PAU

PAN AFRICAN UNIVERSITY

INSTITUTE FOR BASIC SCIENCES, TECHNOLOGY AND INNOVATION

A multiplicative bias reduction for nonparametric approach and the two sample problem in a sample survey

KEMTIM TAMBOUN STEPHANE

MS300-0005/16

A Thesis summited to Pan African University Institute of Science, Technology and Innovation in partial fulfillment of the requirement for the award of the degree of Master of Science in Mathematics (Statistic option) of the Pan African University.

DECLARATION

This thesis is my original work and has not been submitted to any other university for examination .

signature Date

Kemtim Tamboun Stephane

MS300-0005/16

This thesis has been summitted for examination with our approval as university supervisor.

Signature Date

Prof.Romanus Odhiambo Otieno

Jomo Kenyatta University of Agriculture and Technology

Signature Date

Dr. Thomas Mageto

Jomo Kenyatta University of Agriculture and Technology

DEDICATION

I would like to dedicate this work to the strongest and toughest woman i know, my mum, Nganhou Micheline. Also to the almighty GOD because of His patience, guidance and unshakable faith in me.

ACKNOWLEDGEMENT

My first appreciation goes to my supervisors Professor Romanus Odhiambo Otieno and Doctor Thomas Mageto for guiding me through this work. Also, a lot of thanks to the African Union for putting this program in place, providing financial and educational means to African graduate to carry out scientific research with the main goal of developing Africa. Lastly but not the least, thanks to my family for their support.

List of Tables

List of Figures

ABSTRACT

This study carries on the problem of nonparametric estimation of finite population total using multiplicative bias correction technique for two sample problem in a sample survey is considered. Let two separate surveys collect related information on a single population U. Consider a situation where we want to best combine data from the two surveys to yield a single set of estimates of a population quantity ("population parameter") of interest. This thesis presents a multiplicative bias reduction estimator for nonparametric regression to two sample problem in a sample survey. The approach consists of applying a multiplicative bias correction to an estimator. The multiplicative bias correction method which was proposed, by Linton & Nielsen, 1994, assures a positive estimate and reduces the bias of the estimate with negligible increase in variance. Even as we apply this method to the two sample problem in a sample survey, we found out through the study of it asymptotic properties that it was asymptotically unbiased and statistically consistent. Furthermore, an empirical study was carried out to compare the performance of the developed estimator with the existing ones. The theoretical and empirical results led to the conclusion that the multiplicative bias corrected estimator can be highly recommended for two sample problem in sample survey when estimating the finite population total.

Contents

Chapter 1

INTRODUCTION

1.1 Background of Study

Sometimes, it happens that two separate surveys gather related information on a variable of interest of a population, U, having perhaps distinct designs and different mode of sampling. It becomes very important on how to combine the data from the two surveys.

Take as example, the students of the sub-regional institute of statistics and applied economics (ISSEA), and those of the polytechnic institute, both in different ways and with different interest collected data on unemployment in Cameroon. Researchers at the national institute of statistics(Cameroon) are faced with the following problem; how can the data from these two distinct surveys be joined together to produce a single set of data and have a better representation of the population.

Some researchers have been looking into these problems for several years. The approach to this problem has been in different ways; one of which involve getting estimates of the two surveys separately and using the inverse of the estimated variances as weights to weigh them together as seen in Merkouris (2004). Changbao (2004) went further by using empirical likelihood method to combine information from multiple survey. Another option to this consist of putting the two data sets in a single data set, taking into account the weight on individual sample units. Developed by Dorfman (2008) are some of these methods which include; the pseudo-likelihood, missing information principle and iterated post-stratified estimator. After simulations on two different population, it was concluded that, in neither population the design based ways of combining data yield good results. The iterated post-stratified estimator becomes a very promising non-parametric way to combined data from two sources.

Dorfman (2009) used the Nonparametric regression, which is the model-based sampler's method of choice when there is serious doubt about the suitability of a linear or other simple parametric model for the survey data at hand. The nonparametric regression supersedes the need for use of design weights and standard design-based weights. Recognition of this is especially helpful in confronting problems in sampling situations where design weights are missing or questionable.

This study uses kernel smoothers, especially the Nadaraya Watson smoother. However, estimators based on Nadaraya Watson smoothing weights are normally biased in small samples and at boundary points.

There exist alternative techniques of reducing the bias. For a detailed review see Marron and Härdle (1986), Bierens (1987), Muller and Stadtmuller (1987), Linton and Nielsen (1994) and Fan (1992). These methods improve the performance of nonparametric regression at points of large curvature. But in this framework, we consider a multiplicative bias correction approach to nonparametric regression to have an estimate with a smaller bias than existing ones.

1.2 Problem Statement

Two sample problems look at the best way on how to combine data from two, distinct, surveys to yield a single set of estimates of a population quantity of interest.

One such method is the nonparametric regression estimation method where the population total denoted by T can be written as the sum of the proportion truly observed, and proportion that is not observed, but which can be estimated using a corresponding auxiliary information. Many ways are used for doing the nonparametric regression estimation. The basic idea of all of them is that the auxiliary variable x provides some measure of the nearness of points so that we take as an estimate a weighted sum $\hat{m}(x_j) = \sum w_{ij} y_i$. Where w_{ij} depends on the distance of x_i to x_j. One particular version of this is Nadayara-Watson kernel estimation Dorfman (2009). Note that most kernel smoothers have boundary problems and require modifications at the boundary points. In particular, towards the boundary points, the bias of the estimators decreases at the cost of an increasing variance.

So, I am proposing a multiplicative bias reduction technique to reduce the bias of the kernel smoother. The objective of this method is to have a smaller bias than the existing methods with a negligible increase in variance

1.3 Justification of Study

Estimators for sample survey's two sample problem under nonparametric regression have been proven by Dorfman (2009) to be better, that is more efficient than those resulting from design-based.

This study made use of kernel smoothers, especially the Nadaraya Watson smoother.

However, estimators based on Nadaraya Watson smoothing weights are normally biased in small samples and at boundary points. Hence this study can be extended by considering a multiplicative bias reduction technique which is available in nonparametric regression literature. It is hypothesized that such a procedure will give a more bias robust estimator of the population total than the existing estimators.

1.4 Objectives of the Study

1.4.1 Main Objective

T0 develop a multiplicative bias reduction estimator for Nonparametric approach and the two sample problem in a sample survey.

1.4.2 Specific Objectives

1. To develop an estimator of finite population total based on multiplicative bias reduction technique.
2. To derive the asymptotic properties of the estimator.
3. To compare the performance of the proposed estimator to the existing ones.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, we review the different studies that have been made so far in relation to combining two or more samples in sample survey. The chapter highlights some very important contributions in the study of two sample problem, but also in the multiplicative bias reduction technique.

2.2 Different Approaches to two sample problem in sample survey

Recent years have seen a growing interest within statistical organizations in combining comparable information collected independently from multiple surveys of the same population or multiple samples in the same survey for the purpose of producing efficient estimator totals for common target.

In this regard, Renssen and Nieuwenbroek (1997) proposed adjusting the general regression estimator to meet consistency requirements (i.e weights of both surveys produce the same estimates for the unknown population totals of the common variable) by considering common variables as additional auxiliary variables. But this method had already been proposed by Zieschang (1990). The advantage of Renssen and Nieuwenbroek (1997) method over Zieschang is that, it is more general. It can be extended to more than two surveys unlike zieschang.

Merkouris (2004) further studies on this by investigating the composition of regression estimators within an extended framework of optimal regression with a focus on the efficiency of derived estimators and special emphasis on the practicality of competing composite regression methods.

Some methods used during the years were summarized in Dorfman (2009) as follows;

2.2.1 Design-based approaches

Mixture approach:

In this design-based approach, we take the estimates of each sample and weight them together by the inverse of their estimated variance.That is,

$$\hat{T}_{mix} = \hat{v}_1^{-1}\hat{T}_1 + \hat{v}_2^{-1}\hat{T}_2$$

Where $\hat{T}_k$ are estimators such as Horvitz-Thomson estimators and $\hat{v}_k$ are the estimated variances. Where $(k \in 1, 2)$

In this approach, we can not identify duplicates. We have a problem of overlap units in the samples.

Using overall Inclusion probabilities:

In this approach , the first step is to get the overall inclusion probability

$$\pi_i = \pi_{1i} + \pi_{2i} - \pi_{1i}\pi_{2i}$$

Then this can be incorporated into a horvitz-Thomson estimator, that is,

$$\hat{T}_\pi = \sum_{i \in s1 \cup s2} \pi_i^{-1} y_i$$

or in a Hajak estimator, that is,

$$\hat{T}_\pi = \frac{N \sum_{i \in s1 \cup s2} \pi_i^{-1} y_i}{\sum_{i \in s1 \cup s2} \pi_i^{-1}}$$

Using the overall inclusion probabilities has an advantage in that, the units occurring in both samples appear once in the joint sample but sampling probabilities can be very complex, too expensive and time consuming at times to deal with.

2.2.2 Model-base approach

Dorfman(2008) considered the use post-stratified weights to form an estimator of the form

$$\hat{T}_{ps} = \sum_h N_h n_h^{-1} \sum_{i \in h \cap (s1 \cup s2)} y_i$$

Where h are suitable chosen strata

N_h number of population units in h stratum and

n_h number of unique units in the combined sample that fall in the stratum h.

This estimator does not incorporate sampling probabilities

2.2.3 Nonparametric regression estimation

One motivation for Dorfman (2009) for using nonparametric regression estimators for two sample problem in sample survey is to avoid worrying about the complexities of inclusion probabilities. Many ways are available for doing nonparametric regression estimation. The basic idea of all is that the auxiliary variable x provides some measure of nearness of points, so that we take as an estimate a weighted sum.

Assume the following model

$$E(Y_i|X_i = x_i) = m(x_i)$$

$$cov(Y_i, Y_j|X_i = x_i, X_j = x_j) = \left\{ {}^{\sigma^2(x_i), i=j}_{0, i \neq j} \right.$$

Where $m(x_i)$ and $\sigma^2(x_i)$ are twice continuously differentiable functions that they are lipschitz functions. With these assumptions on $m(x_i)$ and $\sigma^2(x_i)$ then Dorfman (2009) considered estimating $m(x_i)$ and $\sigma^2(x_i)$ non-parametrically. He used kernel weights of the form

$$K(\frac{x_i - x_j}{h})$$

where K is a symmetric density function and h a chosen scaling factor(bandwidth). $m(x_i)$ is estimated using $\hat{m}(x_i) = \sum_{i \in s} w_{ij} y_i$. The simplest version of it is the Nadaraya-Watson weight where

$$w_{ij} = \frac{K(\frac{x_i - x_j}{h})}{\sum_{i \in s} K(\frac{x_i - x_j}{h})}$$

Overall, non parametric regression is a suitable way to handle the two sample problem. But the Nadaraya Watson weight used in the non parametric estimator faces problems of biasness. It is known that kernel based estimators also suffer from boundary problem resulting in significant biases. Procedures are needed that can minimize such bias.

2.3 Bias reduction in Nonparametric regression

All nonparametric smoothing methods are generally biased. There are many approaches to reducing the bias, but most of them do so at the cost of an increase in the variance of the estimator.

There exist a number of alternative bias reduction techniques for nonparametric regression, including; Bartlett's higher order kernels, the jacknife estimator of Marron and Härdle (1986), the double bandwidth estimator of Bierens (1987), the variable bandwidth scheme of Muller and Stadtmuller (1987), and the local polynomial regression estimator considered in Fan (1992). At points of large curvature, these methods can considerably improve the performance of non parametric regression estimators. The above procedures can be viewed as performing an additive bias reduction. The additive bias correction have the unfortunate side effect of possibly generating a non-positive estimate.

An attractive alternative to the linear bias correction is the multiplicative bias correction pioneered by Linton and Nielsen (1994), because the multiplicative correction does not alter the sign of the regression function. For more studies on multiplicative bias correction in density estimation, we refer the reader to Hirukawaa (2014) and references there in. In Hengartner et al. (2009), they studied the asymptotic properties of the resulting estimate from a multiplicative bias corrected nonparametric smoothers and prove that this estimator has zero asymptotic bias and the same asymptotic variance as the local linear estimate.

This study , wishes to make use of multiplicative bias reduction technique to develop a new bias robust estimator for a two sample survey problem due to the good asymptotic properties shown by this estimator in Hengartner et al. (2009).

Chapter 3

METHODOLOGY

3.1 INTRODUCTION

3.1.1 Overview of a Multiplicative bias reduction technique

Suppose that there is random sample of bivariate (x_i, y_i) for $i = 1, ..., n$ and the relationship between X and Y is represented by the model

$$Y_i = h(x) + \epsilon_i$$

Where $h(.)$ is a smooth function and the error ϵ_i satisfies the following

$$E(\epsilon_i) = 0,\ cov(\epsilon_i, \epsilon_j) = \{^{\sigma^2 \text{ for i=j}}_{0 \text{ otherwise}}$$

The principal objective of the multiplicative bias corrected technique is to correct the insufficiences of the kernel smoother that is the bias problem at the boundaries. Given a pilot smoother of the regression function

$$\tilde{h}(x) = \sum_{j=1}^{n} w_{xj} Y_j$$

The inverse relative estimation error of the smoother at each of the observations is given by $\frac{h(x)}{\tilde{h}(x)}$.

A noisy estimate, that is the unexplained variation or randomness that is found within a given data sample or formula, of this quantity,$\frac{h(x)}{\tilde{h}(x)}$, is given by te ratio

$$\beta(x) = \frac{Y_j}{\tilde{h}(X_j)}$$

Smoothing the noisy estimate $\beta(x)$ leads to

$$\tilde{\beta}(x) = \sum_{j=1}^{n} w_{xj} \beta(x)$$

$\tilde{\beta}(x)$ gives a better estimate for the inverse of the relative estimation error at each particular observation and can therefore be used as a multiplicative correction of the pilot smoother.

$$\hat{h}(x) = \tilde{\beta}(x)\tilde{h}(x) \tag{3.1}$$

For both $\tilde{h}(x)$, and $\tilde{\beta}(x)$, we use the same weighting scheme;

$$w_{xj} = \frac{1}{nh} K(\frac{x - X_j}{h})$$

where

h is the bandwidth

K is a probability density function, symmetric about zero.

n is the sample size

3.1.2 Bandwidth selection

Choosing a large bandwidth reduces the variance but simultaneously increases the bias of the estimate. Similarly a choice of a small bandwidth mitigates the bias but leads to an increase in the variance of the estimate. A natural way to mitigate this trade-of is to choose a bandwidth that minimizes the mean squared error of the estimate.

Various options will be considered for choosing a scaling factor that is a bandwidth.These options include;

(i) Implement biased cross-validation(bcv).

(ii) Implement unbiased cross-validation(ucv).

(iii) Implements a rule-of-thumb for choosing the bandwidth of a Gaussian kernel density estimator(ndr0).

(iv) Can use a more common variation given by Scott (1992)(ndr).

3.2 Proposed Estimator of Finite Population Total

Consider a finite population, $U = 1, 2, ..., N$ and let $y_1, y_2, ..., y_n$ represent the combined random sample drawn from the population using different sampling techniques. Suppose that to each of these $y_i's$, there is an auxiliary information $x_1, x_2, ..., x_n$.

Let consider the following model;

$$E(Y_i|X_i = x_i) = h(x_i)$$

$$cov(Y_i, Y_j|X_i = x_i, X_j = x_j) = \begin{cases} {\scriptstyle \sigma^2(x_i), i=j} \\ \\ {\scriptstyle 0, i \neq j} \end{cases}$$

Where $h(x_i)$ and $\sigma^2(x_i)$ are twice continuously differentiable functions (i.e lipschitz continuous). With these assumptions on $h(x_i)$ and $\sigma^2(x_i)$, one can estimate $h(x_i)$ and $\sigma^2(x_i)$ non-parametrically.

Let $\epsilon_i = Y_i - h(X_i)$ be i.i.d. with zero mean, and variance σ^2. We can refer to this set-up as the weak model. In this scheme, we can ignore which of the original samples, the $Y_i's$ are available from.

Usually in the computation of finite population total,we have the formula given by

$$T = \sum_{i \in U} y_i = \sum_{i \in s} \frac{1}{\pi_i} y_i + \sum_{j \in r} y_j \qquad (3.2)$$

Where, s refers to the sample and r refers to the nonsampled part of the population. since the values of the sample part is known, the process of estimating the finite population total is equivalent to predicting the nonsample part of the population.

A consistent prediction of the nonsample values is given by the smoother in equation (3.1). We define an estimate of the finite population ($\hat{T}_{MBC}$) total by

$$\hat{T}_{MBC} = \sum_{i \in s} \frac{y_i - \hat{h}(x_i)}{\pi_i} + \sum_{j \in r} \hat{h}(x_j) \tag{3.3}$$

Where

π_i is the inclusion probability

$\hat{h}(x_i)$ is the multiplicative bias corrected estimator.

3.3 Poperties of Proposed Estimator

3.3.1 Assumptions

The following assumptions are made in the estimation of $\hat{h}(x_i)$.

(i) The regression function is bounded and strictly positive, that is, $0 \leq a \leq h(x) \leq b$ for all x

(ii) The regression function is twice continuously differentiable everywhere.

(iii) ϵ has finite fourth moments and has a symmetric distribution around zero.

(iv) The bandwidth h is such that,$h \to 0$, and $nh \to \infty$, $(nh)^2 \to \infty$ as $n \to \infty$

3.3.2 Asymptotic Unbiasedness of the proposed Estimator

We want to show that $E(\hat{T}_{MBC} - T) \to 0$ as $n \to \infty$. Under the model based estimator, the bias of the estimator $\hat{T}_{MBC}$ is defined as follows;

$$E[\hat{T}_{MBC} - T] = E[\hat{T}_{MBC}] - E[T] \tag{3.4}$$

Now, we have the expected value of the proposed estimator for the finite population total given by;

$$\begin{aligned} E[\hat{T}_{MBC}] &= E[\sum_{i \in s} \frac{y_i - \hat{h}(x_i)}{\pi_i} + \sum_{j \in r} \hat{h}(x_j)] && (3.5) \\ &= E[\sum_{i \in s} \frac{y_i - \hat{h}(x_i)}{\pi_i}] + E[\sum_{j \in r} \hat{h}(x_j)] && (3.6) \\ &= \sum_{i \in s} \frac{1}{\pi_i} E(y_i - \hat{h}(x_i)) + \sum_{U|s} E(\hat{h}(x_j)) && (3.7) \end{aligned}$$

$E(\hat{h}(x_j))$ is obtained by analysing the individual terms of the stochastic approximation of $\hat{h}(x)$. Let us then establish the stochastic approximatiom of $\hat{h}(x)$ as shown by (Hengartner 2009).

From (3.1),

$$\hat{h}(x) = \tilde{\beta}(x)\tilde{h}(x) \tag{3.8}$$

$$= \sum_{j=1}^{n} w_{xj} \frac{Y_j}{\tilde{h}(X_j)} \tilde{h}(x) = \sum_{j=1}^{n} w_{xj} \frac{\tilde{h}(x)}{\tilde{h}(X_j)} Y_j \tag{3.9}$$

$$= \sum_{j=1}^{n} w_{xj} R_j(x) Y_j \text{ where } R_j(x) = \frac{\tilde{h}(x)}{\tilde{h}(X_j)} \tag{3.10}$$

Let define, $\bar{h} = E(\tilde{h}(x)|X_1, X_2, ..., X_n)$ then we can express $R_j(x)$ as

$$\begin{aligned}
R_j(x) &= \frac{\tilde{h}(x)}{\tilde{h}(X_j)} \\
&= (\frac{\bar{h}(x)}{\bar{h}(X_j)}) * (\frac{\tilde{h}(x)}{\bar{h}(x)}) * (\frac{\tilde{h}(X_j)}{\bar{h}(X_j)})^{-1} \\
&= (\frac{\bar{h}(x)}{\bar{h}(X_j)}) * (\frac{\tilde{h}(x) - \bar{h}(x) + \bar{h}(x)}{\bar{h}(x)}) * (\frac{\tilde{h}(X_j) - \bar{h}(X_j) + \bar{h}(X_j)}{\bar{h}(X_j)})^{-1} \\
&= (\frac{\bar{h}(x)}{\bar{h}(X_j)}) * (\frac{\tilde{h}(x) - \bar{h}(x)}{\bar{h}(x)} + 1) * (\frac{\tilde{h}(X_j) - \bar{h}(X_j)}{\bar{h}(X_j)} + 1)^{-1} \\
&= (\frac{\bar{h}(x)}{\bar{h}(X_j)}) * (R(x) + 1) * (R(X_j) + 1)^{-1}
\end{aligned}$$

Through the series expansion ,

$$\begin{aligned}
(R(X_j) + 1)^{-1} &= \frac{1}{R(X_j) + 1} = \frac{1}{1 - (-R(X_j))} = \sum_{n=0}^{\infty} [-R(X_j)]^n \\
&= 1 - R(X_j) + R(X_j)^2 + \dots \\
R_j(x) &= \frac{\bar{h}(x)}{\bar{h}(X_j)} * [1 + R(x) - R(X_j) + r_j(x, X_j)] \text{ is an approximation of the quantity R.}
\end{aligned}$$

Replacing both Y_j and R_j in (3.10), we obtain

$$\begin{aligned}\hat{h}(x) &= \sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}[1+R(x)-R(X_j)+r_j(x,X_j)](h(X_j)+\epsilon_j)\\ &= \sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}h(X_j)+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}(\epsilon_j+h(X_j)(R(x)-R(X_j))+\\ &\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}(R(x)-R(X_j))\epsilon_j+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}r_j(x,X_j)(h(X_j)+\epsilon_j)\end{aligned}$$

Using the assumption $nh \to \infty$ the remainder term turns to zero in probability and the expression reduces to;

$$\hat{h}(x) = \sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}h(X_j)+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}(\epsilon_j+h(X_j)(R(x)-R(X_j))+ \tag{3.11}$$

$$\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}(R(x)-R(X_j))\epsilon_j+o_p(\frac{1}{nh}) \tag{3.12}$$

To solve equation (3.7), we need to find $E(\hat{h}(x_j))$ hence,

$$E(\hat{h}(x_j)) = E[\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}h(X_j)+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}(\epsilon_j+h(X_j)(R(x)-R(X_j))+ \tag{3.13}$$

$$\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}(R(x)-R(X_j))\epsilon_j+o_p(\frac{1}{nh})] \tag{3.14}$$

$$= \sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}E(h(X_j))+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}E(\epsilon_j)+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}h(X_j)E(R(x)-R(X_j)) \tag{3.15}$$

$$+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}(R(x)-R(X_j))E(\epsilon_j)+o_p(\frac{1}{nh}) \tag{3.16}$$

$$= \sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}h(X_j)+0+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}h(X_j)E(\frac{\tilde{h}(x)}{\bar{h}(x)}-\frac{\tilde{h}(X_j)}{\bar{h}(X_j)})+0+o_p(\frac{1}{nh}) \tag{3.17}$$

since $E(\epsilon_j = 0)$

$$E(\hat{h}(x_j)) = \sum_{j=1}^{n} w_{xj} \frac{\bar{h}(x)}{\bar{h}(X_j)} h(X_j) + o_p(\frac{1}{nh}) \text{ since } \bar{h}(x) = E(\tilde{h}(x)) \tag{3.18}$$

Hence,

$$E[\hat{T}_{MBC}] = \sum_{i\in s} \frac{1}{\pi_i} E(y_i) - (\sum_{j=1}^{n} w_{xj} \frac{\bar{h}(x)}{\bar{h}(X_i)} h(X_i) + o_p(\frac{1}{nh})) + \sum_{U|s} (\sum_{j=1}^{n} w_{xj} \frac{\bar{h}(x)}{\bar{h}(X_j)} h(X_j)) + o_p(\frac{1}{nh}) \tag{3.19}$$

The above expression can be reduced by considering a limited Taylor series of $\dfrac{h(X_j)}{\bar{h}(X_j)}$ about a point x. Hence

$$\frac{h(X_j)}{\bar{h}(X_j)} = \frac{h(x)}{\bar{h}(x)} + (X_j - x)(\frac{h(x)}{\bar{h}(x)})' + (X_j - x)^2(\frac{h(x)}{\bar{h}(x)})'' + o_p(1)$$

Now, substituting the first two terms in (3.19) gives

$$E[\hat{T}_{MBC}] = \sum_{i\in s} \frac{1}{\pi_i} E(y_i) - E(\hat{h}(x_i)) + \sum_{U|s} (\sum_{j=1}^{n} w_{xj} \bar{h}(x) (\frac{h(x)}{\bar{h}(x)} + (X_j - x)(\frac{h(x)}{\bar{h}(x)})') + o_p(\frac{1}{nh})$$

But $\sum_{j=1}^{n} w_{xj} = 1$ and $\sum_{j=1}^{n} (X_j - x) w_{xj} = 0$, therefore

$$E[\hat{T}_{MBC}] = \sum_{i\in s} \frac{1}{\pi_i} E(y_i) + \sum_{U|s} \sum_{j=1}^{n} w_{xj} h(x) + o_p(\frac{1}{nh}) \tag{3.20}$$

Furthermore,

$$E(T) = \sum_{i \in s} \frac{1}{\pi_i} E(y_i) + \sum_{j \in r} E(y_j) \tag{3.21}$$

$$= \sum_{i \in s} \frac{1}{\pi_i} \bar{y} + \sum_{j \in r} h(x) \tag{3.22}$$

Hence the asymptotic bis of the estimator is given by

$$BIAS(\hat{T}_{MBC}) = E(\frac{\hat{T}_{MBC} - T}{N}) \tag{3.23}$$

$$= \frac{1}{N}(\sum_{U|s} \sum_{j=1}^{n} w_{xj} h(x) - \sum_{U|s} h(x) + o_p(\frac{1}{nh})) \tag{3.24}$$

The bias of $\hat{T}_{MBC}$ will be of order $o_p(\frac{1}{nh})$. Thus it converges to zero at a faster rate compared to the existing non-parametric estimators which generally converge at the rate $o_p(h^2)$.

3.3.3 Asymptotic Variance of the Proposed Estimator

The variance of the finite population total is given by;

$$Var[\hat{T}_{MBC}] = Var[\sum_{i \in s} \frac{y_i - \hat{h}(x_i)}{\pi_i} + \sum_{j \in r} \hat{h}(x_j)] \tag{3.25}$$

$$= Var[\sum_{i \in s} \frac{y_i - \hat{h}(x_i)}{\pi_i}] + Var[\sum_{j \in r} \hat{h}(x_j)] \tag{3.26}$$

$$= \sum_{i \in s} (\frac{1}{\pi_i})^2 Var(y_i - \hat{h}(x_i)) + \sum_{U|s} Var(\hat{h}(x_j)) \tag{3.27}$$

Firstly,

$$Var(\hat{h}(x_j)) = Var(\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}[1 + R(x) - R(X_j) + r_j(x, X_j)](h(X_j) + \epsilon_j)) \tag{3.28}$$

Using the assumption $nh \to \infty$, the remainder terms converge to zero in probability. Therefore $r_j(x, X_j)(h(X_j) + \epsilon_j) = o_p(\frac{1}{nh})$ and equation (3.28) reduces to

$$Var(\hat{h}(x_j)) = Var(\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}[1 + R(x) - R(X_j)](h(X_j) + \epsilon_j) + o_p(\frac{1}{nh})) \tag{3.29}$$

Truncating the binomial expansion at the first term yields

$$Var(\hat{h}(x_j)) = Var(\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}y_j) + o_p(\frac{1}{(nh)^2}) \tag{3.30}$$

$$= \sum_{j=1}^{n}(w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)})^2\sigma^2(x_j) + o_p(\frac{1}{(nh)^2}) \tag{3.31}$$

Simplify equation (3.31) expression by considering the first and second part of the Taylor series of $\frac{\sigma^2(x_j)}{\bar{h}^2(X_j)}$. So we obtain

$$Var(\hat{h}(x_j)) = \sum_{j=1}^{n}(w_{xj})^2\sigma^2(x_j) + o_p(\frac{1}{(nh)^2})$$

Therefore,

$$Var[\hat{T}_{MBC}] = \sum_{i\in s}(\frac{1}{\pi_i})^2\sigma^2(x_i) + \sum_{U}\sum_{j=1}^{n}(w_{xj})^2\sigma^2(x_j) + o_p(\frac{1}{(nh)^2}) \tag{3.32}$$

Thus the asymptotic variance is given by

$$Var(\frac{\hat{T}_{MBC}}{N}) = \frac{1}{N^2}\sum_{i\in s}(\frac{1}{\pi_i})^2\sigma^2(x_i) + \frac{1}{N^2}\sum_{U}\sum_{j=1}^{n}(w_{xj})^2\sigma^2(x_j) + o_p(\frac{1}{(nh)^2}) \tag{3.33}$$

This implies that $\hat{T}_{MBC}$ is more efficient than the usual non-parametric regression estimator proposed by Dorfman (1992)

3.3.4 Asymptotic Mean Square Error

The asymptotic mean square error of the estimator $\hat{T}_{MBC}$ is given by

$$MSE[\hat{T}_{MBC}] = Var[\hat{T}_{MBC}] + [Bias(\hat{T}_{MBC})]^2$$

$$MSE[\hat{T}_{MBC}] = \frac{1}{N^2}\sum_{i\in s}(\frac{1}{\pi_i})^2\sigma^2(x_i)+\frac{1}{N^2}\sum_{U}\sum_{j=1}^{n}(w_{xj})^2\sigma^2(x_j)+0_p(\frac{1}{(nh)^2})+[\frac{1}{N}\sum_{i\in s}\bar{y}+o_p(\frac{1}{nh})]^2 \tag{3.34}$$

As $n \to \infty$ and $h \to \infty$, the $MSE[\hat{T}_{MBC}]$ turns to 0 indicating that, the proposed estimator is statistically consistent.

Chapter 4

EMPIRICAL STUDY

4.1 Population

In this section, the theory developed in the previous section was applied using a set of simulation studies, with a mix of survey designs, and employing various approaches to selecting the best bandwidths. We employ a population U of countries in the world of size, N=188, with auxiliary variable x =gross national index(GNI) and variable of interest y =human development index(HDI), of interest is the population total of the HDI, $y = \sum_{l \in U} y_l$.

Figure 4.1 below shows the scatter diagram of the population. Where HDI is on the vertical axis and GNI on the horizontal axis, where there exist a quadratic relationship between the two variables.

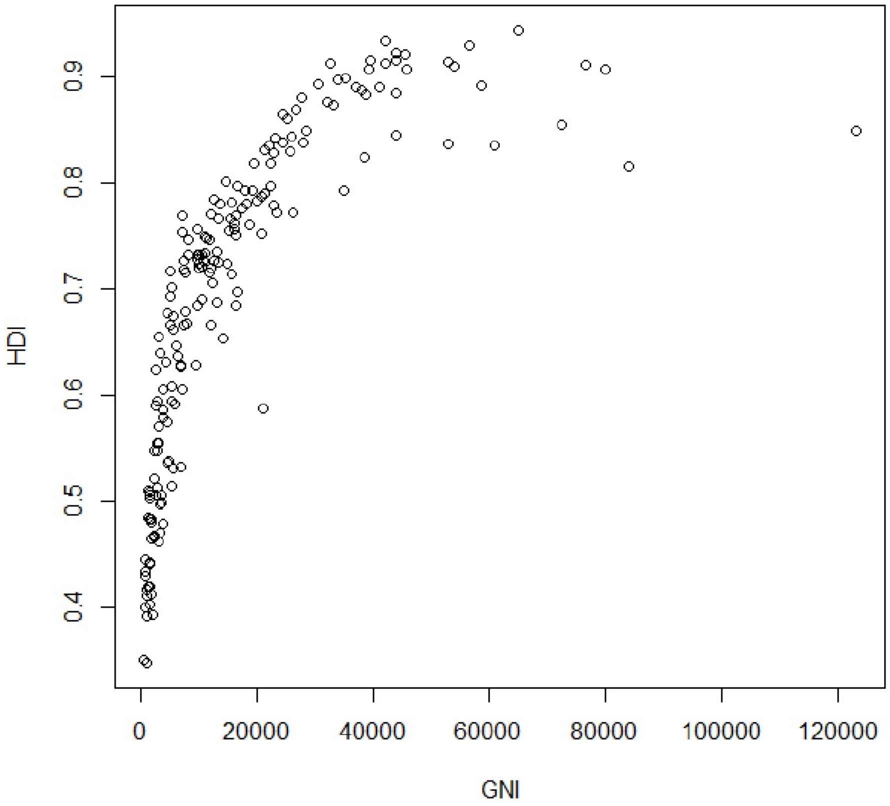


Figure 4.1: Scatter Graph

We suppose, for each run of the experiment that two samples are taken:

Sample 1(s_1): srswor(Simple random sampling without replacement)($n_1 = 32$)

Sample 2(s_2): stratsrs(Stratified simple random sampling)- four strata equal in each, and 8 units taken at random in each, so that $n_2 = 32$. The total experiment consists of 500 runs of pairs of samples. Table (4.1) below gives the estimators considered;

Table 4.1: Proposed and Existing Estimators

Estimator	**Formular**	**Comment**
Non parametric(NP) Regression	$\hat{T}_{NP} = \sum_{i \in s} y_i + \sum_{j \in r} \hat{h}(x_j)$	
Nonparametric(NPT) regression, twiced	$\hat{T}_{NPT} = \sum_{i \in s} \frac{y_i - \hat{h}(x_i)}{\pi_i} + \sum_{j \in U} \hat{h}(x_j)$	π_i = **Inclusion probabilities**
Multiplicative(MBC) Bias Corrected	$\hat{T}_{MBC} = \sum_{i \in s} \frac{y_i - \hat{h}^*(x_i)}{\pi_i} + \sum_{j \in U} \hat{h}^*(x_j)$	π_i = **Inclusion probabilities**

For an estimator $\hat{T}$ we considered three measures of relative success across the 500 runs:

(I) Unconditional relative bias measured

Bias=$\sum_{runs}(\hat{T} - T)/T$

(II) Unconditional relative root mean square error.

rmse=$\sqrt{(\sum_{runs}(\hat{T} - T))^2}/T$

4.1.1 Results

Results obtained from comparing estimators at table (4.1) are tabulated in the table 4.2;

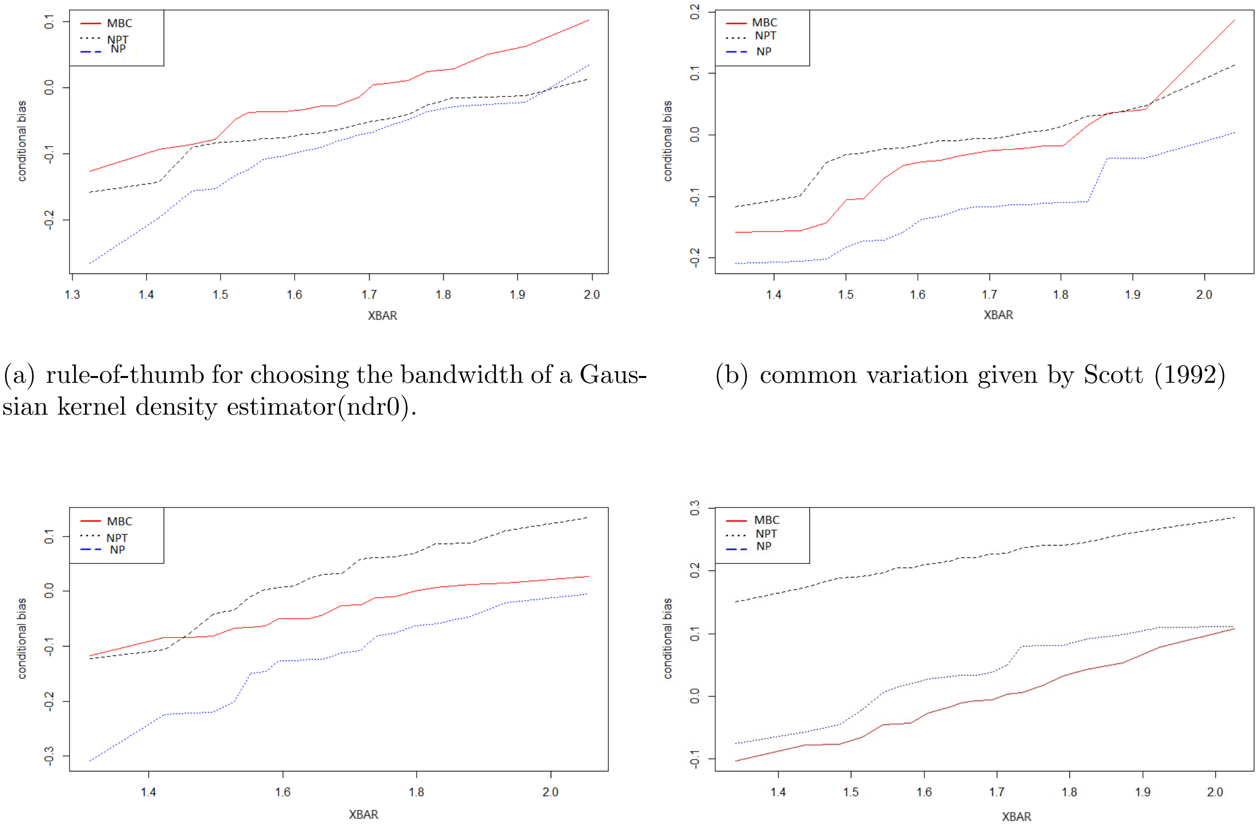
Table 4.2: Empirical results of different estimators

Estimators	**Methods of choosing Bandwidth**	**Bias/T**	10rmse/T
NP(one sample)	**ndr**	0.25	19.63
	ndr0	0.26	20.14
	bcv	0.11	20.71
	ucv	0.37	17.10
NP(s1Us2)	**ndr**	0.01	10.5
	ndr0	0.01	10.49
	bcv	0.45	11.19
	ucv	0.05	8.22
NPT	**ndr**	0.05	9.93
	ndr0	0.24	10.32
	bcv	0.39	10.83
	ucv	0.09	8.54
MBC	**ndr**	0.20	10.23
	ndr0	0.02	9.97
	bcv	0.23	10.17
	ucv	0.01	8.20

From the results obtained, we observe that the unbiased cross validation(ucv) approach is a viable means of selecting bandwidth as it gives the lowest bias and root mean square error across all the estimators. The proposed estimator to the two sample problem gives better estimates of the population total compared to those realized using the estimator proposed by Dorfman (1992),NP, and Dorfman (2009),NPT.

Further more, we study the conditional performances of the selected estimators. 500 samples obtained were sorted by the values of the mean of the auxiliary variable and put in 25 groups each containing 20 values. We then compute the bias and root mean square error of each group. The plots of conditional performances against the average of the sorted mean auxiliary variable. We then report the behaviour of the conditional bias for the different bandwidth.

Figures 4.2 and 4.3 indicate the conditional bias and conditional root mean square respectively on the y-axis ploted against $\bar{x}_i$(XBAR) on the x-axis 2, with each of the plot drawn at different bandwidth.



(a) rule-of-thumb for choosing the bandwidth of a Gaussian kernel density estimator(ndr0).

(b) common variation given by Scott (1992)

(c) biased cross-validation(bcv).

(d) unbiased cross-validation(ucv)

Figure 4.2: Plots indicating the conditional biases of three estimators.

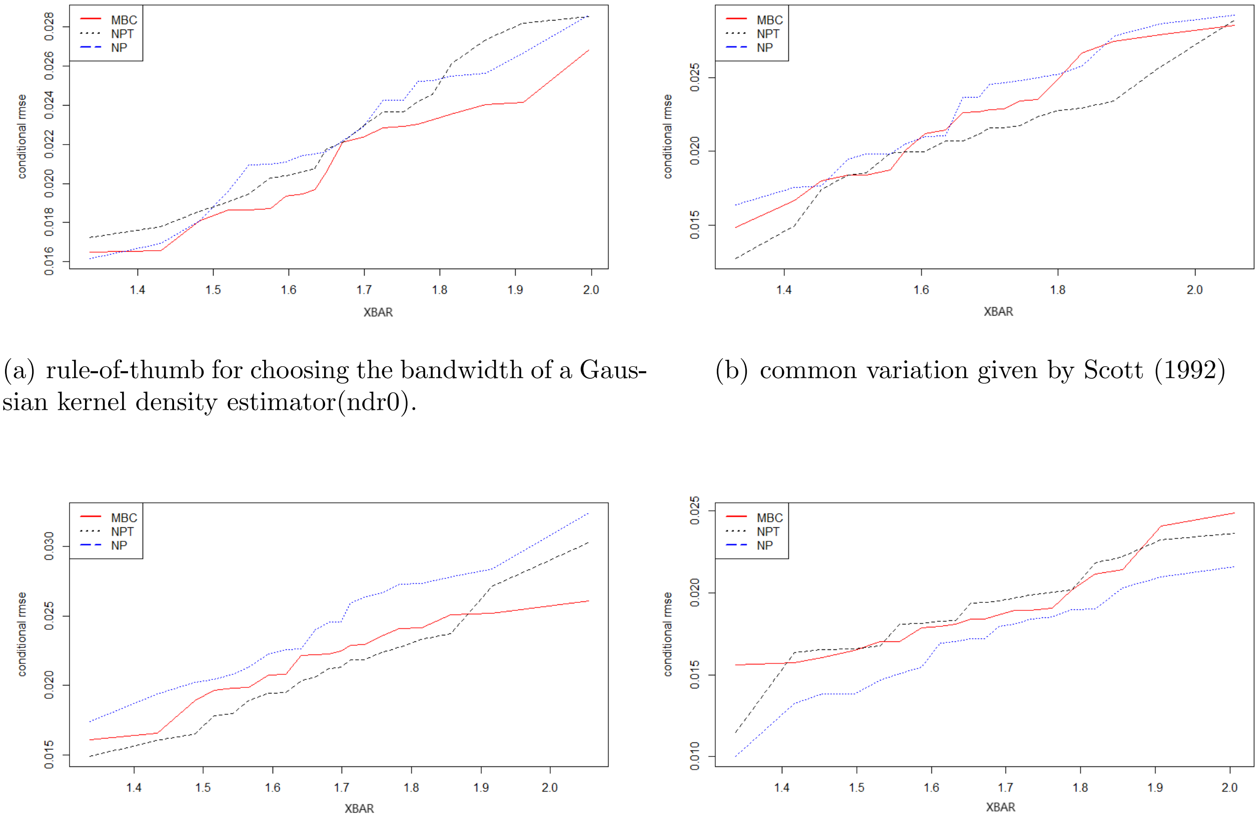
The population mean of auxiliary variable x_i was found to be $\bar{x_i} = 1.701$. At the balancing point that is $\bar{x_i} = 1.701$;

Fig 4.2(a) shows that the bias of our MBC estimator is nearer to zero than the other estimators. Moving towards the lower tail of the estimator, it is still the best estimate but towards the upper end, the other two estimators perform better.

Fig 4.2(b) shows that the NPT estimator is the closest to zero hence the best while our proposed estimator comes in as second. Toward the lower end of our estimator, the NPT estimator still performs better while at the upper tail, the other two estimators perform better than our estimator.

Fig 4.2(c) shows that the bias of our proposed estimator is nearer to zero at the lower tail of the estimator, upper tail of the estimator and at the point itself. So, the MBC estimator is the best compared the others estimators umder this bandwidth.

Fig 4.2(d) shows that the bias of proposed estimator, MBC estimator, is the nearest to zero, followed by the NP estimator. At the lower tail , the bias of the NP estimator is closest to zero and the upper tail, the MBC estimator is the closest to zero.



(a) rule-of-thumb for choosing the bandwidth of a Gaussian kernel density estimator(ndr0).

(b) common variation given by Scott (1992)

(c) biased cross-validation(bcv).

(d) unbiased cross-validation(ucv)

Figure 4.3: Plots indicating the conditional root mean square error of three estimators.

At the balancing point,the MBC estimator of fig 4.3(a) has the lowest rmse value, which indicate a better fit compared to the other two estimators. Moving towards the end of the tails, it is still the estimator with the lowest rmse.

Fig 4.3(b) shows that at the balancing point, our proposed estimator has the second lowest rmse value, it rmse is lowest at it approaches the upper tail compared to the other estimators. The same trend is observe in fig 4.3(c) and fig 4.3(d)

Chapter 5

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

The aim of this study was to develop an estimator with the lowest bias for the finite population total using the multiplicative bias corrected approach to non parametric regression. This study reveals that the proposed estimator is more efficient than the modified nonparametric estimator(NPT). With a suitable bandwidth selection(ucv), the proposed estimator has the smallest bias and root mean square error values. It has therefore proven to be efficient in resolving the biases with the values at the boundary that is associated with the existing nonparametric smoothers.

Under the conditional bias plots, it is concluded that, the proposed estimator out-performs the two currently used estimators in terms of conditional biases especially with the unbiased cross-validation and the biased cross-validation method of selecting bandwidth. This trend persist in the case of conditional root mean square error.

5.2 Recommendation

In this work, we apply the multiplicative bias corrected technique to all the values of the regression. Where as, bias in nonparametric regression occur mostly at the values found at the boundaries of the curves, So, it be interesting to know if by choosing a quantile and applying the multiplicative bias corrected technique at those tails, it will produce a better estimate of the population total than the one developed in this thesis.

References

Bierens, H. J. (1987). Kernel estimators of regression functions. In *Advances in econometrics: Fifth world congress*, volume 1, pages 99–144.

Changbao, W. (2004). Combining information from multiple surveys through the emperical likelihood method. *The Canadian journal of Statistics*, 32(1):15–26.

Dorfman, A. H. (1992). Nonparametric regression for estimating totals in finite populations. In *Proceedings of the Section on Survey Research Methods*, pages 622–625. American Statistical Association Alexandria, VA.

Dorfman, A. H. (2008). The two sample problem. In *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods.*

Dorfman, A. H. (2009). Nonparametric regression and the two sample problem. In *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*, pages 277–270.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American statistical Association*, 87(420):998–1004.

Hengartner, N., Matzner-Løber, E., Rouviere, L., and Burr, T. (2009). Multiplicative bias corrected nonparametric smoothers. *arXiv preprint arXiv:0908.0128.*

Linton, O. and Nielsen, J. P. (1994). A multiplicative bias reduction method for nonparametric regression. *Statistics & Probability Letters*, 19(3):181–187.

Marron, J. S. and Härdle, W. (1986). Random approximations to some measures of accuracy in nonparametric curve estimation. *Journal of Multivariate Analysis*, 20(1):91–113.

Masayuki Hirukawaa, M. (2014). Nonnegative bias reduction methods for density estimation using asymmetric kernels. *ComputationalStatisticsandDataAnalysis*, 92(75):112–123.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99(468):1131–1139.

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric methods in survey sampling. In *New Developments in Classification and Data Analysis*, pages 203–210. Springer.

Muller, H.-G. and Stadtmuller, U. (1987). Variable bandwidth kernel estimators of regression curves. *The Annals of Statistics*, pages 182–201.

Renssen, R. H. and Nieuwenbroek, N. J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92(437):368–374.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*, pages 1215–1230.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, 92(439):1049–1062.