IMPUTATION BASED ON LOCAL POLYNOMIAL REGRESSION FOR NONMONOTONE NONRESPONDENTS IN LONGITUDINAL SURVEYS

PYEYE SARAH

(MS300-0004/15)

A Thesis submitted to Pan African University Institute of Science, Technology and innovation in partial fulfillment of the requirement for the award of the degree of Master of Science in Mathematics (Statistics Option)

2017

DECLARATION

Student's declaration

This thesis report is my original work and has not been submitted to any other University for Examination.

Signature..... Date.....

Pyeye Sarah

Supervisors' Declaration

This thesis report has been submitted for examination with my approval as a University supervisor.

Signature..... Date.....

Professor Romanus Odhiambo

Jomo Kenyatta University of Agriculture and Technology, Kenya

This thesis report has been submitted for examination with my approval as a University supervisor.

Signature..... Date: .....

Professor Leo Odongo

Kenyatta University, Kenya

DEDICATION

This thesis is dedicated to the Almighty God, my beloved husband, parents, children, sisters, brothers, guardians and friends

ACKNOWLEDGMENTS

I thank the Almighty God who has enabled me to accomplish this work and wish to extend my sincere appreciation to my supervisors Prof. Romanus Odhimabo and Prof. Leo Odongo for their effort, dedicated guidance, encouragement, devotion, tolerance and all the necessary support rendered to me especially in the field of sampling and regression modeling to make this study a success.

I wish to acknowledge my dear husband Mr Kadedesya Stephen, my children Brighton and Malcom and my parents Mr and Mrs Mooma for their continued love, encouragement and patience while doing this research.

I wish to acknowledge all members of staff of Pan African University Institute of Science, Technology and innovation especially in of the department of Mathematics and Jomo Kenyatta University of Agriculture and Technology, the host University for granting me the opportunity to conduct my study. I wish to acknowledge Mr Were Festus and Mr Kilunda Charles for the moral support in the research process.

I wish to acknowledge my fellow PAUISTI 2015 students especially in mathematics for their moral support during the study and above all, i wish to acknowledge African Union for their financial support

CONTENTS

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

AIDS	Acquired Immune Deficiency Syndrome
ARVs	Anti-Retro viral Drugs
GSE	Generalized Smoothing Estimator
GRE	Generalized Regression Estimator
GAMLSS	Generalized Additive Models for Location, Scale, and Shape
GLM	Generalized Linear Models
HIV	Human Immune Virus
LM	Linear Model
LPR	Local Polynomial Regression
MAR	Missing At Random
MCAR	Missing Completely At Random
MNAR	Missing Not At Random
SRSWR	Simple Random Sampling with Replacement
SRSWOR	Simple Random Sampling without Replacement
MSE	Mean Square Error
%RB	Percentage Relative Bias
RAB	Relative Absolute Bias

ABSTRACT

In this study, the problem of nonrespondents in longitudinal survey's data is considered. The study focuses on the imputation for the longitudinal survey data which often has nonignorable nonrespondents. Local linear regression is used to impute the missing values of and then the estimation of the time-dependent finite populations means. The estimation of the time dependent means was based on the assumption that the nonresponse mechanism is last past value dependent. The asymptotic unbiasedness and consistency of the proposed estimator are investigated. The imputation for the nonmonotone nonrespondents is done multiple times through simulation and the simulation study is carried out to asses the best performing estimator of the time-dependent finite populations means. Comparisons between different parametric and nonparametric estimators are performed based on the bootstrap standard deviation, mean square error and percentage relative bias. The simulation results show that local linear regression estimator yields good properties.

CHAPTER 1

INTRODUCTION

1.1 Background

Longitudinal surveys is one of the special cases of the repeated surveys that refer to a type of sampling surveys done repeatedly over time on the same sampled units. In such surveys, data which are rich in information about the specific sampled unit can be obtained and thus suitable for various purposes. Missing data are a problem because almost all standard statistical methods require complete information for all the variables included in the analysis. A relatively few missing observations on some variables lead to reduction on the sample size of the units considered in the analysis which can significantly affect the precision of the confidence interval, weakens the statistical power and the population parameter estimates maybe biased.

When an observation for a sample unit is missing, then, the unit is defined as non-response. In Economic and clinical surveys for example, there are sensitive questions that are investigated by most researchers, such as income status of people and HIV/AIDs health status respectively. In order to do studies on such issues, longitudinal surveys play an important role in data collection and thus data analysis. In longitudinal surveys, the nonresponses often occur in a nonmonotone pattern which requires to be appropriately catered for before fitting any model in order to make meaningful inference about the given population.

Nonmonotone nonresponse exists because most times, not all the sampled units information is got at all determined time points of the surveys and yet they always

return to the survey for instance patients taking anti-retroviral drugs (ARVS) may miss a specific date and after some time point, they come again to collect the drugs. According to Robins et al. (1995), the case of monotone nonresponse is called ignorable missingness whereas that for nonmonotone nonrespondents is the nonignorable missingness. Rubin (1976) and Little and Rubin (2002) classified missing data into three categories based on missing mechanisms i.e. MCAR (missing completely at random), MAR (missing at random) and MNAR (missing not at random). Due to the existence of nonrespondents in real life surveys, three possible solutions were suggested by Kadilar and Cingi (2008), that is; Ignoring the missing observations, to sub-sample the nonrespondents or to impute the missing values. In this study, we opt to cater for the missing values that often occur in longitudinal surveys using imputation approach to intuitively fill in these missing values.

Over time, various imputation models have been developed and they have been used to overcome quite a number of challenges caused by missing data. However, some shortcomings still exist such as biasedness and inefficiency of estimators. This is because imputation models have different assumptions in both parametric and nonparametric contexts.

Parametric methods like maximum likelihood estimation have limitations such as sensitivity to model misspecification while nonparametric regression is motivated by providing more robust and flexible way of studying the relationship between variables (Dorfman, 1992). Some of the methods used by Xu et al. (2008) are simple linear regression imputation and Nadaraya-Watson technique. From their simulation results, it was found that the simple linear regression imputation approach has the weakness of producing biased estimates even when the responses at a particular time (including previous values) are correctly specified. On the other hand, Nadaraya-

Watson technique of (Nadaraya, 1964) and (Watson, 1964) used in the imputation of missing values in the longitudinal data has some weaknesses of producing a large design bias and boundary effects that give unreliable estimates for inference.

As shown by Hastie and Loader (1993) and Wand and Jones (1995), a rival for Nadaraya-Watson technique is the local linear regression estimator which was found to produce unbiased estimates without boundary effects. Cai (2001) studied the weighted Nadaraya-Watson method and was concerned with the limitations of the method such as consistency, asymptotic normality and the interior and boundary point effects. In his study, he found that local linear regression is much better than the weighted Nadaraya-Watson method as it produces asymptotically unbiased estimates without boundary effects. Moreover, Fan and Gijbels (1996) also found that the local linear regression estimator (introduced by Stone (1977)) has desirable properties.

In order to overcome the limitations of Nadaraya-Watson estimator, we derive a local linear regression estimator in the imputation of the nonresponndents in a longitudinal data set. The asymptotic properties (unbiasedness and consistency) of the proposed estimator are investigated. Comparisons between various estimators (parametric and nonparametric) are performed based on the bootstrap standard deviation, mean square error and percentage relative bias. A simulation study is conducted to asses the best performing estimator of the finite population mean.

1.2 Statement of the problem

Longitudinal surveys stand to be hailed for their undeniable significance in statistics. While they are regarded to be better and reliable in informing about various features of a study unit, they suffer from monotone and intermittent patterns of

missing data. This is often as a result of inaccessibility to or deliberate refusal of respondents to provide information thus the occurrence of nonresponses. A statistical technique called imputation is one of the approaches used to intuitively fill in these missing values. Over time, various imputation models have been developed and they have been used to overcome quite a number of challenges caused by missing data. However, some short comings still exist such as biasedness and inefficiency of estimators used in imputation. The problem of missing values especially those with intermittent patterns still demand solution and that may be the reason as to why Eubank (1988) rephrases the controversies of differing imputation models into "Let the data speak for itself". This is due to the fact that most of the imputation models used have different assumptions in both parametric and nonparametric contexts. In longitudinal surveys, the nonresponses often occur in a nonmonotone pattern and yet researchers require to use the whole data collected to make genuine inferences about a population under consideration. The missing values therefore need to be imputed in the most simple and appropriate form in which the estimators are unbiased and or inefficient. Some methods such as censoring approach, simple linear regression and Nadaraya-Watson techniques were used by Xu et al. (2008). Results of the simulation study using the simple linear regression imputation approach has the weakness of producing biased estimates even when the responses at a particular time, including previous values, are correctly specified. Besides the simple linear regression imputation, Nadaraya-Watson technique, though simple in application and have its estimated values of the regression function being within the range of the variables, this imputation method has got some weaknesses. The Nadaraya-Watson technique is limited as it produces a large design bias and boundary effects that do give non reliable estimates for inference. This therefore creates the mathematical

gap, in which imputation methods need to be improved or a new approach to be developed. Local polynomial regression imputation, an extension of the kernel-based method, may be used in place of Nadaraya-Watson technique since it eliminates the boundary effects and gives consistent estimates.

1.3 Research Objectives

1.3.1 General objective

The main objective of this study is to develop an imputation method based on local polynomial regression for nonmonotone nonrespondents in longitudinal surveys and determine its asymptotic properties.

1.3.2 Specific Objectives

1. To derive an imputation regression estimator of finite population mean based on local polynomial regression.
2. To determine the asymptotic properties of the estimators
3. To carry out simulation study on the estimators.

1.4 Justification of the study

Sample surveys is an important field of study in statistics. Through sample surveys, statisticians as well as other researchers are able to obtain data rich in information about a given population and carry out estimation of particular population parameters by drawing samples (sub populations) from populations of interest. One of problems with longitudinal data are the missing observations because, almost all statistical methods allow complete information to make statistical analysis. As a result, the precision of confidence intervals is affected , statistical power reduces and the

population parameter estimates may be biased. Also, the imputation models commonly used such as the multiple imputation model and doubly robust methods are sensitive to model misspecification leading to biased and less precise results.Since local polynomial fitting has got attractive theoretical properties compared to the local constant estimator, imputation of the missing data by means of the local polynomial regression estimator yields reliable results. Using the LPR in the imputation process produce asymptotically unbiased estimates and hence overcomes the great limitation of the kernel estimator. Since statisticians and researchers in other areas get best results if such an estimator is used, the results obtained may be of great importance as it will ensure appropriate inferences about finite populations parameters in longitudinal surveys, hence they can be used for purposes of policy planning, development and implementation in sectors of the economy for instance, healthy policies, education, industry and production.

1.5 Significance of the study

The study has contributed easy and reliable approach to analysis of longitudinal survey data with nonrespondents. Using local linear regression in the imputation process helps researchers to find a set of parameter estimators that fit an imputation model best for prediction of missing values in longitudinal surveys. The advantage of local linear or in general, LPR over other nonparametric regression techniques is that it takes into consideration the tail distribution (boundary points) which is not commonly the case for other techniques. This study has yielded more reliable estimates of missing values as demonstrated in the simulation result. The results in this study will go a long way in addressing the deficiencies suffered by other methods which have been used before.

CHAPTER 2

LITERATURE REVIEW

2.1 Imputation model based on local polynomial regression and finite population mean estimation .

Imputation is one of the reliable options that researchers in both social sciences and other related research use before data analysis and interpretation of results are carried out for any sample survey data. Imputation is the process of filling in missing values (nonrespondents which are either unit or item) in order to make reliable statistical inferences. Nonresponse of the sampled units may be due to personal intention not to respond, failing to locate the unit and or failure of the unit to participate in the survey as a result of various reasons. Other causes of unit nonresponse include the poor or incomplete sampling frame.

In the study by Cai (2001) studies were done about the nonparametric regressions using the weighted Nadaraya-Watson method. The author was concerned with the limitations of the method of the weighted Nadaraya-Watson such as consistency , asymptotic normality and the interior and boundary points effects. In comparison of their method with the linear regression approach, they found out that the local linear regression produces unbiased estimates without boundary effects, see; Hastie and Loader (1993), and Wand and Jones (1995). Local linear regression and its theoretical properties were investigated in comparison with the binning Nadaraya-Watson which is a kernel smoothing estimator. Fan and Gijbels (1992) noted that other nonparametric methods such as the Gasser-Muller estimators which are kernel

smoother and the weighted Nadaraya-Watson method have their results converge very slowly at the boundaries. The results therefore, according to Wand and Jones (1995) the Local linear regression has a great merit especially in the theoretical properties desired for both the interior and near the boundaries. However, the Binning Nadaraya-Watson estimator was good for implementation but the results produced were biased and also had the boundary value effects hence their recommendation for the use of the local polynomial regression. Local polynomial regression is one of the most popular smoothing nonparametric techniques and is considered to be the kernel generalized smoother used in estimation of the population mean, therefore, in the research, we to use it for imputing the nonmonotone nonrespondents.

Yu and Li (2011) did the imputation of non-ignorable nonresponses for income in Taiwan and it was noticed that the nonresponse rates during personal investigations are increasingly high due to deliberate refusal of respondents to provide information and fear for social security. In their study, three stage stratified sampling technique was employed and then the panel survey was done yearly for six year period. Their results were obtained after employing the two step generalized estimating equations method but they were not evaluated, hence recommended to do the validity of their method in comparison to the pattern -mixture method and likelihood based model among others, see Little (1993) and Laird (1988) respectively.

Shahab et al. (2014)) carried out a study of Dual imputation model for incomplete longitudinal data. The authors used the comparative approach of the multiple imputation method and the doubly robust weighting-based method. They integrated the two methods and came up with a new imputation model which was used to fill in the missing values on which estimates were done and conclusions were made. The results showed that the application of their proposed method in statistical soft-

ware is simple and the estimates are unbiased with the assumption that one of the imputation models is specified. Besides, their method does not guarantee accurate results for inference. Also, using very large samples limits the use of their proposed method since, you ought to use the stratification (of about five strata) per time point of imputation and hence demanding further modification of the new method or development of a better method.

According to Shahab et al. (2014), the small samples give correct values compared to the variable numbers under consideration, this makes imputation model to have cases of over parameterization of the design- model and practically, it is rare to have this. Thus, this kind of imputation model may give plausible results which needs extra care in order to construct a good model. The nonmonotone nonresponse clearly makes the statistical analysis complicated if the data is not well managed. We propose to carryout imputation using the local polynomial regression procedures which for years has been popularly used in many other areas as a result of its attractive adaptation of edge effects and bias reduction as stipulated in the book of Fan and Gijbels (1996) chapter three. Non parametric regression is one of the approaches used to give relationship between the regressors and regressand and guide in the exploratory analysis for improving the functional models. According to Sarndal et al. (1992)), the imputation regression estimator is referred to as the Generalized Smoothing Estimator (GSE) which is an extension of the Generalized Regression Estimator (GRE).

According to Jong et al. (2014), they proposed an imputation technique based on generalized additive models for location, scale, and shape (GAMLSS) which is more flexible in combination with a multiple imputation than standard parametric imputation models usually provided by software packages. By adopting their proposed

method, misspecification of imputation models and thus invalid inference is less likely. In a simulation study, performance of a newly proposed imputation method was investigated through comparison of their new imputation method with standard methods , for example, the generalized linear model and two versions of predictive mean matching. However, the linear model (LM) and the Generalized linear model (GLM) parametric regression models posed restrictions on the functional form of the conditional mean and variance of the variable with unobserved values. These restrictions may lead to inconsistent estimation of the parameters of scientific interest, and evantually lead to invalid multiple imputation inferences. The sensitivity of multiple imputation methods to deviations from their distributional assumptions was also investigated using simulations, where the parameters of scientific interest are the coefficients of a linear regression model, and values in predictor variables are missing at random.

Although imputation methods based on predictive mean matching are virtually unbiased, they suffer from moderate under coverage, even in the experiment where all variables are jointly normal thus, GAMLSS method features to have better coverage than currently available methods. The results of their simulation study implied that the proposed GAMLSS method works also well in larger models. Therefore, it is expected that imputation methods that jointly estimate the conditional expectation and conditional variance using nonparametric techniques offer better performance thus a motivation for our study.

Compared with kernel estimators, local polynomial fittings is said to have the advantages of being adaptive to both random and fixed designs, and can adjust boundary biases automatically as stated in the work of Fan and Gijbels (1996), Ruppert and Wand (1994) and Cheng et al. (1997). Therefore, for nonignorable missing values,

the proposed imputation method is based on LPR which may cater for the nonlinear pattern of the nonmonotone nonrespondents by extension of the Generalized smoothing estimator (GSE). The LPR is expected to improve on the theoretical properties of the kernel estimator by using higher order polynomial as a local approximation of the unknown smoothing function. LPR does an extension of the kernel estimation to polynomial fit at the point of interest (focal point). The parameters of LPR are; the band width h , the kernel (weighted) function (w_i) and the order (p) of the polynomial of regression.

Rueda and Sánchez-Borrego (2009) studied about estimation of the population mean using non parametric methods such as the generalized kernel smoother. However, the nonrespondents in their survey were not included in the model, instead the non sampled units were predicted. Their results were evaluated through comparison of the design-based and the model-based model (super population model). It was practically evident that the results based on the local polynomial regression (their proposed method) under model- based approach yielded the best estimates as earlier shown by Breidt and Opsomer (2000). It is on this basis, we propose to carry out longitudinal data imputation for nonmonotone nonrespondents using the local polynomial regression approach. Through related literature, Godambe (1955) and Little (2004), the model-based population parameter estimations is recommended during the sample selections and design.

2.1.1 Nadaraya-Watson Estimator.

In the estimation of finite population parameters such as total or mean, many researchers like Anthony (1999), have been attracted to using the nonparametric procedures. This is due to the various merits exhibited by the nonparametric methods such as being distribution free and being robust during application. Nonparametric

regression such as the Nadaraya-Watson estimator (Xu et al. (2008)) was used in the imputation for nonrespondents before estimation of population mean. This estimator is sometimes called the kernel estimator or local constant estimator and it helps in the smoothing of the scatter plots. Consider a non-parametric regression model

$$Y_i = f(x_i) + \varepsilon_i \tag{2.1}$$

where $f(x_i)$ is the regression function and ε_i refer to the errors which are assumed to be independently and identically distributed with mean zero and a constant variance independent of x_i.

The Nadaraya-Watson estimator is the solution obtained by minimizing the weighted least squares problem. Thus, for

$$w_i(x) = K\{\frac{x - x_i}{h}\} \tag{2.2}$$

the estimator was given by

$$\hat{f}_n(x) = \sum_{i=1}^{n} [s_i(x)Y_i] \tag{2.3}$$

where $s_i(x) = \frac{w_i(x)}{\sum_{i=1}^{n} w_i(x)}$

Hence

$$\hat{f}_n(x) = \sum_{i=1}^{n} \frac{K\left\{\frac{x-x_i}{h}\right\}}{\sum K\left\{\frac{x-x_i}{h}\right\}} Y_i = \sum_{i=1}^{n} W_i(x)Y_i \tag{2.4}$$

where

$$W_i(x) = \frac{K\left\{\frac{x-x_i}{h}\right\}}{\sum K\left\{\frac{x-x_i}{h}\right\}}$$

After getting the estimator, as stated in the work of Dorfman (1992), the Kernel estimator is then used in the prediction before estimation of the finite population

total/mean. Using a sample drawn from the finite population, p. The actual population total was estimated using the sample total given by

$$y = \sum_{i=1}^{N} y_i \tag{2.5}$$

Equation (2.5) gives the sum of all sampled and non-sampled units. For s sampled units, equation (2.5) becomes

$$y = \sum_{i\epsilon s} y_i + \sum_{i\epsilon(p-s)} y_i \tag{2.6}$$

Basically, the idea in this estimation was to get the non sampled units. Now, using nonparametric regression defined by the model $Y_i = f(x_i)+\varepsilon_i$ in the prediction of the unknown units of nonsampled units, we consider some conditions and assumptions. According to Fan (1993), the kernel function satisfy the following conditions;

$K(u) \geq 0$ for all u thus it is positive, $\int K(u)du = 1$ thus, it is a continuous distribution function, $\int uK(u)du = 0$, K is symmetric about zero

$\int u^2K(u) = k^2 < \infty$ and $\int_{-\infty}^{\infty} \{K(u)\}^2 du < \infty$

Using the Kernel function $K(u)$, the weighted kernel function $W_i(x)$ is used to define the Nadaraya-Watson Estimator, hence

$$\hat{Y} = \sum_{i\epsilon s} y_i + \sum_{i\epsilon(p-s)} \hat{f}(x_i). \tag{2.7}$$

In the work of Fan and Gijbels (1992), local linear fitting was done in their study for the univariate case and their results showed that there was reduced design bias , good adaptation to boundary effects. Unlike the fair extensive research done about the local constant fit, it's application has been limited due to the biased estimator

especially for points on the boundaries. More information about the use of local polynomial fitting, see Ruppert and Wand (1994) and Hastie and Loader (1993).

In the study by Singh et al. (2016), two imputation techniques were suggested using auxiliary information followed by two class of point estimators for estimating the population mean following MCAR missing mechanism under a SRSWR scheme. The minimum biases and mean square errors of the proposed class of estimators were determined up to the first order of approximation. It was established theoretically and empirically that the proposed class of estimators and methods of imputation used performs best compared to other estimators like the mean and Ratio imputation methods based on five populations . It was also shown that their proposed point estimators were more efficient than some existing point estimators. However, According to the existing literature, Singh and Horn (2000), the point estimators compared to for evaluation have their theoretical and empirical properties showing that they produce asymptotically unbiased population mean with large precision. This lays a ground for more investigations for the appropriate imputation methods and point estimators.

2.2 Asymptotic properties of an estimator

The asymptotic properties are commonly determined in order to evaluate theoretically the properties of an estimator. The properties commonly considered for a good estimator include; Unbiasedness, Consistency, Efficiency and normality in distribution. When a sample is taken from a population, it is used to estimate the properties of the population, but the size of the sample taken must be a good representation of the population of interest. During asymptotic properties determination, we consider the assumption that as the sample size increases, the more closer to the exact value of the estimate will be. Now, using the nonparametric estimator for the population

total, we consider the assumption as $n \to \infty$ and the smoothing parameter $h \to 0$ in order to obtain the expression for whether the estimator is unbiased or not. In the work of Cheng (1994), the population mean estimator under some conditions was asymptotically normal in distribution with mean zero and constant condition variance. Some well known results on asymptotic normality of estimators include Erdasos and Razenyi (1959), and Scott and Wu (1981) for simple random sampling without replacement, Krewski and Rao (1981) and Bickel and Freedman (1984) for stratified random sampling, and Hazajek (1964) for unequal probability sampling without replacement.

CHAPTER 3

METHODOLOGY

3.1 Introduction

In this chapter, the proposed techniques that have been used for the achievement of the objectives of the study are discussed. The solution to the problem of nonrespondents in surveys are presented by imputation techniques to fill in the missing values of the survey data. In our study, a nonparametric technique based on the local linear regression is used to predict the monotone and nonmonotone nonrespondents. The imputed data is then used in the estimation of the finite population total and or mean and the asymptotic properties of the finite population total or mean estimator are determined. The performance of the proposed estimator is studied through a simulation study where comparison to the simple linear regression imputation method and Nadaraya-Watson imputation approach is done.

3.1.1 Assumptions and notations

1. All sampled units be observed on the first time ($t = 1$) remain in the sample till the final time $t = T$. Let y be the variable of interest so that $y_{i,t}$ is the value of y for the i^{th} population unit at time point t. Also, let the total time points for the longitudinal surveys be T such that; $t = 1, 2, \ldots, T$.

2. Assuming that the first inclusion probability, π_i depends on the population values of the response variables and explanatory variables at the first occasion only and the response indicator function $I_{i,t}$ for the data values is defined by;

$$I_{i,t} = \begin{cases} 1; & y_{i,t} - observed \\ 0; & y_{i.t} - unobserved \end{cases} \qquad t = 1, 2, \cdots, T \tag{3.1}$$

3. The prediction process is past last value dependent (last value dependent non-response mechanism) and the vectors $(y_{i,1} \cdots y_{i,T}, I_{i,1}, \cdots, I_{i,T})$ are independently distributed from the superpopulation under the model-assisted approach for $t = 2, \cdots, T$ and $i = 1, 2, \ldots, N$, thus;

$$P(I_{i,t} = 1 | y_{i,1}, \cdots, y_T, I_{i,1,} \cdots I_{i,t-1,} I_{i,t+1}, \cdots I_{i,T}) = P(I_{i,t} = 1 | y_{i,t-1}) \tag{3.2}$$

where P is the probability with respect to the superpopulation.

4. The vector $(y_{1,} \cdots, y_T)$ follows the Markov chain for longitudinal survey data without missing values

$$L(y_{i,t} | y_{i,t-1,} I_{i,t} = 0, I_{i,t-1} = 1) = L(y_{i,t} | y_{i,t-1,} I_{i,t-1} = 1) \tag{3.3}$$

which is the conditional distribution of the nonresponse y_t given the past respondent value y_{t-1} with the nonresponse y_t having the second finite moment i.e. $E(y_t^2) < \infty$ for unit i.

5. The population is divided into a fixed number of imputation classes, which are basically unions of some small strata. This was the case in the paper of Xu et al. (2008) and Shao et al. (2012), and in this study, only a single imputation class is considered.

3.2 Description of Population and Samples

In this study, model-assisted approach was considered for survey measurements of a finite population P from which a sample is drawn based on stratified SRSWOR design. Without loss of generality, in order to accomplish the goal described in (3.1) a bivariate random variable (X, Y) with probability density function $f(x, y)$ is used. Let A and B be two independent samples of size n_A and n_B respectively with n_A and n_B being fixed by survey design. The sample of size n_A have missing values while the one of n_B has its records complete. This situation is also considered in statistical matching, see Pier et al. (2008) where missingness is induced by survey design and can be considered deterministic. Suppose B_n is the simple random sample of size n from the finite population of size N. In this way, the sample consists of two parts: B_r and B_{n-r} where B_r is the set of all respondents in the sample and B_{n-r} is the set of all non respondents in the sample. In this study, the imputation of y_i is based on the nonparametric estimation of the regression function via the local linear estimators. Making assumption (5), imputation is carried out within each imputation class with the study variable from a population unit following a super population.

3.3 Imputation Process

Suppose we are estimating a regression model with multiple independent variables, impute missing $y_{i,t}$ with a $y^*_{i,t}$ randomly generated from the estimated conditional expectation denoted by $\hat{\varphi}(y_{i,t}|y_{i,t-1})$. In this study, we adopt the regression model such that the main task is to predict the correct missing values with in the sample for $t \geq 2$ following the last past value dependent mechanism. According to Rubin (1987), the joint distribution of bivariate random variables (X, Y) is preserved when the missing value, $y_{i,t}$ is imputed by the conditional distribution of Y given X .

Therefore, the conditional mean imputation approach for the single imputation is considered.

Let

$$\varphi_{i,t,t-1}(y_{t-1}) = E(y_{i,t}|y_{i,1,\cdots},y_{i,t-1}, I_{i,1} = \cdots = I_{i,t-1} = 1, I_t = 0) \tag{3.4}$$

be the conditional expectation with respect to the superpopulation for unobserved value $y_{i,t}$ with observed value $y_{i,t-1}$ for $t \geq 2$

It is clear that when $\varphi_{i,t,t-1}$ is known, the imputed value of unobserved $y_{i,t}$ is given by; $y^*_{i,t} = \varphi_{i,t,t-1}(y_{t-1})$. In cases where $\varphi_{i,t,t-1}(y_{t-1})$ in equation (3.4) is unknown for nonmonotone nonrespondents, we use assumption (3) such that equation (3.4) is equivalent to

$$\varphi_{i,t,t-1}(y_{t-1}) = E(y_{i,t}|y_{i,1,\cdots},y_{i,t-1}, I_{i,1} = \cdots = \ I_{i,t-1} = 1, I_t = 1) \tag{3.5}$$

Using equation (3.4), we are limited to do estimation by regressing the nonrespondents $y_{i,t}$ on the observed values $y_{i,t-1}$ based on the longitudinal survey data, therefore, the equivalent equation (3.5) which allows estimation using data from all subjects having $y_{i,t}$ observed and $y_{i,t-1}$ observed is used. Imputation of the nonrespondents is therefore done using $\varphi_{i,t,t-1}(y_{t-1})$ of equation (3.5), (related work in Xu et al. (2008) and Shao et al. (2012) articles) thus, using several survey data in regression fitting. According to Paik (1997), imputing nonresponses using (3.5) was done for monotone case and their approach is easy to apply if the conditional expectation, $\varphi_{t,t-1}(x)$ in (3.4) has a linear relationship with x. However, for the nonmonotone nonrespondents case in longitudinal surveys, the pattern of nonrespondents may be linear or nonlinear. Adopting the concept of nonparametric method in Cheng (1994),

here, the local linear regression estimator (LL) of $\varphi_{t,t-1}(x)$ is $\hat{\varphi}_{t,t-1}(x)$.

Associated with each $y_{i,t}$ are the known $x_{i,t,q}$;$q = 1, \cdots, Q$, of q explanatory (auxiliary) variables. To make the notations and writings simple, we do make the index t silent and write with a single subscript i , thus $y_{i,t}$ is written as y_i. As in the works of Cheng (1994)and Wang and Rao (2002), under the model -assisted approach, the response indicator I_i and the variable of interest y_i are conditionally independent given $X = x_i$.

The regression imputation model η is given by the relation

$$y_i = m(x_i) + \varepsilon_i \tag{3.6}$$

such that ε_i 's are residuals which are assumed to be independent normally distributed with mean zero and variance $\sigma^2(x_i)$, $\sigma^2(\cdot)$ is a smooth function which is strictly positive.

Then, it is clear that

$$E_\eta(y_i/X = x_i) = m(x_i) \tag{3.7}$$

$m(x_i)$ - the unknown regression function of y_i given x_i with $x_i = (x_{i1}, \cdots, x_{iT})$, the known auxiliary data of the sample unit i and $m(\cdot)$ is a smooth function of a single auxiliary variable x

$$Cov(y_i, y_j|X = x_i, X = x_j) = \begin{cases} \sigma^2(x_i) & i = j \\ 0 & otherwise \end{cases} \tag{3.8}$$

According to Fan and Gijbels (1996) y_i and x_i are regressed using the weighted least squares equation. To obtain the estimate of $m(x_i)$ at y_{t-1} and its derivatives, we use

the weighted local polynomial fitting by taking the assumption that the regression function with $(p+1)th$ derivatives at a point say $x = x_0$ exists and are continuous. Now, the imputation model can be re-written as,

$$y_i = m_{y_{t-1}}(x_i) + \varepsilon_i \tag{3.9}$$

where approximation of $m_{y_{t-1}}(x_t)$ about y_{t-1} is done following the Taylor series expansion, thus, for the observed last value $y_{t-1} = x_0$, we have

$$m_{y_{t-1}}(x_i) \approx m(x_0) + m'(x_0)(x_i - x_0) + \frac{m''(x_0)(x_i - x_0)^2}{2!} + \cdots + \frac{m^p(x_0)}{p!}(x_i - x_0)^p \tag{3.10}$$

$$\approx \beta_0 + \beta_1(x_i - x_0) + \beta_2(x_i - x_0)^2 + \cdots + \beta_p(x_i - x_0)^p \tag{3.11}$$

where $\beta_0 = m(x_0)$; $\beta_1 = m'(x_0)$; $\beta_2 = \frac{m''(x_0)}{2!}$; $\beta_p = \frac{m^p(x_0)}{p!}$, in general $\beta_j = \frac{m^j(x_0)}{j!}$) for $j = 0, 1, \cdots, p$, x_i in the neighborhood of x_0 and m^j denoting the j^{th} derivative of the function m .

Substituting equation (3.11) into equation(3.9), we get

$$y_i = \beta_0 + \beta_1(x_i - x_0) + \beta_2(x_i - x_0)^2 + \cdots + \beta_p(x_i - x_0)^p + \epsilon_t \tag{3.12}$$

Estimation of the unknown parameters in equation (3.11) is done by evaluating the solution to the weighted least squares problem and computed using equation (3.12) to minimize the weighted least squares method given by

$$\sum_{i=1}^{n} w_i(x)\epsilon_i^2 = \sum_{t=1}^{n} \left[y_i - \sum_{j=0}^{p} \beta_j (x_i - x_0)^j \right]^2 w_i(x) \tag{3.13}$$

Here, the local kernel weight denoted by $w_i(x)$ is defined as;

$$w_i(x) = K \left\{ \frac{(x_i - x_0)}{h} \right\} \tag{3.14}$$

where h is the bandwidth, some times called the smoothing parameter and in equation (3.13), it is used to determine the size of the neighborhood of x_i around x_0 (local neighborhood) and thus, h controls the degree of smoothing.

K is the kernel function which should be strictly positive and $K_h(\cdot)$ controls the weights, x_0 is the point of focus and x_i being the random variable (covariates) with design matrix centered at past last value and j is the order of the local polynomial . Now, denote the solution to equation (3.13) as $\hat{\beta}$ such that $\hat{m}(x_0) = \hat{\beta}(x)$, so the LPR estimator is obtained by minimizing equation (3.13).

Let

$$S = \sum_{i=1}^{n} \left[y_i - \sum_{j=0}^{p} \beta_j (x_i - x_0)^j \right]^2 w_i(x) \tag{3.15}$$

Accordingly, for $j = 0$,

$$S = \sum_{i=1}^{n} \{y_i - \beta_0\}^2 w_i(x)$$

Differentiating with respect to β_0 and equationg to zero, we get;

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} w_i(x) y_i}{\sum_{i=1}^{n} w_i(x)} \tag{3.16}$$

Thus, $\hat{y}_i = \hat{\beta}_0$

where

$$\hat{y}_i = \hat{m}_0(x) = \frac{\sum_{i=1}^{n} w_i(x) y_i}{\sum_{i=1}^{n} w_i(x)} \tag{3.17}$$

Equation (3.17) is the Nadaraya-Watson estimator

Using the local constant estimator, the conditional expectation given by $\hat{\varphi}(y_{t-1})$ is used to impute the missing values, see Xu et al. (2008). It is defined by

$$\hat{\varphi}_{t,t-1}(y_{t-1}) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right) \omega_{\mathbf{i}} \mathbf{I_i} y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right) \omega_i \mathbf{I_i}} \tag{3.18}$$

where ω_i is the weight according to the sampling design and

$$\mathbf{I_{t,t-1,i}} = \begin{cases} 1, & I_{t,i} = 1, I_{t-1,i} = 1, \\ 0, & otherwise \end{cases} \qquad t = 2, \cdots, T \tag{3.19}$$

Similarly for $j = 1$,

$$S = \sum_{i=1}^{n} \left\{ y_i - \beta_0 - \beta_1 (x_i - x_0) \right\}^2 w_i(x) \tag{3.20}$$

Differentiating equation (3.20) with respect to β_0 and equating to zero, we get

$$\beta_0 \sum_{i=1}^{n} w_i(x) + \beta_1 \sum_{i=1}^{n} w_i(x)(x_i - x_0) = \sum_{i=1}^{n} y_i w_i(x) \tag{3.21}$$

Now, differentiating equation (3.20) with respect to β_1 ,we get;

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} \{y_i - \beta_0 - \beta_1(x_i - x_0)\}(x_i - x_0)w_i(x) \tag{3.22}$$

Equating to zero equation (3.22) and simplifying equation(3.22), we have;

$$\beta_0 \sum_{i=1}^{n} w_i(x)(x_i - x_0) + \beta_1 \sum_{i=1}^{n} w_i(x)(x_i - x_0)^2 = \sum_{i=1}^{n} y_i w_i(x)(x_i - x_0) \tag{3.23}$$

Solving equations (3.21) and (3.23) simultaneously, we first make β_0 the subject, thus,

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i w_i(x) - \beta_1 \sum_{i=1}^{n}(x_i - x_0)w_i(x)}{\sum_{i=1}^{n} w_i(x)} \tag{3.24}$$

Substituting β_0 in equation (3.24) in equation (3.23), we get

$$\left[\frac{\sum_{i=1}^{n} y_i w_i(x) - \beta_1 \sum_{i=1}^{n}(x_i - x_0)w_i(x)}{\sum_{i=1}^{n} w_i(x)}\right] \sum w_i(x)(x_i - x_0)$$

$$+ \beta_1 \sum_{i=1}^{n} w(x)(x_i - x_0)^2 = \sum_{i=1}^{n} y_i w_i(x)(x_i - x_0) \tag{3.25}$$

Opening the brackets, we have

$$\left\{ \begin{array}{c} \sum_{i=1}^{n} y_i w_i(x) \sum_{i=1}^{n} w_i(x)(x_i - x_0) - \beta_1 \left(\sum_{i=1}^{n} w_i(x)(x_i - x_0)\right)^2 \quad + \\ \beta_1 \sum_{i=1}^{n} w(x)(x_i - x_0)^2 \sum_{i=1}^{n} w_i(x) \end{array} \right\}$$

$$= \sum_{i=1}^{n} y_i w_i(x)(x_i - x_0) \sum_{i=1}^{n} w_i(x) \tag{3.26}$$

Collecting like terms and factorizing β_1 on the left hand side of equation (3.26), we have

$$\left\{ \beta_1 \left[\sum_{i=1}^{n} w(x)(x_i - x_0)^2 \sum_{i=1}^{n} w_i(x) - \left(\sum_{i=1}^{n} (x_i - x_0) w_i(x) \right)^2 \right] \right\}$$

$$= \left\{ \sum_{i=1}^{n} y_i w_i(x)(x_i - x_0) \sum_{i=1}^{n} w_i(x) - \sum_{i=1}^{n} y_i w_i(x) \sum_{i=1}^{n} w_i(x)(x_i - x_0) \right\} \tag{3.27}$$

Clearly,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i w_i(x)(x_i - x_0) \sum_{i=1}^{n} w_i(x) - \sum_{i=1}^{n} y_i w_i(x) \sum_{i=1}^{n} w_i(x)(x_i - x_0)}{\sum_{i=1}^{n} w(x)(x_i - x_0)^2 \sum_{i=1}^{n} w_i(x) - \left(\sum_{i=1}^{n} (x_i - x_0) w_i(x) \right)^2} \tag{3.28}$$

Defining $S_j(x) = \sum_{i=1}^{n} w_i(x)(x_i - x_0)^j$ and $T_j(x) = \sum_{i=1}^{n} y_i w_i(x)(x_i - x_0)^j$, we have

$S_0(x) = \sum_{i=1}^{n} w_i(x)$, $S_1(x) = \sum w_i(x)(x_i - x_0)$, $S_2(x) = \sum_{i=1}^{n} w_i(x)(x_i - x_0)^2$

$T_0(x) = \sum_{i=1}^{n} y_i w_i(x)$, $T_1(x) = \sum_{i=1}^{n} y_i w_i(x)(x_i - x_0)$

Substituting $S_j(x)$ in equation (3.28), we get

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i w_i(x)(x_i - x_0) S_0(x) - \sum_{i=1}^{n} y_i w_i(x) S_1(x)}{S_2(x) S_0(x) - \left(S_1(x) \right)^2} \tag{3.29}$$

$$\hat{\beta}_1 = \sum_{i=1}^{n} \left\{ \frac{(x_i - x_0) S_0(x) - S_1(x)}{S_2(x) S_0(x) - \left(S_1(x) \right)^2} \right\} w_i(x) y_i \tag{3.30}$$

Substituting $T_j(x)$ in equation (3.30), we get

$$\hat{\beta}_1 = \frac{S_0(x) T_1(x) - S_1(x) T_0(x)}{S_2(x) S_0(x) - S_1(x)^2} \tag{3.31}$$

$$\hat{\beta}_1 (x_i - x_0) = \sum_{i=1}^{n} \left\{ \frac{(x_i - x_0) S_0(x) - S_1(x)}{S_2(x) S_0(x) - \left(S_1(x) \right)^2} \right\} w_i(x) y_i \left[(x_i - x_0) \right] \tag{3.32}$$

Substituting equation (3.28) into equation (3.24), we get

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i w_i(x) - [\sum_{i=1}^{n}(x_i - x_0)w_i(x)] \frac{\sum_{i=1}^{n} y_i w_i(x)(x_i - x_0) \sum_{i=1}^{n} w_i(x) - \sum_{i=1}^{n} y_i w_i(x) \sum_{i=1}^{n} w_i(x)(x_i - x_0)}{\sum_{i=1}^{n} w(x)(x_i - x_0)^2 \sum_{i=1}^{n} w_i(x) - \left(\sum_{i=1}^{n}(x_i - x_0)w_i(x)\right)^2}}{\sum_{i=1}^{n} w_i(x)} \tag{3.33}$$

Expanding and simplify the numerator in equation (3.33), equation 3.33 becomes

$$\beta_0 = \frac{\left\{ \begin{array}{c} \sum_{i=1}^{n} y_i w_i(x) \sum_{i=1}^{n} w_i(x)(x_i - x_0)^2 \sum_{i=1}^{n} w_i(x) - \\ \sum_{i=1}^{n} y_i w_i(x)(x_i - x_0) \sum_{i=1}^{n} w_i(x) \sum_{i=1}^{n}(x_i - x_0)w_i(x) \end{array} \right\}}{\sum_{i=1}^{n} w_i(x)(x_i - x_0)^2 \left[\sum_{i=1}^{n} w_i(x)\right]^2 - \left(\sum_{i=1}^{n} w_i(x)(x_i - x_0)\right)^2 \sum_{i=1}^{n} w_i(x)} \tag{3.34}$$

Factorizing $\sum_{i=1}^{n} w_i(x)$ on the right hand side of equation (3.34), we get

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i w_i(x) \sum_{i=1}^{n} w(x)(x_i - x_0)^2 - \sum_{i=1}^{n} y_i w_i(x)(x_i - x_0) \sum_{i=1}^{n} w_i(x)(x_i - x_0)}{\sum_{i=1}^{n} w(x)(x_i - x_0)^2 \sum_{i=1}^{n} w_i(x) - \left(\sum_{i=1}^{n} w_i(x)(x_i - x_0)\right)^2} \tag{3.35}$$

Substituting $S_j(x)$ in equation (3.35), we get

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i w_i(x) S_2(x) - \sum_{i=1}^{n} y_i w_i(x)(x_i - x_0) S_1(x)}{S_2(x) S_0(x) - S_1(x)^2} \tag{3.36}$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left\{ \frac{S_2(x) - (x_i - x_0) S_1(x)}{S_2(x) S_0(x) - S_1(x)^2} \right\} y_i w_i(x) \tag{3.37}$$

Substituting $T_j(x)$ in equation (3.37), we get;

$$\hat{\beta}_0 = \frac{S_2(x)T_0(x) - T_1(x)S_1(x)}{S_2(x)S_0(x) - S_1(x)^2} \tag{3.38}$$

For $j = 1$, equation (3.12) becomes

$$\hat{y} = \hat{\beta}_0 + (x_i - x_0)\hat{\beta}_1 \tag{3.39}$$

where

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left\{ \left[\frac{S_2(x) - (x_i - x_0)\, S_1(x)}{S_2(x)S_0(x) - S_1(x)^2} \right] w_i(x) y_i \right\} \tag{3.40}$$

$$\hat{\beta}_1 = \sum_{i=1}^{n} \left\{ \left[\frac{(x_i - x_0)S_0(x) - S_1(x)}{S_2(x)S_0(x) - (S_1(x))^2} \right] w_i(x) y_i \right\} \tag{3.41}$$

Thus, the local linear estimator for the regression function $m(x)$ is given by

$$\hat{m}_1(x) = \hat{\beta}_0 + (x_i - x_0)\hat{\beta}_1 \tag{3.42}$$

$$\hat{m}_1(x) = \sum_{i=1}^{n} \left\{ \frac{S_2(x) - S_1(x)\,(x_i - x_0)}{S_2(x)S_0(x) - S_1(x)^2} \right\} w_i(x) y_i + (x_i - x_0) \sum_{i=1}^{n} \left\{ \frac{(x_i - x_0)S_0(x) - S_1(x)}{S_2(x)S_0(x) - (S_1(x))^2} \right\} w_i(x) y_i \tag{3.43}$$

Using the idea of Xu et al. (2008), the estimator, $\hat{m}_1(x)$ is given by the conditional expectation, $\hat{\varphi}(y_{t-1})$ which is used to impute the missing values, i.e.

$$\hat{\varphi}_{t,t-1}(y_{t-1}) = \sum_{i=1}^{n} \left\{ \frac{[S_2(x) - S_1(x)\,(x_i - x_0)]\,\omega_i \mathbf{I_i}}{[S_2(x)S_0(x) - S_1(x)^2]\,\omega_i \mathbf{I_i}} \right\} w_i(x) y_i$$

$$+ (x_i - x_0) \sum_{i=1}^{n} \left\{ \frac{[(x_i - x_0)\, S_0(x) - S_1(x)]\, \omega_i \mathbf{I_i}}{[S_2(x) S_0(x) - S_1(x)^2]\, \omega_i \mathbf{I_i}} \right\} w_i(x) y_i \tag{3.44}$$

where ω_i, is the weight according to the survey design (survey weight) and $\mathbf{I_{t,t-1,i}}$ is as defined earlier.

3.3.1 The local polynomial regression estimator

In general case, we define;

$$\beta(x_0) = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \; y_i = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \text{ and } X = \begin{pmatrix} 1 & (x_1 - x_0) & \cdots & (x_1 - x_0)^p \\ 1 & (x_2 - x_0) & \cdots & (x_2 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (x_n - x_0) & \cdots & (x_n - x_0)^p \end{pmatrix}$$

$\beta(x_0)$ is regression coefficient vector, y_i refer to the N-vector of $y_i's$ in the finite population of size N and X is $n \times (p+1)$ design matrix of known auxiliary data of unit $i \epsilon n$,

$W = diag\left\{K\left(\frac{x_i - x_0}{h}\right)\right\}$, its the $(n \times n)$ diagonal matrix of weights.

Now, Let $\hat{\beta}_j(x_i)$ for $j = 0, 1, \cdots, p$ and $i = 1, 2. \cdots, n$ represent the solution to the least squares problem in equation (3.15). If $(X^T W X)$ is a singular matrix, then , $\hat{\beta}_j(x)$ is given by

$$\hat{\beta}(x_0) = (X^T W X)^{-1} X^T W Y \tag{3.45}$$

Equation (3.45) becomes

$$\hat{\beta}(x_0) = \begin{pmatrix} S_0(x_0) & S_1(x_0) & \cdots & S_p(x_0) \\ S_1(x_0) & S_2(x_0) & \cdots & S_{p+1}(x_0) \\ \vdots & \vdots & \vdots & \vdots \\ S_p(x_0) & S_{p+1}(x_0) & \cdots & S_{2p}(x_0) \end{pmatrix}^{-1} \begin{pmatrix} T_0(x_0) \\ T_1(x_0) \\ \vdots \\ T_p(x_0) \end{pmatrix} = S_j^{-1}(x_0) T_j(x_0) \tag{3.46}$$

where $S_n^{-1} T_n$ is a $p \times p$ matrix define by

$$S_j(x_0) = \sum_{i=1}^{n} (x_i - x_0)^j K\left(\frac{x_i - x_0}{h}\right)$$

for $j = 0, 1, \cdots, 2p$

and for $j = 0, 1, \cdots, p$

$$T_j(x_0) = \sum_{i=1}^{n} (x_i - x_0)^j K\left(\frac{x_i - x_0}{h}\right) y_i$$

Let e_i be a vector of appropriate length with 1 (one) in the i^{th} position and zero's elsewhere.

Then,

$$\hat{\beta}_j(x_0) = e_i^T \hat{\beta}_j(x_0) \tag{3.47}$$

$e_i = (0, 0, \cdots, 0, 1, 0, \cdots, 0)^T$ with 1at the i^{th} position.

The local polynomial estimator $\hat{m}_p(x)$ is the intercept term $\hat{\beta}_j(x)$ of the weighted least squares estimator. The LPR estimator of the imputation model in equation (3.6) is

$$\hat{m}_p(x) = e_i^T (X^T W X)^{-1} X^T W Y. \tag{3.48}$$

Since, under the last value dependent mechanism, the conditional expectation $\varphi_{t,t-1}(y_{t-1})$ is estimated by the local polynomial regression estimator, thus,

$$\hat{\varphi}_{t,t-1}(y_{t-1}) = e_i^T S_j(x_0)^{-1} T_j(x_0) \tag{3.49}$$

The resulting weighted least-squares regression fit solution is derived from the sum of squared residuals as used for local linear estimator in the equations above and prediction can be done. Using the nonparametric method for nonlinear pattern, the regression function $\varphi_{t,t-1}(x)$ for $t = 2, 3, \cdots, T$ as stated in the work of Xu et al. (2008), the kernel estimator of the regression $\varphi_{t,t-1}(x_0)$ is obtained from the local polynomial regression model.

Under the regression imputation model derived above, missing values may be replaced by the best linear unbiased predicted value as the correct response value. By the prediction theory, see Anthony (1999) and Kyuseong (2000), the best linear unbiased predictor for $m(x_i)$ is;

$$\hat{m}_p(x_r) = e_i^T (X^T W X)^{-1} X^T W Y. \tag{3.50}$$

So, the non respondents are imputed and the imputed data set is given by ;

$$y_t^{\#} = \begin{cases} y_t & ; I_t = 1 \qquad i \epsilon B_r \\ e_i^T (X^T W X)^{-1} X^T W Y & ; I_t = 0 \qquad i \epsilon B_{n-r} \end{cases} \tag{3.51}$$

Using the imputed data, the population mean estimation is done. The sample mean

of the imputed data is given by

$$\bar{y}_I = \frac{1}{n}\left\{\sum_{i\epsilon B_r} y_{i,t} + \sum_{i\epsilon B_{n-r}} \hat{m}_p(x_r)\right\} \tag{3.52}$$

3.3.2 Estimation of the finite population means using the imputed data

In this study, we consider a finite population from which samples are drawn. Before estimation of the population parameters, imputations of missing data is carried out and imputed data set is then used to do standard statistical analyses

Suppose that the survey measurements are $y_{1,}y_{2,}\cdots,y_N$ on the variables $B_1, B_{2,}\cdots,B_N$ respectively and a simple random sample without replacement, B_n, of size n is selected from a finite population, p of size N. Under SRSWOR, each unit has got equal chances of being selected. The sample consists of two parts: B_r and B_{n-r} , where B_r is the set of all respondents in the survey and B_{n-r} is the set of all non respondents. The missing observations of the sample unit $y_{i,t}$, for $t \geq 2$ are considered. Imputation of the missing value $y_{i,t}$ for $i\epsilon B_{n-r}$ and $t \geq 2$ is done and then a complete data set is produced which is then used in the estimation of finite population means.

Let $\bar{Y}_t$ be the finite population mean at time point, t for $t = 1, 2, \cdots, T$

The mean of the finite population is given by;

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \tag{3.53}$$

The value to be imputed for the non respondent is denoted by $\hat{y}_{i,t}$ such that the imputed data computed from the imputation approach is given as;

$$y_{i,t}^{\#} = \begin{cases} y_{i,t}, & i\epsilon B_r \quad observed \quad value \\ y_{i,t}^{*}, & i\epsilon B_{n-r} \quad imputed \quad value \end{cases} \tag{3.54}$$

Now, using the imputed data, the estimator of the finite population total is the sample total of the imputed data denoted by y_I and is given by;

$$y_I = \sum_{i\epsilon S} \left[y_{i,t} I_i + (1 - I_i)\, y_{i,t}^{*}, \right] \tag{3.55}$$

Thus, using the imputed data, the estimator of the finite population mean is the sample mean of the imputed data denoted by $\bar{y}_I$ and is given by;

$$\bar{y}_I = \sum_{i\epsilon s} \omega_i y_{i,t}^{\#} \tag{3.56}$$

We assume that, the survey weights denoted by ω_i are defined such that for any set in the sample,s, i.e. $i\epsilon s$,

$$E_s \left(\sum_{i\epsilon s} \omega_i y_i \right) = \sum_{i=1}^{N} y_i \tag{3.57}$$

for each $i\epsilon p$.

The imputed values are then treated as if they were observed such that both observed and the imputed are used to estimate the population mean.;

Therefore, sample mean for the imputed data becomes;

$$\bar{y}_I = \left\{ \sum_{i\epsilon B_r} w_i y_{i,t} + \sum_{i\epsilon B_{n-r}} w_i y_{i,t}^{*} \right\} \tag{3.58}$$

We note that the same weight due to sampling design is used in equation (3.58) for all units in the sample. So;

$$\bar{y}_I = \frac{1}{n}\left\{\sum_{i\epsilon B_r} y_{i,t} + \sum_{i\epsilon B_{n-r}} y^*_{i,t}\right\} \tag{3.59}$$

Using the sample data, prediction of the nonrespondents is done based on the local polynomial regression estimator. Under the nonmonotone nonresponse mechanism assumption for the total time points, $t = 1, \ldots, T.$,

$$\hat{\bar{y}}_t = \frac{1}{n}\left(\sum_{i\epsilon B_r} y_i + \sum_{i\in B_{n-r}} y^*_i\right) \tag{3.60}$$

where B_r is the respondents set, B_{n-r} is the non respondents set and y^*_i is equal to the value imputed if the indicator $I_i = 0$ and y_i is observed value if $I_i = 1$.

Thus, according to the nonparametric approach of Cheng (1994), we have the kernel estimator given by equations (3.61) and the local linear estimator given by equation (3.62),

$$y^*_i = \hat{\varphi}_{t,t-1}(y_{t-1}) = \frac{\sum_{i\epsilon S} K\left(\frac{x-y_{i,t-1}}{h}\right)\omega_i \mathbf{I_{i,t,t1}} y_{i,t}}{\sum_{i\epsilon s} K\left(\frac{x-y_{i,t-1}}{h}\right)\omega_i \mathbf{I_{i,t,t-1}}} \tag{3.61}$$

$$y^*_i = \hat{\varphi}_{t,t-1}(y_{t-1}) = \sum_{i=1}^{n}\left\{\frac{[S_2(x) - S_1(x)\,(x_i - x_0)]\,\omega_i \mathbf{I_{i,t,t-1}}}{[S_2(x)S_0(x) - S_1(x)^2]\,\omega_i \mathbf{I_{i,t,t-1}}}\right\} w_i(x) y_i$$

$$+ (x_i - x_0)\sum_{i=1}^{n}\left\{\frac{[(x_i - x_0)\,S_0(x) - S_1(x)]\,\omega_i \mathbf{I_{i,t,t-1}}}{[S_2(x)S_0(x) - S_1(x)^2]\,\omega_i \mathbf{I_{i,t,t-1}}}\right\} w_i(x) y_i \tag{3.62}$$

$$\text{where } \mathbf{I_{i,t,t-1}} = \begin{cases} 1, & I_{i,t} = 1, I_{i,t-1} = 1, \quad for\ t = 2, \cdots, T \\ 0, & otherwise \end{cases}$$

$$\hat{\bar{y}}_t = \frac{1}{n}\left(\sum_{i\epsilon B_r} y_i + \sum_{i\in B_{n-r}}\left[\frac{\sum_{i\epsilon S} w_i(x)\omega_i \mathbf{I_{i,t,t1}} y_{i,t}}{\sum_{i\epsilon s} w_i(x)\omega_i \mathbf{I_{i,t,t-1}}}\right]\right) \tag{3.63}$$

$$\hat{\bar{y}}_t = \frac{1}{n}\left\{\sum_{i\epsilon B_r} y_i + \sum_{i\epsilon B_{n-r}}\sum_{i=1}^{n}\left\{\frac{\left[S_2(x) - S_1(x)\left(x_i - x_0\right)\right]\omega_i \mathbf{I_{i,t,t-1}}}{\left[S_2(x)S_0(x) - S_1(x)^2\right]\omega_i \mathbf{I_{i,t,t-1}}}\right\} w_i(x) y_i\right\}$$

$$+ \frac{1}{n}\left\{\sum_{i\epsilon B_{n-r}}\left(x_i - x_0\right)\sum_{i=1}^{n}\left\{\frac{\left[\left(x_i - x_0\right)S_0(x) - S_1(x)\right]\omega_i \mathbf{I_{i,t,t-1}}}{\left[S_2(x)S_0(x) - S_1(x)^2\right]\omega_i \mathbf{I_{i,t,t-1}}}\right\} w_i(x) y_i\right\} \tag{3.64}$$

Under superpopulation model, the finite population mean are estimated from the x variables with a view that, the local polynomial regression estimator shall be used in determining the nonrespondents mean.

3.4 Asymptotic properties of the estimator

In order to investigate the asymptotic properties of the proposed estimator, we use the Taylor series approximation of the fitted values of the smooth function $\hat{m}(x_i)$. In derivation process, we adopt a set of regularity conditions as established by Krewski and Rao (1981), Cheng (1994). According to Cheng (1994), the development of asymptotic theory is given by the concept of a sequence of finite populations $\{P_\nu\}_{\nu=1}^{\infty}$ with ν strata in P_ν. We assume that there is a sequence of finite populations and the corresponding sequence of samples. The finite population P indexed by ν is assumed to be a member of the sequence of the populations. The sample size denoted by n_ν and the population size denoted by N_ν approach infinity as $\nu \to \infty$. Assuming uni-

form response and that size m_ν of the nonrespondents set B_{n-r} satisfy the condition $\frac{m_\nu}{n_\nu} \to \alpha < 1$. For easy notation, the subscript ν will be ignored in the subsequent work. All limiting processes will be understood as $\nu \to \infty$ such that the following regularity conditions below are satisfied;

Denote f to be a probability density function (pdf) of X and define $g(x) = p(x)f(x)$ where $p(x)$ is defined by;

$$p(x) = P(I = 1|Y, X) = P(I = 1|X) \tag{3.65}$$

We take g and f to have bounded second derivatives

(i)The Kernel density function K is a bounded and twice continuously differentiable symmetric function.

Let $K(u) = K(-u) \cdot 1(u < 0) + K(u) \cdot 1(u \geq 0)$. This shows that the kernel function $K(\cdot)$ is symmetric, bounded, nonnegative and continuous on the interval $[-1, 1]$. It also satisfies the conditions below; $k_0 = \int K(u)du = 1$, $k_1 = \int uK(u)du = 0$, $k_2 = \int u^2K(u)$, $k_2 < \infty$ and $\int_{-\infty}^{\infty} \{K(u)\}^2 du < \infty$

(ii)The regression function $m(\cdot)$ is at least twice continuously differentiable every where in the neighborhood of x_0

(iii)The sample survey variable of interest has a finite second moments bounded on the interval $(0, 1)$, thus, $E(y^2) < \infty$ and $E(P(x)) = P(I = 1)$ with $E(\sigma^2(x)|P(x)) < \infty$

(iv)The conditional variance $\sigma^2(x_i) = Var(y_i|X = x_i)$ is bounded and continuous.

Theorem

Assuming conditions $(i)-(iv)$ and also the assumptions described in section (3.1.1) of this chapter hold. Also, suppose B_n is simple random sample with sample size n and g is an ignorable response mechanism. Then under the regression imputation model η (equation (3.6)), the sample mean of the imputed data, $\bar{y}_I$ is asymptotically unbiased and consistent for the population mean $\bar{Y}_t$.

Proof

In order to prove the theorem, we expect to obtain two main results, that is ; sample mean with nonrespondents imputed by the local constant estimator and sample mean with nonrespondents imputed by the local linear estimator. Then, the results of Lemma 1 and Lemma 2 are used in the proof of the theorem.

Proof of the theorem using the local constant estimator

Lemma 1

The bias of $\hat{m}_0(x)$ is given by

$$Bias\left(\hat{m}_0(x)\right) = \frac{1}{2}m''(x_0)h^2k_2 + o(h^2) \tag{3.66}$$

Under the regularity conditions in section 3.4, $Bias\left(\hat{m}_1(x)\right) \longrightarrow 0$ as $h \to 0$ and $n \to \infty$.

Proof of Lemma 1

For $j = 0$, the estimator is given by

$$\hat{m}_0(x) = \sum_{i=1}^{n} w_i(x)y_i \tag{3.67}$$

where $w_i(x) = K\left(\frac{x_i - x_0}{h}\right)$

Expectation of the local constant estimator, $\hat{m}_0(x)$ is given by

$$E\left(\hat{m}_0(x)\right) = \sum_{i=1}^{n} w_i(x) E\left(y_i\right) \tag{3.68}$$

$$= \sum_{i=1}^{n} w_i(x) m\left(x_i\right) \tag{3.69}$$

Using the Taylor series expansion of $m(x_i)$ about x_0 , we have

$$m(x_i) = m(x_0 + uh) \tag{3.70}$$

where $u = \frac{x_i - x_0}{h}$ so that $uh = (x_i - x_0)$ with $m(x_i)$ being a smooth curve which is at least twice continuously differentiable

$$m(x_i) \approx m(x_0) + uhm'(x_0) + \frac{1}{2}u^2h^2m''(x_0) + \cdots$$

$$m(x_i) \approx m(x_0) + m'(x_0)(x_i - x_0) + \frac{m''(x_0)(x_i - x_0)^2}{2} + \cdots$$

$$E\left(\hat{m}_0(x)\right) = \sum_{i=1}^{n} w_i(x) \left\{ m(x_0) + m'(x_0)(x_i - x_0) + \frac{m''(x_0)(x_i - x_0)^2}{2} + \cdots \right\} \tag{3.71}$$

Using Theorem 3 of Fan and Gijbels (1996), and the work of Eubank and Speckman (1993) and Masry (1996), under the assumption that the auxiliary variables are fixed uniform design points on the interval $(0, 1)$. we have

$$E\left(\hat{m}_0(x)\right) = \int K(u)m(x_0)du + \int m'(x_0)(x_i - x_0)K(u)du + \int \frac{1}{2}m''(x_0)(x_i - x_0)^2K(u)du + o(h^2) \tag{3.72}$$

$$E\left(\hat{m}_0(x)\right) = \int K(u)m(x_0)du + \int m'(x_0)huK(u)du + \int \frac{1}{2}m''(x_0)h^2u^2K(u)du + o(h^2) \tag{3.73}$$

$$E\left(\hat{m}_0(x)\right) = m(x_0)\int K(u)du + m'(x_0)h\int uK(u)du + \frac{1}{2}m''(x_0)h^2\int u^2K(u)du + o(h^2) \tag{3.74}$$

Using the regularity conditions of a symmetric kernel function $K(u)$, we get

$$E\left(\hat{m}_0(x)\right) = m(x_0) + \frac{1}{2}m''(x_0)h^2\int u^2K(u)du + o(h^2)$$

$$E\left(\hat{m}_0(x)\right) = m(x_0) + \frac{1}{2}m''(x_0)h^2k_2 + o(h^2)$$

where $k_2 = \int u^2K(u)du$

Now,

$$Bias\left(\hat{m}_0(x)\right) = \frac{1}{2}m''(x_0)h^2k_2 + o(h^2) \tag{3.75}$$

Assuming that $h \to 0$ as $n \to \infty$, $m(.)$ is bounded as stated in condition (ii) and $k_2 < \infty$

$$Bias\left(\hat{m}_0(x)\right) \to 0 \tag{3.76}$$

Lemma 2

The asymptotic variance of $\hat{m}_0(x)$ **is given by**

$$Var_{asy}\left(\hat{m}_0(x)\right) = \sum_{i=1}^{n} \frac{d_k}{nh}\sigma^2(x_0) \tag{3.77}$$

where $d_k = \int K^2(u)du$,

Proof of Lemma 2

From the expression of the estimator

$$\hat{m}_0(x) = \sum_{i=1}^{n} w_i(x)y_i \tag{3.78}$$

$$Var\left(\hat{m}_0(x)\right) = Var_{\eta}\left(\sum_{i=1}^{n} w_i(x)y_i\right) \tag{3.79}$$

$$Var\left(\hat{m}_0(x)\right) = \sum_{i=1}^{n} w_i^2(x)Var(y_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i(x)Cov(y_{i,}y_j) \tag{3.80}$$

since $Cov(y_{i,}y_j) = 0$, we have

$$Var\left(\hat{m}_0(x)\right) = \frac{\sum_{i=1}^{n} w_i^2(x)\sigma^2(x_i)}{\sum_{i=1}^{n} w_i(x)} \tag{3.81}$$

Expanding $\sigma^2(x_i)$ using Taylor series approximation, we get

$$\sigma^2(x_i) \approx \sigma^2(x_0) + \sigma^{2'}(x_0)(x_i - x_0) + \frac{\sigma^{2''}(x_0)(x_i - x_0)^2}{2} + \cdots \tag{3.82}$$

$$\sigma^2(x_i) \approx \sigma^2(x_0) + \sigma^{2'}(x_0)uh + \frac{1}{2}\sigma^{2''}(x_0)u^2h^2 + \cdots \tag{3.83}$$

Substituting $\sigma^2(x_i)$ in equation (3.83) into equation (3.81)

$$Var\left(\hat{m}_0(x)\right) = \sum_{i=1}^{n} w_i^2(x)\left\{\sigma^2(x_0) + \sigma^{2'}(x_0)uh + \frac{1}{2}\sigma^{2''}(x_0)u^2h^2 + \cdots\right\} \tag{3.84}$$

$$Var\left(\hat{m}_0(x)\right) = \sum_{i=1}^{n} w_i^2\left[\sigma^2(x_0)\right] + \sum_{i=1}^{n} w_i^2\left[\sigma^{2'}(x_0)uh\right] + \sum_{i=1}^{n} w_i^2\left[\frac{1}{2}\sigma^{2''}(x_0)u^2h^2\right] + \cdots \tag{3.85}$$

But, $w_i^2(x) = \left\{K\left(\frac{x_i - x_0}{h}\right)\right\}^2 = K^2\left(\frac{x_i - x_0}{h}\right)$

$$Var\left(\hat{m}_0(x)\right) = \int K^2(u)\sigma^2(x_0)du + \int K^2(u)\left[\sigma^{2'}(x_0)uh\right]du$$

$$+ \int K^2(u)\left[\frac{1}{2}\sigma^{2''}(x_0)u^2h^2\right]du + o(h^2) \tag{3.86}$$

Assuming that the second derivative of $\sigma^2(x_0)$ is bounded and continuous , $h \to 0$ and $nh \to \infty$ as $n, N \to \infty$ then, the variance of the estimator is asymptotically estimated by

$$Var_{asy}\left(\hat{m}_0(x)\right) = \frac{d_k}{nh}\sigma^2(x_0) \tag{3.87}$$

where $d_k = \int K^2(u)du$,

The Mean Square Error (MSE) **of the estimator** $\hat{m}_0(x)$

$$MSE\left(\hat{m}_0(x)\right) = \left\{Bias\left(\hat{m}_0(x)\right)\right\}^2 + Var\left(\hat{m}_0(x)\right) \tag{3.88}$$

Now, substituting the results of the $Bias\left(\hat{m}_0(x)\right)$ and $Var\left(\hat{m}_0(x)\right)$, we get

$$MSE(\hat{m}_0(x)) = \left\{\frac{1}{2}m''(x_0)h^2k_2 + o(h^2)\right\}^2 + \frac{1}{nh}\sigma^2(x_0)d_k \tag{3.89}$$

which is the asymptotic expression of MSE. $MSE\left(\hat{m}_1(x)\right) \to 0$ under the assumption $h \to 0$ and $nh \to \infty$ as $n \to \infty$, $m(\cdot)$ is bounded as stated in condition (ii) and $k_2 < \infty$,

Consequently, $\hat{m}_0(x)$ is asymptotically unbiased and consistent.

Proof of the theorem

Consider the population total $Y_t = \sum_{i=1}^{N} y_i$ having the sample set and the non sampled set such that

$$Y_t = \sum_{i\epsilon B_n} y_i + \sum_{i\epsilon B_{N-n}} y_i \tag{3.90}$$

The population total can be rewritten as

$$Y_t = \left(\sum_{i\epsilon B_r} y_{i,t} + \sum_{i\epsilon B_{(n-r)}} y^*_{i,t}\right) + \sum_{i\epsilon B_{N-n}} y_i \tag{3.91}$$

For writing simplicity, denote $i\epsilon B_r$, $i\epsilon B_{(n-r)}$ and $i\epsilon B_{(N-n)}$ for the respondents and nonrespondents by $i\epsilon r$, $i\epsilon(n-r)$ and $i\epsilon(N-n)$ respectively throughout the following work.

From equation (3.6), the estimator for finite population total is given by

$$y_I = \sum_{i\epsilon r} y_i + \sum_{i\epsilon(n-r)} y_i^* \tag{3.92}$$

Let $\hat{Y}_t = y_I$ such that

$$\hat{Y}_t = \sum_{i\epsilon r} y_i + \sum_{i\epsilon(n-r)} y_i^* \tag{3.93}$$

Replacing y_i^* by the corresponding nonparametric estimator $\hat{m}(x_i)$, we have;

$$\hat{Y}_t = \sum_{i\epsilon r} y_i + \sum_{i\epsilon(n-r)} \hat{m}(x_i) \tag{3.94}$$

Bias of $\hat{Y}_t$

Consider the difference,

$$\hat{Y}_t - Y_t = \left(\sum_{i\epsilon r} y_i + \sum_{i\epsilon(n-r)} \hat{m}(x_i)\right) - \left(\sum_{i\epsilon r} y_i + \sum_{i\epsilon(n-r)} y_i\right) - \sum_{i\epsilon N-n} y_i$$

$$\hat{Y}_t - Y_t = \sum_{i\epsilon(n-r)} (\hat{m}(x_i) - y_i) - \sum_{i\epsilon N-n} y_i \tag{3.95}$$

Introducing $m(x)$ in equation (3.95), we get

$$\hat{Y}_t - Y_t = \sum_{i\epsilon(n-r)} ([\hat{m}(x_i) - m(x_i)] + [m(x_i) - y_i]) - \sum_{i\epsilon N-n} y_i \tag{3.96}$$

Taking the conditional expectation on both sides, we have

$$E\left(\hat{Y}_t - Y_t | X = x_i\right) = \sum_{i\epsilon(n-r)} \left(E_\eta \left[\hat{m}(x_i) - m(x_i) | X = x_i\right] + \sum_{i\epsilon(n-r)} E_\eta \left[m(x_i) - y_i | X = x_i\right] \right)$$

$$- \sum_{i\epsilon N-n} E_\eta \left(y_i | X = x_i\right) \tag{3.97}$$

Clearly, $\sum_{i\epsilon(n-r)} E_\eta \left[m(x_i) - y_i | X = x_i\right] = 0$ since $\sum_{i\epsilon(n-r)} E_\eta \left(y_i | X = x_i\right) = m(x_i)$

Now,

$$E_\eta \left(\hat{Y}_t - Y_t | X = x_i\right) = \sum_{i\epsilon(n-r)} E_\eta \left[\hat{m}(x_i) - m(x_i) | X = x_i\right] - \sum_{i\epsilon N-n} E_\eta(y_i) \tag{3.98}$$

$$E_\eta \left(\hat{Y}_t - Y_t | X = x_i\right) = \sum_{i\epsilon(n-r)} E_\eta \left[\hat{m}(x_i) - m(x_i) | X = x_i\right] - \sum_{i\epsilon N-n} m(x_i) \tag{3.99}$$

According to Fan and Gijbels (1996) y_i and x_i are regressed using the weighted least squares equation. Approximation of $m_{y_{t-1}}(x_t)$ about y_{t-1} is done following the Taylor series expansion, thus; letting the last observed value (point of focus) $y_{t-1} = x_0$, we have

$$\hat{m}(x_i) - m(x_0) \approx m'(x_0)(x_i - x_0) + \frac{1}{2} m''(x_0)(x_i - x_0)^2 + \cdots \tag{3.100}$$

Taking conditional expectation on both sides, we have

$$E_\eta\left(\hat{m}(x_i) - m(x_0)\right) = E_\eta\left(m'(x_0)(x_i - x_0) + \frac{1}{2}m''(x_0)(x_i - x_0)^2 + \cdots\right) \quad (3.101)$$

Let $u = \frac{x_i - x_0}{h} \Rightarrow (x_i - x_0) = uh$

$$E_\eta\left(\hat{m}(x_i) - m(x_0)\right) = E_\eta\left(uhm'(x_0) + \frac{1}{2}(uh)^2 m''(x_0) + \cdots\right) \quad (3.102)$$

Theorem 3 of Fan and Gijbels (1996) , under conditions $(i) - (iv)$, and$h \to 0$ and $nh \to \infty$ as $n \to \infty$ for each ν such that the right hand side of equation (3.101) becomes

$$E_\eta\left(uhm'(x_0) + \frac{1}{2}(uh)^2 m''(x_0) + \cdots\right) \to m'(x_0)\int (uh)K(u)du$$

$$+ \frac{1}{2}m''(x_0)\int (uh)^2 K(u)du + o(h^2) \quad (3.103)$$

$$= m'(x_0)h\int uK(u)du + \frac{1}{2}m^{''}(x_0)h^2\int u^2K(u)du + o(h^2) \quad (3.104)$$

Since $\int uK(u)du = 0$ and $k_2 = \int u^2K(u)du$

Therefore,

$$E_\eta\left(uhm'(x_0) + \frac{1}{2}(uh)^2 m''(x_0) + \cdots\right) \to \frac{1}{2}h^2 m^{''}(x_0)k_2 + o(h^2) \quad (3.105)$$

Substituting the expression of equation (3.105) into equation (3.99)

$$E_\eta\left(\hat{Y}_t - Y_t|X = x_i\right) \to \sum_{i\epsilon(n-r)}\left(\frac{1}{2}h^2 m^{''}(x_0)k_2 + o(h^2)\right) - \sum_{i\epsilon N-n}\bar{Y} \tag{3.106}$$

$$E_\eta\left(\hat{Y}_t - Y_t|X = x_i\right) \to \sum_{i\epsilon(n-r)}\left(\frac{1}{2}h^2 m^{''}(x_0)k_2 + o(h^2)\right) - (N-n)\bar{Y} \tag{3.107}$$

Making the assumption that as $n, N \to \infty$, $N - n \to 0$, then, the non sampled units tend to zero in equation (3.107), thus

$$E_\eta\left(\hat{Y}_t - Y_t|X = x_i\right) \to \sum_{i\epsilon(n-r)}\left(\frac{1}{2}h^2 m^{''}(x_0)k_2 + o(h^2)\right) \tag{3.108}$$

Under the assumptions that $h \to 0$, as $n \to \infty$, $m(\cdot)$ is bounded as stated in condition (ii) and $k_2 < \infty$ then,equation (3.108) becomes

$$E_\eta\left(\hat{Y}_t - Y_t|X = x_i\right) \to \sum_{i\epsilon(n-r)}[0] \tag{3.109}$$

$$E_\eta\left(\hat{Y}_t - Y_t|X = x_i\right) \to 0 \tag{3.110}$$

Clearly,it shows in equation (3.110) that the sample total of the imputed data is an asymptotically unbiased estimator of the finite population total.

Variance of $\hat{Y}_t$ using $\hat{m}_0(x_i)$

Variance of the $\hat{Y}_t$ is estimated using the Variance of the $\hat{Y}_t - Y_t$

Now, the difference,

$$\hat{Y}_t - Y_t = \left(\sum_{i\epsilon r} y_i + \sum_{i\epsilon(n-r)}\hat{m}(x_i)\right) - \left(\sum_{i\epsilon r} y_i + \sum_{i\epsilon(n-r)} y_i\right) - \sum_{i\epsilon N-n} y_i$$

$$\hat{Y_t} - Y_t = \sum_{i\epsilon(n-r)} (\hat{m}(x_i) - y_i) - \sum_{i\epsilon N-n} y_i \tag{3.111}$$

$$\hat{Y_t} - Y_t = \sum_{i\epsilon(n-r)} \hat{m}(x_i) - \sum_{i\epsilon(n-r)} y_i - \sum_{i\epsilon N-n} y_i$$

$$\hat{Y_t} - Y_t = \sum_{i\epsilon(n-r)} \sum_{i\epsilon r} w_i(x) y_i - \sum_{i\epsilon(n-r)} y_i - \sum_{i\epsilon N-n} y_i$$

Taking Variance on both sides, we get

$$Var\left\{\hat{Y_t} - Y_t\right\} = Var\left\{ \sum_{i\epsilon(n-r)} \sum_{i\epsilon r} w_i(x) y_i - \sum_{i\epsilon(n-r)} y_i - \sum_{i\epsilon N-n} y_i \right\} \tag{3.112}$$

$$= \sum_{i\epsilon(n-r)} \sum_{i\epsilon r} w_i^2(x) Var\left(y_i\right) - \sum_{i\epsilon(n-r)} Var\left(y_i\right) - \sum_{i\epsilon N-n} Var\left(y_i\right)$$

$$= \sum_{i\epsilon(n-r)} \sum_{i\epsilon r} w_i^2(x) \sigma_i^2(x) - \sum_{i\epsilon(n-r)} \sigma_i^2(x) - \sum_{i\epsilon N-n} \sigma_i^2(x) \tag{3.113}$$

Using Taylor series approximation, we have

$$Var\left\{\hat{Y_t} - Y_t\right\} = \sum_{i\epsilon(n-r)} \sum_{i\epsilon r} w_i^2(x) \left\{ \sigma^2(x_0) + \sigma^{2'}(x_0) uh + \frac{1}{2} \sigma^{2''}(x_0) u^2 h^2 + \cdots \right\} -$$

$$\sum_{i\epsilon(n-r)} \left\{ \sigma^2(x_0) + \sigma^{2'}(x_0)uh + \frac{1}{2}\sigma^{2''}(x_0)u^2h^2 + \cdots \right\} - \sum_{i\epsilon N-n} \sigma_i^2(x) \tag{3.114}$$

Under the assumptions that $h \to 0$, $nh \to \infty$ as $n \to \infty$, $m(.)$ is bounded as stated in condition (II) and $k_2 < \infty$. we have also considered the assumption that $n, N \to \infty$ and $N - n \to 0$, implying that the non sampled units diminishes. Thus,

$$Var\left\{\hat{Y}_t - Y_t\right\} = \sum_{i\epsilon(n-r)} \sum_{i\epsilon r} w_i^2(x)\sigma^2(x_0) + \sum_{i\epsilon(n-r)} \sum_{i\epsilon r} w_i^2 \frac{1}{2}\sigma^{2''}(x_0)u^2h^2$$

$$- \sum_{i\epsilon(n-r)} \sigma^2(x_0) + \sum_{i\epsilon(n-r)} \frac{1}{2}\sigma^{2''}(x_0)u^2h^2 + \cdots \tag{3.115}$$

Using the results of the local

$$Var\left\{\hat{Y}_t - Y_t\right\} = \sum_{i\epsilon(n-r)} \sum_{i\epsilon r} w_i^2(x)\sigma^2(x_0) - \sum_{i\epsilon(n-r)} \sigma^2(x_0) \tag{3.116}$$

Using the results of the variance of $\hat{m}_0(x)$, the asymptotic expression of the variance of the estimator becomes;

$$Var\left\{\hat{Y}_t - Y_t\right\} \approx \frac{1}{nh} \sum_{i\epsilon(n-r)} \sum_{i\epsilon r} w_i^2(x)\sigma^2(x_0) \tag{3.117}$$

$$= \sum_{i\epsilon(n-r)} \frac{d_k}{nh}\sigma(x_0) \tag{3.118}$$

where $d_k = \int K^2(u)du$

The Mean Square Error (MSE) **of the estimator** $\hat{Y}_t$ **using** $\hat{m}_0(x_i)$ **results**

$$MSE\left(\hat{Y}_t\right) = \left\{Bias\left(\hat{Y}_t\right)\right\}^2 + Var_\eta\left(\hat{Y}_t\right) \tag{3.119}$$

Thus,

$$MSE\left[\hat{Y}_t\right] = \left\{\frac{1}{2}m''(x_0)h^2k_2 + o(h^2)\right\}^2 + \sum_{i\epsilon(n-r)} \frac{1}{nh} d_k\sigma^2(x_0) \tag{3.120}$$

which is the asymptotic expression of MSE. $MSE\left(\hat{m}_1(x)\right) \to 0$ under the assumption $h \to 0$ and $nh \to \infty$ as $n \to \infty$, $m(\cdot)$ is bounded as stated in condition (ii) and $k_2 < \infty$,

Consequently, $\hat{m}_0(x)$ is asymptotically unbiased and consistent. Therefore,using the notion of the Law of Large Numbers, we get the weak consistency which is a desirable property

Lemma 3

The bias of $\hat{m}_1(x)$ is given by

$$Bias\left(\hat{m}_1(x)\right) = uhm'(x_0) + \frac{h}{2k_2}\left(hk_2^2 + (uh)k_3\right)m''(x_0) \tag{3.121}$$

Under the regularity conditions in section 3.4, $Bias\left(\hat{m}_1(x)\right) \longrightarrow 0$ as $h \to 0$ and $n \to \infty$.

Proof of Lemma 3

From

$$\hat{m}_1(x) = \sum_{i=1}^{n}\left\{\left(\frac{s_2(x) - (x_i - x_0)\, s_1(x)}{s_2(x)s_0(x) - s_1(x)^2}\right) w_i(x)y_i\right\}$$

$$+ (x_0 - x_i) \sum_{i=1}^{n} \left\{ \left(\frac{(x_i - x_0)s_0(x) - s_1(x)}{s_2(x)s_0(x) - (s_1(x))^2} \right) w_i(x) y_i \right\} \tag{3.122}$$

Equation (3.122) can be written as

$$\hat{m}_1(x) = \sum_{i=1}^{n} w_i^*(x) y_i + (x_0 - x_i) \sum_{i=1}^{n} w_i^{**}(x) y_i \tag{3.123}$$

where $w_i^*(x) = \left(\frac{s_2(x) - (x_i - x_0)s_1(x)}{s_2(x)s_0(x) - s_1(x)^2} \right) w_i(x)$, $w_i^{**}(x) = \left(\frac{(x_i - x_0)s_0(x) - s_1(x)}{s_2(x)s_0(x) - (s_1(x))^2} \right) w_i(x)$

and $w_i(x) = K(\frac{x_i - x_0}{h})$

Expectation of $\hat{m}_1(x)$,

$$E\left[\hat{m}_1(x)\right] = \sum_{i=1}^{n} w_i^*(x) E\left[y_i\right] + (x_0 - x_i) \sum_{i=1}^{n} w_i^{**}(x) E\left[y_i\right] \tag{3.124}$$

$$E\left[\hat{m}_1(x)\right] = \sum_{i=1}^{n} \left\{ \left(\frac{s_2(x) - (x_i - x_0)\, s_1(x)}{s_2(x)s_0(x) - s_1(x)^2} \right) w_i(x) E\left[y_i\right] \right\} +$$

$$(x_0 - x_i) \sum_{i=1}^{n} \left\{ \left(\frac{(x_i - x_0)s_0(x) - s_1(x)}{s_2(x)s_0(x) - (s_1(x))^2} \right) w_i(x) E\left[y_i\right] \right\} \tag{3.125}$$

$$E\left[\hat{m}_1(x)\right] = \sum_{i=1}^{n} \left\{ \left(\frac{s_2(x) - (x_i - x_0)\, s_1(x)}{s_2(x)s_0(x) - s_1(x)^2} \right) w_i(x) m(x_i) \right\}$$

$$+ (x_0 - x_i) \sum_{i=1}^{n} \left\{ \left(\frac{(x_i - x_0)s_0(x) - s_1(x)}{s_2(x)s_0(x) - (s_1(x))^2} \right) w_i(x) m(x_i) \right\} \tag{3.126}$$

Using Taylor series expansion,

$$m(x_i) \approx m(x_0) + (x_i - x_0)m'(x_0) + \frac{1}{2}(x_i - x_0)^2 m''(x_0) + \cdots$$

Equation (3.126) becomes

$$E\left[\hat{m}_1(x)\right] = \sum_{i=1}^{n} \left\{ \frac{S_2(x) - (x_i - x_0)\, S_1(x)}{S_2(x)S_0(x) - S_1(x)^2} w_i(x) \left[m(x_0) + m'(x_0)(x_i - x_0) + \frac{m''(x_0)(x_i - x_0)^2}{2} + \cdots \right.\right.$$

$$\left.\left. + (x_0 - x_i) \sum_{i=1}^{n} \left\{ \frac{(x_i - x_0)S_0(x) - S_1(x)}{S_2(x)S_0(x) - (S_1(x))^2} w_i(x) \left[m(x_0) + m'(x_0)(x_i - x_0) + \frac{m''(x_0)(x_i - x_0)^2}{2} + \cdots \right] \right.\right. \tag{3.127}$$

$$E\left[\hat{m}_1(x)\right] = \frac{S_2(x)}{S_2(x)S_0(x) - S_1(x)^2} \left\{ S_0(x)m(x_0) + S_1(x)m'(x_0) + \frac{1}{2}S_2(x)m''(x_0) \right\}$$

$$- \frac{S_1(x)}{S_2(x)S_0(x) - S_1(x)^2} \left\{ S_1(x)m(x_0) + S_2(x)m'(x_0) + \frac{1}{2}S_3(x)m''(x_0) \right\}$$

$$+ \frac{(x_0 - x_i)S_0(x)}{S_2(x)S_0(x) - S_1(x)^2} \left\{ S_1(x)m(x_0) + S_2(x)m'(x_0) + \frac{1}{2}S_3(x)m''(x_0) \right\}$$

$$- \frac{(x_0 - x_i)S_1(x)}{S_2(x)S_0(x) - S_1(x)^2} \left\{ S_0(x)m(x_0) + S_1(x)m'(x_0) + \frac{1}{2}S_2(x)m''(x_0) \right\} \tag{3.128}$$

Collecting like terms, we get

$$E\left[\hat{m}_1(x)\right] = \left\{ \frac{\left[S_2(x)S_0(x) - s_1^2(x)\right] + \left[(x_0 - x_i)\left(S_0(x)S_1(x) - S_1(x)S_0(x)\right)\right]}{S_2(x)S_0(x) - S_1(x)^2} \right\} m(x_0)$$

$$+\left\{\frac{[S_1(x)S_2(x)-S_1(x)S_2(x)]+[(x_0-x_i)\,(S_0(x)S_2(x)-S_1^2(x))]}{S_2(x)S_0(x)-S_1(x)^2}\right\}m^{'}(x_0)$$

$$+\frac{1}{2}\left\{\frac{[S_2^2(x)-S_1(x)S_3(x)]+[(x_0-x_i)\,(S_0(x)S_3(x)-S_1(x)S_2(x))]}{S_2(x)S_0(x)-S_1(x)^2}\right\}m^{''}(x_0) \tag{3.129}$$

$$E\left[\hat{m}_1(x)\right]=\left\{\left[\frac{S_2(x)S_0(x)-S_1^2(x)}{S_2(x)S_0(x)-S_1^2(x)}\right]m(x_0)+\left[(x_0-x_i)\left(\frac{S_0(x)S_1(x)-S_1(x)S_0(x)}{S_2(x)S_0(x)-S_1^2(x)}\right)m(x_0)\right]\right\}$$

$$+\left\{\left[\frac{S_1(x)S_2(x)-S_1(x)S_2(x)}{S_2(x)S_0(x)-S_1^2(x)}\right]m^{'}(x_0)+\left[(x_0-x_i)\left(\frac{S_0(x)S_2(x)-S_1^2(x)}{S_2(x)S_0(x)-S_1^2(x)}\right)m^{'}(x_0)\right]\right\}$$

$$+\left\{\left[\frac{[S_2^2(x)-S_1(x)S_3(x)]}{S_2(x)S_0(x)-S_1^2(x)}\right]\frac{m^{''}(x_0)}{2}+\left[(x_0-x_i)\left(\frac{S_0(x)S_3(x)-S_1(x)S_2(x)}{S_2(x)S_0(x)-S_1^2(x)}\right)\frac{m^{''}(x_0)}{2}\right]\right\} \tag{3.130}$$

Clearly, we have

$$E\left[\hat{m}_1(x)\right]=\Big\{m(x_0)+(x_0-x_i)m^{'}(x_0)\Big\}$$

$$+\left\{\frac{[S_2^2(x)-S_1(x)S_3(x)]+[(x_0-x_i)\,(S_0(x)S_3(x)-S_1(x)S_2(x))]}{2\,[S_2(x)S_0(x)-S_1(x)^2]}\right\}m^{''}(x_0) \tag{3.131}$$

Now, the

$$Bias\,(m(x)) = (x_0 - x_i)m^{'}(x_0)$$

$$+\left\{\frac{[S_2^2(x) - S_1(x)S_3(x)] + [(x_0 - x_i)\,(S_0(x)S_3(x) - S_1(x)S_2(x))]}{2\,[S_2(x)S_0(x) - S_1(x)^2]}\right\} m^{''}(x_0) \quad (3.132)$$

Making the assumption that x_i 's are fixed design points on the interval $(0, 1)$ as pointed out in the work of Mageto (2008) and Eubank and Speckman (1993). According to Masry (1996), also the expression $S_j(x) = \sum_{i=1}^{n}(x_i - x_0)^j w_i(x) \equiv nh^{j+1}k_j + o(nh^{j+3})$ almost everywhere for $x\epsilon(0, 1)$.

For $S_j(x) = nh^{j+1}k_j + o(nh^{j+3})$

We have

$S_0(x) = nhk_0 + o(nh^3) = nh + o(nh^3)$; $S_1(x) = nh^2k_1 + o(nh^4) \equiv o(nh^4)$;

$S_2(x) = nh^3k_2 + o(nh^5)$; $S_3(x) = nh^4k_3 + o(nh^6)$; $S_4(x) = nh^5k_4 + o(nh^7)$.

Now,

1. $S_2^2(x) - S_1(x)S_3(x) = [nh^3k_2 + o(nh^5)]^2 - [o(nh^4)]\,[nh^4k_3 + o(nh^6)]$

$= n^2h^6k_2^2 + 2nh^3k_2o(nh^5) + o(n^2h^{10}) - o(nh^4)nh^4k_3 - o(n^2h^{10})$

$= n^2h^6k_2^2 + 2k_2o(n^2h^8) + o(n^2h^{10}) - o(n^2h^8)k_3 - o(n^2h^{10})$

$= n^2h^6k_2^2 + o(n^2h^8)$

2. $S_0(x)S_3(x) - S_1(x)S_2(x) = [nh + o(nh^3)]\,[nh^4k_3 + o(nh^6)] - [o(nh^4)]\,[nh^3k_2 + o(nh^5)]$

$= n^2h^5k_3 + nho(nh^6) + o(nh^3)nh^4k_3 + o(n^2h^9) - o(nh^4)nh^3k_2 - o(n^2h^9)$

$= n^2h^5k_3 + o(n^2h^7) + o(n^2h^7)k_3 + o(n^2h^9) - o(n^2h^7)k_2 - o(n^2h^9)$ since $k_0 = 1$

$= n^2h^5k_3 + o(n^2h^7)$

3. $S_2(x)S_0(x) - S_1(x)^2 = [nh^3k_2 + o(nh^5)]\,[nh + o(nh^3)] - [o(nh^4)]^2$

$= n^2h^4k_2 + nh^3k_2o(nh^3) + o(nh^5)nh + o(n^2h^8) - o(n^2h^8)$

$= n^2h^4k_2 + o(n^2h^6)k_2 + o(n^2h^6) + o(n^2h^8) - o(n^2h^8)$

$= n^2h^4k_2 + o(n^2h^6)$

4. $S_0(x)S_1(x) - S_1(x)S_0(x) = [nh + o(nh^3)]\,o(nh^4) - o(nh^4)\,[nh + o(nh^3)]$

$= o(n^2h^5) + o(n^2h^7) - o(n^2h^5) - o(n^2h^7) = 0$

Where necessary, we substitute the results for (1.), (2.), (3.) and (4.) in the following work.

For equation (4.2), we have

$$E\left[\hat{m}_1(x)\right] = \left\{m(x_0) + (x_0 - x_i)m^{'}(x_0)\right\}$$

$$+ \left\{\frac{[n^2h^6k_2^2 + o(n^2h^8)] + [(x_0 - x_i)\,(n^2h^5k_3 + o(n^2h^7))]}{2\,[n^2h^4k_2 + o(n^2h^6)]}\right\} m^{''}(x_0) \qquad (3.133)$$

Now, Bias of the local linear estimator is expressed as

$$Bias\left(\hat{m}_1(x)\right) = \left\{(x_0 - x_i)m^{'}(x_0)\right\} +$$

$$+ \left\{\left[\frac{n^2h^6k_2^2 + o(n^2h^8)}{[n^2h^4k_2 + o(n^2h^6)]}\right] + \left[(x_0 - x_i)\left(\frac{n^2h^5k_3 + o(n^2h^7)}{[n^2h^4k_2 + o(n^2h^6)]}\right)\right]\right\} \frac{m^{''}(x_0)}{2} \qquad (3.134)$$

Factorizing out h^2 and h for the first component and the second component of $m^{''}(x_0)$ respectively, we have;

$$Bias\left(\hat{m}_1(x)\right) = \left\{(x_0 - x_i)m^{'}(x_0)\right\}$$

$$+ \left\{h^2\left[\frac{n^2h^4k_2^2 + o(n^2h^6)}{[n^2h^4k_2 + o(n^2h^6)]}\right] + \left[(x_0 - x_i)h\left(\frac{n^2h^4k_3 + o(n^2h^6)}{[n^2h^4k_2 + o(n^2h^6)]}\right)\right]\right\}\frac{m^{''}(x_0)}{2} \tag{3.135}$$

$$Bias\left(\hat{m}_1(x)\right) = (x_0 - x_i)m^{'}(x_0) + \left\{h^2k_2^2 + [(x_0 - x_i)hk_3]\right\}\frac{m^{''}(x_0)}{2k_2} \tag{3.136}$$

$$Bias\left(\hat{m}_1(x)\right) = (x_0 - x_i)m^{'}(x_0) + \frac{h}{2k_2}\left(hk_2^2 + (x_0 - x_i)k_3\right)m^{''}(x_0)$$

$$Bias\left(\hat{m}_1(x)\right) = uhm^{'}(x_0) + \frac{h}{2k_2}\left(hk_2^2 + (uh)k_3\right)m^{''}(x_0) \tag{3.137}$$

Under the assumptions that $h^2 \to 0$ as $n \to \infty$, $m(\cdot)$ is bounded as stated in condition (ii) and $k_2 < \infty$, we have the local linear estimator is an asymptotically unbiased estimator.

Lemma 4

The asymptotic expression of the variance of $\hat{m}_1(x)$ is given by

$$Var\left(\hat{m}_1(x)\right) \approx \frac{d_k}{nh}\sigma^2(x_0) \tag{3.138}$$

as $h \to 0$ and $nh \to \infty$; where $d_k = \int K^2(u)du$.

Proof for Lemma 4

Considering the estimator defined by;

$$\hat{m}_1(x) = \sum_{i=1}^{n} \left\{ \left(\frac{S_2(x) - (x_i - x_0)\, S_1(x)}{S_2(x)S_0(x) - S_1(x)^2} \right) w_i(x) y_i \right\}$$

$$+ (x_0 - x_i) \sum_{i=1}^{n} \left\{ \left(\frac{(x_i - x_0)S_0(x) - S_1(x)}{S_2(x)S_0(x) - (S_1(x))^2} \right) w_i(x) y_i \right\} \tag{3.139}$$

$$\hat{m}_1(x) = \sum_{i=1}^{n} w_i^*(x) y_i + (x_0 - x_i) \sum_{i=1}^{n} w_i^{**}(x) y_i \tag{3.140}$$

where $w_i^*(x) = \left(\frac{S_2(x) - (x_i - x_0)S_1(x)}{S_2(x)S_0(x) - S_1(x)^2} \right) w_i(x)$, $w_i^{**}(x) = \left(\frac{(x_i - x_0)S_0(x) - S_1(x)}{S_2(x)S_0(x) - (S_1(x))^2} \right) w_i(x)$ and $w_i(x) = K(\frac{x_i - x_0}{h})$

$$Var(\hat{m}_1(x) = Var \left[\sum_{i=1}^{n} w_i^*(x) y_i + (x_0 - x_i) \sum_{i=1}^{n} w_i^{**}(x) y_i \right] \tag{3.141}$$

$$Var(\hat{m}_1(x) = \left[\sum_{i=1}^{n} w_i^{*2}(x) Var\,(y_i) + (x_0 - x_i)^2 \sum_{i=1}^{n} w_i^{**2}(x) Var\,(y_i) \right] +$$

$$+ \sum_{i=1}^{n} \sum_{j=1}^{n} 2 w_i^*(x) w_i^*(x) Cov(y_{i,} y_j) + (x_0 - x_i) \sum \sum_{j=1}^{n} 2 w_i^{**}(x) w_i^{**}(x) Cov(y_{i,} y_j) \tag{3.142}$$

$$Var(\hat{m}_1(x) = \sum_{i=1}^{n} w_i^{*2}(x) Var\,(y_i) + (x_0 - x_i)^2 \sum_{i=1}^{n} w_i^{**2}(x) Var\,(y_i) \tag{3.143}$$

$$Var(\hat{m}_1(x) = \sum_{i=1}^{n} w_i^{*2}(x) \sigma_i^2(x) + (x_0 - x_i)^2 \sum_{i=1}^{n} w_i^{**2}(x) \sigma_i^2(x) \tag{3.144}$$

Using $S_0(x) = nhk_0 + o(nh^3)$; $s_1(x) = nh^2 k_1 + o(nh^4) \equiv o(nh^4)$; $S_2(x) = nh^3 k_2 + o(nh^5)$; $S_3(x) = nh^4 k_3 + o(nh^6)$and $S_4(x) = nh^5 k_4 + o(nh^7)$.

$$w_i^{*2}(x) = \left(\frac{S_2(x)w_i(x) - (x_i - x_0)\, S_1(x)w_i(x)}{S_2(x)S_0(x) - S_1(x)^2} \right)^2$$

$$= \left(w_i(x) \frac{nh^3k_2 + o(nh^5) + x_0 o(nh^4) - x_i o(nh^4)}{n^2h^4k_2 + o(n^2h^6)} \right)^2$$

$$= \left(w_i(x) \frac{nh^3k_2 + o(nh^5)}{n^2h^4k_2 + o(n^2h^6)} \right)^2$$

Dividing up and down by nh , we get;

$$= \left(\frac{1}{nh} w_i(x) \frac{nh\,[nh^3k_2 + o(nh^5)]}{n^2h^4k_2 + o(n^2h^6)} \right)^2$$

$$\approx \left\{ \frac{1}{nh} w_i(x) \frac{(n^2h^4k_2 + o(n^2h^6))}{(n^2h^4k_2 + o(n^2h^6))} \right\}^2$$

$$= \frac{1}{n^2h^2} w_i^2(x) \tag{3.145}$$

$$w_i^{**2}(x) = \left(\frac{(x_i - x_0)S_0(x)w_i(x) - S_1(x)w_i(x)}{S_2(x)S_0(x) - (S_1(x))^2} \right)^2 \tag{3.146}$$

Introducing $S_0(x)S_1(x) - S_0(x)S_1(x)$ up and down in equation (3.146), we have;

$$w_i^{**2}(x) = \left(\frac{[S_0(x)S_1(x) - S_1(x)S_0(x)]\,[(x_i - x_0)S_0(x)w_i(x) - S_1(x)w_i(x)]}{S_2(x)S_0(x) - (S_1(x))^2\,[S_0(x)s_1(x) - S_1(x)S_0(x)]} \right)^2$$

$$w_i^{**2}(x) = \left(\frac{[S_0(x)S_1(x) - s_1(x)s_0(x)]\,[(x_i - x_0)s_0(x) - S_1(x)]}{S_2(x)Ss(x) - (S_1(x))^2\,[S_0(x)S_1(x) - S_1(x)S0(x)]} w_i(x) \right)^2$$

$$= \left(\frac{nh}{nh} \frac{[o(n^2h^5) + o(n^2h^7) - 0(n^2h^5) - o(n^2h^7)]\,[(x_i - x_0)\,(nh + o(nh^3)) - o(nh^4)]}{[n^2h^4k_2 + o(n^2h^6)]\,[o(n^2h^5) + o(n^2h^7) - o(n^2h^5) - o(n^2h^7)]} w_i(x) \right)^2$$

$$\approx \left\{ \frac{1}{nh} w_i(x) \frac{(o(n^2h^5) + o(n^2h^7) - o(n^2h^5) - o(n^2h^7))}{(n^2h^4k_2 + o(n^2h^6))} \right\}^2 \to 0$$

Therefore,

$$Var(\hat{m}_1(x) = \left[\sum_{i=1}^{n} w_i^{*2}(x)\sigma_i^2(x) + (x_0 - x_i)^2 \sum_{i=1}^{n} (0).\sigma_i^2(x) \right] \tag{3.147}$$

$$Var(\hat{m}_1(x) = \sum_{i=1}^{n} w_i^{*2}(x)\sigma_i^2(x) \tag{3.148}$$

Using the result of equation (3.145), we have;

$$Var(\hat{m}_1(x) = \sum_{i=1}^{n} \frac{1}{n^2h^2} w_i^2(x)\sigma_i^2(x) \tag{3.149}$$

$$Var\left(\hat{m}_1(x)\right) = \frac{1}{n^2h^2} \sum_{i=1}^{n} w_i^2(x) \left\{ \sigma^2(x_0) + \sigma^{2'}(x_0)uh + \frac{1}{2}\sigma^{2''}(x_0)u^2h^2 + \cdots \right\} \tag{3.150}$$

Simplifying equation (3.150), we get;

$$Var\left(\hat{m}_1(x)\right) = \frac{1}{n^2h^2} \int K^2(u)du\sigma^2(x_0) + \frac{h}{n^2h^2} \int uK^2(u)du\sigma^{2'}(x_0)$$

$$+ \frac{1}{2}\sigma^{2''}(x_0) \frac{h^2}{n^2h^2} \int K^2(u)u^2du + o(h^2) \tag{3.151}$$

Assuming that the second derivative of $\sigma^2(x_0)$ is bounded and continuous, $h \to 0$ and $nh \to \infty$ as $n, N \to \infty$,

$$Var_{asy}\left(\hat{m}_1(x)\right) \approx \sum_{i=1}^{n} \frac{1}{nh} d_k \sigma^2(x_0) \tag{3.152}$$

where $d_k = \int K^2(u)du$.

Mean square error (MSE) of $(\hat{m}_1(x))$

Finally, we have

$$MSE\left(\hat{m}_1(x)\right) = \left\{Bias\left(\hat{m}_1(x)\right)\right\}^2 + Var\left(\hat{m}_1(x)\right) \tag{3.153}$$

$$MSE\left(\hat{m}_1(x)\right) = \left\{(x_0 - x_i)m^{'}(x_0) + \frac{h}{2k_2}\left(hk_2^2 + (x_0 - x_i)k_3\right)m^{''}(x_0)\right\}^2 + \frac{1}{nh}\sigma^2(x_0)d_k \tag{3.154}$$

which is the asymptotic expression of MSE. $MSE\left(\hat{m}_1(x)\right) \to 0$ under the assumption $h \to 0$ and $nh \to \infty$ as $n \to \infty$, $m(\cdot)$ is bounded as stated in condition (ii) and $k_2 < \infty$,

Consequently, $\hat{m}_1(x)$ is asymptotically unbiased and consistent.

Proof of theorem using the local linear estimator

From the general formula for the finite population total given by;

$$Y_t = \sum_{i \epsilon B_n} y_i + \sum_{i \epsilon B_{N-n}} y_i \tag{3.155}$$

The population total can be rewritten as

$$Y_t = \left(\sum_{i \epsilon B_r} y_i + \sum_{i \epsilon B_{(n-r)}} y_i \right) + \sum_{i \epsilon B_{N-n}} y_i \tag{3.156}$$

For writing simplicity, denote $i \epsilon B_r$, $i \epsilon B_{(n-r)}$ and $i \epsilon B_{(N-n)}$ for the respondents and nonrespondents by $i \epsilon r$, $i \epsilon (n-r)$ and $i \epsilon (N-n)$ respectively throughout the following work.

From equation (3.60), the estimator for finite population total is given by

$$y_I = \sum_{i \epsilon r} y_i + \sum_{i \epsilon (n-r)} y_i^* \tag{3.157}$$

Let $\hat{Y}_t = y_I$ such that

$$\hat{Y}_t = \sum_{i \epsilon r} y_i + \sum_{i \epsilon (n-r)} y_i^* \tag{3.158}$$

Using the local linear estimator as the predictor for the missing data, we replace y_i^* by the corresponding estimator, $\hat{m_1}(x_i)$, thus

$$\hat{Y}_t = \sum_{i \epsilon r} y_i + \sum_{i \epsilon (n-r)} \hat{m_1}(x_i) \tag{3.159}$$

Taking expectation on both sides of equation (3.159), we get

$$E\left(\hat{Y}_t\right) = \sum_{i \epsilon r} E\left(y_i\right) + \sum_{i \epsilon (n-r)} E\left(\hat{m_1}(x_i)\right) \tag{3.160}$$

Substituting for the $E\left(\hat{m_1}(x_i)\right)$ in equation (3.160) using the results of Lemma 3, we have

$$E\left(\hat{Y}_t\right) = \sum_{i\epsilon r} m(x_i) + \sum_{i\epsilon(n-r)} m(x_0) + (x_0 - x_i)m^{'}(x_0)$$

$$+ \sum_{i\epsilon(n-r)} \left\{ \frac{[S_2^2(x) - S_1(x)S_3(x)] + [(x_0 - x_i)\left(S_0(x)S_3(x) - S_1(x)S_2(x)\right)]}{2\left[S_2(x)S_0(x) - S_1(x)^2\right]} \right\} m^{''}(x_0) \tag{3.161}$$

$$E\left(\hat{Y}_t\right) = \sum_{i\epsilon r} m(x_i) + \sum_{i\epsilon(n-r)} m(x_0) + \sum_{i\epsilon(n-r)} (x_0 - x_i)m^{'}(x_0) + \sum_{i\epsilon(n-r)} \left(\frac{S_2^2(x) - S_1(x)S_3(x)}{S_2(x)s_0(x) - S_1(x)^2} \right) \frac{m^{''}(x_0)}{2}$$

$$+ (x_0 - x_i) \sum_{i\epsilon(n-r)} \left(\frac{S_0(x)S_3(x) - S_1(x)S_2(x)}{S_2(x)s_0(x) - S_1(x)^2} \right) \frac{m^{''}(x_0)}{2} \tag{3.162}$$

$$E\left(\hat{Y}_t\right) = \left\{ \sum_{i\epsilon r} m(x_i) + \sum_{i\epsilon(n-r)} m(x_0) + \sum_{i\epsilon(n-r)} (x_0 - x_i)m^{'}(x_0) \right\}$$

$$+ \left\{ \sum_{i\epsilon(n-r)} \left\{ \frac{n^2h^6k_2^2 + o(n^2h^8)}{n^2h^4k_2 + o(n^2h^6} + (x_0 - x_i) \left(\frac{n^2h^5k_3 + o(n^2h^7)}{n^2h^4k_2 + o(n^2h^6} \right) \right\} \frac{m^{''}(x_0)}{2} \right\} \tag{3.163}$$

$$E\left(\hat{Y}_t\right) = \left\{ \sum_{i\epsilon r} m(x_i) + \sum_{i\epsilon(n-r)} m(x_0) + \sum_{i\epsilon(n-r)} (x_0 - x_i)m^{'}(x_0) \right\}$$

$$+\left\{\sum_{i\epsilon(n-r)}\left[\left\{h^2\frac{(n^2h^4k_2^2+o(n^2h^6))}{n^2h^4k_2+o(n^2h^6}+(x_0-x_i)h\left(\frac{n^2h^4k_3+o(n^2h^6)}{n^2h^4k_2+o(n^2h^6}\right)\right\}\frac{m^{''}(x_0)}{2}\right]\right\} \tag{3.164}$$

$$E\left(\hat{Y}_t\right)=\left\{\sum_{i\epsilon r}m(x_i)+\sum_{i\epsilon(n-r)}m(x_0)+\sum_{i\epsilon(n-r)}(x_0-x_i)m^{'}(x_0)\right\}$$

$$+\left\{\sum_{i\epsilon(n-r)}\left(\frac{h^2k_2^2+(x_0-x_i)hk_3}{k_2}\right)\frac{m^{''}(x_0)}{2}\right\} \tag{3.165}$$

$$E\left(\hat{Y}_t\right)=\left\{\sum_{i\epsilon r}m(x_i)+\sum_{i\epsilon(n-r)}m(x_0)+\sum_{i\epsilon(n-r)}(x_0-x_i)m^{'}(x_0)\right\}$$

$$+\left\{\sum_{i\epsilon(n-r)}\left[h\left(\frac{hk_2^2+(x_0-x_i)k_3}{k_2}\right)\right]\frac{m^{''}(x_0)}{2}\right\} \tag{3.166}$$

Bias of the finite population total estimator $\hat{Y}_t$ using $\hat{m_1}(x_i)$ results

Consider,

$$E\left(\hat{Y}_t\right)-(Y_t)=\left\{\sum_{i\epsilon r}m(x_i)+\sum_{i\epsilon(n-r)}m(x_0)+\sum_{i\epsilon(n-r)}(x_0-x_i)m^{'}(x_0)\right\}$$

$$+\left\{\sum_{i\epsilon(n-r)}\left[h\left(\frac{hk_2^2+(x_0-x_i)k_3}{K_2}\right)\right]\frac{m^{''}(x_0)}{2}\right\}$$

$$\left\{ -\sum_{i\epsilon r} y_i - \sum_{i\epsilon(n-r)} m(x_0) - \sum_{i\epsilon(n-r)} (x_i - x_0)m'(x_0) - \sum_{i\epsilon(n-r)} \frac{m''(x_0)(x_i - x_0)^2}{2} \right\} \tag{3.167}$$

$\sum_{i\epsilon(n-r)} (x_i - x_0)m'(x_0) \to 0$ in equation (3.167) under the kernel density conditions.

Subtracting equation (3.159) from equation (3.163) gives

$$Bias\left(\hat{Y}_t\right) = \sum_{i\epsilon(n-r)} (x_0 - x_i)m^{'}(x_0) + \left\{ \sum_{i\epsilon(n-r)} \left[h\left(\frac{hk_2^2 + (x_0 - x_i)k_3}{k_2} \right) \right] \frac{m^{''}(x_0)}{2} \right\}$$

$$- \left\{ \sum_{i\epsilon(n-r)} \frac{m''(x_0)(x_i - x_0)^2}{2} \right\} \tag{3.168}$$

$$Bias\left(\hat{Y}_t\right) = \sum_{i\epsilon(n-r)} \left\{ (x_0 - x_i)m^{'}(x_0) + \left[h\left(\frac{hk_2^2 + (x_0 - x_i)k_3}{k_2} \right) - \left((x_i - x_0)^2\right) \right] \frac{m^{''}(x_0)}{2} \right\} \tag{3.169}$$

Variance of the estimator, $\hat{Y}_t$ using results of Lemma 4.

Variance of $\hat{Y}_t$ is given by the variance of the error term given by $\hat{Y}_t - Y_t$, that is

$$Var\left(\hat{Y}_t\right) = Var\left(\hat{Y}_t - Y_t\right) \tag{3.170}$$

$$= Var\left(\sum_{i\epsilon(n-r)}(\hat{m}_1(x_i) - y_i) - \sum_{i\epsilon N-n} y_i\right) \tag{3.171}$$

$$= Var\left(\sum_{i\epsilon(n-r)}\hat{m}_1(x_i) - \sum_{i\epsilon(n-r)} y_i - \sum_{i\epsilon N-n} y_i\right) \tag{3.172}$$

$$= \sum_{i\epsilon(n-r)} Var\left(\hat{m}_1(x_i)\right) - \sum_{i\epsilon(n-r)} Var\left(y_i\right) - \sum_{i\epsilon N-n} Var\left(y_i\right) \tag{3.173}$$

Substituting for $Var\left(\hat{m}_1(x_i)\right)$ from equation (3.148), we have

$$Var\left(\hat{Y}_t\right) = \sum_{i\epsilon(n-r)}\sum_{i=1}^{n} w_i^{*2}(x)\sigma_i^2(x) - \sum_{i\epsilon(n-r)}\sigma_i^2(x) - \sum_{i\epsilon N-n}\sigma_i^2(x) \tag{3.174}$$

$$= \sum_{i\epsilon(n-r)}\sum_{i=1}^{n}\frac{1}{n^2h^2}w_i^2(x)\sigma_i^2(x) - \sum_{i\epsilon(n-r)}\sigma_i^2(x) - \sum_{i\epsilon N-n}\sigma_i^2(x) \tag{3.175}$$

$$\approx \frac{1}{nh}d_k\sigma^2(x_0) - (n-r)\sigma_i^2(x) - (N-n)\sigma_i^2(x) \tag{3.176}$$

where $d_k = \int K^2(u)du$.

Assuming that $\sigma_i^2($x is at least twice continuously differentiable and $n, N \to \infty$ such that $(n-r) \to 0$, $(N-n) \to 0$, then,

$$Var_{asy}\left(\hat{Y}_t\right) \approx \frac{1}{nh} d_k \sigma^2(x_0)$$

The mean square error (MSE) of $\hat{Y}_t$

Finally, we have

$$MSE\left(\hat{Y}_t\right) = \left\{Bias\left(\hat{Y}_t\right)\right\}^2 + Var\left(\hat{Y}_t\right) \tag{3.177}$$

$$MSE\left[\hat{Y}_t\right] = \left\{\sum_{i\epsilon(n-r)} \left\{(x_0 - x_i)m^{'}(x_0) + \left[h\left(\frac{hk_2^2 + (x_0 - x_i)k_3}{K_2}\right) - \left((x_i - x_0)^2\right)\right]\frac{m^{''}(x_0)}{2}\right\}\right\}^2$$

$$+ \frac{1}{nh}\sigma^2(x_0)d_k \tag{3.178}$$

which is the asymptotic expression of the MSE for $\hat{Y}_t$.$MSE\left(\hat{Y}_t\right) \to 0$ as $h \to 0$ and $nh \to \infty$, and thus $\hat{Y}_t$ is consistent. Consequently, $\hat{\bar{y}}_t$ is asymptotically unbiased and consistent.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

Results of simulation are tabulated and interpretation of the results is done. To make analysis of the results easy, we do comparisons of the estimator using the values obtained for the performance measures. Also, we consider the analysis of results for each time point and for each data set with respect to the sample size and a general conclusion is obtained.

4.2 Description of longitudinal data

In this section, theory developed in the previous sections of chapter three is tested using simulated data. Here, a study of the finite population mean estimators based on four measures of performance (percentage relative bias (%RB), MSE and bootstrap standard deviation (SD bootstrap)) is carried out. Simulations and computations of the finite population mean estimators were done using R software (R version 3.2.3 (2015-12-10)) based on 1000 runs. For the the local linear and local constant estimators, the Gaussian kernel with a fixed bandwidth of $h = 0.75$ was used. To fit the nonparametric regression, the loess function in R was used. For comparison purposes, we used complete data as our main reference in the evaluation of the performance of the estimators (Proposed local linear estimator, local constant estimator and the simple linear regression estimator). In this simulation study, a sample of size $n = 1000$ was considered for each of the two data sets. The longitudinal data for each

of the sampled units is of size $T = 4$ that is, $t = 1, 2, 3, 4$. This will yield 2^3 different patterns of the longitudinal data with each of respondent and nonrespondent values being denoted by 1 and 0 respectively at different time points. The assumption that all sampled units are observed at the first time point is used.

Longitudinal data was generated according to two models:

(i) In model 1, simulation of $(y_i, i = 1, 2, 3, 4)$ is done from a multivariate normal distribution with the means for the 4 time points as $1.33, 1.94, 2.73, 3.67$ respectively and the covariance matrix following the AR(1) model with standard error 1 and correlation coefficient 0.9.

(ii) In model 2, simulation of $(log(y_i), i = 1, 2, 3, 4)$ is done from a multivariate normal distribution with the means for the 4 time points as $1.33, 1.94, 2.73, 3.67$ respectively and the covariance matrix following the AR(1) model with standard error 1 and correlation coefficient 0.9.

Using assumption (1) of the imputation process, we may generate the nonrespondents following the nature of the models (i) and (ii) using equations (4.1) and (4.2) respectively for nonresponse under missing at random mechanism.

$$P(I_t = 0|y_{t-1}) = \frac{exp\{2 - 1.3y_{t-1}\}}{1 + exp\{2 - 1.3y_{t-1}\}} \tag{4.1}$$

$$P(I_t = 0|y_{t-1}) = \frac{exp\{3 - 0.8y_{t-1}\}}{1 + exp\{3 - 0.8y_{t-1}\}} \tag{4.2}$$

In order to obtain the nonmonotone pattern in the simulated data, we used the predetermined unconditional probabilities of Shao et al. (2012).

Pattern type	Nonresponse pattern	Normal / Log-normal data	Total Probability
Monotone	1 0 0 0	0.062	0.181
	1 1 0 0	0.043	
	1 1 1 0	0.076	
Nonmonotone	1 0 0 1	0.113	0.494
	1 0 1 0	0.071	
	1 0 1 1	0.186	
	1 1 0 1	0.124	
Complete data	1 1 1 1	0.325	0.325

Table 4.1: Probabilities of nonresponse patterns for $t = 4$

Bootstrap variance estimation

The following steps were used to obtain the bootstrap variance.

1. We constructed a pseudo population by replicating the sample of size 1000 through 1000 simulation runs.

2. A simple random sample of size 200 is drawn with replacement from the pseudo population .

3. We applied the simple linear regression, local constant and local linear regression imputation models to impute the missing y_i's of the sample.

4. Repeating the steps 2 and 3 for a large number of times ($B = 1000$) to obtain $\hat{Y}_I^{(1)}, \cdots, \hat{Y}_I^{(B)}$ where $\hat{Y}_I^{(b)}$ is the analog of $\hat{Y}_I$, for the b-th bootstrap sample.

Step 5: Obtain the bootstrap variance of $\hat{Y}_I$, by

$V_{boot}(\hat{Y}_I) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{Y}_I^{(b)} - \hat{Y}_I^{(\cdot)}\right)^2$

Where $\hat{Y}_I^{(\cdot)}$ is the mean bootstrap analog of $\hat{Y}_I$, given by

$\hat{Y}_I^{(\cdot)}, = \frac{1}{B}\sum_{b=1}^{B}\hat{Y}_I^{(b)}$

4.3 A simulation study

The results of the simulation study are summarized in Table 4.2 and Table 4.3.

Method (Estimator used)	Measure (parameter)	$t=1$	$t=2$	$t=3$	$t=4$
Complete data analysis	Mean	1.328918	1.939003	2.729671	3.66934
	Standard deviation	1.000342	1.000168	0.9997156	1.000435
	%RB	0.0	0.0	0.0	0.0
	MSE	1.001018	1.000666	0.9997697	1.001196
	SD bootstrap	0.6667591	0.6666357	0.6666357	0.6675065
	Length of confidence interval	0.10490764	0.10477944	0.1040547	0.1037104
Local Linear Regression	Mean		1.938469	2.729698	3.669843
	Standard deviation		0.9948414	0.9926485	0.9932463
	%RB		0.0003101247	0.004607907	0.003463886
	MSE		0.9900532	0.9857052	0.9868784
	Length of confidence interval		0.10391248	0.10386226	0.10293502
	SD bootstrap		0.6606914	0.6600272	0.6597972
Nadaraya-Watson	Mean		1.938513	2.688752	3.658198
	Standard deviation		0.9804571	0.995685	0.9812671
	%RB		0.002618051	-1.49819	-0.3079823
	MSE		0.9616402	0.9934076	0.963356
	Length of confidence interval		0.10384772	0.1061238	0.10381222
	SD bootstrap		0.9807754	0.9967448	0.9815455
Simple linear regression	Mean		1.939073	2.729775	3.669467
	Standard deviation		0.9952188	0.9928367	0.9926948
	%RB		0.003486382	0.003859931	0.003474327
	MSE		0.9908072	0.9860761	0.9857896
	Length of confidence interval		0.10391578	0.10386978	0.10292522
	SD bootstrap		0.9952162	0.9938139	0.993223

Table 4.2: Simulated results for mean estimation (normal case)

In terms of standard deviation, at $t = 2$, the Nadaraya-Watson estimator had the least values of 0.9804571 followed by the local linear estimator with standard deviation values 0.9948414, and then the simple linear regression estimator has the largest standard deviation value of 0.9952188. Looking at these values closely, the local linear standard deviation value is very close to that of the simple linear regression estimator and to that of the complete data which is 1.000168 as compared to Nadaraya-Watson estimator standard deviation value. At $t = 3$, the local linear estimator had the least standard deviation value followed by the simple linear regression estimator and then the Nadaraya-Watson estimator . At $t = 4$, it is shown that the Nadaraya-Watson estimator had the least standard deviation value of 0.9812671 followed by the simple linear regression estimator standard deviation value given by 0.9926948 and then the local linear estimator had the largest standard deviation value of 0.9932463. In comparison with the complete data standard deviation value

of 1.0000435, we notice that the local linear estimator performed better than the other two estimators with its value very close to that of the complete data.

In terms of the percentage relative bias (%RB), at time point $t = 2$, it can be seen that the local linear estimator has the least value of 0.0003101247 followed by the Nadaraya-Watson estimator with %RB value of 0.002618051 and then the simple linear regression estimator with %RB value of 0.003486382 which was the largest value. This shows that the nonparametric estimators performed better than the parametric estimator basing on their values of %RB. At time points $t = 3$, observe that the the simple linear regression with %RB value of 0.003859931 has the least value followed by the local linear estimator with %RB value of 0.004607907 and the Nadaraya-Watson estimator performs worst with the largest %RB value given by -1.4981. The %RB values of the local linear estimator and the simple linear regression estimator are very much closer to zero than those for the other estimators. At time point $t = 4$, observe that the local linear estimator has the least value of %RB given by 0.003463886 followed by the simple linear regression estimator with %RB value of 0.003474327 and the Nadaraya-Watson estimator performed worst with large %RB 0f -0.3079823. Through comparisons based on %RB with reference to the complete data, the local linear estimator has its %RB values approaching zero.

In terms of MSE, at time points $t = 2$, the Nadaraya-Watson estimator has the least MSE value of 0.9619855 followed by the local linear estimator MSE values 0.9908072 and lastly the simple linear regression which has the largest values of MSE given by 0.9904082 . At time point $t = 3$, the the local linear estimator has the least values of MSE followed by the simple linear regression estimator and lastly the Nadaraya-Watson estimator which has the largest MSE value implying that the local linear estimator performed well. At time point $t = 4$, Nadaraya-Watson estimator has the

least MSE value of 0.9620454 followed by the simple linear regression estimator with MSE value of 0.9854251 and lastly the local linear estimator which has the largest MSE value of 0.9857896.

In terms of the bootstrap standard deviation, it can be seen in table (4.2) that the local linear estimator performs best at all the three time points $t = 2$, $t = 3$, and $t = 4$ since it had the least bootstrap standard deviation values of 0.6588623, 0.655473and 0.658257 respectively and its values were even smaller than those of the complete data given by 0.6659541, 0.6659541 and 0.658257 in order of increasing time. Nadaraya-Watson estimator bootstrap standard deviation values were 0.9793316, 0.9936533 and 0.9797415 and Simple linear regression estimator has the largest values of bootstrap standard deviation given by 0.9952162 , 0.9938139 and 0.993223 for $t = 1, t = 2$ and $t = 3$ respectively. This implies that the results got with local linear estimator are the best for reliable inferences. From 4.2 of results, it is also shown that the bootstrap variance values of the local linear estimator are more close to those of the Nadaraya-Watson estimator than the simple linear regression estimator. Also, the simple linear regression and Nadaraya-Watson estimators are competing interchangeably in terms of performance for the bootstrap samples.

Method (Estimator used)	Measure (parameter)	$t=1$	$t=2$	$t=3$	$t=4$
Complete data analysis	Mean	1.330963	1.94061	2.731046	3.671122
	Standard deviation	1.000228	0.9999145	0.9998701	1.000415
	%RB	0.0	0.0	0.0	0.0
	MSE	1.000779	1.000138	1.000068	1.001156
	Length of confidence interval	0.8453392	1.4676012	3.469394	8.674622
	SD bootstrap	0.6658951	0.6659541	0.6659541	0.6662738
Local Linear Regression	Mean		1.940391	2.731393	3.671548
	Standard deviation		0.9950302	0.9927199	0.9925087
	%RB		-0.006115805	0.001946422	0.003121577
	MSE		0.9904082	0.9858397	0.9854251
	Length of confidence interval		1.4669038	3.468194	8.694804
	SD bootstrap		0.6588623	0.655473	0.658257
Nadaraya-Watson	Mean		1.940298	2.689957	3.660124
	Standard deviation		0.9806438	0.9958007	0.9805938
	%RB		-0.0109425	-1.506794	-0.3052104
	MSE		0.9619855	0.9936533	0.9620454
	Length of confidence interval		1.4634962	3.459618	8.658362
	SD bootstrap		0.9793316	0.9938614	0.9797415
Simple linear regression	Mean		1.940518	2.731128	3.671224
	Standard deviation		0.9948923	0.9928891	0.9925527
	%RB		-0.004716414	0.002994436	0.002771179
	MSE		0.9901363	0.9861755	0.9855044
	Length of confidence interval		1.463783	3.467702	8.670952
	SD bootstrap		0.9940906	0.9909141	0.9916702

Table 4.3: Simulated results for mean estimation (log-normal case)

In terms of the standard deviation, at time point $t = 2$, it can be seen that the Nadaraya-Watson estimator has the least standard deviation value of 0.9806438 followed by the simple linear regression with a standard deviation value of 0.9948923 and the local linear regression performed worst with a standard deviation value of 0.9950302. But in comparison with the standard deviation of the complete data of 0.9999145, the local linear estimator has the best results. At time point $t = 3$, it is seen that the local linear has the lowest values of the standard deviation (0.9927199) followed by the values of simple linear regression and the Nadaraya-Watson estimator performed worst with a standard deviation value of 0.9958007. At time point $t = 4$, observe that the Nadaraya-Watson estimator has the least value of standard deviation (0.9805938) followed by the local linear estimator with a standard deviation value of 0.9925087 and the simple linear regression estimator which has the largest value (0.9925527).

In terms of the percentage relative bias (%RB), at time points t = 2 and t = 4, observe that the simple linear regression estimator has the least %RB values (-0.004716414 and 0.02771179) followed by the local linear estimator with %RB values -0.006115805 and 0.003121577 and the NadarayaWatson estimator has the biggest %RB values (-0.0109425 and 0.3052104) respectively. Based on these aforementioned results, it is viable to choose the best estimator as the local linear estimator which handles both linear and nonlinear models. At time points t = 3, observe that the local linear estimator has the least %RB value followed by simple linear regression estimator and lastly the Nadaraya-Watson. This implies that, for n = 1500 , the local linear estimator has the smallest bias close to zero as for the complete data, hence the best estimator compared to others.

In terms of the MSE, at time points t = 2 and t = 4, Nadaraya-Watson estimator has the least values of MSE given by 0.9619855 and 0.9620454 respectively, followed by the simple linear regression estimator and lastly the local linear estimator which has the largest values of MSE given by 0.9904082 and 0.9854251 respectively. At time point t = 3, the the local linear estimator has the least MSE value of 0.9858397 followed by simple linear regression estimator with MSE value 0.9861755 and Nadaraya-Watson estimator has the biggest value of MSE given by 0.9936533 implying that local linear estimator performed well at time point t = 3 .

In terms of the bootstrap variance, bootstrap standard deviation values are used. It is seen that from 4.3, the local linear estimator performs the best at all the three time points since it has the least bootstrap standard deviations of 0.6588623, 0.655473and 0.658257 for $t = 1$, $t = 2$ and $t = 3$ respectively and these values are even smaller than those of the complete data given by 0.6659541, 0.6659541 and 0.658257 in order of increasing time. It is can be seen that the bootstrap standard deviations of the

local linear estimator are more close to those of the Nadaraya-Watson estimator than the simple linear regression estimator which has large bootstrap standard deviation values.

4.3.1 Discussion of Results

Basically, two main results were obtained in both theory and practical aspects. The first result examines the performance of the imputation estimators (local constant and local linear estimators) in the estimation of the time dependent finite population means. In this, we were able to explore the asymptotic properties of bias and consistency both theoretically and using the simulated data. Basing on the theoretical derivations which was summarised by the Theorem and its proof, it was found that imputation for the nonrespondents using both local constant estimator (Nadaraya-Watson estimator) and the local linear estimator (local polynomial of degree one) produce asymptotically unbiased results except for the asymptotic expressions for the variance. As compared to the results of Anthony (1999), the simulation results have shown that estimation of the finite population mean using local linear regression as the best estimator for the nonrespondents in the longitudinal surveys without making assumption on the response mechanism which create missing values in the longitudinal data. Estimation of the time dependent finite population parameters for data with missing values was found to be effective using local linear regression estimator than using the Nadaraya-Watson estimator. These results are in line with the results of Cai (2001). The behaviour of the results in terms of the Mean Square Error are justified by the selection of the band width

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

Generally, nonrespondents in any survey data has a significant impact on the bias and the variance of the estimators and therefore, before using such data in statistical inference, imputation with an appropriate technique ought to be done. In this study, the main objective was to obtain an imputation method based on local polynomial regression for nonmonotone nonrespondents in longitudinal surveys and determine its asymptotic properties. In this way, the local linear regression estimator (proposed estimator) was used in the prediction of the missing values. The assumption that all the sampled units were observed at the baseline and remained in the survey was considered throughout the study. Comparing the parametric and nonparametric methods, nonparametric methods performed better than the parametric methods. Among the nonparametric methods, the local linear estimator was the best estimator as it behaved better than the Nadaraya-Watson estimator in terms of %RB. Also, the local linear estimator values of %RB for normal data were smaller than for the log-normal case. In terms of the bootstrap standard deviation, the local linear estimator performs best at all the three time points since it has the least bootstrap standard deviations values for the two data sets. Generally, the local linear estimator performs relatively well and in particular in the normal data. We conclude that use of the nonparametric estimators seem plausible in both theoretical and practical scenarios.

5.2 Recommendations

1. In this study, a fixed bandwidth and a strong correlation were used. The impact of the use of a data-driven or differing bandwidth and weak correlation needs to be investigated.

2. The error terms were assumed to have a constant variance (i.e. independent of the survey variable) in the simulation study. There is still need to carry out investigation of the performance of the estimators when the variance of the error terms is a function of the survey variable.

REFERENCES

Anthony, N. C. (1999). *Nonparametric prediction in survey sampling and its application to nonresponse problem.* PhD thesis, Montreal, Quebec.

Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified Sampling. *The Annals of Statistics*, 12(2):470–482.

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

Cai, Z. (2001). Weighted Nadaraya- Watson Regression Estimation. *Statistics & probability letters*, 51(3):307–318.

Cheng, M. Y., Fan, J., and Marron, J. S. (1997). On automatic boundary corrections. *The Annals of Statistics*, 25(4):1691–1708.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, 89(425):81–87.

Dorfman, A. H. (1992). Nonparametric regression for estimating totals in finite population. Proceeding section of survey methodology. *Journal of American Statistical Association*, (kk):622–625.

Erdasos, P. and Razenyi, A. (1959). On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hungar. Acad. Sci.*, 4:49–61.

Eubank, R. L. (1988). *Spline smoothing and nonparametric regression.* New York.

Eubank, R. L. and Speckman, L. (1993). Local Polynomial Fitting in Adopted to the Autoregressive Context for Modelling Nonlinear Time Series under Some Missing Conditions. *Journal of American Statistical Association*, 88(424):1287–1301.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *Ann. Statist*, 21(1):196–216.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, 20(4):2008–2036.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman and Hall, London.

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *J R Stat Soc Ser B*, 17:269–278.

Hastie, T. J. and Loader, C. (1993). Local regression: automatic kernel carpentry (with discussion). *Statist. Sci*, 8:120–143.

Hazajek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite populations. *Ann. Math. Statist*, 35:1491–1523.

Jong, R. D., Buuren, V. S., and Spiess, M. (2014). Multiple Imputation of Predictor Variables Using Generalized Additive Models. *Communications in Statistics*, 45:1–18.

Kadilar, C. and Cingi, H. (2008). Estimators for the Population Mean in the Case of Missing Data. *Communications in Statistics*, 37(14):2226–2236.

Krewski, D. and Rao, J. N. K. (1981). Inference From Stratified Samples: Properties

of the Linearization, Jackknife and Balanced Repeated Replication Methods. *The Annals of Statistics*, 9(5):1010–1019.

Kyuseong, K. (2000). Variance Estimation Under Regression Imputation Model. *Department of Computer Science and Statistics*, 592-597.

Laird, N. M. (1988). Missing data in longitudinal studies. *Stat. Med*, 7:305–315.

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *J Am Stat Assoc*, 99(446):546–556.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *J. ASA*, 88:125–134.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Chichester, New York.

Mageto, O. T. (2008). *Robust Estimation of the Finite Population Total Using Local Polynomial Regression*. PhD thesis, Jomo Kenyatta University of Agriculture and Technology.

Masry, E. (1996). Multivariate Local Polynomial regression for Time series. Uniform Strong Consistence and Rates. *Journal of Time Series Analysis*, 17:571–599.

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

Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of American Statistical Association*, 92:1320–1329.

Pier, L. C., Daniela, M., and Mauro, S. (2008). Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators. *Computational Statistics and Data Analysis*, 53(2):354–365.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression method for repeated outcomes in the presence of missing data. *Journal of American Statistical Association*, 90:106–121.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc, New York.

Rueda, M. and Sánchez-Borrego, I. R. (2009). A predictive estimator of finite population mean using nonparametric regression. *computational statistics*, 24(1):1–14.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22(3):1346–1370.

Sarndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Scott, A. J. and Wu, C. F. J. (1981). On the asymptotic distribution of ratio and regression estimators. *J. Amer. Statist. Assoc. 76, 98-102*, 76:98–102.

Shahab, J., Laurence, F. E., and Buuren, S. V. (2014). Dual imputation model for incomplete longitudinal data. *Journal of Mathematical and Statistical Psychology*, 67(2):197–212.

Shao, J., Klein, M., and Xu, J. (2012). Imputation for nonmonotone nonresponse in

the survey of industrial research and development. *Survey methodology*, 38(2):143–155.

Singh, R., Verma, H. K., Sharma, P., Rajesh, S., Hemant, K. V., and Prayas, S. (2016). Estimation of Population Mean Using Exponential Type Imputation Technique for Missing Observations. *Journal of Modern Applied Statistical Methods*, 15(1):358–372.

Singh, S. and Horn, S. (2000). Compromised imputation in survey sampling. *Metrika*, 51(3):266–276.

Stone, J. C. (1977). Consistent Nonparametric Regression. *The annals of statistics*, 5(4):595–620.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall, London.

Wang, M. and Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *Annals of Statistics*, 30(3):896–924.

Watson, G. S. (1964). Smooth regression analysis. *Sankhy: The Indian Journal of Statistics*, 26(4):359–372.

Xu, J., Shao, J., Palta, M., and Wang, L. (2008). Imputation for nonmonotone last-value-dependent nonrespondents in longitudinal surveys. *Survey methodology*, 34(2):153–162.

Yu, R. R. and Li, L. A. (2011). Imputation of non-ignorable nonresponses for income: analysis of a panel study on Taiwan. *Qual Quantuant*, 45(4):875–884.