Scientific Research Publishing

Open Journal of Statistics, 2016, 6, 1138-1154
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Imputation Based on Local Linear Regression for Nonmonotone Nonrespondents in Longitudinal Surveys

Sarah Pyeye[1], Charles K. Syengo[1], Leo Odongo[2], George O. Orwa[3], Romanus O. Odhiambo[3]

[1]Pan African University Institute for Basic Sciences, Technology and Innovation, Nairobi, Kenya
[2]Department of Statistics and Actuarial Science, Kenyatta University, Nairobi, Kenya
[3]Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
Email: srhpyeye@gmail.com, kilundac@gmail.com, odongo.leo@ku.ac.ke, gorwa@jkuat.ac.ke, romanusemod@yahoo.com

How to cite this paper: Pyeye, S., Syengo, C.K., Odongo, L., Orwa, G.O. and Odhiambo, R.O. (2016) Imputation Based on Local Linear Regression for Nonmonotone Nonrespondents in Longitudinal Surveys. *Open Journal of Statistics*, **6**, 1138-1154.
http://dx.doi.org/10.4236/ojs.2016.66092

Received: October 13, 2016
Accepted: December 20, 2016
Published: December 27, 2016

Abstract

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 and then the estimation of the time-dependent finite populations means. The asymptotic properties (unbiasedness and consistency) of the proposed estimator are investigated. Comparisons between different parametric and nonparametric estimators are performed based on the bootstrap standard deviation, mean square error and percentage relative bias. A simulation study is carried out to determine the best performing estimator of the time-dependent finite population means. The simulation results show that local linear regression estimator yields good properties.

Keywords

Longitudinal Survey, Nonmonotone, Nonresponse, Imputation, Nonparametric Regression

1. Introduction

Longitudinal surveys 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. While longitudinal surveys 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 after having participated in the surveys thus the occurrence of

nonresponses.

Missing data are a problem because nearly all standard statistical methods presume complete information for all the variables included in the analysis. Using data with missing values leads to reduction in sample size which significantly affects the precision of the confidence interval, statistical power reduce and biased population parameter estimates. 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 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 like sensitivity to model misspecification while nonparametric methods are more robust and flexible [1]. Some of the methods used by [2] 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 [3] and [4] 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 [5] and [6], a rival for Nadaraya-Watson technique is the local linear regression estimator which was found to produce unbiased estimates without boundary effects. [7] 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, [8] also found that the local linear regression estimator (introduced by [9]) 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 nonresponnndents 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 determine the best performing estimator of the finite population mean.

2. Assumptions and Notations

1) All sampled units are observed on the first time point $(t=1)$ and remain in the sample till the final time $t=T$. The variable of interest $y_{i,t}$ is the value of y for the i^{th} unit at time point t.

2) The prediction process is past last value dependent and the vectors

$\left(y_{i,1},\cdots,y_{i,T},I_{i,1},\cdots,I_{i,T}\right)$ are independently and identically distributed (i.i.d) from the superpopulation under the model-assisted approach.

For $t=2,\cdots,T$ and $i=1,2,\cdots,N$ and the response indicator function $I_{i,t}$ is

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

3) The vector $\left(y_1,\cdots,y_T\right)$ follows the Markov chain for longitudinal survey data without missing values

$$L\left(y_{i,t}\mid y_{i,t-1},I_{i,t}=0,I_{i,t-1}=1\right)=L\left(y_{i,t}\mid y_{i,t-1},I_{i,t-1}=1\right) \tag{2}$$

4) We assume that the population P is divided into a fixed number of imputation classes, which are basically unions of some small strata.

3. Regularity Conditions

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

$$p\left(x\right)=P\left(I_{i,t}=1\mid Y,X\right)=P\left(I_{i,t}=1\mid X\right) \tag{3}$$

and g and f have bounded second derivatives

i) The Kernel function K is a bounded and twice continuously differentiable symmetric function on the interval $\left[-1,1\right]$, and such that $k_0=\int K\left(u\right)\mathrm{d}u=1$, $k_1=\int uK\left(u\right)\mathrm{d}u=0$, $k_2=\int u^2K\left(u\right)$, $k_2<\infty$ and $\int_{-\infty}^{\infty}\left\{K\left(u\right)\right\}^2\mathrm{d}u<\infty$.

ii) The regression function $m\left(\cdot\right)$ is at least twice continuously differentiable everywhere in the neighborhood of x_0.

iii) The sample survey variable of interest has a finite second moment bounded on the interval $\left(0,1\right)$. Thus $E\left(y^2\right)<\infty$.

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

4. Methodology

4.1. Imputation Process

Considering the case of the last past value, we do impute for missing value $y_{i,t}^{*}$ by the value obtained through the prediction procedure. But according to [10], the joint distribution of bivariate random variables (X,Y) is preserved when the missing value, Y is imputed by the conditional distribution of Y given X. Therefore, considering the conditional mean imputation approach for the single imputation.

Let

$$\varphi_{i,t,t-1}\left(y_{t-1}\right)=E\left(y_{i,t}\mid y_{i,t-1},I_{i,1}=0,I_{i,t-1}=1\right) \tag{4}$$

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 therefore clear that when $\varphi_{i,t,t-1}$ is known, then the imputed value of unobserved $y_{i,t}$ is given by $y_{i,t}=\varphi_{i,t,t-1}\left(y_{t-1}\right)$. In cases where $\varphi_{i,t,t-1}\left(y_{t-1}\right)$ in Equation (4) is unknown, for nonmonotone nonrespondents, we employ the last value dependent

mechanism.

Under assumption (2), we have

$$\varphi_{i,t,t-1}\left(y_{t-1}\right)=E\left(y_{i,t}\mid y_{i,t-1},I_{i,t}=1,I_{i,t-1}=1\right) \tag{5}$$

Using Equation (4), we are limited to do estimation by regressing the nonrespondents y_t on the observed values y_{t-1} based on the longitudinal survey data, therefore, we apply the equivalent Equation (5) which allows estimation using data from all subjects having observed y_t and observed y_{t-1}. Then, the imputation of the nonrespondents is done using $\varphi_{i,t,t-1}\left(y_{t-1}\right)$ in Equation (5) and under the last value dependent assumption, we are able to use auxiliary survey data in regression fitting. According to [11], imputing nonresponses using (5) was done for monotone case and their approach is easy to apply if the conditional expectation say, $\varphi_{t,t-1}\left(x\right)$ in (4) has a linear relationship with *x*. Adopting the concept of nonparametric method in [12], here, the local linear estimator of $\varphi_{t,t-1}\left(x\right)$ is $\hat{\varphi}_{t,t-1}\left(x\right)$. Let $y_{i,t}$ be the variable of interest for the *i*-th unit at time *t* where $i=1,\cdots,N$ and $t=1,\cdots,T$. Associated with each $y_{i,t}$ are the known $x_{i,t,q}$, $q=1,\cdots,Q$, of *q* auxiliary variables. To make the notations and writings simple, we relax the index *t* and write with a single subscript *i*, thus $y_{i,t}$ is written as y_i.

The regression imputation model η is given by the relation

$$y_i=m\left(x_i\right)+\varepsilon_i \tag{6}$$

such that ε_i's are residuals which are assumed to be independently normally distributed with mean zero and variance $\sigma^2\left(x_i\right)$.

It is clear that

$$E\left(y_i/X=x_i\right)=m\left(x_i\right) \tag{7}$$

$$Cov\left(y_i,y_j\mid X=x_i,X=x_j\right)=\begin{cases}\sigma^2\left(x_i\right) & i=j\\ 0 & \text{otherwise}\end{cases} \tag{8}$$

where $m\left(x_i\right)$ is an unknown regression function which is a smooth function of *x*.

To obtain the estimator of $m\left(x_i\right)$ at y_{t-1} and its derivatives, we use the weighted local polynomial fitting by assuming that the regression function with $\left(p+1\right)th$ derivatives at a point, say $x=x_0$, exists and are continuous.

We can rewrite the imputation model (6) as

$$y_i=m_{y_{t-1}}\left(x_i\right)+\varepsilon_i \tag{9}$$

where approximation of $m_{y_{t-1}}\left(x_i\right)$ about y_{t-1} is done following the Taylor series expansion.

The kernel weight given as

$$w_i\left(x\right)=K\left\{\frac{\left(x_i-x_0\right)}{h}\right\} \tag{10}$$

where *h* is the bandwidth and *K* is the kernel function which should be strictly positive and $K_h\left(\cdot\right)$ controls the weights, x_0 is the point of focus and x_i being the covariate

with design matrix centered at past last value and j is the order of the local polynomial.

Let

$$S=\sum_{i=1}^{n}\left[y_i-\sum_{j=0}^{p}\beta_j\left(x_i-x_0\right)^j\right]^2 w_i\left(x\right) \tag{11}$$

Accordingly, for $j=0$,

$$\hat{m}_0\left(x\right)=\frac{\sum_{i=1}^{n}w_i\left(x\right)y_i}{\sum_{i=1}^{n}w_i\left(x\right)} \tag{12}$$

Equation (12) is the Nadaraya-Watson estimator.

With estimator the $\hat{m}_0\left(x\right)$, the conditional expectation given by $\hat{\varphi}\left(y_{t-1}\right)$ is used to impute the missing values, *i.e.*

$$\hat{\varphi}_{t,t-1}\left(y_{t-1}\right)=\frac{\sum_{i=1}^{n}K\left(\frac{x_i-x_0}{h}\right)\omega_i\mathbf{I}_i y_i}{\sum_{i=1}^{n}K\left(\frac{x_i-x_0}{h}\right)\omega_i\mathbf{I}_i} \tag{13}$$

where ω_i is the survey weight and

$$\mathbf{I}_{t,t-1,i}=\begin{cases}1, & I_{t,i}=1, I_{t-1,i}=1,\\ 0, & \text{otherwise}\end{cases}\quad \text{for } t=2,\cdots,T \tag{14}$$

Similarly for $j=1$,

$$S=\sum_{i=1}^{n}\left\{y_i-\beta_0-\beta_1\left(x_i-x_0\right)\right\}^2 w_i\left(x\right) \tag{15}$$

Minimizing S with respect to β_0 and β_1 in Equation (15) and solving for β_0 and β_1, we get

$$\hat{\beta}_0=\frac{\sum_{i=1}^{n}y_i w_i\left(x\right)\sum_{i=1}^{n}w\left(x\right)\left(x_i-x_0\right)^2-\sum_{i=1}^{n}y_i w_i\left(x\right)\left(x_i-x_0\right)\sum_{i=1}^{n}w_i\left(x\right)\left(x_i-x_0\right)}{\sum_{i=1}^{n}w\left(x\right)\left(x_i-x_0\right)^2\sum_{i=1}^{n}w_i\left(x\right)-\left(\sum_{i=1}^{n}w_i\left(x\right)\left(x_i-x_0\right)\right)^2} \tag{16}$$

and

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}y_i w_i\left(x\right)\left(x_i-x_0\right)\sum_{i=1}^{n}w_i\left(x\right)-\sum_{i=1}^{n}y_i w_i\left(x\right)\sum_{i=1}^{n}w_i\left(x\right)\left(x_i-x_0\right)}{\sum_{i=1}^{n}w\left(x\right)\left(x_i-x_0\right)^2\sum_{i=1}^{n}w_i\left(x\right)-\left(\sum_{i=1}^{n}\left(x_i-x_0\right)w_i\left(x\right)\right)^2} \tag{17}$$

Defining:

$S_j\left(x\right)=\sum_{i=1}^{n}w_i\left(x\right)\left(x_i-x_0\right)^j$ and

$T_j\left(x\right)=\sum_{i=1}^{n}y_i w_i\left(x\right)\left(x_i-x_0\right)^j$, Thus:

Using $S_j\left(x\right)$, in Equation (17), we obtain

$$\hat{\beta}_1=\sum_{i=1}^{n}\left\{\frac{\left(x_i-x_0\right)S_0\left(x\right)-S_1\left(x\right)}{S_2\left(x\right)S_0\left(x\right)-\left(S_1\left(x\right)\right)^2}\right\}w_i\left(x\right)y_i \tag{18}$$

and with $T_j(x)$, in Equation (17), it yields

$$\hat{\beta}_1 = \frac{S_0(x)T_1(x) - S_1(x)T_0(x)}{S_2(x)S_0(x) - S_1(x)^2} \tag{19}$$

Similarly, using $S_j(x)$, in Equation (16) gives

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left\{ \frac{S_2(x) - (x_i - x_0)S_1(x)}{S_2(x)S_0(x) - S_1(x)^2} \right\} y_i w_i(x) \tag{20}$$

and with $T_j(x)$, Equation (16) becomes

$$\hat{\beta}_0 = \frac{S_2(x)T_0(x) - T_1(x)S_1(x)}{S_2(x)S_0(x) - S_1(x)^2} \tag{21}$$

The local linear estimator for the regression function $m_1(x)$ is now given by:

$$\hat{m}_1(x) = \hat{\beta}_0 + (x_i - x_0)\hat{\beta}_1 \tag{22}$$

Substituting for $\hat{\beta}_0$ (from Equation (20)) and $\hat{\beta}_1$ (from Equation (18)) in Equation (22) gives,

$$\hat{m}_1(x) = \sum_{i=1}^{n} \left\{ \frac{S_2(x) - S_1(x)(x_i - x_0)}{S_2(x)S_0(x) - S_1(x)^2} \right\} w_i(x) y_i + (x_i - x_0) \sum_{i=1}^{n} \left\{ \frac{(x_i - x_0)S_0(x) - S_1(x)}{S_2(x)S_0(x) - (S_1(x))^2} \right\} w_i(x) y_i \tag{23}$$

With estimator, $\hat{m}_1(x)$, the conditional expectation given by $\hat{\varphi}(y_{t-1})$ is used to impute the missing values, *i.e.*

$$\begin{aligned} \hat{\varphi}_{t,t-1}(y_{t-1}) = & \sum_{i=1}^{n} \left\{ \frac{\left[S_2(x) - S_1(x)(x_i - x_0) \right] \omega_i \mathbf{I}_i}{\left[S_2(x)S_0(x) - S_1(x)^2 \right] \omega_i \mathbf{I}_i} \right\} w_i(x) y_i \\ & + (x_i - x_0) \sum_{i=1}^{n} \left\{ \frac{\left[(x_i - x_0)S_0(x) - S_1(x) \right] \omega_i \mathbf{I}_i}{\left[S_2(x)S_0(x) - S_1(x)^2 \right] \omega_i \mathbf{I}_i} \right\} w_i(x) y_i \end{aligned} \tag{24}$$

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

4.2. 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, imputation process is done. 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. 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 $i \in B_{n-r}$ 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 $\overline{Y}_t$ be the finite population mean at time point, t for $t = 1, 2, \cdots, T$.

The value to be imputed for the non respondent is denoted by $y_{i,t}^*$ such that the imputed data is given as

$$y_{i,t}^{\#} = \begin{cases} y_{i,t}, & i \in B_r \quad \text{observed value} \\ y_{i,t}^{*}, & i \in B_{n-r} \quad \text{imputed value} \end{cases} \tag{25}$$

The mean of the finite population is given by

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

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

$$y_I = \sum_{i \in B_n} \left[y_{i,t} I_i + (1 - I_i) y_{i,t}^{*} \right] \tag{26}$$

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

$$\bar{y}_I = \sum_{i \in B_n} \omega_i y_{i,t}^{\#} \tag{27}$$

Assuming that for each $i \in B_n$

$$E_s\left(\sum_{i \in B_n} \omega_i y_i \right) = \sum_{i=1}^{N} y_i \tag{28}$$

for each $i \in P$.

The imputed values are treated as if they were observed such that both observed and the imputed are used in the estimation of the population mean:

Sample mean for the imputed data becomes

$$\bar{y}_I = \left\{ \sum_{i \in B_r} w_i y_{i,t} + \sum_{i \in B_{n-r}} w_i y_{i,t}^{*} \right\} \tag{29}$$

Note that the same weight due to sampling design is used in Equation (29) for all units in the sample.

$$\bar{y}_I = \frac{1}{n}\left\{ \sum_{i \in B_r} y_{i,t} + \sum_{i \in B_{n-r}} y_{i,t}^{*} \right\} \tag{30}$$

for $t = 1, \cdots, T$.

Since t is used as a constant variable, Equation (30) is re-written as

$$\hat{\bar{y}}_t = \frac{1}{n}\left(\sum_{i \in B_r} y_i + \sum_{i \in B_{n-r}} y_i^{*} \right) \tag{31}$$

As for [12], the local constant estimation for the nonrespondents in Equation (31) is obtained as:

$$y_i^{*} = \hat{\varphi}_{t,t-1}(y_{t-1}) = \frac{\sum_{i \in S} K\left(\dfrac{x - y_{i,t-1}}{h} \right) \omega_i \mathbf{I}_{i,t,t1} y_{i,t}}{\sum_{i \in s} K\left(\dfrac{x - y_{i,t-1}}{h} \right) \omega_i \mathbf{I}_{i,t,t-1}} \tag{32}$$

and the local linear estimation for the nonrespondents, y_i^{*} in Equation (31) is given by:

$$y_i^* = \hat{\varphi}_{t,t-1}(y_{t-1}) = \sum_{i=1}^{n} \left\{ \frac{\left[S_2(x) - S_1(x)(x_i - x_0)\right]\omega_i \mathbf{I}_{i,t,t-1}}{\left[S_2(x)S_0(x) - S_1(x)^2\right]\omega_i \mathbf{I}_{i,t,t-1}} \right\} w_i(x) y_i$$
$$+ (x_i - x_0) \sum_{i=1}^{n} \left\{ \frac{\left[(x_i - x_0)S_0(x) - S_1(x)\right]\omega_i \mathbf{I}_{i,t,t-1}}{\left[S_2(x)S_0(x) - S_1(x)^2\right]\omega_i \mathbf{I}_{i,t,t-1}} \right\} w_i(x) y_i \quad (33)$$

Clearly, y_i^* in Equation (31) is substituted by Equation (32) and Equation (33) for use of local constant estimator and local linear regression estimator respectively.

5. Asymptotic Properties of the Estimator

In the derivation of the asymptotic properties, we use the set of regularity conditions. According to [12], the asymptotic theory development is provided by the concept of a sequence of finite populations $\{P_\nu\}_{\nu=1}^{\infty}$ with ν strata in P_ν. It is assumed 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 $\nu \to \infty$. The uniform response and the size m_ν of the nonrespondents set B_{n-r} satisfy the condition $\frac{m_\nu}{n_\nu} \to \alpha < 1$. All limiting processes will be understood as $\nu \to \infty$ such that the regularity conditions are satisfied. For easy notation, the subscript ν will be ignored in the subsequent work.

Theorem 1. *Assuming the regularity conditions* (i)-(iv) *and also the assumptions in section* 2 *hold. Then under the regression imputation model* η, (6), *the estimator,* $\hat{\bar{y}}_t$ *in Equation* (31), *is asymptotically unbiased and consistent for the population mean* $\bar{Y}_t$.

Proof. 1) Bias of $\hat{Y}_t$.

The general formula for the finite population total is given by:

$$Y_t = \sum_{i \in B_n} y_i + \sum_{i \in B_{N-n}} y_i \quad (34)$$

where B_n and B_{N-n} are the sampled and the non sampled sets respectively.

Equation (34) can be decomposed as

$$Y_t = \left(\sum_{i \in B_r} y_i + \sum_{i \in B_{n-r}} y_i \right) + \sum_{i \in B_{N-n}} y_i \quad (35)$$

For simplicity, denote B_r, B_{n-r} and B_{N-n} by r, $(n-r)$ and $(N-n)$ respectively throughout the remaining work.

From Equation (31), the estimator for the finite population total is given by

$$\hat{Y}_t = \sum_{i \in r} y_i + \sum_{i \in n-r} \hat{m}_1(x_i) \quad (36)$$

Now consider the difference,

$$\hat{Y}_t - Y_t = \left(\sum_{i \in r} y_i + \sum_{i \in n-r} \hat{m}_1(x_i) \right) - \left(\sum_{i \in r} y_i + \sum_{i \in n-r} y_i \right) - \sum_{i \in N-n} y_i$$

$$\hat{Y}_t - Y_t = \sum_{i\in n-r}\left(\hat{m}_1(x_i) - y_i\right) - \sum_{i\in N-n} y_i \tag{37}$$

$$\hat{Y}_t - Y_t = \sum_{i\in n-r}\left(\left[\hat{m}_1(x_i) - m_1(x_i)\right] + \left[m_1(x_i) - y_i\right]\right) - \sum_{i\in N-n} y_i \tag{38}$$

Taking expectation on both sides of Equation (38), we have

$$E\left(\hat{Y}_t - Y_t\right) = \sum_{i\in n-r} E\left(\hat{m}_1(x_i) - m_1(x_i)\right) + \sum_{i\in n-r} E\left(m_1(x_i) - y_i\right) - \sum_{i\in N-n} E(y_i) \tag{39}$$

Clearly, $\sum_{i\in n-r} E\left(m(x_i) - y_i\right) = 0$ since $E(y_i) = m(x_i)$.

Now,

$$E\left(\hat{Y}_t - Y_t\right) = \sum_{i\in n-r} E\left(\hat{m}_1(x_i) - m_1(x_i)\right) - \sum_{i\in N-n} E(y_i) \tag{40}$$

$$E\left(\hat{Y}_t - Y_t\right) = \sum_{i\in n-r} E\left(\hat{m}_1(x_i) - m_1(x_i)\right) - \sum_{i\in N-n} m(x_i) \tag{41}$$

Assuming $n, N \to \infty$ such that $(N-n) \to 0$, then $\sum_{i\in N-n} m(x_i) \to 0$ in Equation (41) and hence,

$$E\left(\hat{Y}_t - Y_t\right) \approx \sum_{i\in n-r} E\left(\hat{m}_1(x_i) - m_1(x_i)\right) \tag{42}$$

But from Lemma 1 (see **Appendix**),

$$E\left(\hat{m}_1(x)\right) = m_1(x_0) + uhm_1'(x_0) + \frac{h}{2k_2}\left(hk_2^2 + (uh)k_3\right)m_1''(x_0) \tag{43}$$

where $uh = x_i - x_0$.

Thus the bias of $\hat{Y}_t$ becomes

$$Bias\left(\hat{Y}_t\right) = \sum_{i\in n-r}\left\{\left(x_0 - x_i\right)m_1'(x_0) + h\left(\frac{hk_2^2 + \left(x_0 - x_i\right)k_3}{k_2}\right)\frac{m_1''(x_0)}{2}\right\} \tag{44}$$

2) Variance of $\hat{Y}_t$.

The variance of $\hat{Y}_t$ is given by the variance of the error term $\hat{Y}_t - Y_t$. That is,

$$Var\left(\hat{Y}_t\right) = Var\left(\hat{Y}_t - Y_t\right) \tag{45}$$

$$= Var\left(\sum_{i\in n-r}\left(\hat{m}_1(x_i) - y_i\right) - \sum_{i\varepsilon N-n} y_i\right) \tag{46}$$

$$= \sum_{i\in n-r}\sum_{i=1}^{n} w_i^{*2}(x)\sigma^2(x_i) - \sum_{i\in n-r}\sigma^2(x_i) - \sum_{i\in N-n}\sigma^2(x_i) \tag{47}$$

$$Var_{asy}\left(\hat{Y}_t - Y_t\right) \approx \frac{1}{nh}d_k\sigma^2(x_0) - (n-r)\sigma^2(x_i) - (N-n)\sigma^2(x_i) \tag{48}$$

Thus,

$$Var_{asy}\left(\hat{Y}_t - Y_t\right) \approx \sum_{i\in n-r}\frac{1}{nh}d_k\sigma^2(x_0) \tag{49}$$

for sufficiently large n such that $(N-n) \to 0$ and $(n-r) \to 0$; where $d_k = \int K^2(u)du$.

3) Mean square error (*MSE*) of $\hat{Y}_t$.

Finally, we have

$$MSE\left[\hat{Y}_t\right] = \left[Bias\left(\hat{Y}_t\right)\right]^2 + Var\left(\hat{Y}_t\right) \tag{50}$$

$$MSE\left[\hat{Y}_t\right] = \left\{\sum_{i\in n-r}\left\{\left(x_0 - x_i\right)m'\left(x_0\right) + h\left(\frac{hk_2^2 + \left(x_0 - x_i\right)k_3}{K_2}\right)\frac{m''\left(x_0\right)}{2}\right\}\right\}^2 + \sum_{i\in n-r}\frac{1}{nh}\sigma^2\left(x_0\right)d_k \tag{51}$$

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.

6. Simulation Study

6.1. Description of Longitudinal Data

In this section, 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 (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 $n = 1500$ was considered. 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 non-respondent values being denoted by 1 and 0 respectively at different time points.

Longitudinal data was generated according to two models:

1) In model 1, simulation of $\left(y_i, i = 1, 2, 3, 4\right)$ 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.

2) In model 2, simulation of $\left(\log\left(y_i\right), i = 1, 2, 3, 4\right)$ 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.

In order to obtain the nonmonotone pattern in the simulated data, we used the predetermined unconditional probabilities of [13] shown in **Table 1**.

6.2. 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 1500 times through 1000 simulation runs.

Table 1. Probabilities of nonresponse patterns for $t=4$.

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

2) A simple random sample of size 200 was 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.

5) Obtain the bootstrap variance of $\hat{Y}_I$ by the formula $V_{boot}\left(\hat{Y}_I\right)=\frac{1}{B}\sum_{b=1}^{B}\left(\hat{Y}_I^{(b)}-\hat{Y}_I^{(.)}\right)^2$ where $\hat{Y}_I^{(.)}$ is the mean bootstrap analog of $\hat{Y}_I$, given by $\hat{Y}_I^{(.)}=\frac{1}{B}\sum_{b=1}^{B}\hat{Y}_I^{(b)}$.

6.3. Results and Discussion

The results of this simulation study are summarized in Table 2 and Table 3.

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 followed by the Nadaraya-Watson estimator and then the simple linear regression estimator, which was the largest value of %RB.

At time point $t=3$, observe that the the simple linear regression estimator has the least %RB value compared to that of the local linear estimator and the Nadaraya-Watson estimator performed worst with the largest %RB. 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 %RB value followed by the simple linear regression estimator and the Nadaraya-Watson estimator performed worst. 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 point $t=2$, Nadaraya-Watson estimator has the least values followed by the local linear estimator and lastly the simple linear regression estimator which has the largest values. At time point $t=3$, the local linear estimator has the least values of *MSE* followed by the simple linear regression estimator and lastly

Table 2. Simulated results for mean estimation (normal case).

Method	Quantity	$t=1$	$t=2$	$t=3$	$t=4$
Complete data	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
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
	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
	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
	SD bootstrap		0.9952162	0.9938139	0.993223

Table 3. Simulated results for mean estimation (log-normal case).

Method	Quantity	$t=1$	$t=2$	$t=3$	$t=4$
Complete data	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
	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
	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
	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
	SD bootstrap		0.9940906	0.9909141	0.9916702

the Nadaraya-Watson estimator which has the largest *MSE* value. At time points $t=4$, Nadaraya-Watson estimator has the least values of *MSE* followed by the simple linear regression estimator and lastly the local linear estimator which has the largest *MSE* value.

In terms of the bootstrap standard deviation, it can be seen that the local linear estimator performs the best at all the three time points $t=2$, $t=3$, and $t=4$ in which its values are even lower than those of the complete data implying that the results got with the local linear estimator are the best. The simple linear regression and Nadaraya-Watson estimators are competing interchangeably in terms of performance for the bootstrap samples.

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 followed by the local linear estimator and the Nadaraya-Watson estimator has the biggest %RB values. 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*, followed by the simple linear regression estimator and lastly the local linear estimator which has the largest values of *MSE*. At time point $t=3$, the the local linear estimator has the least values implying that it performed well at time point $t=3$.

In terms of the bootstrap standard deviation, observe from Table 3 that the local linear estimator performs the best at all the three time points since it has the least bootstrap standard deviations and these values are even smaller than those of the complete data in order of increasing time.

From Table 3 of results, 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.

7. Conclusion

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 linear regression for nonmonotone nonrespondents in longitudinal surveys and determine its asymptotic properties. Comparing the parametric and nonparametric methods, nonparametric methods performed better than the parametric methods. This was evident from the *MSE* and %RB values in both the normal and log-normal data. 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. In terms of the bootstrap standard deviation, the local linear estimator performs the best at all the three time points since it has the least bootstrap standard deviations 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.

Acknowledgements

We wish to thank the African Union Commission for fully funding this research.

References

[1] Dorfman, A.H. (1992) Nonparametric Regression for Estimating Totals in Finite Population. Proceeding Section of Survey Methodology. American Statistical Association Alexandria, VA, 622-625.

[2] Xu, J., Shao, J., Palta, M. and Wang. L. (2008) Imputation for Nonmonotone Last-Value-Dependent Nonrespondents in Longitudinal Surveys. *Survey Methodology*, **34**, 153-162.

[3] Nadaraya, E.A. (1964) On Estimating Regression. *Theory of Probability and Its Applications*, **9**, 141-142.

[4] Watson, G.S. (1964) Smooth Regression Analysis. *Sankhy*: *The Indian Journal of Statistics*, **26**, 359-372.

[5] Hastie, T.J. and Loader, C. (1993) Local Regression: Automatic Kernel Carpentry (with Discussion). *Statistical Science*, **8**, 120-143.

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

[7] Cai, Z. (2001) Weighted Nadaraya-Watson Regression Estimation. *Statistics & Probability Letters*, **51**, 307-318.

[8] Fan, J. and Gijbels, I. (1996) Local Polynomial Modelling and Its Applications. Chapman and Hall, London.

[9] Stone, C.J. (1977) Consistent Nonparametric Regression. *The Annals of Statistics*, **3**, 595-620.

[10] Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Inc., New York. https://doi.org/10.1002/9780470316696

[11] 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.

[12] Cheng, P.E. (1994) Nonparametric Estimation of Mean Functionals with Data Missing at Random. *Journal of the American Statistical Association*, **89**, 81-87.

[13] Shao, J., Klein, M. and Xu, J. (2012) Imputation for Nonmonotone Nonresponse in the Survey of Industrial Research and Development. *Survey Methodology*, **38**, 143-155.

[14] Masry, E. (1996) Multivariate Local Polynomial Regression for Time Series. Uniform Strong Consistence and Rates. *Journal of Time Series Analysis*, **17**, 571-599.

Appendix

LEMMA 1. The bias of $\hat{m}_1(x)$ is given by

$$Bias\left(\hat{m}_1(x)\right) = uhm'(x_0) + \frac{h}{2k_2}\left(hk_2^2 + (uh)k_3\right)m''(x_0) \tag{52}$$

Under the regularity conditions in section 3, $Bias\left(\hat{m}_1(x)\right) \to 0$ as $h \to 0$ and $n \to \infty$.

PROOF OF LEMMA 1.

Proof. From Equation (23),

$$\hat{m}_1(x) = \sum_{i=1}^{n} w_i^*(x) y_i + (x_0 - x_i)\sum_{i=1}^{n} w_i^{**}(x) y_i \tag{53}$$

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

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

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

$$E\left[\hat{m}_1(x)\right] = \sum_{i=1}^{n} w_i^*(x) E[y_i] + (x_0 - x_i)\sum_{i=1}^{n} w_i^{**}(x) E[y_i] \tag{54}$$

$$E\left[\hat{m}_1(x)\right] = \{m(x_0) + (x_0 - x_i)m'(x_0)\} + \left\{\frac{\left[S_2^2(x) - S_1(x)S_3(x)\right] + \left[(x_0 - x_i)\left(S_0(x)S_3(x) - S_1(x)S_2(x)\right)\right]}{2\left[S_2(x)S_0(x) - S_1(x)^2\right]}\right\} m''(x_0) \tag{55}$$

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

$$Bias\left(\hat{m}_1(x)\right) = (x_0 - x_i)m'(x_0) + \left\{\frac{\left[S_2^2(x) - S_1(x)S_3(x)\right] + \left[(x_0 - x_i)\left(S_0(x)S_3(x) - S_1(x)S_2(x)\right)\right]}{2\left[S_2(x)S_0(x) - S_1(x)^2\right]}\right\} m''(x_0) \tag{56}$$

For fixed design points of x_i's on the interval $(0,1)$, the expression $S_j(x) = \sum_{i=1}^{n}(x_i - x_0)^j w_i(x) \equiv nh^{j+1}k_j + o\left(nh^{j+3}\right)$ almost everywhere, see [14].

Now,

1) $$S_2^2(x) - S_1(x)S_3(x) = \left[nh^3k_2 + o\left(nh^5\right)\right]^2 - \left[o\left(nh^4\right)\right]\left[nh^4k_3 + o\left(nh^6\right)\right] = n^2h^6k_2^2 + o\left(n^2h^8\right)$$

2) $$S_0(x)S_3(x) - S_1(x)S_2(x) = \left[nh + o\left(nh^3\right)\right]\left[nh^4k_3 + o\left(nh^6\right)\right] - \left[o\left(nh^4\right)\right]\left[nh^3k_2 + o\left(nh^5\right)\right] = n^2h^5k_3 + o\left(n^2h^7\right)$$

3) $$S_2(x)S_0(x) - S_1(x)^2 = \left[nh^3k_2 + o\left(nh^5\right)\right]\left[nh + o\left(nh^3\right)\right] - \left[o\left(nh^4\right)\right]^2 = n^2h^4k_2 + o\left(n^2h^6\right)$$

4) $S_0(x)S_1(x)-S_1(x)S_0(x)=\left[nh+o\left(nh^3\right)\right]o\left(nh^4\right)-o\left(nh^4\right)\left[nh+o\left(nh^3\right)\right]=0$

Equation (56) becomes

$$Bias\left(\hat{m}_1(x)\right)=\left(x_0-x_i\right)m'\left(x_0\right)+\left\{h^2k_2^2+\left[\left(x_0-x_i\right)hk_3\right]\right\}\frac{m''\left(x_0\right)}{2k_2} \tag{57}$$

Letting $u=\frac{x_i-x_0}{h}\Rightarrow uh=x_i-x_0$.

Hence, the bias of $\hat{m}_1(x)$ can be re-written as

$$Bias\left(\hat{m}_1(x)\right)=uhm'\left(x_0\right)+\frac{h}{2k_2}\left(hk_2^2+(uh)k_3\right)m''\left(x_0\right) \tag{58}$$

and hence the result.

LEMMA 2. The asymptotic expression of the variance of $\hat{m}_1(x)$ is given by

$$Var\left(\hat{m}_1(x)\right)\approx\frac{d_k}{nh}\sigma^2\left(x_0\right) \tag{59}$$

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

PROOF OF LEMMA 2.

Proof. Using Equation (23),

$$Var\left(\hat{m}_1(x)\right)=\sum_{i=1}^{n}w_i^{*2}(x)Var\left(y_i\right)+\left(x_0-x_i\right)^2\sum_{i=1}^{n}w_i^{**2}(x)Var\left(y_i\right) \tag{60}$$

since $Cov\left(y_i,y_j\right)=0$.

It follows that

$$Var\left(\hat{m}_1(x)\right)=\sum_{i=1}^{n}w_i^{*2}(x)\sigma_i^2(x)+\left(x_0-x_i\right)^2\sum_{i=1}^{n}w_i^{**2}(x)\sigma_i^2(x) \tag{61}$$

where

$$w_i^{*2}(x)\approx\frac{1}{n^2h^2}w_i^2(x) \tag{62}$$

and

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

as $n\to\infty$.

Thus,

$$Var\left(\hat{m}_1(x)\right)=\frac{1}{n^2h^2}\sum_{i=1}^{n}w_i^2(x)\sigma^2\left(x_i\right) \tag{63}$$

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

$$Var\left(\hat{m}_1(x)\right)\approx\frac{d_k}{nh}\sigma^2\left(x_0\right) \tag{64}$$

where $d_k=\int K^2(u)\mathrm{d}u$. Hence the result.

***MSE* of** $\hat{m}_1(x)$

From LEMMA 1 and 2, the *MSE* of $\hat{m}_1(x)$ becomes

$$MSE\left[\hat{m}_1(x)\right] = \left\{(x_0 - x_i)m'(x_0) + \frac{h}{2k_2}\left(hk_2^2 + (x_0 - x_i)k_3\right)m''(x_0)\right\}^2 + \frac{d_k}{nh}\sigma^2(x_0) \quad (65)$$