and their extensions. Asymmetric loss functions such as linear exponential LINEX and modified LINEX loss functions are widely used loss functions Zellner (1986) and their followers.

Note that we computed the predictive distribution without reference to the loss function.

In non-Bayesian framework, typically, the model is trained by minimizing the empirical loss(or risk). On contrary, in the Bayesian paradigm there is a clear difference between the loss function and the likelihood function (which is used for training) . The likelihood function portrays how the noisy observations are deviated from the underlying noise free model. For predictive Gaussian distribution, the mean and the median coincide, hence for any symmetric loss function and predictive Gaussian distribution, we always get $\hat{y}$ as the mean of the posterior predictive distribution. However, in most of the practical problems the loss functions may be asymmetric, and point predictions $\hat{y}$ may be obtained directly from the posterior predictive distribution. Readers can found detailed notes on treatment of decision theory in Berger (2013).

3.3 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. The detail about variance estimation has already been discussed in Section 2.5. We utilize model selection criteria which indirectly provide estimate of error variance σ^2 for obtaining estimate for mean squared prediction error of $\hat{\tau}_B(y)$. It can be seen that the variance of error term given in (3.2.12) depends on error variance σ^2 and the auxiliary data from the whole population and the prior parameters. When we have known sub-matrix of the basis function for the non-sampled part as well and the prior parameters, we need estimate for σ^2 only for estimating prediction variance of $\hat{\tau}_B(y)$. Hence a sample estimate for the prediction error variances given in (3.2.12) can be expressed as:

$$\hat{V}_M\left(e(\hat{\tau}_{_B})\right) = \hat{\sigma}^2\left[\gamma_{\bar{s}}^T\gamma_{\bar{s}} + \gamma_{\bar{s}}^T\Phi_{\bar{s}}(I_M - \Lambda_{sm})A_{sm}^{-1}\left(I_M - \Lambda_{sm}\right)^T\Phi_{\bar{s}}^T\gamma_{\bar{s}}\right]. \tag{3.3.1}$$

Estimation of σ^2 based on residuals is a routine practice in literature of statistical inference. The estimate is taken from the sampled observations or a part of observations (training set) provide a good measure for average noise in the study variable. The residuals under Bayesian framework is different from that of the residuals obtained under frequentist approach.

Consider the residual sum of square as cost function

$$
\begin{aligned}
C^* &= \left(y_s - \hat{y}_s\right)^T \left(y_s - \hat{y}_s\right) \\
&= y_s^T P^* y_s - \sigma^2 \Phi_s Q_s^{*-1} \Sigma_0^{-1} \mu_0,
\end{aligned}
$$

where $P^* = I_M - \Phi_s Q_s^{*-1} \Phi_s^T$ is the projection matrix which is symmetric and idempotent. This expression contains the hyper parameters and the unknown parameter σ^2 when appropriate prior information is available then it is logical to proceed with an iterative method by selecting some initial value of σ^2. When prior is generated from Gaussian with mean vector zero then the last term in C^* vanishes. An estimator of prediction error variance of the estimator $\hat{\tau}_B(y)$ is obtained under residual method by replacing $\hat{\sigma}^2_{res}$ by $\hat{\sigma}^2$ in (3.3.1), where

$$
\hat{\sigma}^2_{res} = \frac{1}{n-M} y_s^T P^{*2} y_s. \tag{3.3.2}
$$

Another most widely used model selection criteria is unbiased estimate of variance (UEV) which is similar to residual variance obtained by replacing the number of total parameter by the number of effective parameter in the denominator. The UEV estimator of σ^2, is given by

$$
\hat{\sigma}^2_{UEV} = \frac{1}{n-M^{**}} y_s^T P^{*2} y_s, \tag{3.3.3}
$$

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 (Zheng and Agresti, 2000) as discussed early. In following subsections, we provide some alternative variance estimation methods. We extend these methods for estimating the error variance in estimation of finite population parameter $\tau(y)$ under Bayesian approach. The projection matrix, say P^*, plays a key role in obtaining the estimate for σ^2 using above mentioned methods. For obtaining estimates for σ^2, we use following model selection criteria.

1. Cross Validation (CV)

2. Generalized Cross Validation (GCV)

3. Final Prediction Error or Akaike's Information Criterion (AIC)

4. Bayesian Information Criterion (BIC).

3.3.1 Estimate based on Cross Validation

The LOO and k-fold cross validation methods are already discussed in Section 2.5. The prediction error variance of the estimator $\hat{\tau}(y)$ in LOO is obtained by replacing $\hat{\sigma}^2$ by $\hat{\sigma}^{*2}_{loo}$ in (3.3.1).

$$\hat{\sigma}^{*2}_{loo} = \frac{1}{n} y_s^T P^* \left\{ \mathrm{diag}(P^*) \right\}^{-2} P^* y_s. \tag{3.3.4}$$

The LOO cross validation is time consuming a tiresome task. To avoid this lengthy process the GCV method is alternative way which also provides a more compact variance estimator as follow.

3.3.2 Estimate based on Generalized Cross validation

The diagonal matrix $\mathrm{diag}(P^*)$ makes LOO mathematically inappropriate. Its alternate, GCV introduced by Golub et al. (1979), is more convenient and is obtained by replacing the matrix $\mathrm{diag}(P^*)$ by the average of the diagonal elements multiplied by the identity matrix I_n of order n i.e. $\mathrm{trace}(P^*/n)I_n$. An estimator for the prediction error variance of $\hat{\tau}(y)$ under GCV is obtained by replacing $\hat{\sigma}^2$ by $\hat{\sigma}^{*2}_{GCV}$ in (3.3.1), where $\hat{\sigma}^{*2}_{GCV}$ is

$$\hat{\sigma}^{*2}_{GCV} = \frac{n y_s^T P^{*2} y_s}{\left\{ \mathrm{trace}(P^*) \right\}^2} \tag{3.3.5}$$

GCV is the one of widely used model selection criteria which includes an adjustment to the average of expected squared prediction error (ESPE) over the training set. It is equivalent to standard residual method given in (3.3.2), when

$$\frac{n}{\left\{ \mathrm{trace}(P^*) \right\}^2} = \frac{1}{n - M^{**}},$$

where $M^{**} = n - \mathrm{sum}(diag(P^*))$ 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 $\mathrm{trace}(P^*)$ as

$$\hat{\sigma}^{2}_{GCV} = \frac{n y_s^T P^{*2} y_s}{\left(n - M^{**} \right)^2}. \tag{3.3.6}$$

3.3.3 Estimate based on Final Prediction Error (FPE)

Mallows's C_p (Mallows, 1973) is a statistic which assess the fit of a regression model that is estimated via ordinary least squares. We extended this method for estimating error variance under Bayesian approach. A brief discussion on FPE is discussed in Section 2.5. An estimator for prediction error variance of $\hat{\tau}_B(y)$ under FPE is obtained by replacing $\hat{\sigma}^2$ by $\hat{\sigma}^{*2}_{FPE}$ in (3.3.1), where $\hat{\sigma}^{*2}_{FPE}$, an alternative version of Mallows's C_p (James et al., 2013), and is defined as:

$$\hat{\sigma}^{*2}_{FPE} = \frac{1}{n}\left(y_s^T P^{*2} y_s + 2M^{**}\hat{\sigma}^{*2}_{res}\right) = \frac{n+M^{**}}{n-M^{**}}\frac{y_s^T P^{*2} y_s}{n}, \tag{3.3.7}$$

where M^{**} is the effective number of parameters.

3.3.4 Estimate Based on Bayesian Information Criterion (BIC)

The BIC developed by (Schwarz et al., 1978), is a Bayesian argument on maximum likelihood of the data. An estimator of variance of prediction error for $\hat{\tau}_B(y)$ based on BIC is obtained by substituting $\hat{\sigma}^{*2}_{BIC}$ by $\hat{\sigma}^2$ in (3.3.1), where $\hat{\sigma}^{*2}_{BIC}$ is defined as:

$$\begin{aligned}\sigma^{*2}_{BIC} &= \frac{1}{n}\left(y_s^T P^{*2} y_s + \ln(n) M^{**}\hat{\sigma}^2_{res}\right)\\ &= \frac{n+M^{**}\left(\ln(n)-1\right)}{n-M^{**}}\frac{y_s^T P^{*2} y_s}{n},\end{aligned} \tag{3.3.8}$$

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

All the mentioned estimators of the variance of prediction error 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 $\sigma^2_{abc} = \Gamma_{abc} y_s^T P^{*2} y_s / n$ and have following natural ordering

$$\Gamma_{UEV} \leq \Gamma_{FPE} \leq \Gamma_{GCV} \leq \Gamma_{BIC} \tag{3.3.9}$$

The factors Γs are approximated by using Taylor's series as already mentioned in Chapter 2.

$$\Gamma_{UEV} = \frac{n}{n-M^{**}} = 1 + \frac{M^{**}}{n} + \frac{M^{*2}}{n} + \frac{M^{*3}}{n} + \ldots.$$

$$\begin{aligned}\Gamma_{FPE} &= \frac{n+M^{**}}{n-M^{**}} = 1+\frac{2M^{**}}{n}+\frac{2M^{*2}}{n^2}+\frac{2M^{*3}}{n^3}+\ldots.\\ \Gamma_{GCV} &= \frac{M^{*2}}{\left(n-M^{**}\right)^2} = 1+\frac{2M^{**}}{n}+\frac{3M^{*2}}{n^2}+\frac{4M^{**3}}{n^3}+\ldots.\\ \Gamma_{BIC} &= \frac{n+\left(\ln(n)-1\right)M^{**}}{n-M^{**}} = 1+\ln(n)\left(\frac{M^{**}}{n}+\frac{M^{*2}}{n}+\frac{M^{**3}}{n}+\ldots.\right)\end{aligned}$$

Hence the estimators of $\hat{\sigma}^2$ obtained through different model selection criteria and can be ranked according to the factor Γ as:

$$\hat{V}\left(e(\hat{\tau}_B)\right)_{UEV} \leq \hat{V}\left(e(\hat{\tau}_B)\right)_{FPE} \leq \hat{V}\left(e(\hat{\tau}_B)\right)_{GCV} \leq \hat{V}\left(e(\hat{\tau}_B)\right)_{BIC}, \tag{3.3.10}$$

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

3.4 Model Selection

Although forward selection is an algorithm of non-linear type, still it has many advantages and listed in Section 2.6. In forward selection, at each step the model grows by one basis function. To see the effect of increase in model with a new basis function, we introduce the incremental operations (see Appendix B). We see the effect of adding a new predictor on the bias and prediction error variance of $\hat{\tau}_B(y)$ in this section. The model-bias of Bayesian estimator $\hat{\tau}_B(y)$ with $(M+1)$ basis functions can be written as:

$$\begin{aligned}B_M[\hat{\tau}_B(y)]_{m+1} =&\sigma^2\Phi_{\bar{s}m}Q_{sm}^{*-1}\Sigma_{0m}^{-1}\left(\mu_{0m}-\beta_m\right)+\frac{\sigma^2}{\triangle_1}\left[\left\{\Phi_{\bar{s}m}Q_{sm}^{*-1}\Phi_{sm}^T\phi_{s(m+1)}\phi_{s(m+1)}^T\Phi_{sm}Q_{sm}^{*-1}\right.\right.\\ &+\left.\left\{\Phi_{\bar{s}m}\phi_{s(m+1)}^T\Phi_{sm}Q_{sm}^{*-1}+\phi_{\bar{s}(m+1)}\right\}\sigma_{0(m+1)}^{-2}\left(\mu_{0(m+1)}-\beta_{m+1}\right)\right].\end{aligned}$$

The absolute change in bias of estimator $\hat{\tau}_B(y)$ when an additional basis function is added to the model

$$\begin{aligned}\left|B_M\left[\hat{\tau}_B(y)\right]_m - B_M\left[\hat{\tau}_B(y)\right]_{m+1}\right| =&\frac{\sigma^2}{\triangle_1}\gamma_{\bar{s}}^T\left[\left\{\Phi_{\bar{s}m}Q_{sm}^{*-1}\Phi_{sm}^T\phi_{s(m+1)}\phi_{s(m+1)}^T\Phi_{sm}Q_{sm}^{*-1}\right.\right.\\ &-\left.\phi_{\bar{s}(m+1)}Q_{sm}^{*-1}\Phi_{sm}^T\phi_{s(m+1)}\right\}\Sigma_{0m}^{-1}\left(\mu_{0m}-\beta_m\right)\\ &-\left\{\Phi_{\bar{s}m}\phi_{s(m+1)}^T\Phi_{sm}Q_{sm}^{*-1}-\phi_{\bar{s}(m+1)}\right\}\sigma_{0(m+1)}^{-2}\left(\mu_{0(m+1)}\right.\\ &\left.\left.-\beta_{m+1}\right)\right].\end{aligned}$$

The variance of the prediction error of $\hat{\tau}_B(y)$ after adding a new basis function is as follow:

$$\begin{aligned} V_M\big(e(\hat{\tau}_{_B})\big)_{m+1} = \sigma^2 \big[\gamma_{\bar{s}}^T \gamma_{\bar{s}} + \gamma_{\bar{s}}^T \Big(& \Phi_{\bar{s}m} \Lambda_{11}^{**} \Phi_{\bar{s}m}^T + \phi_{\bar{s}(m+1)} \Lambda_{21}^{**} \Phi_{\bar{s}m}^T + \Phi_{\bar{s}m} \Lambda_{12}^{**} \phi_{\bar{s}(m+1)}^T \\ & + \phi_{\bar{s}(m+1)} \Lambda_{21}^{**} \phi_{\bar{s}(m+1)}^T \Big) \gamma_{\bar{s}} \big], \end{aligned} \tag{3.4.1}$$

where $\Lambda_{s(m+1)}$ is already defined in (B.2) (see Appendix B). The decrement in variance or increment in efficiency IE_B is defined by

$$\begin{aligned} IE_B &= V_M\big(e(\hat{\tau}_{_B})\big)_m - V_M\big(e(\hat{\tau}_{_B})\big)_{m+1} \\ &= -\sigma^2 \gamma_{\bar{s}}^T \Big[\Phi_{\bar{s}m} \Big(\Lambda_{11}^{-1} \triangle^{-1} A_{sm}^{-1} \Phi_{sm}^T \phi_{s(m+1)} \phi_{s(m+1)}^T \Phi_{sm} A_{sm}^{-1} \Lambda_{11}^{-1} \\ &\quad + \Lambda_{12}^{-1} A_{21}^{-1} \Lambda_{11}^{-1} + (\Lambda_{11}^{-1} A_{12}^{-1} + \Lambda_{12}^{-1} A_{22}^{-1}) \Lambda_{12}^{-1} \Big) \Phi_{\bar{s}m}^T + \phi_{\bar{s}(m+1)} \Lambda_{21}^{**} \Phi_{\bar{s}m}^T \\ &\quad + \Phi_{\bar{s}m} \Lambda_{12}^{**} \phi_{\bar{s}(m+1)}^T + \phi_{\bar{s}(m+1)} \Lambda_{21}^{**} \phi_{\bar{s}(m+1)}^T \Big] \gamma_{\bar{s}} \end{aligned}$$

The detail description of the expression IE_B and its derivation are available in (B.1), (B.3) and (B.4) (see Appendix B). A positive value of IE_B provides the evidence of increase in efficiency on using an additional basis function. In this way, for reaching to a suitable model for predicting non-sampled value in finite populations one can take advantage from the value of IE_B. The incremental operators helps in computing the inverse of matrices more quickly which leads to quick decision in choosing models. In next section, we see the empirical evidences and effect of changing prior parameters and design parameters on efficiency.

3.5 Simulations

Two simulation studies are conducted to evaluate the error variance of the proposed estimator of $\tau(y)$ ($\gamma_i = 1$ for all $i \in \mathcal{U}$) and to find expected values of the estimated error variance of $\hat{\tau}_B(y)$. For this purpose, firstly, we provide a simulation study using artificially generated population and fitting basis functions. Secondly, we perform repeated sampling to obtain design-based properties of the estimator and estimated error variance of the estimator $e(\hat{\tau}_B)$ using the Hospitals data from Valliant (2000). The simulation study is performed using following algorithm.

(i). To constitute a population showing non-linear behavior, draw two independent vectors

u^* and v^* of length $N = 1000$ each with uniform $U(0,1)$. 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 $N(0,5)$ respectively. We generate the vector of the study variable y as a unique function $y = \sin(2\pi x) + e$. 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 frequentist design-bias and design-expected prediction error to see the behavior of the proposed estimator $\hat{\tau}_B(y)$.

(ii). For fixed n, we divide the generated data $df(y,x,x^2,x^3,...x^{M-1})$ (where M is the number of basis function including intercept 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 (β and σ^2) under Bayesian approach. For known variance structure, the prior for parameters β is considered Gaussian with hyper parameters $\mu_0 = (0\ \ 0\ \ 0)^T$ and variance $\Sigma_0 = \sigma_0^2 I_M$ (σ_0^2 is taken as different proportions of σ^2 obtained from population). After estimating β and fitting Bayesian regression, the estimated values of σ^2 are obtained under different model selection criteria discussed in Section 3.5.

(iii). Further, we evaluate the the proposed estimator $\hat{\tau}_B(y)$ (with $\gamma_i = 1$ for all $i \in \mathcal{U}$) and the estimated variance of $\hat{\tau}_B(y)$ under different formula.

(iv). Repeat Steps (ii) and (iii), 20,000 time to design-expected prediction error (i.e. bias) and design-expected squared prediction error of $\hat{\tau}_B(y)$ ($\gamma_i = 1$ for all $i \in \mathcal{U}$) and expected values of the estimated variance of $\hat{\tau}_B(y)$ for different combinations of M, n and variance ratio $\sigma_0^2/\sigma^2 = v_r$ (say).

For bootstrapping, we take first 203 hospitals from hospital data given in (Valliant, 2000, Appendix B Page 424) as the study population. The number of beds (x) in each hospital is considered as the predictor for the number of patients discharged (y). Repeated sampling, as early mentioned for hypothetical population, is performed to study the design-based properties of total estimator and estimated error variance. Expected squared prediction error

ESPE are obtained as:

$$\text{ESPE}_B = \frac{1}{20,000} \sum_{sim} \left\{ e(\hat{\tau}_{_B}) \right\}^2$$

respectively for $v_r = 0, 0.1, 0.3, 0.5$ (the choice 0 represents classical estimator), where the $\sum_{sim}$ is used to denote that summation is taken over all 20,000 simulated samples. Further the design expectation of estimated prediction error variance of $\hat{\tau}_B(y)$ are obtained through respective formula after averaging over all selected samples. We use polynomial basis functions of different orders with intercept and without intercept.

Table 3.1.: Simulated results for linear basis function model under Bayesian approach

			Expected estimated variances				
n	v_r	ESPE	RES	UEV	FPE	GCV	BIC
				$M = 1$			
	0	33400.39	225969	228251.52	230534.03	230557.09	236480.37
100	0.1	31738.1	22616.37	22839.12	23061.86	23064.06	23642.15
	0.3	28761.16	22615.5	22827.65	23039.79	23041.78	23592.47
	0.5	26179.46	22615.55	22818.07	23020.58	23022.39	23548.17
	0	28629.98	100934.55	101441.76	101948.97	101951.52	103621.91
200	0.1	28141.07	20188.89	20289.07	20389.24	20389.74	20719.66
	0.3	27210.92	20188.89	20286.61	20384.33	20384.8	20706.63
	0.5	26339.65	20188.98	20284.36	20379.74	20380.19	20694.33
				$M = 2$			
	0	15400.784	224291.06	228868.43	233445.8	233539.21	245370.62
100	0.1	9879.433	22242.37	22685.19	23128	23136.82	24281.61
	0.3	3149.375	22240.55	22664.07	23087.58	23095.65	24190.92
	0.5	378.7548	22242.83	22650.27	23057.7	23065.17	24119.14
	0	16730.15	100153.4	101165.08	102176.73	102186.95	105513.47
200	0.1	18394.1	19934	20132.88	20331.76	20333.74	20987.72
	0.3	21693.64	19934.4	20128.67	20322.95	20324.84	20963.73
	0.5	24929.27	19935.41	20125.49	20315.58	20317.4	20942.55
				$M = 3$			
	0	14537.493	225482.17	232455.85	239429.52	239645.2	257597.13
100	0.1	8776.893	22052.92	22723.25	23393.57	23413.95	25139.89
	0.3	2192.927	22050.6	22700.61	23350.62	23369.78	25044.01
	0.5	54.50894	22052.69	22685.86	23319.03	23337.21	24968.54
	0	18371.14	100359.67	101887.99	103416.31	103439.58	108457.19
200	0.1	20138.01	19856.06	20155.88	20455.71	20460.24	21444.62
	0.3	23633.48	19856.45	20151.54	20446.62	20451.01	21419.9
	0.5	27052.27	19857.46	20148.24	20439.02	20443.27	21398.09
				$M = 4$			
	0	31220.73	228790.27	238323.19	247856.12	248253.3	272691.02
100	0.1	11316.72	21870.02	22765.19	23660.36	23697	25992.42
	0.3	16630.67	21865.76	22734.32	23602.89	23637.39	25865.64
	0.5	12156.44	21868.72	22716.55	23564.37	23597.24	25773.09
	0	17187.2	100787.16	102844.04	104900.92	104942.9	111685.16
200	0.1	18518.49	19779.89	20180.07	20580.24	20588.34	21900.15
	0.3	21062.11	19780.53	20174.39	20568.24	20576.09	21867.31
	0.5	23446.66	19782.14	20170.45	20558.75	20566.37	21839.51

Table 3.2.: Simulated results for proportional basis function model under Bayesian approach

			Expected estimated variances				
n	v_r	ESPE	RES	UEV	FPE	GCV	BIC
$M=2$							
100	0	182127	122370.54	123606.61	124842.68	124855.16	128062.84
	0.1	182076.6	22515.85	22743.25	22970.65	22972.95	23563.08
	0.3	181975.7	22515.85	22743.19	22970.53	22972.83	23562.79
	0.5	181875	22515.85	22743.13	22970.41	22972.7	23562.51
200	0	145846.8	59467.49	59766.33	60065.16	60066.66	61050.8
	0.1	145831.5	20106.51	20207.54	20308.57	20309.08	20641.8
	0.3	145801	20106.51	20207.52	20308.54	20309.05	20641.73
	0.5	145770.5	20106.51	20207.51	20308.51	20309.02	20641.66
$M=3$							
100	0	211862.6	124781.42	127328	129874.54	129926.51	136508.76
	0.1	211714.9	22318.67	22774	23229.33	23238.62	24415.55
	0.3	211420.5	22318.67	22773.7	23228.73	23238	24414.15
	0.5	211127.2	22318.67	22773.4	23228.12	23237.39	24412.75
200	0	153348.4	59927.89	60533.23	61138.56	61144.67	63135.14
	0.1	153316.5	20025.28	20227.52	20429.76	20431.81	21096.83
	0.3	153252.8	20025.28	20227.46	20429.64	20431.68	21096.49
	0.5	153189.3	20025.28	20227.39	20429.51	20431.55	21096.16
$M=4$							
100	0	113100.2	157002.21	161857.95	166713.69	166863.87	179363.7
	0.1	113130.7	22113.28	22796.99	23480.71	23501.85	25261.9
	0.3	113191.4	22113.28	22796.59	23479.9	23501.02	25260.05
	0.5	113251.8	22113.28	22796.2	23479.11	23500.2	25258.21
200	0	80991.21	71881.62	72976.27	74070.91	74087.58	77681.4
	0.1	80969.58	19939.16	20242.77	20546.38	20551	21547.77
	0.3	80926.37	19939.17	20242.7	20546.24	20550.86	21547.39
	0.5	80883.23	19939.17	20242.63	20546.09	20550.71	21547.02

Table 3.3.: Bootstrapped results for linear basis function model under Bayesian approach

			Expected estimated variances				
n	v_r	ESPE	RES	UEV	FPE	GCV	BIC
				$M = 1$			
50	0	6.675544	1115.06466	1137.82108	1160.5775	1161.04192	1204.08831
	0.1	34627.4285	504.38388	504.41178	504.43968	504.43968	504.49302
	0.3	34753.6887	505.23336	505.24269	505.25202	505.25202	505.26987
	0.5	34779.02438	505.40382	505.40943	505.41503	505.41503	505.42574
100	0	197.1884	375.17234	378.96196	382.75158	382.78986	392.62419
	0.1	17279.87453	336.4681	336.48666	336.50522	336.50522	336.55357
	0.3	17393.99011	337.58962	337.59585	337.60208	337.60208	337.61831
	0.5	17416.95969	337.81542	337.81916	337.8229	337.8229	337.83265
				$M = 2$			
50	0	379.38659	842.41857	877.51935	912.62012	914.08265	979.73361
	0.1	648.78412	203.74991	207.81552	211.88113	211.96225	219.65466
	0.3	979.69691	204.85014	208.75659	212.66304	212.73754	220.13226
	0.5	1348.43147	206.79044	210.56796	214.34549	214.4145	221.5682
100	0	633.51469	280.12628	285.84314	291.56	291.67667	306.4534
	0.1	720.06036	136.72265	138.08985	139.45705	139.47073	143.01885
	0.3	838.58638	136.918	138.25534	139.59267	139.60574	143.07666
	0.5	960.89769	137.28418	138.59557	139.90695	139.91948	143.32333
				$M = 3$			
50	0	275.66837	845.48914	899.45653	953.42392	956.86865	1056.61082
	0.1	794.55001	202.24116	209.70549	217.16982	217.44534	231.4418
	0.3	1719.64414	207.99479	214.61123	221.22766	221.43819	233.87844
	0.5	2345.03606	212.70029	218.89488	225.08947	225.26993	236.93367
100	0	627.25487	280.21823	288.88477	297.55131	297.81935	320.12913
	0.1	794.37098	136.06582	138.67091	141.27599	141.32587	148.06268
	0.3	1100.72528	137.68441	140.0744	142.46439	142.50588	148.69073
	0.5	1337.53684	139.52428	141.77933	144.03439	144.07084	149.90919
				$M = 4$			
50	0	289.43681	868.14337	943.6341	1019.12483	1025.68924	1163.46484
	0.1	641.80331	199.46403	209.36151	219.25899	219.75016	238.1832
	0.3	782.94262	200.50475	209.5589	218.61305	219.02192	235.92479
	0.5	825.38163	200.85444	209.67224	218.49005	218.87717	235.34988
100	0	622.21693	283.01128	294.80342	306.59556	307.08689	337.31608
	0.1	799.03316	135.48904	139.02182	142.55461	142.64672	151.75811
	0.3	905.32574	136.14561	139.32538	142.50515	142.57942	150.789
	0.5	944.41637	136.44744	139.50386	142.56028	142.62875	150.52278

Table 3.4.: Bootstrapped results for proportional basis function model under Bayesian approach

			Expected estimated variances				
n	v_r	ESPE	RES	UEV	FPE	GCV	BIC
$M = 2$							
50	0	501.5782	676.1110	689.9092	703.7074	703.9890	730.0899
	0.1	649.0696	203.7508	207.8135	211.8762	211.9572	219.6442
	0.3	979.9622	204.8507	208.7562	212.6617	212.7361	220.1290
	0.5	1348.7263	206.7911	210.5680	214.3449	214.4139	221.5665
100	0	662.6404	242.9413	245.3952	247.8492	247.8740	254.2421
	0.1	720.2201	136.7232	138.0884	139.4537	139.4673	143.0105
	0.3	838.7315	136.9183	138.2550	139.5917	139.6047	143.0740
	0.5	961.0491	137.2845	138.5955	139.9065	139.9190	143.3218
$M = 3$							
50	0	246.1663	768.1118	800.1164	832.1211	833.4546	893.3147
	0.1	794.7564	202.2413	209.7043	217.1673	217.4427	231.4368
	0.3	1720.0348	207.9963	214.6120	221.2278	221.4383	233.8773
	0.5	2345.4660	212.7025	218.8966	225.0907	225.2711	236.9339
100	0	609.6638	264.5690	269.9684	275.3678	275.4780	289.4340
	0.1	794.4958	136.0658	138.6701	141.2744	141.3242	148.0591
	0.3	1100.9179	137.6847	140.0743	142.4640	142.5055	148.6895
	0.5	1337.7592	139.5251	141.7799	144.0347	144.0711	149.9089
$M = 4$							
50	0	290.9959	827.6359	880.4637	933.2915	936.6635	1034.2995
	0.1	641.9711	199.4649	209.3609	219.2570	219.7480	238.1784
	0.3	783.0522	200.5055	209.5589	218.6123	219.0211	235.9227
	0.5	825.4590	200.8550	209.6723	218.4896	218.8767	235.3485
100	0	627.0282	275.2336	283.7459	292.2583	292.5216	314.4345
	0.1	799.1815	135.4894	139.0215	142.5537	142.6458	151.7555
	0.3	905.4390	136.1462	139.3256	142.5049	142.5792	150.7878
	0.5	944.5021	136.4479	139.5041	142.5602	142.6287	150.5220

Tables 3.1. and 3.2. provide the efficiency comparison for the hypothetically generated data while Tables 3.3. and 3.3. present the same comparison for the hospital data obtained from Valliant (2000). First and second columns of each table give the choices of sample size n and the variance ratio v_r. The behavior of ESPE differs for different choices of the number of basis functions M in Table 3.1.. For example, for $M = 1$ (Table 3.1.) the ESPE tends to decrease with increase in v_r when $n = 100$. While for $M = 2,3,4$, the ESPE for

similar combination the ESPE drastically decreases with increase in v_r when $n = 100$ and increases with increase in v_r for $n = 200$. For fixed choice of v_r one can distill from Table 3.1. that ESPE tends to decline when sample size rises. This is because the fact that the design bias of the estimator $\hat{\tau}_B(y)$ and/or variance become smaller with increase in sample size. Which leads to the property of consistency of the estimator. Although the biases of the proposed estimator is computed but not reported here for the sake of space. Table 3.2. presents the results for the proportional basis function models (models without intercept) in similar manner as discussed above. For example, with $M = 4$, $n = 100$, the ESPE values are 161857.9 for the classical estimator (i.e. for $v_r = 0$) . While ESPE values are 22797, 22796.6 and 22796.2 for $v_r = 0.1$, $v_r = 0.3$ and $v_r = 0.5$ respectively. This implies change in prior variance even by a large amount does not alter the ESPE value. Same interpretation can be made for other choices of M. The values of ESPE are observed almost stable for all choices of the variance ratio v_r. On the other hand, ESPE tends to inclined with increase in sample size n. For bootstrapped results (see Tables 3.3. and 3.4.), on can observe that ESPE rises with increase in v_r for all choices of M and n. However for smaller n the increase in ESPE is higher than the increment with larger n. For example for $M = 4$ in Table 3.3., ESPE goes from 275.66 to 794.55 for $n = 50$, while it goes from 627 to 794.36 for $n = 100$. The ESPEs are observed higher for $n = 100$ as compared to $n = 50$ this contradicts the result given in simulation study.

The expected values of estimated error variance of $\hat{\tau}_B(y)$ under different model selection criteria are observed from left to right, in ascending order, of each table. The residual method dispense smaller values of the expected estimates as compared to other estimators. The highest expected variance is observed for BIC. For numerical study with $M = 2$, $n = 50$ and $v_r = 0$ the estimated variances are 676.11, 689.9, 703.7, 703.99 and 730.1 which satisfies the inequality given in (3.3.10). Additionally, we can infer the following relation from our empirical evidences $\hat{V}(e(\hat{\tau}_B))_{RES} \leq \hat{V}(e(\hat{\tau}_B))_{UEV} \leq \hat{V}(e(\hat{\tau}_B))_{FPE} \leq \hat{V}(e(\hat{\tau}_B))_{GCV} \leq \hat{V}(e(\hat{\tau}_B))_{BIC}$. Tables 3.1.–3.4. also provide the evidence that estimated variances decrease with increase in v_r. 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. The unbiasedness and consistency of the variance estimators are good tools in this regard. However, these properties are not discussed in this study as our goal was basically construction of estimator τ and discussing problems associated with its estimation and inference about the finite population in model-based setting.

3.6 Conclusion

Unlike to classical school of thought, practitioners in Bayesian statistics emphasizes on utilization of prior information about the parameter in estimating their posterior parameters. This chapter extended the basis function approach for estimating finite population parameter τ under Bayesian point of view. Results (ESPE and Expected estimated error variance) under Bayesian approach is then compared with results obtained without using prior information ($\sigma^2 = 0$ versus $\sigma_0^2 = 0.1\sigma^2$, $0.3\sigma^2$ and $0.5\sigma^2$). The result in all tables depicts the superiority of estimators under Bayesian paradigm. Increase in σ_0^2 results in decrease in variance but introduce bias in the estimator and makes a trade-off in expected squared prediction error (ESPE). Different variance estimators were established for estimating prediction error variance. We also established an ordering between the expected values of estimated variance which is also proved in simulation studies $\hat{V}(e(\hat{\tau}))_{RES} \leq \hat{V}(e(\hat{\tau}))_{UEV} \leq \hat{V}(e(\hat{\tau}))_{FPE} \leq \hat{V}(e(\hat{\tau}))_{GCV} \leq \hat{V}(e(\hat{\tau}))_{BIC}$. Tables 3.1.–3.4. also provide the evidence that estimated variances decrease with increase in v_r. Similar statement can be made for the relation between estimated variances and sample size (except for $M = 4$ in Table 3.4.).

Chapter 4

Model-Based Estimation of Parameters Sub-Sampling Non-respondents

4.1 Outline

The problem of handling non-ignorable non-response has been typically addressed under the design-based approach using the well-known sub-sampling technique introduced by Hansen and Hurwitz (1946). Alternatively, the model-based paradigm emphasizes on utilizing the underlying model relationship between the outcome variable and one or more covariate(s) whose population values are known prior to the survey. This chapter utilizes the model relationship between the study variable and covariate(s) for handling non-ignorable non-response and obtaining an unbiased estimator for the population total under the sub-sampling technique. In this chapter, a model unbiased linear predictor for the population total in presence of non-ignorable non-response is proposed assuming unit non-response. The sub-sampling technique introduced by Hansen and Hurwitz (1946) is used to obtain samples under a fixed sampling design. We provide a revision of model-based approach for estimation of finite population total in Section 4.2. Our proposed estimator and its properties under assumed model are given in Section 4.3. Some shortcomings of the proposed estimation technique and their possible solutions are discussed in Section 4.4. A numerical study with real data set and a Monte Carlo simulation are respectively provided in Sections 4.5 and 4.6. A discussion with concluding remarks is given in Section 4.7.

4.2 Model-based Estimation of Population Total

Again consider a finite population of N distinct units $U=\{1,2,..i..,N\}$. Let $y=\left(y_i, i\in U\right)$ be the vector of the realized values of a stochastic vector $Y=\left(Y_i, i\in U\right)$ of order $(N\times 1)$ and $x=\left(x_{ij}, i\in U, j=0,1,2,...,p\right)$ be a matrix of (p+1) auxiliary variables whose values are assumed to be known for every unit in U. We start with multiple linear regression model $Y=x\beta+\varepsilon$, where $\beta=\left(\beta_0,\beta_1,....,\beta_p\right)^T$ and $\varepsilon=\left(\varepsilon_1,\varepsilon_2,....,\varepsilon_N\right)^T$ be the vectors of regression coefficients and the random error terms respectively. Let $s=\{1,2,3,...,n\}$ be a member of S of all possible samples of size n that can be drawn from U using some sampling design (SD). Further, the random vector of the study variable Y, the known auxiliary matrix x and the random error vector ε are splitted into sampled (s) and non-sampled $(\bar{s})$ as: $Y=\left(Y_s;Y_{\bar{s}}\right)^T$, $x=\left(x_s;x_{\bar{s}}\right)^T$ and $\varepsilon=\left(\varepsilon_s;\varepsilon_{\bar{s}}\right)^T$, where $\bar{s}=U-s$. The population total t_y (which is assumed to be random under model-based approach) is expressed as $t_y=\gamma_s^T Y_s+\gamma_{\bar{s}}^T Y_{\bar{s}}$, where $\gamma=(\gamma_i, i\in U)$ is the vector containing 1's for every units in population. For obtaining population mean γ are taken as vector of $1/N$ for all units. For further statistical inference about the estimated parameter assumption of normally distributed error term is also necessary specially in case of small sample sizes. After observing y_s as the realized values of Y_s, the problem is to predict sub-vector $Y_{\bar{s}}$ using the information contained in the sample and the auxiliary information through model relationship between the study variable and the auxiliary variable(s). Under linear population model, a predictor for $Y_{\bar{s}}$ is $x_{\bar{s}}\hat{\beta}$, where the vector $\hat{\beta}=\left(\hat{\beta}_0,\hat{\beta}_2,....,\hat{\beta}_p\right)^T$ is the solution of the normal equations $x_s^T x_s\hat{\beta}=x_s^T y_s$ which is obtained by minimizing the sum of squared residuals. The model-based estimator given in Mukhopadhyay (1993) is

$$\hat{t}_y=\gamma_s^T y_s+\gamma_{\bar{s}}^T x_{\bar{s}}\hat{\beta}. \tag{4.2.1}$$

Note that the total estimator given in (4.2.1) works only when error terms are iid with zero mean and constant variance Bellhouse (1987). $\hat{t}_y$ posses all the properties with respect to the model as the predictor of $y_{\bar{s}}$ does Royall and Cumberland (1981). When all OLS assumptions fulfill the estimator $\hat{t}_y$ is model unbiased with the model-variance after averaging over all possible sample of same SD.

$$E_D\{V_M(\hat{t}_y)\}=\sigma^2 E_D\left[\gamma_{\bar{s}}\{x_{\bar{s}}\left(x_s^T x_s\right)^{-1}x_{\bar{s}}^T+I_{N-n}\}\gamma_{\bar{s}}^T\right], \tag{4.2.2}$$

where the subscript D is used to show that the expectation is applied with respect to SD and I_{N-n} is the identity matrix of order $(N-n)\times(N-n)$. It is noteworthy that model-variance of $\hat{t}_y$ itself is assessed instead of computing prediction error variance like in other chapters. Setting $p=0$, the linear regression model reduces to homogeneous population model i.e. $Y=x_0\beta_0+\varepsilon$, where x_0 is vectors of 1's. Care should be taken while selecting a suitable set of predictors which comes under the domain of variable selection (inclusion and exclusion) Rawlings et al. (2001). Moreover, when variance of the error term depends on some function of the auxiliary variable(s), weighted least square (WLS) estimator is preferred for estimating β as alternative to OLS. Moreover, if the number of regressors exceeds the number of observations in the sample then ridge regression is preferred (Draper and Smith, 2014; Bellhouse, 1987, pp 313-323). We discuss these problems for our proposal later in Section 4.4.

4.3 Model-based Estimation of Population Total in Presence of Non-response

In voluntary surveys, a common threat to the validity of the survey estimates is the problem of non–response. Different surveys possess different response rates, the surveys that ask questions which seem interesting and relevant to the respondents are tend to achieve the highest response rates. In recent years, response rates have been declined even in popular surveys, and, as a consequence, worries about non-response bias have been increased. As we discussed in introduction section that non-response is considered as problematic only if the population of non-respondents is an informative sample of the total sample. Unfortunately, this appears almost in majority of practical applications. In household surveys, for instance, there is a lot of evidence that non-respondents are often younger than respondents, and that women are more likely to persuade to take part than men. Similarly, response rates are also tend to be lower in deprived areas than the areas with abundance of facilities. All of these examples show that the pattern of achieved samples for surveys mostly do not reflect the population that is meant to represent very well. These surveys typically may over-represent women, and the persons elder than certain age. And often under-represent those living in less developed cities and deprived areas. When values of such demographic variable(s) are known for whole target population, we can stratify the population as the respondents and the non-respondents. The problem is then to choose a variable which more accurately stratifies the population as respondents and non-respondents. Suppose that R is a stratification vector defined as $R = \left(R_i, i \in U\right)$, where $R_i = 1(0)$ according to the ith unit belongs to the population of respondents (non-respondents). In case of missing completely at random (MCAR), non-response factor R and the study variable Y are uncorrelated and one can ignore the non-response or just apply different imputation techniques Holt and Elliot (1991). When the stratification variable R is related to the study variable Y, the model for the respondents differs from that of the non-respondents such as in above example the population models may differ among men and women, youngers and elders and deprived and settled areas. To capture this difference, we specify the model of respondents and non-respondents in the population separately according to the values of R such that

$$Y_1 = x_1\beta_1 + \varepsilon_1 \text{ for } R_i = 1 \tag{4.3.1}$$

$$Y_2 = x_2\beta_2 + \varepsilon_2 \text{ for } R_i = 0 \text{ for } i \in U, \tag{4.3.2}$$

where β_1 and β_2 are the vectors of regression coefficients corresponding to the respondents and the non-respondents respectively. Consequently, we get sub-populations U_1 and U_2 such that $U = U_1 \cup U_2$, where U_1 and U_2 are the subsets of U denoting populations of respondents and non-respondents with sizes N_1 and N_2 respectively. It is assumed that the error terms are independently and identically distributed (IID) with means $E_M(\varepsilon_1) = E_M(\varepsilon_2) = 0$ with model variances $V_M(\varepsilon_1) = \sigma_1^2 I_{N1}$, and $V_M(\varepsilon_2) = \sigma_2^2 I_{N2}$, where I_{N_1} and I_{N_2} are the identity matrix of order N_1 and N_2 respectively. Separation of model is straight forward when we have exact knowledge about the occurrence of non-response and a related stratification variable which is almost impossible in real world problem. As it is not possible to have such information that separates the underlying model exactly into the respondents and the non-respondents. One way to overcome this problem may be to use two phase sampling for obtaining information on stratification variable. Under two-phase sampling, we select a larger sample on first phase and observe the stratification variable (i.e. respondents are marked as respondents according to their behavior to respond the first phase survey are observe such factor which cause non-response) and estimate the proportions of units fall in sub-populations i.e. $\lambda_1 = N_1/N$ and $\lambda_2 = N_2/N$. These information then can be used at second phase for estimating population parameters of the study variable. Before going toward our proposal, we discuss the estimation of population total without sub-sampling non-respondents which help us in knowing how the non-response creates biasedness in estimation of total.

4.3.1 Estimation of Total Without Sub-sampling

For a sample s of size n assume that only n_1 units respond while remaining n_2 units don't respond. The prediction problem given in Section 4.2 becomes $t_y = \gamma_{s_1}^T Y_{s_1} + \gamma_{\bar{s}_1}^T Y_{\bar{s}_1} + \gamma_2^T Y_2$, where $\gamma_{s_1}^T$, $\gamma_{\bar{s}_1}^T$, and γ_2^T, are vectors of weights associated with n_1 respondents, $N_1 - n_1$ non-sampled units from responding population, and N_2 units from non-responding population respectively. Further $\gamma_{\bar{s}_1}^T Y_{\bar{s}_1} + \gamma_2^T Y_2$ is unknown and can be predicted using sample at hand and the auxiliary information for the non-responded and non-sampled values. A predictive estimator for population total based on respondents only, can be found as follow:

$$\hat{t}_{y1} = \gamma_{s_1}^T y_{s_1} + \gamma_{\bar{s}_1}^T x_{\bar{s}_1} \hat{\beta}_1 + \gamma_2^T x_2 \hat{\beta}_1, \tag{4.3.3}$$

where $\hat{\beta}_1$ is the vector of OLS estimates of β_1 based on n_1 respondents. The model bias of $\hat{t}_{y1}$ is

$$B_M(\hat{t}_{y1}) = \gamma_{s_1}^T x_2 \left(\beta_1 - \beta_2\right). \tag{4.3.4}$$

See Appendix C for proof. $\hat{t}_{y1}$ is unbiased estimate of t_y if the vectors of coefficients for the responding and non-responding sub-populations are same i.e. $\beta_1 = \beta_2$, this is equivalent to regression imputation. This situation occurs when behavior of the responding and the non-responding populations are same allowing us to ignore the non-response just as reduced sample size. We obtain model mean squared error (M-MSE) of the total estimator $\hat{t}_{y1}$ as:

$$\begin{aligned} MSE_M(\hat{t}_{y1}) =& \left\{B_M(\hat{t}_{y1})\right\}^2 + V_M(\hat{t}_{y1}) \\ =& \left\{B_M(\hat{t}_{y1})\right\}^2 + \sigma_1^2\left(n_1 + \gamma_{\bar{s}_1}^T x_{\bar{s}_1} \left(H_{s_1}\right)^{-1} x_{\bar{s}_1}^T \gamma_{\bar{s}_1}\right) \\ & + \sigma_2^2\left(\gamma_2^T x_2 \left(H_{s_1}\right)^{-1} x_2^T \gamma_2\right). \end{aligned} \tag{4.3.5}$$

The subscript M shows that expectation is applied over model. The model-mean squared error (M-MSE) given in (4.3.5) depends on random sample under designed-based point of view. Consequently, it varies with sampling fluctuations. To obtain a fix value, we apply expectation with respect to SD.

4.3.2 Estimation of Total With Sub-sampling

As we already discussed, there are several approaches for handling the problem of non-response in sample literature. A suitable approach may be chosen according to the type of non-response (full or partial), the accessibility of the auxiliary variable(s) and the validity of the underlying response model for handling the problem. In general, re-weighting is used to deal with full (non-availability of units) non-response. Imputation is preferably applied for dealing with partial non-response although it can be applied for full non-response if appropriate auxiliary information is available. Re-weighting eliminates or at least reduces total non-response bias (Särndal, 2007; Holt and Elliot, 1991). While the sub-sampling method introduced by Hansen and Hurwitz (1946) provides a good adjustment for non-response bias and yield unbiased estimator for the population mean when the non-response variable R is significantly correlated with the survey outcome.

In this study, we develop a model-based estimator for population total by adjusting non-response using sub-sampling procedure. As the models described in (4.3.1) and (4.3.2) have different parameters it is inevitable to obtain information about both sub-populations. The sample information obtained from respondents alone leads to biased estimate for the population total of the whole population. For estimating the relationship between the study and the auxiliary variables for the population of non-respondents and estimating total, we need some information from non-respondents as well. The sampling mechanism in Section 4.3 is based on the respondents from first sample which don't provide any information about the population model of non-respondents. The sub-sampling introduced by Hansen and Hurwitz (1946) is the best alternative to handle such situation of non-response which assumes the mode of data collection on first round was inexpensive and then a more stronger mode of interview is employed for sub-sampling non-respondents. The rationale behind taking a sub-sample instead of following all non-respondent is the fact that taking information from all non-respondents by using stronger mode of interview increases survey cost. Sometime randomized response techniques (more expensive and complex method) are applied to gather information on second call Ahmed et al. (2017) . The method assumes sub-sampling $\acute{n}_2 = \frac{n_2}{k}$ $(k > 1)$ units from n_2 units selected and not respond on first round, using some stronger mode of interview (face to face survey, telephonic survey etc). The estimation process covers two prediction problems (i) predicting $N_1 - n_1$ non-sampled units from the sample taken from the first round using model given in (4.3.1) and (ii) predicting $N_2 - n_2$ (non-sampled)+$n_2 - \acute{n}_2$ (non-responded) units on the basis of sample obtained on second round using the model relationship given in (4.3.2). Let $\acute{s}_2$ be the sub-sample of size $\acute{n}_2$ selected from s_2 and $\acute{\bar{s}}_2 = U_2 - \acute{s}_2$ be the set representing non-sampled values from the population of non-respondents. Now the outcome vector for respondents is further partitioned as $Y_1 = \left(Y_{s_1} : Y_{\bar{s}_1}\right)^T$ and for non-respondents $Y_2 = \left(Y_{\acute{s}_2} : Y_{\acute{\bar{s}}_2}\right)^T$. The matrix x, the vector γ and the random error vector ε are also partitioned into sampled and non-sampled parts in same way. The population total of the study character is now expressed as $t_y = \gamma^T_{s_1} Y_{s_1} + \gamma^T_{\bar{s}_1} Y_{\bar{s}_1} + \gamma^T_{\acute{s}_2} Y_{\acute{s}_2} + \gamma^T_{\acute{\bar{s}}_2} Y_{\acute{\bar{s}}_2}$ after replacing known values of the response units, we have $t_y = \gamma^T_{s_1} y_{s_1} + \gamma^T_{\bar{s}_1} Y_{\bar{s}_1} + \gamma^T_{\acute{s}_2} y_{\acute{s}_2} + \gamma^T_{\acute{\bar{s}}_2} Y_{\acute{\bar{s}}_2}$. The problem is to predict $\gamma^T_{\bar{s}_1} Y_{\bar{s}_1} + \gamma^T_{\acute{\bar{s}}_2} Y_{\acute{\bar{s}}_2}$. The first part is predicted on the basis of sample obtained on first round along with model given in (4.3.1) and the second part is predicted on the basis of sample obtained on second round and the model given in (4.3.2). Under the sub-sampling technique a linear unbiased predictor for t_y is

$$\hat{t}^*_y = \gamma^T_{s_1} y_{s_1} + \gamma^T_{\bar{s}_1} x_{\bar{s}_1} b_r + \gamma^T_{\acute{s}_2} y_{\acute{s}_2} + \gamma^T_{\acute{\bar{s}}_2} x_{\acute{\bar{s}}_2} \hat{\beta}_2, \tag{4.3.6}$$

where $\gamma^T_{s_1}$, $\gamma^T_{\bar{s}_1}$,$\gamma^T_{\acute{s}_2}$ and $\gamma^T_{\acute{\bar{s}}_2}$ are the vectors of known weights for the values corresponding to the groups mentioned in subscripts. The estimates of model parameters β_1 and β_2 are obtained by solving the normal equations $\left(H_{s_1}\right)\hat{\beta}_1 = x^T_{s_1} y_{s_1}$ and $\left(H_{\acute{s}_2}\right)\hat{\beta}_2 = x^T_{\acute{s}_2} y_{\acute{s}_2} = H_{\acute{s}_2}$ respectively, where $H_{s_1} = x^T_{s_1} x_{s_1}$ and $H_{\acute{s}_2} = x^T_{\acute{s}_2} x_{\acute{s}_2}$ are the Hessian matrix for the first round sample and sub-sample respectively. The well-known GMT provides the evidence that the OLS estimators are the best linear unbiased estimators (BLUE) of the parameters β_1 and β_2 when the observations obtained on first round sample s_1 and the second round sample $\acute{s}_2$ follows two different population models with independently and identically distributed error terms. From design-based point of view, the selection of sub-sample $\acute{s}_2$ depends on the selection of s_1, hence the assumption of independence is no more valid. To proceed we need the assumption of independence of model only. The separation of population as the respondents and the non-respondent is based on the values of R which is already discussed in previous sections. The role of the variable R is same as the role of stratification variable in stratified sampling which is merely used to separate populations into respondents and non-respondents. Hence more correlation between the non-response factor (R) and the study variable is a requirement for using the sub-sampling approach. The case of low correlation between the study variable and the non-response variable can be handled through weighting adjustment and imputation techniques discussed in literature. However the literature of sub-sampling technique reveals that the efficiency of the sub-sampling estimator is not affected by this correlation. But in case of presence of significant correlation proceeding with just respondents on first call may produce invalid and inconsistent statistical inference.

Note that respondents on first sample always represent the responding population U_1. While the non-respondents on first sample may or may not represent the population of the non-respondent U_2 as it depends on the degree of relationship between R and Y and the nature of occurrence of non-response (whether it is ignorable or not). The model bias of $\hat{t}_y^*$ is derived in Appendix C, and given by

$$B_M(\hat{t}_y^*) = \gamma_{\bar{s}_1}^T \left[x_{\bar{s}_1}\beta_1 - x_{\bar{s}_1}\beta_1\right] + \gamma_{\acute{\bar{s}}_2}^T \left[x_{\acute{\bar{s}}_2}\beta_2 - x_{\acute{\bar{s}}_2}\beta_2\right] = 0. \tag{4.3.7}$$

$\hat{t}_y^*$ is model unbiased if all of the OLS assumptions are satisfied for the populations of the respondents and non-respondents. Assuming unbiasedness model variance of the total estimator under non-response is obtained as

$$V_M(\hat{t}_y^*) = n_1\sigma_1^2 + \acute{n}_2\sigma_2^2 + \sigma_1^2\gamma_{\bar{s}_1}^T x_{\bar{s}_1}\left(H_{s_1}\right)^{-1} x_{\bar{s}_1}^T W_{\bar{s}_1} + \sigma_2^2\gamma_{\acute{\bar{s}}_2}^T x_{\acute{\bar{s}}_2}\left(H_{\acute{s}_2}\right)^{-1} x_{\acute{\bar{s}}_2}^T W_{\acute{\bar{s}}_2} \tag{4.3.8}$$

Taking expectation with respect to SD we get

$$\begin{aligned} E_D\left\{V_M(\hat{t}_y^*)\right\} = & N_1\sigma_1^2 + N_2\sigma_2^2 + E_{D_1}\Bigg[\sigma_1^2\gamma_{\bar{s}_1}^T x_{\bar{s}_1}\left(H_{s_1}\right)^{-1} x_{\bar{s}_1}^T W_{\bar{s}_1} \\ & + \sigma_2^2 E_{D_2}\left(\gamma_{\acute{\bar{s}}_2}^T x_{\acute{\bar{s}}_2}\left(H_{\acute{s}_2}\right)^{-1} x_{\acute{\bar{s}}_2}^T W_{\acute{\bar{s}}_2}\right)\Bigg], \end{aligned} \tag{4.3.9}$$

where E_{D_1} and E_{D_2} are expectations with respect to SD used for selecting first sample and sub-sample respectively. The first component of the expected model-variance depends on the error variances while the second component depends on the inverse of the matrix for the first sample and the sub-sample. Hence, for smaller variance the population units with larger sampled values of all included covariates should be prefer. Chambers and Clark (2012) provided a detail discussion on optimum selection of units under different population models.

4.4 Estimation of Total Under Ill-Conditioned Regression

While applying linear regression model for predicting the non-sampled values from the population of non-respondents the number of input variables (regressors) may greatly exceeds the number of observations i.e. $\acute{n}_2 < (p+1)$ as we are sub-sampling a relatively small portion of non-respondents. In such situations, fitting the full model to the non-respondents without penalization will result in wider prediction intervals, and the normal equations may not have trivial solution as the matrix $H_{\acute{s}_2}$ does not possess the full rank property. It is not possible to estimate the parameters of the model when $H_{\acute{s}_2}$ is singular i.e. not of full rank. This situation is called super-collinearity or ill-conditioning. The problem of super-collinearity can be solved using ridge regression which is already discussed in Chapters 2 and 3. To get an estimate for β_2, when there is super-collinearity in x_2, we use ad-hoc fix method proposed by Hoerl and Kennard (1970) for resolving singularity of $H_{\acute{s}_2}$. We simply replace $H = H_{\acute{s}_2}$ by $H(v) = H_{\acute{s}_2} + vI_{p+1}$ with $v \in [0, \infty]$. The scalar v is called tuning parameter or penalty parameter. A clearly defined estimator for β_2 obtained even for high-dimensional data matrix $(\acute{n}_2 \leq p)$ for a strictly positive v is $\hat{\beta}_2(v) = H(v)^{-1} x_{\acute{s}_2}^T y_{\acute{s}_2}$. Using $\hat{\beta}_2(v)$ in (4.3.6), we obtain a partially ridge regression (PRR) estimator (as the concept of ridge regression is used for non-responding part only) for population total which is given by

$$\hat{t}_y^* = \gamma_{s_1}^T y_{s_1} + \gamma_{\bar{s}_1}^T x_{\bar{s}_1} \hat{\beta}_1 + \gamma_{\acute{s}_2}^T y_{\acute{s}_2} + \gamma_{\bar{\acute{s}}_2}^T x_{\bar{\acute{s}}_2} \hat{\beta}_2(v). \tag{4.4.1}$$

The expressions for model-bias and expected model-MSE of the PRR estimator of the total in presence of non-response are obtained by replacing $H(v)$ by H in (4.3.8) and (4.3.9). Following Vinod and Ullah (1981) a range for v in which the model-MSE of $x_{\bar{\acute{s}}_2} \hat{\beta}_2(v)$ is smaller than the model-variance of $x_{\bar{\acute{s}}_2} \hat{\beta}_2$ is

$$0 < v < \frac{2}{[-min(0, \psi_2)]}, \tag{4.4.2}$$

where ψ_2 is the minimum eigen-value of the matrix $\left(H_{\acute{s}_2}\right)^{-1} - \frac{\beta_2 \beta_2^T}{\sigma_2^2}$. PRR is also applicable for predicting non-sampled respondents when $n_1 < (p+1)$ leading to super-collinearity in the respondents.

Another major problem that arises in estimation of so called superpopulation parameters is the violation of assumption of homoscedasticity. In presence of heteroscedasticity, we can write

$$V_M\big(Y_1|x_1\big) = \sigma_1^2 V_1 \;\text{ for }\; R = 1 \tag{4.4.3}$$

$$V_M\big(Y_2|x_2\big) = \sigma_2^2 V_2 \;\text{ for }\; R = 0, \tag{4.4.4}$$

where $V_1 = diag(V_{1ii}, i \in U_1)$ and $V_2 = diag(V_{2ii}, i \in U_2)$ units specific variances for respondents and non-respondents respectively. Here $V_{1ii} = V_M(Y_{1i}|x_{1i}) = \upsilon(x_{1i})$ and $V_{2ii} = V_M(Y_{2i}|x_{2i}) = \upsilon(x_{2i})$, where x_{1i} and x_{2i} are the vectors of the auxiliary variables corresponding to the ith unit in respondents and non-respondents respectively. In such situations, OLS estimators for the regression coefficients may have higher variances. If we have information about the variance structure for the populations of respondents and non-respondents (assuming zero correlation between the units), we can adopt weighted least square (WLS) method of estimation. The WLS estimators of β_1 and β_2 are $\hat{\beta}_{1wls} = \big(x_{s_1}^T V_{s_1}^{-1} x_{s_1}\big)^{-1} x_{s_1}^T V_{s_1}^{-1} y_{s_1}$ and $\hat{\beta}_{2wls} = \big(x_{\acute{s}_2}^T V_{\acute{s}_2}^{-1} x_{\acute{s}_2}\big) x_{\acute{s}_2}^T V_{\acute{s}_2}^{-1} y_{\acute{s}_2}$ respectively, where

$$V_1 = \begin{bmatrix} V_{s_1} & 0 \\ 0 & V_{\bar{s}_1} \end{bmatrix} \;\text{ and }\; V_2 = \begin{bmatrix} V_{\acute{s}_2} & 0 \\ 0 & V_{\acute{\bar{s}}_2} \end{bmatrix}.$$

The sub-matrices are also diagonal assuming zero correlation between the error terms corresponding to the respobndents and the non-respondents. A WLS estimator for t_y in presence of non-response is obtained by replacing $\hat{\beta}_{1wls}$ and $\hat{\beta}_{2wls}$ by $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively in (4.3.6). It is assumed that the variance structures of the responding and non-responding population are known and depend on covariates whose values are known for each population unit. In practice, for many types of data set, the structure of weights (inverse of variance) is usually unknown, so one has to perform an ordinary least squares (OLS) regression first to estimate the variance structure and obtain estimates for the population regression coefficients after performing an iterative process which is commonly known as generalized least square (GLS).

4.5 Numerical Study

A real data set taken from Yeh et al. (2009) is applied to investigate the behavior of our proposed model-based estimator. The data consist of $N = 748$ blood donors on following variables:
y=Monetary total blood donated in c.c., x_1=Time (months since first donation),
x_2=Recency (months since last donation) and x_3=Frequency (total number of donation). Considering the above 748 blood donors as our population of interest, we select a sample of size $n = 100$ using SRSWOR. The scatter plot matrix between the variables in the sample selected on first call and the sub-sample collected on the second call represents the relationship between the variables in the population of respondents and non-respondents for response rates $\lambda_2 = 0.4$ (Figures 4.1 and 4.2) and $\lambda_2 = 0.4$ (see Figures 4.3 and 4.4). Figure 1 shows the relation between the study variable y and the predictors x_1, x_2, and x_3 for the sampled respondents which shows that the study variable y, is highly related to x_3 and moderately related to x_1 but weakly related to x_2. Figure 4.2 portrays the relationship between variables for the sub-sampled non-respondents which is different from the relationship in Figure 4.1 which shows the relevancy of the data to our proposed sampling mechanism. One can observe the similar relationship between the variables for $\lambda = 0.2$ from upper triangle of Figures 4.3 and 4.4. Hence our proposal works here as the relationship between the total monetary blood donated and its three determinants have different relationship for the population of the respondents and the sub-population of the non-respondents which is the main assumption of our data collection mechanism. We select half ($k = 2$) of the non-respondent selected on first call for sub-sampling on second call.

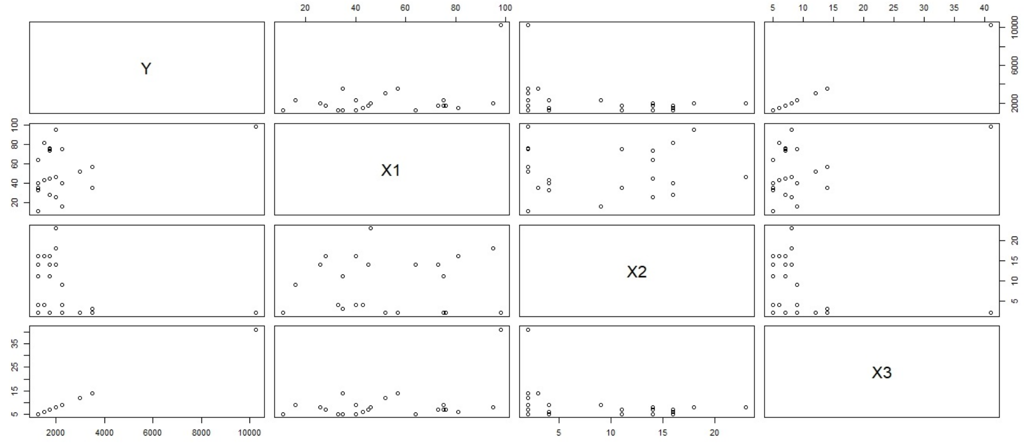

Figure 4.1: Behavior of non-respondents with $\lambda_2 = 0.4$

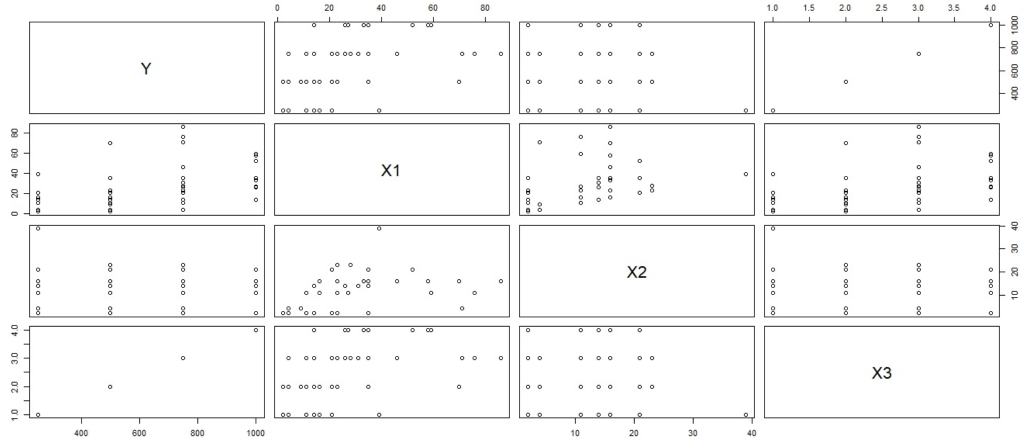

Figure 4.2: Behavior of respondents with $\lambda_2 = 0.4$

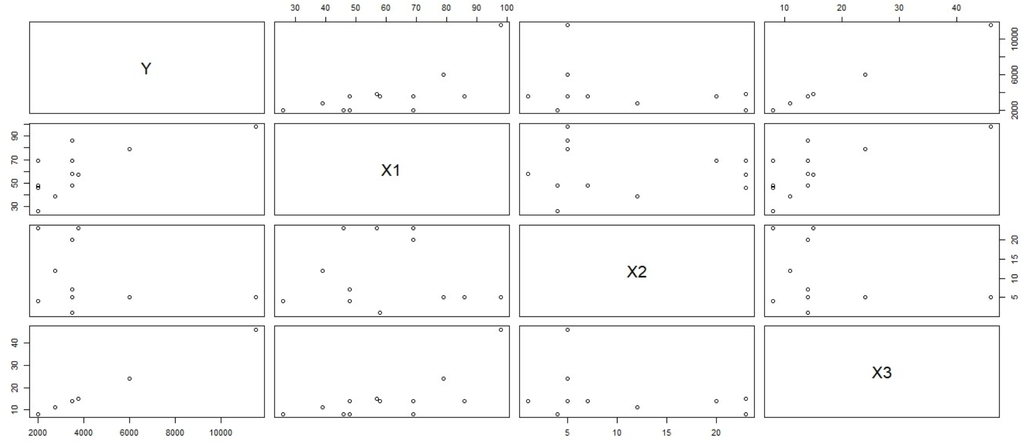

Figure 4.3: Behavior of non-respondents with $\lambda_2 = 0.2$

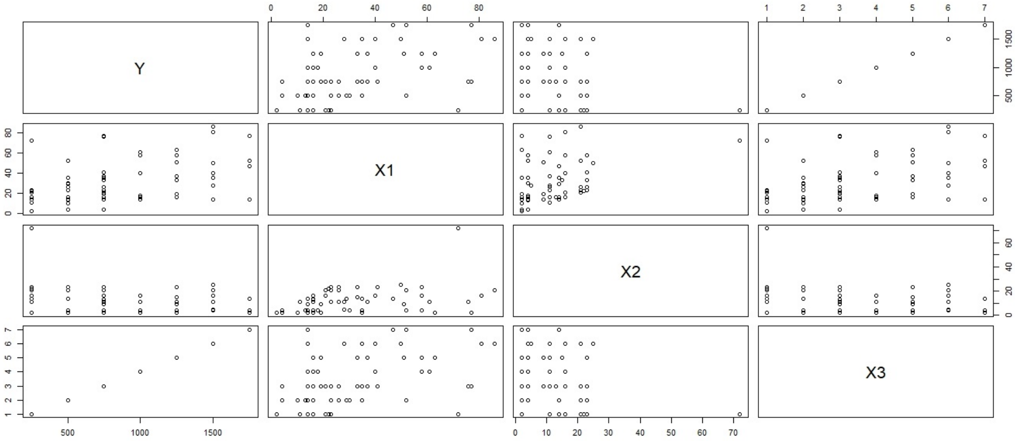

Figure 4.4: Behavior of respondents with $\lambda_2 = 0.2$

Further to see the magnitude of the prediction error, we provide a bootstrap sampling procedure taking different non-response rate (say λ_2) in the population. We generate a new variable R associated with each 748 cases which posses value 1 if the ith unit has an outcome greater than the λ_2th percentile of all the y values in the data set otherwise zero.

(i). A sample of size n (i.e. n=100, 200) is taken from the data using simple random sampling without replacement and divide them into the respondents and the non-respondents according to the value of R and observe n_1 and n_2.

(ii). Select a sub-sample of size $\acute{n}_2 = \frac{n_2}{k}$ (taking k=2,4) from n_2 non-respondents again using SRWOR and compute the estimator using information obtain from first and second samples. We take $p = 2$ to avoid the problem of super-collinearity in our situation.

(iii). Repeat Step ii, 2000 times to get expected value from the sub-sampling. The sub-sampling does not alter results of $\hat{T}_1$ as it is based on sample from respondents only.

(iv). Repeat Steps (i)-(iii), 10,000 times to obtain a stable value of prediction variance and bias for both estimators.

The relative design-based bias and variances are computed as follows:

$$RB(\hat{t}_{y1}) = E_{D1}E_{D2}\left[\frac{\hat{t}_{y1} - t_y}{t_y}\right] \tag{4.5.1}$$

$$RMSE(\hat{t}_{y1}) = E_{D1}E_{D2}\left[\frac{\hat{t}_{y1} - t_y}{t_y}\right]^2 \tag{4.5.2}$$

The RB and RMSE for the Hansen and Hurwitz (1946)-type estimator for the population total are obtained by replacing $\hat{t}_{y1}$ by $\hat{t}_y^*$ in Equations (4.5.1) and (4.5.2) respectively. Table 4.1. provides relative bias (RB) and relative mean squared error (RMSE) of the total estimator based on the sample on first call for different combinations of n, λ_2 and k.

Table 4.1.: Relative bias and MSE of $\hat{t}_{y1}$ and $\hat{t}_y^*$ for different combinations of n, k and λ_2

n	k	$RB(\hat{t}_{y1})$	$RB(\hat{t}_y^*)$	$RMSE(\hat{t}_{y1})$	$RMSE(\hat{t}_y^*)$
			$\lambda_2 = 0.5$		
100	2	-0.56463	0.03325	0.32017	0.01014
	4		0.01626		0.02009
200	2	-0.56646	0.04592	0.32130	0.00575
	4		0.03085		0.00996
			$\lambda_2 = 0.25$		
100	2	-0.36311	0.01615	0.13352	0.00766
	4		-0.00332		0.02661
200	2	-0.36719	0.02492	0.13554	0.00371
	4		0.01449		0.00732
			$\lambda_2 = 0.10$		
100	2	-0.19545	-0.00931	0.04023	0.42009
	4		-0.02266		0.47775
200	2	-0.19806	-0.00607	0.04005	0.03924
	4		-0.00239		0.04046

The results in Table 4.1. are reported assuming non-response rate λ_2 at 50%, 25%, and 10%. RB of both estimators go to zero as non-response rate falls toward zero which assures that for full response it vanishes while the sub-sampling method produce ignorable bias as compared to direct method which is the attractive feature of this method. Further from Table 4.1., one can observe that RMSE is smaller in case of sub-sampling non-respondents, i.e. taking interview of additional non-respondents through some stronger mode of interview, for every choices of λ_2. RB and RMSE of $\hat{t}_y^*$ tend to increase with decrease in non-response rate in the population which shows that our proposed technique works well for higher non-response rates as compared to smaller ones. RB and RMSE of the model-based total estimator go down while increasing sub-sample size $\acute{n}_2$ (decreasing k) as expected. Further, this error decreases when population has smaller non-response rate λ_2. In next section, we provide a simulation study to provide a detailed picture of the performance of estimators in terms of design bias and mean squared error.

4.6 Simulation Study

To see the long run behavior of the proposed estimators in terms of bias and efficiency, a simulation study, generating a hypothetical population, is conducted. Following Najarian et al. (2013), a matrix $z = (z_{ij}, i = 1,2,3,...,N, j = 1,2,...,p)$ with p variate each generated from $N(100,1)$, has been constructed with N=10,000 observations. The ijth element of the auxiliary matrix x is computed as $x_{ij} = (1-\rho)^{0.5} \times z_{ij} + \rho \times z_{ij}$, where ρ is the degree of linear relationship between x and z to be fixed in advance. The vector of the study variable (y) is then obtained by using the relationship $y = x\omega + \varepsilon$, where ω is the vector of coefficients which are computed as the averaged eigen vectors corresponding to the eigen values of $H = x^T x$ that are greater than unity and $\varepsilon \sim N(0, \sigma^2 I_N)$ is randomly generated error term. It is assumed that the variance is of homoscedastic nature with constant diagonal σ^2. We fix σ^2 at 0.01, 0.1 and 1. The data consist of $(y, 1_N, x, R_i)$, where 1_N is the vector of 1's. R_i takes value 1 if the ith value of variable y falls in a threshold lower than $(1-\lambda_2)$th quantile in the population, where λ_2 is non-response rate in the population. In real life, we suggest to choose R in form of some observable covariates or latent variables. The simulation study is conducted in following three steps.

(i). Take a random sample of size n from the population generated through the mechanism described above and split it into n_1 respondents and n_2 non-respondents according to the values of R_i.

(ii). Select a sub-sample of size $\acute{n}_2$ from n_2 non-respondents for fix k.

(iii). Estimate the population total (t_y) using estimated models from samples obtained on Steps 1 and 2.

(iv) Simulate Steps (ii)–(iii), 500 times and average the values of estimates.

(v). Repeat Steps (i)–(iv), 2000 times to obtain prediction errors to obtain 2000 estimated values.

The relative bias and mean squared error of the proposed total estimators are computed using the formula given in Equations (4.5.1) and (4.5.2) respectively after removing the denominators as the generated values are already standardized. The subscript v is used for the results where prediction is performed using PRR.

Tables C.1-C.3 (see Appendix C) provide the bias of the PPR estimator and MSE of both estimators for different combinations of σ^2, λ_2, ρ, n and k in nested order. We obtain results for $p = 5$ and $p = 8$ but the result for $p = 5$ is not reported here for the sake of space. Tables C.1, C.2 and C.3 provide the prediction error measures (B and MSE) for $\sigma^2 = 0.01$, $\sigma^2 = 0.1$ and $\sigma^2 = 1$ respectively. From Tables C.1–C.3 (see Appendix C), one can see that the bias of the PRR total estimator tends to increase with increase in k . This implies selecting a smaller sub-sample increases the bias in estimation due to sampling error although this bias depends on the magnitude of the tuning parameter v. MSE of the total estimator under multiple regression and PRR both increase with increase in k which shows that MSE of the estimators grows with smaller sub-samples from non-respondents. The PRR total estimator is more sensitive to the change in k, in terms of MSE, as the optimum value of the tuning parameter v is estimated from sub–sample. In practice v might be computed using data available from previous surveys of the same population or through expert judgment. The estimation methods of v by minimization of prediction error are available in Najarian et al. (2013). Moreover, whatever model we use for prediction, the MSE values of the total estimators depend on the sample size of respondents and sub-sample of non-respondents. The simulated results are provided for sample size 100, 150 and 250 with sub-sample size inversely proportional to $k = 1.5$, $k = 2$ and $k = 3$. It can be noticed that the MSE values are increasing with increase in k. Comparing two portions of Tables C.1–C.3, we observe that the MSE of proposed estimators fall when non-response rate increases which conflicts the efficiency property of the Hansen and Hurwitz (1946) estimator. The reason is the use of separate models and increasing λ_2 from 0.2 to 0.4 implies (i.e. we are using Model (2) for 40 % of the data) which is the main contribution of our proposal in terms of increased precision. Apart from the design parameters, the data generating process also effects the efficiency of the total estimator which can be seen from three different column-panels (for three different choices of the parameter ρ) assuming that the correlation between the variables X and Z are same for all choices of j of Tables C.1–C.3.

4.7 Conclusion

This chapter is concerned with utilization of model relationship between the outcome variable and one or more covariate(s) for efficient estimation of population total of the outcome variable in surveys with non-ignorable non-response. A model-based version of Hansen and Hurwitz (1946) sub-sampling technique is suggested which assumes that the responding and non-responding population have different models. This assumption may hold for majority of real world situations where the occurrence of non-response is observable like a stratification variable. In public health surveys, the non-response occurrence is based on the gender, ethical affiliation, age and other demographic factors of the respondents. In such situations, respondents and non-respondents may have different models. The method assumes that a stratification variable is available to divide the population into respondents and non-respondents which is difficult to obtain in most of real surveys although a two phase sampling method can provide a better stratification variable to divide the population into respondents and non-respondents. It is shown that under linear population model (linear in parameter as well as in variables), the total estimator with sub-sampling is model-unbiased and has smaller model-variance as compared to predictive estimator based on sampled respondents only. The linearity assumption emphasizes on linear in parameters but not restricted to the linearity in variable. Polynomial regression models are also useful for handling non-response in demographic surveys using age as the predictor. The problem of non-response can be well handled using polynomial regression models which is not included in this dissertation. While sub-sampling non-respondents the number of observations may become smaller than the number of regressors included in the model leading to problem of super-collinearity. To cope with super-collinearity problem, we suggest a version of ridge regression named, called PRR, for predicting the non-sampled non-respondents. WLS and GLS are suggested for obtaining estimates of the regression coefficients for respondents and non-respondents when error terms for at least one model is of heteroscedastic nature. To confirm mathematical expressions a numerical study with blood transfusion data has been carried out with a simulated study. The suggested method is applicable to telephonic or web household surveys where households are first contacted with email or telephone call and then non-respondents are followed via face to face surveys where it seems logical to select a sub-sample of non-respondents through more expensive mode (face to face).

Chapter 5

Model Based Estimation of Parameters Under Ranked Set Sampling

5.1 Outline

As we already discussed in previous chapters that the utilization of superpopulation models for estimation of finite population parameters is an advantageous practice, when it is easy to recognize the relationship between the study variable and one or more auxiliary variables. This chapter is concerned with estimation of finite population total under ranked set sampling without replacement (RSSWOR), a modified version of RSS, by utilizing model relationship, specially gamma population model (GPM) between the study variable and the auxiliary variable. We opt the single predictor case as the ranking mechanism is possible for two correlated variables (outcome and regressor). The RSSWOR sampling procedure, especially, aids in collecting data from a continuous production process. This chapter presents a discussion on Gamma Population Model (GPM) for estimation of finite population total in simple random sampling in Section 5.2. RSSWOR is employed under model-based approach for estimation of total in Section 5.3. Comparison of the proposed estimators are made with existing ones using Monte Carlo (MC) experiment in Section 5.4. Section 5.5 concludes the chapter.

5.2 Model-Based Estimation Under SRSWOR

Remind that Y and X be the study and the auxiliary variables respectively corresponding to the units in population $U = \{U_i; i = 1, 2, ..., N\}$. Further U consists of two mutually exclusive sets s (set of sampled elements) and $\bar{s}$ (set of non-sampled elements) having n and $(N - n)$ elements respectively. We assume following population models for prediction:

- $y_i = \mu + \varepsilon_i$ (Homogeneous population model, HPM)
- $y_i = \beta x_i + \varepsilon_i x_i^{\gamma^*}$ (Gamma population model, GPM),

where y_i, x_i and ε_i are the ith population values corresponding to the study variable Y, auxiliary variable X and the random error term ε respectively. The random error term ε_i is identically independently distributed with zero mean and constant variance. Further μ and β are unknown constant to be estimated using sample data. Here, γ^* is the rate parameter as variation in Y varies with this rate, it may also be unknown but it is chosen in advance using expert judgment or from pilot surveys with cross validation. A lot of works on prediction under linear population model is available in literature of model-based estimation. In this chapter, we first briefly discuss the estimation of population total under HPM and GPM.

5.2.1 Homogeneous Population Model (HPM)

Under HPM, we have the relationship $y_i = \mu + \varepsilon_i$, which assumes that there is no auxiliary variable at design stage or/and estimation stage. We can express population total as:

$$t_y = \sum_{i \in s} y_i + \sum_{i \in \bar{s}} y_i. \tag{5.2.1}$$

The notations $\sum_{i \in s}$ and $\sum_{i \in \bar{s}}$ show that the summation is applied over the sample s and $\bar{s}$ respectively. The expansion estimator $\hat{t}_y^E$ suggested by Chambers and Clark (2012), is given by

$$\begin{aligned}\hat{t}_y^E &= \sum_{i \in s} y_i + E(\hat{t}_{y\bar{s}} | y_i, i \in s) = n\bar{y}_s + (N - n)\bar{y}_s \\ &= N\bar{y}_s,\end{aligned} \tag{5.2.2}$$

where $\hat{t}_{y\bar{s}} = \sum_{i\in\bar{s}} y_i$. The prediction variance of $\hat{t}_y$, is given by

$$V_M(\hat{t}_y^E - t_y) = \sigma^2 \big(N-n\big)\big(\frac{N}{n}\big), \tag{5.2.3}$$

where $\sigma^2 = \frac{1}{N}\sum_{i\in U}(y_i-\mu)^2$. Proof of Equation (5.2.3) can be found in Chambers and Clark (2012).

5.2.2 Gamma Population Model (GPM)

When population under study is heterogeneous then the estimator given in Equation (5.2.2) may not work well. One possible way to overcome this deficiency is stratification but in some occasions it is difficult to stratify the population according to certain stratification variable(s) e.g stratifying units in production process may cause destruction of units. In such situation, a best way to handle the problem of heterogeneity is to search for an auxiliary variable which has some correlation with the study variable. GPM deals with such problems by controlling for variance in the study variate Y, when there is a proportional relationship between the study variable with some auxiliary variable whose values for all population units are available in advance. Another condition that must holds in such model is that the marginal distribution of sampled and non-sampled values of Y for a given value of the auxiliary variable should be same. In other words by conditioning on X, we obtain a non-informative sample Chambers and Clark (2012). Under GPM, we have a relationship $y_i = \beta x_i + \varepsilon_i x_i^{\gamma^*}$ between Y and X. A BLUP for t_y is given by

$$\begin{aligned} \hat{t}_y^g &= \hat{t}_{ys} + E_M\big(\hat{t}_{y\bar{s}}|y_i, i\in s; x_i, i\in U\big) \\ &= \hat{t}_{y_s} + bt_{x\bar{s}}, \end{aligned} \tag{5.2.4}$$

where $b = \sum_{i\in s} c_i y_i$ and $c_i = \frac{x_i^{1-2\gamma^*}}{\sum_{i\in s} x_i^{2-2\gamma^*}}$ for $i = 1,2,\ldots,n$. The conditional expectation of $\hat{t}_y^g$ for given sample information is

$$E_M(\hat{t}_y^g|x_i, i\in s) = \beta x_i = \mu \text{ (say)}. \tag{5.2.5}$$

This reveals that for fixed values of X, $\hat{t}_y$ is unbiased conditioning on values of X with conditional variance

$$\begin{aligned} V_M(\hat{t}_y^g|x_i, i \in U) &= V_M(\hat{t}_{ys}) + t_{x\bar{s}}^2 V_M(\sum_{i\in s} c_i y_i) \\ &= \sigma^2 \sum_{i\in s} \left(1 + \lambda^2 x_i^{2-4\gamma^*}\right), \end{aligned} \tag{5.2.6}$$

where $\lambda = \frac{\hat{t}_{y\bar{s}}}{\sum_{i\in s} x_i^{2-2\gamma^*}}$. The variance goes down when larger values of X are selected in the sample. Comparing Equations (5.2.6) with (5.2.3), we see that $V_M(\hat{t}_y^g|x_i, i \in U) < V_M(\hat{t}_y)$ if

$$n + \lambda \sum_{i\in s} x_i^{1-2\gamma^*} < \left(N - n\right)\left(\frac{N}{n}\right). \tag{5.2.7}$$

The unbiasedness and efficiency properties are computed with respect to model although the total estimator with gamma population under design-based approach is biased.

5.3 Model-Based Estimation Under RSSWOR

To obtain a more accurate data set, McIntyre (1952) proposed ranked set sampling assuming that ranking a small sets of units is economical while taking actual measurement from a large sample is costly. This section provides an application of ranked set sampling scheme to model-based approach after making some modification and discussion on estimation of population total in RSS by assuming HPM and GPM. Consider a finite population U generated from a superpopulation with mean $\mu_{(i)}$ and variance $\sigma_{(i)}^2$ for the ith ordered random variable $y_{(i)}$ for $i \in U$. For any given underlying superpopulation model:

(i) We take sub-populations of size N_j for $j = 1, 2, ...t$ from a superpopulation such that $N = \sum_{j=1}^{t} N_j$, where t is the number of cycles or time frame. It is also assume that every sub-population are large enough to select m^2 units from them i.e. $N_j > m^2$. The concept of so called sub-populations are defined just for taking larger sets to ensure that sampling is without replacement. For a valid statistical inference this division must be at random and independent with the survey variable.

(ii) Select m^2 units from each sub-populations i.e. units produced at same time are on same day can be taken as sub-population in production process.

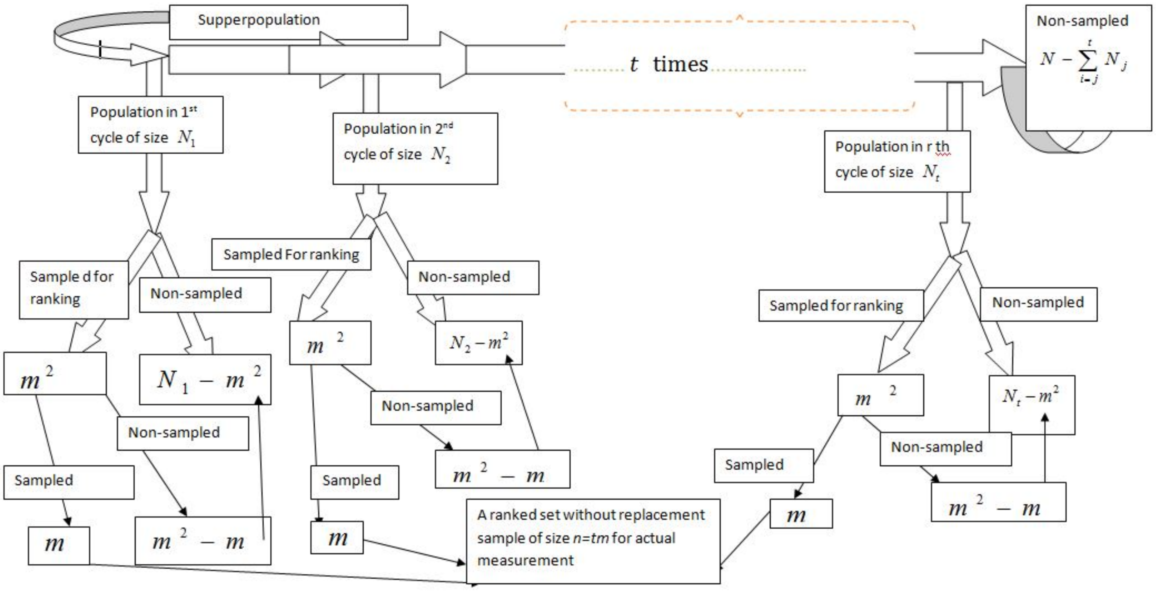

Figure 5.1: Sampling Mechanism of RSSWOR

(iii) Divide each m^2 units in m sets each of size m and rank each set within itself according to some ranking mechanism.

(iv) Select the ith ranked unit from the ith set for $i = 1,2,3,...,m$, and $j = 1,2,..,t$. In this way a ranked set sample without replacement of size tm is obtained. An illustration of RSSWOR scheme is provided in Figure 5.1.

Figure 5.1 explains our sampling scheme assuming that a finite population of size N is coming from a large superpopulation with specified mean and variance. Top stream of Figure 5.1 shows the continuous population. From the finite population of size N units, we consider t different cycles with sizes N_1, N_2, ...,N_t randomly, leaving $N-\sum_{j=1}^{t} N_j$ as non-sampled. For example, in a production process (for quality control) one might considered units produced in 20 days as a finite population, then, we take $t = 8$ randomly selected days as cycles. In this way, we are left with t so called sub-populations. From each sub-populations, we then select m^2 units for ranking leaving $(N_j - m^2)$ units from each sub-population as non-sampled. Finally, applying ranked set sampling for selecting m units from each cycle by returning remaining $m^2 - m$ non-sampled units. The total non-sampled units are found from three different stages which can be seen from Figure 5.1.

$$\begin{aligned}\text{Non-sampled} &= \text{Non-sampled at Stage-1} + \text{Non-sampled at Stage-2}\\ &\quad + \text{Non-sampled at Stage-3}\end{aligned}$$

$$= N - \sum_{j=1}^{t} N_j + \sum_{j=1}^{t} (N_j - m^2)$$

$$+ \sum_{j=1}^{t} (m^2 - m) = N - tm.$$

Let s be the set of tm units selected using the above mechanism and $\bar{s}$ be the set of units which are not in s. A ranked set sample s can be defined as $s = \left\{y_{1(1)1}, \ldots, y_{m(m)1}, \ldots y_{1(1)2}, \ldots, y_{m(m)2}, \ldots\ldots y_{1(1)t}, \ldots y_{m(m)t}\right\}$. When a finite population is partitioned into t mutually exclusive random sets, we have $N = \sum_{j=1}^{t} N_j$. Further when ranking is done with respect to some covariate, we use square brackets in the subscripts of ordered values to denote judgment ranked units. We assume judgment ranking in our proposal.

5.3.1 RSSWOR under HPM

For the ith population value of the study variable Y, we have $y_{(i)} = \mu_{[i]} + \varepsilon_{[i]}$ for $i \in U$, where $\varepsilon_{[i]}$ for all $i \in U$ are i.i.d with zero mean and variance $\sigma^2_{[i]}$. Hence $E_M(y_{[i]}) = 0$, $V_M(y_{[i]}) = \sigma^2_{[i]}$ and $Cov(y_{[i]}, y_{[j]}) = 0$ for $i \neq j$, when $y_{[i]}$ and $y_{[j]}$ are taken from different ranked sets. The condition of zero mean for error term is true only as ranking is performed on some variable other than the study variable. Hence the ranking is supposed as judgment ranking rather than perfect ranking. Consider a predictor for population total given in (5.2.1)

$$\hat{t}^E_{y[rss]} = \hat{t}_{y[rss]s} + \hat{t}_{y\bar{s}}, \tag{5.3.1}$$

where $\hat{t}_{y[rss]s} = \sum_{j=1}^{t} \sum_{i \in s} y_{i[i]j}$ and $\hat{t}_{y\bar{s}} = \sum_{i \in U} y_i - \sum_{j=1}^{t} \sum_{i \in s} y_{i[i]j}$. The problem is to predict $\hat{t}_{y\bar{s}}$ using information at hand such that (i) $E_M(\hat{t}^E_{y[rss]} - t_y) = 0$, the prediction error, and (ii) $E_M(\hat{t}^E_{y[rss]} - t_y)^2$, the squared prediction error, is minimum. $\hat{t}^E_{y[rss]}$ can be expressed as a linear combination of ranked data as:

$$\hat{t}^E_{y[rss]} = \sum_{i \in s} w_{[i]} y_{i[i]}. \tag{5.3.2}$$

For easy of computation, we take $t = 1$, i.e. only one cycle is performed.

$$\hat{t}^E_{y[rss]} - t_y = \sum_{i \in s} w_{[i]} y_{i[i]} + \sum_{i \in s} y_{i[i]} - \sum_{i \in s} y_{i[i]} - t_y$$

$$= \sum_{i \in s} (w_{[i]} - 1) y_{i[i]} - \hat{t}_{y\bar{s}}, \tag{5.3.3}$$

where $(w_{[i]} - 1) = u_{[i]}$ (say) is the prediction weight for the ith non-sampled unit. Taking expectation of Equation (5.3.3), we have

$$E_M(\hat{t}^E_{y[rss]} - t_y) = \sum_{i \in s} u_{[i]} \mu_{[i]} - (N-m)\mu, \tag{5.3.4}$$

$\hat{t}^E_{y[rss]}$ will be unbiased when $\sum_{i \in s} u_{[i]} \mu_{[i]} = (N-m)\mu$. Similarly, variance of $\hat{t}^E_{y[rss]} - t_y$ can be found as

$$\begin{aligned} V_M(\hat{t}^E_{y[rss]} - t_y) &= V_M(\sum_{i \in s} u_{[i]} y_{i[i]} - \hat{t}_{y\bar{s}}) \\ &= V_M(\sum_{i \in s} u_{[i]} y_{i[i]}) + V_M(\hat{t}_{y\bar{s}}) \\ V_M(\hat{t}^E_{y[rss]} - t_y) &= \sum_{i \in s} u^2_{[i]} \sigma^2_{[i]} + (N-m)\sigma^2. \end{aligned} \tag{5.3.5}$$

As the sampled and non-sampled values are independent so the covariance term on the right side of Equation (5.3.5) is zero. The value of u_i that provide unbiased estimate of $\hat{t}^E_{y[rss]}$ is $u_{[i]} = \frac{N-m}{m}$. Moreover, the second term in variance expression is $(N-m)\sigma^2$ as there is no-ranking on non-sampled data. Inserting the value of $u_{[i]}$ in variance expression, we get

$$\begin{aligned} V_M(\hat{t}^E_{y[rss]} - t_y) &= \sum_{i \in s} (\frac{N-m}{m})^2 \sigma^2_{[i]} + (N-m)\sigma^2 \\ &= \frac{N}{m}(N-m)\sigma^2 - (\frac{N-m}{m})^2 \sum_{i \in s} \delta^2_{[i]} \\ &= V_M(\hat{t}_y - t_y) - (\frac{N-m}{m})^2 \sum_{i \in s} \delta^2_{[i]}, \end{aligned} \tag{5.3.6}$$

where $\delta_{[i]} = (\mu_{[i]} - \mu)$ and $(m\sigma^2 - \sum_{i \in s} \delta^2_{[i]}) = \sum_{i \in s} \sigma^2_{[i]}$. From (5.2.3) and (5.3.6), it is clear that $\hat{t}^E_{y[rss]}$ is always more efficient than $\hat{t}_y$.

5.3.2 RSSWOR under GPM

Under GPM, the ith population value of the study variable Y is expressed as $y_{[i]} = x_{[i]}\beta + x^{\gamma^*}_{[i]}\varepsilon_{[i]}$ for $i \in U$, where $E_M(y_{[i]}) = x_{[i]}\beta$, $V_M(y_{[i]}) = x^{2\gamma^*}_{[i]} \sigma^2_{[i]}$ and $Cov(y_{[i]}, y_{[j]}) = 0$ for $i \neq j$, when $y_{[i]}$ and $y_{[j]}$ are taken from different ranked sets. It is also assumed that ranking is performed on the study variable itself (based on personal judgment or some other mechanism). A best predictor for $\hat{t}_{y\bar{s}}$ is $E_M\left(\hat{t}_{y\bar{s}}|y_{[i]}, i \in s; x_{[i]}, i \in U\right)$ (see Chambers and Clark, 2012) for detail.

$$\begin{aligned}\hat{t}^{g}_{y[rss]} =& \hat{t}_{y[rss]s} + E_M\left(\hat{t}_{y\bar{s}}|y_{[i]}, i \in s; x_{[i]}, i \in U\right)\\ \hat{t}^{g}_{y[rss]} =& \hat{t}_{y[rss]s} + \sum_{i\in\bar{s}} x_{[i]}\beta. \end{aligned} \tag{5.3.7}$$

In (5.3.7), β is assumed to be unknown. The value $x_{[i]}$ for $i \in \bar{s}$ denotes the ranking of non-sampled values of the auxiliary variable x. However such ranking is difficult in practical situations. One can proceed by replacing x_i by $x_{[i]}$ for $i \in s$. A best linear unbiased predictor (BLUP) $\hat{\beta}$ for β is obtained by minimizing following sum of squared error for sample data with respect to β.

$$\sum_{i\in s} e^2_{i[i]} = \sum_{i\in s} x^{-2\gamma^*}_{i[i]}\left(y_{i[i]} - x_{i[i]}\beta\right). \tag{5.3.8}$$

which is given by $\hat{\beta}_{[rss]} = \sum_{i\in s} q_{[i]} y_{i[i]}$, where $q_{[i]} = \frac{x^{1-2\gamma^*}_{i[i]}}{\sum_{i\in s} x^{2-2\gamma^*}_{i[i]}}$ and the resulting estimator is

$$\hat{t}^{g}_{y[rss]} = \hat{t}_{y[rss]s} + \sum_{i\in\bar{s}} x_{[i]}\hat{\beta}_{[rss]}.$$

Inserting the value of $\hat{\beta}_{[rss]}$ and after some simplification, we get

$$\hat{t}^{g}_{y[rss]} = \sum_{i\in s}\left(1 + \vartheta x^{1-2\gamma^*}_{i[i]}\right) y_{[i]}, \tag{5.3.9}$$

where $\vartheta = \frac{t_{x\bar{s}}}{\sum_{i\in s} x^{2-2\gamma^*}_{i[i]}}$. It is easy to show that $\hat{t}^{g}_{y[rss]}$ is unbiased and its variance, is given by

$$\begin{aligned} V_M(\hat{t}^{g}_{y[rss]}|x_{[i]}, i \in U) = Var\left(\sum_{i\in s}\vartheta^{*}_{[i]} y_{[i]}\right) &= \sum_{i\in s}\vartheta^{*2}_{[i]}\sigma^2_{[i]}\\ &= \sigma^2\sum_{i\in s}\vartheta^{*2}_{[i]} - \sum_{i\in s}\vartheta^{*2}_{[i]}\delta^2_{[i]}, \end{aligned} \tag{5.3.10}$$

where $\vartheta^{*}_{[i]} = 1 + \vartheta x^{1-2\gamma^*}_{i[i]}$. We can also express (5.3.10) as

$$V_M(\hat{t}^{g}_{y[rss]}|x_{[i]}, i \in U) = V_M(\hat{t}^{g}_{y}|x_{[i]}, i \in U) - \sum_{i\in s}\vartheta^{*2}_{[i]}\delta^2_{[i]},$$

where $\delta_{[i]} = \mu_{[i]} - \mu$. This provides that $\hat{t}^{g}_{y[rss]}$ is more efficient than its counterpart in simple random sampling.

GPM considered in this section is a general population model for the situations where the values of the study variable generated from a stochastic process is proportional to the corresponding values of the auxiliary variable. Further the variation in Y depends on the value of X^{γ^*}, where γ^* is the rate parameter which controls how much the variation in Y depends on X. Chambers and Clark (2012) suggested to choose the value of gamma between 0 and 1. Ratio population model is a special case of the GPM for $\gamma^* = \frac{1}{2}$. We can derive BLUP in RSSWOR for ratio population model by inserting $\gamma^* = \frac{1}{2}$. In practical situations, the value of γ^* can be guessed by observing on scatter plot or through the value of correlation coefficient between X and Y. Similarly, setting $\gamma^* = 0$ and adding intercept term the GPM reduces to LPM. In subsequent section, we use real life data set for checking efficiency of the proposed estimators for population total.

5.4 Monte Carlo (MC) Study

For the purpose of efficiency comparison, we use MC experiment by generating hypothetical data on variable X and obtaining Y using the relationship $Y = \rho^2 X + X^{\gamma^*} e$ for $\gamma^* = 0.3, 0.5, 0.8$, where e is an i.i.d error term, normally distributed with zero mean and variance σ^2 with $\rho = 0.7$. The data on X is generated from gamma distribution assuming different combinations of parameters a and b. Figure 5.2 provides different shapes of gamma distribution for the given combinations of parameters. A ranked set sampling without replacement procedure is obtained by using steps given in Section 3. The estimators for sample total under ranked set sampling with replacement for HPM and GPM models are obtained. For efficiency comparison, we also obtain a SRSWOR of size $n = tm$. Repeat the sampling process 10,000 times to obtain bias and relative of the total estimators. The absolute biases of the total estimators are obtained under designed-based point of view as the unbiasedness is conditioned on X. The relative efficiency of of the suggested estimators, are given by

$$RE_r = \frac{V_M(\hat{t}_y)}{MSE\left(\hat{t}^{g}_{y}\right)}, \quad RE_{rss} = \frac{Var(\hat{t}_y)}{MSE\left(\hat{t}^{E}_{y[rss]}\right)}$$

and

$$RE_{R.rss} = \frac{Var\left(\hat{t}^{E}_{y[rss]}\right)}{MSE\left(\hat{t}^{g}_{y[rss]}\right)}.$$

Tables D.1–D.5 (see Appendix D) provide the relative efficiency (RE) and absolute bias (AB) of the proposed estimators. Different sections of Tables D.1–D.5 are constructed for gamma distribution $G(a,b)$ for different combinations of a and b.

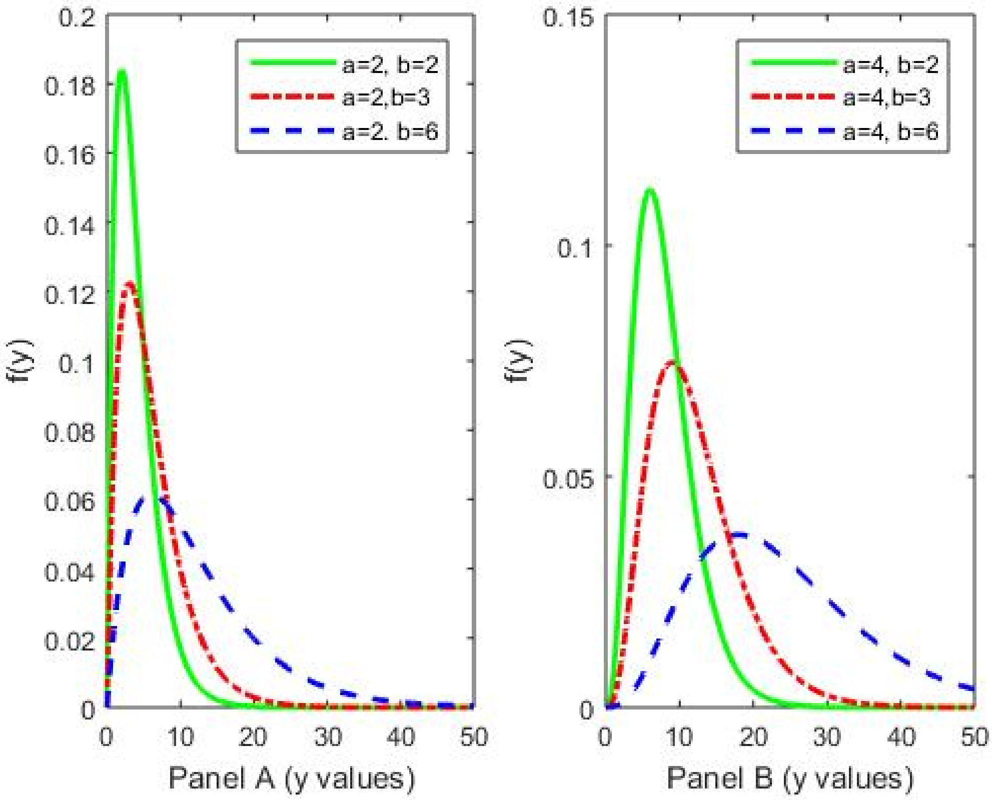

Figure 5.2: Effect of distribution parameters on efficiency

We can interpret the result in following ways

(i). It is obvious that the relative efficiency of $\hat{t}^{g}_{y[rss]}$ and $\hat{t}^{g}_{y}$ both are high when $\gamma^{*} = 1/2$ as compared to the RE for other choices of gamma. It suggest to use the proposed estimator in case of proportional relationship between the two variables with $\gamma^{*} = 1/2$.

(ii). The relative efficiency of the estimator depends on the shape of population from which X is generated. If the ratio a/b increases then the performance of the proportional model increases as compared to homogeneous population model.

(iii). It can be seen from different sections of Tables D.1–D.5 i.e. $G(2,6)$ has lowest efficiency and $G(4,2)$ has highest efficiency relative to their competitors with other combinations. In other words, it can be inferred that relative performance of GPM model is high for skewed populations as compared to homogeneous population models.

(iv) In case of fat tail distribution, the predictors under GPM even perform worst than its counterparts under HPM for both SRSWOR and RSSWOR.

- It can also be noticed that relative efficiencies (REs) i.e. RE_R , RE_{rss} and RE_{rss} are all increasing functions of the set size (m) and the number of cycles (t).

(v). The last two columns of Tables D.1–D.5 (see Appendix D) provide absolute biases of the total estimators under gamma population in SRSWOR and RSSWOR. Absolute bias of the total estimator decreases with increase in set sizes m and number of cycle t under RSSWOR scheme.

(vi). Absolute biases are relatively smaller in case of $\gamma^* = 1/2$ under ratio population model.

5.5 Concluding Remarks

A new version of ranked set sampling for obtaining a without replacement sample under gamma population model (general form of proportional population model) has been introduced. Figure 5.1 presented a picture of the RSSWOR which assumes that the finite population coming from an infinite superpopulation via some stochastic process with finite mean and variance. It is also assumed that the population can be generated from different points i.e. cycles and the m sets taken from one cycle is totally different from the m in other cycles for insuring without replacement. After selecting a sample using RSSWOR, the model relationship between the study variable and the auxiliary variable used to predict the non-sampled values while obtaining a point predictor for the population total. The mathematical expressions and Monte-Carlo experiment both support the superiority of the predictor under RSSWOR over the total predictor under SRSWOR for GPM as well as HPM. Hence, the suggested predictors may perform well for process controls for constructing control charts as in such situations, where we have high dimensional data in the sense of number of observations. It is also applicable in social surveys which we conduct on social media, where one deals with a large population with unending size.

Chapter 6

Model Based Estimation of Parameters in Domains

6.1 Outline

In studies of subpopulations (domains) under post-stratified sampling, it is not easy to obtain an acceptable precision for domain specific estimates due to insufficient sample sizes within certain domains. Indirect estimation using available auxiliary data from whole population under model-based approach is a common practice known as small area estimation and has been practiced for last few decades. To get more reliable results in domains with smaller sample sizes, in this chapter, we use ranked set sampling without replacement (RSSWOR) assuming that ranking smaller sets is easy and cheap. On the basis of RSSWOR, we can either increase precision for fixed sample size or get same level of precision as of the simple random sampling for smaller sample size. We establish domain specific direct estimators under homogeneous and gamma population models. A detailed Monte Carlo (MC) experiment is carried out to observe the design-based efficiency of the estimators. In this chapter, we incorporate the auxiliary information from whole population to obtain efficient estimates for small area total following Chambers and Clark (2012). We then extend the proposed estimators to ranked set sampling frame work for further improvement in efficiency following Ahmed and Shabbir (2019b). In Section 6.2, we provide statistical basis for our proposal from the existing literature and cover estimation of domain total under simple random sampling without replacement (SRSWOR). Extensions of the proposed estimators to RSSWOR are given in Section 6.3. Sections 6.4 and 6.5 are devoted with efficiency comparison and concluding remarks respectively.

6.2 Domain Estimation Under Simple Random Sampling

Let $U = \{1, 2, 3, ...N\}$ be the set of serial numbers attached to the units in a finite population of size N. Further Y and X be the study variable and the auxiliary variable with values y_i and x_i corresponding to the ith population unit for all $i \in U$. The population consists of mutually exhaustive domains whose membership are assumed to be unknown prior to the survey. Such as in demographic studies information regarding age groups, ethical affiliation or women having specific behavior toward having child might be domains of concern. The domain membership variable for the kth domain can be defined as d_{ki} which possess values 1 if ith unit belongs to kth domain, 0 otherwise. The domain size is defined as $N_k = \sum_{i \in U} d_{ki}$, where $\sum_{i \in U}$ denotes summation over population U. In typical situations, domain membership is observable for sample only. However, the distribution of units among domain for non-sampled part of the population is unknown. Even, we don't have information about N_k (the number of units falling in domain k). We consider the homogeneous population model (HPM) and gamma population model (GPM) for the study variable in the kth domain. It is also assumed that the domain membership variable follows Bernoulli distribution with parameter $\theta_k = N_k/N$.

6.2.1 Domain Estimation Under HPM

The population values of y are assumed to be independently distributed in domain k. The homogeneous model for domain k can be written as

$$y_i = \mu_k + \varepsilon_{ki} \quad \text{for } \; i = 1, 2, ..., N \; \text{ and } \; k = 1, 2, ...h, \tag{6.2.1}$$

where μ_k is domain specific mean for domain k, ε_{ki} random error term independently distributed with zero mean and variance σ_k^2 for kth domain and h is the number of domains. The distribution of y_i is conditional on d_{ki}. The domain membership variables d_{ki}'s are already defined as independently distributed Bernoulli random variables. Now the mean and variance of $d_{ki}y_i$ can be obtained as

$$E_M(d_{ki}y_i) = E_M(y_i|d_{ki}=1)p(d_{ki}=1) = \mu_k\theta_k \tag{6.2.2}$$

and

$$V_M(d_{ki}y_i) = \theta_k\sigma_k^2 + \theta_k(1-\theta_k)\mu_k^2 = \sigma_k^{*2}(\textbf{say}). \tag{6.2.3}$$

The covariance between $d_{ki}y_i$ and $d_{ki}y_i$ for $i \neq j \; \forall \;\; i,j \in U$ is zero as y_i (conditionally) and d_{ki} both are independent random variables. Let s be a simple random sample taken with out replacement from the population U with size n. Following Chambers and Clark (2012) the expansion estimator for t_{yk}, is given by

$$\hat{t}_{yk}^E = \frac{N}{n}\sum_{i\in s} d_{ki}y_i = N\hat{\theta}_k\bar{y}_k, \tag{6.2.4}$$

where $\bar{y}_k = \frac{\sum_{i\in s} d_{ki}y_i}{n}$ is the sample mean for the kth domain and $\hat{\theta}_k = \frac{n_k}{n}$ is estimate of θ_k. The derivation of $\hat{t}_{yk}^E$, is given in next subsection.

6.2.2 Derivation of Model Expectation and Variance of the Product *DY*

$$\begin{aligned}
E_M(d_{ki}y_i) &= \sum_y\sum_{d_k} d_{ki}y_i p(y_i,d_{ki}) = \sum_y\sum_{d_k} d_{ki}y_i p(y_i|d_{ki}=1)p(d_{ki}=1)\\
&= p(d_{ki}=1)\sum_y y_i p(y_i|d_{ki}=1)\\
&= p(d_{ki}=1)E_M(y_i|d_{ki}=1)\\
&= \theta_k\mu_k.
\end{aligned}$$

Similarly for variance, we need

$$\begin{aligned}
E_M(d_{ki}^2y_i^2) &= \sum_y\sum_{d_k} d_{ki}^2y_i^2 p(y_i,d_{ki}) = \sum_y\sum_{d_k} d_{ki}^2 p(y_i|d_{ki}=1)p(d_{ki}=1)\\
&= p(d_{ki}=1)\sum_y y_i p(y_i|d_{ki}=1)\\
&= p(d_{ki}=1)E_M(y_i^2|d_{ki}=1)
\end{aligned}$$

$$= \theta_k \left[\mu_k^2 + \sigma_k^2\right].$$

Substituting $E_M(d_{ki}y_i^2)$ in variance formula, we get

$$\begin{aligned} V_M(d_{ki}y_i) &= E_M(d_{ki}^2 y_i^2) - \left\{E_M(d_{ki}y_i)\right\}^2 \\ &= \theta_k\sigma_k^2 + \theta_k(1-\theta_k)\mu_k^2. \end{aligned} \tag{6.2.5}$$

Theorem 1:

The expansion estimator $\hat{t}_{yk}^E$ is unbiased, in model-based sense, for domain total t_{yk} with prediction error variance

$$V_M(\hat{t}_{yk}^E - t_{yk}) = \frac{N(N-n)}{n}\left(\theta_k\sigma_k^2 + \theta_k(1-\theta_k)\mu_k^2\right). \tag{6.2.6}$$

Proof of bias and variance of $\hat{t}_{yk}^E$

To obtain a best linear unbiased estimator for population total say t_{yk}, we write the expansion estimator as the linear combination $d_{ki}y_i$ i.e. $\hat{t}_{yk}^E = \sum_{i\in s} w_i d_{ki} y_i$, where w_i are the weights assigned to ith sampled unit. Further

$$\begin{aligned} \hat{t}_{yk}^E &= \sum_{i\in s} d_{ki}y_i + \sum_{i\in s}\left(w_i - 1\right)d_{ki}y_i \\ &= t_{yks} + \sum_{i\in s} w_i^* d_{ki}y_i, \end{aligned}$$

where $w_i^* = (w_i - 1)$ is the prediction weight corresponding to ith sampled unit for predicting non-sampled part of the population and $t_{yks} = \sum_{i\in s} d_{ki}y_i$. The corresponding domain specific population total $t_{yk} = t_{yks} + t_{ykr}$, where t_{yks} and t_{ykr} are the domain total for sampled and non-sampled units respectively.

The prediction error for $\hat{t}_{yk}^E$ is obtained as: $\hat{t}_{yk}^E - t_{yk} = \sum_{i\in s} w_i^* d_{ki}y_i - t_{ykr}$.

The model bias of the $\hat{t}_{yk}^E$ is

$$\begin{aligned} E_M(\hat{t}_{yk}^E - t_{yk}) &= \sum_{i\in s} w_i^* E_M(d_{ki}y_i) - \sum_{i\in \bar{s}} E_M(d_{ki}y_i) \\ Bias(\hat{t}_{yk}^E) &= \theta_k\mu_k\left[\sum_{i\in s} w_i^* - (N-n)\right]. \end{aligned}$$

The bias vanishes only if

$$\sum_{i\in s} w_i^* = (N-n). \tag{6.2.7}$$

The model variance of prediction error can be obtained as:

$$V_M(\hat{t}_{yk}^E - t_{yk}) = \sum_{i\in s} w_i^{*2} V_M(d_{ki}y_i) + \sum_{i\in \bar{s}} V_M(d_{ki}y_i) - 2Cov(\sum_{i\in s} w_i^*(d_{ki}y_i), \sum_{i\in \bar{s}}(d_{ki}y_i))$$
$$V_M(\hat{t}_{yk}^E) = \sigma_k^{*2}\big[\sum_{i\in s} w_i^{*2} + (N-n)\big].$$

The covariance term in variance expression vanishes as the covariance between $d_{ki}y_i$ and $d_{ki}y_i$ for all $i \neq j$ is zero. The problem is to minimize $V_M(\hat{t}_{yk}^E - t_{yk})$ with respect to w_i^* subject to constrain given in (6.2.7). The Lagrangian function L can be written as

$$L = \sum_{i\in s} w_i^{*2} - 2\lambda\big[\sum_{i\in s} w_i^* - (N-n)\big] \tag{6.2.8}$$

By minimizing (6.2.8), we get $w_i^* = \frac{N-n}{n}$ and $w_i = \frac{N}{n}$. The expansion estimator $\hat{t}_{yk}^E = \frac{N}{n}\sum_{i\in s} d_{ki}y_i$ with variance

$$V_M(\hat{t}_{yk}^E - t_{yk}) = \sigma_k^{*2}\big[\frac{(N-n)^2}{n} + (N-n)\big]$$
$$= \frac{N(N-n)}{n}\sigma_k^{*2}.$$

The expansion estimator $\hat{t}_{yk}^E$ has two attractive features one is BLUP property with respect to model and the other is the compensation for unknown domain size.

6.2.3 Domain Estimation Under GPM

The HPM only uses sample information about the study variable itself for predicting non-sampled values of the study variable. While GPM, on the other hand, assumes that both the mean and variance of the study variable depend on some covariate who's values are known for every unit in the population. Typically, in design-based estimation, we assume that the information about some parameter of that covariate is available. The gamma population model for domain k is defined as

$$y_i = \beta_k x_i + x_i^{\gamma^*}\varepsilon_{ki} \quad \text{for } \ i = 1,2,...,N \quad \text{and} \ \ k = 1,2,...h, \tag{6.2.9}$$

where γ^* is a real constant assumed to be known or guessed in advance and β_k is the domain specific coefficient corresponding to the covariate x. The generalized prediction estimator given in Valliant (2000) for gamma population model can be written as

$$\hat{t}_{yk}^{g} = \sum_{i\in s} d_{ki}y_i + \sum_{\bar{s}} \left(\hat{\beta}_k x_i\right), \tag{6.2.10}$$

where $\hat{\beta}_k$ is the BLUE of β_k and derived in next subsection.

6.2.4 Derivation of $\hat{\beta}_k$

For assuring linearity, we can write $\hat{\beta}_k$ as a linear combination of the domain specific responses $d_{ki}y_i$ as $\hat{\beta}_k = \sum_{i\in s} c_i d_{ki} y_i$, where c_i are the weights associated to ith sampled unit which is obtained as the function of known auxiliary data. The expectation and variance of $\hat{\beta}_k$ can be obtained as

$$E_M(\hat{\beta}_k) = \sum_{i\in s} c_i E_M(d_{ki}y_i) = \beta_k \sum_{i\in s} c_i x_i \tag{6.2.11}$$

For unbiasedness, we must have

$$\sum_{i\in s} c_i x_i = 1 \tag{6.2.12}$$

and

$$V_M(\hat{\beta}_k) = \sum_{i\in s} c_i^2 V_M(d_{ki}y_i) = \sigma_k^{*2} \sum_{i\in s} c_i^2 x_i^{2\gamma^*} \tag{6.2.13}$$

The problem is, then, to minimize $V_M(\hat{\beta}_k)$, specially $\sum_{i\in s} c_i^2$, subject to the constrain given in (6.2.12). The Lagrangian function is again defined as

$$L = \sum_{i\in s} c_i^2 x_i^{2\gamma^*} - 2\lambda\left[\sum_{i\in s} c_i x_i - 1\right] \tag{6.2.14}$$

Differentiating (6.2.14) with respect to c_i and equating to zero results $c_i = \lambda x_i^{1-2\gamma^*}$. Further differentiating (6.2.14) with respect to λ and equating to zero, we get $\sum_{i\in s} c_i x_i = 1$. Solving the two equation results $\lambda = \frac{1}{\sum_{i\in s} x_i^{2-2\gamma^*}}$ and $c_i = \frac{x_i^{1-2\gamma^*}}{\sum_{i\in s} x_i^{2-2\gamma^*}}$. Substituting the value of c_i in $\hat{\beta}_k$, we get

$$\hat{\beta}_k = \frac{\sum_{i\in s} d_{ki} x_i^{1-2\gamma^*} y_i}{\sum_{i\in s} x_i^{2-2\gamma^*}}. \tag{6.2.15}$$

The BLUP for domain total is then obtained as

$$\hat{t}^g_{yk} = \sum_{i\in s} d_{ki} y_i + \frac{\sum_{i\in s} d_{ki} x_i^{1-2\gamma^*} y_i}{\sum_{i\in s} x_i^{2-2\gamma^*}} \sum_{\bar{s}} x_i = \sum_{i\in s} (1+u_i^*) d_{ki} y_i, \tag{6.2.16}$$

where $u_i^* = \frac{x_i^{1-2\gamma^*}}{\sum_{i\in s} x_i^{2-2\gamma^*}} \sum_{\bar{s}} x_i$. The prediction error for $\hat{t}^g_{yk}$, is given by

$$\hat{t}^g_{yk} - t_{yk} = \sum_{i\in s} u_i^* d_{ki} y_i - t_{ykr}, \tag{6.2.17}$$

It is straight forward to show that the estimator given in (6.2.16) is unbiased with prediction error variance given by

$$V_M(\hat{t}^g_{yk} - t_{yk}) = \theta_k(1-\theta_k)\beta_k^2 \kappa_1(x) + \theta_k \sigma_k^2 \kappa_2(x) \tag{6.2.18}$$

where $\kappa_1(x) = \sum_{i\in s} u_i^{*2} x_i^2 + \sum_{i\in \bar{s}} x_i^2$ and $\kappa_2(x) = \sum_{i\in s} u_i^{*2} x_i^{2\gamma^*} + \sum_{i\in \bar{s}} x_i^{2\gamma^*}$.

6.2.5 Derivation of Prediction Error Variance of $\hat{t}^g_{yk}$

The model variance of prediction error of $\hat{t}^g_{yk}$ can be obtained as

$$V_M(\hat{t}^g_{yk} - t_{yk}) = \sum_{i\in s} u_i^{*2} V_M(d_{ki} y_i) + \sum_{i\in \bar{s}} V_M(d_{ki} y_i) - 2Cov\Big(\sum_{i\in s} u_i^*(d_{ki} y_i), \sum_{i\in \bar{s}} (d_{ki} y_i)\Big) \tag{6.2.19}$$

The covariance term in variance expression vanishes as the the covariance between $d_{ki} y_i$ and $d_{ki} y_i$ for all $i \in s$ and $j \in \bar{s}$ is zero. We know under gamma population model $E_M(d_{ki} y_i) = \theta_k \beta_k x_i$. Further, we have

$$V_M(d_{ki} y_i) = E_M(d_{ki} y_i)^2 - \{E_M(d_{ki} y_i)\}^2$$

$$
\begin{aligned}
&= p(d_{ki}=1)E_M\left(y_i^2|d_{ki}=1\right) - \left\{p(d_{ki}=1)E_M\left(y_i|d_{ki}=1\right)\right\}^2 \\
&= \theta_k\left(\beta_k^2 x_i^2 + x_i^{2\gamma^*}\sigma^2\right) - \left(\theta_k\beta_k x_i\right)^2.
\end{aligned}
$$

Substituting expression of $V_M(d_{ki}y_i)$ in (6.2.19) and simplifying, we get

$$
V_M(\hat{t}_{yk}^g - t_{yk}) = \theta_k(1-\theta_k)\beta_k^2\kappa_1(x) + \theta_k\sigma_k^2\kappa_2(x). \tag{6.2.20}
$$

Note that the expectation and variance are applied after conditioning on the auxiliary data which is not shown mathematically. GPM for domain estimation provides a rationale of utilizing the auxiliary data from whole population to certain sub-populations. The idea of domain estimation using different population under matrix approach has been dealt in Chambers and Clark (2012). In this study, we revise the domain estimation problem for single auxiliary variable when ranked set sampling within finite population is performed. The GPM is preferred, when values of the study variable (y) generated from a stochastic process is proportional to the corresponding values of the auxiliary variable (x) which is assumed to be known for whole population prior to the survey or obtained during the survey (under two-phase sampling scheme). For GPM, it is also assumed that the variation in y depends on the value of x^{γ^*}, where γ^* is the rate which controls how much the variation in Y depends on x.

6.3 Domain Estimation Estimation Under RSSWOR

The well known ranked set sampling, introduced by McIntyre (1952), is preferred when it is economical to rank small sets of units while taking actual measurement from a larger sample is expensive. In Chapter 5, ranked set sampling without replacement (RSSWOR) is used under model-based approach after making some modification for estimation of the population total. In this section, we provide RSSWOR for estimation of domain specific totals. Consider a finite population U generated from a superpopulation with mean $\mu_{[i]}$ and variance $\sigma_{[i]}^2$ for ith judgment ordered random variable $y_{[i]}$ for $i \in U$. For judgment ranked unit the respective parameters are subscripted in square brackets. To obtain a RSSWOR from a finite population generated from the superpopulation, we introduce following algorithm.

(i). Divide the finite population into t sub-populations at random with sizes N_l for $l =$

$1, 2, ...t$ such that $N = \sum_{l=1}^{t} N_l$, where t is the number of cycles or sampling interval. Further, it is assumed that every sub-population is large enough to select m^2 units from them i.e. for all l, $N_l > m^2$. The concept of, so called, sub-populations is defined just for obtaining larger sets to ensure that sampling is without replacement. For a valid inference this division must be independent with the study variable. Note that this division of population into t sub-populations is haphazard and is totally different from the concept of domain. The division is merely made for sample selection purpose.

(ii). Select m^2 units from each sub-populations and divide each m^2 units in m sets each of size m.

(iii) Rank each set within itself according to some ranking mechanism different ranking mechanisms are discussed in Ahmed et al. (2017).

(iii). Select the ith ranked unit from the ith set for $i = 1, 2, 3, ..., m$, and $l = 1, 2, .., t$ to obtain a ranked set sample without replacement of size tm after retaining $t(m^2 - m)$ units from the ranked sets.

(iv). Observe $y_{i[i]l}$, $x_{i[i]l}$ and $d_{i[i])l}$ from the final RSSWOR to obtain domain specific estimates. An illustration of RSSWOR for domain estimation is provided in Figure 6.1.

Figure 6.1 explains RSSWOR for domain estimation assuming that a finite population of size N is coming from a large superpopulation with specified mean and variance which is generating via some stochastic process. From the finite population of size N units, we consider t different cycles with sizes N_1, N_2, ...,N_t randomly, leaving $(N - \sum_{l=1}^{t} N_l)$ as non-sampled. In this way, we are left with t so called sub-populations. From each sub-populations, we then select m^2 units for ranking leaving $(N_l - m^2)$ units from each sub-population as non-sampled, where $m$is taken three for pictorial presentation. Finally, applying ranked set sampling for selecting m units from each cycle by returning remaining $(m^2 - m)$ non-sampled units. The total non-sampled units are found from two stages which is illustrated in Figure 6.1.

$$\begin{aligned}\text{Non-sampled} &= \text{Non-sampled at Stage-1} + \text{Non-sampled at Stage-2} \\ &= \sum_{j=1}^{t} (N_l - m^2)\end{aligned}$$

$$+\sum_{j=1}^{t}(m^2 - m) = N - tm.$$

Let s be the ranked set sample of size tm selected using the above mechanism and $\bar{s}$ be the set of units which are not in s. The RSSWOR s can be written as $s = \left\{y_{i[i]j}, x_{i[i]j}, d_{i[i]j}; i = 1, 2, ..., m, j = 1, 2, ..., t\right\}$.

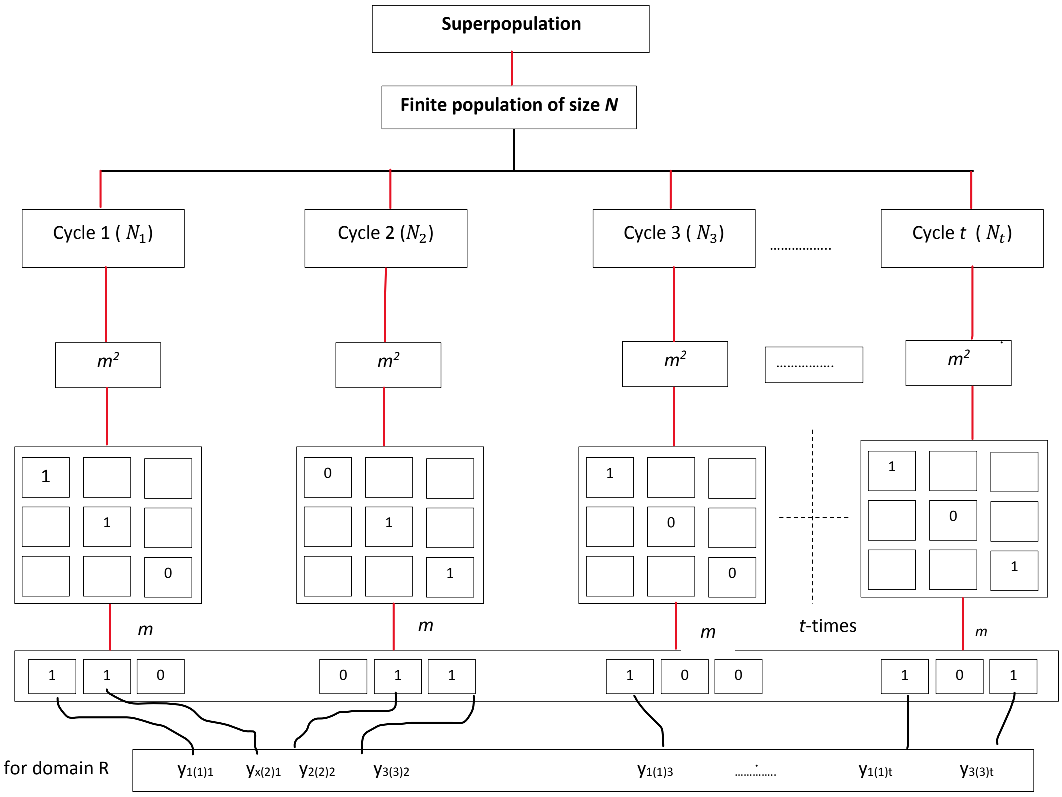

Figure 6.1: Flow chart of selection process of a RSSWOR with t cycles and $m = 3$ for domain specific estimation

6.3.1 Domain Estimation Under HPM Using RSSWOR

Let $y_{[i]}$ be the value of the study variable corresponding to the ith judgmental ranked unit from population N. The values $y_{[i]}$ (for $i = 1, 2, ..., m$) are assumed to be independently distributed in domain k as samples are drawn from independent sub-populations and independent sets. Under HPM the ordered population values can be modeled as:

$$y_{[i]} = \mu_{k[i]} + \varepsilon_{k[i]} \quad \text{for } \; i = 1, 2, ..., N \quad \text{and } \; k = 1, 2, ...h, \tag{6.3.1}$$

where $\mu_{k[i]}$ is domain specific mean of the ith ordered realizations for domain k, $\varepsilon_{k[i]}$ random error term independently normally distributed with zero mean and variance $\sigma^2_{k[i]}$ for the kth domain and h is the number of domains. The condition of zero mean and independence of error term is true only if ranking is performed on some variable other than the study variable. Hence, the ranking is assumed to be judgmental rather than perfect. The domain membership variables denoted as $d_{k[i]}$'s will not be much affected by ranking as it is binary and independently distributed Bernoulli random variables. Now the mean and variance of $d_{k[i]}y_{[i]}$ can be obtained as

$$E_M(d_{k[i]}y_{[i]}) = E_M(y_{[i]}|d_{k[i]} = 1)p(d_{ki} = 1) = \mu_{k[i]}\theta_k \tag{6.3.2}$$

and

$$V_M(d_{k[i]}y_{[i]}) = \theta_k\sigma^2_{k[i]} + \theta_k(1-\theta_k)\mu^2_{k[i]} = \sigma^{*2}_{k[i]}(\text{say}). \tag{6.3.3}$$

The covariance between $d_{k[i]}y_{[i]}$ and $d_{k[j]}y_{[j]}$ for $i \neq j$ is zero for all $i, j \in U$ as $y_{[i]}$ (conditionally) and $d_{k[i]}$ both are independent random variables as units are selected from different sets. Following Chambers and Clark (2012) the expansion estimator for $\hat{t}^E_{yk[rss]}$ under RSSWOR is obtained as:

$$\hat{t}^E_{yk[rss]} = \frac{N}{tm}\sum_{l\in c}\sum_{i\in s} d_{ki[i]l}y_{i[i]l} = \frac{N}{tm}\sum_{l\in c}\sum_{i\in s_k} y_i[i]l = N\hat{\theta}_k\bar{y}_{k[rss]}, \tag{6.3.4}$$

where the notation $\sum_{l\in c}$ represents that summation is taken over cycles. Further, $\bar{y}_{k[rss]} = \frac{1}{tm}\sum_c\sum_{i\in s} d_{ki[i]l}y_{i[i]l}$ is the RSSWOR mean for the kth domain and $\hat{\theta}_k$ is estimate of θ_k obtained through RSSWOR. To obtain a best linear unbiased estimator for population total t_{yk} under RSSWOR, we write the expansion estimator as the linear combinations $d_{ki[i]}y_{i[i]}$ i.e. $\hat{t}^E_{yk[rss]} = \sum_{i\in s} w_{[i]}d_{ki[i]}y_{i[i]}$, where $w_{[i]}$ are the weights associated with the ith ranked unit from the ith set for predicting non-sampled units. Further

$$\begin{aligned}\hat{t}^E_{yk[rss]} &= \sum_{i\in s} d_{ki[i]}y_{i[i]} + \sum_{i\in s}\left(w_{[i]}-1\right)d_{ki[i]}y_{i[i]}\\ &= t_{yk[rss]} + \sum_{i\in s} w^*_{[i]}d_{ki[i]}y_{i[i]},\end{aligned}$$

where $w^*_{[i]} = w_{[i]} - 1$ is the prediction weight corresponding to the ith sampled unit for predicting non-sampled part of the population and $t_{yk[rss]} = \sum_{i\in s} d_{ki[i]} y_i$. The corresponding domain specific population total is $t_{yk} = t_{yk[rss]} + t_{ykr}$, where t_{ykr} is the domain total for non-sampled units. The prediction error for $\hat{t}^E_{yk[rss]}$ is obtained as $\hat{t}^E_{yk[rss]} - t_{yk} = \sum_{i\in s} w^*_{[i]} d_{ki} y_{i[i]} - t_{ykr}$ The model bias of the $\hat{t}^E_{yk[rss]}$ is

$$E_M(\hat{t}^E_{yk[rss]} - t_{yk}) = \sum_{i\in s} w^*_{[i]} E_M(d_{ki[i]} y_{i[i]}) - \sum_{i\in \bar{s}} E_M(d_{ki[i]} y_{i[i]})$$
$$B_M(\hat{t}^E_{yk}) = \theta_k \big[\sum_{i\in s} w^*_{[i]} \mu_{k[i]} - (N-m)\mu_k\big].$$

The bias vanishes only if

$$\sum_{i\in s} w^*_{[i]} \mu_{k[i]} = (N-n)\mu_k. \tag{6.3.5}$$

The model variance of prediction error can be obtained as

$$V_M(\hat{t}^E_{yk[rss]} - t_{yk}) = \sum_{i\in s} w^{*2}_{[i]} V_M(d_{ki[i]} y_{i[i]}) + \sum_{i\in \bar{s}} V_M(d_{ki} y_i) - 2Cov(\sum_{i\in s} w^*_{[i]}(d_{ki[i]} y_{i[i]}), \sum_{i\in \bar{s}}(d_{ki} y_i))$$
$$= \big[\sum_{i\in s} w^{*2}_{[i]} \sigma^{*2}_{k[i]} + (N-n)\sigma^{*2}_k\big].$$

The covariance term in variance expression vanishes as the the covariance between $d_{ki[i]} y_i$ and $d_{kj[j]} y_{j[j]}$ for $i \neq j$ is zero. The problem is to minimize $V_M(\hat{t}^E_{yk[rss]} - t_{yk})$ with respect to $w^*_{[i]}$ subject to restriction given in (6.3.5). The Lagrangian function L contains $\sigma^{*2}_{k[i]}$ and $\mu_{k[i]}$, hence the optimum value involves these unknown parameters which make the estimation complex. One possible solution is to carry with $w^*_{[i]} = N/tm$ and $w_{[i]} = 1 - N/tm$. For single cycle i.e. $t = 1$ $w^*_{[i]} = N/m$ and $w_{[i]} = 1 - N/m$. The expansion estimator can be written as $\hat{t}^E_{yk[rss]} = \frac{N}{m} \sum_{i\in s} d_{ki[i]} y_{i[i]}$ with variance

$$V_M(\hat{t}^E_{yk[rss]} - t_{yk}) = \big[\frac{(N-n)^2}{n} \sum_{i\in s} \sigma^{*2}_{k[i]} + (N-n) \sum_{i\in s} \sigma^{*2}_{k[i]}\big]$$
$$= \frac{N(N-n)}{n} \sigma^{*2}_k - \frac{(N-n)^2}{n} \sum_{i\in s} \delta^{*2}_{k[i]},$$

where $\delta^{*2}_{k[i]} = \theta_k \mu_{k[i]} - \theta_k \mu_k$.

6.3.2 Domain Estimation Under GPM Using RSSWOR

Assuming that there exists a covariate whose values are known for all population so we perform the RSSWOR in same way as performed early by adding the judgmental ranked values on x if ranking is performed based on some other variable or the ranked values of x if ranking is performed on x. The GPM for ranked data in the kth domain k can be defined as

$$y_{[i]} = \beta_k x_{[i]} + x_{[i]}^{\gamma^*} \varepsilon_{k[i]} \quad \text{for } i = 1,2,...,N \quad \text{and} \quad k = 1,2,...h, \tag{6.3.6}$$

where γ^* is a real constant assumed to be known or guessed in advance and β_k is the domain specific coefficient corresponding to the covariate x. The general prediction estimator given in Valliant (2000) for GPM can be written under RSSWOR as

$$\hat{t}^{g}_{yk[rss]} = \sum_{i\in s} d_{ki[i]} y_{i[i]} + \sum_{\bar{s}} \left(\hat{\beta}_{k[rss]} x_i\right), \tag{6.3.7}$$

where $\hat{\beta}_{k[rss]}$ is the BLUE of β_k under RSSWOR which is estimated as follow. For assuring linearity, we can write $\hat{\beta}_{k[rss]}$ as a linear combination of the domain specific responses $d_{ki[i]}y_{i[i]}$ as $\hat{\beta}_{k[rss]} = \sum_{i\in s} c_{i[i]} d_{ki[i]} y_{i[i]}$, where $c_{i[i]}$ is the weight associated to the ith judgment ranked unit in ith set to be obtained as a function of known auxiliary data from sampled as well as non-sampled parts of the population. The model expectation (E_M) and variance (V_M) of $\hat{\beta}_k$ can be obtained as

$$E_M(\hat{\beta}_{k[rss]}) = \sum_{i\in s} c_{i[i]} E_M(d_{ki[i]} y_{i[i]}) = \beta_k \sum_{i\in s} c_{i[i]} x_{i[i]}. \tag{6.3.8}$$

Unbiasedness is obtained only if

$$\sum_{i\in s} c_{i[i]} x_{i[i]} = 1 \tag{6.3.9}$$

and

$$V_M(\hat{\beta}_{k[rss]}) = \sum_{i\in s} c_{[i]}^2 V_M(d_{ki[i]} y_{i[i]}) = \sum_{i\in s} \sigma^{*2}_{k[i]} c_{[i]}^2 x_{i[i]}^{2\gamma^*} \tag{6.3.10}$$

The problem is to minimize $V_M(\hat{\beta}_k)$ subject to the constrain given in (6.3.9). The Lagrangian function is again a function which involves unknown population parameters $\sigma^2_{k[i]}$, hence we can't proceed with Lagrangian's method to obtain BLUE of β. However directly minimizing the sum of squared residuals under OLS method gives $\hat{\beta}_{k[rss]} = \frac{\sum_{i\in s} d_{ki[i]} x_{i[i]}^{1-2\gamma^*} y_{i[i]}}{\sum_{i\in s} x_{i[i]}^{2-2\gamma^*}}$. The BLUP for domain total is then obtained as

$$\hat{t}_{yk}^{g} = \sum_{i\in s} d_{ki[i]} y_{i[i]} + \frac{\sum_{i\in s} d_{ki[i]} x_{i[i]}^{1-2\gamma^*} y_{i[i]}}{\sum_{i\in s} x_{i[i]}^{2-2\gamma^*}} \sum_{\bar{s}} x_i = \sum_{i\in s} (1+u_{[i]}^*) d_{ki[i]} y_{i[i]}, \tag{6.3.11}$$

where $u_{[i]}^* = \frac{x_{i[i]}^{1-2\gamma^*}}{\sum_{i\in s} x_{i[i]}^{2-2\gamma^*}} \sum_{\bar{s}} x_i$. The prediction error for $\hat{t}_{yk}^{g}$ is given by

$$\hat{t}_{yk[rss]}^{g} - t_{yk} = \sum_{i\in s} u_{[i]}^* d_{ki[i]} y_{i[i]} - t_{ykr}. \tag{6.3.12}$$

It is straight forward to show that (6.3.12) is unbiased with prediction error variance, given by

$$V_M(t_{yk[rss]}^{g} - t_{yk}) = \sum_{i\in s} u_{i[i]}^{*2} V_M(d_{ki[i]} y_[i]i) + \sum_{i\in \bar{s}} V_M(d_{ki} y_i) - 2Cov\Big(\sum_{i\in s} u_{i[i]}^{*} (d_{ki[i]} y_{i[i]}), \sum_{i\in \bar{s}} (d_{ki} y_i)\Big). \tag{6.3.13}$$

The covariance term in (6.3.13) vanishes as the the covariance between $d_{ki[i]}y_{i[i]}$ and $d_{kj}y_j$ for $i \neq j$ is zero, where $i \in s$ and $j \in \bar{s}$. We know under GPM, $E_M(d_{ki[i]}y_{i[i]}) = \theta_k \beta_k x_{i[i]}$. Further

$$\begin{aligned}
V_M(d_{ki[i]}y_{i[i]}) =& E_M(d_{ki[i]}y_{i[i]})^2 - \left\{E_M(d_{ki[i]}y_{i[i]})\right\}^2 \\
=& p(d_{ki[i]} = 1) E_M\left(y_{i[i]}^2 | d_{ki[i]} = 1\right) - \left\{p(d_{ki[i]} = 1) E_M\left(y_{i[i]} | d_{ki[i]} = 1\right)\right\}^2 \\
=& \theta_k\left(\beta_k^2 x_{[i]}^2 + x_{[i]}^2 \sigma_{k[i]}^2\right) - \left(\theta_k \beta_k x_{[i]}\right)^2 \\
=& \theta_k \beta_k^2 x_{[i]}^2 + \theta_k x_{[i]}^{2\gamma^*} \left(\sigma_k^2 - \delta_{k[i]}^2\right).
\end{aligned}$$

Substituting $V_M(d_{ki[i]}y_{i[i]})$ in (6.3.13) and simplifying, we get

$$\begin{aligned}
V_M(t_{yk[rss]}^{g} - t_{yk}) =& \theta_k(1-\theta_k)\beta_k^2 \kappa_{[1]}(x) + \theta_k \sigma_k^2 \kappa_{[2]}(x) - \theta_k \sum_{\bar{s}} x_{[i]}^{2\gamma^*} \delta_{k[i]}^2 && (6.3.14) \\
=& V_M(\hat{t}_{yk}^{g} - t_{yk}) - \theta_k \sum_{\bar{s}} x_{[i]}^{2\gamma^*} \delta_{k[i]}^2, && (6.3.15)
\end{aligned}$$

where $\kappa_{[1]}(x) = \sum_{i\in s} u_{i[i]}^{*2} x_{i[i]}^2 + \sum_{i\in \bar{s}} x_i^2$ and $\kappa_{[2]}(x) = \sum_{i\in s} u_{i[i]}^{*2} x_{i[i]}^{2\gamma^*} + \sum_{i\in \bar{s}} x_i^{2\gamma^*}$ Equation (6.3.14) shows the superiority of the domain estimator under ranked set sampling over simple random sampling when units are taken without replacement. The rational behind using GPM as working model is already debated in previous section.

6.4 Simulation Study

For the purpose of efficiency comparison, we conduct a Monte Carlo (MC) experiment by generating a hypothetical population of size $N = 1000$. We generate the cumulative probabilities p_i^* from Uniform $(0,1)$. The values of $X_i =$qgamma(p_i^*, a, b) and $d_i =$qbinom$(p_i^*, 1, \theta_k)$. The values of y_i are obtained using the relationship $y_i = \rho_k^2 X + X_i^{\gamma^*} e_k$ for $\gamma^* = 0.3, 0.5, 0.8$ $i = 1, 2, .., N$, where e is an i.i.d error term, normally distributed with zero mean and variance σ_k^2 i.e. $e_i =$qnorm$(p_i^*, 0, \sigma_k^2)$. Further, qgamma, qbinom, qnorm are used to denote the quantile points corresponding to the cumulative probabilities in Gamma, Binomial and Normal distribution respectively. Note that the term Gamma distribution is quiet different from that of the term used in the GPM. The GPM is named only because a constant γ^* is used in the power of x with error terms. While the Gamma distribution is a widely used parametric distribution used for modeling the continuous random variables with specific nature. The parameters ρ_k and σ_k^2 are assumed to be fixed at 0.8 and 5 respectively for the domain of interest. A RSSWOR procedure is applied according to the algorithm given in Section 6.3, the ranking is performed on the study variable itself for simplicity. The domain specific estimators for population total under RSSWOR are obtained for HPM and GPM with different choices of γ^*, t and m. A SRSWOR of size $n = tm$ is also selected and obtained the corresponding estimators for the purpose of comparison. Repeat both sampling process 5,000 times and obtain the expected squared prediction error (ESPE) of each estimator as a measure of variation. The ESPE is defined as:

$$ESPE(\hat{t}_{yk}) = \frac{1}{\text{sim}} \sum_{\text{sim}} (\hat{t}_{yk} - t_{yk})^2, \qquad (6.4.1)$$

where sim denotes number of repeated samples selected from the population. The ESPE for other estimators can be obtained after replacing $\hat{t}_{yk}$ by the relevant estimators in (6.4.1). The results are given in Tables 6.1. and 6.2.

Table 6.1.: ESPE and relative efficiency of the proposed estimators

			$\theta_k = 0.4$				$\theta_k = 0.6$			
g	t	m	ESPE(h)	ESPE(g)	RE(h)	RE(g)	ESPE(h)	ESPE(g)	RE(h)	RE(g)
					$G(2,2)$					
0.3	5	6	331693.8	353110.8	1.2381	1.1630	309143.9	322310.4	1.6145	1.5486
		8	253265.5	269743.2	1.2776	1.1995	232167.5	241735.7	1.6592	1.5936
		10	209379.2	224181.2	1.2950	1.2095	178806.9	183985	1.7168	1.6685
	8	6	224578.1	237778.4	1.2529	1.1833	212300.6	216776.5	1.5316	1.5000
		8	190889.5	201099.5	1.2082	1.1469	157310.8	159819.4	1.6177	1.5924
		10	163210.3	171288.9	1.2096	1.1525	128920.7	127291.5	1.5613	1.5813
0.5	5	6	356422.4	360932.6	1.2268	1.2114	332074.7	337501.1	1.5653	1.5401
		8	268326.9	269879.3	1.2777	1.2704	241639.6	244941.5	1.6546	1.6323
		10	218210.7	220666.8	1.2949	1.2805	184164.6	186341.9	1.7134	1.6934
	8	6	240028.1	240455.1	1.2303	1.2281	220281.8	223504.3	1.5186	1.4967
		8	195031.5	195487	1.2346	1.2317	163503.3	164515.8	1.5931	1.5833
		10	166873	167336.3	1.2217	1.2183	130675.5	131120.2	1.5655	1.5602
0.8	5	6	441477.4	338738	1.2190	1.5888	416659.2	344536.7	1.5049	1.8200
		8	333367.9	250583.3	1.2550	1.6696	302033.8	245398.5	1.5860	1.9521
		10	263395.9	203821.3	1.2719	1.6437	226452.7	186784.6	1.6486	1.9987
	8	6	291361.7	222221	1.2176	1.5965	270741.3	223717.6	1.4616	1.7689
		8	232180	176981.3	1.2285	1.6116	197273.7	158817.1	1.5460	1.9204
		10	191866.4	148635.7	1.2362	1.5958	152526.9	124921.4	1.5558	1.8995
					$G(2,4)$					
0.3	5	6	150863.75	151060.2	1.2261	1.2245	185436.58	195195.5	1.6690	1.5855
		8	112879.93	114920.5	1.2422	1.2201	133153.72	142912.5	1.7567	1.6367
		10	89289.22	88224.8	1.2886	1.3042	99837.92	104674.2	1.8231	1.7389
	8	6	81425.81	80551.97	1.1710	1.1838	119466.76	123532.5	1.6211	1.5678
		8	59923.2	58262.28	1.2109	1.2454	86093.94	89846.6	1.7230	1.6510
		10	49741.05	49076.52	1.1676	1.1834	66437.85	67725.47	1.7078	1.6754
0.5	5	6	126872.3	128743.95	1.1811	1.1640	155029.78	157980.33	1.6096	1.5796
		8	94889.47	96906.3	1.2116	1.1864	111830.99	114188.98	1.6924	1.6574
		10	75021.07	75749.28	1.2509	1.2389	83646.08	85065.56	1.7567	1.7274
	8	6	78302.49	78947.91	1.2492	1.2390	99659.11	101310.61	1.5679	1.5424
		8	60718.05	61363.78	1.2185	1.2057	72496.54	73060.64	1.6542	1.6415
		10	50506.57	50733.44	1.1769	1.1717	55729.97	55941.56	1.6514	1.6452
0.8	5	6	108138.26	91769.56	1.1499	1.3550	134541.45	110531	1.5188	1.8487
		8	80863.53	69788.35	1.1985	1.3887	97534.7	78939.87	1.5931	1.9683
		10	63011.9	54015.68	1.2442	1.4515	72582.65	59100.69	1.6595	2.0381
	8	6	67933.38	57567.62	1.2042	1.4211	86363.87	70807.01	1.4796	1.8047
		8	52140.77	45164.05	1.1949	1.3795	62949.22	50473.72	1.5576	1.9425
		10	43881.83	38471.23	1.1347	1.2943	48314.63	38886.16	1.5665	1.9464

Table 6.2.: ESPE and RE of the proposed estimators (continued)

			$\theta_k = 0.4$				$\theta_k = 0.6$			
g	t	m	ESPE(h)	ESPE(g)	RE(h)	RE(g)	ESPE(h)	ESPE(g)	RE(h)	RE(g)
					$G(4,2)$					
0.3	5	6	779609.6	772909.2	1.1080	1.1176	617420.7	624585.5	1.4924	1.4753
		8	689467.6	682327	1.0949	1.1063	501409.9	506820.1	1.4829	1.4671
		10	627567.4	618805.8	1.1196	1.1354	416386.4	417869.9	1.4980	1.4927
	8	6	630602.1	623380.5	1.1395	1.1527	469790.1	471951.5	1.3895	1.3832
		8	580623.1	570195.1	1.0858	1.1057	375296.4	374426.6	1.4333	1.4366
		10	561304.2	552892.1	1.0418	1.0577	335271.5	333066.9	1.3502	1.3592
0.5	5	6	835957.1	837672.8	1.1314	1.1290	750443.8	755013.2	1.5296	1.5203
		8	722852.9	725950.9	1.1113	1.1065	583482.4	586896.2	1.5565	1.5474
		10	637584.8	637964.3	1.1413	1.1406	465411.7	467497.9	1.5968	1.5897
	8	6	642117.6	643071.4	1.1662	1.1644	539337.2	542087.4	1.4506	1.4432
		8	566984.4	567459.8	1.1215	1.1206	421467.2	421697.4	1.4966	1.4958
		10	543438.1	543437.5	1.0569	1.0569	360967.6	360358.4	1.4306	1.4331
0.8	5	6	1073696.6	1000057.1	1.1410	1.2250	1146190	1029564	1.5512	1.7269
		8	861808.7	812074.1	1.1522	1.2228	854952.2	771923.9	1.6055	1.7782
		10	720264.6	676716.1	1.1826	1.2587	649367	587369.6	1.6736	1.8502
	8	6	734504.4	691629.7	1.2044	1.2791	643511.8	580876.2	1.4562	1.6132
		8	612651.2	582162.6	1.1574	1.2180	495917.8	453378.1	1.4561	1.5928
		10	554142.3	529417.6	1.0916	1.1426	396054.2	362250.9	1.4413	1.5758

The ESPE and RE for three different varieties of Gamma distribution $G(a,b)$, two choices of t, three choices of m, three values of the constant γ^* and two values of θ_k are presented in Tables 6.1. and 6.2.. One can also observe that the behavior of the domain specific total estimators for all choices of mentioned parameters and constants. Apart from these constants, the efficiency of domain specific estimates also depends on the homogeneity of the domain and the relationship between the study variable and the auxiliary variable within domain. The domain specific correlation can only be observed once a sample is selected. Although the correlation between the study variable and the auxiliary variable for whole population is fixed in advance. The ESPE for both models (HPM and GPM) decreases with increase in the set sizes m keeping other things constant. For fixed set size m, the ESPE for HPM increases with increase in γ^*. For example, in domain with $\theta_k = 0.4$ and $t = 5$, when $m = 10$ MSE(h)$= 20.93 \times 10^4$ for $\gamma^* = 0.3$, 21.82×10^4 for $\gamma^* = 0.5$ and 26.34×10^4. While for $m = 6$ the ESPE(h) values are 33.17×10^4, 35.64×10^4 and 44.14×10^4 respectively for $m = 6$. For GPM, the ESPE is slightly higher for $\gamma^* = 0.5$ as compared to other two choices of

γ^*. The relative efficiency of the domain specific total estimator also increases with increase in set size m for fix choices of other constants. The relative efficiency for both HPM and GPM are higher for larger domains as compared to smaller ones. For G(2,4), $m = 6$, $t = 5$ and $\gamma^* = 0.3$ the RE(h)= 1.23 and RE(g)= 1.22 when $\theta_k = 0.4$. While RE(h)= 1.67 and RE(g)= 1.58 respectively for $\theta_k = 0.6$. Finally, comparing the ESPE and RE for different data generating distributions, it is found that the efficiency of domain specific RSSWOR estimators decreases with decrease in variance of the auxiliary variable x i.e. a/b^2.

6.5 Conclusion

The estimation of sub-population total under a new version of ranked set sampling for obtaining a without replacement sample with GPM (general form of proportional population model) has been dealt in this chapter. Figure 6.1 presented a picture of the RSSWOR sampling layout which assumes that the finite population is coming from an infinite superpopulation via some stochastic process with finite mean and variance. Domain membership variable was observed from selected ranked set sample. The model relationship between the study variable and the auxiliary variable for whole population was used to predict the non-sampled values to establish a domain specific estimator for total. The mathematical expressions and Monte-Carlo experiment both support the superiority of the domain specific predictor under RSSWOR over the total predictor under SRSWOR for GPM as well as HPM. The domain specific estimators are widely used in epidemiology and public health where one need to find total exposure to certain event for different sub-populations. For example, the patients suffering certain disease can be ranked according to their age and total number of infections are recorded from the patients belonging to a certain suburb. More sophisticated small domain estimators can be constructed using multi-level fixed effect and random effect models.

Chapter 7

Application on Pakistan Demographic and Health Survey Data 2017-18

7.1 Outline

Demographic Health Surveys (DHS) contain very useful and detailed information about the demographic characteristics and the factors affecting them. The DHS data can be utilized to predict the average rates of occurrence of vital events for non-sampled part of the country's population using information about appropriate available covariates. In this study, we first see the effect of various socio-demographic factors on births by fitting regression models using Pakistan Demographic Health Survey (PDHS) 2017-18 data. Poisson regression and its extensions (Negative Binomial (NB), Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB)) are used to model the birth data. The births occurred during the time periods of 1 year, 3-years and 5-years are taken as the responses for each model. Both classical and Bayesian estimations are performed for drawing statistical inference about the fertility models. We then show how the model-based approach works in efficiency improvement for estimation of birth rates (Age specific fertility rates (ASFR), total fertility rate (TFR), general fertility rate (GFR) and gross reproduction rate (GRR)) for ever-married women after converting birth data into person-years data. The performance of model-based rates are examined using a boot strapped sampling algorithm. The country and regional level predictive estimates for fertility rates (i.e. TFR, GFR and GRR) are found appropriate. While the ASFR are over-estimated for some age groups and under-estimated for others. In Section 7.2, we give a brief introduction of the application. Section 7.3 provides a review on estimation of fertility indicators from DHS data sets using already developed packages. Section 7.4 delineates some available regression models for number of births. Sections 7.5 and 7.6 give model estimation and results under frequentist and Bayesian frameworks respectively. Section 7.7 applies model-based techniques for estimating birth rates. A short conclusion is given in Section 7.7.

7.2 Introduction

DHS are nationally representative household surveys which have been conducting since 1984 in more than 85 countries. The DHS were basically designed to explore demographic, family planning and fertility data collected in the Contraceptive Prevalence Surveys (CPS) Chamratrithirong et al. (1986) and World Fertility Surveys (WFS) Lightbourne et al. (1982), and to provide a necessary resources for the monitoring and evaluation of vital statistics and health indicators in developing countries. The DHS collect data on a wide range of objectives with a focus on fertility indicators, maternal and child health, reproductive health, nutrition, mortality and health behavior in adults. The main advantages of the DHS are high response rates, employment of qualified and trained interviewers, national coverage, worldwide standardized data collection procedures and consistent material over time, comparable across populations cross-sectionally as well as over time.

7.3 Estimation of Fertility Indicators from Survey Data

The direct estimation methods are widely used to estimate fertility indicators from household surveys. These methods were initially used by the World Fertility Survey (WFS), which conducted from 1972 to 1984, and afterwards by the DHS Program. The approaches utilized by WFS/DHS have been well documented in several articles like Croft et al. (2018); Moultrie et al. (2013); Hill (2013). The approaches have later been used by other household surveys (HS) programs, such as the multiple indicator cluster surveys (MICS). In the direct estimation methods, data about the 3 or 5 years birth history are gathered and used for the calculation of fertility indicators UN (2011).

In DHS surveys, the data on birth histories (month and year of birth of each child), sex of each child and age of each surviving child are used for calculating fertility rates. All the indicators are calculated as occurrences/exposure rates because the rates are calculated from birth histories collected in the survey, where the numerator is the number of births in certain time period usually 3 or 5 years and the denominator is the population exposed to risk of the birth within the reference period. In this section, we discuss the results obtained through DHS.rates package in R for PDHS 2017-18 data. The detail definitions and formula for age-specific fertility rates (ASFR) are given in introduction section (see Chapter 1). The definitions of the fertility indicators given in introduction section and methods of their calculation can be found in the Guide to DHS Statistics Croft et al. (2018). In Table E.1 (see Appendix E), the widely used birth indicator, ASFR per thousand women, is given in Column 3 corresponding to seven standard age groups. In addition to the estimates of ASFR for the seven age groups, the table also provides standard errors (SE), number of exposures (N), weighted number of exposures (WN), design effect (DEFT), relative standard error (RSE) and lower and upper bounds of confidence interval. The results are obtained with the reference period of length 3 years (36 months), the end date of the interview for Islamabad Capital Territory (ICT) was 22 November 2017 and the average reference period was 7 May 2016 in century month code (CMC). Table E.2 (see Appendix E) provides ASFRs for 6 different regions (four provinces, ICT and FATA) excluding Gilgit-Baltistan and Azad Jammu and Kashmir (AJK)[1]. From Table E.2, we see that ASFR corresponding to 15-19 age group is highest for FATA as compared to other regions i.e. 229 births per 1000 women. While ICT has second highest ASFR in age group 15-19 with 219 birth per 1000 women. The lowest early age (15-19) reproductivity is observed in Khayber Pakhtun Khwah (KPK) which is 184 births. For detail visual comparison, readers are referred to stat.compiler page of DHS program website. After computing ASFR the computation of TFR. GFR and GRR is straightforward and the detailed description about the computation of these measures are given in introduction section. Regional level estimates of TFR, their standard error (SE), cell frequencies and 95 % confidence interval for the true TFR's are given in order from left to right columns of Table E.4. We can observe highest total fertility rate in federally administered tribal areas (FATA) (about 6.35 births per women) and lowest in ICT (4.9 births per women). The last row of Table E.4 shows the total fertility at national level. From PDHS 2017-18 dataset, using DHS.rates package, we obtain GFR at sub-national level. Sub-national level estimates of GFR, their standard errors (SE), cell frequencies and 95 %

[1] The results in PDHS report PDHS (2019) are presented by excluding Azad Jammu and Kashmir and Gilgit Baltistan when it refers to Pakistan.

confidence intervals for the true GFR's are given in order from left to right columns of Table E.5 (see Appendix E). We can observe highest fertility rate in FATA (about 226 births per 1000 women) and lowest in ICT (about 163 births per 1000 women). The widest confidence interval for estimated GFR is observed for FATA with around 55 births and the narrowest one for Punjab with around 20 births. The general fertility at national level is given in last row of Table E.5. The GRR are not included in PDHS report as its calculation is not included in DHS.rates package.

7.4 Regression Models for Number of Births

After completing the hectic work on collecting data related to fertility, DHS reports only provide tabulation of cell frequency and visual display of the relationship between the fertility indicators (number of birth during a specified period per women) and other demographic and socio-economic factors (Region, Sex, Marital Status, Age, Education etc). Apart from reaching to a valid statistical inference, estimation of parameters and constructing confidence intervals on the basis of estimated rates, one can also develop regression models for number of births and estimate them to observe the relationship between number of births and its candidate determinants. Birth histories during specific period constitute the primary source for the majority of research on fertility in developing countries like Pakistan. Methods for analyzing births data differ substantially depending on the type of the study (descriptive or explanatory). The ASFRs and TFR found in survey reports and fertility trend estimates are calculated on classical demographic analysis methods. Explanatory studies utilize regression models: logistic regression model Angeles et al. (1998), Poisson regression model Mencarini (1999) and event history methods Raftery et al. (1995). Although the principle has rarely been described in demographic analysis manuals, regression methods, specially Poisson regression, can still be practiced to calculate classical demographic measures such as TFR and other fertility rates Powers and Xie (2008). The main reason of practicing regression models, in fact, makes possible to include variables effecting fertility measures Zeileis et al. (2008). Schoumaker and Hayford (2004) used the 1998–99 Burkina Faso DHS data to explain the estimation of TFR and ASFRs using individual data with Poisson regression. The number of births over 5-years period preceding the interview (variable predefined in the DHS WOMEN recode files PKIR71) is taken as the outcome variable and the 5 years age groups (as dummy variables) in the model controlling for the length of exposure (5-years corresponding to each woman) using a term (offset). Average number of births is obtained by exponentiating the coefficients for each of the seven age groups without introducing intercept, and computing the TFR as the sum of the rates multiplied by five. Fitting an appropriate statistical model gives a clear picture of factors effecting births and prediction of births for non-sampled geographical units. One of the important assumptions of linear regression model is that the model residual error follows a normal distribution. This assumption is often met through transformation of the response variable when a continuous response variable is skewed. However, when the response variable is categorical or discrete but not continuous a simple transformation can't produce normally distributed residual error Gardner et al. (1995); Long (1997). Generalized Linear Models (GLMs) are the models in which outcome variables have a distribution instead of normal which contradict to the Linear regression models, where response variables are assumed to follow normal distribution. Because GLMs have categorical variables of interest such as “Yes' /“No" responses; or belonging to Groups A, B and, therefore, do not have full range i.e. $-\infty$ to $+\infty$. Hence, the relationship between response variable and predictor variable may not be linear.

Let y be the observed response corresponding to a random variable Y whose values are unknown for a finite population of size N indexed as $\mathcal{U} = \{1,2,3,....,N\}$. In matrix notation $y = (y_i, i \in U)$ be the realized stochastic vector of $Y = (Y_i, i \in \mathcal{U})$ under model-based approach. Suppose a sample $s = \{1,2,3,...,n\}$ of size n is drawn from the finite population $\mathcal{U}$ using a 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:

$$Y = X\beta + \varepsilon \tag{7.4.1}$$

where X is the data matrix containing p regressors including intercept and ε be the vector of random error assumed to be distributed normally with mean vector 0 and variance-covariance matrix Σ. Following sub-sections cover some generalized linear regression models which we use as working model for births per women during 1, 3 and 5-years periods prior to interview.

7.4.1 Poisson Regression Model

When the response variable is the number of occurrences of an event, the distribution of counts is discrete and is bound to non-negative integered values. Researchers might face one of the two problems while applying an ordinary linear regression model to such type of data. First, often such count data has positively skewed distribution with many observations having value 0. With a large number of 0's in the data set, one can't transform such skewed distributions into normal. Second, it is quite possible that the regression model produces negative predicted values which contradicts with the theory Cameron and Trivedi (2013). Another alternative method is to use a Poisson regression model or its extensions. The poisson model and its alternates have a number of advantages over an ordinary linear regression model. Poisson regression handles responses with skewed, discrete choice and non-negative values. A Poisson regression model works similar to ordinary linear regression model, with two exceptions. Firstly, it assumes that the error term has Poisson distribution instead of a normal distribution. Secondly, a Poisson regression models the natural logarithm of the response variable as a linear function of the coefficients rather than simply modeling the response variable as a linear function of the regression coefficients. A Poisson Regression model (PRM) is a GLM which is used to model data with response variable as counts. To proceed, it is assumed that the logarithm of mean values (rates) can be modeled into a linear form with some unknown parameters. In Poisson regression, the non-linear relationship is transformed into linear by using the log link function. This is why a Poisson Regression model is called the generalized log-linear model. The mathematical form of PRM is

$$\mathbf{log}\left(y\right) = X\beta \tag{7.4.2}$$

which is equivalent to

$$y = \mathbf{exp}\left(X\beta\right). \tag{7.4.3}$$

The coefficients vector β is calculated using Maximum Likelihood Estimation (MLE) or maximum quasi-likelihood estimation. Let μ be the rate parameter which is also the dispersion parameter of Poisson distribution then we can write

$$\mathbf{log}(\mu) = X\beta \tag{7.4.4}$$

The exponent of the coefficient β_j (jth component of β for $j = 0, 1, 2, ..., p$) for explanatory variables (X_j) thus shows the relationship between the number of births per women for which the explanatory variable has a specified value and the number of birth per women for which the variable has the specified value minus one, all other things remain constant.

7.4.2 Negative Binomial (NB) regression model

The Poisson regression model assumes that the error term (consequently responses for fixed covariate values) has same mean and variance which is not usually true in practice. In many cases, the mean of the error is smaller than its variance. There are two modified versions of the Poisson model that work well for the case of data with over-dispersed error term. In over-dispersed Poisson models, an extra parameter is included which tells how much the variance is larger than the mean. An alternative method for modeling the data with over-dispersed error term is to fit a negative binomial model Ver H. and Boveng (2007). The Negative binomial distribution is a version of the Poisson distribution where the distribution's parameter is considered as a random variable. The variation in this parameter can be considered for a variance of the data that is larger than the mean. In NB distribution, we require more parameterization to get a form which is appropriate to our regression. Following the notation given in Jackman (2000), we parameterize the NB density for ith observation with probability p_i and r. The former is known as the success parameter, and for ith observation it is defined as $p_i = \frac{r}{r+\mu_i}$, where μ_i satisfies the relation given in Equation (7.4.4) as in the Poisson model. The later is the over-dispersion parameter (≥ 0), it is equal to 1 in the Poisson distribution (i.e. there is no over-dispersion). The maximum likelihood estimates of the coefficients are obtained using MASS package in R. The detail about parameter estimation, model assumptions and validity of estimates can be found in (Cameron and Trivedi, 2013, page 326) and Hilbe (2011).

7.4.3 Zero-Inflated Poisson Model

As we already discussed that the count variables often follow a Poisson or one of distributions related to it. The Poisson distribution is based on the assumption that each count is the outcome of the same Poisson process i.e. a random process where events are equally likely and independent. If such count variable is used as the responses of a regression model, one can opt for Poisson regression to estimate the effect of predictors on the number of occurrence of events. But the Poisson regression model has very severe assumptions. One mostly violated assumption is the equality of mean and variance. The cases with responses having large variance and many 0's as well as a few very large values, the negative binomial model as an extension of Poisson handles the extra variance. But sometimes there may exist too many zeros than a Poisson would expect to predict. In such cases, a better option is to use Zero-Inflated Poisson (ZIP) model Atkins et al. (2013). In ZIP models, we assume that some zeros occurred by a Poisson process and some were not even able to have the event occur. Hence two processes work in ZIP: one determines whether the individual is eligible for a non-zero response and the second finds the count of that response for individuals who are eligible. The ZIP model runs two regression models simultaneously. A logistic (or probit) model is used to determine the probability of being eligible for a positive count and a Poisson model is used to model the size of the counts for eligible individuals with positive value. Both models utilize the same predictors, but with separate estimates for their coefficients. In this way, the predictor variables can have quite different effects on the processes. While a ZIP model needs it to be theoretically reasonable that some individuals are not eligible for a count. Zero-Inflation Poisson (ZIP) for response y is defined as:

$$P(y) = \begin{cases} \theta + (1-\theta)\text{Pois}(0|\mu), & \text{if } y = 0; \\ (1-\theta)\text{Pois}(y|\mu), & \text{if } y \geq 0 \end{cases}$$

where θ is the probability of occurring false values (zeros). Hence there are two models coupled together (a mixture model) to give an overall probability:

(i)-when a response is zero (i.e. $y_i = 0$), it is the probability of getting a zero plus the probability of a true value times probability of choosing a value of zero from a Poisson distribution with parameter μ and

(ii)-when a response is greater than 0, it is the probability of a true value times the probability of drawing that value from a Poisson distribution with parameter μ.

This definition indicates that the Poisson parameter μ is same for both the zeros and non-zeros components. The model of zero values are used for essentially investigating whether the likelihood of false zeros is related to the linear predictors. The greater than zero model, then, investigates whether the counts (non-zero responses) are related to the linear predictors. However, typically, we are less interested in modeling determinants of false '0'. It is better that the likelihood of false '0' be unrelated to the linear predictors. For example, if inflated (false '0') are due to issues of detectability (i.e. individuals are present, just not detected), then the detectability is not related to experimental treatments is considered as better. Any detectability issues same across all treatment levels is the most favorite situation. The expected value and the variance of the response y for a ZIP model are given by: $E_M(y_i) = \mu(1-\theta)$ and $V_M(y_i) = \mu(1-\theta) \times (1+\theta\mu^2)$. The model fitting requires an iterative process which is performed using MASS package in R. The detail about derivation and application of ZIP model is available in Cameron and Trivedi (2013).

7.5 Model Estimation Under Frequentist Framework

The birth data file (PKBR71FL.DTA) from PDHS 2017-18 is taken for data analysis. The detail about data collection mechanism, field work, training of staff and pretest has been discussed in PDHS (2019). We include key variables including demographic characteristics, socio-economic status and variables related to family planning as regressors. Before applying the different models to DHS data, we describe variables which we use in the model in Tables 7.1. and 7.2.. The visual display of number of birth for three time periods 1, 3 and 5 years. Panels (a), (b) and (c) of Figure 7.1 provides histogram for births occurred during 1-year, 3-years and 5-years period prior to the survey. Hence all three responses depict highly departure from normality so ordinary linear regression is not possible. One can observe that the number of births during 1-year period includes highest number of zeros than other two counts. The number of births during 5-year period more tends to follow a Poisson distribution without access of zeros.

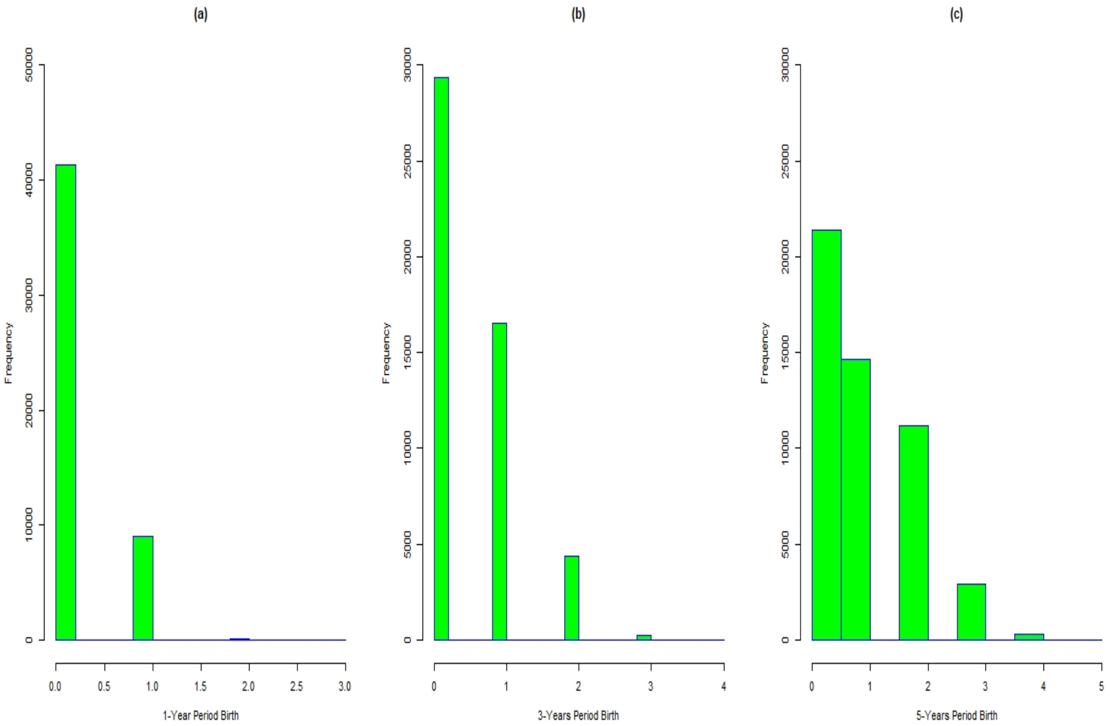

Figure 7.1: Histograms for number of births during different periods

Figure 7.2 provides scatter plot reflecting the relationship between age of women and number of birth during 1, 3 and 5-years periods prior to interview on Panels (a), (b) and (c) respectively. The X-axis consists of responses and Y-axis corresponding frequencies. This relation is also checked for urban and rural areas with different representations (defined on most top-right corners of each diagram). One can infer the behavior of the plot of births versus age is almost same for urban and rural areas.

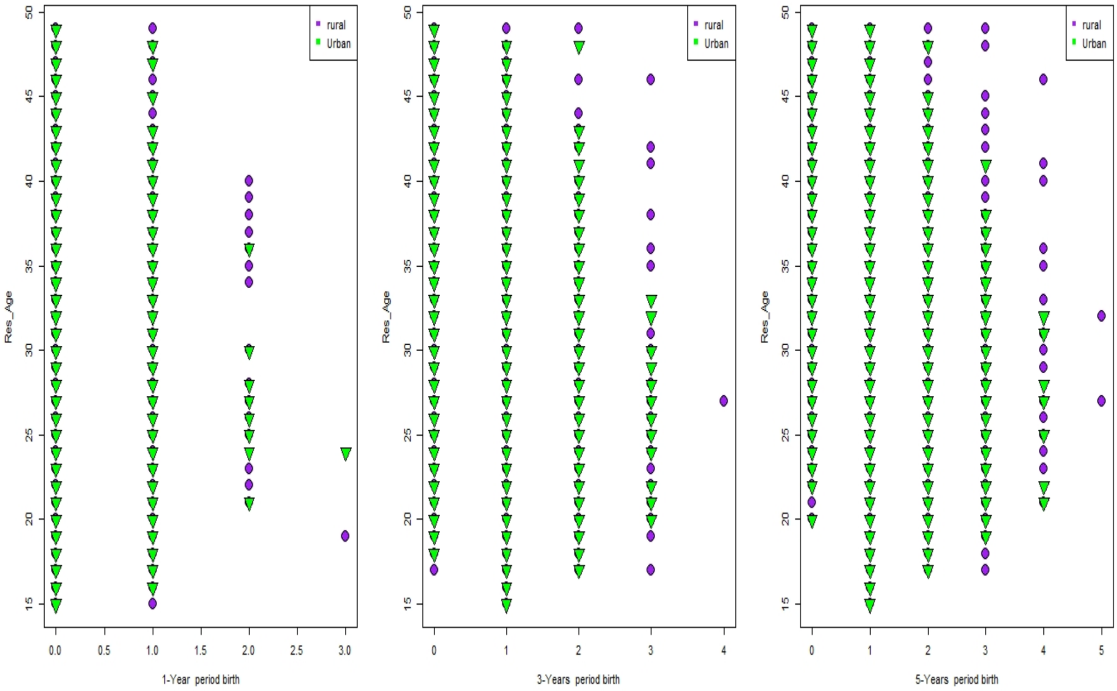

Figure 7.2: Relationship between Age and Births

Ordinary linear regression models usually use ordinary least squares (OLS) technique for the purpose of parameter estimation. For data with count responses, the regression model utilizes maximum likelihood method for estimation of the parameters. It seeks for the values of the regression coefficients with the highest probability (i.e. maximum likelihood) of observing the data at hand. Beaujean and Morgan (2016) provided a particularly understandable introduction of maximum likelihood estimation. The estimates of all coefficients are obtained by using an iterative set of procedures for reaching to the parameter estimates. All the maximum likelihood estimation results are converged and found a unique set of values for each coefficient (parameter). After going through literature and getting evidence from histogram we developed a Poisson regression model for three different periods and results are reported in 3rd column of Tables 7.3. and 7.4..

The results provide sufficient evidence that the estimated coefficients for all variable except of R_Sindh, Prof_tech and Prof_Agr have significant effect on number of 1-year births. The variable Res_Age, Residence_new and Age_husbund shows negative estimated coefficients for Poisson model which supports the argument that the number of births during 1-year period to a woman decreases with the age of couple and also higher for rural areas. The number of births for urban area are $\exp(-0.1021) = 0.903$ times the number of birth in rural areas keeping other factors constant.

Table 7.1.: Variable Description

Variables ID	Variable Name	Variable Description
v020	EMS	Dummy variable: 0=Never married 1=Ever married sample
v001	CL_num	Cluster number (1:528)
v002	HH_num	Household number (1:28)
v003	RL_num	Respondent's line number
v012	Res_Age	Respondent's age at he time of interview 15:49
v024	Region	Region: 1=Punjab, 2=Sindh, 3=KPK, 4=Balochistan, 5=GB, 6=ICT, 7=AJK, 8=FATA
	R_Sindh	Dummy variable: 0=Other Regions, 1=Sindh
	R_KPK	Dummy variable: 0=Other Regions, 1=KPK
	R_Blch	Dummy variable: 0=Other Regions, 1=Balochistan
	R_ICT	Dummy variable: 0=Other Regions, 1=ICT
	R_FATA	Dummy variable: 0=Other Regions, 1=FATA
v025	Residence _new	Dummy variable: 0=Rural, 1=Urban
v106	Edu	Highest educational: 0=No education, 1=Primary, 2=Secondary, 3=Higher
v191	WI	Dummy variable: 0=-ve Wealth Index, 1=+ve Wealth Index i.e. 1 v191>0,0 otherwise
v201	Num_mem	Total children ever born 0:20
v203	Num_daughter	Number of Daughters at home 0:20
v221	Mar_First_int	Marriage to first birth interval (months) 0:350 0=when v221=996 and Negative interval (i.e. birth before marriage)
v312	contraceptive _method	Current contraceptive method 0=Not using, 1=Pill, 2=IUD, 3=Injections, 4=Diaphragm, 5=Male condom, 6=Female sterilization, 7=Male sterilization, 8=Periodic abstinence, 9=Withdrawal, 10=Other traditional, 11=Implants/Norplant, 12=Prolonged abstinence, 13=Lactational amenorrhea (LAM), 14=Female condom, 15=Foam or jelly, 16=Emergency contraception, 17=Other modern method, 18=Standard days method (SDM), 19=Specific method 1, 20=Specific method 2, (m) 99=Missing
	Cont_Pill	Dummy Variable: 1=Pill, 0=otherwise
	Cont_IUD	Dummy Variable: 1=IUD, 0=otherwise
	Cont_Inj	Dummy Variable: 1=Injections, 0=otherwise
	Cont_Female	Dummy Variable: 1=Female sterilization, 0=otherwise
	Cont_Male	Dummy Variable: 1=Male sterilization and Male condom, 0=otherwise
	Cont_with _draw	Dummy Variable: 1=Withdrawal, 0=otherwise
	Cont_other	Dummy Corresponding to remaining categories Leaving "Not Using" as base category
v239	Preg_term Content [2]	Pregnancies terminated before calendar beginning 0=No 1=Yes (m) 9=Missing

Table 7.2.: Variable Description (Continued)

Variables ID	Variable Name	Variable Description
v717	Prof_res	Respondent's Occupation (grouped): 0=Not working, 1=Professional/technical/managerial, 2=Clerical, 3=Sales, 4=Agricultural - self employed, 5=Agricultural - employee, 6=Household and domestic, 7=Services, 8=Skilled manual, 9=Unskilled manual, 98=Don't know (m) 99=Missing
	Prof_tech	Dummy Variable: 1=Professional/technical/managerial or Clerical, 0=otherwise
	Prof_Agr	Dummy Variable: 1=Agricultural - self employed, 0=otherwise
	Prof_Other	Dummy corresponding to remaining categories taking "Not working as base"
v730	Age_husbnd [3]	Husband/partner's age 15:94
v209	Birth_1y	Births in past year (1:4 with 0 for no birth)
v238	Birth_3y	Births in last three years (1:4 with 0 for no birth)
v208	Birth_5y	Births in last five years (1:6 with 0 for no birth)

The R-out put returned estimates of parameters (Est), standard errors (Std), and significance indication. The likelihood value, or its transformations (like AIC or BIC) are then used for comparison of the fitting power of competing models Cameron and Trivedi (2013). The variable on region of the respondent is reconstructed into 5 dummies leaving Punjab as the base category. For interpreting the dummies, we follow recommendations given in Atkins and Gallop (2007). The dummy variable corresponding to the province Khyber Pakhtunkhwa (KPK coded as R_KPK) has negative estimated coefficient (-0.0926) with a standard error of 0.035 indicating lower birth exposure in KPK (R_KPK=1) as compared to Punjab (R_KPK=0) setting all other regressors to zero. While this coefficient is -0.088 and -0.05 with standard errors 0.021 and 0.015 for 3 and 5-years periods respectively for Poisson model. Similarly, one can distill from Tables 7.3., 7.5. and 7.7. that after adjusting for other variables the average numbers of births during 1, 3 and 5-year period are respectively $\exp(-0.0926) = 0.9115$, $\exp(-0.088) = 0.92$ and exp(-0.05) = 0.96 times of the birth in Punjab. The 95 % confidence interval for the true effects are $(-0.0926 \pm 1.96 \times 0.035$, $(-0.088 \pm 1.96 \times 0.021$ and $(-0.05 \pm 1.96 \times 0.015$, i.e.$(-0.1612, -0.024)$, $(-0.12916, -0.04684)$ and (-0.079, -0.021) . The 95% confidence interval for the relative rates (exponentiated estimates) are $(0.851, 0.976)$,

[2] The values of Preg_term are reported after imputing 0 missing cases within 1:13123 and 1 in remaining 37373 cases. Note that this division is made by dividing the data in ratio $4676 \times (50495)/(13317 + 4676)$ and $13317 \times (50495)/(13317 + 4676)$

[3] Missing entries Age_husbnd is imputed by median age of the observed responses

$(0.879, 0.954)$ and $(0.924, 0.98)$.

The regression coefficient associated with wealth index (WI) is approximately -0.13 for all models. As WI is dummy coded, the negative sign shows that the average number of births for those who have positive WI is smaller than for those who have negative WI. Since exp(-0.13)=0.88, i.e. the number of births during 1,3 and 5-years periods for those who have positive WI is 88% of those who have negative WI. Similarly, for all dependent variables, the coefficients for all dummies corresponding to different groups of the contraceptive methods turn negative values denoting the number of births for those who use any one of the contraceptive method has smaller number of birth than those who don't use any of the contraceptive method at all. For example, the average number of births to the women who use Cont_Pill is $\exp(-0.358) = 0.70$ of the births to the women in base category i.e. that the number of births during the 1-year period for those women using Contraceptive pill is 70% of the number of births to the women who don't use any method at all. Further, the ratio of births during 1-year period between those who use male contraceptive and those who use female contraceptive is 0.2383 with $\exp(-0.102 + 0.77885) = 1.97$ showing that the number of births during 1-year period for those who use male contraceptive method have more birth than those who use female contraceptive method. Same interpretation for the coefficients corresponding to contraceptive methods can be done with slight change in the estimated values and standard errors. Further, the average number of 1, 3 and 5-year births (exponentiated coefficients) to the women belonging to agriculture background are respectively 1.07 , 1.06 and 1.123 times higher than those who don't work.

The regression coefficient for dummy corresponding to pregnancy termination (Preg_term_new) is insignificant, hence no interpretation is made. The (1-α)% confidence interval corresponding to each exponentiated coefficient can be constructed after obtaining confidence interval for the coefficients with given standard error in Tables 7.3.-7.8.. The method introduced in Atkins and Gallop (2007) of interpreting dummy variables can't be extended for interpretation of the coefficients corresponding to continuous predictors. For the education level variable, the regression coefficient is 0.095. To see one level change, we put $\triangle = 1$ into the formula of percentage i.e. $100 \times [0.095 \times 1 - 1] = 10$ indicating that there is a 10% (7.6%, for 3-years births 5.6% for 5-years) increase in the expected number of 1-year births for a unit increase in education level (see Atkins and Gallop, 2007). For the number children already living the same household, the regression coefficient is 0.025 with

$100 \times [\exp(0.025 \times 1) - 1] = 25.3$ showing that there is a 25.3% increase in the expected number of 1-year birth for each additional family child. The percentage change in births during 3-year and 5-years periods prior to interview have not been reported here for the sake of space. Similarly, the percentage change in the expected births during 1-year period with change in one unit in the variable Res_age, Age_husbnd, Num_daughter, Mar_First_int are 9.7%, 1.64%, 2.5%, 26.905% and 0.5%. The percentage change in the expected births during 3-years period with change in one unit in the variable Res_age, Age_husbnd, Num_daughter, Mar_First_int are 8.9%, 1.41% 2.14%, 23% and 0.5% respectively. The percentage change in the expected births during 5-years period with change in one unit in the variable Res_age, Age_husbnd, Num_daughter and Mar_First_int are 7.9% , 1.14%, 1.78% and 22.4% respectively.

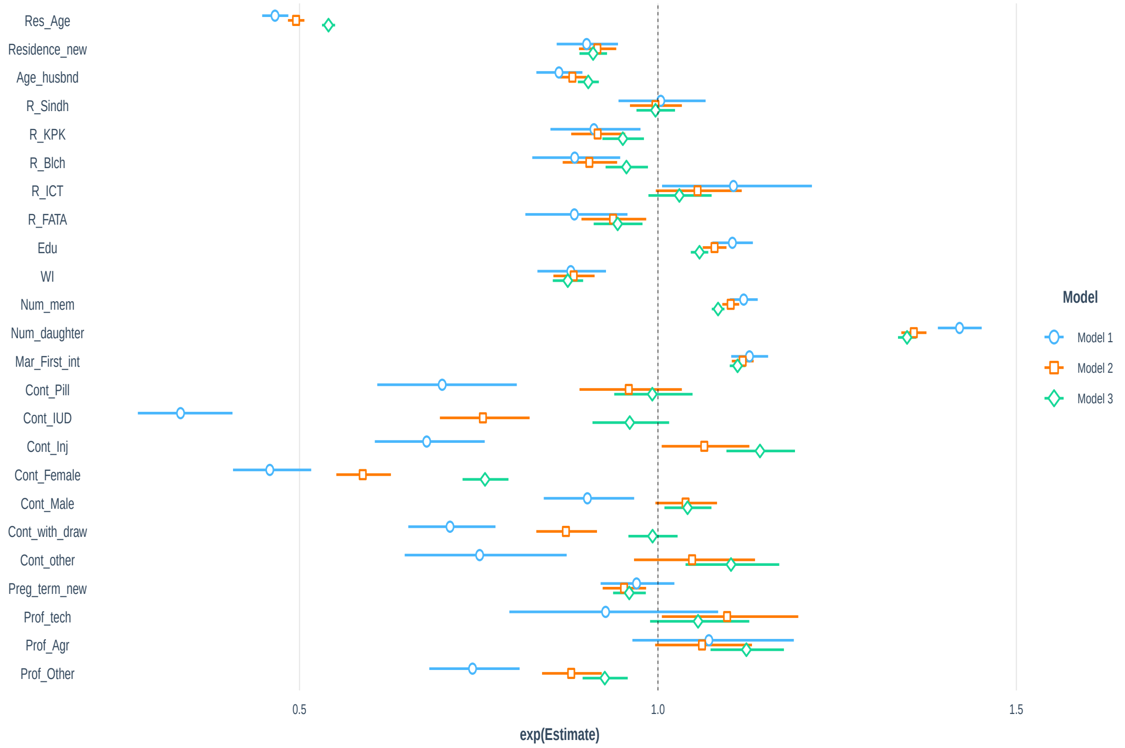

Figure 7.3: Display of exponentiated coefficients for Poisson model

Figure 7.3 displays a picture of the effect of factors, used in this study, on births occurred during 1, 3 and 5-years periods. The bar centers are shown as circles, rectangle and rhombus for representing models with births during 1, 3 and 5-years periods respectively. The length of bars show the variation in estimates while the centers show the exponentiated average change in births. The vertical line on center divides the variables with negative and positive coefficients. The variables with insignificant effects are very close to the vertical line. The bars corresponding to 1-year births for Con_IUD and Con_Female show a highly significant decrement in birth while using these contraceptive methods.

Table 7.3.: Regression model for number of births during 1-year

		Poisson	NB	ZIP-Poisson	ZIP-Inflation
(Intercept)	Est	1.68026	1.70258	0.96482	-18.47578
	Std	0.05967	0.06106	0.06771	1.25260
	Sig	***	***	***	***
Res_Age	Est	-0.10208	-0.10253	-0.07634	0.46323
	Std	0.00267	0.00272	0.00299	0.02889
	Sig	***	***	***	***
Residence_new	Est	-0.10463	-0.10624	-0.06093	0.81639
	Std	0.02421	0.02470	0.02555	0.22871
	Sig	***	***	*	***
Age_husbnd	Est	-0.01655	-0.01682	-0.01034	0.13715
	Std	0.00212	0.00216	0.00222	0.01671
	Sig	***	***	***	***
R_Sindh	Est	0.00368	0.00292	-0.04083	-0.61598
	Std	0.03086	0.03151	0.03289	0.28808
	Sig				*
R_KPK	Est	-0.09266	-0.09318	-0.04247	1.24354
	Std	0.03516	0.03585	0.03735	0.28222
	Sig	**	**		***
R_Blch	Est	-0.12340	-0.12235	-0.22139	-3.97459
	Std	0.03538	0.03608	0.03633	0.63189
	Sig	***	***	***	***
R_ICT	Est	0.10053	0.10079	0.13389	0.32867
	Std	0.04818	0.04915	0.05374	0.38977
	Sig	*	*	*	
R_FATA	Est	-0.12388	-0.12399	-0.13611	-0.40298
	Std	0.04110	0.04194	0.04245	0.48062
	Sig	**	**	**	
Edu	Est	0.09503	0.09517	0.08517	0.01055
	Std	0.01258	0.01285	0.01336	0.10755
	Sig	***	***	***	
WI	Est	-0.12886	-0.13133	-0.06595	0.91400
	Std	0.02777	0.02833	0.02998	0.27589
	Sig	***	***	*	***
Num_mem	Est	0.02464	0.02504	0.01403	-1.12920
	Std	0.00194	0.00199	0.00209	0.09376
	Sig	***	***	***	***
Num_daughter	Est	0.23826	0.23924	0.18919	0.14621
	Std	0.00745	0.00761	0.00798	0.08146
	Sig	***	***	***	.
Mar_First_int	Est	0.00495	0.00497	0.00402	-0.01250
	Std	0.00048	0.00049	0.00050	0.00330
	Sig	***	***	***	***

Table 7.4.: Regression model for number of births during 1-years (Continued)

		Poisson	NB	ZIP-Poisson	ZIP-Inflation
Cont_Pill	Est	-0.35825	-0.36173	-0.36029	1.33779
	Std	0.07083	0.07189	0.07396	0.52684
	Sig	***	***	***	*
Cont_IUD	Est	-1.09660	-1.10041	-1.02717	2.15735
	Std	0.10017	0.10099	0.11326	0.75784
	Sig	***	***	***	**
Cont_Inj	Est	-0.38966	-0.39511	-0.41490	0.84010
	Std	0.05762	0.05860	0.06078	0.58848
	Sig	***	***	***	
Cont_Female	Est	-0.77885	-0.77895	-0.78210	0.27866
	Std	0.06053	0.06100	0.06723	0.35717
	Sig	***	***	***	
Cont_Male	Est	-0.10172	-0.10503	-0.10136	0.90520
	Std	0.03565	0.03641	0.03961	0.35397
	Sig	**	**	*	*
Cont_with_draw	Est	-0.34289	-0.34544	-0.28322	1.70085
	Std	0.04365	0.04432	0.04842	0.32949
	Sig	***	***	***	***
Cont_other	Est	-0.28613	-0.29183	-0.12993	3.33044
	Std	0.07645	0.07784	0.08551	0.49409
	Sig	***	***		***
Preg_term_new	Est	-0.03035	-0.03125	0.00400	0.40885
	Std	0.02704	0.02760	0.02835	0.23123
	Sig				.
Prof_tech	Est	-0.07622	-0.07565	-0.19775	-2.82322
	Std	0.07984	0.08113	0.08993	1.03821
	Sig			*	**
Prof_Agr	Est	0.06870	0.06478	0.11203	0.94149
	Std	0.05354	0.05474	0.05868	0.51905
	Sig			.	.
Prof_Other	Est	-0.29959	-0.30308	-0.14562	2.28478
	Std	0.04332	0.04403	0.04737	0.32065
	Sig	***	***	**	***

Table 7.5.: Regression model for number of births during 3-years

		Poisson	NB	ZIP-Poisson	ZIP-Inflation
(Intercept)	Est	2.3982	2.4188	1.7563	-17.2259
	Std	0.0360	0.0368	0.0417	0.6531
	Sig	***	***	***	***
Res_Age	Est	-0.0939	-0.0945	-0.0674	0.4514
	Std	0.0016	0.0016	0.0018	0.0170
	Sig	***	***	***	***
Residence_new	Est	-0.0877	-0.0889	-0.0499	0.6580
	Std	0.0145	0.0148	0.0153	0.1418
	Sig	***	***	**	***
Age_husbnd	Est	-0.0142	-0.0143	-0.0112	0.0652
	Std	0.0013	0.0013	0.0013	0.0092
	Sig	***	***	***	***
R_Sindh	Est	-0.0039	-0.0034	-0.0352	-0.3242
	Std	0.0185	0.0189	0.0196	0.1608
	Sig			.	*
R_KPK	Est	-0.0879	-0.0886	-0.0804	0.3360
	Std	0.0210	0.0214	0.0223	0.1834
	Sig	***	***	***	.
R_Blch	Est	-0.1010	-0.0985	-0.1754	-1.7996
	Std	0.0214	0.0218	0.0224	0.3302
	Sig	***	***	***	***
R_ICT	Est	0.0519	0.0517	0.0395	-0.3946
	Std	0.0290	0.0296	0.0313	0.2046
	Sig	.	.		.
R_FATA	Est	-0.0649	-0.0640	-0.0646	0.2322
	Std	0.0246	0.0251	0.0255	0.2480
	Sig	**	*	*	
Edu	Est	0.0731	0.0737	0.0547	-0.1968
	Std	0.0076	0.0077	0.0079	0.0627
	Sig	***	***	***	**
WI	Est	-0.1249	-0.1279	-0.0374	1.5051
	Std	0.0166	0.0169	0.0178	0.1555
	Sig	***	***	*	***
Num_mem	Est	0.0211	0.0215	0.0127	-0.7039
	Std	0.0012	0.0012	0.0013	0.0423
	Sig	***	***	***	***
Num_daughter	Est	0.2070	0.2079	0.1567	-0.1655
	Std	0.0045	0.0046	0.0050	0.0504
	Sig	***	***	***	**
Mar_First_int	Est	0.0046	0.0046	0.0033	-0.0166
	Std	0.0003	0.0003	0.0003	0.0023
	Sig	***	***	***	***

Table 7.6.: Regression model for number of births during 3-years (Continued)

		Poisson	NB	ZIP-Poisson	ZIP-Inflation
Cont_Pill	Est	-0.0417	-0.0419	-0.0580	1.0315
	Std	0.0379	0.0386	0.0399	0.2950
	Sig				***
Cont_IUD	Est	-0.2801	-0.2819	-0.3399	-0.2630
	Std	0.0422	0.0428	0.0454	0.4289
	Sig	***	***	***	
Cont_Inj	Est	0.0624	0.0635	-0.0163	-1.7268
	Std	0.0293	0.0299	0.0304	0.6230
	Sig	*	*		**
Cont_Female	Est	-0.5299	-0.5290	-0.4572	0.8415
	Std	0.0330	0.0333	0.0381	0.1868
	Sig	***	***	***	***
Cont_Male	Est	0.0370	0.0368	0.0103	0.3884
	Std	0.0211	0.0216	0.0225	0.1881
	Sig	.	.		*
Cont_with_draw	Est	-0.1376	-0.1380	-0.1762	0.0787
	Std	0.0248	0.0252	0.0265	0.1945
	Sig	***	***	***	
Cont_other	Est	0.0465	0.0459	0.0623	1.1738
	Std	0.0411	0.0419	0.0438	0.2797
	Sig				***
Preg_term_new	Est	-0.0485	-0.0489	-0.0193	0.4537
	Std	0.0162	0.0165	0.0169	0.1465
	Sig	**	**		**
Prof_tech	Est	0.0923	0.0962	0.0179	-1.5429
	Std	0.0442	0.0449	0.0471	0.3359
	Sig	*	*		***
Prof_Agr	Est	0.0596	0.0587	0.1091	1.0635
	Std	0.0325	0.0331	0.0349	0.2628
	Sig	.	.	**	***
Prof_Other	Est	-0.1290	-0.1294	-0.0838	0.6009
	Std	0.0242	0.0246	0.0261	0.1837
	Sig	***	***	**	**

Table 7.7.: Regression model for number of births during 5-years

		Poisson	NB	ZIP-Poisson	ZIP-Inflation
(Intercept)	Est	2.5448	2.5637	2.0000	-16.3950
	Std	0.0269	0.0275	0.0303	0.4825
	Sig	***	***	***	***
Res_Age	Est	-0.0822	-0.0828	-0.0585	0.4309
	Std	0.0012	0.0012	0.0013	0.0129
	Sig	***	***	***	***
Residence_new	Est	-0.0947	-0.0959	-0.0505	0.9489
	Std	0.0108	0.0110	0.0113	0.1055
	Sig	***	***	***	***
Age_husbnd	Est	-0.0114	-0.0115	-0.0089	0.0607
	Std	0.0009	0.0009	0.0010	0.0071
	Sig	***	***	***	***
R_Sindh	Est	-0.0035	-0.0031	-0.0294	-0.1791
	Std	0.0138	0.0140	0.0145	0.1161
	Sig			*	
R_KPK	Est	-0.0502	-0.0505	-0.0598	-0.0199
	Std	0.0156	0.0158	0.0162	0.1294
	Sig	**	**	***	
R_Blch	Est	-0.0450	-0.0427	-0.1130	-1.2969
	Std	0.0158	0.0161	0.0164	0.2068
	Sig	**	**	***	***
R_ICT	Est	0.0293	0.0294	0.0319	-0.0638
	Std	0.0219	0.0222	0.0233	0.1503
	Sig				
R_FATA	Est	-0.0579	-0.0565	-0.0864	-1.0513
	Std	0.0184	0.0188	0.0190	0.2640
	Sig	**	**	***	***
Edu	Est	0.0542	0.0548	0.0345	-0.3233
	Std	0.0057	0.0058	0.0059	0.0489
	Sig	***	***	***	***
WI	Est	-0.1343	-0.1368	-0.0707	1.0432
	Std	0.0124	0.0126	0.0130	0.1105
	Sig	***	***	***	***
Num_mem	Est	0.0176	0.0179	0.0105	-0.6432
	Std	0.0009	0.0009	0.0010	0.0340
	Sig	***	***	***	***
Num_daughter	Est	0.2022	0.2032	0.1525	-0.2735
	Std	0.0033	0.0033	0.0036	0.0372
	Sig	***	***	***	***
Mar_First_int	Est	0.0043	0.0044	0.0030	-0.0182
	Std	0.0002	0.0002	0.0002	0.0016
	Sig	***	***	***	***

Table 7.8.: Regression model for number of births during 5-years (Continued)

		Poisson	NB	ZIP-Poisson	ZIP-Inflation
Cont_Pill	Est	-0.0079	-0.0090	-0.0206	1.2046
	Std	0.0281	0.0286	0.0293	0.2220
	Sig				***
Cont_IUD	Est	-0.0401	-0.0417	-0.1076	-0.5326
	Std	0.0284	0.0289	0.0298	0.2853
	Sig			***	.
Cont_Inj	Est	0.1331	0.1346	0.0423	-4.6319
	Std	0.0213	0.0218	0.0216	0.8425
	Sig	***	***	.	***
Cont_Female	Est	-0.2751	-0.2747	-0.1903	0.9573
	Std	0.0215	0.0218	0.0243	0.1215
	Sig	***	***	***	***
Cont_Male	Est	0.0401	0.0409	-0.0090	-0.1549
	Std	0.0161	0.0164	0.0168	0.1532
	Sig	*	*		
Cont_with_draw	Est	-0.0075	-0.0066	-0.0850	-0.8900
	Std	0.0176	0.0179	0.0187	0.1704
	Sig			***	***
Cont_other	Est	0.0971	0.0979	0.0746	0.5380
	Std	0.0302	0.0308	0.0316	0.2223
	Sig	**	**	*	*
Preg_term_new	Est	-0.0409	-0.0416	-0.0265	0.1889
	Std	0.0121	0.0124	0.0125	0.1082
	Sig	***	***	*	.
Prof_tech	Est	0.0545	0.0570	-0.0239	-1.6005
	Std	0.0334	0.0339	0.0352	0.2942
	Sig		.		***
Prof_Agr	Est	0.1163	0.1160	0.1614	0.9730
	Std	0.0232	0.0237	0.0246	0.1795
	Sig	***	***	***	***
Prof_Other	Est	-0.0771	-0.0772	-0.0381	0.4939
	Std	0.0173	0.0176	0.0185	0.1270
	Sig	***	***	*	***

Note: In Tables 7.3.–7.8., Est stands for estimated coefficients, Std for standard error and Sig for significance, where "***" , "**", "*", "." significance at 0.1%, 1%, 5% and 10% respectively.

Table 7.9.: Comparison of Models

	Poiss	NB	ZIP
		1-year	
Min.	-1	-1	-1.0731
Mean	-0.122	-0.121	0.0000
Max.	51.125	51.793	2.5155
2logLikelihood	-42984.63	-43307.25	-42011.64
AIC	43035	43359	42036.64
		3-years	
Min.	-1	-1	-1.7707
Mean	-0.0978	-0.0969	0.0001
Max.	14.8876	15.0487	3.0787
2logLikelihood	-77067.93	-77564.3600	-75013.42
AIC	77118	77616	75038.42
		5-years	
Min.	-1	-1	-2.6435
Mean	-0.0754	-0.0747	0.0001
Max.	9.4748	9.5761	4.1984
2logLikelihood	-106392.9	-106877.2	-102943.6
AIC	106443	106929	102964.4016

A comparison of three competing models Poisson, NB and ZIP are provided in Table 7.9. for three different responses. The minimum, maximum and mean residuals for each model are reported. The minimum residual is observed at extreme (i.e. -1.0731, -1.7707 and -2.6435) for ZIP model. While maximum residual is observed at extreme (i.e. 51.793, 15.0487 and 9.5761) for NB. Further to see the model performance -2logliklihood and AIC values are also reported. The AIC values are observed smaller for ZIP as compared to other models considered in this study.

7.6 Model Estimation Under Bayesian Framework

In Bayesian inference, one can directly obtain the probability of values of the parameters by finding the area of the posterior distribution to a region on the right of that value, which is equal to the proportion of the values of the parameter in the posterior sample which are larger than that value. We can utilize that information to file the results of Bayesian statistical analysis as means with, so-called, Bayesian credible intervals for estimated parameters. In this way, recently, Bayesian method of estimation and inference have been extensively used.

Table 7.10.: Posterior mean, standard deviation and effective size for regression coefficients

Variables	1-year			3-years			5-years		
	Mean	STD	ESS	Mean	STD	ESS	Mean	STD	ESS
(Intercept)	1.6775	0.0611	38.7081	1.677	0.0611	50.5251	2.5457	0.024	51.1951
Res_Age	-0.1017	0.0021	36.3485	-0.1017	0.0021	24.1892	-0.0819	0.0008	41.7633
Residence _new	-0.1048	0.0228	1534.81	-0.1048	0.0228	571.98	-0.0942	0.0103	639.54
Age_husbnd	-0.0168	0.0016	32.75	-0.0168	0.0016	27.51	-0.0116	0.0006	53.72
R_Sindh	0.0037	0.0301	696.8	0.0037	0.0301	805.95	-0.0037	0.0138	634.94
R_KPK	-0.0934	0.0355	488.83	-0.0934	0.0355	501.22	-0.0508	0.0159	437.95
R_Blch	-0.1249	0.0358	653.91	-0.1249	0.0358	615.68	-0.0453	0.0161	518.26
R_ICT	0.0998	0.0476	1390.18	0.0998	0.0476	1074.85	0.029	0.0217	1318.52
R_FATA	-0.1242	0.0409	708.39	-0.1242	0.0409	792.49	-0.0582	0.0186	702.03
Edu	0.0942	0.0126	485.61	0.0942	0.0126	529.95	0.0544	0.0058	486.34
WI	-0.129	0.0266	470.94	-0.129	0.0266	449.55	-0.1351	0.0122	512
Num_mem	0.0248	0.002	279.92	0.0248	0.002	318.58	0.0176	0.0009	232.89
Num _daughter	0.2379	0.0081	288.75	0.2379	0.0081	318.44	0.2021	0.0032	333.04
Mar_First	0.0049	0.0005	639.47	0.0049	0.0005	464.98	0.0043	0.0002	559.27
Cont_Pill	-0.3625	0.0715	1505.22	-0.3625	0.0715	1603.4	-0.0081	0.0286	1524.9
Cont_IUD	-1.1026	0.1011	1855.87	-1.1026	0.1011	1382.57	-0.041	0.0278	1611.72
Cont_Inj	-0.3891	0.0581	1673.25	-0.3891	0.0581	1374.7	0.133	0.0215	1926.15
Cont_Female	-0.7811	0.0619	1465.15	-0.7811	0.0619	1665.05	-0.2768	0.0219	1393.43
Cont_Male	-0.104	0.0353	1280.29	-0.104	0.0353	1327.15	0.0409	0.016	1231.12
Cont_with _draw	-0.3441	0.0437	1499.29	-0.3441	0.0437	1252.7	-0.0081	0.0178	1417.29
Cont_Other	-0.2888	0.0767	685.28	-0.2888	0.0767	1626.67	0.0964	0.0309	1388.1
Preg_term	-0.0304	0.0281	248.15	-0.0304	0.0281	278.01	-0.0412	0.0127	228.99
Prof_tech	-0.0785	0.0785	1641.41	-0.0785	0.0785	1560.07	0.0538	0.0331	1374.54
Prof_Agr	0.0689	0.0537	1394.11	0.0689	0.0537	1204.17	0.1156	0.0232	1375.04
Prof_Other	-0.3014	0.0422	1585.41	-0.3014	0.0422	1434.15	-0.0773	0.0172	1398.44

Table 7.11.: Model Comparison

	1-years	3-years	5-years
RES-mean	0.00061	0.33003	0.00107
Res-SD	0.13137	0.34963	0.58909
Mean Deviance	43015	77117	106448
penalty	24.08	25.26	25.56
Penalized deviance	43039	77142	106474

One of the areas to focus in applied Bayesian inference is Bayesian regression models. The most important aspect of the Bayesian learning process is explaining a relationship and generalizing it to others, and this section is our attempt to use the Bayesian Linear Regression (BLR) for predicting the outcome for non-sampled set. What we result from the frequentist linear regression is an estimate of the model parameters from only the training data set (the sampled data set in our problem). Our model is informed completely by the sample: in this way, everything that we need to recognize our model is available in the sample data. However, if the sample size is small, one might like to express the estimate as a distribution of possible values of the parameter given the sample information. This is the situation where BLR is needed. As we already discussed when the response variable (consequently residuals) don't follow a Gaussian distribution it is not possible to continue with ordinary linear regression. The argument for using Poison model for the data with count responses has already discussed in previous sections.

In Bayesian model fitting involving Markov Chain Monte Carlo MCMC, the researcher must be worried about the programming errors and the problems occur in estimation routines Hamra et al. (2013). The trade-off to this extra task is that there is huge flexibility in model construction, statistical inference, and assessment of model fit than a frequentist. Aside from these basic programming errors that can make an MCMC algorithm inadequate, there are two main concerns with the employment of the MCMC algorithms: mixing and convergence. Researchers must confirm that the algorithm results a Markov chain that "converges" to the appropriate posterior density and "mixes" well throughout the values of the density Lynch (2007). In this section, we analyze the birth history data using Poisson regression model, previously run under frequentist's point of view in a Bayesian framework, by adding a normal prior on the coefficients of the linear log-mean function as given in (7.4.2) i.e. $\beta_j \sim N(0, 1 \times 10^{-3})$ for all $j = 0, 1, 2, ..., 24$. The analysis is done in R-library (rjags) Just

Another Gibbs Sampler (JAGS) taking three chains. We initialize the model, run the burn in period and the model is updated 100 times. The number of iterations is taken 1000. we take relatively fewer iterations as the number of nodes are 50495 which leads to take a longer duration in running the complete model . After deciding the burn-in period, we run the simulation for the samples that will keep. We re-run the model for checking the convergence and to see the auto-correlation. The deviance information criterion is also obtained for each model.

First we look at the interpretation of posterior means of the coefficients corresponding to each model. The 2nd, 5th and and 8th columns of Table 7.10. give the posterior means of each coefficients, 3rd, 6th and 9th columns provide their standard deviation and 4th, 7th and 10th columns give effective sample sizes (ESS) for models with 1, 3 and 5-years number of births as the response variables. One can notice that the posterior means almost match with the estimated coefficients corresponding to each variable with a slight reduction in standard error. The standard error for the residence (Residence_new) variable with Poisson model with 1-year period births as the response under maximum likelihood estimation is 0.02421 while the corresponding posterior standard deviation for the same coefficient under MCMC is 0.0228. After running MCMC, we need to check whether the estimated coefficients are valid or not. We need to confirm that whether the MCMC sampler covers the parameter space efficiently, i.e. it doesn't accept or reject too many proposals. A detail discussion of MCMC diagnostics can be found in Mengersen et al. (1999), Boone et al. (2014) and Vats et al. (2019) etc. If the MCMC rejects too many proposals, we need a large number of simulations to generate a considerable number of parameter samples. On the other end if a large number of proposals are accepted, we can't found much information about the parent distribution. Trace plots provide an important tool for assessing mixing of a chain. Density plots are smoothed histograms of the samples, i.e. they show the function that we are trying to explore. Figures E.6-E.12 (see Appendix E) show the trace plots corresponding to each coefficients which provide the evidence of presence of randomness (lack of pattern) in data. The trace plots corresponding to intercept β_0, and two coefficients β_1 and β_3 reflect slight lack of randomness while the trace plot corresponding to all other coefficients provide enough evidence of randomness. We provide the trace plot corresponding to the model for the births during 1-year period. The trace plot corresponding to other two models are not reported for the sake of space. Observing Figures E.6-E.12, one can also see the behavior of posterior densities for each coefficient. An alternative way to check for convergence of the estimates is to look at the auto-correlations among the samples obtained from MCMC. The lag-l auto-correlation is the correlation between every sample and the sample l steps before which become smaller as l increases, i.e. considering samples as independent. On the other hand, if this auto-correlation remains constant (high) for higher values of l too, then the situation depicts a higher correlation between every sample and the sample l steps before. The auto correlation plots corresponding to each coefficient for the three models are obtained from MCMC sampling. The plots for the model with 1-year period births as response variable is reported in Appendix.

After looking at the auto-correlation plots given in Figures E.1-E.5 (see Appendix E), we can notice that the auto-correlation goes down for all coefficients with increase in l (“lag", i.e. the x-axis in the plot) which is a good sign. The auto-correlation plot corresponding to intercept and the coefficient corresponding to age indicates presence of auto-correlation. This auto-correlation can be reduced by thinning the MCMC chains, i.e. we discard n samples for every sample that we keep. The thinning of MCMC chain is actually not much useful, unless we want to reduce the memory and storage space in long chains. With this argument one should keep only one out of ten samples instead of thinning the chain, because this is more efficient (with respect to the effective sample size) to run only one chain 10 times as long, it will take 10 times more storage space. A more reliable estimate for burn-in cut-off is through the effective sample size (ESS). An ESS is the number of independent samples with same estimation power as the number of autocorrelated samples. The burn-in contains samples that are not much informative, and if the period of burn-in is estimated to be short enough this will lead to reduction in the ESS. On contrary, if the period of burn-in is estimated to much longer, again cause in reduction of the ESS as informative samples are being isolated. An increase in ESS should be with the optimal estimate of the burn-in are highly recommended in practical estimation procedures. We can assess the burn-in samples by a glance at the trace plots and effective sample sizes. Table 7.10. show that the ESS for each coefficient is enough for all coefficients, except of β_0, β_1 and β_3 which are 38.7, 36.3 and 32.75 respectively, are large enough to continue. The ESS for the coefficient of Cont_Inj is maximum with 1855. Finally, we see the predictive power of our models by checking at deviances. The most widely used tool for checking the predictive power of a model is deviance information criterion (DIC). It is an estimate of the expected predictive error of the models. Table 7.11. gives the mean deviances for the three models. The DIC value is least (i.e. 43015) for the model with number of birth during 1-year period as the response. The penalized deviance for the same model with penalty of 24.08 is 43039 which is slightly larger than the deviance without penalty.

7.7 Model-Based Estimates of Fertility Rates

In PDHS final report PDHS (2019), ASFRs are obtained according to the formula given in Equation (1.2.1) for each age group. Before going into more detail about model-based construction of rates, we remind the data need for construction of rates from PDHS report. In computing ASFR, numerator is obtained by tabulating births according to period of birth (3-years period is taken here according to DHS standard) and the mother age at the time of the birth. The age of child is obtained as the difference between the date of interview and the date of birth, both are taken in century-month code (CMC) format. Counting the births occurred during 1-36 months before the survey (v008 - b3 in DHS format, see variable description in Table 7.1.). Age of mothers are computed, in CMC format, by taking the difference of the date of interview and the date of birth of the mother. Births are then tabulated by age group after converting the ages in years. Similarly, the denominator in ASFR are women-years of exposure, calculated as the sum of the counts of months exposed in the five-year age group during the 3-years time period divided by 12. A woman can expose in several age groups in the given period, with varying length of the period. For a 3-years period a woman will contribute to at most two five-year age groups during the 36-months period. For further details related to allocation of women to the higher age group and lower age group readers are referred to PDHS (2019).

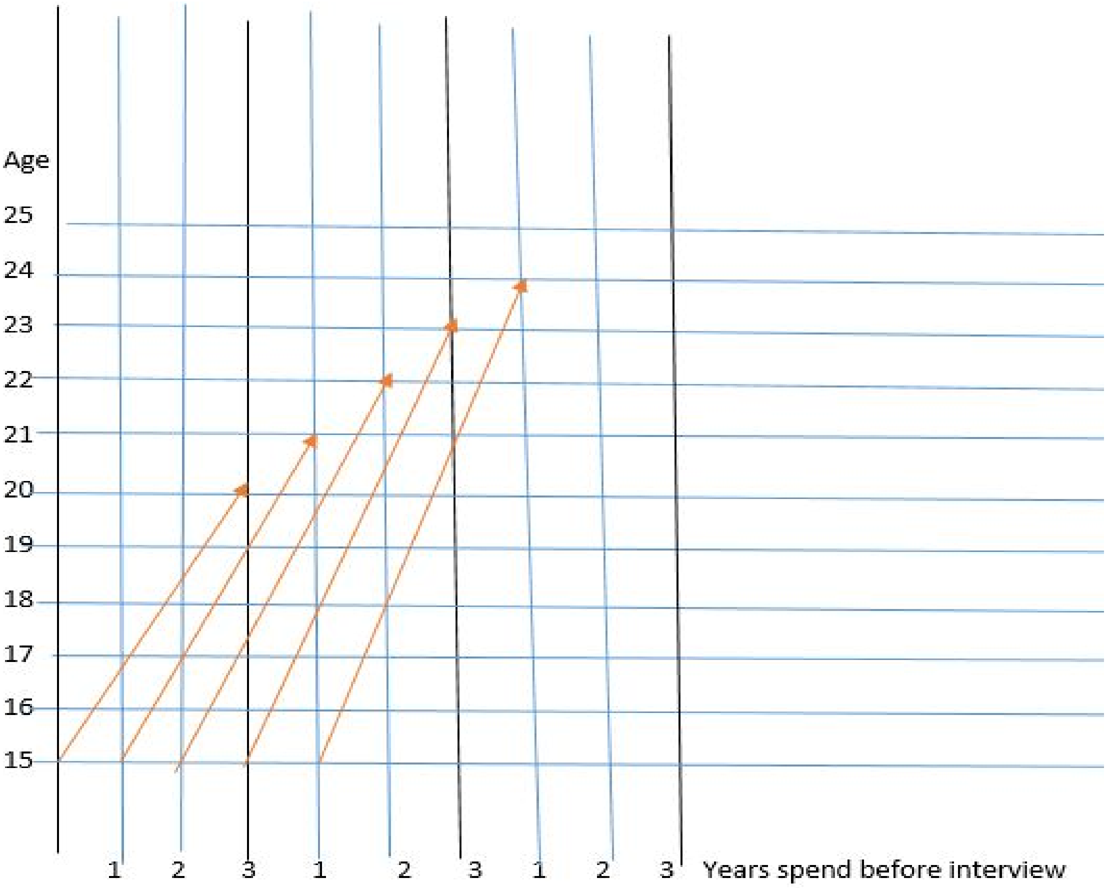

Figure 7.4: Illustration of birth history data on a Lexis diagram

Illustration of birth history data on a Lexis diagram (each birth can fall in either of the categories are given in Figure 7.4, then person-years are calculated according to these categories). Years spent by respondents before interview is shown on X-axis for periods of 3 years. While age of respondents are shown on Y-axis the first arrow begins from 15 age and end at 20 inferring that birth falling in this category, i.e. area between first and second arrows before the first vertical line corresponding to 3, are counted as exposure for group 15-19. Instead of using DHS.rates Package already given by IGME (2018), we developed R codes for estimating ASFR after converting the individual (IR) data to person-year data. The R-codes may be provided as supplementary material to this thesis. For tabulation of person-years data, each woman is tallied twice, once in the lower age group aggregating lower age group exposure and once according to the higher age group summing the exposure she contributes to the higher age group. In computing fertility rates, we use only ever-married samples without taking "all-women factor" (awfactt in DHS code) under model-based approach. Hence interpretation of the rates are done on the basis of birth per ever-married women only.

The total exposure in years in each age group is then the sum of the exposure in each age group tallying from the first and second. After obtaining ASFR it is straightforward to obtain TFR, GFR and GRR using formulae given in Equations (1.2.2), (1.2.3) and (1.2.4).

In this section, we obtain fertility measures i.e. ASFR, TFR, GFR and GRR using predicted response obtained from the regression models after partitioning data into sampled and non-sampled parts. A bootstrap sampling procedure to study the design-based properties of the estimated fertility rates PDHS 2017-18 women re-coded data is given as follow:

(i). Select a simple random sample of size 10,000 from PDHS 2017-18 women re-coded data and partition the data into sampled and non-sampled parts.

(ii). Fit the Poisson, NB, ZIP and ZINB models to the sampled data and obtain the coefficients and residuals.

(iii). Predict the non-sampled part of the response variable by using the fitted models from Step (ii).

(iv). Convert the individual re-code (PKIR71) data into person-year data and obtain fertility rates under the working models.

(v). Repeat Steps (i)–(iv), 10,000 times to obtain expected rates and the corresponding root mean squared errors (RMSEs). Root mean squared error for estimated rate say $\hat{R}$ against the true rate R is given by

$$RMSE(\hat{R}) = \sqrt{\frac{\sum_{sim}\left(\hat{R}-R\right)^2}{10000}}, \tag{7.7.1}$$

where $\hat{R}$ is the predictive estimate of fertility rate obtained from repeated samples and R is the corresponding rate obtained from full data without use of weights (weights corresponding to ever-married sample) the notation $\sum_{sim}$ denotes the summation is taken over all 10,000 repeated samples. Model-based ASFRs for Pakistan at national and sub-national level for ever-married sample with their RMSE are given in Table 7.12., and Tables 7.13.-7.15. respectively under four different working models. The expected ASFR with smallest RMSE among four (obtained under four alternative models) is bolded corresponding to each age group for sub-national and national level. In majority of cases ZIP model gives relatively smaller RMSE at sub-national and national levels. Comparing the estimated ASFR with the

ASFR obtained from full data in last column of Tables 7.12.-7.15., we can noticed that the ASFR for age groups 1, 2, 6 and 7 are upward biased for all models. The expected ASFR with smallest difference (bias) with the ASFR obtained from full data among four (obtained under four alternative models) is underlined corresponding to each age group for sub-national and national level. From ASFR based on full PDHS data, one can observe that for age group 15-19 that the highest birth is observed for FATA and lowest in KPK. Similarly, ASFR can be compared for different regions based on full data. While individual estimates are obtained under model-based approach using four alternative models.

Table 7.12.: Model-based ASFR for ever-married women in Pakistan

AGE		Poisson	NB	ZIP	ZINB	Full Data
		PAKISTAN				
1	MEAN	222.7225	223.7919	**203.6582**	202.6054	178.6971
	RMSE	4.8895	7.0937	3.1837	3.6582	
2	MEAN	275.0806	266.0772	258.5447	**258.7548**	271.7718
	RMSE	2.7795	4.1703	2.2707	2.1496	
3	MEAN	244.4803	244.8132	**250.2329**	247.9442	254.2912
	RMSE	2.5342	3.4166	2.0359	2.1655	
4	MEAN	191.6947	191.7994	**204.7832**	203.1567	202.2433
	RMSE	2.1752	2.8781	1.4647	2.2746	
5	MEAN	114.4192	114.3627	**131.7197**	130.7572	115.6562
	RMSE	2.5423	2.9110	2.2248	1.2844	
6	MEAN	54.6209	54.415	**72.1514**	71.4900	46.4946
	RMSE	2.1848	3.1478	2.9820	1.8025	
7	MEAN	**23.7134**	23.4652	39.6013	37.6739	14.7247
	RMSE	1.1865	1.2567	2.6903	1.9403	

Table 7.13.: Model-based ASFR for ever-married women by region

AGE		Poisson	NB	ZIP	ZINB	Full Data
			Punjab			
1	MEAN	211.3012	212.0825	**189.9279**	193.1613	249.1946
	RMSE	10.1419	9.9172	6.8784	10.1949	
2	MEAN	269.3565	269.8071	**263.3218**	263.9245	231.7774
	RMSE	5.0639	6.6696	5.3568	4.7155	
3	MEAN	250.4719	250.5010	**254.3172**	257.0261	198.0558
	RMSE	5.4656	5.1818	5.1457	5.1563	
4	MEAN	192.0939	191.9749	204.7935	**206.7177**	186.5237
	RMSE	4.2740	6.6887	5.5483	4.3205	
5	MEAN	101.8989	101.6711	**118.7083**	120.3670	127.9033
	RMSE	6.0881	4.7710	3.5397	5.6747	
6	MEAN	44.7504	**44.4477**	63.5791	63.1820	76.0920
	RMSE	5.6576	3.1458	4.6625	3.0235	
7	MEAN	18.9567	**18.6261**	34.1430	35.4604	39.3004
	RMSE	2.1963	1.8623	3.0579	3.7498	
			SINDH			
1	MEAN	229.4928	230.7073	212.4477	**210.0833**	228.2823
	RMSE	19.5323	13.7867	9.9588	7.7385	
2	MEAN	268.9236	260.0987	**250.9072**	253.5122	265.9237
	RMSE	11.2111	5.1890	3.0585	5.9418	
3	MEAN	239.7884	240.3403	**246.1280**	245.5123	230.0445
	RMSE	15.2525	5.8365	4.9307	6.4958	
4	MEAN	187.7445	187.9909	204.9746	**200.4824**	189.7307
	RMSE	4.7523	4.7700	6.7566	4.9784	
5	MEAN	109.7143	109.5498	**130.3497**	126.9330	111.7718
	RMSE	8.3903	7.6144	4.5765	6.2280	
6	MEAN	56.3930	**56.2313**	75.8477	72.9101	61.2671
	RMSE	9.1977	4.4619	5.1601	6.1154	
7	MEAN	24.6972	**24.4979**	40.3120	40.2554	27.8576
	RMSE	3.7597	2.2051	8.5818	3.4375	

Table 7.14.: Model-based ASFR for ever-married women by region (Continued)

AGE		Poisson	NB	ZIP	ZINB	Full Data
			KPK			
1	MEAN	196.3699	196.9613	182.8862	**180.3818**	207.4155
	RMSE	7.4761	8.9657	6.6572	5.8342	
2	MEAN	273.8359	**274.3997**	266.2204	266.6967	270.6913
	RMSE	5.6553	5.3221	7.8679	9.9190	
3	MEAN	263.2431	263.4433	**267.2585**	269.0296	246.0604
	RMSE	5.6614	8.4516	3.5384	9.7035	
4	MEAN	186.1139	186.2141	201.3701	**199.6779**	191.6476
	RMSE	5.4555	6.2886	5.1218	4.8980	
5	MEAN	115.2574	115.1831	**133.5651**	132.0458	101.7487
	RMSE	4.4904	4.9616	4.1421	4.2693	
6	MEAN	45.6254	**45.3674**	62.7137	63.4028	44.1182
	RMSE	5.6872	3.0984	4.3727	4.1459	
7	MEAN	22.0920	21.7557	36.7238	**37.9287**	19.6460
	RMSE	3.6424	5.8643	7.6976	3.5424	
			BALOCHISTAN			
1	Mean	250.9692	252.7714	**227.3190**	228.3367	252.4604
	RMSE	9.9384	10.2157	6.0973	8.9911	
2	Mean	239.4906	240.6887	**228.4109**	229.5668	248.7213
	RMSE	9.9295	6.7784	7.5310	13.6000	
3	Mean	213.5977	214.6399	**212.7518**	214.7551	225.0510
	RMSE	16.6663	9.2602	5.9430	7.4345	
4	Mean	188.6344	189.4245	**194.9166**	195.9774	188.5075
	RMSE	6.9316	10.7594	5.4797	7.5575	
5	MEAN	**126.1279**	126.4127	138.1701	138.6163	119.9730
	RMSE	5.8020	13.6353	6.1309	6.5612	
6	Mean	82.2158	**82.3038**	97.3578	95.8696	76.1724
	RMSE	8.3163	5.6318	8.5431	5.6471	
7	Mean	40.8662	**40.7404**	54.4509	54.7438	44.1215
	RMSE	5.3392	5.1060	5.7093	5.1905	

Table 7.15.: Model-based ASFR for ever-married women by region (Continued)

AGE		Poisson	NB	ZIP	ZINB	Full Data
				ICT		
1	MEAN	266.4663	267.6229	**238.3424**	240.7646	211.9564
	RMSE	18.8398	18.5211	4.2664	15.3548	
2	MEAN	273.4660	265.5485	**254.1284**	253.1415	258.9262
	RMSE	9.6579	7.6228	7.0958	13.6507	
3	MEAN	257.7272	258.4283	**262.2330**	256.3389	225.4344
	RMSE	7.6095	7.0926	4.9718	7.0901	
4	MEAN	236.7798	237.5580	**241.5292**	243.5958	186.5302
	RMSE	8.9159	10.3614	7.5267	8.8469	
5	MEAN	128.4957	128.5902	**139.6517**	144.3909	102.2706
	RMSE	14.1416	6.6153	4.0699	6.7723	
6	MEAN	66.6295	**66.5571**	84.6699	82.0437	48.0665
	RMSE	7.5488	6.6754	7.5503	9.2615	
7	MEAN	22.5053	**22.3162**	36.0057	34.6945	21.6777
	RMSE	7.3344	3.8912	9.5149	3.6871	
				FATA		
1	MEAN	231.6374	232.2293	**218.1838**	215.8759	257.9901
	RMSE	17.4629	21.7423	7.9326	13.8303	
2	MEAN	289.0122	282.2671	**285.6814**	278.5234	278.9904
	RMSE	9.0171	8.3427	5.5339	7.9979	
3	MEAN	243.3209	243.0650	**258.9386**	251.4702	254.7245
	RMSE	13.1805	9.7643	4.9383	12.3537	
4	MEAN	187.9857	187.6520	**202.9574**	199.8495	239.7390
	RMSE	8.1129	9.2729	0.9261	6.3040	
5	MEAN	87.3725	87.0357	**104.0907**	106.042	141.0778
	RMSE	5.2186	5.8885	1.8045	8.5333	
6	MEAN	48.7856	48.3721	**66.2820**	66.9248	71.1328
	RMSE	8.0002	4.5553	4.2805	5.8337	
7	MEAN	18.2021	**17.8044**	33.8488	35.2680	16.1374
	RMSE	4.6030	3.3961	4.6296	7.8904	

The TFR, GFR and GRR for ever-married women are provided at sub-national and nation levels in Tables 7.16., 7.17. and 7.18. respectively. The TFR for ever-married sample is observed highest in ICT with 6.26 and lowest in Punjab with 5.44 per 1000 ever-married women. While the TFR for all women given in PDHS 2017-18 report is observed highest for FATA. The TFR obtained for all women using DHS.rates package are given in Appendix E. The estimates obtained under NB, ZIP and ZINB are more accurate than the ones observed for Poisson which can be noticed from RMSE in Table 7.16. corresponding to each region. The smallest RMSE is observed at national level when ZIP is used to model births i.e. 0.0617. Similarly, predictive estimate for GFR at sub-national and national level are observed highest for ever-married women in KPK which is 174.95.6 (NB as it is more precise than other three estimates) births per 1000 married women. The lowest GFR is observed for Punjab with 162 children per 1000 ever-married women. The RMSE are observed highest when Poisson regression model is used for modeling births for KPK. While for other regions continuing with Poisson, ZIP and NB give almost similar RMSE for estimating GFR at sub-national and national levels. The GFR for full data without model fitting are obtained in last column of Table 7.17.. The GRR is computed using proportion of female births (PF) from all age groups using sex ratio of male to female from full data (PF from census or from administrative records can be used for obtaining sex ratio). The GRR is observed higher in Bolachistan, Punjab and KPK as compared to other regions replacement of 2 or more daughters per women before the death of their mother. RMSE are smaller for Punjab, Sindh and KPK when NB model is used for prediction. While it is smaller for remaining regions when ZIP model is employed.

Table 7.16.: Model-based TFR at regional level

REGION		Poisson	NB	ZIP	ZINB	Full Data
PUNJAB	MEAN	5.4441	**5.4456**	5.6440	5.6992	4.4317
	RMSE	1.0153	0.1107	0.5165	0.5705	
SINDH	MEAN	5.5838	**5.5471**	5.8048	5.7484	4.4317
	RMSE	1.1555	0.1313	0.4532	0.4015	
KPK	MEAN	5.5127	**5.5166**	5.7537	5.7458	4.4317
	RMSE	1.0844	0.0878	0.4440	0.4359	
BOLACHISTAN	MEAN	5.7095	5.7349	**5.7669**	5.7893	4.4317
	RMSE	1.2829	0.2292	0.2282	0.2556	
ICT	MEAN	6.2603	6.2331	**6.2828**	6.2749	4.4317
	RMSE	1.8331	0.6443	0.4199	0.4108	
FATA	MEAN	5.5316	5.4921	5.8499	**5.7698**	4.4317
	RMSE	1.1092	0.7648	0.5838	0.5306	
PAKISTAN	MEAN	5.6337	5.5936	**5.8035**	5.7619	4.4317
	RMSE	1.2025	0.1914	0.0617	0.3910	

Table 7.17.: Model-based GFR at regional level

REGION		Poisson	NB	ZIP	ZINB	Full Data
PUNJAB	MEAN	**161.797**	161.755	170.618	172.124	160.784
	RMSE	2.191	2.195	10.128	11.512	
SINDH	MEAN	164.991	**164.327**	174.105	172.506	164.967
	RMSE	2.460	2.476	10.200	8.753	
KPK	MEAN	225.582	**174.952**	182.789	182.689	174.797
	RMSE	50.894	2.644	8.331	8.369	
BOLACHISTAN	MEAN	**168.603**	169.313	171.639	172.448	168.600
	RMSE	3.284	3.455	4.391	5.342	
ICT	MEAN	**181.956**	181.466	186.118	185.669	179.435
	RMSE	4.095	4.383	8.044	7.659	
FATA	MEAN	**153.197**	152.487	166.547	164.103	153.163
	RMSE	3.635	3.660	13.856	12.113	
PAKISTAN	MEAN	**168.187**	167.459	176.005	174.751	168.165
	RMSE	1.064	1.075	8.821	7.523	

Table 7.18.: Model-Based GRR at Regional Level

REGION		Poisson	NB	ZIP	ZINB	Full Data
PUNJAB	MEAN	2.161824	**2.46518**	2.528332	2.55295	2.704505
	RMSE	0.542681	0.052107	0.212113	0.236151	
SINDH	MEAN	1.970801	**2.51457**	2.591036	2.566196	2.469386
	RMSE	0.504035	0.057286	0.184696	0.161956	
KPK	MEAN	2.161824	**2.506772**	2.562835	2.557729	2.637384
	RMSE	0.475561	0.041888	0.173325	0.168304	
BOLACHISTAN	MEAN	2.161824	2.617896	**2.426287**	2.435872	2.817091
	RMSE	0.655267	0.109318	0.086432	0.09822	
ICT	MEAN	1.970801	2.799685	**2.81312**	2.810767	2.310812
	RMSE	0.368522	0.302086	0.167947	0.165813	
FATA	MEAN	1.970801	2.496049	2.423849	**2.388887**	2.798262
	RMSE	0.83461	0.342369	0.219095	0.195004	
PAKISTAN	MEAN	1.970801	**2.540**	2.610421	2.546088	2.415219
	RMSE	0.444418	0.078	0.026216	0.152864	

7.8 Conclusion

In this chapter, we analyzed the birth history data using three separate models taking 1-year period births for first, 3-years period birth for second and 5-years period births for third models as the responses with 24 regressors. Although Poisson regression model is considered useful for the data with count responses the assumption of equality of mean and variance spoils the inference. To deal with responses having large variance and many 0's as well as a few very large values, we used NB, ZIP and ZINB models as extensions of Poisson model. Comparison of the three methods were made using AIC. The ZIP model gives smaller AIC as it deals with inflated zeros. The ZIP and ZINB model produce approximately same result so we did not report results obtained under ZINB model for inference purpose. We also conducted the estimation of regression models under Bayesian paradigm assuming normal priors for each coefficients including intercept under Poisson regression model. The posterior means were obtained using rjags package in R. The posterior means for each coefficients are closer to classical estimates for Poisson models. Some model diagnostics were used to check the validity of estimation procedure. The model diagnostics indicated positive report of the estimation procedure. Predictive ASFR, TFR, GFR and TRR were obtained using predicted response under above estimated models. Illustration of predictive approach taking bootstrap samples from the PDHS 2017-18 individual recode data was provided. The predictive rates provide efficient results when relevant auxiliary data are available from census at unit level or at cluster level. It is important to mention that the data about majority of regressors considered for prediction in this study is not easy in practical situations at individual level. However, data might be available for clusters through Civil Registration of vital events or from previously conducted surveys. The predictive approach suggested here in constructing rates in case of missing responses on birth and for small area estimation.

Chapter 8

Conclusion of the Study

8.1 Outline

This chapter provides a comprehensive concluding remarks of the study. Based on theoretical and practical research works related to model-based estimation, we provide some theoretical frame work related to parameter estimation, specially population total, in finite population setting under model-based approach. Extensions and possible application were followed by the theoretical framework. An application is provided based on the birth data from PDHS 2017-18. In Section 8.2, we describe main findings of our study via theoretical and applied point of view followed by brief discussion of related works. Section 8.3 gives some recommendation for future work. While Section 8.4 delineates limitations of the study.

8.2 Concluding Remarks

A general framework of model-based approach for estimation of finite population parameter τ (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 $\hat{\tau}$ were derived. For statistical inference about τ, estimation of prediction error variance under different model section criteria i.e. residual, GCV, UEV, FPE and BIC methods (the widely used feature selection criteria in ML) was also studied. Among all competing variance estimators, the estimator obtained under UEV provides minimum variance estimates and the variance estimator under BIC is highest which was shown theoretically as well as via simulation. For the purpose of simulation study and further modifications, we deducted the case of estimation of population total i.e. $\gamma_i = 1$ for all $i \in U$. The study also provides a guideline for model selection in finite population parameter estimation through incremental operators. The model selection is based on a measure, named as increment in efficiency, (IE) which provides a guideline for selecting a model with appropriate number of basis function. Positive value of IE shows increase in efficiency on adding additional basis functions to the feature matrix.

In Chapter 3, prior information about the superpopulation parameters had been incorporated for estimating the finite population parameter with maximum efficiency. The simulated results depict the superiority of estimators of the population total with prior information. Increase in σ_0^2 results in decrease in variance but introduces bias in the estimator and makes a trade-off in expected squared prediction error (ESPE). Estimators for prediction error variance of the Bayesian estimator of τ were also obtained using model selection criteria with the help of projection matrix. We established an ordering of the expected values of estimated variance as $\hat{V}(e(\hat{\tau}))_{RES} \leq \hat{V}(e(\hat{\tau}))_{UEV} \leq \hat{V}(e(\hat{\tau}))_{FPE} \leq \hat{V}(e(\hat{\tau}))_{GCV} \leq \hat{V}(e(\hat{\tau}))_{BIC}$. This ordering was verified through simulated and bootstrapped sampling. Tables 3.1.–3.4. provided the comparison and behavior of estimated variances. It is concluded that all variance estimators are increasing function of the variance ratio v_r on the average. Same statement is justified for the relation between the estimated variances and sample size (except for $M = 4$ in Table 3.4.). Further, ill-conditioning of the regression estimation was also coped with typical regularization method which introduces slight bias in estimates of β's but provides smaller estimate of the variance of the error term and, consequently, smaller estimated variance of the prediction error of $\hat{\tau}$.

Chapter 4 covered a model-based version of Hansen and Hurwitz (1946) sub-sampling technique technique for handling non-ignorable non-response in estimation of finite population parameters (specially total). The method works under the assumption that the responding and non-responding population have different models and the occurrence of non-response is observable like a stratification variable. The sub-sampling technique is suggested for application in the field of public health where the non-response occurrence is often found with respect to gender, ethical affiliation, age and other demographic factors of the respondents. Consequently, respondents and non-respondents might have different models. From Chapter 4, we also conclude that under linear population model (linear in parameter as well as in variables) the total estimator with sub-sampling is model-unbiased and has smaller model-variance as compared to predictive estimator based on sampled respondents only. The linearity assumption emphasizes on linear in parameters but not restricted to the linearity in variable. To cope with ill-conditioning, we adapt a version of ridge regression, called partial ridge regression, for predicting the non-sampled non-respondents. Mathematical expressions are verified via a numerical study with blood transfusion data and an extensive Monte Carlo experiment.

Application of a new ranked set sampling mechanism for the estimation of finite population parameter (specially total) under GPM is another major contribution of this dissertation. In Chapter 5, Figure 5.1 presented a picture of the RSSWOR which assumes that the finite population is coming from an infinite superpopulation via some stochastic process with finite mean and variance. It is also assumed that the population can be generated from different points i.e. cycles and the m sets taken from one cycle is totally different from the m in other cycles for insuring without replacement. The mathematical expressions and Monte-Carlo experiment both supported the superiority of the total estimator under RSSWOR over the competitor under SRSWOR for GPM as well as HPM. The suggested estimators can be recommended for process controls by constructing control charts.

Estimation of sub-population total under a new version of ranked set sampling for obtaining a without replacement sample with GPM (general form of proportional population model) was also dealt in Chapter 6. Figure 6.1 illustrated the RSSWOR sampling algorithm which assumes that the finite population is coming from an infinite superpopulation via some stochastic process with finite mean and variance. Domain membership variable was observed from selected ranked set sample. The model relationship between the study variable and the auxiliary variable for whole population was used to predict the non-sampled values to establish a domain specific estimator for total. The superiority of the domain specific total estimator under RSSWOR over the total estimator under SRSWOR for GPM as well as HPM was shown mathematically as well as through Monte-Carlo experiment. The domain specific estimators is highly recommended to use in epidemiology and public health research where one need to find total exposure to certain event for different sub-populations.

Finally, in Chapter 7, we analyzed the birth history data using three separate models taking 1-year period births for first, 3-years period birth for second and 5-years period births for third models as the responses and 24 regressors. The histograms for the outcomes given in Figure 7.1 recommended to use Poisson regression model with log-link function. Although, Poisson regression model is considered useful for the data with count responses the assumption of equality of mean and variance spoils the inference. To deal with responses having large variance and many 0's as well as a few very large values, we used NB, ZIP and ZINB models as extensions of Poisson model. Comparison of the three methods was made on the basis of AIC values. The ZIP model gives smaller AIC as it deals with inflated zeros. The ZIP and ZINB model produce approximately same result. We also conducted the estimation of regression models under Bayesian paradigm assuming normal priors for each coefficients including intercept. The posterior means were obtained using MCMC via rjags package in R. The posterior means for each coefficients are observed closer to classical estimates for Poisson models. Some model diagnostics were used to check the validity of estimation procedure. The model diagnostics indicated positive report of the estimation procedure. Model based fertility rates including ASFR, TFR, GFR and GRR were obtained using predicted response under the estimated models. We provided illustration of predictive approach through bootstrap sampling from the PDHS 2017-18 individual recode data. The model-based rates provide efficient results when relevant auxiliary data are available from census at unit level or at cluster level. It is important to mention that the data about majority of regressors considered for prediction in this study is not easy in practical situations at individual level. However, data might be available for clusters through civil registration of vital events or from previously conducted surveys. The predictive approach suggested here for constructing rates are helpful in case of missing responses on birth and providing separate estimates to the domains with insufficient sample sizes.

8.3 Recommendations for Future Research

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 multi-level models and small area estimation. Researchers who have interest to work in the same areas are recommended to extend the following topics

1. Estimation of finite population quantities in multi-level models.

2. Incorporation of weights in estimation stage which is a widely accepted strategy for

obtaining reliable estimates in survey sampling domain.

3. Utilization of more advanced machine learning (ML) algorithm for prediction purpose such as random forest model, radial basis function.

4. Further extension of the proposed work to other sampling designs such as stratified and cluster sampling.

5. Application of the model-based framework to the data from other surveys and fields of research.

6. Obtaining more sophisticated small area estimators using fixed effect and random effect models.

8.4 Limitations of the Study

The model-based frame work is used here for the non-linear model in the sense that the response variable depends on some non-linear function of x denoted by $\phi(x)$. Hence the proposed study is not applicable to the regression models with non-linear in parameter. The GLMs used to model the birth data in application section which are converted to linear functions via some link functions and estimated birth rates are obtained through back transformation. A major concern with practicability of the model-based approach is the assumption about error terms. In certain situations, specially for RSSWOR, it is not possible to obtain an IID error term in practical situations. Another limitation can be found in the observability of the non-response factor for splitting the respondents in non-respondents for tackling the situation of non-response under model-based framework. Finally in application section, model-based fertility rates are obtained without incorporating survey weights which leads to biased results. Incorporating survey weights for fertility rates estimation under model-based framework is an open area to work and recommended above.

References

A Alamm, A., Raqab, M. Z., and Madi, M. T. (2007). Bayesian prediction intervals for future order statistics from the generalized exponential distribution. *Journal of the Iranian Statistical Society*, 6(1):17–30.

Abel, G. J., Barakat, B., Samir, K., and Lutz, W. (2016). Meeting the sustainable development goals leads to lower world population growth. *Proceedings of the National Academy of Sciences*, 113(50):14294–14299.

Ahmad, A. A., Mohammad, Z. R., and Mohamed, T. M. (2007). Bayesian prediction intervals for the future order statistics from the generalized exponential distribution. *Journal of The Iranian Statistical Society*, 6(1):17–30.

Ahmed, S. and Shabbir, J. (2019a). Extreme-cum-median ranked set sampling. *Brazilian Journal of Probability and Statistics*, 33(1):24–38.

Ahmed, S. and Shabbir, J. (2019b). On use of ranked set sampling for estimating super-population total: Gamma population model. *Scientia Iranica (accepted on 18 june 2019*, DOI:10.24200/SCI.2019.50976.1946.

Ahmed, S., Shabbir, J., and Gupta, S. (2017). Use of scrambled response model in estimating the finite population mean in presence of non response when coefficient of variation is known. *Communications in Statistics-Theory and Methods*, 46(17):8435–8449.

Aitchison, J. and Dunsmore, I. R. (1975). Statistical prediction analysis. *Cambridge, MA: Cambridge University Press.*

Akaike, H. (1977). On entropy maximization principle,. *Applications of Statistics*, pages 27–41.

Al-Omari, A. I. and Jaber, K. H. (2008). Percentile double ranked set sampling. *Journal of Mathematics and Statistics*, 4(1):60–64.

Al-Saleh, M. F. and Al-Omari, A. I. (2002). Multistage ranked set sampling. *Journal of Statistical planning and Inference*, 102(2):273–286.

Alpaydin, E. (2009). *Introduction to machine learning*. MIT Press, Massachuastts.

Ambler, R., Caplan, D., Chambers, R., Kovacevic, M., and Wang, S. (2001). Combining unemployment benefits data and lfs data to estimate ilo unemployment for small areas: An application of a modified fay-herriot methodî,. *Proceedings of the International Association of Survey Statisticians, Meeting of the International Statistical Institute, Seoul, August 2001*.

Angeles, G., Guilkey, D. K., and Mroz, T. A. (1998). Purposive program placement and the estimation of family planning program effects in tanzania. *Journal of the American Statistical Association*, 93(443):884–899.

Ardilly, P. (1991). Optimum and proportional to size sampling procedure. *Annales of Economics and Statistics*, 23:91–113.

Atkins, D. C., Baldwin, S. A., Zheng, C., Gallop, R. J., and Neighbors, C. (2013). A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychology of Addictive Behaviors*, 27(1):166–177.

Atkins, D. C. and Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: a tutorial on count regression and zero-inflated models. *Journal of Family Psychology*, 21(4):726–735.

Barreto, M. C. M. and Barnett, V. (1999). Best linear unbiased estimators for the simple linear regression model using ranked set sampling. *Environmental and Ecological Statistics*, 6(2):119–133.

Barton, J., Bain, C., Hennekens, C. H., Rosner, B., Belanger, C., Roth, A., and Speizer, F. E. (1980). Characteristics of respondents and non-respondents to a mailed questionnaire. *American Journal of Public Health*, 70(8):823–825.

Beaujean, A. A. and Morgan, G. B. (2016). Tutorial on using regression models with count outcomes using r. *Practical Assessment, Research & Evaluation*, 21.

Bellhouse, D. R. (1987). Model-based estimation in finite population sampling. *The American Statistician*, 41(4):260–262.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Biancolini, M. E. (2017). *Fast radial basis functions for engineering applications*. Springer International Publishing.

Biemer, P. P., Chen, P., and Wang, K. (2013). Using level-of-effort paradata in non-response adjustments with application to field surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1):147–168.

Bierens, H. J. (1988). *The Nadaraya-Watson kernel regression function estimator*. Faculty of Economics and Business Adminstration, Vrije University, Amsterdam.

Böhning, D. (1992). Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, 44(1):197–200.

Boisbunon, A., Canu, S., Fourdrinier, D., Strawderman, W., and Wells, M. T. (2013). Aic, cp and estimators of loss for elliptically symmetric distributions. *arXiv preprint arXiv:1308.2766*.

Boone, E. L., Merrick, J. R., and Krachey, M. J. (2014). A hellinger distance approach to mcmc diagnostics. *Journal of Statistical Computation and Simulation*, 84(4):833–849.

Bouza, C. N. (2002). Ranked set sub sampling the non response strata for estimating the difference of means. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 44(7):903–915.

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4):1026–1053.

Breidt, F. J., Opsomer, J. D., et al. (2008). Endogenous post-stratification in surveys: classifying with a sample-fitted model. *The Annals of Statistics*, 36(1):403–427.

Broomhead, D. S. and Lowe, D. (1988). Radial basis functions, multi-variable functional interpolation and adaptive networks(no. rsre-4148). Technical report, Royal Signals and Radar Establishment Malvern (United Kingdom).

Brown, G., Chambers, R., Heady, P., and Heasman, D. (2001). Evaluation of small area estimation methods – an application to unemployment estimates from the uk lfs. *Proceedings of Statistics Canada Symposium Achieving Data Quality in a Statistical Agency: A Methodological Perspective, CDROM*.

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

Cai, Z., Si, W., Si, S., and Sun, S. (2014). Modeling of failure prediction bayesian network with divide-and-conquer principle. *Mathematical Problems in Engineering*, 2014:1–7.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press.

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

Casady, R. J. and Valliant, R. (1993). Conditional properties of poststratified estimators under normal theory. *Survey Methodology*, 19(2):183–192.

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

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

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

Chamratrithirong, A., Kamnuansilpa, P., and Knodel, J. (1986). Contraceptive practice and fertility in thailand: results of the third contraceptive prevalence survey. *Studies in Family Planning*, 17(6 Pt 1):278–287.

Chen, Z. and Wang, Y.-G. (2004). Efficient regression analysis with ranked-set sampling. *Biometrics*, 60(4):997–1004.

Clair, L. (2017). *Nonparametric Kernel Estimation Methods Using Complex Survey Data*. PhD thesis.

Cochran, W. (1940). 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):262–275.

Copas, A. J. and Farewell, V. T. (1998). Dealing with non-ignorable non-response by using an 'enthusiasm-to-respond'variable. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 161(3):385–396.

Croft, T. N., Marshall, A. M., Allen, C. K., et al. (2018). Guide to dhs statistics. *Rockville, Maryland, USA: ICF*.

Dell, T. and Clutter, J. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, pages 545–555.

Dever, J. A. and Valliant, R. (2010). A comparison of variance estimators for poststratification to estimated control totals. *Survey Methodology*, 36(1):45–56.

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

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

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

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

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

Draper, N. R. and Smith, H. (2014). *Applied regression analysis (Vol. 326).* John Wiley & Sons. doi 10:9781118625590.

Drew, J. D., Singh, M. P., and Choudhry, G. H. (1982). Evaluation of small area techniques for the canadian labour force survey. *Survey Methodology*, 8(1):17—-47.

Elkasabi, M. (2019). Calculating fertility and childhood mortality rates from survey data using the dhs. rates r package. *PloS one*, 14(5):e0216403.

Ericson, W. A. (1967). Optimal sample design with nonresponse. *Journal of the American Statistical Association*, 62(317):63–78.

Falorsi, P. D. and Righi, P. (2016). A unified approach for defining optimal multivariate and multi-domains sampling designs. In *Topics in Theoretical and Applied Statistics*, pages 145–152. Springer Internation Publishing Switzerland.

Fan, G. (1996). *Local Polynomial Modeling and its Applications, London*. Chapman and Hal/CRC Press, London.

Fay, R. E. and Herriot, R. A. (1979). Estimation of income for small places: An application of jamesstein procedures to census data,. *Journal of the American Statistical Association*, 74:268–277.

Fushiki, T. (2011). Estimation of prediction error by using k-fold crossvalidation. *Statistics and Computing*, 21:137–146.

Gardner, W., Mulvey, E. P., and Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed poisson, and negative binomial models. *Psychological Bulletin*, 118(3):392–404.

Giraud, C. (2014). *Introduction to high-dimensional statistics*. Chapman and Hall/CRC.

Godambe, V. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):269–278.

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

Guan, Z., Leung, D. H., and Qin, J. (2018). Semiparametric maximum likelihood inference for nonignorable nonresponse with callbacks. *Scandinavian Journal of Statistics*, 45(4):962–984.

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

Hamra, G., MacLehose, R., and Richardson, D. (2013). Markov chain monte carlo: an introduction for epidemiologists. *International Journal of Epidemiology*, 42(2):627–634.

Hansen, M. H. and Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41(236):517–529.

Haq, A., Brown, J., Moltchanova, E., and Al-Omari, A. I. (2014). Mixed ranked set sampling design. *Journal of Applied Statistics*, 41(10):2141–2156.

Hazlett, C. (2013). A balancing method to equalize multivariate densities and reduce bias without a specification search. *Department of Political Science, 77 Massachusetts Avenue, Cambridge, MA 02139*.

Hedayat, A. and Majumdar, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *Journal of Statistical Planning and Inference*, 44(2):237–247.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hill, K. (2013). Introduction to child mortality analysis. *Tools for D*, page 141.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Holt, D. and Elliot, D. (1991). Methods of weighting for unit non-response. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 40(3):333–342.

Horn, R. and Johnson, C. (1985). *Matrix analysis*. Cambridge University Press.

Huang, G. and Tai-kang (2009). Taking the opportunities of health care reform to develop clinical pharmacy. *Asian Journal of Social Pharmacy4*, 2:65–69.

IGME (2018). United nations inter-agency group for child mortality estimation (un igme). levels & trends in child mortality: Report 2018, estimates developed by the united nations inter-agency group for child mortality estimation. *New York: United Nations Children's Fund*.

Jackman, S. (2000). Estimation and inference are missing data problems: Unifying social science statistics via bayesian simulation. *Political Analysis*, 8(4):307–332.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer, New York.

James, W. and Stein, C. (1992). Estimation with quadratic loss. In *Breakthroughs in statistics*, pages 443–460. Springer.

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

Jochems, A., Deist, T. M., Van Soest, J., Eble, M., Bulens, P., Coucke, P., Dries, W., Lambin, P., and Dekker, A. (2016). Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital–a real life proof of concept. *Radiotherapy and Oncology*, 121(3):459–467.

Khare, B. and Sinha, R. (2009). On class of estimators for population mean using multi-auxiliary characters in the presence of non-response. *Statistics in Transition*, 10(1):3–14.

Khare, B. and Srivastava, S. (1993). Estimation of population mean using auxiliary character in presence of non-response. *National Academy Science Letters*, 16(3):111–114.

Khare, B. and Srivastava, S. (1995). Study of conventional and alternative two phase sampling ratio, product and regression estimators in presence of nonresponse. *Proceedings-National Academy Of Sciences India Section A*, 65:195–204.

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):291–311.

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):92–101.

Knudsen, A. K., Hotopf, M., Skogen, J. C., Øverland, S., and Mykletun, A. (2010). The health status of nonparticipants in a population-based health study: the hordaland health study. *American Journal of Epidemiology*, 172(11):1306–1314.

Lightbourne, J. R., Singh, S., and Green, C. P. (1982). The world fertility survey: charting global childbearing. *Population Bulletin*, 37(1):1–55.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables (Vol. 7)*. Thousand Oaks, CA:Sage.

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

Luc, C. (2016). Nonparametric kernel regression using complex survey data. *Job market paper*.

Lütkepohl, H. (1996). *Handbook of matrices*, volume 1. Wiley Chichester.

Lynch, S. M. (2007). Evaluating markov chain monte carlo algorithms and model fit. In *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*, pages 131–164. Springer.

Mahdizadeh, M. and Zamanzade, E. (2019). Efficient body fat estimation using multistage pair ranked set sampling. *Statistical Methods in Medical Research*, 28(1):223–234.

Mallows, C. L. (1973). Some comments on c p. *Technometrics*, 15(4):661–675.

Masset, E. (2016). Syncmrates: Stata module to compute child mortality rates using synthetic cohort probabilities. *Statistical Software Components S458149*.

McIntyre, G. (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research*, 3(4):385–390.

Mencarini, L. (1999). An analysis of fertility and infant mortality in south africa based on 1993 lsds data. In *Third African Population Conference, African Population in the 21st Century*, pages 109–128. UAPS.

Mengersen, K. L., Robert, C. P., and Guihenneuc-Jouyaux, C. (1999). Mcmc convergence diagnostics: a reviewww. *Bayesian Statistics*, 6:415–440.

Minka, T. (2000). Bayesian linear regression. Technical report, Technical Report, Massachusetts Institute of Technology in Cambridge, MA.

Moore, R. E. (2014). *Reliability in computing: the role of interval methods in scientific computing*, volume 19. Elsevier.

Moultrie, T. A., Dorrington, R., Hill, A., Hill, K., Timæus, I., and Zaba, B. (2013). *Tools for demographic estimation*. International Union for the Scientific Study of Population.

Mukhopadhyay, P. (1993). Estimation of a finite population total under regression models: a review. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 141–155.

Murthy, M. (1964). Product method of estimation. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 69–74.

Muttlak, H. A. (2003). Modified ranked set sampling methods. *Pakistan Journal of Statistics*, 19(3):315–324.

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

Najarian, S., Arashi, M., and Kibria, B. G. (2013). A simulation study on some restricted ridge regression estimators. *Communications in Statistics-Simulation and Computation*, 42(4):871–890.

Nazerfard, E. and Cook, D. J. (2015). Crafft: an activity prediction model based on bayesian networks. *Journal of Ambient Intelligence and Humanized Computing*, 6(2):193–205.

Ohyama, T., Doi, J. A., and Yanagawa, T. (2008). Estimating population characteristics by incorporating prior values in stratified random sampling/ranked set sampling. *Journal of Statistical Planning and Inference*, 138(12):4021–4032.

Orr, M. J. et al. (1996). Introduction to radial basis function networks.

Patil, G., Sinha, A., and Taillie, C. (1995). Finite population corrections for ranked set sampling. *Annals of the Institute of Statistical Mathematics*, 47(4):621–636.

PDHS (2019). Pakistan demographic and health survey 2017-18 key indicators report, national institute of population studies islamabad, pakistan. *The DHS Program ICF Rockville, Maryland, USA*.

Pettit, L. (1986). Diagnostics in bayesian model choice. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 35(2):183–190.

Peytchev, A., Baxter, R. K., and Carley-Baxter, L. R. (2009). Not all survey effort is equal: Reduction of nonresponse bias and nonresponse error. *Public Opinion Quarterly*, 73(4):785–806.

Pfeffermann, D., Bell, P., and Signorelli, D. . (1996). Labour force trend estimation in small areas. *Proceedings of the Annual Research Conference, US Bureau of the Census*, pages 407–431.

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

Powers, D. and Xie, Y. (2008). *Statistical methods for categorical data analysis*. Emerald Group Publishing.

Press, William H, V. W. and Flannery, B. (1992). Numerical recipes in c:the art of scientific computing. *Cambridge University Press*, pages 656–680.

Priya, R. and Thomas, P. Y. (2016). An application of ranked set sampling when observations from several distributions are to be included in the sample. *Communications in Statistics-Theory and Methods*, 45(23):7040–7052.

Purcell, N. I. and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review,*, 48(1):3–18.

Raftery, A. E., Lewis, S. M., and Aghajanian, A. (1995). Demand or ideation? evidence from the iranian marital fertility decline. *Demography*, 32(2):159–182.

Rao, J. (1994). Small area estimation by combining time series and cross sectional data. *Canadian Journal of Statisitcs*, 22:511–528.

Rao, J. (2003). Small area estimation. *New York: Wiley.*

Raqab, M. Z. and Madi, M. T. (2002). Bayesian prediction of the total time on test using doubly censored rayleigh data. *Journal of Statistical Computational and Simulation*, 72:781–789.

Rawlings, J. O., Pantula, S. G., and Dickey, D. A. (2001). *Applied regression analysis: a research tool*. Springer Science & Business Media.

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

Royall, R. M. and Cumberland, W. G. (1981). The finite-population linear regression estimator and estimators of its variance—an empirical study. *Journal of the American Statistical Association*, 76(376):924–930.

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

Salehi, M. and Ahmadi, J. (2015). Estimation of stress-strength reliability using record ranked set sampling scheme from the exponential distribution. *Filomat*, 29(5):1149–1162.

Samawi, H. M. and Muttlak, H. A. (1996). Estimation of ratio using rank set sampling. *Biometrical Journal*, 38(6):753–764.

Sánchez-Borrego, I., Opsomer, J. D., Rueda, M., and Arcos, A. (2014). Nonparametric estimation with mixed data types in survey sampling. *Revista Matemática Complutense*, 27(2):685–700.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33(2):99–119.

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):2758–2765.

Schoumaker, B. (2013). A stata module for computing fertility rates and tfrs from birth histories: tfr2. *Demographic Research*, 28(38):1093–1144.

Schoumaker, B. and Hayford, S. R. (2004). A person-period approach to analysing birth histories. *Population*, 59(5):689–702.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shao, J., Wu, C. J., et al. (1989). A general theory for jackknife variance estimation. *The Annals of Statistics*, 17(3):1176–1197.

Singh, H. P. and Kumar, S. (2008). A regression approach to the estimation of the finite population mean in the presenece of non-response. *Australian & New Zealand Journal of Statistics*, 50(4):395–408.

Sinha, S. K. (1990). On the prediction limits for rayleigh life distribution. *Calcutta Statistical Association Bulletin,*, 39:105–109.

Smith, A. (1986). Some bayesian thoughts on modelling and model choice. *The Statistician*, 35(2):97–101.

Smouse, E. P. (1982). Bayesian estimation of a finite population total using auxiliary information in the presence of nonresponse. *Journal of the American Statistical Association*, 77(377):97–102.

Särndal, C., Swesson, B., and Wretman, J. (1978). Statistical analysis of reliability and life testing model. *New York, NY: Marcel Dekker.*

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

Tikhonov, A. N. and Arsenin, V. I. (1977). *Solutions of ill-posed problems*, volume 14. Vh Winston.

UN (2011). Mortality estimates from major sample surveys: Towards the design of a database for the monitoring of mortality levels and trends. *United Nations Department of Economic and Social Affairs, Population Division, The Technical Paper series; (Technical Paper No. 2011/2).*

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):627–636.

Valliant, R. (2000). *Finite population sampling and inference: a prediction approach*. Number 04; QA276. 6, V3.

Vats, D., Robertson, N., Flegal, J. M., and Jones, G. L. (2019). Analyzing mcmc output. *arXiv preprint arXiv:1907.11680*.

Vehtari, A. and Ojanen, J. (2012). A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228.

Ver H., J. M. and Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772.

Vinod, H. D. and Ullah, A. (1981). *Recent advances in regression methods*, volume 41. Marcel Dekker Incorporated.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

White, I. R., Kalaitzaki, E., and Thompson, S. G. (2011). Allowing for missing outcome data and incomplete uptake of randomised interventions, with application to an internet-based alcohol trial. *Statistics in Medicine*, 30(27):3192–3207.

Whitworth, A., Carter, E., Ballas, D., and Moon, G. (2017). Estimating uncertainty in spatial microsimulation approaches to small area estimation: A new approach to solving an old problem. *Computers, Environment and Urban Systems*, 63:50–57.

Wood, A. M., White, I. R., and Hotopf, M. (2006). Using number of failed contact attempts to adjust for non-ignorable non-response. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):525–542.

Wu, J., Yin, L., and Guo, Y. (2012). Cyber attacks prediction model based on bayesian network. In *Parallel and Distributed Systems (ICPADS), 2012 IEEE 18th International Conference on*, pages 730–731. IEEE.

Yeh, I.-C., Yang, K.-J., and Ting, T.-M. (2009). Knowledge discovery on rfm model using bernoulli sequence. *Expert Systems with Applications*, 36(3):5866–5871.

You, Y. (2008). Small area estimation using area level models with model checking and applications. *Proceedings of the Survey Methods Section, Statistical Society of Canada*.

Yuan, Y. C. (2000). Multiple imputation for missing data: Concepts and new development. In *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference*, volume 267.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in r. *Journal of Statistical Software*, 27(8):1–25.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81(394):446–451.

Zellner, A. (1994). Bayesian and non-bayesian estimation using balanced loss functions. In *In Statistical decision theory and related topics V*, pages 377–390. Springer.

Zhang, L.-C. (2000). Post-stratification and calibration—a synthesis. *The American Statistician*, 54(3):178–184.

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

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

Zheng, H. and Little, R. J. (2004). Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Survey Methodology*, 30(2):209–218.

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

Appendix A

Model Selection in Basis Function Regression

The Projection Matrix

The projection matrix P based on M covariates is defined by

$$P_m = I_M - \Phi_s Q_{sm}^{-1} \Phi_s^T \tag{A.1}$$

where $Q_{sm}^{-1} = \Phi_s^T \Phi_s + \nu I_M$ is the Hessian Matrix (Lütkepohl, 1996) based on Φ_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 Horn and Johnson (1985) for the purpose of matrix inversion

Lemma 1 For a any partitioned square matrix B defined as:

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

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

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

Lemma 2

Let the inverse of matrix $B_0^{-1} \in \mathcal{R}^{m\times m}$, X, $Y^T \in \mathcal{R}^{m\times r}$ and $R \in \mathcal{R}^{r\times r}$ 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

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

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

Decrement in Variance

The Hessian matrix after adding an additional basis function

$$Q_{m+1} = \Phi^T_{s(m+1)}\Phi s(m+1) + vI_{m+1} = \begin{bmatrix} Q_{sm} & \Phi^T_{sm}\phi_{s(m+1)} \\ \phi^T_{s(m+1)}\Phi_{sm} & \phi^T_{s(m+1)}\phi_{s(m+1)} + v \end{bmatrix} \tag{A.2}$$

where $\Phi^T_{s(m+1)} = \begin{bmatrix} \Phi_{sm} & \phi_{s(m+1)} \end{bmatrix}$ and $Q_{sm} = \Phi^T_{sm}\Phi_{sm} + vI_m$. To obtain inverse of the Hessian matrix given in (A.2), we use Lemma 1 with

$$Q^{-1}_{s(m+1)} = \begin{bmatrix} Q^{-1}_{sm} & 0 \\ 0^T & 0 \end{bmatrix} + \triangle^{-1}\begin{bmatrix} Q^{-1}_{sm}\Phi^T_{sm}\phi_{s(m+1)}\phi^T_{s(m+1)}\Phi_{sm}Q^{-1}_{sm} & -Q^{-1}_{sm}\Phi^T_{sm}\phi_{s(m+1)} \\ -\phi^T_{s(m+1)}\Phi_{sm}Q^{-1}_{sm} & 1 \end{bmatrix}, \tag{A.3}$$

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

$$\begin{aligned}
\phi_{\bar{s}(m+1)}Q^{-1}_{\bar{s}(m+1)}\phi^T_{s(m+1)} &= \begin{bmatrix} \Phi_{\bar{s}m} & \phi_{\bar{s}(m+1)} \end{bmatrix} \times \begin{bmatrix} Q^{-1}_{11} & Q^{-1}_{12} \\ Q^{-1}_{21} & Q^{-1}_{22} \end{bmatrix} \begin{bmatrix} \Phi^T_{\bar{s}m} \\ \phi^T_{\bar{s}(m+1)} \end{bmatrix} \\
&= \begin{bmatrix} \Phi_{\bar{s}m}Q^{-1}_{11} + \phi_{\bar{s}(m+1)}Q^{-1}_{21} & \Phi_{\bar{s}m}Q^{-1}_{12} + \phi_{\bar{s}(m+1)}Q^{-1}_{22} \end{bmatrix} \\
&\quad \times \begin{bmatrix} \Phi^T_{\bar{s}m} \\ \phi^T_{\bar{s}(m+1)} \end{bmatrix} \\
&= \Phi_{\bar{s}m}Q^{-1}_{11}\Phi^T_{\bar{s}m} + \phi_{\bar{s}(m+1)}Q^{-1}_{21}\Phi^T_{\bar{s}m} + \Phi_{\bar{s}m}Q^{-1}_{12}\phi^T_{\bar{s}(m+1)}
\end{aligned}$$

$$+\phi_{\bar{s}(m+1)}Q_{22}^{-1}\phi_{\bar{s}(m+1)}^{T}, \tag{A.4}$$

where
$Q_{11}^{-1} = Q_{sm}^{-1} + \triangle^{-1}Q_{sm}^{-1}\Phi_{sm}^{T}\phi_{s(m+1)}\phi_{s(m+1)}^{T}\Phi_{sm}Q_{sm}^{-1}$, $Q_{12}^{-1} = -\triangle^{-1}\,Q_{sm}^{-1}\Phi_{sm}^{T}\phi_{s(m+1)}$, $Q_{21}^{-1} = -\triangle^{-1}\,\phi_{s(m+1)}^{T}\Phi_{sm}Q_{sm}^{-1}$ and $Q_{22}^{-1} = \triangle^{-1}$. 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 $Q_{sm}^{-1} = A_{sm}^{-1}$ for $m = 1, 2, ...M$.

For prediction models with no regularization, we have the variance of $V_M(\hat{\tau})$ with M regressors

$$V_M(\hat{\tau})_m = \big(N-n\big)\sigma^2 + \sigma^2\left[\gamma_{\bar{s}}^{T}\Phi_{\bar{s}m}A_{sm}^{-1}\Phi_{\bar{s}m}^{T}\gamma_{\bar{s}}\right].$$

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

$$V_M(\hat{\tau})_{m+1} = \big(N-n\big)\sigma^2 + \sigma^2\left[\gamma_{\bar{s}}^{T}\Phi_{\bar{s}(m+1)}A_{s(m+1)}^{-1}\Phi_{\bar{s}(m+1)}^{T}\gamma_{\bar{s}}\right].$$

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

$$B_M\big(\hat{\tau}_{ridge}(y)\big)_m = -v\gamma_{\bar{s}}^{T}\Phi_{\bar{s}m}Q_{sm}^{-1}\beta_m. \tag{A.5}$$

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

$$\begin{aligned} B_M\big(\hat{\tau}_{ridge}(y)\big)_{m+1} = -v\gamma_{\bar{s}}^{T}\big[&\Phi_{\bar{s}m}Q_{sm}^{-1}\beta_m + \frac{1}{\triangle}\Phi_{\bar{s}m}Q_{sm}^{-1}\Phi_{sm}\phi_{s(m+1)}\phi_{s(m+1)}\Phi_{sm}Q_{sm}^{-1}\beta_m \\ &+ \phi_{\bar{s}(m+1)}Q_{21}^{-1}\beta_m + \Phi_{\bar{s}m}Q_{12}^{-1}\beta_{m+1} + \phi_{\bar{s}(m+1)}Q_{22}^{-1}\beta_{m+1}\big]. \end{aligned} \tag{A.6}$$

Appendix B

Bayesian Model Selection

The inverse matrix for $(M+1)$ basis functions under Bayesian setting can be written as $Q^{*-1}_{s(m+1)}$

$$Q^{*-1}_{s(m+1)} = \begin{bmatrix} Q^{*-1}_{11} & Q^{*-1}_{12} \\ Q^{*-1}_{21} & Q^{*-1}_{22} \end{bmatrix}, \tag{B.1}$$

where

$$\begin{aligned}
Q^{*-1}_{11} &= Q^{*-1}_{sm} + \frac{1}{\triangle_1} Q^{*-1}_{sm} \Phi^T_{sm} \phi_{s(m+1)} \phi^T_{s(m+1)} \Phi_{sm} Q^{*-1}_{sm}, \\
Q^{*-1}_{12} &= \frac{1}{\triangle_1} \phi^T_{s(m+1)} \Phi_{sm} Q^{*-1}_{sm} \\
Q^{*-1}_{21} &= \frac{1}{\triangle_1} Q^{*-1}_{sm} \Phi^T_{sm} \phi_{s(m+1)} \\
Q^{*-1}_{22} &= \frac{1}{\triangle_1}.
\end{aligned}$$

Further $\triangle_1 = \frac{\sigma^2}{\sigma^2_{0(m+1)}} + \phi^T_{s(m+1)} \left[I_n - \Phi^T_{sm} Q^{*-1}_{sm} \Phi^T_{sm} \right] \phi_{s(m+1)}$. For a model with $(M+1)$ regressors we can write the bias as follow

$$B_M[\hat{\tau}_B(y)]_{m+1} = \gamma^T_{\bar{s}} \Phi^T_{\bar{s}(m+1)} \Lambda_{s(m+1)} \left(\mu_{0(m+1)} - \beta_{m+1} \right),$$

where

$$\Lambda_{s(m+1)} = \sigma^2 Q^{*-1}_{s(m+1)} \Sigma^{-1}_{0(m+1)} = \sigma^2 \times \begin{bmatrix} Q^{*-1}_{11} \Sigma^{-1}_{0m} & Q^{*-1}_{12} \sigma^{-2}_{0(m+1)} \\ Q^{*-1}_{21} \Sigma^{-1}_{0m} & Q^{*-1}_{22} \sigma^{-2}_{0(m+1)} \end{bmatrix} \tag{B.2}$$

$$
\begin{aligned}
B_M[\hat{\tau}_B(y)]_{m+1} =& \sigma^2\gamma_{\bar{s}}\Phi_{\bar{s}(m+1)} \times \begin{bmatrix} Q_{11}^{*-1}\Sigma_{0m}^{-1} & Q_{12}^{*-1}\sigma_{0(m+1)}^{-2} \\ Q_{21}^{*-1}\Sigma_{0m}^{-1} & Q_{22}^{*-1}\sigma_{0(m+1)}^{-2} \end{bmatrix} \times \left(\mu_{0(m+1)} - \beta_{m+1}\right) \\
=& \sigma^2\left[\left(\Phi_{\bar{s}m}Q_{11}^{*-1} + \phi_{\bar{s}(m+1)}Q_{21}^{*-1}\right)\Sigma_{0m}^{-1}\left(\mu_{0m} - \beta_m\right) + \left(\Phi_{\bar{s}m}Q_{12}^{*-1}\right.\right. \\
& \left.\left. + \phi_{\bar{s}(m+1)}Q_{22}^{*-1}\right)\sigma_{0(m+1)}^{-2}\left(\mu_{0(m+1)} - \beta_{m+1}\right)\right] \\
=& \sigma^2\Phi_{\bar{s}m}Q_{sm}^{*-1}\Sigma_{0m}^{-1}\left(\mu_{0m} - \beta_m\right) + \frac{\sigma^2}{\triangle_1}\left[\left\{\Phi_{\bar{s}m}Q_{sm}^{*-1}\Phi_{sm}^T\phi_{s(m+1)}\phi_{s(m+1)}^T\Phi_{sm}Q_{sm}^{*-1}\right.\right. \\
& \left.\left. + \left\{\Phi_{\bar{s}m}\phi_{s(m+1)}^T\Phi_{sm}Q_{sm}^{*-1} + \phi_{\bar{s}(m+1)}\right\}\sigma_{0(m+1)}^{-2}\left(\mu_{0(m+1)} - \beta_{m+1}\right)\right].
\end{aligned}
$$

The absolute change in bias of estimator $\hat{\tau}_B(y)$ when an additional basis function is added to the model

$$
\begin{aligned}
ADB =& \left|B_M\left[\hat{\tau}_B(y)\right]_m - B_M\left[\hat{\tau}_B(y)\right]_{m+1}\right| \\
=& \frac{\sigma^2}{\triangle_1}\gamma_{\bar{s}}^T\left[\left\{\Phi_{\bar{s}m}Q_{sm}^{*-1}\Phi_{sm}^T\phi_{s(m+1)}\phi_{s(m+1)}^T\Phi_{sm}Q_{sm}^{*-1} - \phi_{\bar{s}(m+1)}Q_{sm}^{*-1}\Phi_{sm}^T\phi_{s(m+1)}\right\}\Sigma_{0m}^{-1}\left(\mu_{0m}\right.\right. \\
& \left. - \beta_m\right) - \left\{\Phi_{\bar{s}m}\phi_{s(m+1)}^T\Phi_{sm}Q_{sm}^{*-1} - \phi_{\bar{s}(m+1)}\right\}\sigma_{0(m+1)}^{-2}\left(\mu_{0(m+1)}\right. \\
& \left.\left. - \beta_{m+1}\right)\right].
\end{aligned}
$$

Now for variance of the prediction error, we have

$$
\begin{aligned}
V_M\left(e(\hat{\tau}_B)\right)_{m+1} = \sigma^2\left[\gamma_{\bar{s}}^T\gamma_{\bar{s}} + \gamma_{\bar{s}}^T\Phi_{\bar{s}(m+1)}\left(I_{M+1} - \Lambda_{s(m+1)}\right)A_{s(m+1)}^{-1}\left(I_M\right.\right. \\
\left.\left. - \Lambda_{s(m+1)}\right)^T\Phi_{(rm+1)}^T\gamma_{\bar{s}}\right].
\end{aligned}
$$

After some simplification and using theorem of matrix inversion, we get

$$
\begin{aligned}
V_M\left(e(\hat{\tau}_B)\right)_{m+1} = \sigma^2\left[\gamma_{\bar{s}}^T\gamma_{\bar{s}} + \gamma_{\bar{s}}^T\left(\Phi_{\bar{s}m}\Lambda_{11}^{**}\Phi_{\bar{s}m}^T + \phi_{\bar{s}(m+1)}\Lambda_{21}^{**}\Phi_{\bar{s}m}^T\right.\right. \\
\left.\left. + \Phi_{\bar{s}m}\Lambda_{12}^{**}\phi_{\bar{s}(m+1)}^T + \phi_{\bar{s}(m+1)}\Lambda_{21}^{**}\phi_{\bar{s}(m+1)}^T\right)\gamma_{\bar{s}}\right],
\end{aligned}
\tag{B.3}
$$

where $\Lambda_{s(m+1)}$ is already defined in (B.2) and the inverse of matrix $A_{s(m+1)}$ is obtained by setting $v = 1$ in equation (A.3).

The product $\left(I_{M+1} - \Lambda_{s(m+1)}\right)A_{s(m+1)}^{-1}\left(I_{M+1} - \Lambda_{s(m+1)}\right)^T = \Lambda^{**}$ (say) is expressed as a symmetric matrix defined by

$$\Lambda^{**} = \begin{bmatrix} \Lambda^{**}_{11} & \Lambda^{**}_{12} \\ \Lambda^{**}_{21} & \Lambda^{**}_{22} \end{bmatrix}, \tag{B.4}$$

where

$$\begin{aligned}
\Lambda^{**}_{11} &= \left(\Lambda^{-1}_{11} A^{-1}_{11} + \Lambda^{-1}_{12} A^{-1}_{21}\right)\Lambda^{-1}_{11} + \left(\Lambda^{-1}_{11} A^{-1}_{12} + \Lambda^{-1}_{12} A^{-1}_{22}\right)\Lambda^{-1}_{12}, \\
\Lambda^{**}_{12} &= \left(\Lambda^{-1}_{11} A^{-1}_{11} + \Lambda^{-1}_{12} A^{-1}_{21}\right)\Lambda^{-1}_{21} + \left(\Lambda^{-1}_{11} A^{-1}_{12} + \Lambda^{-1}_{12} A^{-1}_{22}\right)\Lambda^{-1}_{22} \\
\Lambda^{**}_{21} &= \left(\Lambda^{-1}_{21} A^{-1}_{11} + \Lambda^{-1}_{22} A^{-1}_{21}\right)\Lambda^{-1}_{11} + \left(\Lambda^{-1}_{21} A^{-1}_{12} + \Lambda^{-1}_{22} A^{-1}_{22}\right)\Lambda^{-1}_{12} \\
\Lambda^{**}_{22} &= \left(\Lambda^{-1}_{21} A^{-1}_{11} + \Lambda^{-1}_{22} A^{-1}_{21}\right)\Lambda^{-1}_{21} + \left(\Lambda^{-1}_{21} A^{-1}_{12} + \Lambda^{-1}_{22} A^{-1}_{22}\right)\Lambda^{-1}_{22} \\
\Lambda^{**}_{22} &= \left(\Lambda^{-1}_{21} A^{-1}_{11} + \Lambda^{-1}_{22} A^{-1}_{21}\right)\Lambda^{-1}_{21} + \left(\Lambda^{-1}_{21} A^{-1}_{12} + \Lambda^{-1}_{22} A^{-1}_{22}\right)\Lambda^{-1}_{22}
\end{aligned}$$

with

$$\begin{aligned}
\Lambda^{-1}_{11} &= I_M - \Lambda_{sm}\Lambda^{-1}_{12} & &= -\sigma^2 Q^{*-1}_{sm}\sigma^{-2}_{0(m+1)} \\
\Lambda^{-1}_{21} &= -\sigma^2 Q^{*-1}_{21}\Sigma^{-1}_{0m}, \\
\Lambda^{-1}_{22} &= -\sigma^2 Q^{*-1}_{22}\sigma^{-2}_{0(m+1)}
\end{aligned}$$

and Λ^{-1}_{pq} for $p,q = 1,2$ are the entries of the inverse of matrix $A^{-1}_{s(m+1)}$. Further, the sub-matrix Λ^{**}_{11} can be simplified as

$$\begin{aligned}
\Lambda^{**}_{11} &= \Lambda^{-1}_{11} A^{-1}_{sm}\Lambda^{-1}_{11} + \Lambda^{-1}_{11} \triangle^{-1}_{1} A^{-1}_{sm}\Phi^{T}_{sm}\phi_{s(m+1)}\phi^{T}_{s(m+1)}\Phi_{sm}A^{-1}_{sm}\Lambda^{-1}_{11} \\
&\quad + \Lambda^{-1}_{12} A^{-1}_{21}\Lambda^{-1}_{11} + \left(\Lambda^{-1}_{11} A^{-1}_{12} + \Lambda^{-1}_{12} A^{-1}_{22}\right)\Lambda^{-1}_{12} \\
&= \Lambda^{-1}_{11} A^{-1}_{sm}\Lambda^{-1}_{11} + \Lambda^{-1}_{11} \triangle^{-1}_{1} A^{-1}_{sm}\Phi^{T}_{sm}\phi_{s(m+1)}\phi^{T}_{s(m+1)}\Phi_{sm}A^{-1}_{sm}\Lambda^{-1}_{11} + \Lambda^{-1}_{12} A^{-1}_{21}\Lambda^{-1}_{11} \\
&\quad + \left(\Lambda^{-1}_{11} A^{-1}_{12} + \Lambda^{-1}_{12} A^{-1}_{22}\right)\Lambda^{-1}_{12}
\end{aligned}$$

The decrement in error variance is expressed as a new index IE_B

$$\begin{aligned}
IE_B &= V_M\big(e(\hat{\tau}_B)\big)_m - V_M\big(e(\hat{\tau}_B)\big)_{m+1} \\
&= -\sigma^2\gamma^{T}_{\bar{s}}\Bigg[\Phi_{\bar{s}m}\Bigg(\Lambda^{-1}_{11} \triangle^{-1}_{1} A^{-1}_{sm}\Phi^{T}_{sm}\phi_{s(m+1)}\phi^{T}_{s(m+1)}\Phi_{sm}A^{-1}_{sm}\Lambda^{-1}_{11}
\end{aligned}$$

$$+\Lambda_{12}^{-1}A_{21}^{-1}\Lambda_{11}^{-1}+\left(\Lambda_{11}^{-1}A_{12}^{-1}+\Lambda_{12}^{-1}A_{22}^{-1}\right)\Lambda_{12}^{-1}\Bigg)\Phi_{\bar{s}m}^{T}+\phi_{\bar{s}(m+1)}\Lambda_{21}^{**}\Phi_{\bar{s}m}^{T}$$

$$+\Phi_{\bar{s}m}\Lambda_{12}^{**}\phi_{\bar{s}(m+1)}^{T}+\phi_{\bar{s}(m+1)}\Lambda_{21}^{**}\phi_{\bar{s}(m+1)}^{T}\Bigg]\gamma_{\bar{s}}$$

Appendix C

Model based Estimation in Presence of Non-response

Derivation of Bias and MSE $\hat{t}_{y1}$ without sub-sampling

$$
\begin{aligned}
B_M(\hat{t}_{y1}) =& E_M\big(\hat{t}_{y1} - t_y\big) = E_M\big(W_{s_1}^T y_{s_1} + W_{\bar{s}_1}^T x_{\bar{s}_1}\hat{\beta}_1 + W_2^T x_2\hat{\beta}_1 - t_y\big)\\
=& E_M\big(W_{s_1}^T y_{s_1} + W_{\bar{s}_1}^T x_{\bar{s}_1}\hat{\beta}_1 + W_2^T x_2\hat{\beta}_1 - t_y\big)\\
=& E_M\big(W_{\bar{s}_1}^T x_{\bar{s}_1}\hat{\beta}_1 + W_2^T x_2\hat{\beta}_1 - W_{\bar{s}_1}^T Y_{\bar{s}_1} - W_2^T Y_2\big)\\
=& E_M\big(A\hat{\beta}_1 - W_{\bar{s}_1}^T Y_{\bar{s}_1} - W_2^T Y_2\big)\\
=& E_M\big[A\big(H_{s_1}\big)^{-1} x_{s_1}^T y_{s_1} - W_{\bar{s}_1}^T Y_{\bar{s}_1} - W_2^T Y_2\big]\\
=& A\big(H_{s_1}\big)^{-1} x_{s_1}^T E_M(y_{s_1}|x_{s_1}) - W_{\bar{s}_1}^T E_M(Y_{\bar{s}_1}|x_{\bar{s}_1}) - W_2^T E_M(Y_2|x_2)\\
=& A\big(H_{s_1}\big)^{-1} H_{s_1}\beta_1 - W_{\bar{s}_1}^T x_{\bar{s}_1}^T\beta_1 - W_2^T x_2^T\beta_2\\
=& A\big(H_{s_1}\big)^{-1} H_{s_1}\beta_1 - W_{\bar{s}_1}^T x_{\bar{s}_1}^T\beta_1 - W_2^T x_2^T\beta_2\\
B_M(\hat{t}_{y1}) =& W_2^T x_2\big(\beta_1 - \beta_2\big)
\end{aligned}
$$

where $A = W_{\bar{s}_1}^T x_{\bar{s}_1} + W_2^T x_2$. The model variance of $\hat{t}_{y1}$ is derived as

$$
V_M(\hat{t}_{y1}) = V_M\big(W_{s_1}^T y_{s_1} + W_{\bar{s}_1}^T x_{\bar{s}_1}\hat{\beta}_1 + W_2^T x_2\hat{\beta}_1\big).
$$

Under OLS assumptions, we have $V_M(\hat{\beta}_1) = \sigma_1^2\big(H_{s_1}\big)^{-1}$. Inserting this result, we get

$$
V_M(\hat{t}_{y1}) = \sigma_1^2\big(n_1 + W_{\bar{s}_1}^T x_{\bar{s}_1}\big(H_{s_1}\big)^{-1} x_{\bar{s}_1}^T W_{\bar{s}_1}\big) + \sigma_2^2\big(W_2^T x_2\big(H_{s_1}\big)^{-1} x_2^T W_2\big).
$$

The MSE of $\hat{t}_{y1}$, is given by

$$\begin{aligned} MSE_M(\hat{t}_{y1}) =& \left\{B_M(\hat{t}_{y1})\right\}^2 + V_M(\hat{t}_{y1}) \\ =& \left\{B_M(\hat{t}_{y1})\right\}^2 + \sigma_1^2\left(n_1 + W_{\bar{s}_1}^T x_{\bar{s}_1}\left(H_{s_1}\right)^{-1} x_{\bar{s}_1}^T W_{\bar{s}_1}\right) \\ &+ \sigma_2^2\left(W_2^T x_2\left(H_{s_1}\right)^{-1} x_2^T W_2\right). \end{aligned}$$

Derivation of Bias and MSE of $\hat{t}_{y1}$ with Sub-sampling

$$\begin{aligned} B_M(\hat{t}_y^*) =& E_M\big(W_{s_1}^T y_{s_1} + W_{\bar{s}_1}^T x_{\bar{s}_1} b_r + W_{\acute{s}_2}^T y_{\acute{s}_2} + W_{\acute{\bar{s}}_2}^T x_{\acute{\bar{s}}_2} b\hat{e}ta_2 - W_{s_1}^T y_{s_1} - W_{\bar{s}_1}^T Y_{\bar{s}_1} - W_{\acute{s}_2}^T y_{\acute{s}_2} \\ & - W_{\acute{\bar{s}}_2}^T Y_{\acute{\bar{s}}_2}\big) \\ =& E_M\left(W_{\bar{s}_1}^T x_{\bar{s}_1}\hat{\beta}_1 + W_{\acute{\bar{s}}_2}^T x_{\acute{\bar{s}}_2} b\hat{e}ta_2 - W_{\bar{s}_1}^T Y_{\bar{s}_1} - W_{\acute{\bar{s}}_2}^T Y_{\acute{\bar{s}}_2}\right) \\ =& W_{\bar{s}_1}^T\left[x_{\bar{s}_1} E_M\left(\hat{\beta}_1\right) - E_M\left(Y_{\bar{s}_1}\right)\right] + W_{\acute{\bar{s}}_2}^T\left[x_{\acute{\bar{s}}_2} E_M\left(b\hat{e}ta_2\right) - E_M\left(Y_{\acute{\bar{s}}_2}\right)\right] \\ =& W_{\bar{s}_1}^T\left[x_{\bar{s}_1}\beta_1 - x_{\bar{s}_1}\beta_1\right] + W_{\acute{\bar{s}}_2}^T\left[x_{\acute{\bar{s}}_2}\beta_2 - x_{\acute{\bar{s}}_2}\beta_2\right] = 0. \end{aligned}$$

The variance of the estimator, is given by

$$\begin{aligned} V_M(\hat{t}_y^*) =& V_M\Bigg[W_{s_1}^T x_{s_1}\beta_1 + W_{s_1}^T \varepsilon_{s_1} + W_{\bar{s}_1}^T x_{\bar{s}_1}\left(H_{s_1}\right)^{-1} H_{s_1}\beta_1 + W_{\bar{s}_1}^T x_{\bar{s}_1}\left(H_{s_1}\right)^{-1} x_{s_1}^T \varepsilon_{s_1} \\ & + W_{\acute{s}_2}^T x_{\acute{s}_2}\beta_2 + W_{\acute{s}_2}^T \varepsilon_{\acute{s}_2} + W_{\acute{\bar{s}}_2}^T x_{\acute{\bar{s}}_2}\left(H_{\acute{s}_2}\right)^{-1} H_{\acute{s}_2}\beta_2 + W_{\acute{\bar{s}}_2}^T x_{\acute{\bar{s}}_2}\left(H_{\acute{s}_2}\right)^{-1} x_{\acute{s}_2}^T \varepsilon_{\acute{s}_2}\Bigg] \end{aligned}$$

$$V_M(\hat{t}_y^*) = \sigma_1^2\left[n_1 + W_{\bar{s}_1}^T x_{\bar{s}_1}\left(H_{s_1}\right)^{-1} x_{\bar{s}_1}^T W_{\bar{s}_1}\right] + \sigma_2^2\left[\acute{n}_2 + W_{\acute{\bar{s}}_2}^T x_{\acute{\bar{s}}_2}\left(H_{\acute{s}_2}\right)^{-1} x_{\acute{\bar{s}}_2}^T W_{\acute{\bar{s}}_2}\right].$$

Rearranging terms, we get

$$V_M(\hat{t}_y^*) = n_1\sigma_1^2 + \acute{n}_2\sigma_2^2 + \sigma_1^2 W_{\bar{s}_1}^T x_{\bar{s}_1}\left(H_{s_1}\right)^{-1} x_{\bar{s}_1}^T W_{\bar{s}_1} + \sigma_2^2 W_{\acute{\bar{s}}_2}^T x_{\acute{\bar{s}}_2}\left(H_{\acute{s}_2}\right)^{-1} x_{\acute{\bar{s}}_2}^T W_{\acute{\bar{s}}_2}$$

Table C.1: Bias and MSEs with $\sigma^2 = 0.01$ and $p = 8$

			$\rho = 0.5$			$\rho = 0.7$			$\rho = 0.9$		
λ_2	n	k	$Bias(\hat{T}^*_{yv})$	$MSE(\hat{t}^*_y)$	$MSE(\hat{T}^*_{yv})$	$Bias(\hat{T}^*_{yv}$	$MSE(\hat{t}^*_y)$	$MSE(\hat{T}^*_{yv})$	$Bias(\hat{T}^*_{yv)}$	$MSE(\hat{t}^*_y)$	$MSE(\hat{T}^*_{yv})$
0.2	100	1.5	-1.9045	2632.5429	3410.3629	0.1861	2501.6722	3101.5096	-3.5464	3615.4525	4146.0974
		2	-3.0774	15332.5627	18264.6392	-0.9490	14016.4277	16670.444	0.8504	22387.5100	24270.7100
		3	-4.7323	63897.550	67396.2853	-6.8877	17623.5750	18316.748	-6.1958	102069.3970	105702.5531
	150	1.5	-0.1320	182.0770	181.8869	-0.0860	177.5572	177.5734	-0.0015	0.0115	0.0004
		2	-0.1403	237.1414	238.3950	-0.2664	233.9038	233.6031	0.8081	200.9532	283.5728
		3	0.1429	998.3503	1041.6369	-4.7654	11861.0556	15667.4757	-4.5923	12698.1924	14579.8670
	200	1.5	-0.2404	124.6444	124.7199	0.0072	0.0472	0.1311	-0.0021	0.0776	0.0007
		2	-0.2501	141.0534	140.9511	0.0491	13.0456	30.5921	-0.0023	0.0682	0.0006
		3	-0.2724	225.3042	230.9373	-2.1225	756.4574	1453.1401	-1.2813	556.3236	888.3302
0.2	100	1.5	-0.0017	0.1004	0.0005	-0.5134	309.4638	309.6424	-0.0082	285.7515	286.0365
		2	-0.1018	2.0960	16.3657	-0.5683	399.1314	399.0764	-0.3330	368.5770	369.1162
		3	-1.3227	1081.2742	5169.5272	-0.1887	703.9439	703.9108	-0.3322	681.8513	686.7257
	150	1.5	-0.0010	0.0189	0.0002	-0.3027	176.2154	176.2077	-0.3701	169.2124	169.2489
		2	-0.0015	0.0676	0.0003	-0.2734	222.5153	222.6088	-0.1682	203.0450	203.0715
		3	-0.0023	0.4645	0.0010	-0.5970	336.7227	337.0286	-0.7045	315.0578	314.8267
	200	1.5	-0.0003	0.0057	0.0001	-0.0077	136.3723	136.3636	-0.1989	124.4898	124.5128
		2	-0.0010	0.0167	0.0002	0.0665	166.4728	166.5163	-0.1507	151.5523	151.5717
		3	-0.0016	0.1027	0.0004	0.3372	237.1700	237.1535	-0.0540	217.1780	217.2875

Table C.2: Bias and MSEs with $\sigma^2 = 0.1$ and $p = 8$

			$\rho = 0.5$			$\rho = 0.7$			$\rho = 0.9$		
λ_2	n	k	$Bias(\hat{T}^*_{yv})$	$MSE(\hat{t}^*_y)$	$MSE(\hat{T}^*_{yv})$	$Bias(\hat{T}^*_{yv}$	$MSE(\hat{t}^*_y)$	$MSE(\hat{T}^*_{yv})$	$Bias(\hat{T}^*_{yv)}$	$MSE(\hat{t}^*_y)$	$MSE(\hat{T}^*_{yv})$
0.2	100	1.5	0.0527	761.5122	6.0175	-0.1304	712.9594	32.4517	0.1087	779.5891	101.4044
		2	0.3897	3072.4058	597.1933	-0.1384	2957.2573	884.0712	0.3682	3096.5782	960.3027
		3	-4.1239	18395.7554	12510.7762	-2.1398	18856.2093	13943.7068	0.4578	17867.0923	16000.5730
	150	1.5	-0.0018	102.4919	0.0004	-0.0013	93.5663	0.0004	-0.0018	98.2044	0.0004
		2	-0.0031	320.7221	0.0007	-0.0021	307.4636	0.0007	-0.0031	312.4716	0.0007
		3	-0.0853	2119.3799	393.5367	-0.1952	1936.9254	120.3755	0.2440	2012.8939	584.2767
	200	1.5	-0.0010	31.2990	0.0002	-0.0007	27.9423	0.0003	-0.0006	30.3235	0.0003
		2	-0.0019	103.3161	0.0003	-0.0014	83.5846	0.0003	-0.0012	96.0962	0.0003
		3	-0.0032	501.3613	0.0008	-0.0027	543.5329	0.0007	-0.0024	486.4563	0.0008
0.4	100	1.5	-0.0012	34782.8600	0.0002	-0.0024	35656.4700	0.0002	-0.0013	36537.1700	0.0002
		2	-0.0022	76289.3900	0.0003	-0.0031	73245.7900	0.0003	-0.0032	335.3555	0.0008
		3	-0.0042	202559.5000	0.0012	-0.0052	184955.9000	0.0009	-0.0069	1811.6840	0.0031
	150	1.5	-0.0055	272320.5567	0.0016	-0.0064	247252.1500	0.0012	-0.0094	-21830.7495	0.0043
		2	-0.0070	356208.8767	0.0021	-0.0078	321901.8650	0.0015	-0.0122	-39193.4925	0.0058
		3	-0.0085	440097.1967	0.0027	-0.0092	396551.5800	0.0019	-0.0150	-56556.2355	0.0073
	200	1.5	-0.0100	523985.5167	0.0032	-0.0106	471201.2950	0.0022	-0.0178	-73918.9785	0.0087
		2	-0.0115	607873.8367	0.0037	-0.0120	545851.0100	0.0026	-0.0207	-91281.7215	0.0102
		3	-0.0130	691762.1567	0.0043	-0.0134	620500.7250	0.0029	-0.0235	-108644.4645	0.0117

Table C.3: Bias and MSEs with $\sigma^2 = 1$ and $p = 8$

λ_2	n	k	$Bias(\hat{T}^*_{yv})$	$MSE(\hat{t}^*_y)$	$MSE(\hat{T}^*_{yv})$	$Bias(\hat{T}^*_{yv}$	$MSE(\hat{t}^*_y)$	$MSE(\hat{T}^*_{yv})$	$Bias(\hat{T}^*_{yv)}$	$MSE(\hat{t}^*_y)$	$MSE(\hat{T}^*_{yv})$
0.2		1.5	-0.92480	61377.38088	2732.37605	-1.18547	64137.80092	2116.32394	-1.18547	64137.80092	2116.32394
	100	2	-2.94078	98309.64742	27170.99000	-8.24624	98661.25065	25524.12997	-8.24624	98661.25065	25524.12991
		3	-27.35530	167858.93049	198667.30000	-23.36264	156272.98600	200505.09000	-23.36264	156272.98600	200505.09930
		1.5	-0.00095	32731.19000	0.00010	-0.00101	31548.83000	0.00007	-0.00101	31548.83000	0.00007
	150	2	-0.00206	53798.86000	0.00057	0.66527	50745.12235	273.95232	0.66527	50745.12235	273.95232
		3	-5.45773	98828.28093	18837.24500	-5.31764	95632.41642	20115.99933	-5.31764	95632.41642	20115.99933
		1.5	-0.00050	17985.20000	0.00003	-0.00086	18345.49000	0.00003	-0.00086	18345.49000	0.00003
	200	2	-0.00073	30170.24000	0.00007	-0.00107	32707.66000	0.00007	-0.00107	32707.66000	0.00007
		3	-0.87860	62423.15344	236.95346	-0.58426	63258.35576	189.24416	-0.58426	63258.35576	189.24416
0.4	100	1.5	-0.00054	74174.77000	0.00005	-0.00054	73609.55000	0.00008	-0.26906	296990.10000	186.24990
		2	-0.00123	137269.40000	0.00018	-0.00114	136571.00000	0.00019	-0.00118	142043.00000	0.00012
		3	0.21696	304271.30000	1655.51200	-0.93803	292194.00000	877.38890	-0.26906	296990.10000	186.24990
	150	1.5	-0.00016	27143.28000	0.00002	-0.00049	27164.02000	0.00002	-0.00094	133416.30000	0.00011
		2	-0.00034	54688.47000	0.00004	-0.00062	57434.25000	0.00004	-0.00060	55799.30000	0.00004
		3	-0.00097	138008.00000	0.00014	-0.00116	144321.40000	0.00014	-0.00094	133416.30000	0.00011
	200	1.5	-0.00019	13010.95000	0.00001	-0.00030	12975.12000	0.00001	-0.00070	72677.00000	0.00005
		2	-0.00021	27447.29000	0.00002	-0.00040	27325.85000	0.00002	-0.00039	27794.85000	0.00002
		3	-0.00058	73274.17000	0.00005	-0.00077	75598.08000	0.00006	-0.00070	72677.00000	0.00005

Appendix D

Model-based estimation under RSSWOR

Table D.1: Bias and RE for $\gamma^* = 0.3$

t	m	RE_r	RE_{rss}	$RE_{R.rss}$	AB.srs	AB.rss
			G(2, 2)			
	2	2.7477	1.0162	2.4882	60.4370	74.3182
5	5	8.4049	1.0416	7.8941	13.3828	28.1902
	8	14.2844	1.0674	12.5380	3.6897	15.7625
	2	6.4891	1.0371	6.0193	16.6093	37.0989
10	5	18.7730	1.0968	15.9923	14.8950	12.1114
	8	28.1957	1.1624	25.2679	40.2373	8.4270
			G(2, 3)			
	2	1.9541	1.0162	1.7817	39.3351	49.5385
5	5	5.9852	1.0416	5.5614	9.2946	19.1030
	8	10.0640	1.0673	8.7772	1.7740	10.4089
	2	4.5732	1.0371	4.2278	10.4029	24.9263
10	5	13.3438	1.0967	11.1148	9.1333	8.4639
	8	20.4853	1.1623	17.5920	25.9294	6.0248
			G(2, 6)			
	2	0.7002	1.0162	0.6532	18.2332	24.7587
5	5	2.1652	1.0416	1.9770	5.2064	10.0158
	8	3.5829	1.0665	3.1030	0.1417	5.0553
	2	1.6207	1.0370	1.5014	4.1965	12.7538
10	5	4.7780	1.0962	3.8859	3.3716	4.8163
	8	7.6737	1.1615	6.1247	11.6216	3.6227

Table D.2: Bias and RE for $\gamma^{*} = 0.3$ (Continued)

t	m	RE_r	RE_{rss}	$RE_{R.rss}$	AB.srs	AB.rss
			G(4, 2)			
	2	6.6740	1.0162	6.5199	70.2027	71.4624
5	5	18.6256	1.0414	17.5572	3.7536	26.4852
	8	30.0509	1.0678	28.2431	32.0967	20.8870
	2	14.6985	1.0371	13.7490	14.6898	36.9738
10	5	37.2922	1.0969	34.4925	50.2069	13.9068
	8	52.2661	1.1625	54.6377	96.6559	8.0505
			G(4, 3)			
	2	5.4981	1.0162	5.3625	46.4492	47.4090
5	5	15.4862	1.0413	14.5785	2.2278	16.6957
	8	24.8446	1.0677	23.5442	22.2571	13.2721
	2	12.2151	1.0371	11.3978	9.7364	24.3884
10	5	31.2106	1.0969	28.5025	33.9727	8.7211
	8	45.2124	1.1624	44.7444	64.8684	4.4050
				G(4, 6)		
	2	2.8007	1.0162	2.7367	22.6957	23.3556
5	5	8.0255	1.0412	7.5142	0.7021	6.9062
	8	12.8020	1.0676	12.1435	12.4175	5.6571
10	2	6.3329	1.0371	5.8753	4.7830	11.8029
	5	16.5931	1.0968	14.5284	17.7384	3.5353
	8	26.1513	1.1619	22.3871	33.0809	0.7596

Table D.3: Bias and RE for $\gamma^* = 0.5$

t	m	RE_r	RE_{rss}	$RE_{R.rss}$	B.srs	B.rss
			G(2, 2)			
5	2	4.1635	1.0162	3.8755	26.3349	38.4283
	5	11.7600	1.0416	10.8376	0.2269	14.6806
	8	19.0647	1.0674	16.6432	12.6135	7.0302
10	2	9.0123	1.0371	8.3368	0.3577	19.2577
	5	24.6040	1.0968	20.8750	22.0938	5.2963
	8	34.8054	1.1624	32.5903	45.1165	3.7419
			G(2,3)			
5	2	2.4424	1.0162	2.2918	16.6708	25.5938
	5	6.9412	1.0416	6.3435	0.4148	9.9228
	8	11.2630	1.0673	9.7110	8.0589	4.2787
10	2	5.2569	1.0371	4.8709	0.7021	12.7774
	5	14.6075	1.0967	12.1113	14.1533	3.5794
	8	21.8860	1.1623	19.0112	29.5143	2.5279
			G(2,6)			
5	2	0.6876	1.0162	0.6540	7.0068	12.7592
	5	1.9717	1.0416	1.7872	0.6028	5.1650
	8	3.2106	1.0665	2.7341	3.5042	1.5272
10	2	1.4733	1.0370	1.3731	1.7620	6.2971
	5	4.1606	1.0962	3.3897	6.2128	1.8626
	8	6.6925	1.1615	5.3228	13.9122	1.3139
			G(4,2)			
5	2	11.6353	1.0162	11.3209	30.6179	36.0369
	5	30.9743	1.0414	29.1134	10.1908	12.5094
	8	47.6155	1.0678	46.2887	39.1973	11.9721
10	2	24.7812	1.0371	23.0511	3.4527	20.0803
	5	57.9638	1.0969	55.5549	56.0029	7.7684
	8	72.2816	1.1625	85.5164	98.7919	4.2145

Table D.4: Bias and RE for $\gamma^* = 0.5$ (Continued)

t	m	RE_r	RE_{rss}	$RE_{R.rss}$	B.srs	B.rss
			G(4, 3)			
	2	8.2487	1.0162	8.0154	20.4934	24.2274
5	5	22.1472	1.0413	20.7583	6.4045	7.8630
	8	34.1948	1.0677	33.1922	26.5797	8.0502
	2	17.6948	1.0371	16.4005	1.6694	13.8710
10	5	42.7483	1.0969	39.5368	37.2152	5.1830
	8	57.7674	1.1624	60.4255	65.7651	2.3580
			G(4,6)			
	2	3.2104	1.0162	3.1190	10.3689	12.4179
5	5	8.7104	1.0412	8.1194	2.6182	3.2166
	8	13.6275	1.0676	13.0141	13.9622	4.1283
	2	6.9433	1.0371	6.4016	0.1139	7.6617
10	5	17.7594	1.0968	15.3926	18.4274	2.5975
	8	27.7148	1.1619	23.3248	32.7382	0.5016

Table D.5: Bias and RE for $\gamma^* = 0.8$

t	m	RE_r	RE_{rss}	$RE_{R.rss}$	B.srs	B.rss
			G(2, 2)			
	2	2.9772	1.0162	2.9319	0.7295	11.7680
5	5	7.8813	1.0416	7.3193	6.4027	6.8744
	8	13.0466	1.0674	11.2503	17.6501	2.4392
	2	6.1334	1.0371	5.7045	9.8893	6.0968
10	5	16.3307	1.0968	13.8569	25.1742	3.3341
	8	24.7321	1.1624	22.0765	46.4994	2.8797
				G(2, 3)		
	2	1.3124	1.0162	1.2942	0.7727	9.2188
5	5	3.4763	1.0416	3.2281	2.4282	5.9425
	8	5.8088	1.0673	4.9801	10.6486	1.9791
	2	2.7028	1.0371	2.5220	6.4683	4.7646
10	5	7.3178	1.0967	6.1311	15.0885	3.1778
	8	11.7379	1.1623	9.8098	29.6435	2.7129
				G(2, 6)		
	2	0.2952	1.0162	0.2914	0.8158	6.6696
5	5	0.7815	1.0416	0.7259	1.5462	5.0106
	8	1.3146	1.0665	1.1239	3.6471	1.5191
	2	0.6073	1.0370	0.5688	3.0473	3.4325
10	5	1.6602	1.0962	1.3828	5.0029	3.0215
	8	2.7646	1.1615	2.2174	12.7877	2.5461
				G(4, 2)		
	2	19.0692	1.0162	18.5281	7.1291	4.0326
5	5	48.3596	1.0414	45.5339	24.8539	2.3440
	8	70.4115	1.0678	72.8449	48.5345	1.7598
	2	38.7544	1.0371	36.2560	21.4485	4.0871
10	5	83.1682	1.0969	85.3838	63.7828	0.5404
	8	90.1236	1.1625	127.4162	103.5377	2.0718

Table D.6: Bias and RE for $\gamma^* = 0.8$ (Continued)

t	m	RE_r	RE_{rss}	$RE_{R.rss}$	B.srs	B.rss
				G(4, 3)		
	2	9.4384	1.0162	9.1597	5.7765	2.1612
5	5	24.1982	1.0413	22.6224	16.4848	2.9508
	8	36.5374	1.0677	36.3521	33.6606	0.9746
	2	19.2451	1.0371	17.9593	14.0669	3.0742
10	5	45.4154	1.0969	42.6479	42.8306	0.9760
	8	59.2750	1.1624	64.0248	69.4480	2.4260
				G(4, 6)		
	2	2.5469	1.0162	2.4673	4.4240	0.2897
5	5	6.5779	1.0412	6.1133	8.1157	3.5577
	8	10.2535	1.0676	9.8319	18.7867	0.1893
	2	5.1960	1.0371	4.8419	6.6852	2.0612
10	5	13.2630	1.0967	11.5543	21.8784	1.4116
	8	20.9691	1.1619	17.4367	35.3583	2.7801

Appendix E

Application to PDHS 2017-18 Data

The results for fertility measure obtained using DHS.rates Package are provided. This appendix further provides auto-correlation and trace plots for posterior estimates under MCMC given in Chapter 7. Fertility rate obtained using DHS.rate package for PDHS 2017-18 data are given by Masset (2016).

Table E.1: ASFR for Pakistan using DHS.rates package

Group	AGE	ASFR	SE	N	WN	DEFT	RSE	LCI	UCI
0	15-19	190.258	9.084	3317	3083	1.341	0.048	172.454	208.062
1	20-24	282.748	7.752	6584	6652	1.499	0.027	267.556	297.941
2	25-29	255.235	6.746	7731	7784	1.432	0.026	242.012	268.457
3	30-34	171.735	6.714	6818	7011	1.491	0.039	158.576	184.894
4	35-39	82.322	5.371	5836	5680	1.467	0.065	71.796	92.848
5	40-44	28.854	4.198	4215	4135	1.513	0.145	20.626	37.083
6	45-49	12.226	3.173	2370	2562	1.287	0.26	6.007	18.445

Table E.2: Regional ASFR for Pakistan using DHS.rates package

	AGE	ASFR	SE	N	WN	DEFT	RSE	LCI	UCI
ASFR for Punjab									
0	15-19	189.538	15.602	660	1323	1.045	0.082	158.959	220.118
1	20-24	282.972	12.621	1807	3643	1.239	0.045	258.234	307.709
2	25-29	258.421	10.29	2137	4251	1.125	0.04	238.254	278.588
3	30-34	170.164	10.32	1980	3817	1.204	0.061	149.937	190.391
4	35-39	69.976	7.843	1553	2985	1.191	0.112	54.604	85.347
5	40-44	16.717	5.294	1253	2337	1.468	0.317	6.341	27.093
6	45-49	6.874	3.96	778	1478	1.349	0.576	0	14.635
ASFR for Sindh									
0	15-19	184.087	15.823	735	739	1.117	0.086	153.074	215.101
1	20-24	278.917	11.171	1438	1523	1.052	0.04	257.021	300.812
2	25-29	239.543	12.888	1664	1740	1.326	0.054	214.283	264.803
3	30-34	170.189	12.627	1565	1681	1.395	0.074	145.441	194.938
4	35-39	87.859	11.179	1269	1293	1.392	0.127	65.949	109.769
5	40-44	47.36	10.644	888	900	1.316	0.225	26.498	68.222
6	45-49	17.753	6.75	613	632	1.281	0.38	4.522	30.983
ASFR for KPK									
0	15-19	183.969	17.233	806	701	1.3	0.094	150.194	217.744
1	20-24	292.882	14.141	1271	989	1.186	0.048	265.166	320.599
2	25-29	273.131	14.255	1484	1184	1.337	0.052	245.191	301.07
3	30-34	175.322	13.484	1244	1000	1.284	0.077	148.893	201.751
4	35-39	95.224	12.975	1135	889	1.476	0.136	69.793	120.655
5	40-44	37.885	11.343	791	599	1.542	0.299	15.652	60.118
6	45-49	11.137	8.133	351	291	1.155	0.73	0	27.076
ASFR for Balocahistan									
0	15-19	217.982	37.111	545	211	2.181	0.17	145.245	290.72
1	20-24	250.993	22.41	914	306	1.754	0.089	207.069	294.916
2	25-29	222.037	13.206	1126	400	1.079	0.059	196.153	247.921
3	30-34	187.286	14.884	818	318	1.114	0.079	158.113	216.458
4	35-39	119.357	13.9	929	367	1.3	0.116	92.114	146.601
5	40-44	46.102	12.511	496	180	1.349	0.271	21.581	70.623
6	45-49	51.247	25.528	300	118	1.448	0.498	1.212	101.281
ASFR ICT									
0	15-19	219.308	36.448	171	15	1.106	0.166	147.871	290.745
1	20-24	236.892	16.109	527	50	0.971	0.068	205.318	268.465
2	25-29	285.973	20.229	698	69	1.314	0.071	246.326	325.621
3	30-34	140.04	12.252	656	62	0.96	0.087	116.026	164.055
4	35-39	71.55	10.199	563	54	0.921	0.143	51.562	91.539
5	40-44	22.499	8.245	499	50	1.266	0.366	6.339	38.659
6	45-49	9.365	9.374	208	20	1.426	1.001	0	27.738

Table E.3: Regional ASFR for Pakistan using DHS.rates Package (Continued)

	AGE	ASFR	SE	N	WN	DEFT	RSE	LCI	UCI
				ASFR for FATA					
0	15-19	228.822	22.686	400	95	1.194	0.099	184.358	273.286
1	20-24	332.567	25.344	625	141	1.634	0.076	282.893	382.24
2	25-29	281.624	21.473	622	141	1.292	0.076	239.538	323.71
3	30-34	187.078	26.03	555	132	1.581	0.139	136.06	238.097
4	35-39	138.735	32.019	387	93	1.691	0.231	75.978	201.491
5	40-44	79.84	28.944	289	69	1.634	0.363	23.112	136.569
6	45-49	20.916	13.263	120	23	1.024	0.634	0	46.911

Table E.4: TFR at regional level

	TFR	N	WN
Punjab	4.973	10169	19834
Sindh	5.129	8171	8507
KPK	5.348	7082	5653
Balochistan	5.475	5127	1899
ICT	4.928	3323	321
FATA	6.348	2999	693
Pakistan:	5.117	36871	36907

Table E.5: Regional level GFR for Pakistan using DHS.rates package

	GFR	SE	N	LCI	UCI
Punjab	179.112	5.206	9391	168.908	189.317
Sindh	181.711	6.74	7558	168.501	194.921
KPK	191.7	7.325	6731	177.343	206.057
Balochistan	184.783	9.966	4826	165.249	204.317
ICT	162.631	6.485	3115	149.92	175.341
FATA	226.402	14.097	2878	198.772	254.033
Pakistan:	182.745	3.43	34500	176.023	189.467

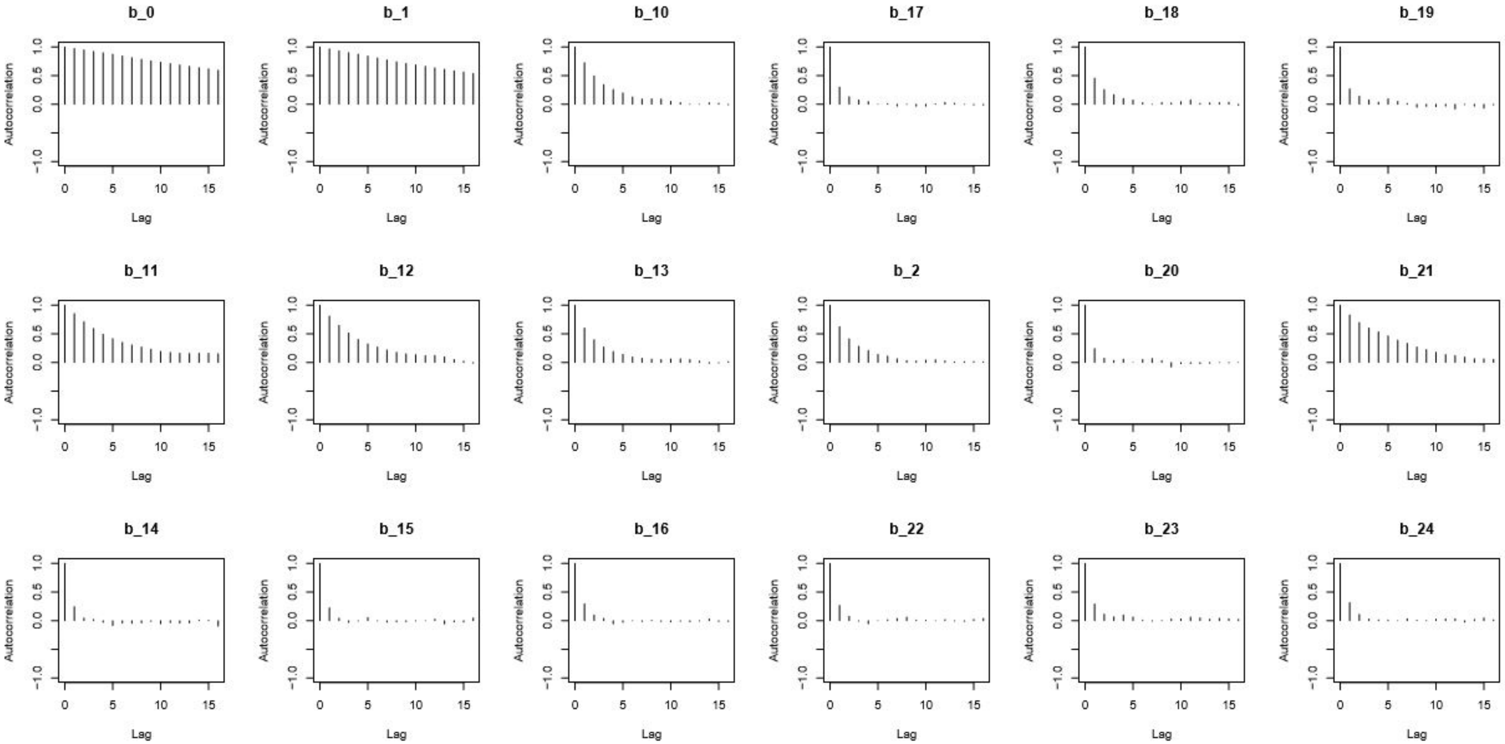

Figure E.1: Auto correlation plot 1

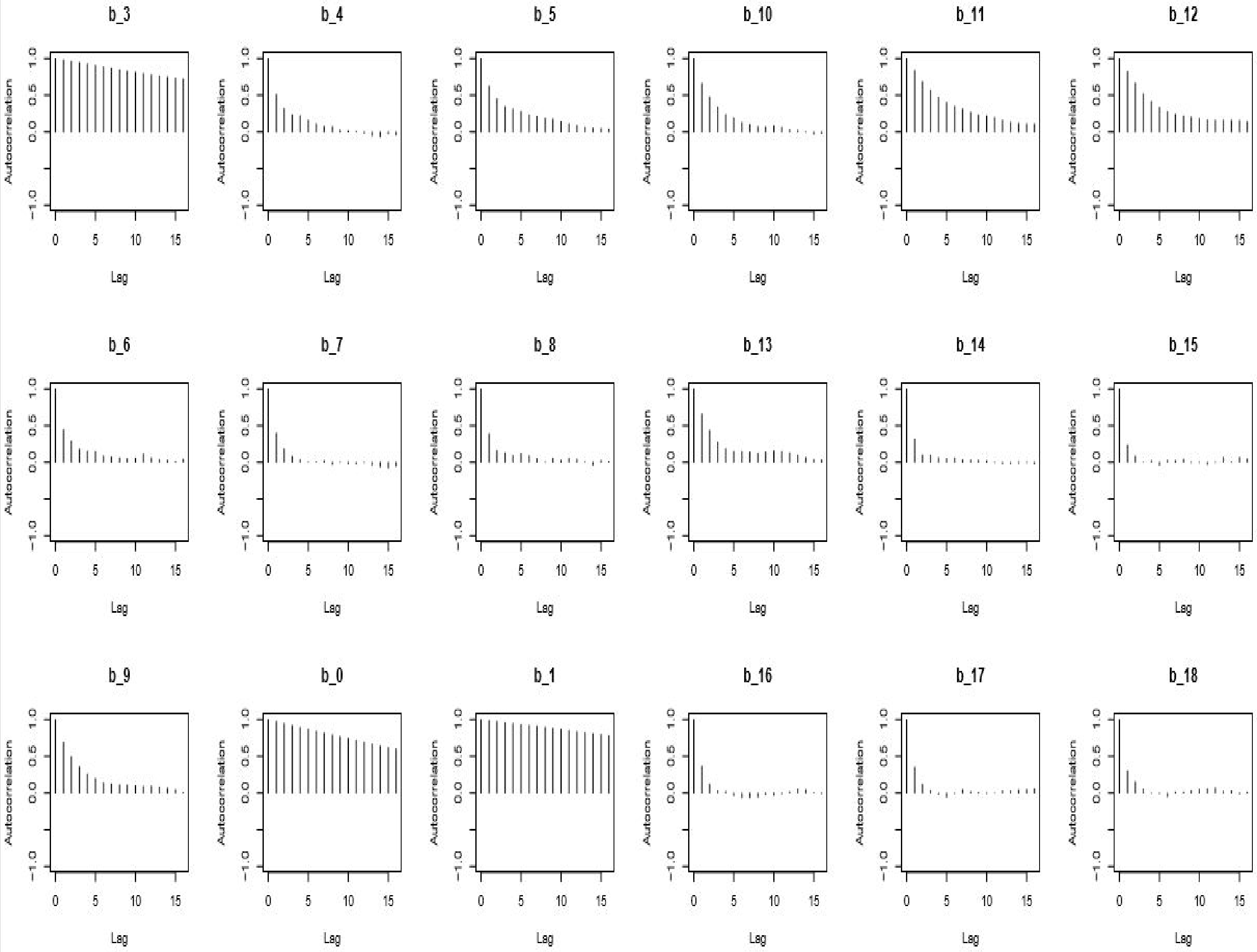

Figure E.2: Auto correlation plot 2

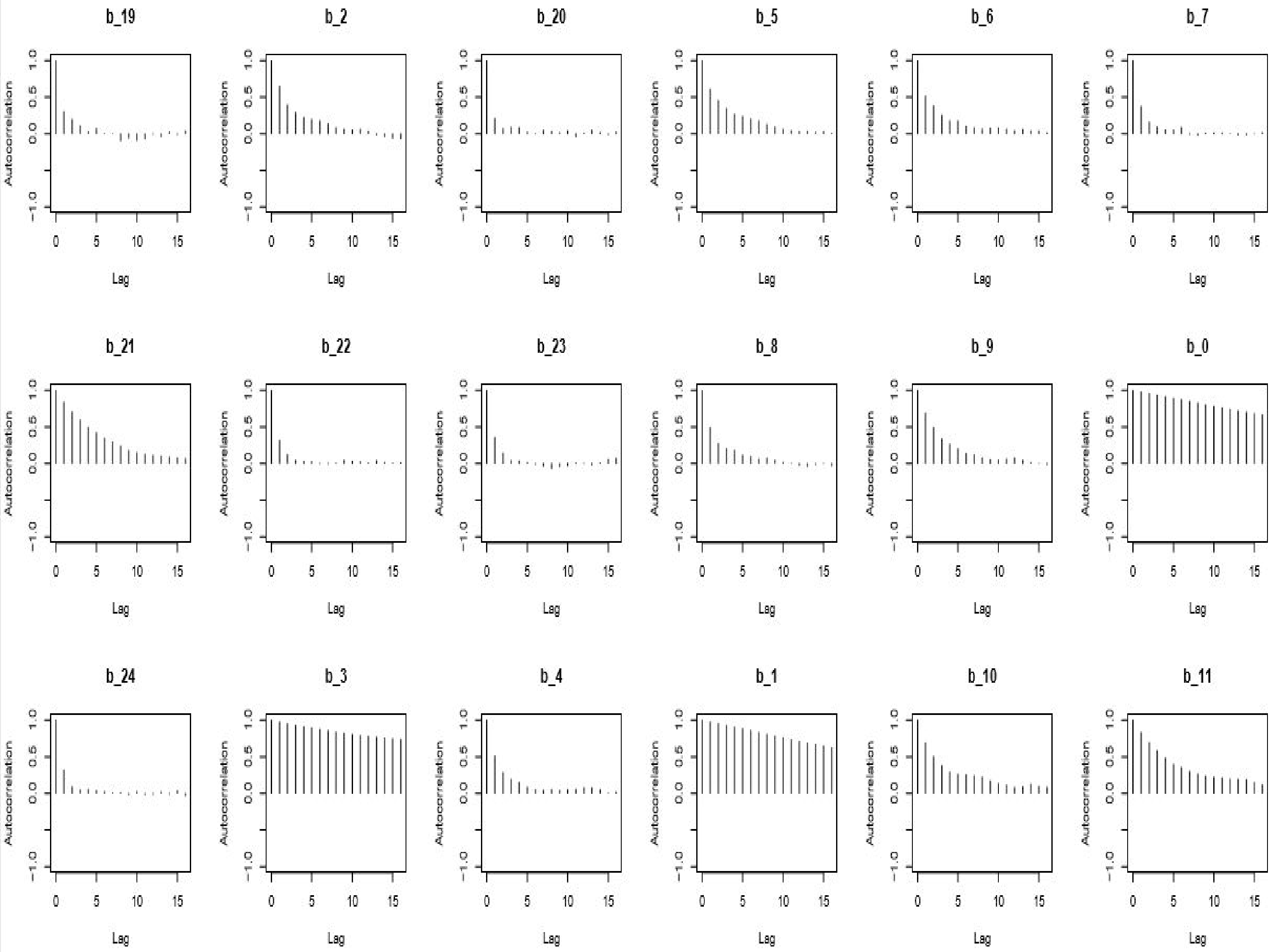

Figure E.3: Auto correlation plot 3

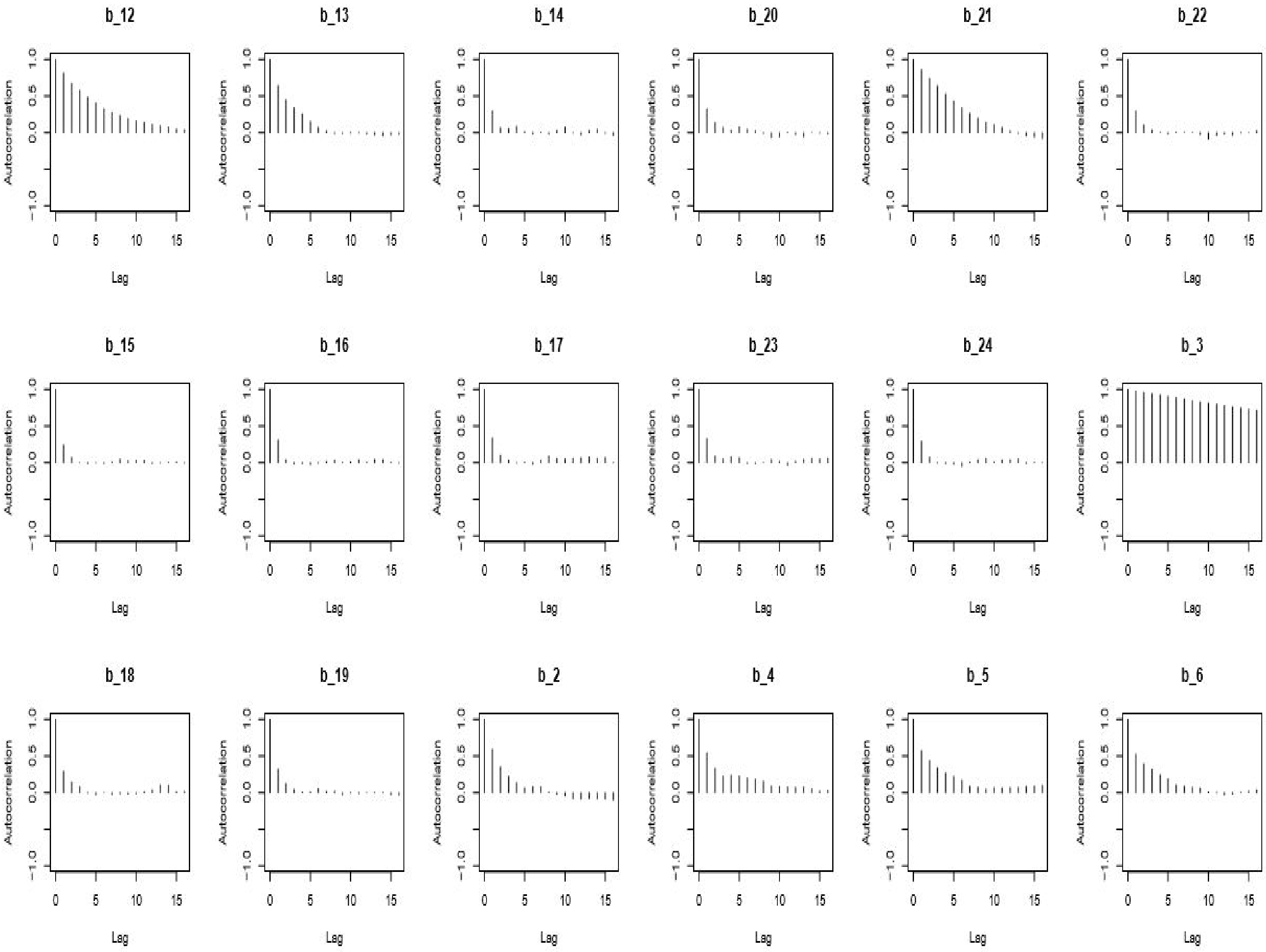

Figure E.4: Auto correlation plot 4

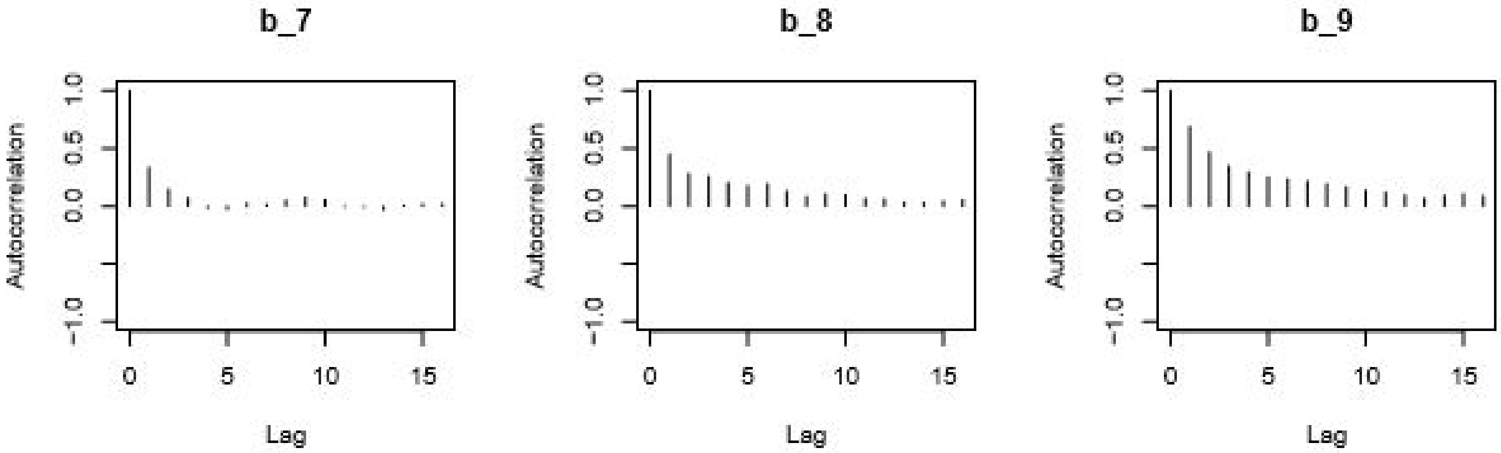

Figure E.5: Auto correlation plot 5

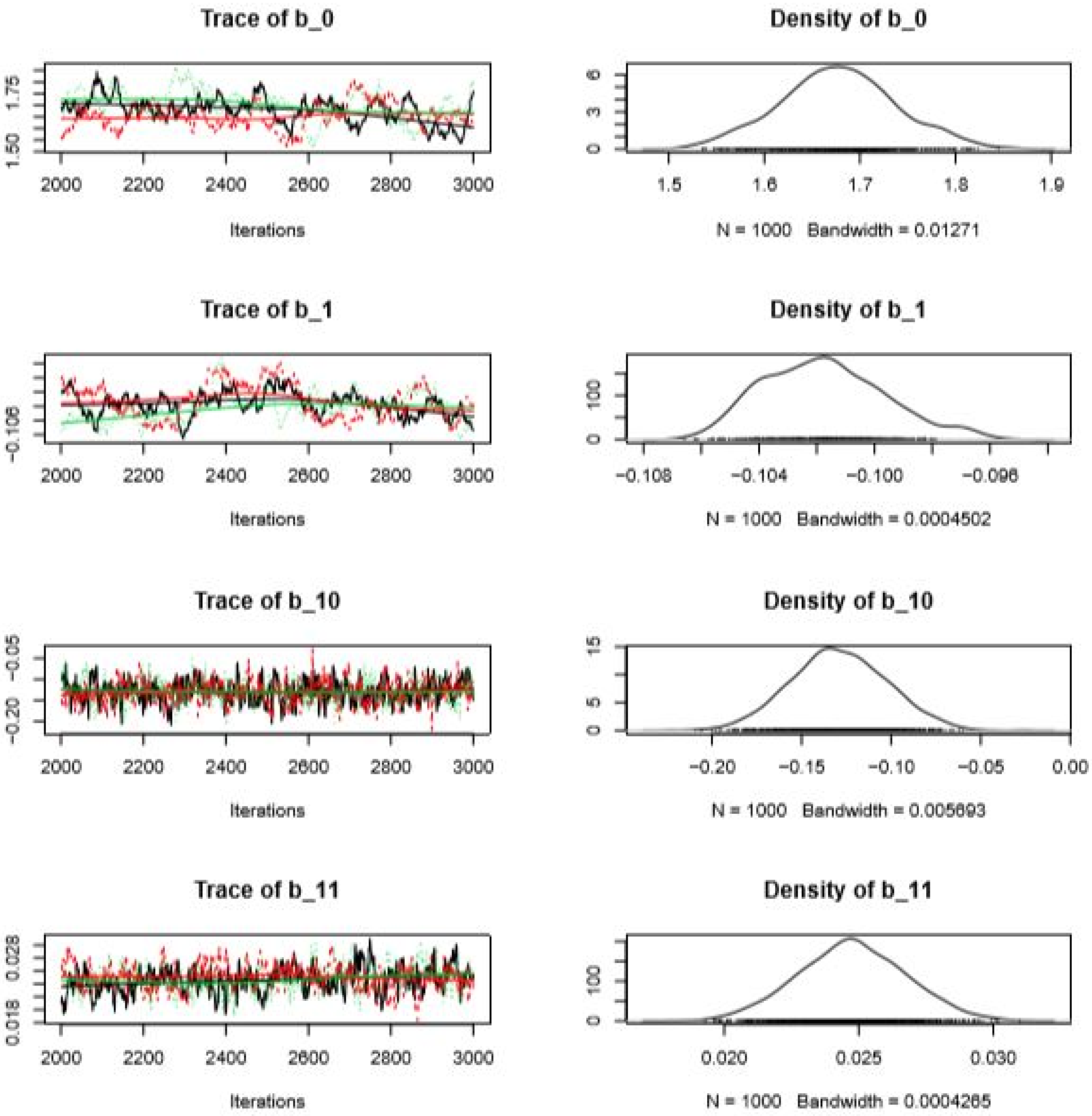

Figure E.6: Trace plot 1-A

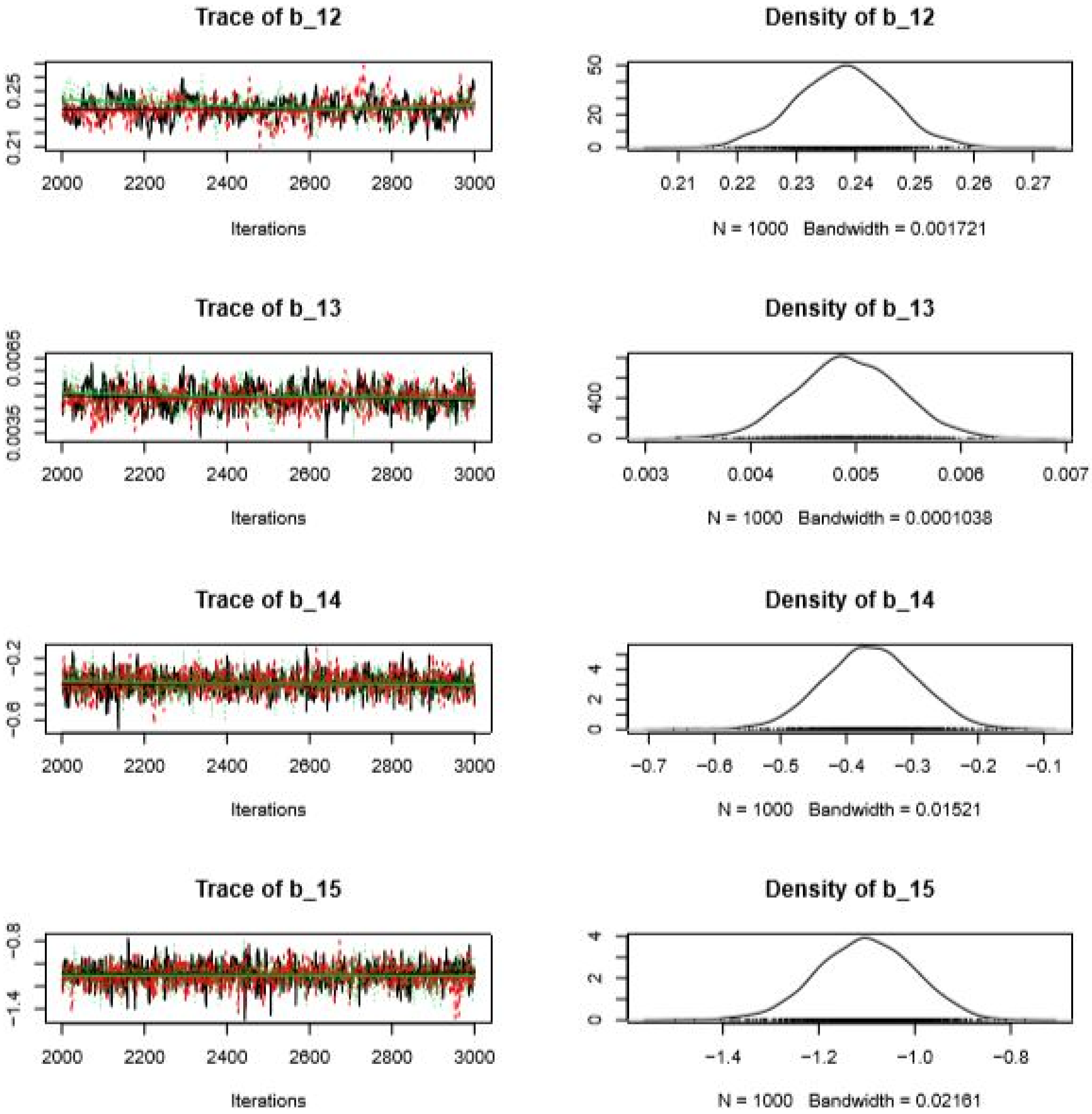

Figure E.7: Trace plot 1-B

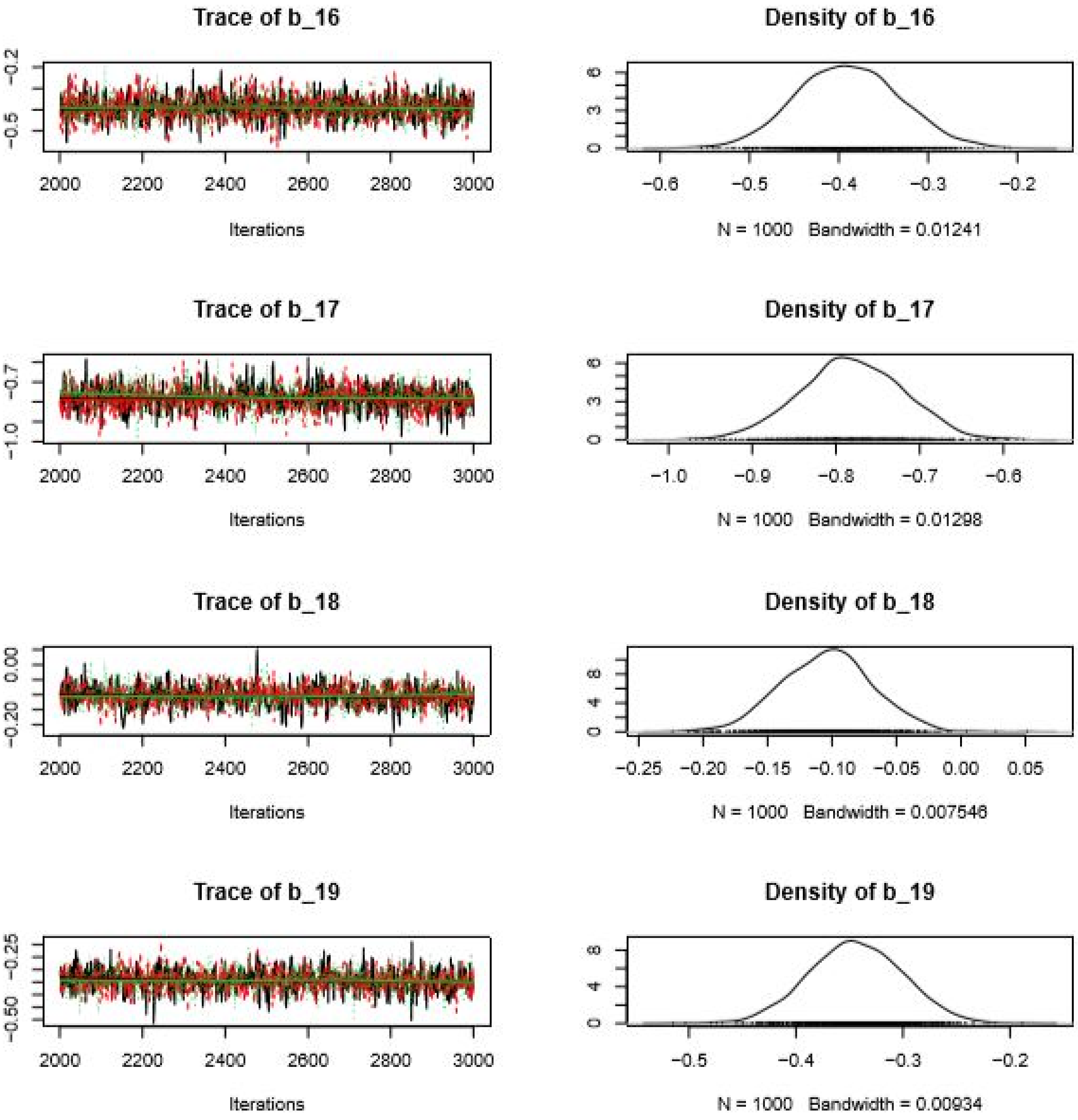

Figure E.8: Trace plot 1-C

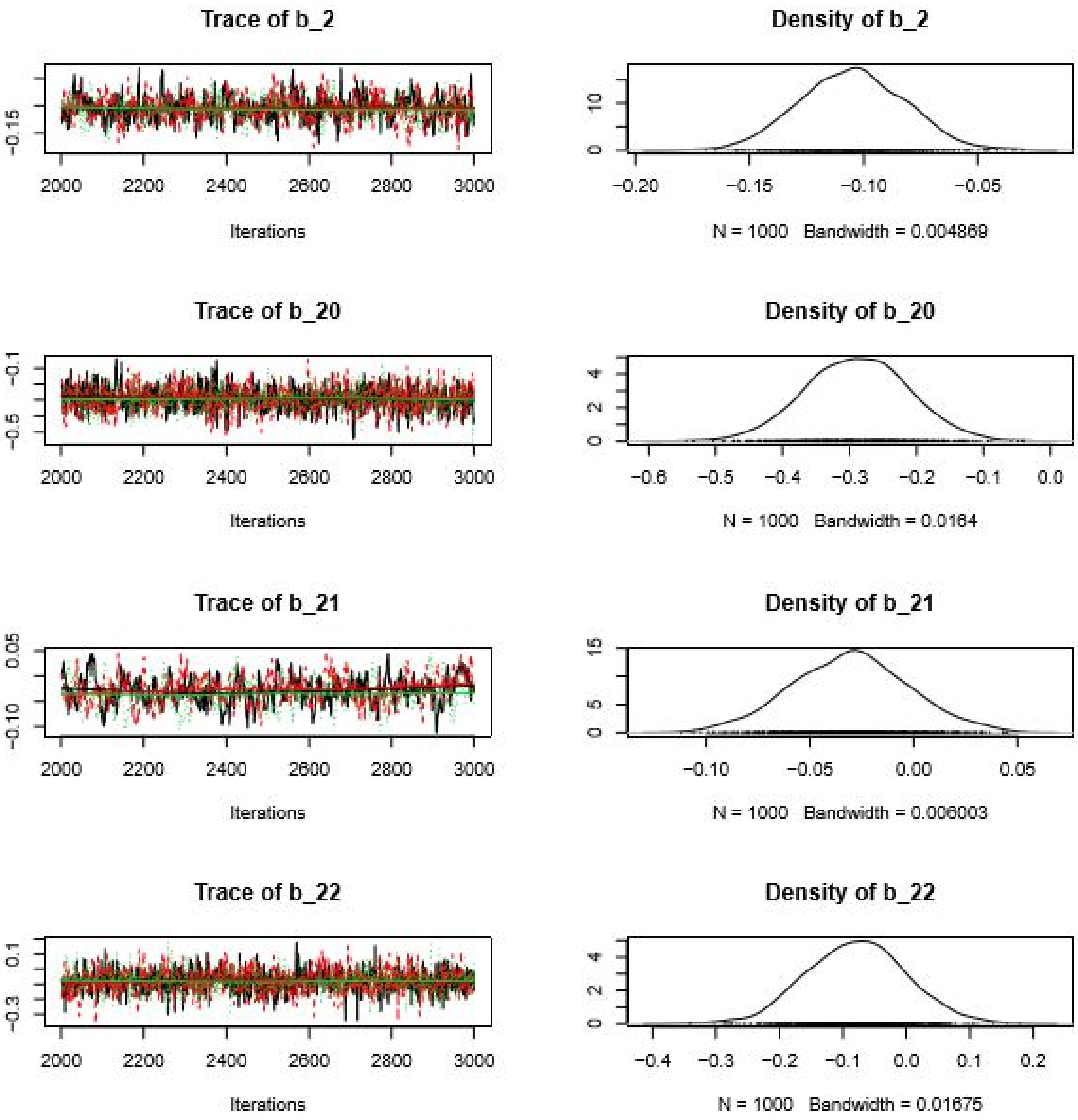

Figure E.9: Trace plot 1-D

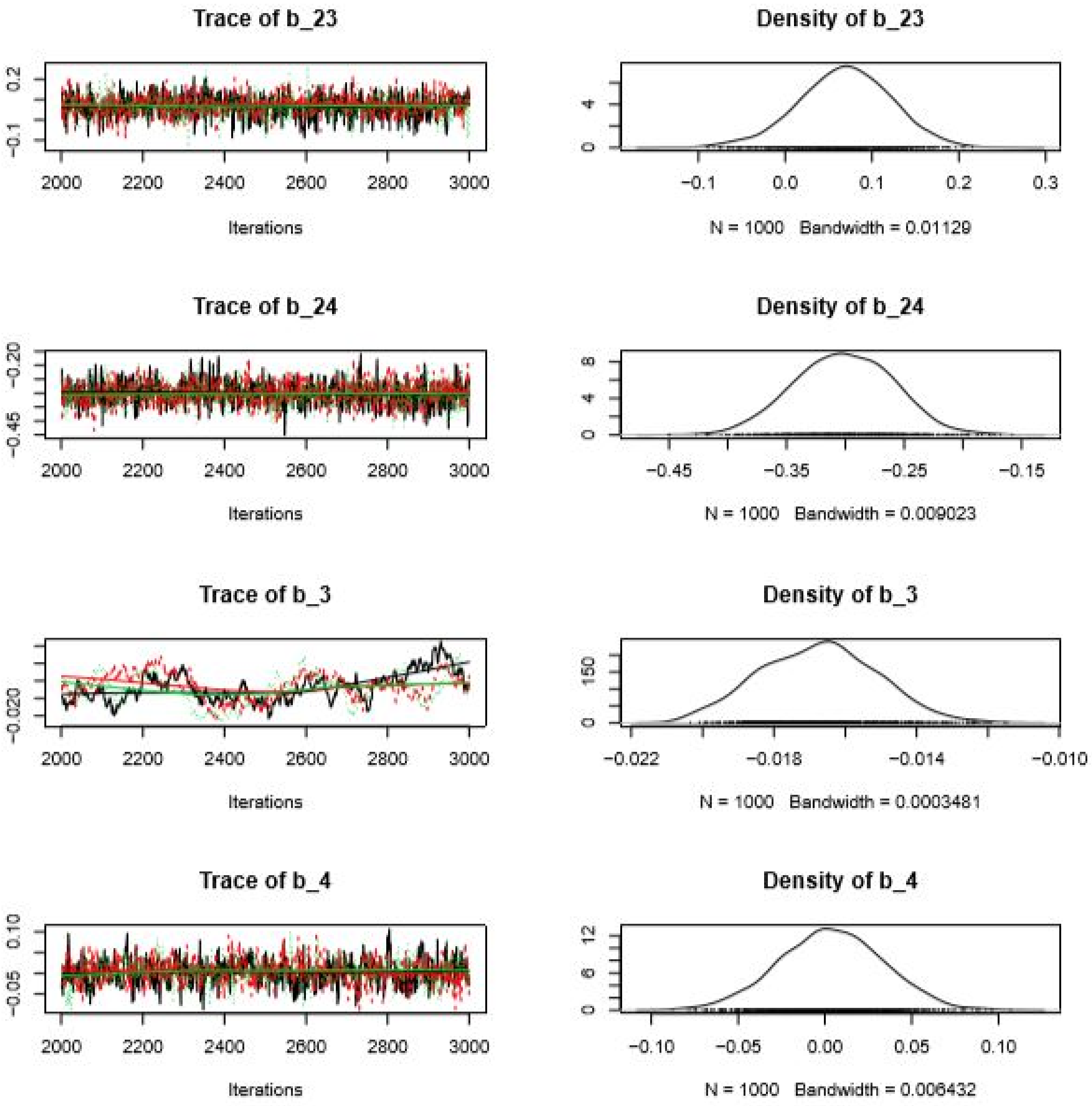

Figure E.10: Trace plot 1-E

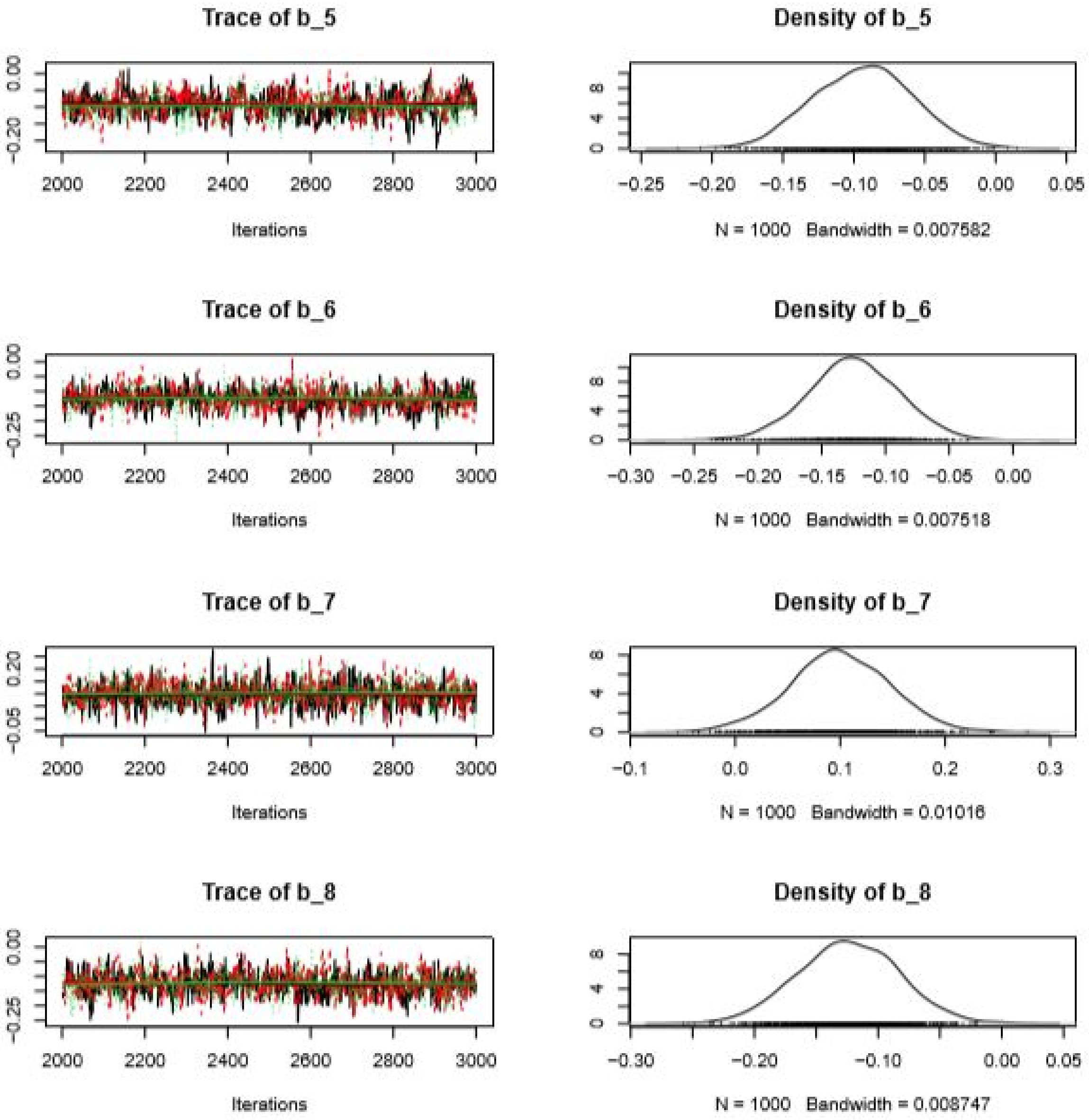

Figure E.11: Trace plot 1-F

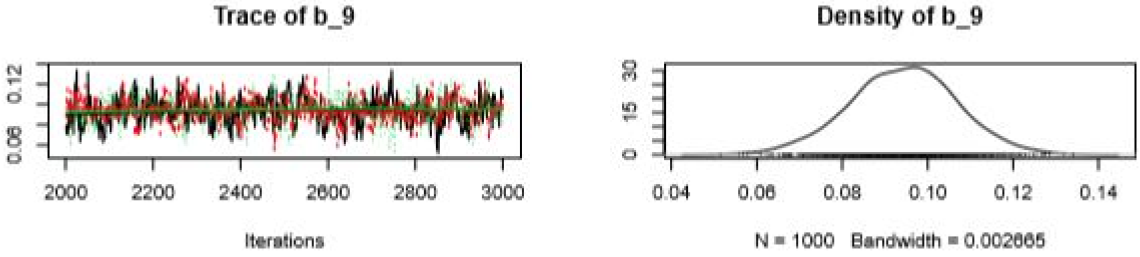

Figure E.12: Trace plot 1-G