NON-PARAMETRIC REGRESSION ESTIMATION OF A FINITE POPULATION TOTAL IN THE PRESENCE OF HETEROSCEDASTICITY

BY

CELESTINE K. INGUTIA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN APPLIED STATISTICS

DEPARTMENT OF MATHEMATICS AND APPLIED STATISTICS

MASENO UNIVERSITY

ABSTRACT

Non parametric regression provides computationally intensive estimation of unknown finite population quantities. Such estimation is usually more flexible and robust than inferences tied to design – probabilities (in design-based inference) or to parametric regression models in (model-based inference). Dorfman [9] used a more general super population model to find a non-parametric regression based estimator for the population total T. He, however, assumed homoscedasticity when constructing his proposed estimator. In his empirical study, he noted that the data showed clear signs of heteroscedasticity. In this study we consider the improvement in the efficiency of Dorfman's non-parametric regression based estimator of a finite population. To do this we incorporate a reasonable assumption of variance structure into the non-parametric regression methodology and use the weighted least squares method to obtain the proposed non-parametric regression based estimator. In our empirical work we have used two kinds of data sets: simulated and secondary data. The simulated data is of two kinds: homoscedastic and heteroscedastic generated with the help of Genstat 8th edition statistical application package. The secondary data was obtained from the internet from the United States Bureau of Labor Statistics. By calculating Dorfman's and our population estimates based on the given data sets using Dorfman's and our proposed estimator's respectively, we have established that our proposed estimator is more efficient than Dorfman's, that is, the efficiency of Dorfman's non-parametric regression based estimator has been improved when we put into account heteroscedasticity.

CHAPTER 1

1.1 INTRODUCTION

1.1.1 Background information

We consider a finite population that consists of N identifiable units, $\underline{U} = \{U_1, U_2, ..., U_N\}$. Associated with each unit of the population is a certain characteristic or variable Y of interest whose values are $\underline{Y} = \{Y_1, Y_2, ..., Y_N\}$. In some cases there may be available values of another variable or variables X for each unit of the population with values of X assumed to be known for every unit of the population.

In most cases interest is not in obtaining the values of Y for each unit of the population, rather interest is in obtaining inference about some function $T(\underline{Y}) = T\{Y_1, Y_2, ..., Y_N\}$of Y called the population parameter. Examples of population parameters are: the population total $\left(T = \sum_{i=1}^{N} Y_i\right)$, the population mean $\left(\overline{T} = \frac{1}{N}\sum_{i=1}^{N} Y_i\right)$, the population variance $\left(\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left[Y_i - \overline{T}\right]^2\right)$, the population ratio $\left(R = \frac{\overline{T}}{\overline{X}}\right)$ and the population distribution function $F(y) = N^{-1}\sum_p I(Y_i \leq y)$. Hence $F(y)$ is useful in estimating population quantiles.

There are two main ways of obtaining information about $T(\underline{Y})$. The first one is to enumerate all the units of the population and then calculate $T(\underline{Y})$. This is called the complete enumeration method or census. This method has a number of drawbacks. The main one is that if the population size, N, is large then this method can be time consuming and costly. An alternative method is to take part of the population, called the sample, make observations on the units constituting the sample and then use the obtained sample values to make inference about $T(\underline{Y})$. This alternative method is preferred over the census method due to the following reasons: the sampling results can be obtained more rapidly and data analyzed much faster due to less time involved, reduced costs and greater accuracy of observations. There are two main problems in a sample survey, namely, the design and inference problems. Hence sample survey is a two-fold problem. The design problem is mainly concerned with the choice of the sample, that is, the method of choosing units of the population to constitute the sample. There are two methods of choosing the sample: subjective sampling and probability sampling. Under the inference problem, we look at how to use the obtained information to make inferences about $T(\underline{Y})$. In this project we focus on the inference problem.

In most sample survey problems, inferences have been done on the population total or mean. Hence we estimate the population total, T of Y. Work done on the distribution function may be found in Chambers, Dorfman, and Wehrly [6] and Dorfman and Hall [10]. In the next section, we look at different approaches of making inference about T.

1.1.2 Approaches to sample survey inference

There are four different approaches to making inference about T: design based, model-based, model-assisted and randomization assisted model-based approaches.

1.1.2.1 Design-based approach

This approach has its origin in Neyman's key paper, Neyman [25]. It has also been discussed in standard survey texts such as Hansen, Hurwitz and Madow [15], Kish [18] and Cochran [7]. See also Royal and Cumberland [27].

The main characteristic of design-based inference is that it is based on the distribution of $I = (I_1, I_2, ..., I_N)$, the set of inclusion indicator variable, where $I_i = 1$ if unit i is included in the sample and $I_i = 0$ if it is not included, with the survey variables Y treated as fixed quantities.

The key concept in this approach is that of design unbiasedness, Chambers [5]. Thus, for any choice of sampling process, S, the weighting process, W, must be such that the frequency weighted average value of $\hat{T}$ over all possible samples generated under S is the actual value of T. In other words, this approach restricts consideration to those weights W which ensure that, irrespective of the particular sample selection method (that is S) used,

$$\mathrm{E}\left(\hat{T} - T/X, Y\right) = 0, \ \forall \text{ values of } X \text{ and } Y. \qquad (1)$$

1.1.2.2 Model-based approach

This approach has been linked with the work of Richard Royall and others. A summary of the philosophy behind this approach is set out in Royall [26]. The model-based approach is based on the assumption that the values of Y can be assumed to be realizations of random variables whose distribution conditional on the known values of

X may be specified through a convenient probability model. For instance, for a simple linear regression model, Y_i is taken to be

$$Y_i = \alpha + \beta x_i + \sigma(x_i)e_i, \; (i = 1,2,\ldots,N) \tag{2}$$

with α and β unknown, $\sigma(x_i)$ known and e_i is normal with mean zero and unknown but constant variance. An appropriate model based estimator is then given by:

$$\hat{T}_{lin} = \sum_s Y_i + \sum_{\tilde{s}} (\hat{\alpha} + \hat{\beta} x_j) \tag{3}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the appropriate weighted least squares estimators of α and βrespectively, $\hat{T}_{lin}$ denotes the model based estimator for the linear regression, and, $\sum_s$ and $\sum_{\tilde{s}}$ denote summations over sample and non-sample values respectively.

A commonly used model for Y expresses the mean and variance of Y as proportional to X. That is

$$\mathrm{E}(Y_i / x_i) = \beta x_i$$

$$Cov(Y_i, Y_j / x_i, x_j) = \begin{cases} \sigma^2 x_i, i = j \\ 0, i \neq j \end{cases} \tag{4}$$

where β and σ^2 are unknown positive constants.

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \frac{[Y_i - \beta x_i]^2}{\sigma^2 x_i}$$

Differentiating the above equation with respect to β and equating to zero we have

$$\frac{ds}{d\beta} = -2\sum_{i=1}^{n} \frac{x_i[Y_i - \beta x_i]}{\sigma^2 x_i} = 0$$

The best linear unbiased estimator of β is

$$\hat{\beta} = \left(\sum_s y_i \Big/ \sum_s x_i\right) = \overline{y_s}/\overline{x_s} \qquad (5)$$

The best linear unbiased predictor of T is the ratio estimator

$$\hat{T}_R = \left(\sum_s y_i \Big/ \sum_s x_i\right) \sum_{i=1}^{N} X_i = N\hat{\beta}\overline{X} \qquad (6)$$

where $\overline{y_s}$ and $\overline{x_s}$ are means of the sample values of Y and X respectively and $\overline{X}$ is the population mean of X. Therefore the ratio estimator is the optimal estimator under the model in equation (4).

Now suppose that there are errors in the assumed model (4). Will the ratio estimator still be unbiased? For example, suppose $\mathrm{E}(Y_i/x_i) \neq \beta x_i$ but $\mathrm{E}(Y_i/x_i) = \alpha + \beta x_i$, or $\mathrm{var}(Y_i/x_i) \neq \sigma^2 x_i$, or still, suppose the random variables Y_i, Y_j are dependent, that is, $\mathrm{cov}(Y_i, Y_j/x_i, x_j) \neq 0, i \neq j$, will the ratio estimator still be unbiased and optimal? Intuitively, we would want to use an estimator that is optimal (or approximately so) under the given model but remains optimal (or approximately so) when there are errors in the model. In other words, we need a robust estimator. This is the major problem that needs to be addressed by researchers using the model-based approach. We look at this approach in chapter 2.

As discussed earlier, the concept of design-unbiasedness is crucial to both the design-based as well as the model-assisted approaches to defining a sampling strategy. However, under the model-based approach this basic requirement is abandoned, Chambers [5]. Since design-unbiasedness is no longer a requirement, the obvious alternative property we require of $\hat{T}$ under this approach is that it be model-unbiased, that is,

$$\mathrm{E}_M\left(\hat{T}-T/S,X\right)=0 \tag{7}$$

where E_M indicate expectation under a given model M.

In other words, the values of the estimation errors $\hat{T}-T$ obtained for all population realizations Y consistent with the actual values of X observed, and the sample S actually obtained, should average out to zero. From now on, we consider the model-based approach to sample survey inference theory, and in the next section we review work that has been done to deal with the robustness problem in model-based surveys.

1.1.2.3 Model-assisted approach

In this approach, practitioners have been willing to use models in order to identify optimal strategies for estimating T. There are two versions of this approach.

- **Model-assisted strategies that are also design-unbiased**

This approach is comprehensively discussed in the text by S ä rndal, Swensson and Wretman [30]. Typically, the approach still assumes that the weighting variable, W at least approximately satisfies (1), that is, the resulting estimator $\hat{T}$ is design-unbiased, or approximately so. More information on this approach can be seen in Chambers [5]. Breidt and Opsomer [1] used the local polynomial kernel as the smoothing tool to develop a design consistent model-assisted estimator of the total. Kim, Breidt and Opsomer [17] have extended this work to two-stage sampling. Other model-assisted estimators have been proposed in S ä rndal, Swensson and Wretman [29,30], Little [22], and Firth and Bennett [13].

- **Model-assisted strategies that are design-unbiased on average**

The requirement that $\hat{T}$ be design-unbiased (or approximately so) is rather strong. An appealing extension of the model-assisted approach, whose motivation follows along the same lines as those leading to the use of the average mean square error, is discussed in Brewer [2]. This replaces the design-unbiasedness requirement by the weaker requirement that the design bias of $\hat{T}$ averages out to zero over possible values of Y. Thus, rather than exact (or approximately exact) design-unbiasedness, one requires average design-unbiasedness, or

$$\mathrm{E}\left(\hat{T}-T/X\right)=\mathrm{E}_M\left(E_D\left(\hat{T}-T/X,Y\right)/X\right)=0 \qquad (8)$$

where E_M and E_D indicate expectations under a given model, M and design, D respectively.

1.1.2.4 Randomization-assisted model-based approach

In this approach one treats model-based inference as the goal of survey sampling, but employs randomization methods to protect against inevitable model failure. For further elaboration on this type of approach, see S $\ddot{a}$ rndal, Swensson and Wretman [29, 30] and Kott [19, 20].

1.2 STATEMENT OF THE PROBLEM

The main interest in model-based approach to statistical survey inference is to overcome the problem of robustness under model misspecifications.

Dorfman [9] used a more general super population model to find a non-parametric regression based estimator for the population total T. He, however, assumed homoscedasticity when constructing his proposed estimator. In his empirical study, he noted that the data showed clear signs of heteroscedasticity. Hence the problem was: To estimate the population total T when there is heteroscedasticity in the data.

1.3 OBJECTIVES OF THE STUDY

Main objective

The main objective of this study was to find a non-parametric regression based estimator of the population total that takes into account heteroscedasticity.

Specific objectives

- To determine the properties of the attained estimator.

- To assess the performance of the estimator as compared to other existing estimators in an empirical study using both secondary and simulated data.

1.4 SIGNIFICANCE OF THE STUDY

The use of a general super population model gave rise to a robust estimator of the population total. This thesis has made use of known heteroscedasticity to innovatively construct a new estimator which can give sound inference in model based surveys. While assessing its performance as compared to Dorfman's in an empirical study, the estimator was found to be better. This study will enable sample survey practitioners who have adopted the model-based approach in survey sampling, to analyze data that is heteroscedastic in nature.

CHAPTER 2

2.1 LITERATURE REVIEW

Model based approach as a sample survey strategy has been discussed in Chambers [5]. This approach has been most strongly linked with the work of Richard Royall and others. An elegant summary of the philosophy behind this approach is set out in Royall [26].

The basic idea in the model-based approach is that it is based on the assumption that the values of Y can be assumed to be realizations of random variables whose distribution conditional on the known values of X may be specified through a convenient probability model. The advantages and disadvantages of this approach have been hotly debated in the sampling theory literature as can be seen in the sequence of papers of Smith [31, 32, and 33].

The model-based approach is used in making inference from sample to population. Given the population total, T, we have

$$T = \sum_{s} Y_i + \sum_{\tilde{s}} Y_i \qquad (9)$$

In this approach a regression model of Y on X is used to predict the non-sample Y's and, by consequence, their total.

In the parametric approach in survey sampling it is assumed that the mean curve has some prespecified functional form, like a line with unknown slope and intercept. The functional form is fully described by a finite set of parameters, see Hardle [14]. The parametric approach, however, has problems, for instance, from equation (4), let us consider the simple case that the regression of Y on X is a straight line with an intercept. That is, the model is:

$$E(Y_i/x_i) = \alpha + \beta x_i$$

$$Var(Y_i/x_i) = \sigma^2(x_i)$$

$$Cov(Y_i, Y_j/x_i, x_j) = 0, i \neq j. \qquad (10)$$

Then under the new model,

$$E(\bar{y}_R) = \frac{\bar{X}}{\bar{x}}\alpha + \beta\bar{X} \qquad (11),$$

while $$E(\bar{Y}) = \alpha + \beta\bar{X} \qquad (12).$$

Clearly the ratio estimator is biased. The problem we are considering above is the robustness problem. Under a given model, we can obtain the optimal estimator. The question we have been considering is this: what happens to this estimator when there are misspecifications in the model? Does it remain unbiased? Does its efficiency remain high? We have shown under the model in equation (4) that in the case of simple linear regression model, the ratio estimator is the optimal estimator. However, if the expected value part of the model is wrong, then the ratio estimator becomes biased. Hence a preselected parametric model might be too restricted or too low-dimensional to fit unexpected features.

The idea of non-parametric regression was first looked at by Nadaraya [24] and Watson [35]. A further reference is Hardle [14]. Other books on non-parametric regression are Wand and Jones [34] and Fan and Gibjels [12].

The non-parametric smoothing approach offers a flexible tool in analyzing unknown regression relationships. The term non-parametric thus refers to the flexible functional form of the regression curve. The non-parametric approach to estimating a regression curve has four main purposes: it provides a versatile method of exploring a general relationship between two variables, it gives predictions of observations yet to be made without reference to a fixed parametric model, it provides a tool for finding spurious observations by studying the influence of isolated points and lastly it contributes a flexible method of substituting for missing values or interpolation between adjacent X values.

Given the concern with robustness, it is natural to consider a non-parametric class of models for ξ, because they allow the models to be correctly specified for much larger classes of functions. Kuo [21], Dorfman [9], Dorfman and Hall [10], Chambers [4] and Chambers, Dorfman and Wehrly [6] have adopted the Superpopulation approach in constructing model-based estimators. Other work on non-parametric regression can be seen in Fan [11] and Kim [16].

Heteroscedastic regression has been studied by a number of researchers. Non-parametric work on heteroscedastic models has mostly been univariate and its aim was to obtain a weighted regression to estimate the mean function more efficiently. This can be seen from Carroll [3], M$\ddot{u}$ller and Stadtm$\ddot{u}$ller[23], Ruppert et al [28] and Dette, Munk and Wagner [8].

Dorfman [9] considered the following general non-parametric regression model for estimating population totals in finite populations:

$$Y_i = \mathrm{m}(x_i) + \sigma(x_i)\, e_i\,,\ (i = 1,2,\ldots,N) \tag{13}$$

where m (.) is a smooth function, e_i is a sequence of independent random variables with mean zero and variance one.

If $\sigma(x_i) = \sigma^2$, a constant, then the model is said to be homoscedastic, that is, the model is said to have constant variance. In situations where $\sigma(x_i)$ is a function of X_i, then the model is said to be heteroscedastic, that is, the variance varies depending on the X_i's.

Dorfman's non-parametric population total estimator is given by

$$\hat{T}_D = \sum_{s} Y_i + \sum_{\tilde{s}} \hat{m}(x_j), \tag{14}$$

where $\hat{m}(x_j) = \sum_{i} w_i(x_j) y_i$ and

$w_i(x_j) = \dfrac{k_b(x_i - x\ )}{\sum_s k_b(x_i - x\ )}$, is the weight associated with the i$^{\text{th}}$ unit of the sample.

Further $k(u)$ is a kernel function, b a scaling factor such that

$$K_b(u) = b^{-1} k\left(u/b\right)$$

The error variance of the population total estimator due to Dorfman [9] is given by

$$\operatorname{var}\left(\hat{T}_D - T/X_p\right) = \sum_i w_i^2 \sigma^2\left(x_i\right) + \sum_j \sigma^2\left(x_j\right) \tag{15}$$

where

$$w_i = \frac{\sum_{\bar{s}} k\left[\frac{x_i - x_j}{b}\right]}{k\left[\frac{x_i - x_j}{b}\right]}$$

and X_p is the population vector of x values.

In his empirical study he compared his proposed estimator with two design-based estimators of the total, namely, the expansion estimator and, a post stratified estimator and as can be noted from his empirical results, he illustrated that his proposed estimator performed well compared to the other estimators of the total.

Though Dorfman [9] began with a very general model, he did not take into account the heteroscedasticity in the data when constructing his proposed estimator. He clearly indicates to us that he noted that the data showed clear signs of heteroscedasticity, which he ignored when constructing his estimator. As a result, in our study we incorporated heteroscedasticity in obtaining a non-parametric regression based estimator using the model in equation (13) as our working model.

CHAPTER 3

THE PROPOSED ESTIMATOR

3.1 INTRODUCTION

In this chapter, the weighted least squares method was used to come up with our proposed estimator. The properties of the estimator, that is, mean and variance were then established.

3.2 METHODOLOGY

We assumed the model in equation (13) as our working model and estimated the population total, T, given in equation (9).

Since $\sum_{s} Y_i$ is known we needed to estimate $\sum_{\tilde{s}} Y_i$ as this contains observations outside the sample. Hence the problem of estimating $T = \sum_{i=1}^{N} Y_i$ is essentially the problem of predicting the sum of unobserved random variables $\sum_{\tilde{s}} Y_i$.

In order to predict $\sum_{\tilde{s}} Y_i$, we obtained $\hat{\sigma}(x_i)$, which is an estimate of a reasonable assumption of variance structure of the observed random variables. We then estimated $\hat{m}(x_j)$, the estimated mean of the unobserved random variables using the weighted least squares method, and finally realized our proposed estimator. To do this, we proceeded as follows:

We rewrite our model in equation (13) as

$$Y_i = m(x_i) + e_i \tag{16}$$

where $\mathrm{E}(e_i) = 0$, var $(e_i) = \sigma(x_i)$ and cov $(e_i, e_j) = 0, i \neq j$.

The objective was to find the optimal estimator of $m(x_i)$. Since this is a heteroscedastic model, we require to apply the weighted least squares method. In this method, we first of all transform the heteroscadastic model into a homoscedastic model then apply the ordinary least squares method to the obtained homoscedastic model to obtain the estimates.

Applying this to the model in equation (16), we get

$$\frac{Y_i}{\sqrt{\sigma(x_i)}} = \frac{m(x_i)}{\sqrt{\sigma(x_i)}} + \frac{e_i}{\sqrt{\sigma(x_i)}} \tag{17}$$

Equation (17) can be written as

$$Y_i^* = m^*(x_i) + e_i^* \tag{18}$$

where $Y_i^* = \dfrac{Y_i}{\sqrt{\sigma(x_i)}}$, $m^*(x_i) = \dfrac{m(x_i)}{\sqrt{\sigma(x_i)}}$ and $e_i^* = \dfrac{e_i}{\sqrt{\sigma(x_i)}}$

Clearly, $\mathrm{E}(e_i^*) = 0$ and

$$\mathrm{var}(e_i^*) = \mathrm{var}\left[\frac{e_i}{\sqrt{\sigma(x_i)}}\right] = \frac{\sigma(x_i)}{\sigma(x_i)} = 1$$

Hence the model in equation (18) is homoscedastic. From equation (18) we have

$$S = \sum_{i=1}^{n} e_i^{*2} = \sum_{i=1}^{n} \left(Y_i^* - m^*(x_i)\right)^2 \tag{19}$$

By replacing Y_i^* and $m^*(x_i)$ in equation (19), we get

$$S = \sum_{i=1}^{n}\left[\frac{Y_i}{\sqrt{\sigma(x_i)}} - \frac{m(x_i)}{\sqrt{\sigma(x_i)}}\right]^2 \tag{20}$$

which implies

$$S = \sum_{i=1}^{n}\frac{[Y_i - m(x_i)]^2}{\sigma(x_i)} \tag{21}$$

Clearly, the minimization of this function with respect to $m(.)$ is complicated by the fact that $\sigma(x_i)$ is unknown. Dorfman assumed that $\sigma(x_i)$ is some constant. We, however, consider the fact that in most cases $\sigma(x_i)$ is not a constant.

One approach is to estimate $\sigma(x_i)$. Intuitively, one would want to use the best estimator of $\sigma(x_i)$. Suppose we consider the simpler model

$$Y_i = \beta x_i + e_i \tag{22}$$

where $\mathrm{E}(e_i) = 0$, $\mathrm{var}(e_i) = \sigma^2 x_i$ and $\mathrm{cov}(e_i, e_j) = 0, i \neq j$. Then

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\frac{[Y_i - \beta x_i]^2}{\sigma^2 x_i} \tag{23}$$

Note that the e's are heteroscedastic.

Differentiating equation (23) with respect to β and equating to zero we have

$$\frac{ds}{d\beta} = -2\sum_{i=1}^{n}\frac{x_i[Y_i - \beta x_i]}{\sigma^2 x_i} = 0 \tag{24}$$

The estimator of β is given by

$$\hat{\beta} = \frac{\sum_s y_i}{\sum_s x_i} \quad \text{, or equivalently}$$

$$\hat{\beta} = \frac{\bar{y}}{\bar{x}} \qquad (25)$$

Now, let

$$r_i = Y_i - \hat{\beta}x_i$$

Then,

$$E(r_i) = E(Y_i - \hat{\beta}x_i)$$

$$= E(Y_i) - x_i E(\hat{\beta}) = \beta x_i - x_i E\left(\frac{\sum_s y_i}{\sum_s x_i}\right), \text{ since } E(y_i) = \beta x_i$$

$$= \beta x_i - x_i \frac{\beta \sum_s x_i}{\sum_s x_i} = 0 \qquad (26)$$

Also

$$\text{var}(r_i) = E(r_i^2) - [E(r_i)]^2$$

$$\Leftrightarrow \text{var}(r_i) = E(r_i^2), \text{ since } [E(r_i)]^2 = 0 \text{, from equation (26).}$$

$$r_i^2 = \left[Y_i - x_i \frac{\sum_s y_i}{\sum_s x_i}\right]^2, \text{ since } \hat{\beta} = \frac{\bar{y}}{\bar{x}} \text{, from equation (25)}$$

$$= Y_i^2 - 2x_iY_i\frac{\sum_s y_i}{\sum_s x_i} + x_i^2\frac{\left(\sum_s y_i\right)^2}{\left(\sum_s x_i\right)^2}$$

$$\mathrm{E}\left(r_i^2\right) = \mathrm{E}\left(Y_i^2\right) - 2x_i\mathrm{E}\left(Y_i\right)\frac{\sum_s y_i}{\sum_s x_i} + x_i^2\frac{\mathrm{E}\left(\sum_s y_i\right)^2}{\left(\sum_s x_i\right)^2}$$

$$= \mathrm{var}\left(Y_i\right) + \mathrm{E}\left(Y_i^2\right) - \frac{2x_i}{\sum_s x_i}\mathrm{E}\left(Y_i\right)\left[Y_i + \sum_{j\neq i} Y_j\right] + \frac{x_i^2}{\left(\sum_s x_i\right)^2}\left[\mathrm{var}\left(\sum_s y_i\right) + \mathrm{E}\left(\sum_s y_i\right)^2\right] \quad (27)$$

From equation (27) we have

$$\mathrm{var}\left(r_i\right) = \sigma\left(x_i\right) + \beta^2x_i^2 - \frac{2x_i}{\sum_s x_i}\left[\sigma\left(x_i\right) + \beta^2x_i^2 + \beta^2x_i\sum_{j\neq i}x_j\right] + \frac{x_i^2}{\left(\sum_s x_i\right)^2}\left[\sum_s\sigma\left(x_i\right) + \beta^2\left(\sum_s x_i\right)^2\right]$$

$$= \sigma\left(x_i\right) + \beta^2x_i^2 - \frac{2x_i}{\sum_s x_i}\left[\sigma\left(x_i\right) + \beta^2x_i\sum_s x_i\right] + \frac{x_i^2}{\left(\sum_s x_i\right)^2}\sum_s\sigma\left(x_i\right) + \beta^2x_i^2$$

$$= \sigma\left(x_i\right) - \frac{2x_i\sigma\left(x_i\right)}{\sum_s x_i} + \frac{x_i^2\sum_s\sigma\left(x_i\right)}{\left(\sum_s x_i\right)^2}$$

$$= \sigma\left(x_i\right) - \frac{2x_i\sigma\left(x_i\right)}{n\bar{x}} + \frac{x_i^2\sum_s\sigma\left(x_i\right)}{n^2\bar{x}^2} \quad (28)$$

From equation (28), as $n \to \infty$, $\frac{2x_i\sigma\left(x_i\right)}{n\bar{x}} \to 0$ and $\frac{x_i^2\sum_s\sigma\left(x_i\right)}{n^2\bar{x}^2} \to 0$, averages of sample values being finite.

hence $\operatorname{var}(r_i) \approx \sigma(x_i)$ *(29)*

This implies that we can therefore use this simple estimator of $\sigma(x_i)$ in the model in equation (21), so that, $\hat{\sigma}(x_i) = (Y_i - \hat{\beta}x_i)^2$

Hence we replace $\sigma(x_i)$ in the model given in equation (21) by $(Y_i - \hat{\beta}x_i)^2$ and derive the estimator of $m(x_j)$.

We are, however, averaging the nearby values of Y_i, where "nearby" is measured in terms of the distances $|x_i - x_j|$. Let $k(u)$ be a symmetric density function, for example, the standard normal density function. For a chosen scaling factor (bandwidth) b, define

$k_b(u) = b^{-1}k[u/b]$, and weights

$$w_i(x_j) = k_b(x_i - x_j) \Big/ \sum_{i=1}^{n} k_b(x_i - x_j)$$

From equation (21) we have

$$S = \sum_{i=1}^{n} \left\{ \frac{w_i(x_j)[Y_i - m(x_i)]^2}{(Y_i - \hat{\beta}x_i)^2} \right\} \tag{30}$$

We expand equation (30) by use of Taylor series expansion at a point x_j as follows

$$S = \sum_{s} \left\{ \frac{w_i(x_j)(Y_i - [m(x_j) + (x_i - x_j)m'(x_j) + \ldots])^2}{(Y_i - \hat{\beta}x_i)^2} \right\} \tag{31}$$

Equation (31) is approximately given by

$$S \approx \sum_{s} \left\{ \frac{w_i(x_j)[Y_i - m(x_j)]^2}{(Y_i - \hat{\beta}x_i)^2} \right\} \tag{32}$$

Differentiating equation (32) with respect to $m(x_j)$ we have

$$\frac{ds}{dm(x_j)} = \sum_{s} \left\{ \frac{(-2)w_i(x_j)[Y_i - m(x_j)]}{(Y_i - \hat{\beta}x_i)^2} \right\} \tag{33}$$

Equating equation (33) to zero and simplifying we get

$$\sum_{s} \left[\frac{w_i(x_j)Y_i}{(Y_i - \hat{\beta}x_i)^2} \right] = \sum_{s} \left[\frac{w_i(x_j)m(x_j)}{(Y_i - \hat{\beta}x_i)^2} \right] \tag{34}$$

From equation (34), the estimation of $m(x_j)$ is found to be

$$\hat{m}(x_j) = \frac{\sum_{s} \left[\frac{w_i(x_j)}{(Y_i - \hat{\beta}x_i)^2} \right] Y_i}{\sum_{s} \left[\frac{w_i(x_j)}{(Y_i - \hat{\beta}x_i)^2} \right]} \tag{35}$$

Our proposed estimator is given by

$$\hat{T}_{np(h)} = \sum_{s} Y_i + \sum_{\tilde{s}} \hat{m}(x_j) \tag{36}$$

where $\hat{m}(x_j)$ is defined in equation (35).

3.3 PROPERTIES OF THE PROPOSED ESTIMATOR

3.3.1 Conditional Bias

The conditional bias of the estimator given in equation (36) is given as follows

$$\mathrm{E}\left[\hat{T}_{np(h)} - T/\underset{-}{X}_{p}\right] = \mathrm{E}\left\{\sum_{\tilde{s}}\left[\hat{m}(x_j) - Y_j\right]\right\} \tag{37}$$

Replacing $\hat{m}(x_j)$ we have,

$$\mathrm{E}\left[\hat{T}_{np(h)} - T/\underset{-p}{X}\right] = \mathrm{E}\left[\sum_{\tilde{s}}\left\{\frac{\sum_{s}(nb)^{-1}\dfrac{k\left[\dfrac{x_i - x_j}{b}\right]Y_i}{\left(Y_i - \hat{\beta}x_i\right)^2}}{\sum_{s}(nb)^{-1}\dfrac{k\left[\dfrac{x_i - x_j}{b}\right]}{\left(Y_i - \hat{\beta}x_i\right)^2}} - Y_j\right\}\right] \tag{38}$$

Equation (38) can be rewritten as

$$\mathrm{E}\left[\hat{T}_{np(h)} - T/\underset{-p}{X}\right] = \sum_{\tilde{s}}\left\{\frac{\sum_{s}(nb)^{-1}k\left[\dfrac{x_i - x_j}{b}\right]\mathrm{E}\left[\dfrac{Y_i}{\left(Y_i - \hat{\beta}x_i\right)^2}\right]}{\sum_{s}(nb)^{-1}k\left[\dfrac{x_i - x_j}{b}\right]\mathrm{E}\left[\dfrac{1}{\left(Y_i - \hat{\beta}x_i\right)^2}\right]} - \mathrm{E}(Y_j)\right\} \tag{39}$$

From equation (39), to a first approximation, we get

$$\mathrm{E}\left[\frac{Y_i}{\left(Y_i - \hat{\beta}x_i\right)^2}\right] = \frac{\mathrm{E}(Y_i)}{\mathrm{E}\left(Y_i - \hat{\beta}x_i\right)^2}$$

In terms of moments,

$$\mathrm{E}(Y_i) = m_1(x_i) \tag{40}$$

and

$$\mathrm{E}\left(Y_i - \hat{\beta}x_i\right)^2 = \mathrm{E}\left(Y_i^2\right) - \frac{2x_i\mathrm{E}\left(Y_i\right)\sum_s y_i}{\sum_s x_i} + \frac{x_i^2\mathrm{E}\left(\sum_s y_i\right)^2}{\left(\sum_s x_i\right)^2} \qquad (41)$$

$$= \mathrm{E}\left(y_i^2\right) - \frac{2x_i\mathrm{E}\left(y_i\right)\left[y_i + \sum_{j\neq i} y_j\right]}{\sum_s x_i} + \frac{x_i^2\mathrm{E}\left[\sum_s y_i^2 + \sum_{i\neq}\sum_j y_i y_j\right]}{\left(\sum_s x_i\right)^2}$$

$$= \mathrm{E}\left(y_i^2\right) - \frac{2x_i\mathrm{E}\left[y_i^2 + y_i\sum_{j\neq i} y_j\right]}{\sum_s x_i} + \frac{x_i^2\mathrm{E}\left[\sum_s y_i^2 + \sum_{i\neq}\sum_j y_i y_j\right]}{\left(\sum_s x_i\right)^2}, \ y_i \text{ and } y_j \text{ are independent.}$$

$$= m_2\left(x_i\right) - \frac{2x_i}{\sum_i x_i}\left[m_2\left(x_i\right) + m_1\left(x_i\right)\sum_{j\neq i} m_1\left(x_j\right)\right] + \frac{x_i^2}{\left(\sum_i x_i\right)^2}\left[\sum_i m_2\left(x_i\right) + \sum_{i\neq}\sum_j m_1\left(x_i\right)m_1\left(x_j\right)\right] \qquad (42)$$

From equation (42) we note that,

$$\sum_{j\neq i} m_1\left(x_j\right) = \sum_i m_1\left(x_i\right) - m_1\left(x_i\right) \qquad (43)$$

$$\sum_{i\neq}\sum_j m_1\left(x_i\right)m_1\left(x_j\right) = \left[\sum_i m_1\left(x_i\right)\right]^2 - \sum_i m_1^2\left(x_i\right) \qquad (44)$$

Hence equation (42) becomes:

$$= m_2\left(x_i\right) - \frac{2x_i}{\sum_i x_i}\left[m_2\left(x_i\right) + m_1\left(x_i\right)\sum_i m_1\left(x_i\right) - m_1^2\left(x_i\right)\right] + \frac{x_i^2}{\left(\sum_i x_i\right)^2}\left[\sum_i m_2\left(x_i\right) + \left[\sum_i m_1\left(x_i\right)\right]^2 - \sum_i m_1^2\left(x_i\right)\right]$$

$$= m_2(x_i) - \frac{2x_i m_1(x_i)\sum_i m_1(x_i)}{\sum_i x_i} + \frac{x_i^2\left[\sum_i m_1(x_i)\right]^2}{\left(\sum_i x_i\right)^2}$$, since the terms of order n go to zero.

Hence simplifying weget

$$E\left(Y_i - \hat{\beta}x_i\right)^2 = \frac{\left(\sum_i x_i\right)^2 m_2(x_i) - 2x_i\left(\sum_i x_i\right)m_1(x_i)\sum_i m_1(x_i) + x_i^2\left[\sum_i m_1(x_i)\right]^2}{\left(\sum_i x_i\right)^2} \qquad (45)$$

Using equations (40) and (45) it implies that,

$$E\left[\frac{Y_i}{\left(Y_i - \hat{\beta}x_i\right)^2}\right] = \frac{\left(\sum_i x_i\right)^2 m_1(x_i)}{\left(\sum_i x_i\right)^2 m_2(x_i) - 2x_i\left(\sum_i x_i\right)m_1(x_i)\sum_i m_1(x_i) + x_i^2\left[\sum_i m_1(x_i)\right]^2} \qquad (46)$$

In the same spirit we have

$$E\left[\frac{1}{\left(Y_i - \hat{\beta}x_i\right)^2}\right] = \frac{E(1)}{E\left(Y_i - \hat{\beta}x_i\right)^2} = \frac{1}{E\left(Y_i - \hat{\beta}x_i\right)^2}$$

$$= \frac{\left(\sum_i x_i\right)^2}{\left(\sum_i x_i\right)^2 m_2(x_i) - 2x_i\left(\sum_i x_i\right)m_1(x_i)\sum_i m_1(x_i) + x_i^2\left[\sum_i m_1(x_i)\right]^2} \qquad (47)$$

Hence equation (39) is given as

$$E\left[\hat{T}_{np(h)} - T / \underset{-p}{X}\right] =$$

$$\sum_{\tilde{s}}\left\{\frac{\sum_i (nb)^{-1} k\left[\frac{x_i - x_j}{b}\right]\frac{\left(\sum_i x_i\right)^2 m_1(x_i)}{\left(\sum_i x_i\right)^2 m_2(x_i) - 2x_i\left(\sum_i x_i\right) m_1(x_i)\sum_i m_1(x_i) + x_i^2\left[\sum_i m_1(x_i)\right]^2}}{\sum_i (nb)^{-1} k\left[\frac{x_i - x_j}{b}\right]\frac{\left(\sum_i x_i\right)^2}{\left(\sum_i x_i\right)^2 m_2(x_i) - 2x_i\left(\sum_i x_i\right) m_1(x_i)\sum_i m_1(x_i) + x_i^2\left[\sum_i m_1(x_i)\right]^2}} - m_1(x_j)\right\} \tag{48}$$

From equation (48), let $\hat{c}_s(x_j)$

$$= \sum_i (nb)^{-1} k\left[\frac{x_i - x_j}{b}\right]\frac{\left(\sum_i x_i\right)^2}{\left(\sum_i x_i\right)^2 m_2(x_i) - 2x_i\left(\sum_i x_i\right) m_1(x_i)\sum_i m_1(x_i) + x_i^2\left[\sum_i m_1(x_i)\right]^2}$$

Then, $E\left[\hat{T}_{np(h)} - T / \underset{-p}{X}\right]$

$$= \sum_{\tilde{s}}\left\{\sum_i (nb)^{-1} k\left[\frac{x_i - x_j}{b}\right]\frac{\left(\sum_i x_i\right)^2}{\left(\sum_i x_i\right)^2 m_2(x_i) - 2x_i\left(\sum_i x_i\right) m_1(x_i)\sum_i m_1(x_i) + x_i^2\left[\sum_i m_1(x_i)\right]^2}\right.$$

$$\left.\left[\hat{c}_s(x_j)\right]^{-1} m_1(x_i) - m_1(x_j)\right\}$$

$$=\sum_{\tilde{s}}(nb)^{-1}\left[\hat{c}_s(x_j)\right]^{-1}\left\{\sum_i k\left[\frac{x_i-x_j}{b}\right]\frac{\left(\sum_i x_i\right)^2}{\left(\sum_i x_i\right)^2 m_2(x_i)-2x_i\left(\sum_i x_i\right)m_1(x_i)\sum_i m_1(x_i)+x_i^2\left[\sum_i m_1(x_i)\right]^2}\right.$$

$$\left. m_1(x_i)-(nb)\left[\hat{c}_s(x_j)\right]m_1(x_j)\right\}$$

$$=\sum_{\tilde{s}}(nb)^{-1}\left[\hat{c}_s(x_j)\right]^{-1}\left\{\sum_i k\left[\frac{x_i-x_j}{b}\right]\frac{\left(\sum_i x_i\right)^2 m_1(x_i)}{\left(\sum_i x_i\right)^2 m_2(x_i)-2x_i\left(\sum_i x_i\right)m_1(x_i)\sum_i m_1(x_i)+x_i^2\left[\sum_i m_1(x_i)\right]^2}\right.$$

$$\left.-(nb)(nb)^{-1}\sum_i k\left[\frac{x_i-x_j}{b}\right]\frac{\left(\sum_i x_i\right)^2 m_1(x_j)}{\left(\sum_i x_i\right)^2 m_2(x_i)-2x_i\left(\sum_i x_i\right)m_1(x_i)\sum_i m_1(x_i)+x_i^2\left[\sum_i m_1(x_i)\right]^2}\right\}$$

$$=\sum_{\tilde{s}}(nb)^{-1}\left[\hat{c}_s(x_j)\right]^{-1}\sum_i k\left[\frac{x_i-x_j}{b}\right]\frac{\left(\sum_i x_i\right)^2}{\left(\sum_i x_i\right)^2 m_2(x_i)-2x_i\left(\sum_i x_i\right)m_1(x_i)\sum_i m_1(x_i)+x_i^2\left[\sum_i m_1(x_i)\right]^2}$$

$$\left\{m_1(x_i)-m_1(x_j)\right\} \qquad (49)$$

3.3.2 Conditional variance

The conditional variance of the estimator given in equation (36) is given by

$$\operatorname{var}\left[\hat{T}_{np(h)} - T / \underline{X}_p\right] = \operatorname{var}\left\{\sum_{\tilde{s}}\left(\frac{(nb)^{-1} k\left[\frac{x_i - x_j}{b}\right]\frac{Y_i}{\left(Y_i - \hat{\beta}x_i\right)^2}}{(nb)^{-1} k\left[\frac{x_i - x_j}{b}\right]\frac{1}{\left(Y_i - \hat{\beta}x_i\right)^2}} - Y_j\right)\right\} \qquad (50)$$

Now, by equation (50)

$$\operatorname{var}\left(Y_j\right) = \mathrm{E}\left(Y_j^2\right) + \left[\mathrm{E}\left(Y_j\right)\right]^2 \qquad (51)$$

In terms of moments equation (51) becomes

$$= m_2\left(x_j\right) - m_1^2\left(x_j\right) \qquad (52)$$

Also

$$\operatorname{var}\left[\frac{Y_i}{\left(Y_i - \hat{\beta}x_i\right)^2}\right] = \mathrm{E}\left[\frac{Y_i}{\left(Y_i - \hat{\beta}x_i\right)^2}\right]^2 - \left\{\mathrm{E}\left[\frac{Y_i}{\left(Y_i - \hat{\beta}x_i\right)^2}\right]\right\}^2$$

$$= \mathrm{E}\left[\frac{Y_i^2}{\left(Y_i - \hat{\beta}x_i\right)^4}\right] - \left\{\mathrm{E}\left[\frac{Y_i}{\left(Y_i - \hat{\beta}x_i\right)^2}\right]\right\}^2 \qquad (53)$$

Now

$$\left(Y_i - \hat{\beta}x_i\right)^4 = Y_i^4 - 4Y_i^3\hat{\beta}x_i + 6Y_i^2\left(\hat{\beta}x_i\right)^2 - 4Y_i\left(\hat{\beta}x_i\right)^3 + \left(\hat{\beta}x_i\right)^4$$

and

$$\mathrm{E}\left(Y_i - \hat{\beta}x_i\right)^4 = \mathrm{E}\left\{Y_i^4 - \frac{4x_{iY_i^3}\sum_s y_i}{\sum_s x_i} + \frac{6x_i^2Y_i^2\left(\sum_s y_i\right)^2}{\left(\sum_s x_i\right)^2} - \frac{4x_i^3Y_i\left(\sum_s y_i\right)^3}{\left(\sum_s x_i\right)^4} + \frac{x_i^4\left(\sum_s y_i\right)^4}{\left(\sum_s x_i\right)^4}\right\} \qquad (54)$$

From equation (54), note that

$$\left(\sum_{s} y_i\right)^2 = \sum_{s} y_i^2 + \sum_{i\neq}\sum_{j} y_i y_j \tag{55}$$

$$\left(\sum_{s} y_i\right)^3 = \sum_{s} y_i \left[\sum_{s} y_j^2 + \sum_{j\neq}\sum_{k} y_j y_k\right]$$

$$= \sum_{s} y_i \sum_{s} y_j^2 + \sum_{s} y_i \sum_{j\neq}\sum_{k} y_j y_k$$

$$= \sum_{s} y_i^3 + \sum_{i\neq}\sum_{j} y_i y_j^2 + \sum_{j\neq}\sum_{k} y_j^2 y_k + \sum_{j\neq}\sum_{k} y_j y_k^2 + \sum_{i\neq}\sum_{j\neq}\sum_{k} y_i y_j y_k$$

$$= \sum_{s} y_i^3 + 3\sum_{i\neq}\sum_{j} y_i y_j^2 + \sum_{i\neq}\sum_{j\neq}\sum_{k} y_i y_j y_k \tag{56}$$

$$\left(\sum_{s} y_i\right)^4 = \left[\sum_{s} y_i^2 + \sum_{i\neq}\sum_{j} y_i y_j\right]\left[\sum_{k} y_k^2 + \sum_{k\neq}\sum_{l} y_k y_l\right]$$

$$= \sum_{i} y_i^2 \sum_{k} y_k^2 + \sum_{i} y_i^2 \sum_{k\neq}\sum_{l} y_k y_l + \sum_{k} y_k^2 \sum_{i\neq}\sum_{j} y_i y_j + \sum_{i\neq}\sum_{j} y_i y_j \sum_{k\neq}\sum_{l} y_k y_l$$

$$= \sum_{i} y_i^4 + \sum_{i\neq}\sum_{k} y_i^2 y_k^2 + \sum_{k\neq}\sum_{l} y_k^3 y_l + \sum_{k\neq}\sum_{l} y_k y_l^3 + \sum_{i\neq}\sum_{k\neq}\sum_{l} y_i^2 y_k y_l + \sum_{i\neq}\sum_{j} y_i^3 y_j + \sum_{i\neq}\sum_{j} y_i y_j^3$$

$$+ \sum_{k\neq}\sum_{i\neq}\sum_{j} y_k^2 y_i y_j + \sum_{j\neq}\sum_{k\neq}\sum_{l} y_j y_k^2 y_l + \sum_{j\neq}\sum_{k\neq}\sum_{l} y_j y_k y_l^2 + \sum_{i\neq}\sum_{k\neq}\sum_{l} y_i y_k^2 y_l + \sum_{i\neq}\sum_{k\neq}\sum_{l} y_i y_k y_l^2 + \sum_{i\neq}\sum_{j\neq}\sum_{k\neq}\sum_{l} y_i y_j y_k y_l$$

$$\tag{57}$$

Equation (57) can be re-written as

$$\left(\sum_{s} y_i\right)^4 = \sum_{i} y_i^4 + \sum_{i\neq}\sum_{k} y_i^2 y_k^2 + 4\sum_{k\neq}\sum_{l} y_k^3 y_l + 6\sum_{i\neq}\sum_{k\neq}\sum_{l} y_i^2 y_k y_l + \sum_{i\neq}\sum_{j\neq}\sum_{k\neq}\sum_{l} y_i y_j y_k y_l$$

Hence equation (54) will be given as

$$E\left(Y_i-\hat{\beta}x_i\right)^4=E\left(Y_i^4\right)-\frac{4x_iE\left(Y_i^3\right)}{\sum_s x_i}\left[y_i+\sum_{j\neq i}y_j\right]+\frac{6x_i^2E\left(Y_i^2\right)}{\left(\sum_s x_i\right)^2}\left[\sum_i y_i^2+\sum_{i\neq}\sum_j y_iy_j\right]$$

$$-\frac{4x_i^3E\left(Y_i\right)}{\left(\sum_s x_i\right)^3}\left[\sum_j y_j^3+3\sum_{i\neq}\sum_j y_iy_j^2+\sum_{i\neq}\sum_{j\neq}\sum_k y_iy_jy_k\right]$$

$$+\frac{x_i^4}{\left(\sum_s x_i\right)^4}E\left[\sum_k y_k^4+\sum_{i\neq}\sum_k y_i^2y_k^2+4\sum_{k\neq}\sum_l y_k^3y_l+6\sum_{i\neq}\sum_{k\neq}\sum_l y_i^2y_ky_l+\sum_{i\neq}\sum_{j\neq}\sum_{k\neq}\sum_l y_iy_jy_ky_l\right]$$

$$(58)$$

Next from equation (58) we note that

$$y_i^3\left[y_i+\sum_{J\neq i}y_j\right]=y_i^4+y_i^3\sum_{j\neq i}y_j \qquad (59)$$

$$y_i^2\left[\sum_i y_i^2+\sum_{i\neq}\sum_j y_iy_j\right]=y_i^2\sum_i y_j^2+y_i^2\sum_{j\neq}\sum_k y_jy_k$$

$$=\sum_i y_i^4+y_i^2\sum_{j\neq i}y_j^2+\sum_{j\neq}\sum_k y_j^3y_k+\sum_{j\neq}\sum_k y_jy_k^3+y_i^2\sum_{j\neq}\sum_k y_jy_k$$

$$=\sum_i y_i^4+y_i^2\sum_{j\neq i}y_j^2+2\sum_{j\neq}\sum_k y_j^3y_k+y_i^2\sum_{j\neq}\sum_k y_jy_k \qquad (60)$$

$$y_i\left[\sum_j y_j^3+3\sum_{i\neq}\sum_j y_iy_j^2+\sum_{i\neq}\sum_{j\neq}\sum_k y_iy_jy_k\right]$$

$$=y_i\sum_j y_j^3+3y_i\sum_{i\neq}\sum_j y_iy_j^2+y_i\sum_{i\neq}\sum_{j\neq}\sum_k y_iy_jy_k$$

$$=\sum_j y_j^4+y_i\sum_{i\neq j}y_j^3+3\sum_{j\neq}\sum_k y_j^2y_k^2+3\sum_{j\neq}\sum_k y_jy_k^3+3y_i\sum_{j\neq}\sum_k y_jy_k^2$$

$$+\sum_{j\neq}\sum_{k\neq}\sum_l y_j^2y_ky_l+\sum_{j\neq}\sum_{k\neq}\sum_l y_jy_k^2y_l+\sum_{j\neq}\sum_{k\neq}\sum_l y_jy_ky_l^2+y_i\sum_{j\neq}\sum_{k\neq}\sum_l y_jy_ky_l$$

$$
\begin{aligned}
&= \sum_{i} y_i^4 + y_i \sum_{j\neq i} y_j^3 + 3\sum_{j\neq}\sum_{k} y_j^2 y_k^2 + 3\sum_{j\neq}\sum_{k} y_j y_k^3 + 3y_i \sum_{j\neq}\sum_{k} y_j y_k^2 \\
&+3\sum_{j\neq}\sum_{k\neq}\sum_{l} y_i^2 y_k y_l + y_i \sum_{j\neq}\sum_{k\neq}\sum_{l} y_j y_k y_l
\end{aligned}
\qquad (61)
$$

This implies that equation (58) becomes;

$$
\mathrm{E}\left(Y_i - \hat{\beta}x_i\right)^4 = \mathrm{E}\left(y_i^4\right) - \frac{4x_i}{\sum_{s} x_i}\mathrm{E}\left[y_i^4 + y_i^3 \sum_{j\neq i} y_j\right]
$$

$$
+\frac{6x_i}{\left(\sum_{s} x_i\right)^2}\mathrm{E}\left[\sum_{i} y_i^4 + y_i^2 \sum_{j\neq i} y_j^2 + 2\sum_{j\neq}\sum_{k} y_j^3 y_k + y_i^2 \sum_{j\neq}\sum_{k} y_j y_k\right]
$$

$$
-\frac{4x_i^3}{\left(\sum_{s} x_i\right)^3}\mathrm{E}\left[\sum_{i} y_i^4 + y_i \sum_{j\neq i} y_j^3 + 3\sum_{j\neq}\sum_{k} y_j^2 y_k^2 + 3\sum_{j\neq}\sum_{k} y_j y_k^3 + 3y_i \sum_{j\neq}\sum_{k} y_j y_k^2\right]
$$

$$
+3\sum_{j\neq}\sum_{k\neq}\sum_{l} y_j^2 y_k y_l + y_i \sum_{j\neq}\sum_{k\neq}\sum_{l} y_j y_k y_l
$$

$$
+\frac{x_i^4}{\left(\sum_{s} x_i\right)^4}\mathrm{E}\left[\sum_{i} y_i^4 + \sum_{i\neq}\sum_{k} y_i^2 y_k^2 + 4\sum_{k\neq}\sum_{l} y_k^3 y_l + 6\sum_{i\neq}\sum_{k\neq}\sum_{l} y_i^2 y_k y_l + \sum_{i\neq}\sum_{j\neq}\sum_{k\neq}\sum_{l} y_i y_j y_k y_l\right]
$$

$$(62)$$

In terms of moments, equation (62) will be re-written as

$$
= m_4\left(x_i\right) - \frac{4x_i}{\sum_{i} x_i}\left[m_4\left(x_i\right) + m_3\left(x_i\right)\sum_{j\neq i} m_1\left(x_j\right)\right]
$$

$$
+\frac{6x_i}{\left(\sum_{i} x_i\right)^2}\left[\sum_{i} m_4\left(x_i\right) + m_2\left(x_i\right)\sum_{j\neq i} m_2\left(x_j\right) + 2\sum_{j\neq}\sum_{k} m_3\left(x_j\right)m_1\left(x_k\right) + m_2\left(x_i\right)\sum_{j\neq}\sum_{k} m_1\left(x_j\right)\right]
$$

$$-\frac{4x_i^3}{\left(\sum_i x_i\right)^3}\left[\sum_i m_4\left(x_i\right)+m_1\left(x_i\right)\sum_{j\neq i} m_3\left(x_j\right)+3\sum_{j\neq}\sum_{k} m_2\left(x_j\right)m_2\left(x_k\right)+3\sum_{j\neq}\sum_{k} m_1\left(x_j\right)m_3\left(x_k\right)\right]$$

$$+3m_1\left(x_i\right)\sum_{j\neq}\sum_{k} m_1\left(x_j\right)m_2\left(x_k\right)+3\sum_{j\neq}\sum_{k\neq}\sum_{l} m_2\left(x_j\right)m_1\left(x_k\right)m_1\left(x_l\right)+m_1\left(x_i\right)\sum_{j\neq}\sum_{k\neq}\sum_{l} m_1\left(x_j\right)m_1\left(x_k\right)m_1\left(x_l\right)$$

$$+\frac{x_i^4}{\left(\sum_i x_i\right)^4}\left[\sum_i m_4\left(x_i\right)+\sum_{i\neq}\sum_{k} m_2\left(x_i\right)m_2\left(x_k\right)+4\sum_{k\neq}\sum_{l} m_3\left(x_k\right)m_1\left(x_l\right)+6\sum_{i\neq}\sum_{k\neq}\sum_{l} m_2\left(x_i\right)m_1\left(x_k\right)m_1\left(x_l\right)\right]$$

$$+\sum_{i\neq}\sum_{j\neq}\sum_{k\neq}\sum_{l} m_1\left(x_i\right)m_1\left(x_j\right)m_1\left(x_k\right)m_1\left(x_l\right) \qquad (63)$$

From equation (63), we note the following

$$\sum_{j\neq i} m_2\left(x_j\right)$$

$$\begin{aligned}&\sum_i m_2\left(x_i\right)=m_2\left(x_i\right)+\sum_{j\neq i} m_2\left(x_j\right)\\ &\Leftrightarrow\sum_{j\neq i} m_2\left(x_j\right)=\sum_i m_2\left(x_i\right)-m_2\left(x_i\right)\end{aligned} \qquad (64)$$

$$\sum_{j\neq}\sum_{k} m_3\left(x_j\right)m_1\left(x_k\right)$$

$$\begin{aligned}&\sum_j m_3\left(x_j\right).\sum_k m_1\left(x_k\right)=\sum_k m_3\left(x_k\right)m_1\left(x_k\right)+\sum_{j\neq}\sum_{k} m_3\left(x_j\right)m_1\left(x_k\right)\\ &\Leftrightarrow\sum_{j\neq}\sum_{k} m_3\left(x_j\right)m_1\left(x_k\right)=\sum_j m_3\left(x_j\right)\sum_k m_1\left(x_k\right)-\sum_k m_3\left(x_k\right)m_1\left(x_k\right)\end{aligned} \qquad (65)$$

$$\sum_{j\neq}\sum_{k} m_1\left(x_j\right)m_1\left(x_k\right)$$

$$\begin{aligned}&\left[\sum_j m_1\left(x_j\right)\right]^2=\left[\sum_j m_1\left(x_j\right)\right]\left[\sum_j m_1\left(x_j\right)\right]=\sum_j m_1^2\left(x_j\right)+\sum_{j\neq}\sum_{k} m_1\left(x_j\right)m_1\left(x_k\right)\\ &\Leftrightarrow\sum_{j\neq}\sum_{k} m_1\left(x_j\right)m_1\left(x_k\right)=\left[\sum_j m_1\left(x_j\right)\right]^2-\sum_j m_1^2\left(x_j\right)\end{aligned} \qquad (66)$$

$$+\sum_{k\neq}\sum_{l} m_1(x_k)m_1(x_l)m_2(x_l)+\sum_{j\neq}\sum_{k\neq}\sum_{l} m_2(x_j)m_1(x_k)m_1(x_l)$$

This implies that

$$\sum_{j\neq}\sum_{k\neq}\sum_{l} m_2(x_j)m_1(x_k)m_1(x_l)=\sum_{j} m_2(x_j)\left[\sum_{k\neq}\sum_{l} m_1(x_k)m_1(x_l)\right]-2\sum_{k\neq}\sum_{l} m_1(x_k)m_2(x_k)m_1(x_l) \tag{70}$$

From equation (70) we have;

$$\begin{aligned}\left[\sum_{k} m_1(x_k)\right]^2 &= \left[\sum_{k} m_1(x_k)\right]\left[\sum_{k} m_1(x_k)\right]\\ &= \sum_{k} m_1^2(x_k)+\sum_{k\neq}\sum_{l} m_1(x_k)m(x_l)\\ &\Leftrightarrow \sum_{k\neq}\sum_{l} m_1(x_k)m_1(x_l)=\left[\sum_{k} m_1(x_k)\right]^2-\sum_{k} m_1^2(x_k)\end{aligned} \tag{71}$$

$$\sum_{k\neq}\sum_{l} m_1(x_k)m_2(x_k)m_1(x_l)$$

$$\begin{aligned}&\sum_{k} m_1(x_k)m_2(x_k).\sum_{l} m_1(x_l)\\ &= \sum_{k} m_1^2(x_k)m_2(x_k)+\sum_{k\neq}\sum_{l} m_1(x_k)m_2(x_k)m_1(x_l)\\ &\Leftrightarrow \sum_{k\neq}\sum_{l} m_1(x_k)m_2(x_k)m_1(x_l)=\sum_{k} m_1(x_k)m_2(x_k)\sum_{l} m_1(x_l)-\sum_{k} m_1^2(x_k)m_2(x_k)\end{aligned} \tag{72}$$

Hence equation (70) becomes

$$\sum_{j\neq}\sum_{k\neq}\sum_{l} m_2(x_j)m_1(x_k)m_1(x_l)=\sum_{j} m_2(x_j)\left[\left[\sum_{k} m_1(x_k)\right]^2-\sum_{k} m_1^2(x_k)\right]-2\sum_{k} m_1(x_k)m_2(x_k)\sum_{l} m_1(x_l)$$

$$+2\sum_{k} m_1^2(x_k)m_2(x_k)$$

$$=\sum_{j} m_2(x_j)\left[\sum_{k} m_1(x_k)\right]^2-\sum_{j} m_2(x_j)\sum_{k} m_1^2(x_k)-2\sum_{k} m_1(x_k)m_2(x_k)\sum_{l} m_1(x_l)+2\sum_{k} m_1^2(x_k)m_2(x_k)$$

$$\sum_{j\neq}\sum_{k\neq}\sum_{l} m_1(x_j)m_1(x_k)m_1(x_l)$$

$$\left[\sum_{j} m_1(x_j)\right]^3=\left[\sum_{j} m_1^2(x_j)+\sum_{j\neq}\sum_{k} m_1(x_j)m_1(x_k)\right]\sum_{i} m_1(x_l)$$

$$=\sum_{l} m_1(x_l)\sum_{j} m_1^2(x_j)+\sum_{l} m_1(x_l)\sum_{j\neq}\sum_{k} m_1(x_j)m_1(x_k)$$

$$=\sum_{j} m_1^3(x_j)+\sum_{j\neq}\sum_{l} m_1(x_l)m_1^2(x_j)+\sum_{j\neq}\sum_{k} m_1^2(x_j)m_1(x_k)+\sum_{j\neq}\sum_{k} m_1(x_j)m_1^2(x_k)$$

$$+\sum_{j\neq}\sum_{k\neq}\sum_{l} m_1(x_j)m_1(x_k)m_1(x_l)$$

$$=\sum_{j} m_1^3(x_j)+3\sum_{l\neq}\sum_{j} m_1(x_l)m_1^2(x_j)+\sum_{j\neq}\sum_{k\neq}\sum_{l} m_1(x_j)m_1(x_k)m_1(x_l)$$

$$\Leftrightarrow\sum_{j\neq}\sum_{k\neq}\sum_{l} m_1(x_j)m_1(x_k)m_1(x_l)=\left[\sum_{j} m_1(x_j)\right]-\sum_{j} m_1^3(x_j)-3\sum_{l\neq}\sum_{j} m_1(x_l)m_1^2(x_j) \quad (73)$$

From equation (73)

$$\sum_{j} m_1^2(x_j).\sum_{l} m_1(x_l)=\sum_{j} m_1^3(x_j)+\sum_{j\neq}\sum_{l} m_1^2(x_j)m_1(x_l)$$

$$\therefore\sum_{j\neq}\sum_{l} m_1^2(x_j)m_1(x_l)=\sum_{j} m_1^2(x_j)\sum_{l} m_1(x_l)-\sum_{j} m_1^3(x_j) \quad (74)$$

This implies that equation (73) becomes;

$$\sum_{j\neq}\sum_{k\neq}\sum_{l} m_1(x_j)m_1(x_k)m_1(x_l) = \left[\sum_j m_1(x_j)\right]^3 - \sum_j m_1^3(x_j) - 3\sum_j m_1^2(x_j)\sum_l m_1(x_l) + 3\sum_j m_1^3(x_j)$$

$$= \left[\sum_j m_1(x_j)\right]^3 + 2\sum_j m_1^3(x_j) - 3\sum_j m_1^2(x_j)\sum_l m_1(x_l)$$

$$\sum_{i\neq}\sum_{j\neq}\sum_{k\neq}\sum_{l} m_1(x_i)m_1(x_j)m_1(x_k)m_1(x_l)$$

$$\left[\sum_i m_1(x_i)\right]^4 = \left[\sum_i m_1^2(x_i) + \sum_{i\neq}\sum_{j} m_1(x_i)m_1(x_j)\right]\left[\sum_k m_1^2(x_k) + \sum_{k\neq}\sum_{l} m_1(x_k)m_1(x_l)\right]$$

$$= \sum_i m_1^2(x_i)\sum_k m_1^2(x_k) + \sum_i m_1^2(x_i)\sum_{k\neq}\sum_{l} m_1(x_k)m_1(x_l) + \sum_k m_1^2(x_k)\sum_{i\neq}\sum_{j} m_1(x_i)m_1(x_j)$$

$$+\sum_{i\neq}\sum_{j} m_1(x_i)m_1(x_j)\sum_{k\neq}\sum_{l} m_1(x_k)m_1(x_l)$$

$$= \sum_i m_1^4(x_i) + \sum_{i\neq}\sum_{k} m_1^2(x_i)m_1^2(x_k) + 4\sum_{k\neq}\sum_{l} m_1^3(x_k)m_1(x_l) + 6\sum_{i\neq}\sum_{k\neq}\sum_{l} m_1^2(x_i)m_1(x_k)m_1(x_l)$$

$$+\sum_{i\neq}\sum_{j\neq}\sum_{k\neq}\sum_{l} m_1(x_i)m_1(x_j)m_1(x_k)m_1(x_l)$$

$$\Leftrightarrow \sum_{i\neq}\sum_{j\neq}\sum_{k\neq}\sum_{l} m_1(x_i)m_1(x_j)m_1(x_k)m_1(x_l) = \left[\sum_i m_1(x_i)\right]^4 - \sum_i m_1^4(x_i) - \sum_{i\neq}\sum_{k} m_1^2(x_i)m_1^2(x_k)$$
$$-4\sum_{k\neq}\sum_{l} m_1^3(x_k)m_1(x_l) - 6\sum_{i\neq}\sum_{k\neq}\sum_{l} m_1^2(x_i)m_1(x_k)m_1(x_l) \qquad (75)$$

From equation (75)

$$\sum_i m_1^2(x_i)\sum_k m_1^2(x_k) = \sum_i m_1^4(x_i) + \sum_{i\neq}\sum_{k} m_1^2(x_i)m_1^2(x_k)$$

$$\Leftrightarrow \sum_{i\neq}\sum_{k} m_1^2(x_i)m_1^2(x_k) = \sum_i m_1^2(x_i)\sum_k m_1^2(x_k) - \sum_i m_1^4(x_i) \qquad (76)$$

$$\sum_{k} m_1^3(x_k)\sum_{l} m_1(x_l) = \sum_{k} m_1^3(x_k)m_1(x_k) + \sum_{k\neq}\sum_{l} m_1^3(x_k)m_1(x_l)$$

$$\Leftrightarrow \sum_{k\neq}\sum_{l} m_1(x_k)m_1(x_l) = \sum_{k} m_1^3(x_k)\sum_{l} m_1(x_l) - \sum_{k} m_1^4(x_k) \qquad (77)$$

$$\sum_{i\neq}\sum_{k\neq}\sum_{l} m_1^2(x_i)m_1(x_k)m_1(x_l)$$

$$\sum_{i} m_1^2(x_i)\left[\sum_{k\neq}\sum_{l} m_1(x_k)m_1(x_l)\right] = \sum_{k\neq}\sum_{l} m_1^3(x_k)m_1(x_l) + \sum_{k\neq}\sum_{l} m_1(x_k)m_1^3(x_l)$$

$$+\sum_{i\neq}\sum_{k\neq}\sum_{l} m_1^2(x_i)m_1(x_k)m_1(x_l)$$

$$\Leftrightarrow \sum_{i\neq}\sum_{k\neq}\sum_{l} m_1^2(x_i)m_1(x_k)m_1(x_l) = \sum_{i} m_1^2(x_i)\left[\sum_{k\neq}\sum_{l} m_1(x_k)m_1(x_l)\right] - 2\sum_{k\neq}\sum_{l} m_1^3(x_k)m_1(x_l)$$

$$= \sum_{i} m_1^2(x_i)\left\{\left[\sum_{k} m_1(x_k)\right]^2 - \sum_{k} m_1^2(x_k)\right\} - 2\left[\sum_{k} m_1^3(x_k)\sum_{l} m_1(x_l) - \sum_{k} m_1^4(x_k)\right]$$

$$\Leftrightarrow \sum_{i\neq}\sum_{k\neq}\sum_{l} m_1^2(x_i)m_1(x_k)m_1(x_l) = \sum_{i} m_1^2(x_i)\left[\sum_{k} m_1(x_k)\right]^2 - \sum_{i} m_1^2(x_i)\sum_{k} m_1^2(x_k) - 2\sum_{k} m_1^3(x_k)\sum_{l} m_1(x_l$$

$$-2\sum_{k} m_1^4(x_k) \qquad (78)$$

And therefore equation (75) becomes:

$$\sum_{i\neq}\sum_{j\neq}\sum_{k\neq}\sum_{l} m_1(x_i)m_1(x_j)m_1(x_k)m_1(x_l) = \left[\sum_{i} m_1(x_i)\right]^4 - \sum_{i} m_1^4(x_i) - \sum_{i} m_1^2(x_i)\sum_{k} m_1^2(x_k)$$

$$+\sum_{i} m_1^4(x_i) - 4\sum_{k} m_1^3(x_k)\sum_{l} m_1(x_l) + 4\sum_{k} m_1^4(x_k) - 6\sum_{i} m_1^2(x_i)\left[\sum_{k} m_1(x_k)\right]^2$$

$$+6\sum_{i} m_1^2(x_i)\sum_{k} m_1^2(x_k) + 12\sum_{k} m_1^3(x_k)\sum_{l} m_1(x_l) + 12\sum_{k} m_1^4(x_k)$$

$$=\left[\sum_i m_1(x_i)\right]^4-6\sum_i m_1^2(x_i)\left[\sum_k m_1(x_k)\right]^2+5\sum_i m_1^2(x_i)\sum_k m_1^2(x_k)+8\sum_k m_1^3(x_k)\sum_l m_1(x_l)+16\sum_k m_1^4(x_k)$$

$$\therefore \mathrm{E}\left[Y_i-\hat{\beta}x_i\right]^4=m_4(x_i)-\frac{4x_i}{\sum_i x_i}\left[m_4(x_i)+m_3(x_i)\sum_i m_1(x_i)-m_3(x_i)m_1(x_i)-2\sum_k m_3(x_k)m_1(x_k)\right]$$

$$+\frac{6x_i}{\left(\sum_i x_i\right)^2}\Bigg[\sum_i m_4(x_i)+m_2(x_i)\sum_i m_2(x_i)-m_2^2(x_i)+2\sum_j m_3(x_j)\sum_k{}^{m_1(x_k)}+m_2(x_i)\left[\sum_j m_1(x_j)\right]^2$$

$$-m_2(x_i)\sum_j m_1^2(x_j)\Bigg]$$

$$-\frac{4x_i^3}{\left(\sum_i x_i\right)^3}\Bigg[\sum_i m_4(x_i)+m_1(x_i)\sum_i m_3(x_i)-m_1(x_i)m_3(x_i)+3\left[\sum_j m_2(x_j)\right]^2-3\sum_j m_2^2(x_j)$$

$$+3\sum_j m_3(x_j)\sum_k m_1(x_k)-3\sum_k m_3(x_k)m_1(x_k)+3m_1(x_k)\sum_j m_1(x_j)\sum_k m_2(x_k)-3m_1(x_i)\sum_k m_1(x_k)m_2(x_k)$$

$$+3\sum_j m_2(x_j)\left[\sum_k m_1(x_k)\right]^2-3\sum_j m_2(x_j)\sum_k m_1^2(x_k)-6\sum_k m_1(x_k)m_2(x_k)\sum_l m_1(x_l)+6\sum_k m_1^2(x_k)m_2(x_k)$$

$$+m_1(x_i)\left[\sum_j m_1(x_j)\right]^3+2m_1(x_i)\sum_j m_1^3(x_j)-3m_1(x_i)\sum_j m_1^2(x_j)\sum_l m_1(x_l)\Bigg]$$

$$+\frac{x_i^2}{\left(\sum_i x_i\right)^4}\Bigg[\sum_i m_4(x_i)+\left[\sum_i m_2(x_i)\right]^2-\sum_i m_2^2(x_i)+4\sum_k m_3(x_k)\sum_l m_1(x_l)-4\sum_l m_3(x_l)m_1(x_l)$$

$$+6\sum_j m_2(x_j)\left[\sum_k m_1(x_k)\right]^2 - 6\sum_j m_2(x_j)\sum_k m_1^2(x_k) - 12\sum_k m_1(x_k)m_1(x_k)\sum_l m_1(x_l) + 12\sum_k m_1^2(x_k)m_2(x_k)$$

$$+\left[\sum_i m_1(x_i)\right]^4 - 6\sum_i m_1^2(x_i)\left[\sum_k m_1(x_k)\right]^2 + 5\sum_i m_1^2(x_i)\sum_k m_1^2(x_k) + 8\sum_k m_1^3(x_k)\sum_l m_1(x_l) + 16\sum_k m_1^4(x_k)\Big]$$

$$= m_4(x_i) - \frac{4x_i}{\sum_i x_i} m_4(x_i) - \frac{4x_i}{\sum_i x_i} m_3(x_i)\sum_i m_1(x_i) + \frac{4x_i}{\sum_i x_i} m_3(x_i)m_1(x_i) + \frac{6x_i^2}{\left(\sum_i x_i\right)^2}\sum_i m_4(x_i)$$

$$+\frac{6x_i^2}{\left(\sum_i x_i\right)^2} m_2(x_i)\sum_i m_2(x_i) - \frac{6x_i^2}{\left(\sum_i x_i\right)^2} m_2^2(x_i) + \frac{12x_i^2}{\left(\sum_i x_i\right)^2}\sum_j m_3(x_j)\sum_k m_1(x_k) - \frac{12x_i^2}{\left(\sum_i x_i\right)^2}\sum_k m_3(x_k)m_1(x_k)$$

$$+\frac{6x_i^2}{\left(\sum_i x_i\right)^2} m_2(x_i)\left[\sum_j m_1(x_j)\right]^2 - \frac{6x_i^2}{\left(\sum_i x_i\right)^2} m_2(x_i)\sum_j m_1^2(x_j) - \frac{4x_i^3}{\left(\sum_i x_i\right)^3}\sum_i m_4(x_i) - \frac{4x_i^3}{\left(\sum_i x_i\right)^3} m_1(x_i)\sum_i m_3($$

$$+\frac{4x_i^3}{\left(\sum_i x_i\right)^3} m_1(x_i)m_3(x_i) - \frac{12x_i^3}{\left(\sum_i x_i\right)^3}\left[\sum_j m_2(x_j)\right]^2 + \frac{12x_i^3}{\left(\sum_i x_i\right)^3}\sum_j m_2^2(x_j) - \frac{12x_i^3}{\left(\sum_i x_i\right)^3}\sum_j m_3(x_j)\sum_k m_1(x_k)$$

$$+\frac{12x_i^3}{\left(\sum_i x_i\right)^3}\sum_k m_3(x_k)m_1(x_k) - \frac{12x_i^3}{\left(\sum_i x_i\right)^3} m_1(x_i)\sum_j m_1(x_j)\sum_k m_2(x_k) + \frac{12x_i^3}{\left(\sum_i x_i\right)^3} m_1(x_i)\sum_k m_1(x_k)m_2(x_k)$$

$$-\frac{12x_i^3}{\left(\sum_i x_i\right)^3}\sum_j m_2(x_j)\left[\sum_k m_1(x_k)\right]^2 + \frac{12x_i^3}{\left(\sum_i x_i\right)^3}\sum_j m_2(x_j)\sum_k m_1^2(x_k) + \frac{24x_i^3}{\left(\sum_i x_i\right)^3}\sum_k m_1(x_k)m_2(x_k)\sum_l m_1(x_l)$$

$$-\frac{24x_i^3}{\left(\sum_i x_i\right)^3}\sum_k m_1^2(x_k)m_2(x_k) - \frac{4x_i^3}{\left(\sum_i x_i\right)^3} m_1(x_i)\left[\sum_j m_1(x_j)\right]^3 - \frac{8x_i^3}{\left(\sum_i x_i\right)^3} m_1(x_i)\sum_j m_1^3(x_j)$$

$$+\frac{12x_i^3}{\left(\sum_i x_i\right)^3} m_1(x_i)\sum_j m_1^2(x_j)\sum_l m_1(x_l) + \frac{x_i^4}{\left(\sum_i x_i\right)^4}\sum_i m_4(x_i) + \frac{x_i^4}{\left(\sum_i x_i\right)^4}\left[\sum_i m_2(x_i)\right]^2 - \frac{x_i^4}{\left(\sum_i x_i\right)^4}\sum_i m_2^2(x_i)$$

$$+\frac{4x_i^4}{\left(\sum_i x_i\right)^4}\sum_k m_3(x_k)\sum_l m_1(x_l) - \frac{4x_i^4}{\left(\sum_i x_i\right)^4}\sum_l m_3(x_l)m_1(x_l) + \frac{6x_i^4}{\left(\sum_i x_i\right)^4}\sum_j m_2(x_j)\left[\sum_k m_1(x_k)\right]^2$$

$$-\frac{6x_i^4}{\left(\sum_i x_i\right)^4}\sum_j m_2\left(x_j\right)\sum_k m_1^2\left(x_k\right)-\frac{12x_i^4}{\left(\sum_i x_i\right)^4}\sum_k m_1\left(x_k\right)m_2\left(x_k\right)\sum_l m_1\left(x_l\right)+\frac{12x_i^4}{\left(\sum_i x_i\right)^4}\sum_k m_1^2\left(x_k\right)m_2\left(x_k\right)$$

$$+\frac{x_i^4}{\left(\sum_i x_i\right)^4}\left[\sum_i m_1\left(x_i\right)\right]^4-\frac{6x_i^4}{\left(\sum_i x_i\right)^4}\sum_i m_1^2\left(x_i\right)\left[\sum_k m_1\left(x_k\right)\right]^2+\frac{5x_i^4}{\left(\sum_i x_i\right)^4}\sum_i m_1^2\left(x_i\right)\sum_k m_1^2\left(x_k\right)$$

$$+\frac{8x_i^4}{\left(\sum_i x_i\right)^4}\sum_k m_1^3\left(x_k\right)\sum_l m_1\left(x_l\right)+\frac{16x_i^4}{\left(\sum_i x_i\right)^4}\sum_k m_1^4\left(x_k\right)$$

$$=m_4\left(x_i\right)-\frac{4x_i}{\sum_i x_i}m_3\left(x_i\right)\sum_i m_1\left(x_i\right)+\frac{12x_i^2}{\left(\sum_i x_i\right)^2}\sum_j m_3\left(x_j\right)\sum_k m_1\left(x_k\right)+\frac{6x_i^2}{\left(\sum_i x_i\right)^2}m_2\left(x_i\right)\left[\sum_j m_1\left(x_j\right)\right]^2$$

$$-\frac{12x_i^3}{\left(\sum_i x_i\right)^3}\sum_j m_2\left(x_j\right)\left[\sum_k m_1\left(x_k\right)\right]^2-\frac{4x_i^3}{\left(\sum_i x_i\right)^3}m_1\left(x_{i.}\right)\left[\sum_j m_1\left(x_j\right)\right]^3+\frac{x_i^4}{\left(\sum_i x_i\right)^4}\left[\sum_i m_1\left(x_i\right)\right]^4 \tag{79}$$

since the terms of order n go to zero.

And

$$\left[\mathrm{E}\left(Y_i-\hat{\beta}x_i\right)^2\right]^2=m_2\left(x_i\right)\left[m_2\left(x_i\right)-\frac{2x_i}{\sum_i x_i}m_1\left(x_i\right)\sum_i m_1\left(x_i\right)+\frac{x_i^2}{\left(\sum_i x_i\right)^2}\left[\sum_i m_1\left(x_i\right)\right]^2\right]$$

$$-\frac{2x_i}{\sum_i x_i}m_1\left(x_i\right)\sum_i m_1\left(x_i\right)\left[m_2\left(x_i\right)-\frac{2x_i}{\sum_i x_i}m_1\left(x_i\right)\sum_i m_1\left(x_i\right)+\frac{x_i^2}{\left(\sum_i x_i\right)^2}\left[\sum_i m_1\left(x_i\right)\right]^2\right]$$

$$+\frac{x_i^2}{\left(\sum_i x_i\right)^2}\left[\sum_i m_1\left(x_i\right)\right]^2\left[m_2\left(x_i\right)-\frac{2x_i}{\sum_i x_i}m_1\left(x_i\right)\sum_i m_1\left(x_i\right)+\frac{x_i^2}{\left(\sum_i x_i\right)^2}\left[m_1\left(x_i\right)\right]^2\right] \tag{80}$$

In terms of moments, equation (80) will be

$$=m_2^2\left(x_i\right)-\frac{4x_i}{\sum_i x_i}m_1\left(x_i\right)m_2\left(x_i\right)\sum_i m_1\left(x_i\right)+\frac{2x_i^2}{\left(\sum_i x_i\right)^2}m_2\left(x_i\right)\left[\sum_i m_1\left(x_i\right)\right]^2+\frac{4x_i^2}{\left(\sum_i x_i\right)^2}m_1^2\left(x_i\right)\left[\sum_i m_1\left(x_i\right)\right]^2$$

$$-\frac{4x_i^3}{\left(\sum_i x_i\right)^3}m_1\left(x_i\right)\left[\sum_i m_1\left(x_i\right)\right]^3+\frac{x_i^4}{\left(\sum_i x_i\right)^4}\left[\sum_i m_1\left(x_i\right)\right]^4 \qquad (81)$$

This implies that equation (53) will be

$$=\frac{m_2\left(x_i\right)\left(\sum_i x_i\right)}{\left(\sum_i x_i\right)^4 m_4\left(x_i\right)-4x_i\left(\sum_i x_i\right)^3 m_3\left(x_i\right)\sum_i m_1\left(x_i\right)+6x_i^2\left(\sum_i x_i\right)^2\left\{m_2\left(x_i\right)\left[\sum_j m_1\left(x_j\right)\right]^2 +2\sum_j m_3\left(x_j\right)\sum_k m_1\left(x_k\right)\right\}-4x_i^3\left(\sum_i x_i\right)\left\{m_1\left(x_i\right)\left[\sum_j m_1\left(x_j\right)\right]^3+3\sum_j m_2\left(x_j\right)\left[\sum_k m_1\left(x_k\right)\right]^2\right\} +x_i^4\left[\sum_i m_1\left(x_i\right)\right]^4}$$

$$-\frac{m_1^2\left(x_i\right)\left(\sum_i x_i\right)^4}{\left(\sum_i x_i\right)^4 m_2^2\left(x_i\right)-4x_i\left(\sum_i x_i\right)^3 m_1\left(x_i\right)m_2\left(x_i\right)\sum_i m_1\left(x_i\right)+2x_i^2\left(\sum_i x_i\right)^2\left\{m_2\left(x_i\right)\left[\sum_i m_1\left(x_i\right)\right]^2 +2m_1^2\left(x_i\right)\left[\sum_i m_1\left(x_i\right)\right]^2\right\}-4x_i^3\left(\sum_i x_i\right)m_1\left(x_i\right)\left[\sum_i m_1\left(x_i\right)\right]^3+x_i^4\left[\sum_i m_1\left(x_i\right)\right]^4}$$

$$\frac{m_2(x_i)\left(\sum_i x_i\right)^4}{a}-\frac{m_1^2(x_i)\left(\sum_i x_i\right)^4}{p} \qquad (82)$$

by letting

a to be equal to

$$\left(\sum_i x_i\right)^4 m_4(x_i) - 4x_i\left(\sum_i x_i\right)^3 m_3(x_i)\sum_i m_1(x_i) + 6x_i^2\left(\sum_i x_i\right)^2 \left\{m_2(x_i)\left[\sum_j m_1(x_j)\right]^2\right.$$

$$\left. +2\sum_j m_3(x_j)\sum_k m_1(x_k)\right\} - 4x_i^3\left(\sum_i x_i\right)\left\{m_1(x_i)\left[\sum_j m_1(x_j)\right]^3 + 3\sum_j m_2(x_j)\left[\sum_k m_1(x_k)\right]^2\right\}$$

$$+x_i^4\left[\sum_i m_1(x_i)\right]^4$$

And p to be equal to

$$\left(\sum_i x_i\right)^4 m_2^2(x_i) - 4x_i\left(\sum_i x_i\right)^3 m_1(x_i)m_2(x_i)\sum_i m_1(x_i) + 2x_i^2\left(\sum_i x_i\right)^2 \left\{m_2(x_i)\left[\sum_i m_1(x_i)\right]^2\right.$$

$$\left. +2m_1^2(x_i)\left[\sum_i m_1(x_i)\right]^2\right\} - 4x_i^3\left(\sum_i x_i\right)m_1(x_i)\left[\sum_i m_1(x_i)\right]^3 + x_i^4\left[\sum_i m_1(x_i)\right]^4$$

Equation (53) becomes;

$$\operatorname{var}\left[\frac{Y_i}{\left(Y_i - \hat{\beta}x_i\right)^2}\right] = \frac{\left(\sum_i x_i\right)^4\left[pm_2(x_i) - am_1^2(x_i)\right]}{ap} \tag{83}$$

Similarly,

$$\operatorname{var}\left[\frac{1}{\left(Y_i - \hat{\beta}x_i\right)^2}\right] = \frac{\left(\sum_i x_i\right)^4\left[p - a\right]}{ap} \tag{84}$$

And

$$\operatorname{var}\left(Y_j\right) = \mathrm{E}\left(Y_j^2\right) - \left[\mathrm{E}\left(Y_j\right)\right]^2$$

$$= m_2\left(x_j\right) - m_1^2\left(x_j\right) \tag{85}$$

We let (83), (84) and (85) be denoted by v_i, w_i and f_j respectively.

Equation (50), can be written as

$$\operatorname{var}\left[\hat{T}_{np(h)} - T \mid \underline{X}_p\right] = \sum_{\bar{s}} \left\{ \frac{\left(\sum_s (nb)^{-1} k\left[\frac{x_i - x_j}{b}\right]\right)^2 vi}{\left(\sum_s (nb)^{-1} k\left[\frac{x_i - x_j}{b}\right]\right)^2 w_i} + f_j \right\} \tag{86}$$

CHAPTER 4

EMPIRICAL WORK

4.1 INTRODUCTION

In this section, three simulation studies were performed. Two on artificially constructed populations consisting of homoscedastic and heteroscedastic data. The other on a more realistic population derived at the United States Bureau of Labour Statistics.

4.2 PROCEDURE

In the first population that consisted of homoscedastic data, 400 data points were generated according to the model

$$Y_i = ax_i^3 + bx_i^2 + cx_i + e_i \tag{87}$$

with the $e_i \sim N(0,\sigma^2), x_i \sim U[0,1]$ mutually independent and independent across i. We used values $a = 1, b = -1.5, c = \frac{2}{3}, and\sigma = 0.02$. Similarly, for the second population that consisted of heteroscedastic data, 400 data points were generated according to the model in equation (87) but this time with $e_i \approx N\left[0,\sigma^2\right]$, and $x_i, a, b, c,$ and σ defined as above. We used Genstat 8th edition statistical application package for generating the data points. Dorfman chose to use 150 samples of size 60 each drawn by simple random sampling, and so, we choose to do the same for each of these populations. He too showed that a bandwidth of 0.09 gave a curve which seemed to reflect the underlying structure of the given population. As a result, we put in use the suggested bandwidth. For each sample, we calculated our proposed nonparametric regression based estimator and Dorfman's estimator. The estimates were calculated using a standard normal kernel.

We also considered a population consisting of $N = 400$ occupations. The data was taken from the United States Bureau of Labor Statistics' May 2005 National Occupational Employment and Wage estimates. The variable of interest Y is the total number of workers in each occupation; x is the total wages paid to workers in the selected group of occupations. From this population, 150 samples of size 60 each were taken using simple random sampling and our proposed and Dorfman's estimators calculated for each sample, using a standard normal kernel.

In addition, we calculated the variance of the two estimators: error variance due to Dorfman given in equation (15), and, the error variance due to the proposed estimator given in equation (86), using the three data sets to find out which one has minimum variance.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 RESULTS

Table 1 and *2* below give summary results in the form of the average relative error (ARE)

$$\sum_{r=1}^{150} T^{-1}\left(\hat{T}_r - T\right)\Big/150 \text{ and the root average squared error (RASE), } \left\{\sum_{r=1}^{150}\left(\hat{T}_r - T\right)^2 \Big/150\right\}^{1/2}$$

respectively where $\hat{T}_r$ is one of the estimators of T computed for sample r.

Table 1 (ARE)

ESTIMATOR	HOMOSCEDASTIC DATA	HETEROSCEDASTIC DATA	WAGE DATA
Our proposed estimator	**-0.11768**	**0.101641**	**6.37527E-07**
Dorfman's estimator	**-0.24842**	**0.267556**	**1.27524E-06**

Table 2 (RASE)

ESTIMATOR	HOMOSCEDASTIC DATA	HETEROSCEDASTIC DATA	WAGE DATA
Our proposed estimator	**0.39568**	**0.30508**	**5**
Dorfman's estimator	**0.83526**	**0.80308**	**10**

Table 3 below give results of the error variances

Table 3 (ERROR VARIANCE)

ESTIMATOR	HOMOSCEDASTIC DATA	HETEROSCEDASTIC DATA	WAGE DATA
Our proposed estimator	**0.00487**	**0.00204**	**91888646416**
Dorfman's estimator	**0.00532**	**0.00312**	**92456822341**

5.2 DISCUSSION

As can be noted from tables *1* and *2* above, our proposed non-parametric regression based estimator is much more efficient, since it has lower ARE and RASE as compared to Dorfman's estimator in each of the given data sets.

From table *3,* the proposed estimator has a lower error variance compared to Dorfman's estimator. This too illustrates that the former is better than the latter.

CONLUSION AND RECOMMENDATIONS

The objectives of this research were; to find a non-parametric regression based estimator of the population total that takes into account heteroscedasticity, to determine the properties of the attained estimator and to assess the performance of the estimator as compared to other existing estimators in an empirical study using both secondary and simulated data. We have derived a new estimator of a finite population total when we take into consideration the fact that variance varies in a given set of population and determined the properties of the attained estimator. Furthermore an empirical study was carried out and the results suggest that the nonparametric regression based estimator of a finite population total taking into account heteroscedasticity is a better improvement on Dorfman's estimator. It is likely to reflect better the actual structure of the data, yielding greater efficiency.

In this study we did not establish the Mean Square Error for the proposed estimator. We therefore suggest further research on this. Further research can also be done on the comparison of these research results with the ones under a restrictive model.

REFERENCES

[1]. Breidt, F. J. and Opsomer, J. D. (2000): Local Polynomial Regression Estimators in Survey Sampling. Annals of Statistics, **28**: 1026-1055.

[2]. Brewer, K. R. W. (1995): *Combining design based and model based inference. Chapter 30 in Business Survey Methods (editors: Cox, Binder, Chinnappa, Christianson, Colledge and Kott)*. New York: John Wiley

[3]. Carroll, R. J. (1982): Adapting for Heteroscedasticity in linear models. Annuals of Statistics, **10**: 1224-1233.

[4]. Chambers, R. L. (1996): Robust case-weighting for multipurpose establishment surveys. Journal of Official Statistics, **12**: 3-32.

[5]. Chambers, R. L. (2003): Which Sample Survey Strategy? A Review of Three Different Approaches. Southampton Statistical Sciences Research Institute, University of Southampton.

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

[7]. Cochran, W. G. (1977): *Sampling Techniques*, 3rd Edition. New York: John Wiley.

[8]. Dette, H., Munk, A. and Wagner, T. (1998): Estimating the variance in nonparametric regression-what is a reasonable choice? Journal of the Royal Statistical Society, **B, 60**: 751-764.

[9]. Dorfman, A. H. (1992): Non-parametric regression for estimating totals in finite populations. Proceedings of the Section on Survey Research Methods.

American Statistical Association, 47-52, 622-625.

[10]. Dorfman, A. H. and Hall, P. (1992): Estimators of the finite population distribution function using nonparametric regression. Annals of Statistics, **21**:1452-1475.

[11]. Fan, J. (1992): Design-adaptive nonparametric regression. Journal of the American Statistical Association, **87**: 998-1004.

[12]. Fan, J. and Gibjels, I. (1996): *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

[13]. Firth, D. and Bennett, K. E. (1998): Robust Models in Probability Sampling. Journal of the Royal Statistical Society, **B, 60**: 3-21.

[14]. Hardle, W. (1990): *Applied Nonparametric Regression Analysis*, Cambridge: Cambridge University Press.

[15]. Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953): *Sampling Survey Methods and Theory*, volumes I and II, New York: John Wiley.

[16]. Kim, J. –Y. (2004): Nonparametric Regression Estimation in Survey Sampling. Ph. D. thesis, Lowa State University.

[17]. Kim, J. –Y., Breidt, F. J. and Opsomer, J. D. (2000): Nonparametric Regression Estimation of Finite Population Totals under Two-Stage Sampling.

[18]. Kish, L. (1965): *Survey Sampling*. New York: John Wiley.

[19]. Kott, P. S. (1990a): Estimating the conditional variance of a design

consistent regression estimator. Journal of Statistical Planning and Inference, **24**: 287-296.

[20]. Kott, S. P. (2002): Randomization- Assisted Model- Based Survey Sampling. A paper prepared for the Fourth Biennial International Conference on Statistics, Probability and Related Areas. Dekalb, Illinois, 1-25.

[21]. Kuo, L. (1998): Classical and prediction approaches to estimating distribution function from survey data. Proceedings of the Section on Survey Research Methods, 280-285. American Statistical Association, Alexandria, VA.

[22]. Little, R. J. A. (1983b): Estmating a finite population mean from unequal probability samples. Journal of the American Statistical Association, **78**: 596-604.

[23]. M *ü* ller, H. G. and Stadtm *ü* ller, U. (1987): Estimation of heteroscedasticity in regression Analysis. Annuals of Statistics, **15**: 610-625.

[24]. Nadaraya, E. A. (1964): On estimating regression. Theory of Prob. And Applic. **9**: 141-142.

[25]. Neyman, J. (1934): On the two different aspects of the representative method. The method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society, **97**: 558-625.

[26]. Royall, R. M. (1976): Current advances in sampling theory: Implications for human observational studies. American Journal of Epidemiology, **104**: 463-474.

[27]. Royall, R. M., and Cumberland, W. G. (1981): An empirical study of the

ratio estimator and estimators of its variance. Journal of American Statistical Association, **76**: 66-77.

[28]. Ruppert, D., Wand, M. P., Hollst, V. and H*ö*ssjer, O. (1997): Local polynomial variance function estimation. Technometrics, **99**:262-273.

[29]. S*ä*rndal, C. –E., Swensson, B. and Wretman, J. H. (1989): The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total. Biometrika, **76**: 527-537.

[30]. S*ä*rndal, C. –E., Swensson, B. and Wretman, J. H. (1992): *Model Assisted Survey Sampling*, Springer Verlag: New York.

[31]. Smith, T. M. F. (1976): The foundations of survey sampling: A review. Journal of the Royal Statistical Society A, **139**: 183-204.

[32]. Smith, T. M. F. (1984): Sample surveys: Present position and potential developments: Some personal views. Journal of the Royal Statistical Society A, **147**: 208-221.

[33]. Smith, T. M. F. (1994): Sample surveys 1975-1990; An age of reconciliation? International Statistical Review, **62**: 3-34.

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

[35]. Watson, G. S. (1964): Smooth regression analysis. Sankhya A, **26**: 359-372.