
DEC. 1995
DECLARATION

This thesis is my original work and has not been presented for degree in any other University

Romanus Odhiambo Otieno

Signature ______________________________

Date: 9/1/96

This thesis has been submitted for examination with my approval as the University Supervisor.

Dr. Charles Wafula

Signature ______________________________

Date: 9/1/96
ABSTRACT

After a sample has been obtained the statistic of interest can be computed. The next step (and a more formidable one) is the assessment of the accuracy (precision) of the resulting statistic. The most commonly used measure of accuracy in the model based surveys is the variance of the prediction error associated with the considered statistic. In general variance are not known and must be estimated using the sampled data.

In this thesis we have proposed new methods for estimating the error variance for finite population sampling. In particular we have considered fixed bandwidth Kernel smoothing of the squared residual and bootstrap technique based on resampling of the residuals.

Analytical and empirical performances of the new variance estimators are studied vis-avis the robust estimators favoured in the current practice.

On average the proposed estimators have better robustness properties than estimators favoured in the current practice. Further more the new estimators have the desired properties of non negativity, simplicity and extend even to cases where some of the current estimators can not be applied.
ACKNOWLEDGEMENTS

As always it is not possible to adequately thank all the friends, teachers and members of my family who assisted and supported this thesis project. I am grateful to many more than I can express here.

For advisory support, Dr. Charles Wafula was everything a supervisor ought to be resourceful patient and reliable.

I would also like to thank Kenyatta University for once again giving me an opportunity to carry out a study that has led to this piece of work.

The empirical results were programmed by Madow and Adongo. I am grateful to them.

The preparation of this thesis would not have been easy if I had not got financial assistance from DAAD (German Academic Exchange services). To them I extend my sincere appreciation.

I cannot forget to thank my dear wife, Awuor who accepted to take care of the entire family while I was studying in the office. Her encouragements and prayers are greatly appreciated.
Contents

1 INTRODUCTION 4

1.1 Description of the Population and the selection scheme 4
1.2 The Sample Survey Problem 6
1.3 The Main Approaches to Sample Survey Problem 7
  1.3.1 The randomization approach 7
  1.3.2 The model-based approach 10
1.4 Literature Review 14
  1.4.1 Unistage sampling 14
  1.4.2 Multistage sampling 20
1.5 Objectives of the Study 22
1.6 Outline of the Thesis 25

2 ROBUST ESTIMATION OF THE ERROR VARIANCE IN UNISTAGE 26
SAMPLING 26

2.1 Introduction 26
2.2 The Working Model and the Error Variance 28
2.3 Nonparametric Estimation of the Error Variance 30
5 AN EMPIRICAL STUDY

5.1 Description of the Study Populations
5.2 Simulation Procedure
  5.2.1 Simulation results for unconditional relative bias REBI(V)
  5.2.2 Simulation results for root average square error RASE(V)
  5.2.3 Simulation results for unconditional coverage probability COPRO(V)
  5.2.4 Simulation results for conditional biases
  5.2.5 Simulation results for conditional coverage probability COPRO(V)
5.3 Concluding Remarks

6 SUMMARY AND FURTHER QUESTIONS

6.1 Epilogue of the Thesis
6.2 Further Questions
Chapter 1

INTRODUCTION

1.1 Description of the Population and the selection scheme

We consider a finite population of \( N \) identifiable units

\[
U = (u_1, u_2, \ldots, u_N)^T
\]

for which some frame is available. Associated with this vector is an unknown vector

\[
Y = (Y_1, Y_2, \ldots, Y_N)^T
\]

of characteristics under study, where \( Y_i \) (\( i=1, \ldots, N \)) is the value of characteristic under study for the \( i \)-th unit of the population. We also write

\[
Y = (Y_s, Y_r)^T
\]

the partition of \( Y \) induced by the selection of units in \( s \), \( s \) being a subset of \( n \) distinct units while \( r \) is the complement of \( s \). We consider a case where there exists an \( N \times q_L \) known matrix

\[
X = (X_1, X_2, \ldots, X_N)^T
\]
The prior data $X$ usually include indicator variables which express sample stratum or cluster membership as well as quantitative variables such as size measurements. The sample $s$ from the finite population $U$ is selected using a known sampling selection scheme. If we let

$$a(s) = (a_1(s), a_2(s), \ldots, a_N(s))^T$$

where

$$a_i(s) = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{otherwise} \end{cases}$$

then the vector $a(s)$ determines which units are selected in sample $s$. In general the rule for evaluating $a(s)$ (i.e. the sample selection scheme) will depend on the prior information $X$, the survey variable $Y$ and the unknown parameter vector $\theta$. In many applications the joint effect of $Y$ and $\theta$ on the selection scheme is generally represented by a vector $W$.

Hence a sample selection scheme can be expressed as

$$p(s, X, W)$$

where $p(s, X, W)$ is such that

$$p(s, X, W) \geq 0$$

$\forall s$ and

$$\sum_{s \in S} p(s, X, W) = 1$$

where $S$ is the set of all possible samples that can be obtained from $U$. For standard random sampling schemes, $a(s)$ is a random variable which depends only on the prior information $X$ and so the selection scheme can be expressed as $p(s \mid X)$, a probability
density function defined on all possible subsets \( s \subseteq S \). This probability is known exactly since both \( s \) and \( X \) are known. The inclusion probabilities (i.e. probability of including the \( i \)-th unit in a sample) \( \pi_i \) \((i = 1, \ldots, N)\), can be deduced from \( p(s \mid X) \).

### 1.2 The Sample Survey Problem

In most cases the aim of sample survey is to estimate some function \( f(Y) \) of the vector \( Y \) by a statistic \( \hat{f}(Y_s) \). The function \( f(Y) \) can be population total

\[
T = \sum_{j=1}^{N} Y_j;
\]

the population mean

\[
\bar{Y} = \frac{T}{N}
\]

the population variance

\[
\sigma^2 = \frac{\sum_{i \in U} (Y_i - \bar{Y})^2}{N}
\]

or any function of these finite population parameters.

A statistical inference about these quantities is called descriptive inference. In this inference if all the units were observed, then there would be no uncertainty in the inference.

Sometimes, however, the objectives of sample survey go beyond estimation of finite population quantities, at the time the sample was drawn. For example, for many multivariate analysis problems, inferences about wider population are paramount. Here the concern is: given a finite population, how do we use the available information to make inference about the infinite population that is believed to have generated the finite population? This problem is more akin to the usual estimation problem in the conventional statistics. In the finite population, a solution to this problem has been to assume certain
probability model that is believed to have generated the finite population from the infinite population. The problem then is to estimate the unknown parameters in the assumed model. Once the unknown parameters have been estimated, then "prediction" about functions of variables from the nonsample units can be made. This kind of inference is called analytic inference. These are essentially inferences about parameters of underlying probability distribution that is assumed to have generated the finite population. A typical example is the estimation of parameters in a regression model.

In conclusion, therefore, sample surveys deal with estimation of unknown function from a finite population and the subsequent inferences about these unknown functions. The term 'function' is taken to mean descriptive function as in descriptive surveys or parameters as in analytic surveys.

1.3 The Main Approaches to Sample Survey Problem

The literature on estimation and inference in sample survey is extensive and complex, but generally distinguishes between two approaches: randomization approach and model-based approach. In this section we describe the main features of the two approaches.

1.3.1 The randomization approach

In the randomization approach the values of $Y$ are assumed to be unknown constants and the only probabilities are $p(s \mid X)$. If $p(s \mid X)$ is not a random selection scheme then randomization inference is not possible. With random sampling, the probability distribution, (i.e. $p(s \mid X)$) is known exactly and does not depend on any unknown parameters, nor on the unknown constants. Since the values of $Y$ do not index $p(s \mid X)$, they can not
be interpreted as parameters and so interpretation of \( Y \) as a set of unknown constants which are neither random variables nor parameters seems justified. Since \( X \) and \( p(s \mid X) \) are known, inferences about this distribution are irrelevant. The only way to employ \( p(s \mid X) \) is to take averages or expectations. If for a statistic \( \hat{f}(Y_s) \)

\[
E_p\left(\hat{f}(Y_s)\right) = f(Y)
\]

where

\[
E_p\left(\hat{f}(Y_s)\right) = \sum_{s \in S} p(s \mid X)\hat{f}(Y_s),
\]

then we say that \( \hat{f}(Y_s) \) is \( p \)-unbiased, but if

\[
E_p\left(\hat{f}(Y_s)\right) \neq f(Y),
\]

then we can say nothing for the bias and any measure of efficiency will depend on the unknown population values \( Y_i, i \in r \).

The concentration on unbiasedness in the randomization approach now appears natural because there is really no other property that can be examined, given the comparison will depend on the unknown constants and hence can not be determined. Certain estimates (e.g. ratio or regression estimates) are in fact biased estimates, but motivation for their usage is in the fact that they have smaller mean square error (mse) than that of the unbiased estimates for certain population. The decision to use these estimates, however, does not rest on any principle of statistical inference related to randomization distribution but on the statistician’s knowledge about the structure of the population which also influences the choice of the sampling.

Besides unbiased estimation under repeated sampling, interests in randomization ap-
proach has also been concerned with the development of relationship between an estimator and its standard error. If the sample is large then in most cases this relationship is established by an appeal to central limit theorem. If the population is finite, then an appeal to asymptotic results is not possible directly but it is to be hoped that approximate normality will hold if the sample size is large.

The main arguments in favour of randomization as a method of inference seem to be three folds.

1. the need to specify a model relating $Y$ and $X$ is avoided.

2. the method is robust in the sense that, since no explicit model is assumed, the inference made from a sample are insensitive to the violation to the assumptions that would have been made in any model. This is achieved here through the use of sampling with known probabilities and consistent estimators, and using a sample large enough so that central limit theorem holds. Hansen, Madow and Tepping (1983) base their defence of the randomization approach on these grounds.

3. the method seems to work in the case of the numerous surveys and is now publicly acceptable.

However the approach does have its drawbacks. If the sampling selection scheme $p(s | X)$ is not a random selection scheme then no randomization inference is possible: the $\pi(i = 1 \ldots, N)$ would be unknown prior to the survey. For example in quota sampling, selection depends on some of the values of $Y$ and $X$. For the problem of nonresponse, common to all social surveys, the selection is of an unknown form dependent on $Y$ and $X$ and possibly other unknown variables. A second argument against the randomization approach is that it is easy to construct population for which central limit theorem does
not apply (Smith (1979)) and so the validity of this inference requires a restriction of the parameter space of possible values of $Y$. This restriction needs to be specified for each population, and so it is not true that randomization inference is totally free from assumption. Thirdly, true population values are rarely available and so determining whether the method 'works' is not easy. Smith (1983) gives the example of public opinion polls when the true population value is known shortly after the sample has been selected, and concludes that since the Gallup polls (employing non-random samples) do as well as National opinion polls (employing random samples) then if random sampling 'works' so does quota sampling.

A further argument against randomization approach is that the randomization distribution is defined on a given sampling frame for a fixed set of measurements of $Y$. Therefore only descriptive inference can be made to this frame. This is one major reason why useful estimation techniques such as maximum likelihood estimation procedure, ordinary least squares estimation technique, etc. can not be applied in this approach. Strictly speaking, a randomization approach to analytical inference is not possible without additional model-type assumptions relating the finite population parameter to the parameter of the analytical interest.

1.3.2 The model-based approach

In recent years much literature has appeared challenging the randomization approach and model-based approach closely related to that in the traditional statistical inference has been proposed to overcome some of the problems in randomization approach. The model-based approach (prediction approach) starts from the assumption that the unknown vector $Y$ in the finite population can be treated as a random variable. Thus
randomness is introduced directly into $Y$ values. The first problem is thus to construct a stochastic model for $Y$ (normally called probability model) that incorporates within it all the population information contained in the prior data, $X$. If $X$ contains indicators for sample membership, stratum, or cluster membership then this information should be incorporated into the model. Once a model has been chosen then a sample selection scheme consistent with that model and taking other consideration such as cost will be employed to draw a sample, $s$. The sample is thus viewed as the outcome of a two-step sampling procedure in which the finite population is generated at the first step and the sample at the second step. The inference is then based on the probability distribution specified by the model, which we call the $\xi$-distribution. Once the design has been chosen and the sample (it need not be a probability sample) obtained, then the selection probabilities do not feature in the model-based inference. This approach looks attractive conceptually because it embraces both descriptive and analytic inferences.

Unlike randomization approach where interest centres on unbiased estimation under repeated sampling, here the concerns are: unbiased estimation given one sample, the choice of the model and finally the robustness of the given estimator whenever the assumptions in the model fail. A key point to observe is that model-based approach depends heavily on assumptions and, if used naively, will not be robust. Royall and his co-researchers recognized this problem at an early stage and in a series of papers (see Royall and Cumberland (1981), and references there in) they have considered various designs and estimators (strategies) which will make their inferences more robust. For example, if the population is stratified then models are fitted within the strata. If there is a relationship between a survey variable $Y$ and a prior (design) variable $X$ then a ratio or regression estimator is employed. Royall and his co-researchers advise that this estimator should be protected.
against the misspecification of the mean by using a balanced sample and against misspecification of the variance structure by using a robust variance estimator. This combined strategy will often give designs and estimators similar to those used in the randomization approach.

One model-based approach for descriptive surveys was developed by Royall in a series of papers during 1970's. Least squares prediction was used to obtain estimators (or predictors) of finite population totals under variety of assumed model representing the underlying population structure. The basis of this method is to "predict" functions of the variables in the unobserved units in the population from the observed values in the sample by employing the \( \xi \)-distribution (see Royall (1970, 1971), Royall and Herson (1973)). Royall (1976) gives further discussion of this approach.

As an example, the following simple regression model

\[
E(Y_i) = \beta x_i, \text{cov}(Y_i, Y_i) = \begin{cases} 
\sigma^2 x_i & \text{if } i = j \\
0 & \text{otherwise}
\end{cases}
\]

is used whenever the ratio estimator \( \hat{T}_R \), is used to estimate the population total \( T \), where

\[
\hat{T}_R = \frac{\bar{y}}{\bar{x}} \sum_{i=1}^{N} x_i
\]

\( \bar{y}, \bar{x} \) are sample means of \( y_i, x_i \) respectively. A sample is selected from the finite population using a specified random selection scheme, \( P(s | X) \) (e.g. simple random sampling) which is then ignored in the inference stage. Expectation and variances of the predictor are then taken with respect to the \( \xi \)-distribution. Inferences about \( f(Y) \) are made via this distribution. A statistic \( f(Y_s) \) is said to be \( \xi \)-unbiased if

\[
E_\xi(f(Y_s)) = E_\xi(f(Y)).
\]
For example $\widehat{T}_R$ is $\xi$-unbiased under the above model since

$$E_\xi(\widehat{T}_R) = E_\xi(T).$$

Even when the sampling scheme and estimator are similar under both approaches, inferences from a model will often differ from randomization inferences. Under a model, when the prior variable $X$ is known, the selection rule $p(s \mid X)$ can be ignored and inferences made conditional on the units in the sample. Randomization inferences are unconditional and average over all samples that might have been drawn. There are no principles which lead to conditional inference in the randomization approach. This is one of the major problems which must be faced. Further within the framework of the randomization distribution generated by a random sampling scheme randomization inferences will be valid statistically for many finite populations. Likewise within the framework of an agreed statistical model, inferences will be valid statistically. The question is: which set of inferences has greatest scientific validity, which is the most relevant to the user of the work that has been done in the survey problem we consider the following survey data?

In a model-based approach, an attempt is made to construct a stochastic model which, given prior data $X$, could have generated the finite population under consideration. It is known that such a model can never be correct but only represents an approximation to the truth, a working hypothesis. In the randomization approach, a sampler chooses a selection scheme based on the prior data $X$. If the scheme captures the relationship between $Y$ and $X$ then inferences will be precise, variances will be small. If the scheme fails to capture any relationship then the variances will be large. But regardless of the relationship of the survey and auxiliary variables inferences based on the randomization
principles will always be valid statistically as long as approximate normality holds. Thus it is impossible to talk about (address the question of) robustness, it has no meaning, all one can do is to discuss the unbiasedness of an estimator assuming one has all the possible samples from the population.

The approach to be followed in this study is the model-based approach. In particular our concern will be that of estimating the error variance of \( \hat{f}(Y_s) \) under the model-based approach. Such estimators are extremely useful for setting confidence intervals and measuring confidence levels associated with the point estimator in use. Besides it can also be used for planning future surveys. We review the methods that have been developed for this important problem in section 1.4. Section 1.5 gives the objectives of this study while section 1.6 gives the outline of the thesis.

1.4 Literature Review

1.4.1 Unistage sampling

For reviewing the work that has been done in the above problem we consider the following model:

\[
E(Y) = X\beta \\
\text{Var}(Y) = W\sigma^2
\]

where \( \beta \) is a vector of q unknown regression coefficients and \( W\sigma^2 \) is a diagonal covariance matrix. The function \( f(Y) \) that has received a lot of interest in the literature is the population total \( T = \sum_{i=1}^{N} y_i \). Let \( \hat{T} \) be an estimator of \( T \). Without loss of generality, the vector \( Y \) is arranged so that the first \( n \) elements correspond to the sample units and \( Y, X \)
and \( W \sigma^2 \) are partitioned according to sample and non sample units:

\[
\begin{align*}
Y &= \begin{pmatrix}
Y_s \\
Y_r
\end{pmatrix} \\
X &= \begin{pmatrix}
X_s \\
X_r
\end{pmatrix} \\
W &= \begin{pmatrix}
W_s & 0 \\
0 & W_r
\end{pmatrix}.
\end{align*}
\]

The Best Linear Unbiased Estimator (BLUE) of \( T \) under (1.1) Royall (1976) is

\[
\hat{T}_B = \lambda_s^T Y_s + \lambda_r^T X_r A^{-1} X_s^T W_s^{-1} Y_s
\]

\[
= \lambda_s^T Y_s + \lambda_0^T b
\]

where \( b = A^{-1} X_s^T W_s^{-1} Y_s \), \( \lambda_r^T = \lambda_s^T X_r \), \( A = X_s^T W_s^{-1} X_s \) and \( \lambda_s^T, \lambda_r^T \) are vectors of the form \((1, \ldots, 1)\) having lengths \( n \) and \( N - n \) respectively. The error variance of \( \hat{T}_B \) under (1.1) is given as:

\[
\text{Var}_\varepsilon(\hat{T}_B - T) = (\lambda_0^T A^{-1} \lambda_0 + \lambda_r^T W_r \lambda_r) \sigma^2.
\]

Under the prediction approach interest is in the estimation of (1.2). One obvious way of obtaining an unbiased estimator of (1.2) is to use the weighted least squares residual technique. However, obtaining an unbiased estimator is not enough. It is well recognized that no model can be claimed to represent reality perfectly. So if one obtains an unbiased estimator of (1.2) assuming model (1.1) is correct, what would happen to such an estimator if the model is incorrect in some of its specifications? This is one of the most important
questions that underly current research on variance estimation methods in the model-based approach.

A close look at (1.1) reveals that at least three misspecifications can occur in the model:

1. There could be an error in specifying the variance structure. For example, the variance matrix could be $V$ which is not proportional to $W$.

2. There could be an error in the specification of the expected value of $Y$. For example, instead of $E(Y)$ being a linear function, this relationship can be non-linear.

3. Finally, (1.1) further assumes that the variance-covariance matrix of $Y$ is diagonal. However, it is well known that data that follow a time series sequence can violate this important assumption quite significantly.

Ideally then a key research objective in model-based approach has been to obtain an unbiased estimator under (1.1) which remains unbiased or approximately unbiased under (i.e. bias robust to the) possible misspecifications in the $\xi$-distribution.

The above problem was first considered by Royall and Eberhardt (1975) in the context of the ratio estimator and a special case of model (1.1). Royall and Cumberland (1978) considered this problem too but for the general linear regression model (1.1). In constructing robust variance estimators for finite populations, these authors primarily considered misspecifications in the variance structure. They supposed that the correct model is (1.3) and not (1.1):

\begin{align*}
E(Y) &= X\beta \\
\text{var}(Y) &= V\sigma^2
\end{align*}
where $V$ is not proportional to $W$.

We note that $\text{var}(T_B - T)$ in (1.2) has two components. The second depends on the non-sample units only, while the first depends on sample units as well. Royall and Cumberland showed that it is possible to obtain an estimator of the first component which is unbiased under (1.1) and which remains unbiased or approximately so under model (1.3). They give three estimators for this component. To estimate the second component, they suggested three estimators which are strictly unbiased under model (1.1). Adding the estimators of the two components they obtain three estimators of (1.2) which are strictly unbiased under (1.1) and asymptotically unbiased under model (1.3). This method of estimating (1.2) is called direct method of variance estimation.

A main criticism of the above papers was that they were mainly theoretical with no empirical backing. Even their theoretical conclusions are suspect: they are based mainly on asymptotic theory. However, it is well known that such results do not reflect, quite generally, the true performances of the estimators in finite populations. For example, the robustness of the proposed estimators to deviations from variance structure in model (1.1) are suspect especially in small samples. This point became clearer from subsequent empirical studies of the new estimators alongside the estimator based on the least squares technique and the classical variance estimator popularly employed in the randomization approach. Authors such as Royall and Cumberland (1981), Wu and Deng (1983) have given an extensive and convincing exposure and analysis of the empirical properties of the variance estimators. The main three conclusions are critical of both the randomization and the prediction approach and are:

1. Classical estimator of the error variance does not provide robust inference.
2. the variance estimator based on least square theory is inadequate.

3. The estimators proposed by Royall and Cumberland (1978) are robust only in specific families of models (see Royall and Cumberland rejoinder to Smith(1981)'s criticisms). In fact in section 4.2 of their paper, Royall and Cumberland (1978) assert emphatically that these variance estimators are only reliable in samples that are well balanced on the auxiliary. The authors seem to be setting a precedence that these estimators may not be robust in all situations. Even in their theoretical investigations of the properties of the estimators, the authors admit that the new estimators may not give reliable confidence intervals in some situations.

The third point became clearer from Royall and Cumberland (1988) study in which these authors investigated the conditional coverage properties of the confidence intervals associated with the various variance estimators. The performance of the variance estimators were less satisfactory in some of the populations (counties 70 and cities). The work of Royall and his co-authors (1975, 1978) also considered the robustness of the derived estimators to the misspecification of the regression line $E(Y) = X\beta$. For example, they considered situation for which the expected regression line is not $X\beta$ but is represented by $X\beta + Z\Omega$ where $Z$ is an $N \times q$ matrix of rank $q$ whose columns are linearly independent of those of $X$. $\Omega$ is a vector of $q$ unknown constants. It is shown that such a misspecification renders the variance estimators biased with an undesired consequence of conservative coverage performance in some cases.

We emphasize that Royall and co-authors (1975, 1978) studied robustness of variance estimators to the misspecifications of the regression line and variance structure. The question as to what happens to the estimators if the components of the survey vector are dependent...
has been considered by Wafula (1988) and Odhiambo (1991). The results suggest that the usual variance estimators are sensitive to the abuse of the assumption of independence. These studies, however, never suggested alternative robust procedures to handle this common problem.

Outside the direct variance estimation method, very little work on new methods as applied to model-based variance estimation has been made. However, model-based resampling technique of Do and Kokic (1992) is worth quoting here. These authors have suggested an application of bootstrap technique to variance estimation for model-based surveys. It must be remembered that a lot of studies employing bootstrap procedure have been done (Wu and Rao (1988), Sitter (1992)). However, these studies have been mainly design-based. The work of Do and Kokic seems to be a radical departure from this tendency.

More specifically, Do and Kokic have considered the estimation of the confidence intervals for the population total when the vector of weights are ridge-type. We note that, unlike Royall and Eberhardt (1975) and Royall and Cumberland (1975, 1978), these authors are considering robustness problem in the framework where $X$ is not of full rank, a crucial requirement in model (1.1). An outstanding result of this study is that the proposed estimator has promising coverage properties. A criticism of this study is that the new estimator was not compared at all with a direct model-based estimator favoured in the current practice (see Dunstan and Chambers (1986)). In general, the properties of a new estimator should never be appraised without comparing them with the estimator favoured in the current practice.
1.4.2 Multistage sampling

The above contributions to finite population variance estimation considered populations in which sampling units are elements. It was Royall (1986) that employed the prediction approach to surveys in which sampling units are clusters.

In this method Royall extended the direct method of variance estimation to two stage cluster sampling. The study considered populations which can be organized in \( N \) clusters, each cluster being of size \( M_i \) (\( i = 1, \ldots, N \)). Associated with the \( j \)-th unit (\( j = 1, \ldots, M_i \)) of the \( i \)-th cluster is a variate of interest, \( y_{ij} \). A sample is taken in two stages. A random sample \( s \) of size \( n \) first stage units is taken. Then from the \( i \)-th sampled cluster a second stage sample \( s_i \) of size \( m_i \) is taken. For estimating total

\[
Y = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}
\]

Royall (1986) considered the following general class of estimators:

\[
\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} u_i \hat{y}_i
\]

where \( \hat{y}_i = M_i \bar{y}_i \), \( \bar{y}_i = \frac{\sum_{j=1}^{M_i} y_{ij}}{M_i} \). Assuming variables in different clusters are uncorrelated, Royall obtained the error variance of \( \hat{Y} \) as

\[
Var(\hat{Y} - Y) = \frac{N^2}{n^2} \sum_{s} u_s^2 Var(\hat{y}_i) - \frac{2N}{n} \sum_{s} u_s Cov(y_i, \hat{y}_i) + \sum_{i=1}^{N} Var(y_i)
\]

where \( y_i = \sum_{j=1}^{M_i} y_{ij} \). Formula (1.4) clearly indicates that the error variance has three components. To obtain a robust variance estimator, Royall simply obtained estimators of these three components under different models. An unbiased estimator of the first component is obtained under the following model

\[
E(y_{ij}) = \mu
\]
For the second component, the following model was used

\[
E(y_{ij}) = \mu \\
\text{Cov}(y_{ij}, y_{kl}) = \begin{cases} 
\sigma_i^2 & i = k, j = l \\
\rho_i \sigma_i^2 & i = k, j \neq l \\
0 & \text{otherwise}
\end{cases}
\]

Finally, for the last term the following model was employed:

\[
E(y_{ij}) = \mu \\
\text{Cov}(y_{ij}, y_{kl}) = \begin{cases} 
\sigma^2 & i = k, j = l \\
\rho \sigma^2 & i = k, j \neq l \\
0 & \text{otherwise}
\end{cases}
\]

Combining the pieces of estimators, he obtained the following estimator for model-based complex surveys:

\[
Vo = \frac{N^2}{n^2} \sum_{i \in s} u_i (u_i - f) \hat{V}_i + \frac{N}{n} \sum_{i \in s} u_i M_i (1 - f_i) \frac{R^2}{f_i}
\]

\[
+ \left( \frac{1}{N} \sum_{i=1}^N M_i^2 - \frac{1}{n} \sum_{i \in s} u_i M_i^2 \right) \theta_2
\]

where

\[\hat{V}_i = r_i (1 - g_i / n)^{-1} - \frac{M_i^2 (1 - g_i / n)^{-1}}{M_i^2} \left( u_i^2 M_i^2 (1 - g_i / n)^{-1} \sum_{j \in s} u_j^2 M_j^2 (1 - g_j / n)^{-1} \right) \sum_{j \in s} u_j^2 M_j^2 (1 - g_j / n)^{-1} \]

\[r_i = \frac{\hat{y}_i - M_i \hat{Y} / N \hat{M}}{N \hat{M} / \hat{M}}, g_i = 2u_i M_i / \hat{M}, \hat{M} = \frac{N}{n} \sum_{i=1}^N M_i / \bar{M} = \frac{1}{n} \sum_{i \in s} M_i, R_i^2 = \frac{\sum_{s_i} (y_{ij} - \bar{y}_{s_i})^2 / \gamma n_i - 1}{} \]

\[\hat{\theta} = \frac{N \sum_{i \in s} (\hat{V}_i - M_i, S_i^2 / f_i)}{\sum_{i \in s} M_i^2} \quad f_i = \frac{n}{N}, \quad \hat{m}_i = \frac{M_i}{M_i} \cdot \]

Clearly if model (1.7) is assumed then an estimator \( V_0 \) will be robust to certain departures
in the covariance structure of this model. However, a number of questions and criticisms arise out of Royall’s procedure. In the first place, the procedure requires that the number of first stage sampling units should be greater than $g_i$ for every $i$ in $s$. For many surveys this is a very restrictive condition. In the special case of the ratio estimator (i.e. if $u_i = \frac{M_i}{m_s}$) this requirement implies that $n > 2M_i/\overline{m_s}$ for every $i$ in $s$. If $M_i$'s vary markedly, it will require a large value of $n$ to ensure non-negativity of the corresponding variance estimator. Further criticism of the procedure is in its application to stratified two stage cluster sampling. A case in point is the selection plan which picks two clusters at the first stage from each stratum. The procedure will not work here, yet this is a very useful design in complex surveys, especially in small area estimation. Finally a look at Royall’s models immediately raises the robustness questions regarding the estimator $V_0$. How robust is this estimator to those practical cases in which model (1.7) does not apply?

Royall evaded this crucial question by considering large samples under stable conditions. For many survey problems, like in small area estimation, it is known that estimators of variance based on such theory do not give a good measure of standard error of the associated point estimator (Chaudhuri(1994)). Another question not addressed by Royall is the robustness of $V_0$ to the misspecification of the covariance structure between the clusters. In many applications units in adjacent clusters tend to be autocorrelated. For these cases $V_0$ may not be appropriate.

### 1.5 Objectives of the Study

We note the following about past research on prediction approach to survey sampling
1. The linear least square prediction theory has been extensively developed when the units are single elements. Available literature reveal that the method of least squares technique is not a plausible method and it is not worth pursuing further. The direct method of variance estimation is currently the method of choice for many model-based surveys. However, several pertinent questions and criticisms regarding this method remain unsolved. For example the question of robustness of the estimators especially from the small samples, as is usually the case in many practical surveys, remains unattended. In small samples several questions can be raised.

Consider the following cases: suppose that one wants to estimate the variance of the estimator using a set of measurements whose spread is not of a polynomial type but something less regular, less defined. How does one proceed? Or suppose that one is totally not sure of the spread of the measurements, then how does one proceed? Clearly for these cases one can not use the direct method of variance estimation.

2. The direct method of variance estimation can partially be described as a semiparametric method where one uses a more general (robust) model to derive estimators of the dominant term in the error variance, and then use a more specified model to derive estimators of lower order terms of the error variance. One area that has not been exploited is the estimation of the two components by employing a general model at both stages. For estimation of the dominant term of the error variance one can apply the procedure of Royall and Cumberland (1978). Procedures are, however, not available for estimation of the lower order terms when the working model contains parameters whose functional forms are not specified.
3. the model assumed by Royall (1986) does not represent some of the clustered populations met in practice. For example, there are situations in which population elements are spread over a large geographical area, and clusters are created by subdividing the area. In this case, it might be unreasonable to assume that adjacent elements which happen to fall in different clusters are uncorrelated. For these populations, the estimators proposed by Royall (1986) may be inappropriate.

4. The bootstrap method has so far proved to be a competitive estimation technique in the design-based surveys (Kovar, Rao, and Wu (1988), Rao and Wu (1988), Sitter (1992)). The superiority of the confidence intervals based on this procedure over the normal theory intervals is now a common knowledge (see for example Abramovitch and Singh (1985), Efron (1987)). However, development of this technique in the model-based surveys has been, to a large extent, a neglected area. The work of Do and Kokic (1992) is only applicable to multicollinearity cases considered by Dunstan and Chambers (1986). How the bootstrap procedure (as discussed both in the design-based approach and in the study of Do and Kokic) can be re-designed so as to be applicable to other model-based surveys remain an open research problem.

The work to be presented in the subsequent chapters of this thesis will address some of the above problems. In particular, the main objectives of this thesis are:

1. derive and study nonparametric variance estimators for one stage and two stage samplings.

2. develop and study model-based bootstrap technique for two stage cluster sampling.

3. examine the empirical properties of the derived estimators.
1.6 Outline of the Thesis

This thesis is organized into six chapters. Chapter two gives the nonparametric procedure and some other robust procedures for estimating the error variance of the ratio estimator when the sampling units are single elements. We examine some properties of these estimators in the same chapter. Chapter three is an extension of nonparametric variance estimation to two-stage cluster sampling. Bootstrap technique is generally considered as a nonparametric procedure. We suggest and study theoretically a model-based bootstrap technique in chapter four. In chapter five we perform an empirical study of all the estimators developed in chapter three and four. We conclude in chapter six with a summary and some suggestions for future directions for research.
Chapter 2

ROBUST ESTIMATION OF THE ERROR VARIANCE IN UNISTAGE SAMPLING

2.1 Introduction

In this chapter we consider the problem of robust estimation of the error variance of the popular ratio estimator under unistage sampling. Past attempts to this problem have presupposed an existence of a superpopulation model which is assumed to have generated the finite population. For estimating the second component of (1.2), Royall and Cumberland (1978), for example, assumed that \( \text{Var}(Y) = W\sigma^2 \) (see (1.4)). It is known that such a model can never be correct, but only represents an approximation, a working hypothesis.

The alternative approach is to suppose that no prior knowledge exists of the functional
form of $\text{Var}(Y)$ i.e. the structure of $\text{Var}(Y)$ is less marked and trial and testing prove fruitless; or suppose that the functional form of $\text{Var}(Y)$ changes within a region of interest. How does one proceed to estimate the error variance of $\hat{T}$ in these situations?

In the sequel we follow nonparametric procedure in estimating the error variance. The term nonparametric will thus be used to refer to the flexible functional form of $\text{Var}(Y)$. There are other notions of 'nonparametric statistics' which refer mostly to distribution-free methods. In the present context, neither the error distribution nor the functional form of the variance is prespecified. One of our great motivation for using this technique lies in its prediction potential. It gives prediction of observations yet to be made without reference to a fixed parametric model. The method is thus very flexible and potentially robust: properties which are yet to be exploited in finite population variance estimation.

Parallel work on distribution function may be found in Chambers, Dorfman, Wehrly (1992). Smith and Njenga (1992) have also employed a similar procedure in an attempt to overcome lack of robustness of estimates of the covariance matrix of survey variables using a general regression model. Recently Chambers, Dorfman, Wehrly (1993) and Dorfman (1994) employed the same procedure to estimate the finite population total, $T$.

One common finding of these authors is that the method may pay off in terms of increased robustness of the resulting estimator.

In section 2.3 we use this technique to estimate (1.2) in the special case when $q = 1$.

Section 2.2 introduces a working model and the error variance to be estimated. We consider variance estimation for the simple ratio estimator in section 2.4. In the context of a uniform kernel we discuss the analytical performances of these estimators in section 2.5.

We consider empirical performances of the variance estimators in section 2.6.
2.2 The Working Model and the Error Variance

We assume, without loss of generality, that the population labels are ordered by increasing $x$ so that $a < x_1 < x_2 < \ldots, < x_n \leq b \in s$ and $a < x_{n+1} < x_{n+2} < \ldots, < x_N \leq b \in r = U - s$, where $[a, b]$ is a closed interval of the real line. Having observed the sample (i.e. $s$) and all the prior values $x_i$'s, all the expectation, variance and covariance expressions will be understood to be conditional quantities with the $x_i$'s being the conditioning parameter.

For estimation of (1.2) we use the following model

$$E(Y_i) = \beta x_i$$

$$\text{Cov}(Y_i, Y_j) = \begin{cases} \sigma^2(x_i), & i = j \\ 0, & \text{otherwise} \end{cases}$$

The functional form of $\sigma^2(x_i)$ is unspecified although it is assumed to be Lipschitz continuous.

Unlike (1.1) this model leaves unspecified the structure of $\text{Var}_\xi(Y_i | x_i)$. It is hoped that starting with a general structure like this one, may result in an estimator which is more flexible to the varying specifications of the variance model. That is such an estimator may be robust.

If inferences are to be made conditional on the observed sample $s$, then the variance to be estimated is defined by

$$\text{Var}_\xi(\hat{T} - T) = E_\xi((\hat{T} - T) - E_\xi(\hat{T} - T))^2$$

If $\hat{T}$ is the Best Linear Unbiased Estimator (BLUE) of the population total $T$, then (2.2) reduces to

$$\text{Var}_\xi(\hat{T}_{\text{BLUE}} - T) = E_\xi(\hat{T}_{\text{BLUE}} - T)^2$$

28
Under model (2.1), $\hat{T}_{BLUE} = \sum_{s} y_{i} + \beta \sum_{r} x_{i}$, where $\hat{\beta} = \frac{\sum_{s} \frac{y_{i} x_{i}}{\sigma^{2}(x_{i})}}{\sum_{s} \frac{1}{\sigma^{2}(x_{i})}}$. Hence

\begin{equation}
\hat{T}_{BLUE} - T = \hat{\beta} \sum_{r} x_{i} - \sum_{r} y_{i}.
\end{equation}

From (2.4), the error variance of $\hat{T}_{BLUE}$ under the working model is

\begin{equation}
\text{Var}_{\xi}(\hat{T}_{BLUE} - T) = \frac{\sum_{r} x_{i}^{2}}{\sum_{s} \frac{1}{\sigma^{2}(x_{i})}} + \sum_{r} \sigma^{2}(x_{i}).
\end{equation}

This variance decreases as $\sum_{s} \frac{x_{i}^{2}}{\sigma^{2}(x_{i})}$ increases. Ideally (2.5) implies that if model (2.1) is correct, $\hat{T}_{BLUE}$ is optimal and the optimal sample consists of $n$ units whose $x_{i}$'s values are the largest and are homogenous. This homogeneity requirement can be achieved if, at the design stage, appropriate sampling techniques are employed. Stratified random sampling, for example, may give samples satisfying such requirements. However, if such a sample is used departures from model (2.1) may have disastrous effect on the point estimator $\hat{T}_{BLUE}$. For example, if $E(Y_{i}) = \beta_{0} + \beta_{1} x_{i}$, then

\begin{equation}
E_{\xi}(\hat{T}_{BLUE} - T) = \beta_{0} \left( \frac{\sum_{r} x_{i} \sum_{s} \frac{x_{i}}{\sigma^{2}(x_{i})}}{\sum_{s} \frac{1}{\sigma^{2}(x_{i})}} - N + n \right).
\end{equation}

This bias vanishes when

\[ \sum_{r} x_{i} = \left( \frac{\sum_{s} \frac{x_{i}^{2}}{\sigma^{2}(x_{i})}}{\sum_{s} \frac{1}{\sigma^{2}(x_{i})}} \right) (N - n) \]

i.e.

\[ \bar{x}_{r} = \frac{\sum_{s} \frac{x_{i}^{2}}{\sigma^{2}(x_{i})}}{\sum_{s} \frac{1}{\sigma^{2}(x_{i})}}. \]

When $\sigma^{2}(x_{i}) = \sigma x_{i}$, the bias becomes

\[ \frac{\beta_{0}(\bar{x}_{r} - \bar{x}_{s})}{\bar{x}_{s}} (N - n) \]

This result was obtained by Royall and Cumberland (1978) in the context of the the ratio estimator $\hat{T}_{R}$. 

29
We consider the problem of robust estimation of (2.5) in the presence of misspecification of the variance structure. Whereas the first moment $E_\xi(Y_i \mid x_i)$ can be modelled accurately, accurate modelling of $Var_\xi(Y_i \mid x_i)$ is difficult. Hence, not surprisingly, incorrect specification of $Var_\xi(Y_i \mid x_i)$ is by far the most common and severe form of misspecification in variance estimation in finite population sampling.

### 2.3 Nonparametric Estimation of the Error Variance

In this section nonparametric procedure for estimating the error variance is proposed. In particular we have proposed to estimate the error variance of $\hat{T}_{\text{BLUE}}$ using fixed bandwidth kernel smoothing techniques which, of late, have become quite popular in density and curve estimations. Our method is based on the simple fact that the squared residual

$$(y_i - \hat{\beta}x_i)^2$$

is approximately an unbiased estimator of $Var(y_i) = \sigma^2(x_i)$ (Horn, Horn, Duncan (1975), Royall and Cumberland (1978)). In fact

$$E_\xi[(y_i - \hat{\beta}x_i)^2] = \sigma^2(x_i) + O(n^{-1}) = \sigma^2(x_i) + \epsilon_i.$$ 

It follows that

$$\sigma^2(x_i) \approx (y_i - \hat{\beta}x_i)^2.$$ 

Thus estimation of the error variance can be achieved via the squared residuals. By exploiting the smoothing properties of the function $\sigma^2(x_i)$, local fits or local smoothing of the residuals might be used to get an unbiased or approximately unbiased estimator of the error variance.
Let 
\[ \hat{e}_i^2 = (y_i - \hat{\beta} x_i)^2 \]
be a naive estimator of \( \text{Var}(y_i) \) (i.e. \( \sigma^2(x_i) \)). To get a more improved estimate of \( \sigma^2(x_i) \), where \( i \in U \), we smooth
\[ \hat{e}_j^2 = (y_j - \hat{\beta} x_j)^2 \]
, \( j \in s \), and \( x_j, y_j \) are sample points close to \( x_i, y_i \) respectively. The closeness of \( x_j, y_j \) to \( x_i, y_i \) is measured by the distance \( |x_i - x_j| \).

Let \( W_h(x_i, x_j) \) be a linear smoothing function and \( h \) be a bandwidth parameter which specifies the amount of smoothing to be done. Applying this smoother to \( \hat{e}_j^2 \) we get an estimator of \( \sigma^2(x_i) \) as
\[ (2.6) \quad \sigma_{np}^2(x_i) = \sum_{j \in s} W_h(x_i, x_j) \hat{e}_j^2 \]
where \( i = 1, \ldots, N \) and \( j = 1, \ldots, n \). Hence a nonparametric estimator of the variance of \( \hat{T}_{BLUE} - T \) is given by
\[ (2.7) \quad V_{np}(\hat{T}_{BLUE} - T) = \frac{(\sum_{k \in r} z_k)^2}{\sum_{i \in r} \sigma_{np}^2(x_i)} + \sum_{i \in r} \sigma_{np}^2(x_i). \]

A number of smoothing functions have been suggested in the literature. An exposition of the properties of these functions is given in Silverman (1985). The best known and widely used of these functions are:

1. smoothing splines (Wahba (1975)),

2. \( K - \text{Nearest Neighbour}(K - N.N) \) (Mack (1981)),

3. kernel smoothers with subtypes (a) Priestley Chao type (Priestley and Chao (1972), Gasser and Muller (1979)) (b) Nadaraya-Watson type (Nadaraya (1964), Watson (1964)).
None of these smoothing functions is uniformly best. Kernel smoothers, however, have been found to have optimal minimax properties (Gasser and Engel (1990)). Consequently, in this study we focus on the kernel smoothing functions of the Nadaraya-Watson type and the Priestley-Chao type. Both of these smoothing functions are associated with the kernel function:

\[ K \left( \frac{x_i - x_j}{h} \right). \]

More specifically we will employ the following two kernel smoothers:

1. the Priestley-Chao (PC) weight represented by

\[ W_h(x_i, x_j) = \left( \frac{x_j - x_{j-1}}{h} \right) K \left( \frac{x_i - X_j}{h} \right), \]

2. the Nadaraya-Watson (NW) weight represented by

\[ W_h(x_i, x_j) = \frac{K \left( \frac{x_i - X_j}{h} \right)}{\sum_{j \in s} K \left( \frac{x_i - X_j}{h} \right)}, \]

The PC weight was suggested by Priestley and Chao (1972) while NW was independently suggested by Nadaraya (1964) and Watson (1964) and later on discussed by others in the context of density and curve estimations.

Substituting \( W_h(x_i, x_j) \) in (2.7) with (2.8) gives

\[ \sigma^2_{PC}(x_i) = \sum_{j \in s} \left( \frac{x_j - x_{j-1}}{h} \right) K \left( \frac{x_i - X_j}{h} \right) \bar{e}_j^2, \]

while an estimator of \( \sigma^2(x_i) \) based on NW weights is given by

\[ \sigma^2_{NW}(x_i) = \sum_{j \in s} \frac{K \left( \frac{x_i - X_j}{h} \right)}{\sum_{j \in s} K \left( \frac{x_i - X_j}{h} \right) \bar{e}_j^2}. \]
Hence two kernel-based estimators of the variance of $T_{\text{BLUE}} - T$ are

\begin{align*}
V_{PC} & = \frac{(\sum_{i \in r} x_i)^2}{\sum_{i \in s} \sigma_{PC}^2(x_i)} + \sum_{i \in r} \sigma_{PC}^2(x_i) \\
V_{NW} & = \frac{(\sum_{i \in r} x_i)^2}{\sum_{i \in s} \sigma_{NW}^2(x_i)} + \sum_{i \in r} \sigma_{NW}^2(x_i).
\end{align*}

We note the following:

1. As in curve estimation, the kernel weighing procedure is to average the nearby values of $\hat{e}_i^2 = (y_i - \hat{\beta}x_i)^2$ which is then used to estimate $\sigma^2(x_i)$. Here 'nearby' is measured in terms of the distance $|x_i - x_j|$. 

2. The Priestley-Chao estimate can be viewed as a weighted average of the observed $\hat{e}_j^2, j \in s$. Unlike the usual weighting situations, the sum does not equal to one, but is only an approximation:

\[ \sum_{j \in s} \frac{x_j - x_{j-1}}{h} K\left(\frac{x_i - X_i}{h}\right) \approx \int_{-\infty}^{\infty} K(u)du = 1. \]

3. The Nadaraya-Watson estimate, however, has this desired property (i.e. the sum of weights is exactly equal to 1).

One further difference between the weights is that the Priestley-Chao weights are only applicable to cases in which the auxiliary is restricted to some interval (usually $(0, 1]$), while the Nadaraya-Watson is applicable to all situations. Hence if one is to employ the Priestley-Chao weight in the circumstances under which the data are not restricted to a particular interval then it is necessary that the auxiliary data are transformed to the required interval before any analysis is implemented.
2.3.1 Application to the ratio estimator

The above ideas are now employed to the particular case of the ratio estimator:

\[ \hat{T}_R = \frac{\mathbf{\bar{y}}}{\mathbf{\bar{x}}} \sum_{i=1}^{N} x_i = \hat{R} \sum_{i=1}^{N} x_i . \]

This estimator is BLUE under model (2.1) when \( \sigma^2(x_i) = \sigma^2 x_i \). This is a simple regression model through the origin. In practical situations the model is used in situations where the expected value of \( Y \) is proportional to \( x \), and the variance of \( Y \) is also proportional to \( x \). In sample surveys such models are used especially in analysis of proportions. Under (2.1), \( E_\xi(\hat{T}_R - T) = 0 \). That is \( \hat{T}_R \) is unbiased. Its error variance is given by

\[ \text{Var}_\xi(\hat{T}_R - T) = \left( \frac{\sum_{i \in s} x_i}{\sum_{i \in s} x_i} \right)^2 \sum_{i \in s} \sigma^2(x_i) + \sum_{i \in r} \sigma^2(x_i) \]

where \( \text{Var}_\xi(\cdot) \) denotes the error variance of \( \hat{T}_R \). Under some mild conditions, the two components of (2.12) are of order \( O\left( \frac{N^2}{n} \right) \) and \( O(N - n) \) respectively. Some researchers have used these observations to concentrate on the robust estimation of the first component only and neglecting the other component as insignificant (e.g. Valliant (1987)). Others such as Royall and Eberhardt (1975) and Royall and Cumberland (1978) estimate the first component using a more general prediction framework, but since the second component is asymptotically insignificant, they use a prediction model which attaches great parametric restriction to the variance structure.

For the ratio estimator \( \hat{T}_R \), the squared residual is defined by

\[ r_i^2 = (y_i - \hat{R}x_i)^2 . \]

The model expectation of (2.13) under (2.1) is

\[ E_\xi(r_i^2) = \sigma^2(x_i) \left( 1 - \frac{2x_i}{nx_s} + \frac{x_i^2}{(nx_s)^2} \sum_{i \in s} \sigma^2(x_i) \right) . \]
Clearly this is equal to $\sigma^2(x_i) + O(n^{-1})$ under some mild conditions. This implies that it is possible to obtain an asymptotically robust estimate of the error variance given in (2.12) via the squared residuals defined in (2.13). From (2.8) and (2.9) we obtain two estimators of $\sigma^2(x_i)$ as

$$\sigma^2_{pc}(x_i) = \sum_{j \in s} \left( \frac{x_j - x_{j-1}}{h} \right) K \left( \frac{x_i - X_j}{h} \right) r_j^2$$

and

$$\sigma^2_{nw}(x_i) = \sum_{j \in s} \sum_{\ell \in r} K \left( \frac{x_i - X_j}{h} \right) r_j^2.$$ 

Thus the proposed estimators of the error variance of the ratio estimator are

$$V_{pc} = \frac{\left( \sum_{i \in r} x_i \right)^2}{\left( \sum_{i \in r} x_i \right)^2} \sum_{i \in s} \sigma^2_{pc}(x_i) + \sum_{i \in r} \sigma^2_{pc}(x_i)$$

and

$$V_{nw} = \frac{\left( \sum_{i \in r} x_i \right)^2}{\left( \sum_{i \in r} x_i \right)^2} \sum_{i \in s} \sigma^2_{nw}(x_i) + \sum_{i \in r} \sigma^2_{nw}(x_i).$$

### 2.3.2 The Asymptotic mean Square error and the bandwidth choice for the variance estimators

For studying the asymptotic properties of these estimators the following assumptions will prove useful:

1. $\sigma^2(x_i)$ is 2 times continuously differentiable.

2. $\sigma^2(x_i) \geq c_0 > 0$ for all $x_i \in [a, b]$

3. $|\sigma^2(x_i) - \sigma^2(x_j)| \leq L|x_i - x_j|^\theta$ for some $\theta \in (0, 1)$ L is a lipschitz constant

4. $K(u) = 0$ for all $|u| \geq 1$

5. $K(u)$ is an even function.
\[|k(u)-k(v)| \leq M|u-v|^{\alpha}, \quad u, v \in (0,1) \text{ for some } \alpha \in (0,1),\]

M is a constant.

\[\int_{-\infty}^{\infty} k(u) \, du = 1, \quad \int_{-\infty}^{\infty} u k(u) \, du = 0, \quad \int_{-\infty}^{\infty} u^i k(u) \, du \neq 0, \quad \text{for } i=2\]

\[E_{\xi}(Y_j^4) = \mu_4(x_j) < \infty.\]

In the following theorem, we let

\[E_{\xi}(Y_j^2) = \mu_2(x_j), \quad a=0, \quad b=1\]

**Theorem 1.** If in addition to the above assumptions, \(x_j\)'s, \((j \in \mathbb{S})\) form an everywhere dense set in \([0,1]\) whose elements are equispaced then \(V_{nw}\) and \(V_{pc}\) are consistent if \(h \to 0, \, nh \to \infty\) as \(n \to \infty\).

**Proof:** We will prove the two conditions:

(i) asymptotic unbiasedness of the variances of these variance estimators.

(ii) that the limiting variances of these variance estimators vanish when the sample size become large.

**Asymptotic Unbiasedness of** \(V_{pc}\)

\[
\lim_{n \to \infty} E_{\xi}(V_{pc}) = \lim_{n \to \infty} \left\{ \sum_{i \in r} x_i^2 \left( \frac{1}{\sum_{i \in s} x_i} \right)^2 \sum_{i \in s} E_{\xi} \sigma_{pc}^2(x_i) + \sum_{i \in r} E_{\xi} \left( \sigma_{pc}^2(x_i) \right) \right\}
\]
But
\[
E_\xi \left( \sigma^2 \left( x_1 \right) \right) = \sum_{j \in s} \left( \frac{x_j - x_{j-1}}{h} \right) k \left( \frac{x_j - x_{j-1}}{h} \right) E_\xi \left( r_j^2 \right).
\]

It is known that
\[
E_\xi \left( r_j^2 \right) = \sigma^2 (x_j) + O \left( n^{-1} \right)
\]
subject to some mild conditions (see Royall and Cumberland(1978)).

Hence
\[
E_\xi \left( \sigma^2 \left( x_1 \right) \right) = \frac{1}{h} \sum_{j \in s} \left\{ \int_{x_{j-1}}^{x_j} k \left( \frac{x_j - x_{j-1}}{h} \right) \sigma^2 (x) \, dt \right\}
\]

\[
= \frac{1}{h} \sum_{j \in s} \left\{ \int_{x_{j-1}}^{x_j} \left[ k \left( \frac{x_j - x_{j-1}}{h} \right) + k \left( \frac{x_j - x_{j-1}}{h} \right) \right] \sigma^2 (t) \, dt \right\}
\]

\[
+ \int_{x_{j-1}}^{x_j} k \left( \frac{x_j - x_{j-1}}{h} \right) \sigma^2 (t) \, dt + \int_{x_{j-1}}^{x_j} k \left( \frac{x_j - x_{j-1}}{h} \right) \left[ \sigma^2 (x_j) - \sigma^2 (t) \right] \, dt.
\]

That is
\[
\left| E_\xi \left( \sigma^2 \left( x_1 \right) \right) \right| \leq \frac{1}{h} \sum_{j \in s} \int_{x_{j-1}}^{x_j} k \left( \frac{x_j - x_{j-1}}{h} \right) \sigma^2 (t) \, dt
\]

\[
\leq \frac{M}{h} \sum_{j \in s} \int_{x_{j-1}}^{x_j} \left[ \frac{x_j - x_{j-1}}{h} \right] \sigma^2 (t) \, dt
\]

\[
+ \frac{L}{h} \sum_{j \in s} \int_{x_{j-1}}^{x_j} \left[ \frac{x_j - x_{j-1}}{h} \right] \sigma^2 (t) \, dt.
\]

\[
s \sup \left\{ \sigma^2 (x_j) \right\} \sum_{j \in s} \left( \frac{x_j - x_{j-1}}{h} \right)^{\alpha+1} + \frac{k(u)}{(\beta+1) h} \sum_{j \in s} \frac{1}{\alpha+1}
\]

\[
\left( u = \frac{x_j - x_{j-1}}{h} \right).
\]
Clearly, the RHs goes to zero if \( n h \to \infty \), as \( n \to \infty \).

Hence

\[
E \left[ \sigma_{pc}^2(x_i) \right] = \frac{1}{h} \sum_{j \in S} \int_{x_{i-1}}^{x_i} k \left( \frac{x}{h} - t \right) \sigma^2(t) \, dt .
\]

Now on expanding \( \sigma^2(t) \) about \( x_i \), using Taylor's expansion, we obtain

\[
E \left[ \sigma_{pc}^2(x_i) \right] = \sigma^2(x_i) + \frac{d^2}{dx_i^2} \left[ \sigma^2(x_i) \right] \frac{h^2}{2}
\]

where

\[
d_k = \int_{x_{i-1}}^{x_i} u^2 k(u) \, du \quad u = \frac{(x_i - t)}{h} .
\]

Hence the bias of \( \hat{V}_{pc} \) is

\[
B_i(\hat{V}_{pc}) = \xi_{pc} = \xi_{pc} \cdot \mu_{pc} - \text{Var}_{\xi_{pc}} (T - \hat{T})
\]

\[
= \frac{h^2}{2} \left[ \sum_{i \in R} \frac{x_i}{X} \right]^2 \sum_{i \in S} \frac{d^2}{dx_i^2} \left[ \sigma^2(x_i) \right] + \frac{h^2}{2} \sum_{i \in R} \frac{d^2}{dx_i^2} \left[ \sigma^2(x_i) \right].
\]

It is now easily seen that in estimating the population mean, \( \hat{V}_{pc} \) will give a relative bias of
\[
\frac{V_{pc}}{\text{Var}_{\xi}(T_r - T)} = h^2 \frac{d^2}{2} \left\{ \sum_{i \in \varepsilon} \frac{dx_i^2}{\sum_{i \in \varepsilon} \sigma_i^2} \right\} \left\{ \sum_{i \in \varepsilon} \frac{x_i^2}{\sum_{i \in \varepsilon} \sigma_i^2} \right\} - 2
\]

From this, we conclude that \( \text{E}(V_{pc}) = \text{Var}_{\xi}(T_r - T) \)
when \( h \to 0 \), as \( n \to \infty \).

Next, we consider limiting Variance of \( V_{pc} \).

Now

\[
\text{Var}_{\xi}(V_{pc}) = \left( \sum_{s \in \varepsilon} \sum_{i \in \varepsilon} \text{cov}(\sigma_i^2(x_i), \sigma_{pc}^2(x_k)) \right) + 2 \left( \sum_{i \in \varepsilon} \sum_{k \in \varepsilon} \text{cov}(\sigma_i^2(x_i), \sigma_{pc}^2(x_k)) \right)
\]

Also

\[
\text{cov}(r_j^2, r_j^2) = \text{cov}(y_j^2 - 2Rx_j, y_j^2 - 2Rx_j, y_j^2 - 2Rx_j, y_j^2 - 2Rx_j, y_j^2 - 2Rx_j, y_j^2 - 2Rx_j, y_j^2 - 2Rx_j, y_j^2 - 2Rx_j)
\]
and

\[
\text{Cov}(y_k y_{k'}, y_{v'} y_v) = \begin{cases} 
\text{cov}(y_k^2, y_{k'}^2) E_x(y_k) & \text{if } v = u = k, k \neq L \\
\text{var}(y_v y_v) & \text{if } v = k, u = L, L = k, v \neq u \\
\text{cov}(y_k^2, y^2 L) E_x(y) & \text{if } v = k = L, u \neq k \\
\text{var}(y_v^2) & \text{if } v = k = u = L \\
0 & \text{otherwise}
\end{cases}
\]

Using the foregoing expression i.e. \(\text{cov}(y_k y_{k'}, y_{v'} y_v)\), it can be shown that for all \(j \neq L\), expressions:

\[
\text{cov}(y_j^2, y_{L}^2), \text{cov}(y_j^2, y_{r}^2), \text{cov}(y_{L}^2, y_{L}^2), \text{cov}(y_{r}^2, y_{r}^2)
\]

and \(\text{cov}(y_j^2, y_{L}^2)\) are all of order \(O\left(\frac{1}{n}\right)\).

Next, with the linearity of the covariance (in its components) we get, for \(j \neq L\), that

\[\text{cov}(r_j^2, r_L^2) = O\left(\frac{1}{n}\right); \quad (n \to \infty).\]

For \(j = L\), we get

\[
\text{cov}(r_j^2, r_L^2) = \text{cov}(y_j^2, y_{L}^2) + 4x_j^2 \text{cov}(y_{L}^2, y_{L}^2) - 4x_j \text{cov}(y_j^2, y_{L}^2) + O\left(\frac{1}{n}\right)
\]

(because the expressions \(\text{cov}(y_j^2, R^2), \text{cov}(y^2, R^2)\) and \(\text{cov}(R^2, R^2)\) are all of order \(O\left(\frac{1}{n}\right), (n \to \infty).\) )

Thus

\[
\text{cov}(r_j^2, r_L^2) = \text{var}(r_j^2) = \text{var}(y_j^2) - \frac{4x_j^2}{(\sum x_i^2)^2} \sum_{i \in s, k \in s} \text{cov}(y_{L}^2, y_{L}^2) - \frac{4x_j^2}{(\sum x_i^2)^2} \sum_{i \in s, k \neq L} \text{cov}(y_{L}^2, y_{L}^2)
\]

40
\[
- \frac{4x_L^2}{(\sum x_i^2)_{i \in S}} \sum_{i \in S} \text{cov}(y_i^2, y_i^2) + O(\frac{1}{n}) \quad ; \quad (n \to \infty)
\]

Now
\[
\frac{4x_L^2}{n \times S} \sum_{i \in S} \text{cov}(y_i^2, y_i^2) = -4\beta x_L \text{cov}(y_L, y_L^2)
\]
and
\[
\frac{4x_L^2}{(\sum x_i^2)_{i \in S} \sum_{i, k \in S} \text{cov}(y_i, y_k, y_k) = 4\beta^2 x_L^2 \text{var}(y_L^2)
\]

Therefore
\[
\text{var}(r^2) = \text{var}(y_L^2) - 4x_L \beta \text{cov}(y_L, y_L^2)
\]

\[
+ 4\beta^2 x_L^2 \text{var}(y_L)
\]

\[
= \text{var}(y_L^2 - 2\beta y_L)
\]

\[
= \text{var}(y_L^2)[1 + O(1)]
\]

\[
= [\mu_4(x_L) - \mu_2^2(x_L)] [1 + O(1)]
\]

where \(\mu_2(x) = E (y_L^2 | x_L = x)\).

Hence
\[
| \text{cov}(\sigma_{pc}^2 (x_i), \sigma_{pc}^2 (x_j)) = \sum_{i \in S} [\mu_4(x_L) - \mu_2^2(x_L)] [1 + O(1)] \left( \frac{x_i - x_j}{h} \right) \left( \frac{x_j - x_i}{h} \right) \left( \frac{k - x_j}{nh} \right) | \leq O(\frac{1}{n}) \quad , \quad (n \to \infty)
\]
i.e.

\[
\text{cov}(\sigma^2_{pc}(x), \sigma^2_{pc}(x)) = \sum_{i \in E \setminus S} [\mu(x) - \mu(x)] \{1 + O(1)\} x
\]

\[
\frac{1}{nh} \left( \frac{x_{j} - x_{j'}}{h} \right) k \left( \frac{x_{j} - x_{j'}}{h} \right) k \left( \frac{x_{j} - x_{j'}}{h} \right)
\]

\[
= \frac{1}{nh} \left\{ \mu(x) - \mu(x) \right\} \{1 + O(1)\} \int_{0}^{1} k \left( \frac{x_{j} - x_{j'}}{h} \right) k \left( \frac{x_{j} - x_{j'}}{h} \right) dt
\]

\[
= \frac{1}{nh} \left\{ \mu(x) - \mu(x) \right\} \{1 + O(1)\} \int_{0}^{1} k(u) k \left( \frac{x_{j} - x_{j'}}{h} \right) du = V_{ik}.
\]

Thus

\[
\lim_{n \to \infty} \text{Var} \left\{ V_{pc} \right\} = \sum_{i, k \in E \setminus S} V_{ik} \left( \frac{x_i}{s_i} \right)^4 + \sum_{i \in E \setminus S} \sum_{k \in E \setminus S} V_{ik} \left( \frac{x_i}{s_i} \right)^2
\]

Hence

\[
\lim_{n \to \infty} \frac{V_{pc}}{\text{var} (\hat{T}_R - T)} = \sum_{i, k \in E \setminus S} V_{ik} \left( \frac{x_i}{s_i} \right)^2 + \sum_{i \in E \setminus S} \sum_{k \in E \setminus S} V_{ik} \left( \frac{x_i}{s_i} \right)^2
\]

\[
\sum_{i, k \in E \setminus S} V_{ik} \left( \frac{x_i}{s_i} \right)^4 + \sum_{i \in E \setminus S} \sum_{k \in E \setminus S} V_{ik} \left( \frac{x_i}{s_i} \right)^2
\]

i.e.

the asymptotic mean square error (AMSE) of

\[
\text{AMSE} \left( \hat{T}_R - T \right) = \frac{V_{pc}}{\text{Var} (\hat{T}_R - T)}
\]

is

42
Indeed, AMSE \( (\cdot) \) goes to zero if the conditions of the theorem hold. This implies that

\[
\lim_{n \to \infty} \text{Var}_\xi \left( V_{pc} \right) \to 0
\]

if \( nh \to \infty \) as \( n \to \infty \). Thus the limiting variance of \( V_{pc} \) vanishes when the conditions of the theorem hold.

Next, we derive the AMSE of \( V_{nw} \).

We first note that

\[
\sigma^2_{nw}(x_i) = \frac{1}{nh} \sum_{j \in s} k \left( \frac{x_i - x_j}{h} \right)^2 - \left( \frac{1}{nh} \sum_{j \in s} k \left( \frac{x_i - x_j}{h} \right) \right)^2
\]

\[
= \frac{1}{nh} \sum_{j \in s} k \left( \frac{x_i - x_j}{h} \right)^2 ;
\]
since, as \( n \to \infty \),
\[
\frac{1}{nh} \sum_{j \in S} k \left( \frac{x_i - x_j}{h} \right) = \sum_{j \in S} \int_{x_{j-1}}^{x_j} k \left( \frac{x - x_j}{h} \right) \frac{dt}{h}
\]
\[
= \int_0^1 k(u) \, du = 1.
\]

Hence
\[
E \left[ \sigma^2_{nw}(x_i) \right] = \sum_{j \in S} \int_{x_{j-1}}^{x_j} k \left( \frac{x_i - x_j}{h} \right) \sigma^2(x_j) \frac{dt}{h}
\]
\[
= \sum_{j \in S} \int_{x_{j-1}}^{x_j} k \left( \frac{x_i - x_j}{h} \right) \sigma^2(x_j) \frac{dt}{h}
\]
\[
= \sum_{j \in S} \int_{x_{j-1}}^{x_j} k \left( \frac{x - t}{h} \right) \sigma^2(t) \, dt.
\]
This expression had been shown previously to be approximately equal to
\[
\sigma^2(x_i) + \frac{d}{dx_i} \left( \frac{h}{2} \right) + \frac{d^2}{dx_i^2} \left( \sigma^2(x_i) \right) \quad (\text{See page 37}).
\]

Therefore,
\[
E \left[ \sigma^2_{nw}(x_i) \right] = \sigma^2(x_i) + \frac{d}{dx_i} \left( \frac{h}{2} \right) + \frac{d^2}{dx_i^2} \left( \sigma^2(x_i) \right).
\]
Also
\[
\text{Cov} \left[ \sigma^2_{nw}(x_i), \sigma^2_{nw}(x_j) \right] = \sum_{j \in S} \sum_{k \in S} \frac{1}{2h^2} k \left( \frac{x_i - x_j}{h} \right) k \left( \frac{x_k - x_j}{h} \right) \text{Cov}(r_i^2, r_j^2)
\]
\[
= \frac{1}{nh} \sum_{j \in S} \int_{x_{j-1}}^{x_j} k \left( \frac{x - t}{h} \right) k \left( \frac{x_k - t}{h} \right) \text{cov}(r_i^2, r_j^2) + O \left( \frac{1}{n} \right)
\]
We now conclude that

\[
S_E'(V_{nw}) = \left[ \sum_i \frac{x_i}{s_i} \right]^2 \frac{s_i^2(\sigma^2(x_i))}{\sum_i \sigma^2(x_i)} + \frac{h^2}{2} \frac{d^2}{dx^2} \left[ \sigma^2(x) \right] + \sum_{i \in r} \left[ \sigma^2(x_i) + \frac{h^2}{2} \frac{d^2}{dx^2} \left[ \sigma^2(x) \right] \right]
\]

and

\[
\text{Var}_\xi(V_{nw}) = \left( \sum_i \frac{x_i}{s_i} \right)^4 \sum_{i,k \in s} V_{ik} + 2 \left( \sum_{i \in r} \frac{x_i}{s_i} \right)^2 \sum_{i \in s} \sum_{k \in r} V_{ik} + \sum_{i,k \in r} V_{ik}
\]

Hence the AMSE of \( V_{nw} \) is AMSE \( \left( \frac{V_{nw}}{\text{Var}_\xi(T_R - T)} \right) = \text{AMSE} \left( \frac{V_{nw}}{\text{Var}_\xi(T_R - T)} \right) \)

\[
\frac{1}{nh} \left[ \sum_{i,k \in s} \frac{V_{ik}}{\sigma^2(x_i) \sigma^2(x_k)} \right] + \frac{1}{4} \left[ \sum_{k,i \in s} \frac{d^2}{dx^2} \left[ \sigma^2(x) \right] \frac{d^2}{dx^2} \left[ \sigma^2(x_k) \right] \right]
\]

Clearly AMSE \( \left( \frac{V_{nw}}{\text{Var}_\xi(T_R - T)} \right) \to 0 \) if \( h \to 0 \), \( nh \to \infty \)

45
as \( n \to \infty \). This result illustrates among other consequences, that

AMSE of \( V_{nw} \) vanishes whenever \( h \to 0, nh \to \infty \) and \( n \to \infty \). Further, on comparing the AMSE of \( V_{pc} \) and \( V_{nw} \), it is noted that the two expressions are identical. An important implication of this result is that under the asymptotic conditions of the theorem, \( V_{nw} \) and \( V_{pc} \) are equivalent.

Let

\[
C = \left[ \frac{\sum_{i, k \in S} V_{ik}}{\sum_{i, k \in S} \sigma^2(x_i) \sigma^2(x_k)} \right]
\]

and

\[
B = \left( \frac{d_k}{2} \right)^2 \sum_{i, k \in \mathcal{R}} \frac{d^2}{dx_i^2} \left( \frac{\sigma^2(x_i)}{\sigma^2} \right) \frac{d^2}{dx_k^2} \left( \frac{\sigma^2(x_k)}{\sigma^2}\right)
\]

Then

\[
\text{AMSE} \left[ \frac{V_{nw}}{\text{Var}_\xi(T_{R})} \right] = \text{AMSE} \left[ \frac{V_{pc}}{\text{Var}_\xi(T_{R})} \right]
\]

\[
\text{AMSE} \left[ \frac{V_{nw}}{\text{Var}_\xi(T_{R})} \right] = \frac{h^4 B}{4} + \frac{C}{nh}
\]

Hence an optimal bandwidth that minimizes AMSE (\( \cdot \)) is

\[
h_{\text{opt}} = \left[ \frac{C}{nB} \right]^{\frac{1}{5}}.
\]
It is now easy to conclude that the AMSE of 
\[
\frac{V_{nw}}{\text{Var}_{z}(\hat{T}_R - T)} \quad \text{and} \quad \frac{V_{pc}}{\text{Var}_{z}(\hat{T}_R - T)}
\]
cannot be smaller than 
\[
\frac{AMSE_{opt}}{\text{Var}_{z}(\hat{T}_R - T)} V_{nw} = AMSE_{opt} \frac{V_{pc}}{\text{Var}_{z}(\hat{T}_R - T)} = 0 \left( (nh)^{-1} \right)
\]
Following Silverman (1986), \( h \sim n^{-1/5} \). Thus the optimal local rate of convergence of the variance estimators is \( O(n^{-4/5}) \). It is interesting to compare this rate with that of the parametric estimators which is usually \( O(n^{-1}) \). Remarks: we note that though the nonparametric estimators are bias robust, they are less efficient than estimators based on a correctly specified variance model.

2.4 Alternative Estimators of the Error Variance

In the above paragraph we have seen that nonparametric method gives an estimator of the error variance which is less efficient than estimator based on the parametric model. However a criticism against the direct method of variance estimation is that it estimates the second component of (2.12) assuming that (1.1) is correct. In practice the model may be misspecified, or it may be known that the model changes quite often within a region of interest. For these cases it would be more appropriate to employ estimation procedures that employ general model given in (2.1) to estimate the second component of (2.12).

In particular, to capture both the strong efficiency of the direct variance estimation procedure and the good bias robustness property of the nonparametric procedure, we are proposing to estimate (2.12) by the following two procedures:

1. estimate the first component or (2.12) using Royall and Cumberland (1978) and then estimate the second component using the nonparametric method.

2. estimate the first component using Chew (1970)'s procedure and then estimate the second component using nonparametric procedure.
Recall that the ratio estimator is BLUE when

$$\sigma^2(x_i) = \sigma^2 x_i.$$ 

Hence

$$E(r_j^2) = \sigma^2 x_j (1 - k_j)$$

where $k_j = \frac{x_j}{n \bar{x}_s}$. Clearly $r_j^2$ is a biased estimator of $\sigma^2(x_j)$. Royall and Cumberland’s method consists essentially of correcting for this potential bias arising from the working model. Thus an unbiased estimator of $Var(Y_j | x_j)$ when $\sigma^2(x_j) = \sigma^2 x_j$ is $r_j^2(1 - k_j)^{-1}$.

Hence an estimator of the error variance based on the Royall and Cumberland (1978) and nonparametric procedures is

$$V_{r,n} = \left(\frac{(N - n)\bar{x}_r}{n\bar{x}_s}\right)^2 \sum_{j \in s} r_j^2 (1 - k_j)^{-1} + \sum_{i \in r} \sum_{j \in s} W_h(x_i, x_j) r_j^2.$$ 

The second method based on Chew (1970)’s procedure also makes use of the residuals.

Note

$$E(\xi(r_j^2)) = \sigma^2(x_j)(1 - 2k_j) + \frac{k_j^2}{\sigma^2(x_j)} \sum_{j \in s} \sigma^2(x_j)).$$

This expectation is taken via model (2.1). Dropping the expectation operator we get

$$r_j^2 = \sigma^2(x_j)(1 - 2k_j) + \frac{k_j^2}{\sigma^2(x_j)} \sum_{j \in s} \sigma^2(x_j)).$$

Clearly there will be $n$ equations and $n$ unknown $\sigma^2(x_j)$,

and hence the equation can be solved. After solving this system of equations it can be shown that a new estimator based on Chew and nonparametric procedures is

$$V_{ch,n} = \left(\frac{(N - n)\bar{x}_r}{(n\bar{x}_s)^2(1 + \Delta_s)}\right)^2 \sum_{j \in s} r_j^2 (1 - 2k_j)^{-1} + \sum_{i \in r} \sum_{j \in s} W_h(x_i, x_j) r_j^2$$

where

$$\Delta_s = \sum_{j \in s} \frac{x_j^2}{(1 - 2k_j)(n \bar{x}_s)^2}.$$
In the next section performances of the four proposed estimators (i.e. $V_{nw}$, $V_{pc}$, $V_{ch,n}$, $V_{r,n}$) are derived and compared. We also include the following robust estimators:

$$V_{ch} = \frac{N(N-n)\bar{X}_r}{(n\bar{x}_s)^2(1 + \Delta_s)} \sum_{j \in s} r_j^2(1 - 2k_j)^{-1}$$

and

$$V_D = \frac{N(N-n)\bar{X}_r}{(n\bar{x}_s)^2} \sum_{j \in s} r_j^2(1 - k_j)^{-1}$$

The estimator $V_D$, an estimator based on Royall and Cumberland (1978)'s procedure, is equivalent (and in many cases superior) to many current estimators of (2.11) which are favoured in practice. $V_{ch}$ is an estimator that was derived by Royall and Cumberland (1978), by employing the covariance matrix estimation procedure proposed by Chew (1970). Unlike $V_D$, $V_{ch}$ has not been studied extensively. A comparative study employing this estimator alongside other estimators was first given by Wafula (1988). Unlike all the other estimators of (2.11), Chew's procedure estimates the first component of (2.11) strictly unbiasedly under any specification of the variance model. We note, however, that though it has good bias robustness properties this estimator can take (assume) a negative value whenever $2x_i > n\bar{x}_s$ for all units in the sample. This, indeed, is the greatest impediment in the application of this procedure.

### 2.5 Analytical Performances of the Variance Estimators

The analytical performances of the variance estimators have been assessed using their biases. In deriving the biases it is assumed, without loss of generality, that the kernel function $K(.)$ corresponds to the uniform kernel and that the bandwidth $h$, is large. In
particular we have considered the kernel:

\[
K\left(\frac{x_i - x_j}{h}\right) = \begin{cases} 
\frac{1}{2} & \text{if } |\frac{x_i - x_j}{h}| \leq 1 \\
0 & \text{otherwise}
\end{cases}
\]

For any variance estimator, \(V\), we define and denote its bias under model (2.1) as:

\[
B_{\xi}(V) = \mathbb{E}_{\xi}^{(v)}\text{Var}_{\xi}(\hat{T}_R - T).
\]

The biases of the variance estimators are

\[
B_{\xi}(V_{ch}) = \frac{(N - n)\bar{x}_r}{n\bar{x}_s^2} \sum_{j \in s} \sigma^2(x_j) - \sum_{j \in r} \sigma^2(x_j)
\]

\[
B_{\xi}(V_D) = \frac{(N - n)\bar{x}_r}{n\bar{x}_s^2} \sum_{j \in s} \sigma^2(x_j) - \sum_{j \in r} \sigma^2(x_j)
\]

\[
+ \frac{N(N - n)\bar{x}_r^2}{n^2\bar{x}_s^2} \left( \sum_{j \in s} \frac{k_j^2 \sum_{j \in s} \sigma^2(x_j) - k_j\sigma^2(x_j)}{1 - k_j} \right)
\]

\[
B_{\xi}(V_{r,n}) = \frac{(N - n)}{n} \sum_{j \in s} \sigma^2(x_j) - \sum_{j \in r} \sigma^2(x_j)
\]

\[
+ \left( \frac{(N - n)^2\bar{x}_r^2}{n^2\bar{x}_s^2} \right) \left( \sum_{j \in s} \frac{k_j^2 \sum_{j \in s} \sigma^2(x_j) - k_j\sigma^2(x_j)}{1 - k_j} \right)
\]

\[
- \frac{(N - n)}{n} \left( \sum_{j \in s} k_j\sigma^2(x_j) \right) - k^2 \sum_{j \in s} \sigma^2(x_j)
\]

\[
B_{\xi}(V_{nw}) = \frac{(N - n)}{n} \sum_{j \in s} \sigma^2(x_j) - \sum_{j \in r} \sigma^2(x_j)
\]

\[
+ \frac{(N - n)}{n} \left( \frac{(N - n)\bar{x}_r^2}{n\bar{x}_s^2} + 1 \right) \left( \sum_{j \in s} [k_j^2 \sum_{j \in s} \sigma^2(x_j) - k_j\sigma^2(x_j)] \right)
\]

\[
- \frac{(N - n)}{n} \left( \frac{(N - n)\bar{x}_r^2}{n\bar{x}_s^2} + 1 \right) \sum_{j \in s} k_j\sigma^2(x_j)
\]

\[
B_{\xi}(V_{ch,n}) = \frac{(N - n)}{n} \sum_{j \in s} \sigma^2(x_j) - \sum_{j \in r} \sigma^2(x_j)
\]

\[
+ \frac{(N - n)}{n} \left( \sum_{j \in s} [k_j^2 \sum_{j \in s} \sigma^2(x_j) - k_j\sigma^2(x_j)] \right)
\]

50
The above bias expressions indicate that under stable conditions, i.e. if both the sample and the population averages remain constant as both the sample and the population become large, then

\[
| B_\xi(V_{ch}) | \leq | B_\xi(V_{ch,n}) | \leq | B_\xi(V_{nw}) | = | B_\xi(V_{rn}) | \leq | B_\xi(V_D) |
\]

The behaviour of \( V_{pc} \) is similar to that of \( V_{nw} \) when \( h = \frac{1}{2} \). One further observation to make is that in the open interval (0,1), \( k_i = O(n^{-1}) \). This means that for large samples, the quantity \( k_i \approx 0 \). In this case the biases are

\[
B_\xi(V_{ch}) = \frac{(N-n)\bar{x}_r}{n\bar{x}_s} \sum_{j \in s} \sigma^2(x_j) - \sum_{j \in r} \sigma^2(x_j)
\]

\[
B_\xi(V_D) = \frac{(N-n)\bar{x}_r}{n\bar{x}_s} \sum_{j \in s} \sigma^2(x_j) - \sum_{j \in r} \sigma^2(x_j)
\]

\[
B_\xi(V_{rn}) = \frac{N-n}{n} \sum_{j \in s} \sigma^2(x_j) - \sum_{j \in r} \sigma^2(x_j)
\]

\[
B_\xi(V_{nw}) = \frac{N-n}{n} \sum_{j \in s} \sigma^2(x_j) - \sum_{j \in r} \sigma^2(x_j)
\]

\[
B_\xi(V_{ch,n}) = \frac{(N-n)}{n} \sum_{j \in s} \sigma^2(x_j) - \sum_{j \in r} \sigma^2(x_j)
\]

\[
B_\xi(V_{pc}) = \frac{(N-n)}{2hn} \sum_{j \in s} \sigma^2(x_j) - \sum_{j \in r} \sigma^2(x_j)
\]

\[
+ \left( \frac{1}{2h} - 1 \right) \left( \frac{(N-n)\bar{x}_r}{n\bar{x}_s} \right)^2 \sum_{j \in s} \sigma^2(x_j)
\]
Thus under stable conditions

$$| B_\xi(V_{ch}) | = | B_\xi(V_{ch.n}) | = | B_\xi(V_{nw}) | = | B_\xi(V_{rn}) | = | B_\xi(V_D) |$$

The above observations have important consequences in model based inferences. First they imply that under stable conditions, some of the new estimators are better than the estimator favoured in the current practice. Infact, at the lower tail (i.e. $\bar{x}_r \geq \bar{x}_s$, which are the most common in practice) the new estimators may out perform $V_{ch}$ and $V_D$. Another notable consequence is that in large samples the estimators $V_D, V_{r.n}, V_{nw}, V_{ch.n}, V_{ch}$ are asymptotically equivalent under some conditions. We remarked in our literature review that the estimator $V_D$ is bias robust in large samples. Hence if one can find an estimator which is asymptotically equivalent to it, then one expects such an estimator to have good bias robustness properties. We are encouraged to note that, with exception of $V_{pc}$, our new estimators not only march $V_D$ under balanced large samples, but can also outperform this estimator under stable conditions. Finally the performance of the estimator $V_{pc}$ will depend very much on the bandwidth parameter, $h$. However, we note that if this parameter is chosen wisely, perhaps by using a certain optimality criterion, then $V_{pc}$ may be equally good as the other estimators.

**special cases**

(i) $\sigma^2(x_1) = \sigma^2 x_1$. The ratio estimator is generally studied when $\sigma^2(x_i) = \sigma^2 x_i$. Under this variance model $\hat{T}_R$ is BLUE. The exact biases of the variance estimators in this case are

- $B_\xi(V_{ch}) = B_\xi(V_D) = 0$
- $B_\xi(V_{r,n}) = (N - n)(\bar{x}_s - \bar{x}_r)\sigma^2 - \frac{(N - n)\bar{x}_s^2}{nx_s} \sigma^2$
- $B_\xi(V_{ch,n}) = (N - n)(\bar{x}_s - \bar{x}_r)\sigma^2 - \frac{(N - n)\bar{x}_s^2}{nx_s} \sigma^2$
A number of interesting results are revealed by these analyses. First, as expected, the estimators $V_{ch}$, $V_{D}$ are strictly unbiased when $\sigma^2(x_i) = \sigma^2 x_i$. Secondly, the estimators $V_{ch,n}$, $V_{r,n}$ have equal biases. The similarity in performances of $V_{ch}$, $V_{D}$ is a consequence of the asymptotic equivalence of these two estimators. This is also true for $V_{ch,n}$, $V_{r,n}$.

Finally under stable conditions, the above results indicate that

$$B_\xi(V_{ch}) = B_\xi(V_{D}) = 0 < B_\xi(V_{ch,n}) = B_\xi(V_{r,n}) < B_\xi(V_{nw})$$

That is the new estimators are all biased when the working variance model is $\sigma^2(x_i) = \sigma^2 x_i$, while the estimators favoured in the current practice are unbiased. This trend should not be a surprise. It is well recognized that estimators based on kernel methods are essentially biased. Although model (2.1) is quite general, it is not compatible with the model for which $\sigma^2(x_i) = \sigma^2 x_i$. Hence the new estimators are essentially biased when the working variance model is $\sigma^2(x_i) = \sigma^2 x_i$. An immediate consequence of this is that it does not pay, at least from the unbiasedness criterion, to employ the new estimators under this model.

\[ B_\xi(V_{nw}) = (N - n)(\bar{x}_s - \bar{x}_r)\sigma^2 - \frac{(N - n)(\bar{x}_s')^2}{n\bar{x}_s} - \frac{((N - n)^2\bar{x}_r^2)}{n^2\bar{x}_s^2}\]

\[ B_\xi(V_{pc}) = (N - n)\left(\frac{\bar{x}_s}{2h} - \bar{x}_r\right)\sigma^2 - \frac{(N - n)(\bar{x}_s')^2}{2hn\bar{x}_s} - \frac{((N - n)^2\bar{x}_r^2)}{2hn^2\bar{x}_s^2} + \left(\frac{1}{2h} - 1\right) \left(\frac{(N - n)(\bar{x}_r')^2}{n\bar{x}_s}\right)\]

where

$$\bar{x}_s^{(i)} = \frac{\sum_{j \in s} x_j^i}{n}.$$
We note, however, that our motivation for suggesting these alternative estimators arose from robustness concerns. How robust are these new estimators vis-a-vis the estimators favoured in the current practice? We have seen that the new estimators are biased when the working variance model is \( \sigma^2(x_i) = \sigma^2 x_i \). In studying the bias robustness of these estimators this potential bias will be carried over to the case of misspecified model.

Thus to assess the bias robustness of these estimators we will have to 'bias correct' them under the working model. That is we first remove the bias resulting from the the working model and then proceed to evaluate the robustness of these estimators when there is a model misspecification.

As an illustration we consider the following case, which is a slight deviation from the one we have just considered.

(ii) \( \sigma^2(x_1) = a + \sigma^2 x_1 \): In this case the biases are

\[
B_{\xi}(V_{ch}) = (N - n) \left( \frac{\bar{x}_r - \bar{x}_s}{\bar{x}_s} \right)a
\]

\[
B_{\xi}(V_D) = (N - n) \left( \frac{\bar{x}_r - \bar{x}_s}{\bar{x}_s} \right)a
\]

\[
+ \frac{N(N - n)\bar{x}_r}{n^2\bar{x}_s^2} \left( \frac{\bar{x}_s^{(2)}}{\bar{x}_s^2} - 1 \right) + \frac{\bar{x}_s^{(3)}}{n\bar{x}_s^3} - \frac{\bar{x}_s^{(2)}}{n\bar{x}_s^2} \left( \frac{\bar{x}_s^{(4)}}{n^2\bar{x}_s^4} - \frac{\bar{x}_s^{(3)}}{n^2\bar{x}_s^3} \right) + \ldots \ldots \ldots a
\]

\[
B_{\xi}(V_{r,n}) = (N - n)(\bar{x}_s - \bar{x}_r)\sigma^2 - \frac{(N - n)\bar{x}_s^{(2)}}{n\bar{x}_s} \sigma^2
\]

\[
+ \left( \frac{N - n}{n} \right) \left( \frac{\bar{x}_s^{(2)}}{\bar{x}_s^2} - 2n \right) a
\]

\[
+ \left( \frac{N - n}{n} \right)^2 \left( \frac{\bar{x}_s^{(2)}}{\bar{x}_s^2} - 1 \right) + \frac{\bar{x}_s^{(3)}}{n\bar{x}_s^3} - \frac{\bar{x}_s^{(2)}}{n\bar{x}_s^2} \left( \frac{\bar{x}_s^{(4)}}{n^2\bar{x}_s^4} - \frac{\bar{x}_s^{(3)}}{n^2\bar{x}_s^3} \right) + \ldots \ldots \ldots a
\]

\[
B_{\xi}(V_{ch,n}) = (N - n)(\bar{x}_s - \bar{x}_r)\sigma^2 - \frac{(N - n)\bar{x}_s^{(2)}}{n\bar{x}_s} \sigma^2
\]

\[
+ \left( \frac{N - n}{n} \right) \left( \frac{\bar{x}_s^{(2)}}{\bar{x}_s^2} - 2 \right) a
\]

54
\[ B_\xi(V_{nu}) = (N - n)(\bar{x}_s - \bar{x}_r)\sigma^2 - \frac{(N - n)x_s^{(2)}}{n\bar{x}_s} \sigma^2 \]
\[ - \left( \frac{(N - n)^2x_r^2}{n^2\bar{x}_s^2} \right) \frac{x_s^2}{x_s} \sigma^2 \]
\[ + \left( \frac{N - n}{n} \right) \left( \frac{(N - n)x_r^2}{n\bar{x}_s^2} + 1 \right) \left( \frac{x_s^{(2)}}{\bar{x}_s^2} - 2 \right) a \]

\[ B_\xi(V_{pc}) = (N - n)(\frac{\bar{x}_s}{2h} - \bar{x}_r)\sigma^2 - \frac{(N - n)x_s^{(2)}}{2hn\bar{x}_s} \sigma^2 \]
\[ + \left( \frac{1}{2h} - 1 \right) \left( \frac{(N - n)x_r^2}{n\bar{x}_s^2} \right) \sigma^2 \]
\[ + \left( \frac{N - n}{2hn} \right) \left( \frac{(N - n)x_r^2}{n\bar{x}_s^2} + 1 \right) \frac{x_s^2}{\bar{x}_s} - \sigma^2 \]
\[ + \left( \frac{N - n}{2h} \right) \left( \frac{(N - n)x_r^2}{n\bar{x}_s^2} + 1 \right) \left( \frac{x_s^{(2)}}{\bar{x}_s^2} - 2 \right) a \]
\[ (N - n)(\frac{1}{2h} - 1) \left\{ \frac{N - n}{n} \frac{x_r^2}{\bar{x}_s^2} + 1 \right\} a \]
\[ + \frac{N - n}{2hn} \left\{ \frac{N - n}{n} \frac{x_r^2}{\bar{x}_s^2} + 1 \right\} a \]

Clearly the estimators \( V_{ch} \) and \( V_D \) are no longer unbiased, as was in the first case. The bias expressions of the new estimators consist of the biases under the working model considered in case (i) and those emerging from the variance model misspecification. On subtracting this working model bias, one obtains the following 'bias corrected' biases:

\[ B_t(V_{ch}) = (N - n) \left( \frac{\bar{x}_r - \bar{x}_s}{\bar{x}_s} \right) a \]

\[ B_t(V_{RD}) = (N - n) \left( \frac{\bar{x}_r - \bar{x}_s}{\bar{x}_s} \right) a \]
\[ + \frac{N(N - n)\bar{x}_r}{n^2\bar{x}_s^2} \left( \frac{x_s^{(2)}}{\bar{x}_s^2} - 1 \right) + \frac{x_s^{(3)}}{n\bar{x}_s^3} - \frac{x_s^{(2)}}{n\bar{x}_s^2} + \frac{x_s^{(4)}}{n^2\bar{x}_s^4} - \frac{x_s^{(3)}}{n^2\bar{x}_s^3} + \ldots \ldots, a \]

\[ B_t(V_{rn}) = + \left( \frac{N - n}{n} \right) \left( \frac{x_s^{(2)}}{\bar{x}_s^2} - 2n + 1 \right) a \]
\[ + \left( \frac{N - n}{n} \right)^2 \left( \frac{x_s^{(2)}}{\bar{x}_s^2} - 2 \right) + \frac{x_s^{(3)}}{n\bar{x}_s^3} - \frac{x_s^{(2)}}{n\bar{x}_s^2} + \frac{x_s^{(4)}}{n^2\bar{x}_s^4} - \frac{x_s^{(3)}}{n^2\bar{x}_s^3} + \ldots \ldots, a \]

\[ B_t(V_{ch,n}) = + \left( \frac{N - n}{n} \right) \left( \frac{x_s^{(2)}}{\bar{x}_s^2} - 2 \right) a \]

55
Under stable conditions, it is clear that

\[ B_{\xi}(V_{n\omega}) = \frac{N-n}{n} \left( \frac{N-n}{nx_s^2} + 1 \right) \left( \frac{\overline{x}_s^{(2)}}{\overline{x}_s^2} - 2 \right) a \]

\[ B_{\xi}(V_{pc}) = \frac{2hn}{(N-n)} \left( \frac{N-n}{nx_s^2} + 1 \right) \left( \frac{\overline{x}_s^{(2)}}{\overline{x}_s^2} - 2 \right) a \]

\[ + (N-n) \left( \frac{1}{2hn} - 1 \right) \left( \frac{(N-n)x_r^2}{n^2x_s^2} + 1 \right) a \]

These results indicate that the move from the variance model considered in case (i) to another slightly different model has a more significant impact on \( V_D \) than it has on the other estimators. The results reveal that \( V_D \) is a very unstable estimator: slight deviation from the working variance model will influence it greatly. An immediate consequence from this observation is that this estimator is not appropriate when there is a misspecification in the working model.

We also observe the following points:

1. Although \( V_{ch} \) and \( V_D \) are strictly unbiased when \( \sigma^2(x_i) = \sigma^2 x_i \), this is not the case when the model is misspecified. In this case \( V_{ch} \) performs far better than \( V_D \), indicating that it is more bias robust than \( V_D \).

2. The crossbreed estimators \( V_{ch.n}, V_{r.n} \) are respectively more bias robust than \( V_{ch}, V_D \). This indicates that using nonparametric method to estimate the less dominant component of the error variance has a significant gain over the usual parametric estimation technique.

3. Using balanced samples may give notable gain in the robustness of all the estimators. In this case \( V_{ch} \) will be strictly unbiased, while the other estimators
may have less bias (under some mild conditions) than their biases under unbalanced samples.

4. \( V_{pc} \) will have similar robustness properties as \( V_{nw} \) when \( h \approx 0.5 \). There is some hope, therefore, that a wise choice of the bandwidth parameter may lead to appreciable bias robustness from this estimator under uniform kernel.

5. As is common with estimators based on local smoothing, it is noted that the new estimators will be more biased and less bias robust if the estimation is being done at the boundary. There are, however, kernels meant for these edge problems, and if need be can be employed to provide robust versions of these new estimators.

6. The estimator \( V_{ch} \) may take a negative value. This is also true of \( V_{ch,n} \), although with a less frequency. Hence in practical applications these estimators may not give expected good results. Consequently for applications where there is a suspected variance model misspecification we recommend \( V_{nw} \) and \( V_{r,n} \) as estimators of choice.

2.6 Empirical Study

The properties of the six variance estimators are studied in nine populations. The natural populations are described in table 1. The artificial populations were generated as follows: a finite population of 500 values of \( x_i \)'s were generated by random sampling from a uniform \((0,5)\) distribution. For each value of \( x_i \) a corresponding value of an error term, \( e_i \), was obtained from \( N(0,1) \) distribution. The survey values for the first artificial population (AP1) were obtained using the formula

\[
y_i = 100x_i + \sqrt{x_i}e_i; \quad i = 1, \ldots, 500
\]
<table>
<thead>
<tr>
<th>Population</th>
<th>Source</th>
<th>y</th>
<th>x</th>
<th>N</th>
<th>(\rho(x,y))</th>
<th>(\rho_y)</th>
<th>(\rho_z)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1989 UN Report book</td>
<td>Expenditures in hospital per day</td>
<td>Number of beds in a hospital</td>
<td>38.00</td>
<td>0.40</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>Daily Nation of 17.7.93</td>
<td>Exchange rate on 17.7.93</td>
<td>Exchange rate on 17.7.92</td>
<td>34</td>
<td>0.75</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>3</td>
<td>Kenya statistical abstract 1977</td>
<td>Number of people employed in towns,1975</td>
<td>Number of people employed in towns,1970</td>
<td>65</td>
<td>0.91</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>4</td>
<td>Kenya statistical abstract 1977</td>
<td>No.of births and deaths in compulsory registration area,1977</td>
<td>No.of births and deaths in compulsory registration area,1970</td>
<td>74</td>
<td>0.92</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>5</td>
<td>Kenya statistical abstract 1977</td>
<td>Number of people employed in industries,1976</td>
<td>Number of people employed in industries,1973</td>
<td>159</td>
<td>0.93</td>
<td>0.11</td>
<td>0.10</td>
</tr>
<tr>
<td>6</td>
<td>Kenya statistical abstract 1976</td>
<td>Size of imported articles in 1975</td>
<td>NSize of imported articles in 1970</td>
<td>94</td>
<td>0.97</td>
<td>0.20</td>
<td>0.23</td>
</tr>
<tr>
<td>7</td>
<td>Montgomery (1985) Regional sales</td>
<td>Regional expenditure</td>
<td>20</td>
<td>0.99</td>
<td>0.60</td>
<td>0.55</td>
<td></td>
</tr>
</tbody>
</table>
while the survey values for the second population \((AP2)\) were obtained using the formula

\[ y_i = 100x_i + x_ie_i; \quad i = 1, \ldots, 500. \]

We note that the parameter values for \((AP1)\) are not very different from those in the model in which the ratio estimator is known to perform well. Observe that for this population

\[ E(Y_i \mid X_i = x_i) = 100x_i; \quad \text{Var}(Y_i \mid X_i = x_i) = x_i. \]

\(AP2\) is an example of variance model mispecification. It's parameters are

\[ E(Y_i \mid X_i = x_i) = 100x_i; \quad \text{Var}(Y_i \mid X_i = x_i) = x_i^2. \]

The general method used was to draw 650 simple random samples with \(n = 32\) for populations 3, 4, 5, 6 and \(n = 8\) for populations 1, 2, 7. For \(AP2\) and \(AP1\), 1000 simple random samples of size \(n = 50\) were drawn.

To ensure that \(V_{ch}\) did not take negative values, we rejected sample points for which

\[ 2x_i \geq n\bar{x}. \]

A Gaussian Kernel,

\[ K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \]

was used in the study. The bandwidth parameter employed in the study was that which minimized the mean square error (MSE) and satisfied Silverman\((1986)\)'s condition:

\[ \frac{\sigma}{4n^{1.5}} \leq h \leq \frac{3\sigma}{2n^{1.5}}. \]

Since Priestley Chao weights are generally employed in a fixed interval, an interval \([0,1]\) for the \(x\) values was used.

For each population, we calculated the population total: \(T = \sum_{i=1}^{N} y_i\), the total for the \(x\)
values $X = \sum_{i=1}^{N} x_i$. Further for each sample, we calculated the ratio estimate $\hat{T}_R$; the prediction error $E = \hat{T}_R - T$; the variance estimates $V_D, V_{ch}, V_{ch.n}, V_{nw}, V_{r.n}, V_{pc}$; the $t$ statistic: $t = \frac{E}{\sqrt{V}}$ for all the variance estimators.

The unconditional behaviour of the variance estimators were studied by taking the average of the corresponding quantities among all the samples: $MSE = \frac{\sum E^2}{G_p}$ (where $G_p$ is the total number of samples taken from a given population, $\sum$ indicates sum over all the the $G_p$ samples drawn from the given population); the bias of the variance estimators were computed as

$$B(V) = \frac{\sum V}{G_p} - MSE$$

The root mean square error for the variance estimators were computed as

$$RMSE(V) = \left\{ \frac{\sum (V - MSE)^2}{G_p} \right\}^{\frac{1}{2}}$$

Also for each sample and each variance estimator we calculated the interval

$$\left( T_R - t_{\frac{a}{2}(n-1)}\sqrt{V}, T_R + t_{1-\frac{a}{2}(n-1)}\sqrt{V} \right)$$

where $t_{a(r)}$ is the $100a$ percent point of the student $t$ distribution with $r$ degrees of freedom.

The coverage probability associated with each variance estimator was then determined by calculating the percentage of the $G_p$ $t$-statistics that satisfy $|t| \leq 1.96$ for populations $AP1, AP2, 3, 4, 5, 6$ and $|t| \leq 2.365$ for the remaining populations. Summary results are given in tables 2, 3, 4.
2.6.1 Results for real populations

**root mean square error (RMSE)**

1. The estimators $V_{ch}$ and $V_{nw}$ are the best or nearly the best in terms of minimizing $MSE$. That is, for planning future surveys, analysis based on these estimators would predict outcome of future surveys better than the other estimators.

2. Between $V_{r,n}$ and $V_{ch,n}$ there does not exist a clear winner. The $RMSE$ associated with these estimators are very close to each other.

3. The estimators $V_{pc}$ and $V_{D}$ are the worst in terms of minimizing $RMSE$. The poor performance of $V_{D}$ was also noted by Wu and Deng (1983).

**Table 2** Unconditional Values of Root Mean Square Error (RMSE) $\div 1000$

<table>
<thead>
<tr>
<th>Variance</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>AP1</th>
<th>AP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{pc}$</td>
<td>831.37</td>
<td>264.54</td>
<td>738.50</td>
<td>215.14</td>
<td>381.99</td>
<td>762.00</td>
<td>147.05</td>
<td>254.40</td>
<td>594.00</td>
</tr>
<tr>
<td>$V_{ch}$</td>
<td>170.35</td>
<td>7.86</td>
<td>21.50</td>
<td>47.82</td>
<td>105.36</td>
<td>39.04</td>
<td>39.28</td>
<td>35.00</td>
<td>45.80</td>
</tr>
<tr>
<td>$V_{D}$</td>
<td>204.33</td>
<td>20.60</td>
<td>66.70</td>
<td>68.14</td>
<td>128.44</td>
<td>574.8</td>
<td>24.30</td>
<td>52.1</td>
<td>60.0</td>
</tr>
<tr>
<td>$V_{r,n}$</td>
<td>198.34</td>
<td>9.77</td>
<td>65.56</td>
<td>58.60</td>
<td>121.41</td>
<td>530.17</td>
<td>21.02</td>
<td>62.10</td>
<td>56.00</td>
</tr>
<tr>
<td>$V_{ch,n}$</td>
<td>195.01</td>
<td>9.41</td>
<td>59.90</td>
<td>59.77</td>
<td>124.26</td>
<td>551.30</td>
<td>22.37</td>
<td>63.1</td>
<td>67.00</td>
</tr>
<tr>
<td>$V_{nw}$</td>
<td>181.76</td>
<td>9.19</td>
<td>56.04</td>
<td>52.03</td>
<td>109.83</td>
<td>540.73</td>
<td>19.22</td>
<td>64.15</td>
<td>68.00</td>
</tr>
</tbody>
</table>

**Bias**

1. With the exception to $V_{pc}$ (which always overestimates $MSE$) the other estimators either overestimate or underestimate $MSE$. 

61
2. There does not exist an outright winner among the estimators considered here. However, the estimator $V_D$ is, in many cases, an estimator of choice whenever it underestimates $MSE$. For other cases $V_{ch,n}$ and $V_{r,n}$ are, in most cases, the best estimators. The most significant result here is that the three estimators: $V_D, V_{ch,n}, V_{r,n}$, are generally the best for these populations, indicating that most of these populations follow the superpopulation structure that is optimal for the ratio estimation.

3. As is expected from theory, $V_{r,n}$ appears to be a compromise between the estimators $V_D$ and $V_{nw}$. The bias of this estimator is always sandwiched between those of $V_D$ and $V_{nw}$.

Table 3  Unconditional Values of the Bias of the Variance Estimators $\times$ 1000

<table>
<thead>
<tr>
<th>Variance Estimator</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>AP1</th>
<th>AP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{pc}$</td>
<td>59.55</td>
<td>171.57</td>
<td>170.30</td>
<td>107.86</td>
<td>160.85</td>
<td>342.30</td>
<td>132.80</td>
<td>100.50</td>
<td>20.41</td>
</tr>
<tr>
<td>$V_{ch}$</td>
<td>-93.41</td>
<td>4.87</td>
<td>24.45</td>
<td>-67.08</td>
<td>-103.31</td>
<td>-310.50</td>
<td>-34.03</td>
<td>74.00</td>
<td>10.60</td>
</tr>
<tr>
<td>$V_D$</td>
<td>16.81</td>
<td>4.81</td>
<td>-3.45</td>
<td>2.29</td>
<td>-6.38</td>
<td>-22.62</td>
<td>14.33</td>
<td>0.52</td>
<td>10.90</td>
</tr>
<tr>
<td>$V_{r,n}$</td>
<td>8.03</td>
<td>1.26</td>
<td>3.95</td>
<td>-1.74</td>
<td>-8.74</td>
<td>-32.15</td>
<td>8.17</td>
<td>2.50</td>
<td>9.50</td>
</tr>
<tr>
<td>$V_{ch,n}$</td>
<td>-2.14</td>
<td>0.87</td>
<td>-5.42</td>
<td>-6.17</td>
<td>-6.49</td>
<td>-28.40</td>
<td>6.48</td>
<td>2.70</td>
<td>8.40</td>
</tr>
<tr>
<td>$V_{nw}$</td>
<td>-10.92</td>
<td>3.41</td>
<td>1.26</td>
<td>-8.59</td>
<td>-22.87</td>
<td>-124.61</td>
<td>7.10</td>
<td>4.00</td>
<td>12.54</td>
</tr>
</tbody>
</table>
4. The estimator $V_{pc}$ and $V_{ch}$ are the worst, with $|\text{bias}_{V_{ch}}| < |\text{bias}_{V_{pc}}|$. The poor performance of $V_{ch}$ is a surprise. In theory this estimator is supposed to give very good results. It appears that rejection of some points (i.e. those for which $2x_i \geq n\bar{x}_s$) make this estimator very susceptible to bias and hence inefficient.

**coverage probability**

1. With exception to $V_{pc}$, the coverage probabilities are all less than the nominal value. The result for $V_{pc}$ can be attributed to the conservative values of its bias.

2. There is no clear estimator of choice under this criterion. More generally, for most of these populations, we have the order $V_D, V_{r,n}, V_{ch,n} > V_{ch} > V_{pc}$ (where $>$ indicates 'covers the interval better than'). That is $V_D, V_{r,n}, V_{ch,n}$ are the best estimators with none of them being a clear winner. The estimator $V_{nw}$ has also performed remarkably well and is a close competitor to the winning estimators. In fact, all the four estimators seem to have almost equal confidence coefficients.

3. A result that stands out so clearly from the comparison of the $RMSE$ results and those of the coverage probabilities is that estimators like $V_{ch}$, $V_{nw}$ that perform well under $RMSE$ do not fare well for giving reliable confidence interval. Thus these estimators are not good for statistical inference purposes. We do emphasize, however, that some surveys are purely for planning future surveys and for these surveys estimators that minimise $RMSE$ are the best.
### Table 4  Unconditional Values of Coverage Probability of the 95% Confidence Interval

<table>
<thead>
<tr>
<th>Variance Estimator</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>AP1</th>
<th>AP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{pc}$</td>
<td>100.00</td>
<td>96.00</td>
<td>99.40</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>$V_{ch}$</td>
<td>84.70</td>
<td>70.00</td>
<td>17.00</td>
<td>29.20</td>
<td>30.00</td>
<td>10.70</td>
<td>83.70</td>
<td>60.00</td>
<td>63.00</td>
</tr>
<tr>
<td>$V_D$</td>
<td>91.50</td>
<td>87.40</td>
<td>90.60</td>
<td>87.90</td>
<td>85.50</td>
<td>89.70</td>
<td>95.50</td>
<td>95.00</td>
<td>92.00</td>
</tr>
<tr>
<td>$V_{r,n}$</td>
<td>91.20</td>
<td>87.50</td>
<td>91.20</td>
<td>87.60</td>
<td>85.00</td>
<td>89.70</td>
<td>95.70</td>
<td>93.60</td>
<td>93.00</td>
</tr>
<tr>
<td>$V_{ch,n}$</td>
<td>91.20</td>
<td>87.50</td>
<td>89.50</td>
<td>87.00</td>
<td>85.30</td>
<td>89.70</td>
<td>95.70</td>
<td>93.80</td>
<td>93.31</td>
</tr>
<tr>
<td>$V_{nw}$</td>
<td>90.20</td>
<td>86.90</td>
<td>86.90</td>
<td>86.00</td>
<td>85.00</td>
<td>89.50</td>
<td>95.20</td>
<td>92.36</td>
<td>92.41</td>
</tr>
</tbody>
</table>

#### 2.6.2 Results for artificial populations

The bias and root mean square error are given in tables 2 and 3. For $AP1$, it is evidently clear that the direct variance estimator, i.e. $V_D$, performs exceedingly better than all the other estimators. It is followed by the estimators $V_{r,n}$ and $V_{ch,n}$ in that order (although the difference in performances of these two estimators is very small). Another notable result is that the nonparametric estimators $V_{pc}$ and $V_{nw}$ lag far behind, depicting once again that nonparametric methods are not all that good as intuition would like us to believe. Our results indicate that estimators based on such methods may be highly biased (look at the performance of $V_{pc}$). It is also to be noted that the exact performances of these estimators is hard to envisage: they depend critically on the width of the smoothing window and the choice of such a width remains a problem to statisticians.

The good performance of $V_D$ is a consequence arising from the model used to generate $AP1$. For that model we indicated analytically that $V_D$ may be hard to beat! However the poor performance of $V_{ch}$ came as a surprise. It is expected (analytically) that this
...estimator should give very good results under the simple regression model, yet its empirical performances under criteria of bias and RMSE are the other way round.

A move to $AP2$ reverses the trend of the performances described above. In particular the estimator $V_D$ lags behind the new estimators $V_{ch,N}$ and $V_r,n$, showing the vulnerability of this estimator to variance model misspecification. This result came out so clearly in our analytical study. It was remarked in that section that though $V_D$ is the best estimator under the simple regression model, it generally gives poor results under variance model misspecifications. The good performances of the winning estimators were also inferred in the same section. Another result which is implicit in the table of biases is that a comparison of the bias results for $AP1$ and $AP2$ reveals that $V_{ch,n}, V_r,n$ and $V_{nw}$ are the most bias robust to the misspecification of the variance model. This again supports the conjecture that estimators based on the nonparametric methods are generally the best from robustness point of view.

Confidence interval coverage performances are given in table 4. The estimator $V_D$ can be seen to perform strongly in $AP1$, attaining a coverage equal to the nominal value. The new estimators $V_{ch,n}, V_r,n$ and $V_{nw}$ gave under coverages, but all the same gave quite impressive performances. The worst performer was $V_{ch}$, and this can be linked to its poor performance under bias criterion. Another point to note is that $V_{ch,n}$ and $V_r,n$ give almost similar performances.

The performances of the variance estimators in population $AP2$ reflect the same trend as those observed under the biases and RMSE. The coverage probabilities of most of the new estimators have not undergone substantial changes from their values in $AP2$. There is thus an illustration of distributional robustness of these estimators.
Chapter 3

NONPARAMETRIC VARIANCE ESTIMATION IN TWO STAGE CLUSTER SAMPLING

3.1 Introduction

Given a finite population, \( P \), organized in \( N \) identifiable clusters, each of size \( M_i \) units, we consider the variance estimation problem for two stage cluster sampling. A first stage sample, \( s \), of \( n \) clusters is obtained using simple random sampling without replacement (srs wor). If the \( i - th \) cluster is in \( s \), we obtain a second stage sample, \( s_i \), of size \( m_i \) subunits from its \( M_i \) units using srs wor. It is assumed that all the cluster sizes are known.
Let \( y_{ij}(i = 1, \ldots, N; j = 1, \ldots, M_i) \) be the value of characteristic of interest for the \( j \)-th subunit in the \( i \)-th cluster. When necessary it will be assumed that to each \( y_{ij} \) there exists some prior value, \( x_{ij} \), positively correlated with it. Then for estimating the population total

\[
Y = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}
\]

we consider a general class of estimators

\[
\hat{Y} = \frac{N}{n} \sum_{s} u_{i} \bar{y}_{i}
\]

where \( \bar{y}_{i} = M_{i} \bar{y}_{s_{i}}; \bar{y}_{s_{i}} = \frac{1}{m_{i}} \sum_{s} y_{ij} \) and \( \sum_{s_{i}} \sum_{s} \) indicate sums taken over the first stage and second stage sampling units respectively. The weights \( u_{i} \) are chosen so as to provide unbiased estimators of (3.1). That is, if \( \xi \) denotes the distribution of the working model, then

\[
E_{\xi}(\hat{Y}) = E_{\xi}(Y)
\]

where \( E_{\xi} \) is the expectation operator taken with respect to the considered superpopulation model. We assume throughout that \( u_{i} \)'s are independent of \( y_{ij} \)'s and that \( y_{ij} \)'s have a constant mean, \( \mu \). Hence (3.3) implies that \( \sum_{s} u_{i} M_{i} = n\bar{M} \), where \( \bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_{i} \).

The class of estimators defined in (3.2) includes most estimators commonly considered in two stage cluster surveys, e.g.,

\[
u_{i} = \frac{\bar{M}}{\bar{M}_{s}} \quad \text{(the ratio estimator where } \bar{M}_{s} = \frac{\sum M_{i}}{n})
\]

After a sample (i.e. \( s \)) has been selected the measurements \( y_{ij}(i \in s, j \in s_{i}) \) are known and the statistic \( \hat{Y} \) can easily be computed for a specified \( u_{i} \). As in unistage sampling a basic requirement is that a measure of precision be provided for the computed statistic. Such a measure is useful for statistical inference and for planning future surveys.

The most commonly used measure of precision is the variance of the computed statistic.
In general variances are not known and must be estimated from the available sample survey data.

In this chapter we extend the nonparametric variance estimation developed in chapter two to complex surveys. In particular we consider two stage cluster sampling and construct a nonparametric variance estimator. We follow the prediction approach to the estimation of the variance of the prediction error (i.e. the error variance). The organization of the subsequent sections of this chapter are as follows: we give the working model and derive the error variance in section 2. We also give a brief review of the available study. In section 3 we suggest an alternative estimator and discuss its properties.

### 3.2 The Working Model and the Error Variance

It is assumed that the current finite population comes from an infinite population with the variables $Y_{ij}$'s satisfying a linear model of the form

$$Y_{ij} = \mu + e_{ij}$$  \hspace{1cm} (3.4)

in which $\mu$ is a constant, $e_{ij}$'s are random errors with

$$\text{Cov}(e_{ij}, e_{kj}) = \begin{cases} \sigma_i^2 & \text{if } i = k, j = l \\ 0 & \text{otherwise} \end{cases}$$

This model asserts that variables in different clusters are uncorrelated while those within a cluster are identically and independently distributed.

Under (3.4) $\hat{Y}$ is $\xi$-unbiased and its prediction error is given as

$$\hat{Y} - Y = \frac{N}{n} \sum_s \hat{y}_i - \sum_s y_i - \sum_r y_i$$  \hspace{1cm} (3.5)
where $y_i = \sum_{j=1}^{M_i} y_{ij}$ and $r$ is the complement of $s$. The variance of (3.5) under (3.4) is

\begin{align*}
\text{Var}(\hat{Y} - Y) &= \frac{N^2}{n^2} \sum_{s} u_i^2 \text{var}(\hat{y}_i) - \frac{2N}{n} \sum_{s} u_i \text{cov}(y_i, \hat{y}_i) + \sum_{i=1}^{N} \text{var}(y_i) \\
&= \frac{N^2}{n^2} \sum_{s} u_i^2 \frac{M_i^2}{m_i} \sigma_i^2 - \frac{2N}{n} \sum_{s} u_i M_i \sigma_i^2 + \sum_{i=1}^{N} M_i \sigma_i^2.
\end{align*}

A key research concern in model-based surveys has been robust estimation of the error variance. To this end, Royall (1986) has suggested a robust estimator of (3.6). By estimating the three components of (3.6) unbiasedly under different superpopulation models, Royall (1986) gave the following estimator of the error variance

\begin{align*}
V_o &= \frac{N^2}{n^2} \sum_{s} u_i (u_i - f) \hat{v}_i + \frac{N}{n} \sum_{s} u_i M_i (1 - f_i) R_i^2 f_i^{-1} \\
&\quad + \left\{ \sum_{i=1}^{N} M_i^2 - \frac{N}{n} \sum_{s} u_i M_i^2 \right\} \theta
\end{align*}

where

\begin{align*}
f &= \frac{n}{N}, f_i = \frac{m_i}{M_i}, R_i^2 = (m_i - 1)^{-1} \sum_{s_i} (y_{ij} - \bar{y}_{s_i})^2, \\
\hat{v}_i &= \frac{(M_i M)^2 (1 - g_i)}{n^2} \left(1 + \frac{u_i^2 M_i^2}{M_i^2 n^2} (1 - g_i)\right)^{-1} A_i = \left(1 - g_i\right)^{-1}, \\
g_i &= 2 \frac{M_i u_i}{M}, r_i = \hat{y}_i - \frac{\hat{Y}}{NM} M_i
\end{align*}

and

$$\theta = \sum_{s} \{ \hat{v}_i - M_i R_i^2 f_i^{-1} \}.$$

One of the appealing properties of $V_o$ is that its leading term is strictly unbiased under a general model (1.5). Hence it is asymptotically robust. However, it has two significant
drawbacks. First it gives a negative estimate when \( g_i \) is not less than \( n \) for every \( i \in s \) (Chew, 1970). Secondly, when \( n \) is small this estimator can be very unstable. In particular when \( n = 2 \), \( \hat{V}_i \) is not defined so that \( V_o \) is indeterminate in this special case. Hence it is required that \( n > 2 \). Obviously this eliminates large scale complex surveys, especially stratified multistage design in which 2 clusters are sampled from each stratum.

This is a very common design in small area estimation which is a rapidly growing research topic.

We suggest a new estimator using nonparametric regression in the next section. The new estimator has the appealing properties of robustness, nonnegativity and is applicable even when \( n = 2 \).

3.3 Kernel Estimation of the Error Variance

The estimation of (3.6) is based on the squared residuals:

\[ r_i^2 = \left( \hat{y}_i - \frac{\hat{Y}}{NM_i} M_i \right)^2. \]

Let \( t_i = \frac{M_i}{M_G} \), where \( M_G \) is the largest cluster size in the population. Suppose that \( \text{Var}(\hat{y}_i) \) is a smooth function of \( t_i \). In particular let

\[ \text{Var}(\hat{y}_i) = \sigma^2(t_i). \]

Assuming further that \( n, N \to \infty, E(\hat{y}_i^4) \) and \( M_i \) remain finite, then it can be shown that

\[ E_\xi(r_i^2) = \sigma^2(t_i) + O(n^{-1}). \]

The expression (3.7) links the estimation of \( \sigma^2(t_i) \) in an obvious way to the nonparametric regression problem. One immediately sees the problem of estimation of \( \sigma^2(t_i) \) as that
of fitting a smooth curve through the sampled squared residuals. More specifically if we assume that \( t_k \)'s \( (k \in s) \) are regularly spaced in \([0, 1]\) we can define a linear smoothing function by

\[
W_h(t_i, t_k) = \frac{K\left(\frac{t_i - t_k}{h}\right)}{\sum_{k \in s} K\left(\frac{t_i - t_k}{h}\right)},
\]

where \( h \) is an appropriate smoothing parameter, and estimate \( \sigma^2(t_i) \) using

\[
V_i = \sum_{k \in s} \frac{K\left(\frac{t_i - t_k}{h}\right)}{\sum_{k \in s} K\left(\frac{t_i - t_k}{h}\right)} r_k^2
\]

where \( h \) is an appropriate smoothing parameter. A nonparametric estimator of the first component of (3.6) is

\[
N^2 \sum_{i \in s} u_i^2 V_i.
\]

It is important to note that the derivation of (3.8) has been done without specifying the covariance structure of the variables within a cluster. However, in estimating the remaining components of the error variance we insist that the working model applies. Under this model

\[
Cov(y_i, \hat{y}_i) = \frac{M_i}{m_i} \left[ \sum_{s_i} Cov(y_{ij}, y_{ij}) + \sum_{l=m_i+1}^{m_i} \sum_{j=1}^{m_i} Cov(y_{ij}, y_{il}) \right]
\]

\[
= M_i \sigma_i^2 = Var(y_i)
\]

so that an estimator of the second component is given by

\[
-\frac{2}{n} N \sum_{i \in s} u_i f_i V_i.
\]

while the third component is estimated by

\[
\sum_{i \in s} f_i V_i + \sum_{i \in r} V_i.
\]

Adding (3.8), (3.9) and (3.10) gives an estimator of the error variance as

\[
V_{NP2} = \sum_{i \in s} \left( \frac{N^2}{n^2} u_i^2 + f_i \right) V_i - \frac{2}{n} N \sum_{i \in s} u_i f_i V_i + \sum_{i \in r} V_i.
\]
An alternative estimator can be motivated by assuming that \( \sigma_i^2 = \sigma^2 \forall i = k, j = l \).

Suppose that for each \( Y_{ij} \) there exists some auxiliary variable \( X_{ij} \). Let \( x_{ij} \) be the observed value of \( X_{ij} \). Define residuals by

\[
\hat{e}_{ij} = y_{ij} - \sum_{j \in s_i} Z_{ij} y_{ij}
\]

where

\[
Z_{ij} = \frac{K \left( \frac{x_{ij} - x_{il}}{h} \right)}{\sum_{l \in s_i} K \left( \frac{x_{ij} - x_{il}}{h} \right)}.
\]

Let

\[
S_i^2 = \sum_{j \in s_i} \hat{e}_{ij}^2.
\]

Noting that

\[
\hat{e}_{ij}^2 = y_{ij}^2 - 2 \left[ Z_{ij} y_{ij} + \sum_{j' \in s_i, j' \neq j} Z_{ij} y_{ij} y_{ij'} \right] + \sum_{j \in s_i} Z_{ij}^2 y_{ij}^2 + \sum_{j \in s_i, j' \neq j} Z_{ij} Z_{ij'} y_{ij} y_{ij'},
\]

it follows that under the modified model

\[
E_\epsilon(S_i^2) = \sigma^2 \left[ m_i - 2 + m_i \sum_{j \in s_i} Z_{ij}^2 \right]
\]

so that a strictly unbiased estimator of the second and third components of (3.6) are respectively

\[
-2\frac{N}{n} \sum_{i \in s} M_i u_i V_i',
\]

and

\[
\sum_{i \in s} f_i V_i + \sum_{i \in r} M_i V_i'.
\]

where

\[
V_i' = \left[ m_i - 2 + m_i \sum_{j \in s_i} Z_{ij}^2 \right]^{-1} S_i^2.
\]
It is clear that in the special case of uniform weight:

\[
Z_{ij} = \begin{cases} 
\frac{1}{m_i} & \text{if } |x_{ij} - x_{i1}|/h_i \leq 1 \\
0 & \text{otherwise}
\end{cases}
\]

\[
V_i' = \frac{1}{m_i - 1} \sum_{j \in s_i} (y_{ij} - \bar{y}_{si})^2 = R_i^2
\]

which is the usual estimator for \( \text{Var}(y_{ij}) \) given in the classical text books. An advantage of \( V_i' \) over \( R_i^2 \) is that it is more flexible.

Adding (3.8), (3.10) and (3.11) gives an alternative estimator of the error variance as

\[
V_{NP2}' = \sum_{i \in s} \left( \frac{N^2}{n^2} u_i^2 + f_i \right) V_i - \frac{2N}{n} \sum_{i \in s} M_i u_i V_i' + \sum_{i \in r} M_i V_i'.
\]

**Remark:** The estimators \( V_{NP2}' \) and \( V_{NP2} \) are asymptotically equivalent. However, if they are compared in terms of the computing time, \( V_{NP2}' \) will need more time than \( V_{NP2} \). Hence \( V_{NP2}' \) may only be of theoretical interest while for application in finite samples, \( V_{NP2} \) may be preferable.

### 3.3.1 Asymptotic properties of the new variance estimator

We will assume that the finite population under study is a member of a sequence of finite populations \( \{P_k\}_{k=1}^{\infty} \), where the \( k \)-th population is of size \( N_k \) clusters such that \( N_k \geq N_{k-1} \). Let \( \{S_k\}_{k=1}^{\infty} \) be a sequence of corresponding samples where the \( k \)-th sample is of size \( n_k \) and \( n_k \geq n_{k-1} \). Further as \( k \to \infty \), let \( f_k = \frac{n_k}{N_k} \to f = 1 - \delta \), \( N_k > O(n_k^{3-\delta}) \), \( \delta > \frac{1}{2} \), \( r_k = N_k - n_k \to \infty \), \( M_{ki} = O(1) \), where \( M_{ki} \) is the size of the \( i \)-th cluster in the \( k \)-th population. Finally let as \( k \to \infty \) , sample and population averages converge to non-zero constants.

**Theorem 2** Suppose the above asymptotic conditions apply where \( \delta = 1 \) as \( k \to \infty \). Suppose further that
1. \(0 < E(y^2_k) \leq M < \infty\) for all \(k\).

2. The kernel function \(K(.)\) is bounded, symmetric on a support \([-1, 1]\) and is such that
   \[d_k = \int u^2 K(u) du < \infty\]

3. Suppose further that \(\sigma^2(t_i)\) has a bounded second derivative and is such that
   \[\sigma^2(t_i) \geq C_0 > 0 \quad \forall \ t_i \in [0, 1] \quad C_0 \text{ is a constant}.

Assume further that \(t_i\)'s (i \in s) are regularly spaced and let, without loss of generality,
\(0 \leq t_1 \leq t_2 \leq \ldots \leq t_n \leq 1\) for all \(t_i\)'s in the sample. If variables in different clusters are independent then

the asymptotic mean square error of \(\frac{V_{NP2}}{Var(Y - \bar{Y})}\) is

\[
MSE \left( \frac{V_{NP2}}{Var(Y - \bar{Y})} \right) \approx \frac{h^4}{4} \left[ \frac{d_k \sum_{i \in s} u_i^2 \sigma''(t_i)}{\sum_{i \in s} u_i^2 \sigma^2(t_i)} \right]^2 + \frac{1}{nh} \left[ \frac{\sum_{i \in s} \sum_{k \in s} u_i^2 u_k^2 (\mu_4(t_i) - \mu_2^2(t_i) \{1 + o(1)\}) \int K(u)K \left( \frac{u_k - t_i}{h} - u \right) du}{\sum_{i \in s} u_i^2 \sigma^2(t_i)} \right]^2
\]

\[
MSE \left( \frac{V_{NP2}}{Var(Y - \bar{Y})} \right) \to 0
\]

if \(h \to 0\), \(nh \to \infty\)

where

\[\mu_4(t_i) - \mu_2^2(t_i) = Var(\hat{g}_i^2)\]
Remark 1: When clusters are of unequal size, even though $M_1, \ldots, M_N$ are fixed and assumed known, in a strict probabilistic sense the $M_i$'s are really random variables whose realisations depend on which clusters are sampled. Hence the expressions involving the expectation, the variances and the covariances are really conditional statements for given values of $M_i$'s $(i \in s)$. Note that the results obtained here also apply to the unconditional setting.

Remark 2: The theorem applies to broad conditions that include those of the working model as a special case. The theorem thus indicates the bias robustness property of the proposed estimator under broad conditions within a cluster.

Remark 3: That the new estimator is asymptotically unbiased illustrates that it can be a strong competitor to $V_0$. Royall (1986) remarked that any estimator that is asymptotically unbiased under the general model is asymptotically equivalent to $V_0$ and hence robust in the sense of consistency under broad conditions. Indeed this is the main conclusion of the second part of the theorem.

Proof of the theorem

$$E_\xi(V_{NP2}) = \sum_{i \in s} \left( \frac{N^2}{n^2} u_i^2 + f_i \right) E_\xi(V_i) - \frac{2N}{n} \sum_{i \in s} u_i f_i E_\xi(V_i) + \sum_{i \in r} E_\xi(V_i)$$

$$E_\xi(V_i) = \sum_{k \in s} \sum_{k \in s} \frac{K\left(\frac{u_i - t_k}{h}\right)}{K\left(\frac{u_i}{h}\right)} E_\xi(\tau^2_k)$$

$$\approx \frac{1}{h} \sum_{k \in s, t_{k-1}} \int_{t_{k-1}}^{t_k} \left[ K\left(\frac{t_i - t_k}{h}\right)\sigma^2(t_k) - K\left(\frac{t_i - z}{h}\right)\sigma^2(z) \right] dz + \frac{1}{h} \sum_{k \in s} \int_{t_{k-1}}^{t_k} K\left(\frac{t_i - z}{h}\right)\sigma^2(z) dz$$

75
Hence
\[ E_\varepsilon(V_{NP2}) \approx \frac{N^2}{n^2} u_i^2 + f_i \left[ \sigma^2(t_i) + \sigma^2(t_i) h^2 d_k / 2 \right] - 2 \frac{N}{n} \sum_{i \in s} u_i f_i \left[ \sigma^2(t_i) + \sigma^2(t_i) h^2 d_k / 2 \right] \]
\[ + \sum_{i \in r} \left[ \sigma^2(t_i) + \sigma^2(t_i) h^2 d_k / 2 \right] , \]
\[ \sigma \in \mathbb{R} \]

Since
\[ Var(\hat{Y} - Y) = \sum_{i \in s} \frac{N^2}{n^2} u_i^2 \sigma^2(t_i) - 2 \frac{N}{n} \sum_{i \in s} u_i Cov(y_i, \hat{y}_i) + \sum_{i \in P} Var(y_i) , \]
it follows that the bias of \( V_{NP2} \) is
\[ B_\varepsilon(V_{NP2}) = \sum_{i \in s} \frac{N^2}{n^2} u_i^2 \frac{\sigma^2(t_i)}{2} h^2 d_k + O(N) , \]
After dividing this bias by the error variance we obtain the relative bias (RB) as
\[ RB(V_{NP2}) = h^2 \frac{d_k \sum_{i \in s} u_i^2 \sigma^2(t_i)}{2 \sum_{i \in s} u_i^2 \sigma^2(t_i)} + O(f) + O(h^2 f) . \]

Next we derive the asymptotic variance of \( V_{NP2} \). Note
\[ Var(V_{NP2}) = \sum_{i \in s} \sum_{k \in s} \left( \frac{N^2}{n^2} u_i^2 + f_i \right) \left( \frac{N^2}{n^2} u_k^2 + f_k \right) Cov(V_i, V_k) + \frac{4N^2}{n^2} \sum_{i \in s} \sum_{k \in s} f_i f_k u_i u_k Cov(V_i, V_k) \]
\[ + \sum_{i \in s} \sum_{k \in s} Cov(V_i, V_k) + 2 \sum_{i \in s} \sum_{k \in s} \left( \frac{N^2}{n^2} u_i^2 + f_i \right) Cov(V_i, V_k) \]
\[ - 2 \sum_{i \in s} \sum_{k \in s} \frac{N}{n} u_i f_k \left( \frac{N^2}{n^2} u_i^2 + f_i \right) Cov(V_i, V_k) - 2 \frac{N}{n} \sum_{i \in s} \sum_{k \in s} u_i f_i Cov(V_i, V_k) . \]

Now
\[ Cov(V_i, V_k) = \sum_{l \in s} \sum_{j \in s} \left\{ \left( \frac{1}{nh} \right)^2 K(t_i - t_l) K(t_j - t_l) \right\} Cov(r_i^2, r_j^2) \]
\[ \approx \left( \frac{1}{nh} \right)^2 \sum_{l \in s} \sum_{j \in s} K(t_i - t_l) K(t_j - t_l) Cov(r_i^2, r_j^2) . \]

But
\[ Cov(r_i^2, r_j^2) = Cov(\hat{y}_i^2, \hat{y}_j^2) - \frac{2M_l}{NM} Cov(\hat{y}_i, \hat{y}_j) + \left[ \frac{M_l}{NM} \right] ^2 Cov(\hat{y}_i, \hat{y}_j) . \]
Now observe that

\[
\text{Cov}(\hat{y}_j^2, \hat{y}_j^2) = \begin{cases} 
\text{Var}(\hat{y}_j^2) & \text{if } j = 1 \\
0 & \text{otherwise} 
\end{cases},
\]

\[
-2 M_i \frac{M_j}{N M} \text{Cov}(\hat{y}_j^2, \hat{y}_j) = \begin{cases} 
\frac{2M_i}{nM} \sum_{i \in s} u_i \text{Cov}(\hat{y}_i^2, \hat{y}_i) & \text{if } j = 1 \\
O(n^{-1}) & \text{otherwise}
\end{cases},
\]

\[
\left[ \frac{M_i}{nM} \right]^2 \text{Cov}(\hat{y}_j^2, \hat{y}_j^2) = \left[ \frac{M_i}{nM} \right]^2 \sum_{i \in s} \sum_{k \in s} u_i u_k \text{Cov}(\hat{y}_j^2, \hat{y}_j \hat{y}_k) = \\
= \begin{cases} 
O(n^{-1}) & \text{if } j = 1 \\
O(n^{-1}) & \text{otherwise}
\end{cases},
\]

\[
-2 M_i \frac{M_j}{N M} \text{Cov}(\hat{y}_i^2, \hat{y}_j) = \begin{cases} 
\frac{2M_i}{nM} \sum_{i \in s} u_i \text{Cov}(\hat{y}_i^2, \hat{y}_i) & \text{if } j = 1 \\
O(n^{-1}) & \text{otherwise}
\end{cases},
\]

\[
4 \left[ \frac{M_i M_j}{(N M)^2} \right] \text{Cov}(\hat{y}_j \hat{y}_j \hat{y}_i Y) = \begin{cases} 
4 \left[ \frac{M_i}{nM} \right]^2 \sum_{i \in s} \sum_{k \in s} u_i u_k \text{Cov}(\hat{y}_i \hat{y}_j, \hat{y}_i \hat{y}_k) & \text{if } j = 1 \\
O(n^{-1}) & \text{otherwise}
\end{cases},
\]

\[
-2 \left[ \frac{M_i^2 M_j}{(N M)^2} \right] \frac{\text{Cov}(\hat{y}_j \hat{y}_j \hat{y}_j Y^2)}{(N M)^2} = \begin{cases} 
O(n^{-1}) & \text{if } j = 1 \\
O(n^{-1}) & \text{otherwise}
\end{cases},
\]

\[
-2 \left[ \frac{M_i M_j^2}{(N M)^2} \right] \text{Cov}(\hat{y}_j \hat{y}_j Y^2) = \begin{cases} 
O(n^{-1}) & \text{if } j = 1 \\
O(n^{-1}) & \text{otherwise}
\end{cases},
\]
\[ \left( \frac{M_j}{N M} \right)^2 \text{Cov}(\hat{y}_j^2, \hat{y}_j^2) = \left( \frac{M_j}{n M} \right)^2 \sum_{j \in \mathcal{E}} \sum_{k \in \mathcal{E}} u_i u_k \text{Cov}(\hat{y}_i^2, \hat{y}_k^2) \]

\[ = \begin{cases} 
O(n^{-1}) & \text{if } j = l \\
O(n^{-1}) & \text{otherwise}
\end{cases} \]

Adding all the above we obtain

\[ \text{Cov}(r_j^2, r_l^2) = \begin{cases} 
\text{Var}(\hat{y}_j^3) - 4 \mu_j \text{Cov}(\hat{y}_j^2, \hat{y}_j) + 4 M_j^2 \mu^2 \text{Var}(\hat{y}_j) + O(n^{-1}) & \text{if } j = l \\
O(n^{-1}) & \text{otherwise}
\end{cases} \]

That is

\[ \text{Cov}(r_j^2, r_l^2) = \begin{cases} 
\text{Var}(\hat{y}_j^3) (1 + O(1)) + O(n^{-1}) & \text{if } j = l \\
O(n^{-1}) & \text{otherwise}
\end{cases} \]

Hence applying Taylor expansion to \( \text{Var}(\hat{y}_j^3) = \sigma^2(v_j^3) \) about \( t_i \); and as \( n \to \infty \), we obtain

\[ \text{Cov}(V_i, V_k) \approx \frac{1}{nh} \left[ \mu_4(t_i) - \mu_2^2(t_i) \right] \{1 + O(1)\} \int K(u)K \left( \frac{t_i - t_k}{h} - u \right) du. \]

Next we observe that

\[ \frac{4 N^2}{n^2} \sum_{i \in \mathcal{E}} \sum_{k \in \mathcal{E}} f_i f_k u_i u_k \text{Cov}(V_i, V_k) = O(N^2(nh)^{-1}) \]

\[ \sum_{i \in \mathcal{E}} \sum_{k \in \mathcal{E}} \text{Cov}(V_i, V_k) = O((N^3nh)^{-1}) \]

\[ \sum_{i \in \mathcal{E}} \sum_{k \in \mathcal{E}} \left( \frac{N^2}{n^2} u_i^2 + f_i \right) \text{Cov}(V_i, V_k) = O(N^3(n^2h)^{-1}) \]

\[ -2 \sum_{i \in \mathcal{E}} \sum_{k \in \mathcal{E}} \frac{N}{n} u_k f_k \left( \frac{N^2}{n^2} u_i^2 + f_i \right) \text{Cov}(V_i, V_k) = O(N^3(nh)^{-1}) \]
\[-4 \frac{N}{n} \sum_{i \in s} \sum_{k \in S} u_i f_i \text{Cov}(V_i, V_k) = 0 \left( N^2(nh)^{-1} \right) \]

Hence

\[
\text{Var} \left( \frac{V_{NP2}}{\text{Var}(\hat{Y} - Y)} \right) \approx \frac{1}{nh} \left[ \sum_{i \in S} \sum_{k \in S} u_i^2 u_k^2 \left[ \mu_4(t_i) - \mu_2^2(t_i) \right] \{1 + O(1)\} \int K(u)K \left( \frac{t_i - t_k}{h} - u \right) du \right] + O \left( \frac{n}{Nh} \right)
\]

and the asymptotic meansquare error of \( \frac{V_{NP2}}{\text{Var}(\hat{Y} - Y)} \) is

\[
\text{MSE} \left( \frac{V_{NP2}}{\text{Var}(\hat{Y} - Y)} \right) \approx \frac{h^4}{4} \left[ \frac{d_k \sum_{i \in S} u_i^2 \sigma^2(t_i)}{\sum_{i \in S} u_i^2 \sigma^2(t_i)} \right]^2 + \frac{1}{nh} \left[ \sum_{i \in S} \sum_{k \in S} u_i^2 u_k^2 \left[ \mu_4(t_i) - \mu_2^2(t_i) \right] \{1 + O(1)\} \int K(u)K \left( \frac{t_i - t_k}{h} - u \right) du \right]
\]

Clearly this goes to zero if the conditions of the theorem apply.

Let

\[
T = \left[ \frac{d_k \sum_{i \in S} u_i^2 \sigma^2(t_i)}{\sum_{i \in S} u_i^2 \sigma^2(t_i)} \right]^2
\]

and

\[
D = \frac{\sum_{i \in S} \sum_{k \in S} u_i^2 u_k^2 \left[ \mu_4(t_i) - \mu_2^2(t_i) \right] \{1 + O(1)\} \int K(u)K \left( \frac{t_i - t_k}{h} - u \right) du}{\left[ \sum_{i \in S} u_i^2 \sigma^2(t_i) \right]^2}
\]

Then the asymptotic meansquare error of \( \frac{V_{NP2}}{\text{Var}(\hat{Y} - Y)} \) is

\[
\frac{V_{NP2}}{\text{Var}(\hat{Y} - Y)} \approx \frac{h^4 T}{4} + \frac{D}{nh}
\]

An optimal bandwidth that minimizes this expression is

\[
h_{opt} = \left[ \frac{D}{nT} \right]^{\frac{1}{2}}
\]

and if \( h_{opt} \approx O(n^{-1/4}) \) then the local rate of convergence for the asymptotic mean square error of

\[
\frac{V_{NP2}}{\text{Var}_\xi(\hat{Y} - Y)} \] is \( O(n^{-4/5}) \)

Suppose \( n \hat{m}_i = \hat{M}_i \) (i.e. unistage sampling) then we observe \( y_i \) for all \( i \in s \). If \( u_i = \frac{\hat{M}_i}{\hat{m}_i} \),

79
then \( \hat{Y} \) becomes the ordinary ratio estimator

\[
\hat{Y}_R = \frac{\bar{y}_s}{\bar{X}_s} \sum_{i=1}^{N} X_i
\]

where \( y_s = \frac{\sum y_i}{n} \). Under the general model the error variance of \( Y_R \) is

\[
\text{Var}(\hat{Y}_R - Y) = \left[ \frac{(N-n)\bar{M}_r}{nm_s} \right]^2 \sum_{i \in s} \text{Var}(Y_i) + \sum_{i \in r} \text{Var}(Y_i).
\]

A non-parametric estimator of this variance is obtained by substituting \( u_i = \frac{\bar{M}}{m_s} \), in \( V_{NP2} \). This gives

\[
V_{nw} = \left[ \frac{(N-n)\bar{M}_r}{nm_s} \right]^2 \sum_{i \in s} V_i + \sum_{i \in r} V_i,
\]

previously obtained in chapter 2.

### 3.3.2 Dependent clusters

In the preceding section it has been assumed that clusters are independent. The assumption of independence can often be justified in practice as it facilitates obtaining theoretical results of classical interest. Although independence is generally assumed, it is seldom realised in survey practice. Thus the assumption that observations lying in adjacent clusters are uncorrelated is generally not met.

A more realistic view would be to assume an existence of dependence among different clusters and to model the structure of the population so as to incorporate this dependence.

In this section we focus on a dependence structure that is taken to be a function of cluster sizes. In particular we consider the following model

\[
E(\hat{Y}_{ij}) = \mu
\]

90.
where \( p(O) = 1 \); \( p(u) \leq 1 \) for all \( u \). It is assumed further that \( f(M_i, M_k) \) is reasonably smooth. This model asserts that units in different clusters are correlated but imposes no constraints on the variance structure of listing units within a cluster. The units within a cluster are, however, assumed to be independent. The error variance of \( \hat{Y} \) under model (3.13) is

\[
Var(\hat{Y} - Y) = \frac{N^2}{n^2} \left\{ \sum_s u_i^2 Var(\hat{y}_i) + \sum_{i \neq k} \sum_j \sum_j u_i u_k M_i^1 M_k^1 F(X_i, X_k) \right\}
- \frac{2N}{n} \left\{ \sum_s u_i Cov(\hat{y}_i, \hat{y}_i) + \sum_{i \neq k} \sum_j u_i M_i^1 M_k f(M_i, M_k) \right\}
+ \sum_{i \in s} \sum_{k \in r} u_i M_i^1 M_k f(M_i, M_k)
+ \sum_{i=1}^N Var(Y_i) + \sum_{i \neq k} \sum_j M_i^1 M_k^1 F(M_i, M_k).
\]

If \( f(M_i, M_k) = 0 \forall i \neq k \), then (3.14) reduces to (3.6) which is the expression of the error variance when clusters are uncorrelated. The expression (3.14) is a function of \( f(M_i, M_k) \), indicating that dependence among clusters affects the error variance. This implies that the estimators \( V_0 \) (see (1.8)) and \( V_{NP2} \), derived purely by assuming independence between clusters, will be biased and hence may not be appropriate if conditions in (3.13) apply.

Under model (3.13), it is observed that

\[
Y_i = M_i \mu + e_i^1 = F(M_i) + e_i
\]

where \( e_i = \sum_{j=1}^{M_i} e_{ij} \); \( e_{ij} = y_{ij} - \mu \). If it is assumed that \( F(M_i) \) is a smooth function of \( M_i \),
one immediately sees a direct link of estimation of the same function for independent clusters. If this can work, then one can be tempted to treat the problem of estimating (3.14) as that of fitting a smooth curve through estimated variances and covariances at each unit. Now (3.14) can be rearranged and expressed as

\[
\begin{align*}
\text{Var}(\hat{Y} - Y) &= \frac{N^2}{n^2} \left( \sum_s u_i^2 \text{Var}(\hat{y}_i) + \frac{-2N}{n} \sum_s u_i \text{Cov}(\hat{y}_i, y_i) \right) + \sum_{i=1}^N \text{Var}(Y_i) \\
&\quad + \frac{N^2}{n^2} \sum_{i \neq k} \sum_{s} u_i u_k M_i M_k \bar{f}(M_i, M_k) - \frac{2N}{n} \sum_{i \neq k} \sum_{s} u_i M_i M_k \bar{f}(M_i, M_k) \\
&\quad - \frac{2N}{n} \sum_{i \in s} \sum_{k \in r} u_i M_i M_k \bar{f}(M_i, M_k) + \sum_{i \neq k}^{N} \sum_{s} M_i M_k \bar{f}(M_i, M_k).
\end{align*}
\]

\( B \) can be estimated by \( V_{NP2} \) while \( M_i M_k \bar{f}(M_i, M_k) \) can be estimated by fitting a smooth function through the product \( r_i r_k \) where the residuals are taken from \( s \). More particularly if we let

\[
W_h(M_i, M_k) = \frac{K \left( \frac{M_L - M_E}{h} \right)}{\sum L K \left( \frac{M_L - M_E}{h} \right)}
\]

be a smoothing function, where \( M_L = \frac{M_i + M_k}{2} \); \( M_E = \frac{M_i + M_k}{2} \); \( \sum_{l \neq k'} l \); \( i', k' \) indicate sample clusters at the neighbourhood of \( i, k \) respectively. Then \( M_i M_k \bar{f}(M_i, M_k) \) can be estimated by

\[
V(M_i, M_k) = \sum_{L} W_h(M_i, M_k) r_L
\]

where \( r_L = r_i r_{k'} \). Thus an estimator of the error variance is given by
(3.17) \[ V_{NP_1} = V_{NP_2} + \frac{N^2}{n^2} \sum \sum_{i \neq k} u_i u_k V(M_i, M_k) \]
\[ - \frac{2N}{n} \left[ \sum_{i \neq k} \sum_{k \in s} u_i V(M_i, M_k) + \sum_{i \in s} \sum_{k \in s} u_i V(M_i, M_k) \right] \]
\[ + \sum_{i \neq k} \sum_{s} V(M_i, M_k) . \]

In the next section we compare the analytical properties of \( V_{NP_1}, V_{NP_2} \) and \( V_o \). We will require that as \( n, N \) become large both the population and the sample averages be bounded above. Further it will also prove useful to assume that the sample units are dense in the domain of interest. Finally, and without loss of generality, \( K(.) \) shall be assumed to be a uniform kernel function.

3.4 Performances of the Variance Estimators

We shall take model (1.7) as the working model. Under that model,

\[ E[\xi(r_i^2)] = Var(\hat{y}_i) \left( 1 - 2 \frac{M_i u_i}{nM} \right) + \left( \frac{M_i}{nM} \right)^2 \sum_s u_i^2 Var(\hat{y}_i) \]

where

\[ Var(\hat{y}_i) = \frac{M_i^2}{m_i} (1 - \rho + m_i \rho) \sigma^2. \]

Noting further that \( sup \frac{2\alpha}{n} = O(n^{-1}) \) where \( g_i = 3M_i \frac{M}{M^2} \), it follows that

\[ E[\xi(r_i^2)] \approx \frac{M_i^2}{m_i} (1 - \rho + m_i \rho) \sigma^2 \]

as \( n \to \infty \). This implies that under model (1.7)

\[ Var(\hat{y} - y) = \frac{N^2}{n^2} \sum_{i \in s} \frac{M_i^2}{m_i} (1 - \rho + m_i \rho) + O(N) \]

Under the uniform kernel,

\[ E[\xi(V_{NP_2})] = \frac{N^2}{n^2} \sum_{i \in s} \frac{2M_i^2}{m_i} (1 - \rho + m_i \rho) + O(N) \]

83
Hence $V_{NP_2}$ is asymptotically unbiased under the working model. Note also that for $i' \neq k'$, $i'$, $k' \in S$,

$$r_{i't_{k'}} = \left( \frac{\bar{y}_{i'} - \bar{Y}_{M_{i'}}}{N M} \right) \left( \frac{\bar{y}_{k'} - \bar{Y}_{M_{k'}}}{N M} \right)$$

$$= \frac{\bar{y}_{i'} \bar{y}_{k'} - M_{i'} M_{k'}}{n M} - \frac{M_{k'}}{n M} \bar{y}_{i'}^2 + \frac{M_{i'}}{n M} \bar{y}_{k'}^2$$

$$- \left\{ \frac{M_{i'}}{n M} \sum_{k' \neq i'}^n u_{i'} \bar{y}_{i'} \bar{y}_{k'} + \frac{M_{k'}}{n M} \sum_{k' \neq i'}^n u_{k'} \bar{y}_{i'} \bar{y}_{k'} \right\}$$

$$+ \frac{M_{i'} M_{k'}}{(n M)^2} \left[ \sum_{i'}^n u_{i'}^2 \bar{y}_{i'}^2 + \sum_{i'}^n \sum_{k' \neq i'}^n u_{i'} u_{k'} \bar{y}_{i'} \bar{y}_{k'} \right].$$

It implies that

$$E_\xi(r_{i't_{k'}})$$

$$= -\frac{1}{n M} \left[ M_{k'} u_{i'} \text{Var}(\bar{y}_{i'}) + M_{i'} u_{k'} \text{Var}(\bar{y}_{k'}) \right] + \frac{M_{i'} M_{k'}}{(n M)^2} \sum_{s} u_{i'}^2 \text{Var}(\bar{y}_{i'})$$

$$+ \left\{ M_{i'} M_{k'} - \left( \frac{M_{k'} M_{i'}}{n M} + \frac{M_{k'} M_{i'}}{n M} \right) n M + \left( \frac{M_{k'} M_{i'}}{(n M)^2} \right) \left( \sum_{u_{i'}} u_{i'} M_{i'} \right) \right\} \mu^2.$$

Recall that $u_{i'}$'s are constants chosen such that $\bar{Y}$ is unbiased. Hence

$$E_\xi(r_{i't_{k'}})$$

$$= -\frac{1}{n M} \left[ M_{k'} u_{i'} \text{Var}(\bar{y}_{i'}) + M_{i'} u_{k'} \text{Var}(\bar{y}_{k'}) \right] + \frac{M_{i'} M_{k'}}{(n M)^2} \sum_{s} u_{i'}^2 \text{Var}(\bar{y}_{i'})$$

$$= A.$$

Hence

$$E_\xi(V_{NP_1})$$

$$= \frac{N^2}{n^2} \sum_{s} u_{i'}^2 \text{Var}(\bar{y}_{i'}) - \frac{2N}{n} \left[ \sum_{i'}^n \sum_{k' \neq i'}^n u_{i'} u_{k'} A + \frac{N - n}{n - 1} \sum_{i'}^n \sum_{k' \neq i'}^n u_{i'} A \right]$$

$$+ \frac{N(N - 1)}{n(n - 1)} \sum_{i'}^n \sum_{k' \neq i'}^n A.$$
Under the same model

\[ E_\xi(V_0) = \text{Var}_\xi(\bar{Y} - Y). \]

That is, \(V_0\) is strictly unbiased. Thus from the above expressions

\[ E_\xi(V_0) = \text{Var}_\xi(\bar{Y} - Y) \approx E_\xi(V_{NP2}) \leq E_\xi(V_{NP1}). \]  

In conclusion, these results imply that \(V_{NP2}\) is asymptotically unbiased, while \(V_{NP1}\) is biased even in large samples. In particular, \(V_{NP1}\) will overestimate the error variance when stable conditions apply.

### 3.4.1. Effects of the Failure of the Working Model

The working model which has been used to study analytical properties discussed above assumes that clusters are uncorrelated. We will now consider a model that allows for dependence among clusters. In particular, we consider the following model

\[ E_\xi(Y_{ij}) = \mu \]

\[
\begin{align*}
\text{Cov}(Y_{ij}, \bar{Y}_{kl}) &= \begin{cases} 
\sqrt{\text{Var}(Y_{ij})}\sqrt{\text{Var}(Y_{kl})}\rho(M_i - M_k) = f(M_i, M_k) & \text{if } i \neq k \\
\text{Var}(Y_{ij}) & i = k; j = l \\
0 & \text{if } i = k; j \neq l
\end{cases}
\end{align*}
\]

Under this model, it can be shown

\[
E_\xi(r_{ij} r_{kl}) \approx M_{ij} M_{ik} f(M_{ij} M_{ik}) - \frac{1}{nM} \left\{ M_{ij} u_{iv} \text{Var}(\hat{y}_{iv}) + M_{ij} u_{iv} \text{Var}(\hat{y}_{iv}) \right\} + \frac{M_{ij} M_{ij}}{(nM)^2} \sum_s u_s^2 \text{Var}(\hat{y}_{iv})
\]

\[ = M. \]
Also under the same model

\[ E_{\xi}(\tau_{i}^{2}) = \text{Var}(\hat{y}_{i})(1 - \frac{2M_{i}u_{i}}{nM}) + \left(\frac{M_{i}}{nM}\right)^{2} \left\{ \sum_{s} u_{i}^{2}\text{Var}(\hat{y}_{i}) + \sum_{i \in s} \sum_{k \neq i} u_{i}u_{k}f(M_{i}, M_{k}) \right\}. \]

Thus if \( \frac{M_{i}}{nM} \rightarrow 0 \text{ as } n, N \text{ become large} \) it is seen that

\[ E_{\xi}(V_{NP2}) = E_{\xi}(V_{0}) - \frac{N^{2}}{n^{2}} \left\{ \sum_{s} u_{i}^{2}\text{Var}(\hat{y}_{i}) + \sum_{i \in s} \sum_{k \neq i} u_{i}u_{k}f(M_{i}, M_{k}) \right\} + O(N) \]

and

\[ E_{\xi}(V_{NP1}) = E_{\xi}(V_{NP2}) + \frac{N^{2}}{n^{2}} \left\{ \sum_{s} u_{i}^{2}\text{Var}(\hat{y}_{i}) + \sum_{i \in s} \sum_{k \neq i} u_{i}u_{k}f(M_{i}, M_{k}) \right\} \]

\[ + \frac{N(N - 1)}{n(n - 1)} \sum_{i' \in s} \sum_{k' \neq i'} M_{i'M_{k'}}f(M_{i'}, M_{k'}) \]

\[ - \frac{2N}{n} \left\{ \sum_{i' \in s} \sum_{k' \neq i'} u_{i'M_{i'}}M_{k'}f(M_{i'}, M_{k'}) + \frac{N(n - n)}{n(n - 1)} \sum_{i' \in s} \sum_{k' \neq i'} M_{i'M_{k'}}f(M_{i'}, M_{k'}) \right\}. \]

Also

\[ \text{Var}_{\xi}(\hat{Y} - Y) = \frac{N^{2}}{n^{2}} \left\{ \sum_{s} u_{i}^{2}\text{Var}(\hat{y}_{i}) + \sum_{i \in s} \sum_{k \neq i} u_{i}u_{k}f(M_{i}, M_{k}) \right\} \]

\[ - \frac{2N}{n} \left\{ \sum_{i \in s} \sum_{k \neq i} u_{i}M_{i}M_{k}f(M_{i}, M_{k}) + \sum_{i \neq k} M_{i}M_{k}f(M_{i}, M_{k}) \right\} \]

\[ + \sum_{i \in s} \sum_{k \neq i} M_{i}M_{k}f(M_{i}, M_{k}) + O(N). \]

Clearly if we define bias of the variance estimators as

\[ B_{\xi}(V) = E_{\xi}(V) - \text{Var}_{\xi}(\hat{Y} - Y) \]

then, under some stable conditions, it is seen that

\[ |B_{\xi}(V_{NP1})| \leq |B_{\xi}(V_{0})| = |B_{\xi}(V_{NP2})|. \]
This result illustrates that $V_{NP_1}$ is more bias robust to the misspecification of the correlation structure among the clusters than the other estimators. Note that the Relative Bias (REBI) (i.e. $\frac{B_r(V)}{\text{var}(Y-Y)} - 1$) of $V_0, V_{NP_2}$ are given by

$$REBI(V) = \frac{N(N-1)}{n(n-1)} \sum_{i \in S} \sum_{k \neq i} M_i M_k f(M_i, M_k)$$

Thus a condition for $V_0, V_{NP_2}$ to be asymptotically unbiased in this general setting is, besides the usual regularity conditions, $REBI(V)$ should tend to zero as $n, N$ become large. In practice such conditions may not be met and generally these estimators will be negatively biased in such settings. An implication coming from this brief discussion is that estimators such as $V_0, V_{NP_2}$ derived by assuming independence among the clusters are not appropriate in the setting in which there are some indications of dependence among the clusters. For such settings $V_{NP_1}$ could comparatively be a better estimator (from bias robustness criterion) than the others. Finally between $V_{NP_2}$ and $V_0$ we recommend $V_{NP_2}$ for applications since, unlike $V_0$, this estimator cannot take negative value when $n < \frac{2M_{\text{min}}}{nM}$. But more significantly this estimator can be employed even when $n = 2$, a case in which $V_0$ fails. We thus have a wider scope of applications with $V_{NP_2}$ than is possible with $V_0$. 

37
Chapter 4

BOOTSTRAP VARIANCE ESTIMATION FOR MODEL-BASED TWO STAGE CLUSTER SAMPLING

4.1 Introduction

In this chapter we consider a new variance estimation procedure for complex surveys. In particular we consider resampling procedure in which repeated samples are taken from a given initial sample and an estimator constructed from the resamples. Estimators based on such technique have proved useful and competitive under classical frequentist approach.

The two most commonly used resampling techniques for complex surveys data are the
jackknife and the balanced repeated replication (BRR) methods. These methods are often cumbersome to implement or can not be extended to some sampling options. For example, both of these techniques are not applicable when primary sampling units are sampled with replacement.

Another resampling technique which has become quite popular in the current research practice is the bootstrap method. Efron (1982) gave an extensive study of this method in the identically and independently distributed case to obtain standard error estimates and nonparametric confidence intervals for any parameter of interest. Thus bootstrap potentially offers a method of overcoming robustness problem associated with the model-based surveys. Rao and Wu (1988) have shown how to apply this technique in the design based setting. Sitter (1992) also performed similar study with some slight illustrative improvements over that of Rao and Wu (1988).

Extension of bootstrapping to model based surveys have been some what mild and less aggressive. However, the increased efficiency with such approach often makes it the more viable option (Royall and Cumberland (1981)). Further in other sampling situations (e.g. quota sampling) design based estimators are often invalid and the model based alternatives are often the only available choice. Do and Kokic (1992) extended bootstrap estimation to model based surveys in the context of a ridge estimator of the population mean. The extension of this procedure to complex surveys awaits investigation. It would also be of interest to see how such procedure compares with the one favoured in the current practice.

This chapter explores the above problems in the context of the two stage cluster sampling. The proposed procedure is given in section two. In section three we consider a special case and also derive an estimator of the variance for unistage sampling. A brief
summary is given in section four.

4.2 The Proposed Procedure and the New Estimator

We shall assume that the population under study follows a general superpopulation framework described in (1.5). This two stage model was employed by Royall (1986) to derive a robust estimator for the leading term of the error variance. For estimating the variance of the other components of the error variance, Royall (1986) employed a less general superpopulation structures. In this section we will show that it is possible to estimate the same error variance using (1.5) only. The motivation is that such a procedure may be more bias robust than the direct estimator proposed by Royall (1986).

Assume that the estimator \( \hat{Y} \) is used to estimate the population total \( Y \). Then we can obtain the error (i.e., the prediction error) associated with \( \hat{Y} \) as

\[
\hat{Y} - Y = \frac{N}{n} \sum_s u_i \hat{Y}_i - \sum_s y_i - \sum_r y_i.
\]

Now recall that \( y_i = M_i \mu + e_i; y_{si} = m_i \mu + e_{si} \). The prediction error can thus be expressed as

\[
\hat{Y} - Y \approx \sum_s \left( \frac{N}{n} u_i - 1 \right) \frac{M_i}{m_i} e_{si} - \sum_r e_i
\]

where \( e_i = \sum_{j=1}^{M_i} e_{ij}, e_{si} = \sum_{j \in s_i} e_{ij} \) and \( e_{ij} \) is defined in (3.4). Hence under (1.5) the error variance is given as

\[
\text{Var}_\xi(\hat{Y} - Y) \\
\approx \text{Var}_\xi \left( \sum_s \left( \frac{N}{n} u_i - 1 \right) \frac{M_i}{m_i} e_{si} \right) + \text{Var}_\xi \left( \sum_r e_i \right) \\
- 2 \text{Cov}_\xi \left( \sum_s \left( \frac{N}{n} u_i - 1 \right) \frac{M_i}{m_i} e_{si}, \sum_r e_k \right).
\]
For \( i \in s; k \in r \) the covariance component is null and \( e_s, \approx \hat{e}_s, = \hat{y}_i - m_i\hat{\mu} \). Thus

\[
(4.1) \quad \text{Var}_\xi(\hat{Y} - Y) \approx \text{Var}_\xi \left( \sum \frac{N}{n} u_i - 1 \right) \frac{M_i}{m_i} \hat{e}_s, + \text{Var}_\xi \left( \sum e_i \right)
\]

If (4.1) is expanded via model (1.5) then it reduces to (3.1). Estimation of (4.1) is therefore equivalent to estimation of (3.1).

### 4.2.1 The proposed method

The method proposed for estimating (4.1) uses the sample residuals

\[
\{ \hat{e}_i; i \in s \}
\]

as the initial sample. An element

\[
\hat{e}_{bi}; i = 1, \ldots, n
\]

in a bootstrap sample \( b \), is obtained at random from the parent (i.e. initial) sample \( \{ \hat{e}_i; i \in s \} \) with uniform probability \( p_i = \frac{1}{n}; i \in s \). B bootstrap samples are obtained in this way.

For estimating the first component of (4.1) we consider

\[
R = \sum \left( \frac{N}{n} u_i - 1 \right) \frac{M_i}{m_i} \hat{e}_s, = \sum \left( \frac{N}{n} u_i - 1 \right) \hat{e}_i
\]

as the statistic of interest. A corresponding statistic in the bootstrap sample \( b \), is obtained by substituting \( e_i \) with \( \hat{e}_{bi} \). That is

\[
R_b^* = \sum \left( \frac{N}{n} u_i - 1 \right) \hat{e}_{bi}^*
\]

Thus from \( B \) resamples we can generate the following sequence of statistics

\[
(4.2) \quad R_1^*, R_2^*, \ldots, R_B^*
\]
A key point to note is that both the parent sample and the bootstrap sample are assumed to come from the same superpopulation model (1.5). In other words we are assuming that bootstrap residuals trace the same distribution as the initial sample residuals. Taking this into account, it is then straightforward to assert that

$$Var_\xi(R) = Var_\xi(R^*_b) = E_\xi(R^*_b)^2.$$ 

Thus any one of the statistics given in (4.2) will give a model unbiased estimator of $Var_\xi(R)$. However, we choose to estimate this quantity by the average of the squares of the statistics in (3.2). That is, the model based bootstrap variance estimator of $Var_\xi(R)$ is:

$$V_1 = \frac{1}{B} \sum_{b=1}^{B} R^*_b.$$ 

To estimate the second quantity we first note that

$$\sum_r e_k \approx \sum_s \frac{N}{n} u_i \frac{M_i}{m_i} e_{s_i} - \sum_s e_i = \sum_s \left( \frac{N}{n} u_i - 1 \right) \frac{M_i}{m_i} e_{s_i}.$$ 

We further note that

$$Var_\xi \left( \sum_r e_i \right) = O(N - n).$$ 

We can thus estimate $Var_\xi \left( \sum e_i \right)$ via the sequence of statistics given in (4.2). In particular if we let

$$R_1 = \frac{R}{\sqrt{L}}; L = \frac{Nu}{n} - 1,$$

we obtain a sequence of bootstrap analogues of $R_1$ in the $B$ bootstrap samples as

$$(4.3) \quad \frac{R_1^*}{\sqrt{L}}, \frac{R_2^*}{\sqrt{L}}, \ldots, \frac{R_B^*}{\sqrt{L}}.$$
Noting that
\[ E_\xi\left( \frac{R}{\sqrt{L}} \right) = 0, \]
it is seen that
\[ \text{Var}_\xi(R_1) \approx V_2 = \frac{V_1}{L}. \]
Adding \( V_1 \) and \( V_2 \) gives the required model based estimator for the error variance of \( \hat{Y} \) as
\[ V_B = V_1 + V_2 = V_1(1 + \frac{1}{L}). \]

**Remark:** We note that in the design based approach Rao and Wu (1988) recommended that the expected value of \( R_i^* \) should first be estimated by \( \frac{1}{B} \sum_b R_b^* \). For our case, where we are resampling residuals, the expected value of the statistic of interest is zero. Hence we don’t have to estimate this quantity.

**Theorem 3** If \( n, N \) grow so that \( \frac{n}{N} \rightarrow c \), where \( c \) is a fixed constant, and if as this growth occurs \( M_i = O(1) \) and both the sample and the population remain stable then
\[ E_\xi\left( \frac{V_B}{N^2} \right) = \frac{1}{n^2} \sum_s u_i^2 \text{Var}_\xi(\hat{y}_i^*) + O((Nn)^{-1}) \]
and
\[ \text{Var}_\xi\left( \frac{V_B}{N^2} \right) = O(n^{-1}). \]

**Proof** Note that \( V_B = V_1 + O(N) \). Now
\[ V_1 = \frac{1}{B} \sum_b R_b^{2*} = \frac{1}{B} \sum_b \left\{ \sum_s \left( \frac{N}{n} u_i - 1 \right) \frac{M_i^2}{n^2} \hat{e}_{bi}^2 + \sum_{i=1}^n \sum_{k \neq i} \left( \frac{N}{n} u_k - 1 \right) \left( \frac{N}{n} u_i - 1 \right) \frac{M_i M_k}{m_i m_k} \hat{e}_{bi}^* \hat{e}_{ki}^* \right\}. \]
Now observe that
\[ E_\xi(\hat{e}_{bi}^{2*}) = \text{Var}(\hat{y}_{bi}) \left\{ 1 - \frac{2M_i}{nM} \right\} + \frac{M_i^2}{n^2M^2} \sum_s \text{Var}(\hat{y}_{bi}) \]
\[ E_\xi \left( \sum_{i=1}^{n} \sum_{k \neq i} \left( \frac{N}{n} u_i - 1 \right) \left( \frac{N}{n} u_k - 1 \right) \frac{M_i M_k}{m_i m_k} \hat{e}_{bi} \hat{e}_{bki} \right) \]
\[ = \frac{2}{B^2} \sum_{b} \left\{ \sum_{s} \left( \frac{N}{n} u_i - 1 \right)^2 \frac{M_i^2}{m_i^2} \text{Var} \left( \hat{y}_{bi}^* \right) \left\{ 1 - \frac{2M_i}{nM} \right\} + \frac{M_i^2}{n^2 M^2} \sum_s \text{Var} \left( \hat{y}_{bi}^* \right) \right\} . \]

Hence as \( n, N \) become large

\[ E_\xi (V_1) \approx \frac{1}{B^2} \sum_b \left\{ \sum_s \left( \frac{N}{n} u_i - 1 \right)^2 \frac{M_i^2}{m_i^2} \text{Var} \left( \hat{y}_{bi}^* \right) \left\{ 1 - \frac{2M_i}{nM} \right\} + \frac{M_i^2}{n^2 M^2} \sum_s \text{Var} \left( \hat{y}_{bi}^* \right) \right\} . \]

Since \( \hat{y}_{bi}^* = \hat{y}_i^* \), it follows that as \( n, N \) become large the result

\[ E_\xi \left( \frac{V_B}{N^2} \right) \approx \sum_s \left( \frac{1}{n} \right)^2 \text{Var} \left( \hat{y}_i^* \right) \]

follows immediately.

Next we consider the variance of \( \frac{V_B}{N^2} \). First note that

\[ \text{Var}_\xi (V_B) = \frac{1}{B^2} \sum_{b} \sum_{y'} \text{Cov} \left( R_b^{*2}, R_{b'}^{*2} \right) . \]

Now

\[ \text{Cov} \left( R_b^{*2}, R_{b'}^{*2} \right) = \sum_{i \in s} \sum_{i' \in s} \sum_{k \in s} \sum_{k' \in s} \left\{ \left( \frac{N}{n} u_k - 1 \right) \left( \frac{N}{n} u_{k'} - 1 \right) \right\} \text{Cov} \left( \hat{e}_{bi} \hat{e}_{b'i'}, \hat{e}_{b'i} \hat{e}_{b'k} \right) \]

\[ \approx \sum_{i \in s} \sum_{i' \in s} \sum_{k \in s} \sum_{k' \in s} \left\{ \left( \frac{N}{n} u_k - 1 \right) \left( \frac{N}{n} u_{k'} - 1 \right) \right\} \text{Cov} \left( \hat{y}_{bi} \hat{y}_{b'i'}, \hat{y}_{b'i} \hat{y}_{b'k} \right) \]

and noting that

\[ \text{Cov} \left( \hat{y}_{bi} \hat{y}_{b'i'}, \hat{y}_{b'i} \hat{y}_{b'k} \right) = 0 \]

\( \forall b \).
if all indices are different i.e. if \( i \neq k \neq k' \neq i' \), it immediately follows that

\[
\sup \left\{ \text{Cov} \left( R_i^{k^2}, R_{i'}^{k'^2} \right) \right\}
\]

occurs when at most three of the indices are different.

Let \( k = i' \) and suppose \( i \neq k \neq k' \). Then

\[
\text{Cov}(\hat{y}_{i_n}, \hat{y}_{i_{b_k}}, \hat{y}_{i_{b_k}}, \hat{y}_{i_{b_k}}') = \text{Cov}(\hat{y}_{i_{b_k}}, \hat{y}_{i_{b_k}}, \hat{y}_{i_{b_k}}')
\]

\[
= M_{i_{b_k}}M_{i_{b_k'}} \mu \text{Var}(\hat{y}^*_k).
\]

This implies that the leading term of \( \sup \left\{ \text{Cov} \left( R_i^{k^2}, R_{i'}^{k'^2} \right) \right\} \) is

\[
\sup \left\{ \text{Cov} \left( R_i^{k^2}, R_{i'}^{k'^2} \right) \right\} \approx \frac{N^4}{n^4} \sum_{i \in S} \sum_{i' \in S} \sum_{k \in S} \left\{ M_{i_{b_k}}M_{i_{b_k'}} \mu u_i u_{i'} u_k^2 \text{Var}(\hat{y}^*_k) \right\}
\]

\[
= \frac{N^4}{n^4} \left( \sum_{i \in S} M_{i_{b_k}} \right)^2 \mu^2 \sum_{k \in S} u_k^2 \text{Var}(\hat{y}^*_k).
\]

Hence the leading term of \( \text{Var} \left( \frac{E_B}{N^2} \right) \) is

\[
\text{Var} \left( \frac{E_B}{N^2} \right) = \begin{cases} 
\frac{M_{i_{b_k}}^2}{b_n} \sum_{k \in S} u_k^2 \text{Var}(\hat{y}^*_k)/n & \text{if } b = b' \\
\frac{M_{i_{b_k}}^2}{n} \sum_{k \in S} u_k^2 \text{Var}(\hat{y}^*_k)/n & \text{if } b \neq b' 
\end{cases}
\]

\[
= O(n^{-1}).
\]

Remark 1: the theorem indicates that the leading term of \( E_B(V_B) \) captures the leading term of the error variance (see (3.1)). That is, \( V_B \) is asymptotically unbiased. We note that in deriving \( V_B \) no assumptions were made about the variances and covariances of units within a cluster. So the estimator is bias robust to all possible variance-covariance misspecifications that can occur within a cluster. One can see a contrast with the direct method of variance estimation.

Remark 2: it is important to note that we have resampled residuals from parent sample.
of the residuals which are not identically and independently distributed (i.i.d). The consistency of the new variance estimator follows immediately from the theorem. The rate of convergence of the Mean Square Error of $\hat{\text{Var}}$ is $O(n^{-1})$. Note that this is equal to the rate of convergence for the direct variance estimator proposed by Royall (1986). It is also important to observe that this is a stronger rate than that observed in the case of fixed bandwidth kernel estimation of the error variance considered in chapter 2.

**Remark 3** : the proposed procedure extends quite easily to the case of dependent clusters. To see this first note from the previous remarks that the procedure is not influenced by the variance-covariance structure of the units within a cluster. It is tempting to exploit this idea to handle autocorrelated cases. In particular if we form clusters of consecutive observations which are highly correlated and then employ our proposed procedure we can reduce the effect of dependence using such clumping of serially correlated data into a cluster. The assumption is that the correlation between clusters decreases as the difference in time indices become large. Thus the model-based bootstrap estimator proposed here, is potentially insensitive to correlation among clusters. With some modifications, the procedure can incorporate the Moving Block Bootstrap (MBB) which is the latest innovation for handling dependence in the data (Liu and Singh, 1992).

### 4.3 A Special Case: The Ratio Estimator

The ratio estimator $\hat{Y}$ has the required form with $u_i = \frac{M}{m_i}$. The model-based bootstrap variance estimator $\hat{\text{Var}}$ becomes

$$V_B = \frac{1}{B} \sum_{b=1}^{B} T_b^2; T_b = \sum_i \frac{M_i}{m_i} \left( \frac{NM}{nm^s} - 1 \right) \hat{e}_{b,i};$$

$$K = 1 + \frac{1}{L}, \quad L = \frac{(N-n) \tilde{m}}{nm^s}.$$
Under the general model (1.5) the actual error variance is

\[
\text{Var}_\xi(\hat{Y}_R - Y) = \left(\frac{N \bar{M}}{nm_s}\right)^2 \sum_s \text{Var}(\hat{y}_i) - \frac{2N \bar{M}}{nm_s} \sum_s \text{Cov}(y_i, \hat{y}_i) + \sum_{i=1}^{N} \text{Var}(y_i).
\]

Clearly if the conditions of theorem 4.1 hold, then it can be shown that

\[
E_\xi(V_b) = E_\xi(V_{or}) = E_\xi(V_{np2}) = \text{Var}_\xi(\hat{Y}_R - Y) \leq E_\xi(V_{np1})
\]

where \(V_{or} \); \(V_{np2}\) and \(V_{np1}\) are obtained by substituting \(u_i = \frac{M}{m_s}\) in \(V_O\); \(V_{NP2}\) and \(V_{NP1}\) respectively.

When \(m_i = M\) (i.e. a case of unistage sampling) then \(\hat{Y}_R = \hat{T}_R\), the ordinary ratio estimator. For this,

\[
V_b = \frac{N(N - n)\bar{M} \bar{M}_r}{(nm_s)^2} \sum_{b} T^2_b,
\]

where \(T_b^* = \sum_s \hat{e}_{hsi}^*\) with

\[
E_\xi(V_b) = \frac{N(N - n)\bar{M} \bar{M}_r}{(nm_s)^2} \sum_s \left\{ \text{Var}(y_i^*) \left(1 - \frac{2M_i}{nm_s}\right) + \frac{M_i^2}{(nm_s)^2} \sum_s \text{Var}(y_i^*) \right\}.
\]

Hence under the asymptotic conditions of the theorem

\[
E_\xi(V_b) \approx \frac{N(N - n)\bar{M} \bar{M}_r}{(nm_s)^2} \sum_s \text{Var}(y_i^*).
\]

Under similar conditions

\[
E_\xi(V_D) \approx \text{Var}_\xi(\hat{T}_R - T) = \frac{N(N - n)\bar{M} \bar{M}_r}{(nm_s)^2} \sum_s \text{Var}(y_i).
\]

Hence under some conditions on \(y_i^*\)'s

\[
E_\xi(V_b) \approx E_\xi(V_D) \approx \text{Var}_\xi(\hat{T}_R - T).
\]

That is both \(V_b, V_D\) are asymptotically unbiased irrespective of the variance specification of the superpopulation model. These results indicate heteroscedacity bias robustness of
the two estimators.

In particular if we let $Var(y_i) = \sigma^2 M_i$ (a variance structure that allows $T_R$ to be BLUE) then it is observed that

$$E_t(V_b) = \frac{N(N-n)MM_T}{(nm_s)^2} \sum \sigma^2 M_i \left( 1 - \frac{M_i}{nm_s} \right)$$

while

$$E_t(V_D) = Var_t(\hat{T}_R - T) = \frac{N(N-n)MM_T}{(nm_s)^2} \sum \sigma^2 M_i.$$ 

Clearly under this parametric model $V_b$ is asymptotically unbiased whereas $V_D$ is strictly unbiased. A more promising estimator can be motivated by modifying the residuals from the parent sample: divide the original residuals in the sample by $(1 - \frac{M_i}{nm_s})$ and then resample the modified residuals. In this case $V_b$ will be strictly unbiased under the above specified model and will be equivalent to $V_D$ under some conditions on $y_i$'s in the bootstrap sample.

### 4.4 Conclusion

The application of the bootstrap technique to variance estimation for sample survey data from a model-based perspective, has been considered in this chapter. The chapter has specifically focussed on the through the origin model, employing the general point estimator for the population total and finally has considered the special case of the ratio estimator. Throughout the chapter emphasis has been on simplicity and no reference, whatsoever, has been made to probabilistic (classical) concepts and theory. Thus we see our proposed procedure as being directly relevant to an applied statistician who operates in the prediction framework.

We have considered resampling from the initial sample of raw residuals which are not
identically and independently distributed (i.i.d.). Thus the procedure potentially offers a solution to the robustness problems associated with non i.i.d cases. In one of our remarks we have also described how this procedure can be extended to dependent clusters.

Other appealing characteristics of the proposed procedure are: the estimator can not assume a negative value and is applicable to stratified two stage cluster sampling where the number of the first stage sampling units is two. This design is very important in sample surveys, especially those concerned with small domains. It is known that the estimator based on the direct method of variance estimation does not have such appealing characteristics.

We thus see a wider scope and flexibility in using the new procedure than is possible with the usual model based procedure for complex surveys.
Chapter 5

AN EMPIRICAL STUDY

The theory developed in chapters three and four is examined in a fairly extensive empirical study using one real population and five artificial populations. We have specifically considered the case of the ratio estimator (i.e. $u_i = \frac{M_i}{n_i}$). We have accordingly changed our notations of the variance estimators from $V_{NP2}$, $V_{NP1}$, $V_0$, $V_B$ to $V_{np2}$, $V_{np1}$, $V_0$, $V_B$, $V_0$ respectively. Section 5.1 describes the study populations in more details. Section 5.2 outlines the simulation procedure employed and the results of the empirical study. In section 5.3, general conclusions of the chapter are given.

5.1 Description of the Study Populations

The natural population ($NP$) used for this empirical study is a subset of the responding households in a portion of Kenya population census in 1979, compiled by Kenya Central Bureau of Statistics. A subset of 80 divisions was used and the number of responding households in each location of a given division was recorded. In this experiment, division represented a first stage unit (i.e. a cluster) while households represent the listing units.
number of the listing units \((M_i)\) in each cluster was also recorded.

The artificial populations were simulated by assuming a regression model of the form

\[
y_i = y_0 + e_i
\]

where \(i = 1, \ldots, 10000\), \(y_0 = 5000\), \(e_i = \rho e_{i-1} + u_i\) and \(\rho\) is the correlation coefficient and hence satisfies the condition \(|\rho| \leq 1\). \(u_i\)'s are white noise realizations and \(e_0\) was chosen from random normal numbers of unit variance and zero mean. The choice of this autoregressive (AR) model is appropriate since it gives survey values which are autocorrelated. Five values of \(\rho\) were employed. The simulated populations were then partitioned into 80 clusters, each of size \(M_i\). The \(M_i\)'s were assigned randomly and were not necessarily equal. Table 5 gives descriptions of the five simulated populations.

<table>
<thead>
<tr>
<th>Population</th>
<th>No. of Clusters</th>
<th>MSE</th>
<th>(\bar{M})</th>
<th>(\rho)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80</td>
<td>405391.91</td>
<td>9.81</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
<td>0.6521</td>
<td>9.83</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
<td>176890.15</td>
<td>18.15</td>
<td>-1</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>25242.84</td>
<td>10.45</td>
<td>0.5</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>559948.65</td>
<td>12.44</td>
<td>-0.5</td>
</tr>
<tr>
<td>NP</td>
<td>80</td>
<td>45678.44</td>
<td>7.67</td>
<td>-</td>
</tr>
</tbody>
</table>

### 5.2 Simulation Procedure

A total of 1000 two stage samples were selected from each population. At the first stage a simple random sample of \(n\) clusters (\(n\) is specified in table 7) was selected. At the second
stage a simple random sample of four listing units was selected without replacement from each cluster. In all the populations $M_i \geq 5$. In case $M_i \leq 4$ we recommend that all the listing units be selected. The cluster sizes for the artificial populations ranged from 5 to 30, reflecting the kind of variation often found in real populations.

A resampling algorithm due to Yang and Robinson (1986) was employed to generate 200 bootstrap resamples for each of the 1000 original samples selected. Since the estimator $V_{or}$ can take negative value whenever $n < \frac{2M_i}{m_i}$ for all $i \in s$ (Royall 1986), we rejected samples for which this condition was met. A gaussian Kernel was used.

For each population we calculated the population total $\bar{Y} = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$.\hspace{1cm} $\bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i$.

For each sample we calculated the ratio estimate $\hat{Y}_R = \frac{NM}{n \bar{M}} \sum_{i=1}^{s} \hat{Y}_i$, where $\hat{Y}_i = M_i \bar{y}_{si}$, $\bar{y}_{si} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$, $\bar{M}_i = \frac{1}{n} \sum_{i=1}^{N} M_i$; the actual prediction error squared $E^2 = (\hat{Y}_R - \bar{Y})^2$ as well as the four variance estimates $V_{or}, V_{np1}, V_{np2}, V_b$. For $V_{np1}, V_{np2}$ the rule of thumb of Silverman (1986) was used to choose a $h$ that minimized the $\frac{1}{1000} \sum_{s=1}^{1000} (V - MSE)^2$, where $V$ denotes the variance estimator under consideration, $MSE = \frac{1}{1000} \sum_{s=1}^{1000} E^2$, while $\sum$ indicates sum over all the 1000 samples.

For each variance estimate $V$, its unconditional Relative Bias (REBI(V)) was calculated as

$REBI(V) = \frac{\sum_{s=1}^{1000} V}{1000MSE} - 1$

over the 1000 simulated samples and its Root Average Square Error (RASE(V)) as

$RASE(V) = \left[ \frac{1}{1000} \sum_{s=1}^{1000} (V - MSE)^2 \right]^{\frac{1}{2}}$

over the same 1000 samples. The REBI(V) and the RASE(V) are given in tables 6 and 7 respectively. To see how the variance estimators perform under coverage probability criterion, we considered the $t$-statistic.
Table 6 Relative Bias of the Variance Estimators (REBI(V))

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample Size</th>
<th>$V_{or}$</th>
<th>$V_{np1}$</th>
<th>$V_{np2}$</th>
<th>$V_{b}$</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$n=8$</td>
<td>-6.24</td>
<td>1.32</td>
<td>-2.31</td>
<td>-6.64</td>
<td>$10^3$</td>
</tr>
<tr>
<td>2</td>
<td>$n=8$</td>
<td>9.82</td>
<td>31.04</td>
<td>10.41</td>
<td>-11.45</td>
<td>$10^2$</td>
</tr>
<tr>
<td>3</td>
<td>$n=12$</td>
<td>-17.59</td>
<td>6.31</td>
<td>-8.56</td>
<td>-15.84</td>
<td>$10^3$</td>
</tr>
<tr>
<td>4</td>
<td>$n=8$</td>
<td>25.3</td>
<td>1.62</td>
<td>-1.69</td>
<td>-0.91</td>
<td>$10^2$</td>
</tr>
<tr>
<td>5</td>
<td>$n=8$</td>
<td>15.2</td>
<td>19.00</td>
<td>-14.50</td>
<td>-14.46</td>
<td>$10^2$</td>
</tr>
<tr>
<td>NP</td>
<td>$n=8$</td>
<td>-8.67</td>
<td>4.52</td>
<td>-7.52</td>
<td>-6.51</td>
<td>$10^3$</td>
</tr>
</tbody>
</table>
\[ t = \frac{E}{\sqrt{V}} \] and the \((1 - \alpha)\) confidence interval for estimating \(Y\):

\[ (\hat{Y}_R - t^{(n-1)}_{\frac{\alpha}{2}} \sqrt{V}, \hat{Y}_R + t^{(n-1)}_{\frac{\alpha}{2}} \sqrt{V}) \]

where \(t^{(n-1)}_{\frac{\alpha}{2}}\) is the upper \(\frac{1}{2}\alpha\) point of the student \(t\) distribution with \(n - 1\) degrees of freedom (d.f.). The unconditional coverage probability of the above interval associated with the variance estimator \(V\) was computed as the percentage of the 1000 intervals that lie within the above interval.

To see how the performances of the variance estimates depend on \(\bar{m}_s\), we arranged the 1000 samples from each population in order of increasing values \(\bar{m}_s\). We then grouped the samples in 20 sets of 50 so that the first set contains 50 where \(\bar{m}_s\) are smallest, the next set contains the samples with the next 50 smallest \(\bar{m}_s\), and so on. For each of these 20 sets we calculated: the average value of \(\bar{m}_s\) as \(\frac{1}{50}\sum_{1}^{50} \bar{m}_s\); the conditional MSE (CMSE) \(= \frac{1}{50}\sum_{1}^{50} E^2\) and the averages of the variance estimates as \(\bar{V} = \frac{1}{50}\sum_{1}^{50} V\) for all the variance estimators. We then plotted the values of \(CMSE, \bar{V}\) against the average values of \(\bar{m}_s\).

The graphs are given in figures 1-6.

To study the behaviour of the standardized error (i.e. \(t = \frac{E}{\sqrt{V}}\)) in each sample we calculated the percentage of the 50 \(t\)'s in the interval

\[ \left[ -t^{(n-1)}_{\frac{\alpha}{2}}, t^{(n-1)}_{\frac{\alpha}{2}} \right] \]

where \(\alpha = 0.05\). This percentage is called the conditional coverage probability of \(V\).

Results are given in figures 7-12.

5.2.1 Simulation results for unconditional relative bias \(\text{REBI}(V)\)

For the real population the order of performance as measured by the relative bias, from best to worst, is \(V_{np1}, V_{np2}, V_\delta\) and \(V_\sigma\) respectively. Thus, in terms of unbiasedness the
nonparametric estimator $V_{np1}$ is the estimator of choice. For the artificial populations for which $\rho \neq 0$ (i.e. cases of autocorrelated clustered populations :1 , 3 , 4 , 5 ) best performance was achieved by $V_{np1}$. For the other estimators the order $V_{np2} > V_b > V_{or}$ was observed ($>$ indicates "is better than"). A further notable observation in table 6 is that with exception of $V_{np1}$ all the other estimators are predominantly negatively biased.

Further it is also observed that the absolute bias tend to increase with increase in $\rho$.

When $\rho = 0$ i.e.a case of autocorrelated clustered populations, the Royall estimator $V_{or}$ is the best and is followed more closely by $V_{np2} > V_b$. The close performance of $V_{or}$ and $V_{np2}$ was inferred in their analytical comparision.

Comparing the results for cases where $\rho = 0$ with those for which $\rho \neq 0$, one concludes that the new estimators $V_{np1}, V_{np2}, V_b$ are strong competitors to the established estimator, $V_{or}$.

Infact, from bias robustness point of view it is evident that the new estimators are more bias robust to the existence of autocorrelation than $V_{or}$. Another curious observation from the results is that the estimator $V_{np2}$ is the most stable, it tries to maintain its position with very little deviations.

The good performances of $V_{np2}, V_b$ over $V_{or}$ are attributed to the fact that these two estimators were derived using procedures that allow for robust estimation of the first two terms of the error variance. In the case of $V_{or}$ the least square technique was predominantly used to derive estimators of the last two components of the error variance.

5.2.2 Simulation results for root average square error (RASE(V))

Table 7 gives summary results of RASE(V) of the four variance estimators. It is observed that for the real population the estimator $V_{np}$ is the best while $V_{or}$ is the worst. The estimator $V_{or}$ recorded a larger RASE than the model based bootstrap variance estimator.

105
### Table 7  Root Average Square Error of the Variance Estimators (RASE(V))

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample Size</th>
<th>$V_{or}$</th>
<th>$V_{np1}$</th>
<th>$V_{np2}$</th>
<th>$V_{b}$</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$n=8$</td>
<td>35.8</td>
<td>5.11</td>
<td>6.38</td>
<td>18.70</td>
<td>$10^{17}$</td>
</tr>
<tr>
<td>2</td>
<td>$n=8$</td>
<td>208.13</td>
<td>270.40</td>
<td>158.04</td>
<td>234.60</td>
<td>$10^{6}$</td>
</tr>
<tr>
<td>3</td>
<td>$n=12$</td>
<td>10.20</td>
<td>1.32</td>
<td>1.75</td>
<td>2.91</td>
<td>$10^{19}$</td>
</tr>
<tr>
<td>4</td>
<td>$n=8$</td>
<td>20.30</td>
<td>1.42</td>
<td>1.75</td>
<td>2.91</td>
<td>$10^{19}$</td>
</tr>
<tr>
<td>5</td>
<td>$n=8$</td>
<td>8.80</td>
<td>1.22</td>
<td>5.20</td>
<td>4.80</td>
<td>$10^{20}$</td>
</tr>
<tr>
<td>NP</td>
<td>$n=8$</td>
<td>8.18</td>
<td>2.19</td>
<td>3.19</td>
<td>6.80</td>
<td>$10^{16}$</td>
</tr>
</tbody>
</table>
For the artificial populations, however, $V_{np1}$ performed relatively better than the other estimators. The estimator $V_{np2}$ is the second best while $V_{or}$ is the worst performer. Thus if one's interest is on planning future surveys, these results recommend use of the new estimators for those surveys for which there are strong indications that the units of the populations are serially correlated.

5.2.3 Simulation results for unconditional coverage probability (COPRO($V$))

The asymptotic distribution of $\hat{Y}_R$ generated by simple random sampling is standard normal when $\hat{Y}_R$ is standardized by subtracting $Y$ and dividing by $\sqrt{V}$ (Scott and Wu (1981)). If the normal approximation is accurate then the coverage probability of the associated $V$ should be about 95 percent.

Table 8 indicates that the confidence intervals based on $V_{or}$ and $V_{np2}$ both have coverage probability very close to the nominal value for the unautocorrelated artificial population. For the same population, the estimator $V_{np1}$ is highly conservative while $V_s$ gives an undercoverage. The poor performance of $V_{np1}$ can be attributed to its positive bias. For the other populations, the new estimators appear to perform better than $V_{or}$.

Overall, the estimator $V_{np2}$ appears an estimator of choice.

5.2.4 Simulation results for conditional biases

The worth of prediction theory becomes evident when results are examined for the 20 groups defined in terms of $\bar{m}_s$. Figures 4 through 6 show the results for 1000 simple random samples from the five populations. For population 2 it is observed that the estimator $V_{np1}$ tends to be too large when $\bar{m}_s$ is smaller than $\bar{M}$ and decreases when $\bar{m}_s > \bar{M}$. As prediction theory indicates, $V_{np1}$ generally overestimates CMSE, irrespective of
Table 8 Unconditional Coverage Probabilities of the Variance Estimators (COPRO(V) $\alpha = 0.05$)

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample Size</th>
<th>$V_0$</th>
<th>$V_{npl}$</th>
<th>$V_{np2}$</th>
<th>$V_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$n=8$</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>85</td>
</tr>
<tr>
<td>2</td>
<td>$n=8$</td>
<td>94.30</td>
<td>99.00</td>
<td>94.50</td>
<td>70.00</td>
</tr>
<tr>
<td>3</td>
<td>$n=12$</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>87.00</td>
</tr>
<tr>
<td>4</td>
<td>$n=8$</td>
<td>98.50</td>
<td>93.00</td>
<td>92.00</td>
<td>83.00</td>
</tr>
<tr>
<td>5</td>
<td>$n=8$</td>
<td>99.30</td>
<td>90.00</td>
<td>89.00</td>
<td>80.00</td>
</tr>
<tr>
<td>NP</td>
<td>$n=8$</td>
<td>98.00</td>
<td>96.00</td>
<td>96.00</td>
<td>65</td>
</tr>
</tbody>
</table>
the sample size. The performance of $V_{np2}, V_{or}$ in tracking the actual CMSE as $\bar{m}_s$ varies is clearly superior to the other two estimators. Specifically for $\bar{m}_s > \bar{M}$

$$V_{or} > V_{np2} > V_b > V_{np1}$$

while for $\bar{m}_s < \bar{M}$ the order

$$V_{np2} > V_{or} > V_b > V_{np1}$$

is evident. At balance $V_{or}$ is the best and is followed more closely by $V_{np2}$. In general the performances of these two estimators are quite close to each other and tend to support the conjecture that these two estimators are asymptotically equivalent. Another notable results is that the model based bootstrap estimator $V_b$ always underestimate CMSE, a result that supports its conditional coverage performance as seen in figure 9. Finally, although the performance of $V_{np1}$ is quite awful, its trajectory at balance indicates that balancing can reduce its bias.

For the cases where $p \neq 0$ (this appears to include the real population) the shapes of the trajectories were generally the same. All the populations gave CMSE that tended to decrease with $\bar{m}_s$, a result which is expected for many surveys. The order of performance of the variance estimators were generally the same. Hence we give results only for population 3.

The estimator $V_{np1}$ tracks the CMSE reasonably well for each of the 20 groups, a finding which is in general agreement with the analytical result. For $\bar{m}_s < \bar{M}$ the estimator $V_{np1}$ generally overestimates CMSE while at balance (or there about) the conditional bias is quite small and negative. For large $\bar{m}_s$ (i.e.$\bar{m}_s > \bar{M}$) $V_{np1}$ generally overestimates CMSE but still maintains its lead. For large samples estimators $V_{or}, V_{np2}, V_b$ appear to perform almost similarly. They generally track the CMSE from below.
the sample size. The performance of \( V_{np2}, V_{or} \) in tracking the actual CMSE as \( \bar{m}_s \) varies is clearly superior to the other two estimators. Specifically for \( \bar{m}_s > \bar{M} \)

\[
V_{or} > V_{np2} > V_5 > V_{np1}
\]

while for \( \bar{m}_s < \bar{M} \) the order

\[
V_{np2} > V_{or} > V_5 > V_{np1}
\]

is evident. At balance \( V_{or} \) is the best and is followed more closely by \( V_{np2} \). In general the performances of these two estimators are quite close to each other and tend to support the conjecture that these two estimators are asymptotically equivalent. Another notable results is that the model based bootstrap estimator \( V_5 \) always underestimate CMSE, a result that supports its conditional coverage performance as seen in figure 92. Finally, although the performance of \( V_{np1} \) is quite awful, its trajectory at balance indicates that balancing can reduce its bias.

For the cases where \( \rho \neq 0 \) (this appears to include the real population) the shapes of the trajectories were generally the same. All the populations gave CMSE that tended to decrease with \( \bar{m}_s \), a result which is expected for many surveys. The order of performance of the variance estimators were generally the same. Hence we give results only for population 3.

The estimator \( V_{np1} \) tracks the CMSE reasonably well for each of the 20 groups, a finding which is in general agreement with the analytical result. For \( \bar{m}_s < \bar{M} \) the estimator \( V_{np1} \) generally overestimates CMSE while at balance (or there about) the conditional bias is quite small and negative. For large \( \bar{m}_s \) (i.e. \( \bar{m}_s > \bar{M} \)) \( V_{np1} \) generally overestimates CMSE but still maintains its lead. For large samples estimators \( V_{or}, V_{np2}, V_5 \) appear to perform almost similarly. They generally track the CMSE from below.
Fig. 1. Trajectories for natural population

![Graph showing trajectories for natural population.](image-url)
Fig. 2: Trajectories for population 1

![Graph showing trajectories for population 1 with different symbols representing different groups and average values.](image-url)
Fig. 3. Trajectories for population 2
Fig. 4: Trajectories for population 3
Fig. 5 Trajectories for population 4

Average of volt vs. $C_mse$
(Times 10E5)

Group Average of $\bar{M}_S$

- $cmse$
- vnp1
- vor
- vb
- vmp2

8.5 9.5 10.5 11.5 12.5

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

115
A.2.3. Simulation results for conditional coverage probability.

The simulated coverage proportion of the conditional confidence intervals are shown in Figure 8. These illustrate how the various estimates of averages with respect to the population.

Fig. 8. Trajectories for population 5

![Graph showing trajectories for population 5](image-url)

**Average of MS**

- □ cmse + vnp1
- ◇ vor
- ∆ vb
- × vnp2

Group Average of $\overline{MS}$
5.2.5 Simulation results for conditional coverage probability (COPRO(V))

The conditional coverage properties of the confidence intervals are shown in figures 7-12. These illustrate how the variance estimator V performs with respect to the estimation of the confidence interval for a given sample. In population 2, V is seen to be negatively biased (see figure 3) for all \( \bar{m}_s \). Hence this estimator gives conditional intervals which cover \( Y \) less frequently. For extremely low values of \( \bar{m}_s \), the estimators \( V_{or} \) and \( V_{np2} \) produce better confidence intervals than the other two estimators. At balance the estimators \( V_{or} \) and \( V_{np2} \) give coverage probabilities very close to the nominal value. For \( \bar{m}_s > \bar{M} \), the coverage probability of \( V_{np2} \) is better than all the remaining estimators. In most of the samples, the coverage probability of \( V_{np1} \) is above the nominal value. This can be explained by extremely large positive bias associated with his estimator (see figure 3).

The results for most of the remaining populations indicate an exceedingly poor performance (see trajectories for natural population, populations: 1-4).
Fig. 7. Trajectories for Natural population 1

Conditional Coverage
Probabilities

Group Average $\bar{M}_s$

40
50
60
70
80
90
100
110

30
6.8
7
7.2
7.4
7.6
7.8
8
8.2
8.4
8.6

Group Average of $\bar{M}_s$

118
Fig. 8. Trajectories for population 1

Group Average of $\bar{M}_S$

- cvop
- cvnp2
- cvnp1
- cvb
Fig. 9. Trajectories for population 2

Conditional coverage probabilities

Group average of $\bar{m}$,

---

$CV(V_b)$

$CV(V_{np2})$

$CV(V_{or})$

$CV(V_{np1})$
Fig. 10 Trajectories for population 3

Conditional coverage probabilities

CV($V_0$)

CV($V_{np2}$)

CV($V_{or}$)

CV($V_{np1}$)
Fig. 11. trajectories for population 4

Group Average of $\overline{M}_S$

CV(vor)  +  CV(vnp2)  o  CV(vnp1)  △  CV(vb)
5.3 Concluding Remarks

The empirical examples provided here do provide an optimistic picture of what the proposed alternative variance estimation techniques can achieve. An obvious conclusion from these empirical examples is that nonparametric methods and model-based bootstrap techniques can provide estimates of the variance which are more efficient than their counterparts obtained from further study that can explain this apparent superior performance.

Fig. 12 Trajectories for population

Conditional Coverage Probabilities

Group Average of $\bar{M}_S$

- o cvn2
- cvn1
- cvb

123
5.3 Concluding Remarks

The empirical examples provided here do provide an optimistic picture of what the proposed alternative variance estimation techniques can achieve. An obvious conclusion from these empirical examples is that nonparametric method and model based bootstrap technique can give estimators of the variance which are more efficient than the estimator favoured in the current practice. The results based on coverage performance, however, have not been so satisfactory. Further study that can explain this erratic behaviour is needed.
Chapter 6

CONCLUSION AND FURTHER QUESTIONS

6.1 Epilogue of the Thesis

The purpose of this thesis has been the estimation of the error variance of a point estimator $\hat{T}$. The ultimate goal has been to obtain variance estimators that have some good robustness properties under general variance structure of the prediction model. We specifically considered linear estimators for the population total. To investigate the above, we arranged our work in six chapters.

In chapter 1 we outlined the sample survey estimation problem. We reviewed the two main approaches to sample survey problem. It was observed that the two main approaches are philosophically different. Hence, as a consequence, variance estimation problem under the two approaches have been motivated by different concerns. In the prediction approach it was observed that the concern has been on obtaining a robust estimator given
a superpopulation model, while in the classical approach obtaining unbiased estimator under repeated sampling has been the main concern. A review of earlier studies—particularly those following prediction approach revealed that the available model-based variance estimators are not satisfactory. We thus concluded this chapter by "a call for alternative variance estimators."

In chapter 2, we introduced new procedures for estimating the error variance of a given descriptive statistic under unistage sampling. In particular, we suggested two estimators based on nonparametric method and further two estimators based on the marriage of nonparametric method and the direct method of variance estimation. We justified the use of the nonparametric method for variance estimation by establishing the consistency of the proposed estimators under a general heteroscedasticity. It was observed (somehow disappointingly) that this consistency is achieved at the price of decrease in efficiency. In fact, the optimal rate of the convergence for the asymptotic mean square error was shown to be $O(n^{-1})$. Further, we studied the performance of the variance estimators with regards to their biasedness/unbiasedness and robustness on some given prediction models. If the working model for the ratio estimation applies, it was observed that the estimator based on direct method are the best. However, when there are misspecifications in the variance model the new estimators were the best from the bias robustness point of view. In the empirical study of the estimators, it was observed that the simulation results, in many cases, resembled the analytical findings. In particular, results based on RASE criterion indicated that an estimator, derived purely using nonparametric method is the best. For statistical inference, a cross breed estimator and an estimator based on the direct method of variance estimation showed significant performances over the others.
An extension of nonparametric variance estimation procedure of chapter 2 to two stage cluster sampling was considered in chapter 3. We also considered nonparametric estimation of the error variance in the case of dependent clusters. Two new estimators $V_{NP}$ and $V_{NP2}$ were proposed. Under ratio estimation, $V_{NP2}$ is the two stage analogue of a nonparametric estimator suggested in the unistage sampling. Investigation of the sensitivity/robustness of the proposed estimators and the direct variance estimator $V_0$, to the misspecification of the "between cluster" covariance structure revealed that the estimator $V_0$ is biased if variables in different clusters are dependent and that the new estimators had some significant gains over $V_0$. However, for the working model in which the "between cluster" covariance is null, $V_0$ was an estimator of choice followed more closely by $V_{NP2}$ while $V_{NP1}$ was the worst.

A model-based bootstrap variance estimator for two stage cluster sampling was considered in chapter 4. A key principle in the proposed procedure is that the resampled residuals are generated by the same superpopulation model generating the parent sample of the residuals. With this crucial assumption we were able to establish the consistency of the proposed estimator under a general model. Unlike the other estimators, the new estimator has the required advantage of simplicity, which is a central requirement for a good variance estimator in complex surveys. An outstanding result that came out from this chapter is that the asymptotic meansquare of the proposed estimator achieved the same optimal rate of convergence as that of the estimator based on a correctly specified model.

In chapter 5, we examined the comparative empirical properties of the estimators in chapter 3 and 4. On average the empirical examples considered gave an optimistic picture of performance of the proposed variance estimators. In particular the new estimators showed some strong gains in robustness over the estimator favoured in practice. The coverage
6.2 Further Questions

This section briefly describes some areas for further study. These topics have emerged during the work on this thesis.

1) In this study we have considered linear models (i.e. $E(Y_i) = \beta x_i$, for one stage sampling and $E(Y_{i}) = M_i \mu$, for two stage cluster sampling). Using these models we have derived estimators of the variance of linear estimators of the population total. The motivation for using linear estimators is that they are unbiased or approximately so under these models. However, when the models are nonlinear then the estimators of the population total considered here are no longer unbiased. Hence the area of robust prediction (both of the population total and the error variance of the resulting estimators) under nonlinear models appears to be an open area for further study. Chambers (1986) studied the effects of outliers under linear model and suggest some estimators of the total. However, little appears to have been written for robust prediction of the error variance in the presence of nonlinear models (and/or unusual data points). Extension of the model-based bootstrap technique as proposed in chapter four looks feasible and may provide some worthwhile study. Another possibility could be to use weighted bootstrap as in Wu (1986). Another possibility for variance estimation is the trimming technique as applied to residuals.

2) Related to the topic of robust variance estimation is the construction of robust confidence intervals. The empirical results for confidence intervals associated with the point estimators and the new variance estimators have not been so satisfactory. It appears that unbiasedness or approximate unbiasedness of these estimators is not enough to ensure that
confidence intervals have their nominal coverage probabilities. Other factors seem also to
effect some influence on the coverage performance. Several avenues for further studies thus
suggest themselves here. Asymptotic studies of the tail probabilities of confidence inter-
vals computed using the variance estimators suggested here seem to be in order. Skewness
in the y distribution can result in confidence intervals computed using normal approxima-
tion, which too often miss the function of interest either from below or above.

Wyant (1978) gives methods of producing robust intervals for the mean of a distribution
when the distribution itself is positively skewed. It would be of interest to investigate
whether Wyant's procedure improves the coverage probability of the confidence intervals
associated with the new variance estimators proposed in this thesis.

3) An assumption that has been made in chapter 2 and 3 is that there are no jumps
in the variance function $\sigma^2(x_i)$ or $\sigma^2(t_i)$. In survey sampling this eliminates situations
where there are non response. Extensions of the methods developed in this thesis to
such situations still await investigations. Another problem which can occur is when the
set $\{x_j, \ j \in s\}$ is not an everywhere dense subset of $\mathcal{U}$. In particular, how will $\sum_{i \in \mathcal{E}} \text{Var}_i y_i$
be predicted if $y_i$'s ($i \in \mathcal{E}$) happen to be remote (i.e. lie at the edges of the closed interval
$[a, b]$) from the sampled ones?
REFERENCES


Ericsson, B. and Adell, R., (1994). On the Analysis of Life Tables for Dependent obser-


