

**RESAMPLING-BASED VARIANCE ESTIMATORS  
IN RATIO ESTIMATION WITH APPLICATION  
TO WEIGH SCALING**

By

Al-Karim Madatally Ladak

B.Sc., The University of British Columbia, 1988

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF  
THE REQUIREMENTS FOR THE DEGREE OF  
MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES  
THE DEPARTMENT OF STATISTICS

We accept this thesis as conforming  
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September 1990

© Al-Karim Madatally Ladak, 1990

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia  
Vancouver, Canada

Date September 25, 1990

## **Abstract**

Weigh scaling is a method of estimating the total volume of timber harvested from a given region. The implementation of statistical sampling techniques in weigh scaling is described, along with related issues. A review of ratio estimators, along with variance estimators of the classical ratio estimator is conducted. The estimation of the variance of the estimated total volume is considered using jackknife- and bootstrap-based variance estimators. Weighted versions of the jackknife and bootstrap variance estimators are derived using influence functions and Fisher Information matrices. Empirical studies of analytic and resampling-based variance estimators are conducted, with particular emphasis on small sample properties and on robustness with respect to both the homoscedastic variance and zero-intercept population characteristics. With a squared error loss function, the resampling-based variance estimators are shown to perform very well at all sample sizes in finite populations with normally distributed errors. These estimators are found to have small negative biases for small sample sizes and to be robust with respect to heteroscedasticity.

# Contents

Abstract .....	ii
Table of Contents .....	iii
List of Tables .....	vi
List of Figures .....	vii
Acknowledgment .....	viii
<b>1 Statistical Sampling in the Forest Industry in British Columbia</b> .....	<b>1</b>
1.1 The Forest Industry in British Columbia .....	1
1.2 Weigh Scaling .....	2
1.3 Statistical Estimation of Total Volume .....	6
1.4 Implementation Problems .....	9
1.5 Problem of Interest .....	11
<b>2 Estimating a Population Total via Ratio Estimators</b> .....	<b>14</b>
2.1 Estimating a Population Total .....	15
2.2 Some Ratio Estimators .....	17
2.3 Review of Empirical Studies of Ratio Estimators .....	21
2.4 Estimation of the Variance of the Classical Ratio Estimator of a Population Total ...	22
2.5 Summary .....	29

<b>3</b>	<b>Jackknifing in Ratio Estimation</b>	<b>30</b>
3.1	Introduction to Jackknife	31
3.2	Jackknife in Linear Models	33
3.3	Weighted Jackknife in Linear Models	37
3.3.1	Approach Using Influence Functions	37
3.3.2	Approach Using the Fisher Information Matrix	41
3.4	Jackknife Variance Estimates in a Heteroscedastic Linear Model	42
3.4.1	Influence Function-Based Jackknife	44
3.4.2	Information Matrix-Based Jackknife	46
3.5	Weighted Jackknife Variance Estimates in Ratio Estimation	48
3.5.1	Influence Function-Based Approach	48
3.5.2	Information Matrix-Based Approach	50
3.6	Summary	51
<b>4</b>	<b>Bootstrapping in Ratio Estimation</b>	<b>52</b>
4.1	Introduction	52
4.2	Bootstrap Variance Estimates - A Design-Based Approach	56
4.3	Bootstrap Variance Estimates - A Model-Based Approach	57
4.4	Weighted Bootstrap Variance Estimates	61
4.5	Some Comments on Proposed Estimators	63
4.6	Summary	65

<b>5</b>	<b>Sensitivity of Variance Estimates: An Empirical Study</b>	<b>66</b>
5.1	Purpose of Empirical Study	66
5.2	Overview of Empirical Study	67
5.3	Description of Populations	70
5.4	Results from Real Populations	76
5.5	Results from Artificial Populations	79
5.6	Summary	82
<b>6</b>	<b>Conclusions and Recommendations</b>	<b>84</b>
	<b>Bibliography</b>	<b>87</b>

## List of Tables

Table 1.	Empirical Results of Variance Estimators: Population A . . . . .	91
Table 2.	Empirical Results of Variance Estimators: Population B . . . . .	92
Table 3.	Empirical Results of Variance Estimators: Population C . . . . .	93
Table 4.	Empirical Results of Variance Estimators: Population Y00 . . . . .	94
Table 5.	Empirical Results of Variance Estimators: Population Y10 . . . . .	95
Table 6.	Empirical Results of Variance Estimators: Population Y20 . . . . .	96
Table 7.	Empirical Results of Variance Estimators: Population Y01 . . . . .	97
Table 8.	Empirical Results of Variance Estimators: Population Y11 . . . . .	98
Table 9.	Empirical Results of Variance Estimators: Population Y21 . . . . .	99

List of Figures

Figure 1a. Populations 1-9 from B.C. Ministry of Forests . . . . . 100

Figure 1b. Populations 10-18 from B.C. Ministry of Forests . . . . . 101

Figure 2. Scatterplot of Population A . . . . . 102

Figure 3. Scatterplot of Population B . . . . . 103

Figure 4. Scatterplot of Population C . . . . . 104



## Acknowledgment

I would like to thank the Faculty of the Department of Statistics at the University of British Columbia for all their help. In particular, I wish to express my gratitude and indebtedness to my thesis supervisors, Dr. Mohan Delampady and Dr. A. John Petkau, for their generosity of time and effort during the research and writing of this thesis. In many ways, this thesis is as much theirs as it is mine, particularly any mistakes, errors and omissions. I also take this opportunity to thank Dr. A. John Petkau and Dr. James V. Zidek for their guidance, help and inspiration during my studies at the University of British Columbia.

I wish to thank Mr. Patrick Plunkett of the British Columbia Ministry of Forests for his interest in the research of this thesis, and for providing the data which formed the basis of Chapter 5. I also wish to thank Ms. Rabiya Alimohamed for her impeccable typing and for her tolerance of numerous revisions of this thesis.

It is my pleasure to acknowledge the kindness and assistance extended to me by my colleagues at Price Waterhouse. Mr. Robert G. Elton and Mr. R. Martin Roberts took interest in my studies and provided considerable financial and other support, and I take this opportunity to express my deep and sincere gratitude to them for their kindness and generosity. Finally, I would especially like to thank my friend and colleague Dr. Edward J. Mansfield for kindly and generously providing me with his time and guidance both during my studies and during my association with Price Waterhouse. He, along with Dr. A. John Petkau, has significantly contributed to my training as a statistician, and I take this opportunity to express my deep and heartfelt gratitude to them. I hope this thesis is worthy of their support and efforts.

# **Chapter 1**

## **Statistical Sampling in the Forest Industry in British Columbia**

### **1.1 The Forest Industry in British Columbia**

Forestry is the dominant industry in British Columbia, accounting for a major portion of the provincial economic output. The extensive forests which cover much of the province form the basis for this industry. Most of the forested lands are owned by the Government of British Columbia, the management and regulation of these lands being the responsibility of the provincial Ministry of Forests. While the Ministry takes an active interest in all aspects of forestry, it does not engage in the actual harvesting and processing of timber. Rather the Ministry (under several different programs) allows private companies to do this in exchange for payments called stumpage fees.

The method of calculating stumpage fees changes from time to time and can be quite complex, incorporating such factors as silviculture costs, harvesting costs, the species of tree being harvested and the estimated value of the products produced from the harvested timber. However, the central idea behind all of these methods is that the stumpage fees that a private company pays should be proportional to the volume of timber which that company harvests. Consequently, it is extremely important to both the Ministry and the companies that the volume of timber harvested be determined accurately and efficiently.

Because of the differences in timber species and harvesting methods, the forest industry in British Columbia is grouped into two geographic categories: Coast and Interior. In the coast region, the Government requires that the volume of each log be measured separately. The process of measuring the volume of a log is done manually and is known as hand or stick scaling. It is

expensive to hand scale each log; however, coastal timber has a high dollar value and so the costs of scaling are relatively small compared to the value of the logs.

In the interior of the province, timber is of lower quality and dollar value. Consequently, the government does not require the scaling of each log. Rather, all logs are weighed, and the total volume harvested by a company is estimated via a sample of logs which are both weighed and hand scaled. This process is known as weigh scaling. Regardless of the method of scaling, however, the estimate of total volume must be accurate to a level set by provincial legislation.

It is one aspect of weigh scaling which motivates this study, namely that of estimating the total volume of timber harvested by a logging company to within a prescribed level of precision, while using the minimum of resources (i.e. lowest cost). In particular, this study will concentrate on the estimation of the variance of that total volume estimate. This chapter will describe the physical process involved in the logging and scaling of timber and the current method of estimation of total volume used in weigh scaling. Both statistical and non-statistical problems which arise in the process will also be discussed.

## **1.2 Weigh Scaling**

In this section, we will describe the physical process involved in the logging and weigh scaling of harvested timber. In particular, we will concentrate on the sampling aspects of this process.

At a logging site, trees are cut and their branches are removed until just the trunks remain. These logs are then trimmed at the ends. Thereafter, the logs are transported to a central location within the logging area. This process, from the actual cutting to the transport to a central location is called falling, bucking and yarding. Once at this central location, the logs are sorted, tagged

and loaded onto trucks for transport to various locations. Typically, the majority of logs from a logging site are sent to one location. For example, this could be a pulp mill or a sawmill.

When logging trucks arrive at a final destination, they are weighed both loaded and empty, so that the weight of the load may be obtained. Some statistically relevant decisions are made at the weigh scale and thus, the sequence of events occurring when a truck arrives to be weighed will be described.

Before proceeding, it is important to mention that in order to reduce the amount of sampling required the Ministry of Forests allows companies to create strata. The creation of appropriate and homogeneous strata from the population of all truckloads which arrive at a specific location may significantly reduce the sample size requirement. Thus, by creating strata a company can reduce the amount of scaling it must conduct and pay for, without reducing the accuracy of the estimates. The quantity and type of strata are determined by the company, which submits its plan to the Ministry for approval. The strata are not actual physical locations, but rather, species and grade types. The Ministry determines the sampling rate for each stratum and it is the company's responsibility to ensure that it is carried out.

When a logging truck arrives at a mill, it is directed to a weigh scale and weighed. The weigh scale operator, a company employee, classifies the truckload of logs as belonging to one of that company's strata. His decision is based on many factors, including the timbermark, the species makeup of the load, the amount of rot and the presence and extent of pest damage. Once the load is assigned to a stratum (or its appropriate stratum identified depending on one's point of view), the relevant details of the load are entered on a computer attached to the weigh scaling machine. The computer keeps a record of the accumulated number of truckloads belonging to each stratum. Each stratum has a different sampling rate, and the computer signals the weigh scale operator

when a truckload belonging to a particular stratum has to be sampled. If it is to be sampled, the truck is sent to a special site on the location where the logs are unloaded and are subsequently scaled, i.e. the volume of each of the logs is determined by hand scaling. Otherwise the truck is sent to the yard where it unloads its logs together with others belonging to the same stratum.

The overall sampling process is as follows. At the beginning of the year, each logging company identifies the strata it has chosen and the anticipated number of truckloads to be harvested from each of these strata. Companies are free to choose the number and type of strata. Since a company pays for the scaling of its wood, it is in the best interest of the company to choose that combination of number and type of strata which minimizes the amount of sampling required. The Ministry of Forests then determines the number of truckloads that must be sampled from a stratum in order for a prescribed level of precision to be met. Once a sampling rate has been decided by the Ministry, it is the responsibility of the company and its employees to ensure that this rate is achieved. This prescribed level of precision is set by provincial legislation.

Under provincial legislation, the estimate of total volume of wood harvested by a company must be within 1% of the true value, with a confidence level of 95%. In other words, the estimate of total volume must be within 1% of the true value, in 19 out of 20 estimates of total volume of different populations. The Ministry of Forests estimates the total sample size required from each stratum of a company in order that the overall 1% requirement be met. Since a company may have chosen to categorize their harvest into several strata, the Ministry of Forests allocates to each stratum the number of samples which are required from it. The samples are selected via a sampling scheme called block sampling: the random selection of one truck from every block of consecutive  $N_m$  trucks from stratum  $m$ , where  $N_m$  is determined by the Ministry of Forests.

For example, a company may choose to establish three strata and anticipate harvesting 1000, 500 and 200 truckloads per respective stratum. Using variance estimates from similar strata from the previous year, the Ministry of Forests may initially estimate that 100 truckloads should be sampled from the first stratum, 50 from the second, and 20 from the third. (These numbers are just examples to illustrate the sampling scheme.) However, since the variance estimate formula used by the Ministry is based on the assumption that the sample sizes will be large, the Ministry has imposed a minimum sample size constraint of 30 truckloads. In addition, the Ministry may at its discretion, increase the sample size if it feels the initial estimates are too small. Suppose in this example that the Ministry chose to increase each sample size by five, so that the final sample sizes per stratum would be: 105, 55 and 35. This means that for stratum one, one truck will be randomly selected from every successive set of nine trucks (1,000 divided by 105 is approximately nine) and similarly for the other strata. In other words, a block size is determined within which one truck will be chosen using simple random sampling. To maintain the integrity of the sampling process, the company is not to know which truckload will be sampled. To achieve this, the sampling scheme for each stratum is entered onto the computer at the weigh scale, and the computer generates the necessary random numbers to accomplish the random selection of one truck within each block for each stratum.

The hand scaling of logs in the sample loads is carried out by specially trained employees of the company called scalers. The volume measurements obtained from hand scaling are themselves inexact as logs are often curved and bent, and may contain knots and other irregularities. Consequently, if two different scalers measure the same load, they might arrive at two different volume estimates. However, if both scalers are experienced, their estimates should not differ by a large amount. The entire process relies on trust and honesty, but there are occasional spot checks by the Ministry to ensure the scalers do not cheat on their estimates.

At the end of a calendar year, the Ministry uses the weights and volumes of the sampled loads to develop a set of conversion factors relating volume to weight. The estimates of total volume harvested are then made by multiplying the total weight of all truckloads by the appropriate conversion factors.

### 1.3 Statistical Estimation of Total Volume

The previous section provided a brief introduction to the actual physical process whereby a standing tree in the forest is felled, yarded, weighed and possibly scaled. In this section we will describe the currently used analytic formulae to estimate the ratio of conversion from total weight to total volume, the total volume, the variances of the ratios and totals, and other related quantities. All the formulae are derived and discussed in the standard sampling text, Cochran (1977).

For each stratum, the Ministry first estimates the total volume of harvested timber, and then estimates the variability of this estimate. To properly discuss the above statistical estimates, we will first need to establish the notation.

Without loss of generality, we will consider the population of interest to be a particular stratum, the basic unit of which is a truckload. Let  $N$  be the total number of truckloads of timber harvested from a particular stratum during a calendar year, and let  $n$  be the number of truckloads sampled during the year from that stratum. Let  $Y$  be the total volume of timber in a population,

$$Y = \sum_{i=1}^N y_i = N\bar{Y} ,$$

and let  $X$  be the total weight of timber in that population,

$$X = \sum_{i=1}^N x_i = N\bar{X} .$$

Let  $R$  be the population ratio of volume to weight (inverse of the population density),

$$R = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}} ,$$

and let  $f$  be the sampling fraction,

$$f = \frac{n}{N} .$$

The mean volume of the scaled (sampled) truckloads is

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i ,$$

where  $y_i$  is the volume of the  $i^{\text{th}}$  sampled truckload. Similarly, the mean weight of the sampled truckloads is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i ,$$

where  $x_i$  is the weight of the  $i^{\text{th}}$  sampled truckload. The estimate of the population ratio of volume to weight is

$$\hat{R} = \frac{\bar{y}}{\bar{x}} ,$$

and the corresponding estimate of the population total volume is  $\hat{Y} = \hat{R}X = \frac{\bar{y}}{\bar{x}}X$  where  $X$  is the total weight of all truckloads.

A large-sample approximation to the variance of the ratio estimate is

$$Var(\hat{R}) \doteq \left( \frac{1-f}{n\bar{X}^2} \right) \left( \frac{1}{N-1} \right) \sum_{i=1}^N (y_i - Rx_i)^2 ,$$



and, since the population quantity  $\bar{X}$  is known, the natural sample analogue of this is

$$var(\hat{R}) = \left( \frac{1-f}{n\bar{X}^2} \right) \left( \frac{1}{n-1} \right) \sum_{i=1}^n (y_i - \hat{R}x_i)^2 .$$

Since  $\hat{Y} = \hat{R}X = \hat{R}N\bar{X}$ , the corresponding large-sample approximation to the variance of the estimate of the population total is

$$Var(\hat{Y}) = \left( \frac{1-f}{n} \right) \left( \frac{N^2}{N-1} \right) \sum_{i=1}^N (y_i - Rx_i)^2 ,$$

for which the sample analogue is

$$var(\hat{Y}) = \left( \frac{1-f}{n} \right) \left( \frac{N^2}{n-1} \right) \sum_{i=1}^n (y_i - \hat{R}x_i)^2 .$$

To determine the total sample size required for a company, the Ministry of Forests uses "estimates" of the required population parameters. These "estimates" are really anticipatory figures, possibly being based on previous year's data or from a few samples at the beginning of the year. After determining the total sample size required for all strata, sample sizes are determined for each individual stratum by Neyman allocation (Cochran, 1977, p. 99). If the estimated volumes and standard deviations required to implement Neyman allocation are not available, then the total sample size is distributed to the individual strata based on the anticipated number of loads to be harvested from each stratum (proportional allocation). However, because the variance estimates are based on large sample approximations, all sample sizes which are less than 30 are increased to at least 30, and often, to 35. Thus, there is an incentive for companies to select large strata; otherwise they will be required to sample more truckloads (and incur the additional costs associated with scaling these truckloads) so that the variance estimates are accurate.

The total volume harvested by a company is estimated by summing the estimates of the total volume for the  $m$  strata,

$$\hat{Y} = \sum_{i=1}^m \hat{Y}_i$$

and the variance estimates from all strata are added to obtain the variance estimate of the total volume:

$$var(\hat{Y}) = \sum_{i=1}^m var(\hat{Y}_i)$$

The above are the analytic formulae used by the Ministry of Forests. In the next section we will discuss both statistical and non-statistical problems associated with the presently used scaling system.

## 1.4 Implementation Problems

Several practical, implementation type problems arise as a result of difficulties in properly applying the required methods, whereas other problems arise simply because the underlying statistical theory is not correct for the problem at hand. Though this research deals with the statistical theory upon which the scaling system is based, it is nevertheless important to understand some of the implementation issues, simply because the presence of inconsistencies between theory and practice reduces the effectiveness of the scaling system.

The implementation problems can be categorized into two types: system errors and intentional errors. Intentional errors are those errors in sampling, weighing or data collection which are either committed on purpose or allowed to go uncorrected. The Ministry of Forests collects over \$500 million in stumpage fees annually. The sheer magnitude of this figure implies that there may be some groups who will try to reduce the amount they are required to pay. However, these

are legal rather than statistical problems. Consequently, for the purposes of this research, we will not consider these problems.

System errors are those errors arising from the actual implementation of the scaling system.

These errors can be classified into three general categories: misclassification errors, measurement errors and random sampling errors. An alternate way of viewing these errors is: errors due to measuring what should not be measured, errors due to mistakes in measuring what should be measured and the errors arising from having a sample which does not adequately represent the true population.

Misclassification errors are those errors due to misclassifying a truckload of timber into the wrong stratum. Recall that when a truck arrives at a mill, it is assigned to a stratum depending the geographic source of the logs, the species and age mix, amount of rot, amount of pest damage and other criteria. However, this decision is subject to human error. For example, a truckload of mostly hemlock and a bit of fir may be misclassified as belonging to a stratum containing fir only. Outliers in sample data could be due to measurement errors, to misclassification errors, to other errors or they might be a feature of the population. Therefore, it is difficult to ascertain whether misclassification errors constitute a major source of variability. One method of determining the rate of misclassification errors is to search for outliers in the sample data, although this would involve an implicit assumption that outliers within a stratum are due to misclassification errors and are not simply a feature of the population.

Measurement errors are those errors arising primarily from inaccuracies due to the measuring procedure used by scalers in measuring the volume of a truckload and from the errors due to weighing a truckload. As mentioned earlier, two scalers measuring a load of timber will occasionally produce volume estimates with substantial differences. However, the Ministry staff

feel that scaler to scaler differences do not constitute a major source of variability. Consequently, we will disregard scaler to scaler error. The errors due to weigh scales are not likely to be substantial when considered in relation to other sources of error. The weigh scales are electronic and are (presumably) regularly calibrated. Thus, it is quite likely that scaler to scaler variability exceeds weigh scale error, and if we are disregarding the former, we may also disregard the latter. It should be noted that these sources of error are not ignored; we will simply make no attempt to isolate them. All these sources of error are aggregated and considered as the sample to sample variability in the ratio.

Random selection errors are those errors which cause an improper sample selection - aside from intentional "errors". A primary source of errors here are the anticipatory figures used at the beginning of the year to determine the sample size breakdown between the strata. For example, a company may anticipate harvesting a certain number of truckloads from a stratum, and at midyear may realize they have overestimated - that in fact they will harvest fewer truckloads from this stratum. Then, during mid-year updates, the Ministry may reduce the sampling rate for that stratum for the rest of that year. Consequently, the wood harvested during the latter half of the year would be under-represented in the sample from that stratum. And since there often exists a difference between the type of wood harvested during the summer and winter seasons, this may turn out to be a major source of error. Another problem is that truckloads may not arrive at a weigh scale in random order. In other words, the semi-systematic method of sampling used may not be as good as a true simple random sample. As Kish and Frankel (1974) note, "'well-mixed' urns are seldom provided by nature or created by man".

## **1.5 Problem of Interest**

In this thesis, we will concentrate on the estimation of the variance of the estimate of total volume. Recall that in the current method of estimating the variance of an estimate of total volume for a

stratum, the Ministry requires that the sample size exceed a minimum of 30 (see Cochran, 1977, page 153). One question we will address is whether the variance estimator we described earlier is accurate when the sample size is less than 30. We will also investigate alternate estimators of the variance, particularly those estimators which perform well at sample sizes less than 30. We believe that identification of an effective small sample estimator of the variance would benefit both the forest companies of British Columbia and the B.C. Ministry of Forests. The benefit would arise because the minimum sample size requirement could be reduced thereby allowing companies to select smaller and possibly more homogenous strata. Such strata would require fewer samples than larger, more heterogeneous strata and the overall result would be greater efficiency in the weigh scaling process.

The primary motivation for our thesis is the possibility that recently developed resampling techniques might provide accurate and effective variance estimators for small samples. These techniques (technically known as the Bootstrap and the Jackknife) have not been studied in detail in similar problems, but previous empirical studies suggest that these methods might be effective in our situation. Consequently, we will derive estimators based on these new techniques and we will study their small sample behaviour via an empirical study.

In the next chapter, we will review various methods of both estimating the population total volume and of estimating the variance of these estimates. This will be done in a general statistical context. Thereafter, in Chapter 3, we will derive weighted jackknife-based variance estimators. As one author has pointed out, "the weighted jackknife...would seem to be particularly appropriate for sample survey applications" (Smith, 1981). In Chapter 4 we will continue with our derivation of resampling-based estimators, particularly in the context of ratio estimation. It has been suggested that bootstrap techniques are more accurate than jackknife techniques (Efron, 1979), and thus this chapter will attempt to derive bootstrap analogues to the estimators obtained in

Chapter 3. In Chapter 5, we will conduct empirical studies on the performance of the various variance estimators in the context of the types of populations faced by the Ministry of Forests. In addition, we will also study the robustness of the various estimators, particularly with respect to common model assumptions. Finally, in Chapter 6, we will summarize the results of this study and provide specific recommendations concerning the statistical aspects of the weigh scaling process currently used by the Ministry of Forests.

## Chapter 2

### Estimating a Population Total via Ratio Estimators

In Chapter 1, we described and discussed the implementation of the weigh scaling program by the B.C. Ministry of Forests. In this chapter, we will discuss weigh scaling from a statistical point of view, concentrating both on the estimation of total volume via ratio estimators and on the estimation of the variability of these estimators.

The weigh scaling program may be statistically represented as the following estimation problem:

A finite population consists of  $N$  units of two associated variables of interest:  $(X_i, Y_i)$ ,  $i=1,2,\dots,N$

with corresponding population totals  $Y = \sum_{i=1}^N Y_i$  and  $X = \sum_{i=1}^N X_i$ . In our context, the units of the

population are the truckloads and the variables  $X_i, Y_i$  are, respectively, the weight and volume of

the  $i^{th}$  truckload. Having observed a simple random sample of  $n$  pairs of  $X_i, Y_i$  we wish to

estimate both the total volume  $Y$  and the variance of our estimate of total volume,  $Var(\hat{Y})$ , given that we know  $X$ .

In this thesis we will focus on the variance estimation problem given that a particular approach will be taken for the estimation of the population total. That particular approach is to estimate

the population total volume,  $Y$ , by the classical ratio estimator  $\hat{Y}_R = \hat{R}_c X = \frac{\bar{y}}{\bar{x}} \cdot X$ .

However, there are other possible estimators of  $Y$ , and we will briefly discuss some of these before proceeding to our main topic of interest: the various proposed estimators of  $Var(\hat{Y}_R)$ . Our reason

for discussing alternate estimators of  $Y$  besides  $\hat{Y}_R$ , is mainly to place the variance estimation problem in context. The problem we will attempt to address is a small part of a much larger problem: the problem of drawing inference on  $Y$  given a sample of pairs of  $X_i, Y_i$  and given fairly complete information on the covariates  $X_1, \dots, X_N$ .

## 2.1 Estimating a Population Total

There are two common approaches for estimating  $Y$ : One approach ignores covariate information, while the other takes advantage of this information. The simplest estimator of the former approach, known as the mean-per-unit method, first estimates the population average of  $Y_i$ , and then multiplies it by  $N$ . In our context, this amounts to first estimating the average volume per truckload, and then multiplying this by the total number of truckloads. We will denote this estimator by  $\hat{Y}_M = N\bar{y}$ , where  $\bar{y}$  is the average of the observed  $y_1, \dots, y_n$ .

The alternate approach is to make use of covariate information. There are many ways of making use of covariate information, the simplest of which is as follows:  $\hat{Y}_R = \frac{\bar{y}}{\bar{x}}X$ . For obvious reasons,

this is known as the ratio method. In our context, this amounts to first estimating the inverse of the average density and then multiplying this by the total weight,  $X$ , to estimate total volume.

Quite often, a model-based approach is used for estimating a population total. Under this approach, the expectation of  $Y_i$  is assumed to be a function of  $X_i$ . Consequently, it is natural to estimate  $Y$  using our knowledge of  $X_i$ 's. The model typically used is a linear model with specific variance assumptions:



$$Y_i = X_i\beta + e_i, \quad (2.1)$$

where the  $e_i$  are the stochastic components. The  $e_i$  are assumed to have expectation zero, and a common assumption is that their variance given  $x_i$  is  $\sigma^2 x_i$ . If we assume that the  $e_i$  are normally distributed, then the maximum likelihood estimator of  $\beta$  is the classical ratio estimator  $\hat{R}_c = \frac{\bar{y}}{\bar{x}}$ .

Even without the normality assumption, weighted least squares results in  $\beta = \hat{R}_c$ .

An appropriate question to ask here is: under what circumstances is  $\hat{Y}_M$  a better estimator of  $Y$  than is  $\hat{Y}_R$ ? It can be shown using the design-based approach (Cochran, 1977, p. 157) that for simple random sampling with large samples, the ratio estimate  $\hat{Y}_R$  has a smaller variance than  $\hat{Y}_M$  if

$$\rho_{xy} > \frac{1}{2} \left( \frac{S_x}{\bar{X}} \right) \left( \frac{\bar{Y}}{S_y} \right)$$

where  $S_x, S_y$  are the standard deviations in the  $X$  and  $Y$  populations respectively, and  $\rho_{xy}$  is the correlation between  $X_i$  and  $Y_i$  in the population.

For the total volume estimation problem of the B.C. Ministry of Forests, the ratio estimator would be expected to be the better of the two estimators because, in general, the correlation between  $X_i$  and  $Y_i$  is large, while the two coefficients of variation are similar. Hence, we will concentrate on

ratio type estimators of the population total:  $\hat{Y}_* = \hat{R}_* X$ , where the  $*$  indicates a particular method

of estimating  $R = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}}$ .

## 2.2 Some Ratio Estimators

Ratio estimators have been studied for several decades. As early as 1932, a non-symmetric confidence interval for  $R = E(Y_i)/E(X_i)$  was being studied under an infinite population and bivariate normal assumption (Fieller, 1932). During the 1950's, numerous researchers approached the ratio estimation problem through the design-based framework. In this section, we will briefly discuss several ratio estimators.

The most commonly used method of estimating  $R$  is by the classical ratio estimator,  $\hat{R}_c$  which is given by the ratio of the average of the observed  $Y_i$ 's divided by the average of the observed  $X_i$ 's:

$$\hat{R}_c = \frac{\bar{y}}{\bar{x}}, \text{ where } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \text{ and } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

It has been shown that the ratio of the averages is a better estimator of  $R$  than the average of the ratios. If we assume that  $\text{Var}(Y_i) \propto X_i$  and that the relationship of  $Y_i$  and  $X_i$  is a straight line through the origin, then it can easily be shown (by using Lagrange Multipliers) that the estimate of the population total obtained via the classical ratio estimator is the best linear unbiased estimator of the population total under the design-based approach (Cochran, 1977, p. 158); Brewer (1963) and Royall (1970) show this for finite populations. Under the model-based approach, the result is of course, well-known from regression theory.

However, the classical ratio estimator is biased. Using an alternate approach, a design-unbiased estimator of  $R$  was obtained by Hartley and Ross (1954). Their estimator is obtained by correcting

for the bias in the average of the ratios:  $\bar{r} = \frac{1}{n} \sum_{i=1}^n r_i$ , where  $r_i = y_i/x_i$ . Note that:

$$\frac{1}{N} \sum_{i=1}^N r_i(x_i - \bar{X}) = \bar{Y} - \bar{X}E(r_i) = \bar{X}[R - E(r_i)] .$$

Thus, the bias in  $\bar{r}$  can be expressed as

$$E(\bar{r}) - R = E(r_i) - R = - \frac{1}{N\bar{X}} \sum_{i=1}^N r_i(x_i - \bar{X}) . \quad (2.2)$$

By Theorem 2.3 of Cochran (1977), it can easily be seen that an unbiased estimate of the above bias is

$$\left( \frac{-1}{N\bar{X}} \right) \cdot \frac{N-1}{n-1} \sum_{i=1}^n r_i(x_i - \bar{x}) = \left( \frac{-1}{N\bar{X}} \right) \frac{(N-1)n}{n-1} \cdot (\bar{y} - \bar{r}\bar{x}) . \quad (2.3)$$

Then by substituting (2.3) into (2.2),

$$E(\bar{r}) - R = E \left\{ \frac{-1}{N\bar{X}} \cdot \frac{(N-1)n}{n-1} \cdot (\bar{y} - \bar{r}\bar{x}) \right\} .$$

Therefore  $\hat{R}_{HR} = \bar{r} + \frac{(N-1)n}{N\bar{X}(n-1)} (\bar{y} - \bar{r}\bar{x})$  is an unbiased estimator of  $R$ .

In a similar attempt, Mickey (1959) also derived a design-unbiased estimator of the ratio. His estimator was obtained in an effort to find a class of unbiased ratio and regression estimators and

is derived using a "leave-one-out" approach. By estimating and correcting for the bias in the average of the "leave-one-out" estimates of R,

$$\hat{R}_\cdot = \frac{1}{n} \sum_{i=1}^n \frac{n\bar{y} - y_i}{n\bar{x} - x_i} = \frac{1}{n} \sum_{i=1}^n \hat{R}_{\cdot i} ,$$

Mickey obtains the following design-unbiased estimator of R:

$$\hat{R}_M = \hat{R}_\cdot + \frac{(N-n+1)n}{N\bar{X}} (\bar{y} - \hat{R}_\cdot \bar{x}) .$$

Tin (1965) took a different approach to correcting for the bias in  $\hat{R}_\cdot$ . Note that

$$\begin{aligned} \hat{R}_c &= \frac{\bar{y}}{\bar{x}} = \frac{\bar{y}}{\bar{x}} \cdot R^{-1} \cdot R = \frac{(\bar{y} / \bar{Y})}{(\bar{x} / \bar{X})} \cdot R \\ &= R \left[ 1 + \frac{(\bar{y} - \bar{Y})}{\bar{Y}} \right] \left[ 1 + \frac{(\bar{x} - \bar{X})}{\bar{X}} \right]^{-1} \\ &= R \left[ 1 + \frac{(\bar{y} - \bar{Y})}{\bar{Y}} \right] \left[ 1 - \frac{(\bar{x} - \bar{X})}{\bar{X}} + \frac{(\bar{x} - \bar{X})^2}{\bar{X}^2} + \dots \right] \\ &= R \left[ 1 + \frac{(\bar{y} - \bar{Y})}{\bar{Y}} - \frac{(\bar{x} - \bar{X})}{\bar{X}} - \frac{(\bar{y} - \bar{Y})(\bar{x} - \bar{X})}{\bar{Y}\bar{X}} + \frac{(\bar{x} - \bar{X})^2}{\bar{X}^2} + \dots \right] . \end{aligned}$$

Taking expectations leads to

$$\begin{aligned} E(\hat{R}_c) &= R \left[ 1 - \frac{E(\bar{y} - \bar{Y})(\bar{x} - \bar{X})}{\bar{Y}\bar{X}} + \frac{E(\bar{x} - \bar{X})^2}{\bar{X}^2} \right] \\ &= R \left\{ 1 - \left( \frac{1-f}{n} \right) \left[ \frac{S_{xy}}{\bar{X}\bar{Y}} - \frac{S_x^2}{\bar{X}^2} \right] \right\} , \end{aligned}$$

which yields an approximation to the bias in  $\hat{R}_c$ . Replacing all population quantities in the approximation by their natural estimates and subtracting, leads to the estimator suggested by Tin:

$$\hat{R}_T = \hat{R}_c \left\{ 1 + \left( \frac{1-f}{n} \right) \left[ \frac{s_{xy}}{\bar{x}\bar{y}} - \frac{s_x^2}{\bar{x}^2} \right] \right\},$$

$$\text{where } s_{xy} = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

$$s_x^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2, \text{ and } f = \frac{n}{N}.$$

It is interesting to note that Tin's estimator can also be obtained through a one-term Taylor expansion of the denominator of an estimator suggested earlier by Beale (1962):

$$\hat{R}_B = \hat{R}_c \frac{\left[ 1 + \left( \frac{1-f}{n} \right) \frac{s_{xy}}{\bar{x}\bar{y}} \right]}{\left[ 1 + \left( \frac{1-f}{n} \right) \frac{s_x^2}{\bar{x}^2} \right]}.$$

Using a general bias-reducing technique known as the jackknife (which is discussed in more detail in Chapter 3), Durbin (1959) obtained the following estimator of R:

$$\hat{R}_J = n\hat{R}_c - \left( \frac{n-1}{n} \right) \hat{R}.$$

Durbin was able to show that under some distributional assumptions, the jackknife method leads to estimators which are less biased and more efficient than the classical estimator  $\hat{R}_c$ . Rao and Webster (1966) showed that the above "leave-one-out" method is the optimal version of the jackknife for ratio estimation.

Whereas Mickey's method estimates the overall bias with explicit use of covariate information, the jackknife method estimates and attempts to eliminate first order bias without the direct use of covariates. Although both Mickey's estimator and the jackknife estimator use  $\hat{R}$ , the two methods arise from different approaches. Mickey's estimator arises out of a design-based, general methodology to produce unbiased ratio and regression estimators. On the other hand, the jackknife is a general, all-purpose technique designed to eliminate or reduce the bias of an estimator.

### **2.3 Review of Empirical Studies of Ratio Estimators**

Numerous empirical studies have been conducted on the preceding estimators of  $R$ . An empirical study by Hutchison (1971) showed that if  $X_i$ 's are simulated as log normal, and if the errors are simulated as normal with zero mean and variance proportional to  $\sigma^2 x^\lambda$ , where  $\lambda = \{0, 0.5, 1, 1.5, 2\}$ , then the Hartley-Ross estimator is as efficient as any of the other estimators. If the  $X_i$ 's are simulated to have a gamma distribution, Rao and Rao (1971) conclude that Mickey's unbiased estimator is more efficient than the Hartley-Ross estimator; this is especially so when  $\lambda = 0$ . Hutchison (1971) and Rao and Rao (1971) show that though Mickey's estimator is preferable to the Hartley-Ross estimator, it is not as good as some other estimators; in particular, the Beale, Tin and jackknife estimators. The jackknife estimator has been studied on numerous occasions and the general conclusion is that it is a very good estimator (Durbin, 1959; Rao and Webster, 1966; P.S.R.S. Rao, 1969; Rao and Rao, 1971; Hutchison, 1971; and, Rao, 1979). Another pair of estimators which have been shown in the above studies to be accurate and efficient are Tin's and Beale's estimators. Both have bias of order  $n^{-2}$  and have estimated variances comparable to the estimated variance of the jackknife estimator.

It is clear from the above discussion that there are numerous estimators of the ratio. But our objective in this thesis is to consider estimation of the variance in estimating a population total using some form of the ratio estimator. To keep the problem to a manageable scale, we shall restrict our attention to the classical estimate of the ratio,  $\hat{R}_c$ . Henceforth, all references to  $\hat{R}$  and  $\hat{Y}$  will imply estimation via the classical ratio estimator method.

The techniques we shall use in this thesis to develop new variance estimators for the estimate of population total using the classical ratio method could also be applied to develop new variance estimators for other methods of ratio estimation.

## 2.4 Estimation of the Variance of the Classical Ratio Estimator of a Population Total

The problem to be addressed in this thesis is the estimation of the variance of the estimator of total volume given by:

$$\hat{Y} = \hat{R}X = \frac{\bar{y}}{\bar{x}} X .$$

This estimate is simply an estimate of  $R$  scaled by  $X$ , and thus,  $Var(\hat{Y}) = X^2 Var(\hat{R}) = (N\bar{X})^2 Var(\hat{R})$ .

It can easily be shown that the approximate variance of  $\hat{R}$  under simple random sampling is:

$$Var(\hat{R}) \doteq \left( \frac{1-f}{n\bar{X}^2} \right) \cdot \left( \frac{\sum_{i=1}^N (Y_i - RX_i)^2}{N-1} \right) .$$

This approximation has a bias of order  $n^{-1}$ . The exact variance of  $\hat{R}$  cannot be evaluated under the design-based approach because both  $X_i$  and  $Y_i$  are random; the difficulty arises because  $\hat{R}$  contains random variables in both the numerator and the denominator.

Therefore the approximate variance of  $\hat{Y}$  (with a bias of order  $n^{-1}$ ) is:

$$\begin{aligned} \text{Var}(\hat{Y}) &= (N\bar{X})^2 \left( \frac{1-f}{n\bar{X}^2} \right) \left( \frac{\sum_{i=1}^N (Y_i - RX_i)^2}{N-1} \right) \\ &= \left( \frac{N^2(1-f)}{n} \right) \left( \frac{\sum_{i=1}^N (Y_i - RX_i)^2}{N-1} \right). \end{aligned}$$

By estimating the sum of the squared residuals, the natural sample analogue of  $\text{Var}(\hat{R})$  is

$$\text{var}_0(\hat{R}) = \left( \frac{1-f}{n\bar{X}^2} \right) \left( \frac{\sum_{i=1}^n (y_i - \hat{R}x_i)^2}{n-1} \right);$$

and therefore,

$$v_0 = \text{var}_0(\hat{Y}) = \left( \frac{N^2(1-f)}{n} \right) \left( \frac{\sum_{i=1}^n (y_i - \hat{R}x_i)^2}{n-1} \right).$$

In some ratio estimation contexts,  $\bar{X}$  is not known and therefore not available for the evaluation of

$\text{var}_0(\hat{R})$ . Estimation of  $\bar{X}$  by  $\bar{x}$  leads to the alternate estimator of  $\text{Var}(\hat{R})$ :

$$\text{var}_2(\hat{R}) = \left( \frac{1-f}{n\bar{x}^2} \right) \left( \frac{\sum_{i=1}^n (y_i - \hat{R}x_i)^2}{n-1} \right),$$

which in turn leads to



$$v_2 = var_2(\hat{Y}) = \left( \frac{N^2(1-f)}{n} \right) \left( \frac{\bar{X}^2}{\bar{x}^2} \right) \left( \frac{\sum_{i=1}^n (y_i - \hat{R}x_i)^2}{n-1} \right).$$

Results from empirical studies on the performance of  $v_0$  and  $v_2$  as estimators of  $MSE(\hat{R})$  are mixed. In terms of squared error, some studies suggest that  $v_0$  is a better estimator (e.g., Krewski and Chakrabarty, 1981), whereas other studies suggest that  $v_2$  is a better estimator (e.g., Wu and Deng, 1983). It is difficult to ascertain which estimator performs better. However, it is thought that both methods seriously underestimate the  $MSE(\hat{R})$  for smaller sample sizes, with relative bias on the order of -15% (Cochran, 1977; Rao, 1968).

Recall that under the assumption of a linear model where the variance of  $Y_i$  given  $x_i$  is  $\sigma^2 x_i$ ,  $\hat{R}$  is the weighted least squares estimator of  $\beta$ . Assuming the linear model of (2.1) and assuming that the population values  $y_1, \dots, y_N$  are realized values of random variables  $Y_1, \dots, Y_N$ , one obtains as the variance of the usual estimate of population total (Royall, 1970; Royall and Eberhardt, 1975):

$$Var_{wLS}(\hat{Y}) = \sigma^2 \left( \frac{N^2(1-f)}{n} \right) \left( \frac{\bar{X}}{\bar{x}} \right) \left( \frac{N\bar{X} - n\bar{x}}{N-n} \right). \quad (2.4)$$

An unbiased estimator of  $var_{wLS}(\hat{Y})$  is

$$var_{wLS}(\hat{Y}) = \hat{\sigma}^2 \left( \frac{N^2(1-f)}{n} \right) \left( \frac{\bar{X}}{\bar{x}} \right) \left( \frac{N\bar{X} - n\bar{x}}{N-n} \right), \quad (2.5)$$

where

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2 x_i^{-1}.$$

(This linear model approach towards finite population sampling is known as the prediction theory approach. For more details, see Royall, 1970, 1971). This estimate is biased under violations of the variance assumption (Royall and Eberhardt, 1975). The empirical study of Royall and Cumberland (1981) suggests that this estimate of variance consistently and significantly underestimates the variance of the total volume estimate.

Under the prediction theory approach (where  $Y_i$ 's are random and  $x_i$ 's are not), Royall (1971) has shown that

$$E(v_0) = \sigma^2 \left( \frac{N^2(1-f)}{n} \right) \left( 1 - \frac{1}{n} V_x^2 \right) \bar{x} , \quad (2.6)$$

where

$$V_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \bar{x}^{-2} ,$$

and the expectation is with respect to the underlying distribution of the random  $Y_i$ 's, and not the distribution imposed by the sampling design.

Note that whereas (2.4) is a decreasing function of the sample mean  $\bar{x}$ , (2.6) is an increasing function of  $\bar{x}$ . This may explain the poor performance of (2.5) in Royall and Cumberland's (1981) empirical study. Under the prediction theory approach,  $v_0$  is biased and the following estimator removes its bias (Royall and Eberhardt, 1975):

$$var_H(\hat{Y}) = var_0(\hat{Y}) \left( \frac{\bar{X}}{\bar{x}} \right) \left( \frac{N\bar{X} - n\bar{x}}{N-n} \right) \left( 1 - \frac{1}{n} V_x^2 \right)^{-1} .$$

However,  $var_H(\hat{Y})$  is biased (under the prediction theory approach) under violations of the variance assumption. In an attempt to robustify  $var_H(\hat{Y})$  to these violations, Royall and Cumberland (1978) suggest a new estimator which is approximately unbiased for different assumptions of variance:

$$var_D(\hat{Y}) = \left( \frac{N^2(1-f)}{n^2} \right) \left( \sum_{i=1}^n \frac{(y_i - \hat{R}x_i)^2}{\left( 1 - \frac{x_i}{n\bar{x}} \right)} \right) \left( \frac{\bar{X}}{\bar{x}^2} \right) \left( \frac{N\bar{X} - n\bar{x}}{N-n} \right).$$

The empirical study of Royall and Cumberland suggests that  $var_D(\hat{Y})$  has a lower bias than  $var_H(\hat{Y})$ , at least for the populations contained in their study. Wu and Deng's (1983) study suggests that  $var_H(\hat{Y})$  has a lower root mean squared error than  $var_D(\hat{Y})$ . Both these studies present results for the same sample sizes ( $n=32$ ) and for the same populations. Thus, it appears that the bias robustness of  $var_D(\hat{Y})$  is achieved at the expense of increased instability.

Wu (1982) suggested a new estimator which is the geometric mean of  $var_0(\hat{Y})$  and  $var_2(\hat{Y})$ ; that is:

$$v_1 = var_1(\hat{Y}) = \left( \frac{N^2(1-f)}{n} \right) \left( \frac{\bar{X}}{\bar{x}} \right) \left( \frac{\sum_{i=1}^n (y_i - \hat{R}x_i)^2}{n-1} \right).$$

Note that the only difference between  $v_0$ ,  $v_1$  and  $v_2$  is in the value of the exponent of  $(\bar{X}/\bar{x})$ .

Extending this, Wu (1982) suggested using the data to estimate the optimal value of this exponent; optimal in the sense of minimizing the leading terms in the asymptotic expansion of the mean

squared error of  $v_g = \left( \frac{\bar{X}}{\bar{x}} \right)^g \cdot var_0$ . The sample analogue is  $\hat{g}$  as described below. For more detail,

see Wu (1982) and Wu and Deng (1983). This new estimator is

$$var_{\hat{g}}(\hat{Y}) = \left( \frac{N^2(1-f)}{n} \right) \left( \frac{\bar{X}}{\bar{x}} \right)^{\hat{g}} \left( \frac{\sum_{i=1}^n (y_i - \hat{R}x_i)^2}{n-1} \right),$$

where  $\hat{g}$  = sample regression coefficient of  $\frac{\bar{z}_i}{\bar{z}}$  on  $\frac{x_i}{\bar{x}}$  ,

$$\bar{z}_i = (y_i - \hat{R}x_i)^2 - 2(y_i - \hat{R}x_i) \cdot \frac{\sum_{i=1}^n x_i (y_i - \hat{R}x_i)^2}{\sum_{i=1}^n x_i}, \text{ and}$$

$$\bar{z} = \frac{1}{n} \sum_{i=1}^n \bar{z}_i .$$

A similar estimator suggested by Wu and Deng (1983) is obtained by removing the second term in the above  $z_i$ ; then  $z_i$  is simply the squared residual  $(y_i - \hat{R}X_i)^2$ . This leads to:

$$var_{\tilde{g}}(\hat{Y}) = \left( \frac{N^2(1-f)}{n} \right) \left( \frac{\bar{X}}{\bar{x}} \right)^{\tilde{g}} \left( \frac{\sum_{i=1}^n (y_i - \hat{R}x_i)^2}{n-1} \right),$$

where  $\tilde{g}$  = sample regression coefficient of  $\frac{(y_i - \hat{R}x_i)^2}{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{R}x_i)^2}$  on  $\frac{x_i}{\bar{x}}$  .

The empirical study by Wu and Deng (1983) suggests that  $var_{\hat{g}}(\hat{Y})$  is not as good an estimator as  $var_{\tilde{g}}(\hat{Y})$ .

Fuller (1981) suggested a new estimator which was motivated by the following observation: If the variance of  $e_i$  is a function of  $x_i$ , then  $e_i^2$  is correlated with  $(x_i - \bar{X})$ . Thus, Fuller suggests that because the population mean of the  $x$ 's,  $\bar{X}$ , is known, the parameter to estimate should be the variance of  $e_i$  conditional on  $\bar{X}$ . A linear approximation of the conditional variance leads to the following estimator of the variance of  $\hat{Y}$ :

$$var_{REG}(\hat{Y}) = \left( \frac{n^2(1-f)}{n} \right) \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2 - N\hat{b}(\bar{x} - \bar{X}) .$$

$$= var_0(\hat{Y}) + N\hat{b}(\bar{X} - \bar{x}) ,$$

where  $\hat{b}$  = sample regression coefficient of  $(y_i - \hat{R}x_i)^2$  on  $x_i$  .

Wu and Deng's (1983) empirical study suggests that this is a very good estimator.

The jackknife method, which was briefly mentioned in Section 2.2, also provides an estimate of variance. The idea here is to estimate the variance of  $\hat{R}$  by the appropriately scaled sample variance of the "leave-one-out" estimates of the ratio. (This will be discussed in more detail in Chapter 3.) The jackknife estimate of variance is:

$$var_J(\hat{Y}) = \left( \frac{N^2(1-f)}{n} \right) \bar{X}^2 (n-1) \sum_{i=1}^n (\hat{R}_{-i} - \hat{R}_{\cdot})^2 .$$

Both the empirical studies by Royall and Cumberland (1981) and by Wu and Deng (1983) show that the jackknife variance estimate performs very well.

## 2.5 Summary

There are numerous estimators of both the ratio and of the variance of the classical ratio estimator. These estimators are derived from certain assumptions about the type of situations in which the estimators will be applied. For example, Wu's variance estimators assume that the variance of  $e_i$  is  $\sigma^2 x_i$  and that the sample sizes are large enough to ensure approximate validity of the asymptotic expansions. The success of these estimators in any given situation depends upon the particulars of that problem. Because we are interested in applying these estimators in situations with a variety of population characteristics, it may be that resampling-based estimators will perform best over all the populations. Consequently, in the next two chapters we will concentrate upon deriving resampling-based estimators of the variance of the classical ratio estimator.

## **Chapter 3**

### **Jackknifing in Ratio Estimation**

In this chapter, we will concentrate on jackknife variance estimators for the classical ratio estimator. Empirical studies suggest that the jackknife variance estimator presented in Chapter 2 performs reasonably well in estimating the true variance of a ratio estimate. Recently, variations on the jackknife method have been proposed; the new versions are commonly known as weighted jackknife methods. We conjecture that these methods will lead to better variance estimators in ratio estimation.

In Section 1, we will briefly discuss the (unweighted) jackknife. In Section 2, using linear models at first, we will show how and why the unweighted jackknife estimator does not perform well in the presence of imbalance; especially when the linear model, which is already unbalanced, is further perturbed by heteroscedasticity. Linear models will be used because they include a wide range of models, including the commonly used model for ratio estimation (already discussed in Chapter 2), and because they exhibit imbalance of a type which is probably similar to the imbalance in ratio estimation. The imbalance referred to is the absence of identical distributions. That is, the observations have different means and sometimes, different variances.

After discussing the problems arising from the use of the jackknife in unbalanced problems, we will, in Section 3, discuss the weighted jackknife which attempts to reduce some of the difficulties which arise due to imbalance. We will review and expand upon the weighted jackknife variance estimators for linear models suggested in Hinkley (1977) and Wu (1986). Thereafter, in Section 4, we will derive the weighted jackknife estimators for the special case of the linear model discussed in Chapter 2. Finally, in Section 5, we will derive weighted jackknife variance estimators for the

ratio estimator based on the approaches of Hinkley (1977) and Wu (1986). We will derive three weighted jackknife variance estimators, two of which will be shown to be equivalent to common design-based variance estimators.

### 3.1 Introduction to Jackknife

The jackknife estimator of the ratio,  $\hat{R}_J$ , was described in Chapter 2 along with the jackknife estimator of the variance of the classical ratio estimator. In this section, we will present a more thorough and formal description of jackknife estimators.

Let  $\theta$  be the parameter of interest where  $\theta = \theta(F)$  and  $F$  is the distribution of the random variable  $Z$ . Suppose we have an estimator of  $\theta$  based on a random sample  $z_1, \dots, z_n$  denoted by  $\hat{\theta}(z_1, \dots, z_n) = \hat{\theta}$ .

Let  $\hat{\theta}_{-i}$  be the estimate of  $\theta$  based on the sample with the  $i^{\text{th}}$  observation removed:

$\hat{\theta}_{-i} = \hat{\theta}(z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n)$ . Then the pseudovalue for this "leave-one-out" case is defined as (Quenouille, 1956)

$$P_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}.$$

The jackknife estimator is obtained by averaging these pseudovalues:

$$\hat{\theta}_J = P_{\cdot} = \frac{1}{n} \sum_{i=1}^n P_i = n\hat{\theta} - (n-1) \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{-i} = n\hat{\theta} - (n-1)\hat{\theta}_{\cdot}.$$

In general,  $\hat{\theta}_J$  removes the order  $n^{-1}$  term bias (Miller, 1974a) from biased estimators having expectation of the form

$$E(\hat{\theta}) = \theta + \frac{a_1}{n} + O\left(\frac{1}{n^2}\right).$$



This form of the jackknife is often called the balanced jackknife because there is an implicit assumption that each observation has an equal effect on the parameter estimate. Thus, the parameter estimate is the unweighted average of the pseudovalues. This form of the jackknife usually works very well in situations with some natural balance.

The above pseudovalues and the corresponding jackknife estimates were proposed in a more general form in Quenouille (1949, 1956). The original version dealt with breaking up  $n$  observations into  $g$  groups of size  $h$ , so that  $gh = n$ , and deleting the  $i^{\text{th}}$  group of observations to obtain  $\hat{\theta}_{-i}$ . The more general pseudovalues and corresponding estimator are

$$\begin{aligned} P_i &= g\hat{\theta} - (g-1)\hat{\theta}_{-i} , \\ \hat{\theta}_J &= g\hat{\theta} - (g-1)\hat{\theta}_{\cdot} , \\ \text{where } \hat{\theta}_{\cdot} &= \frac{1}{g} \sum_{i=1}^g \hat{\theta}_{-i} . \end{aligned}$$

Quenouille's original motivation for introducing the jackknife estimate was to eliminate bias of order  $n^{-1}$ . He also suggested similar estimates to remove higher order bias; however we will only deal with the above "leave-one-out" version.

Tukey (1958) proposed treating the pseudovalues,  $P_i$ , as independent and identically distributed random variables, thereby obtaining the jackknife estimate of the variance of  $\hat{\theta}_J$ :

$$\begin{aligned} \text{var}(\hat{\theta}_J) &= \{n(n-1)\}^{-1} \sum_{i=1}^n (P_i - P_{\cdot})(P_i - P_{\cdot})^t , \\ &= \left(\frac{n-1}{n}\right) \sum_{i=1}^n (\hat{\theta}_{-i} - \hat{\theta}_{\cdot})(\hat{\theta}_{-i} - \hat{\theta}_{\cdot})^t . \end{aligned}$$

The variance estimator  $var_j(\hat{Y})$  which was discussed in Chapter 2, was simply an application of the above to the classical ratio estimator, with the finite population correction term appended.

In both this and the following chapters, whenever variance estimators are used in finite populations, the finite population correction term,  $(1-f)$ , will be appended to those variance estimators which were originally intended for use in infinite populations.

### **3.2 Jackknife in Linear Models**

The jackknife is effectively designed for balanced problems; problems which are in a sense symmetric. The jackknife estimates are simple averages. Thus, there is an implicit assumption of balance, or symmetry. However, the model-based ratio estimation problem is not symmetric, because the observed  $y_i$ 's are not assumed to be identically distributed; their means, and sometimes variances, are functions of the  $x_i$ 's. Thus, from the model-based approach, one might not expect the jackknife to perform well in estimating the variance of the ratio estimator.

In this section, we will attempt to show the inadequacy of the jackknife in the linear model setting. The linear model will be used because it has an obvious and known imbalance: the  $y_i$ 's are not identically distributed. An additional advantage of using the linear model is that, in a special case, the weighted least squares estimator of  $\hat{\beta}$  is the ratio estimator. For applications of the weighted jackknife in non-linear regression, see Fox, Hinkley and Larntz (1980) and Simonoff and Tsai (1986).

In the following discussion we will use a linear model with equal variance for the observations; that is,  $\text{Var}(Y_i) = \sigma^2$  for all  $i$ . This will be referred to as the homoscedastic case. The heteroscedastic case occurs when  $\text{Var}(Y_i) = \sigma_i^2$ . Unless explicitly mentioned otherwise, we will assume the following model:

$$Y = X\beta + e ,$$

$$\text{where } Y^t = (Y_1, \dots, Y_n) ,$$

$$X^t = (x_1, \dots, x_n) \text{ where } x_i^t = (x_{i1}, \dots, x_{ip}) \\ \text{corresponds to covariate} \\ \text{information on the } i^{\text{th}} \text{ observations} ,$$

$$\beta^t = (\beta_1, \dots, \beta_p) ,$$

$$e^t = (e_1, \dots, e_n) .$$

The rank of  $X$  is assumed to be  $p$  when any row is deleted. We will further assume

$$E(e_i) = 0, \text{Var}(e_i) = \sigma^2 \text{ and } e_i \text{ is uncorrelated with } e_j \text{ if } i \neq j.$$

Let

$$D_0 = X^t X, w_i = x_i^t D_0^{-1} x_i,$$

$$\hat{\beta} = D_0^{-1} X^t Y, \text{ the least squares estimate of } \beta, \text{ and}$$

$$(r_1, \dots, r_n)^t = Y - \hat{Y} = (I - XD_0^{-1} X^t) Y, \text{ the vector of residuals under } \hat{\beta} .$$

The following lemma will be very useful in the derivations of this chapter. It expresses the  $\hat{\beta}_{-i}$ , delete-one-pair estimate of  $\beta$ , as a function of the parameter estimate from the full data set,  $\hat{\beta}$ , minus a function involving  $x_i$  and  $Y_i$ .

**Lemma** Let  $\beta$  be the following function:  $\beta = \beta(X, Y) = (X'X)^{-1}X'Y$ . Then,

$$\begin{aligned}\beta_{-i} &= (X_{-i}'X_{-i})^{-1}X_{-i}'Y_{-i} \\ &= \beta - \frac{D_0^{-1}x_i(Y_i - x_i'\beta)}{1 - w_i} \\ &= \beta - \frac{D_0^{-1}x_i r_i}{1 - w_i},\end{aligned}$$

where  $X_{-i}$  is the matrix obtained by removing  $x_i'$  from the rows of  $X$ .

**Proof:** See Miller (1974b).

Using the above lemma, we can rewrite the pseudovalues as

$$\begin{aligned}P_i &= n\beta - (n-1)\left[\beta - \frac{D_0^{-1}x_i r_i}{1 - w_i}\right] \\ &= \beta + \frac{D_0^{-1}x_i r_i}{1 - w_i} \cdot (n-1).\end{aligned}$$

We can therefore rewrite the jackknife estimate of  $\beta$  as

$$\begin{aligned}\beta_J &= \beta + \left(\frac{n-1}{n}\right)D_0^{-1}\sum_{i=1}^n \frac{x_i r_i}{1 - w_i} \\ &= \beta + \left(\frac{n-1}{n}\right)D_0^{-1}X_*'(Y - \hat{Y}), \text{ where } X_*' = \left(\frac{x_1}{1 - w_1}, \dots, \frac{x_n}{1 - w_n}\right).\end{aligned}$$

For the homoscedastic case it is well known that the least squares estimator,  $\beta$ , is unbiased.

Thus, the bias-reducing property of the jackknife is redundant. Furthermore, since  $Var(\beta)$  is the minimum among all linear unbiased estimators (by the Gauss-Markov theorem), it is evident that  $Var(\beta_J)$  must be greater than or equal to  $Var(\beta)$ . In fact, the above representation of the

pseudovalue and the fact that the least squares estimate  $\beta$  is uncorrelated with the residuals, leads immediately to

$$\begin{aligned}
Var(\beta_J) &= Var(\beta) + \left(\frac{n-1}{n}\right)^2 D_0^{-1} X_*' Var(Y - \hat{Y}) X_* D_0^{-1} \\
&= \sigma^2 D_0^{-1} + \left(\frac{n-1}{n}\right)^2 D_0^{-1} X_*' \sigma^2 (I - X D_0^{-1} X') X_* D_0^{-1} \\
&= \sigma^2 D_0^{-1} \left[ D_0 + \left(\frac{n-1}{n}\right)^2 (X_*' X_* - X_*' X D_0^{-1} X' X_*) \right] D_0^{-1} \\
&= \sigma^2 D_0^{-1} \left[ D_0 + \left(\frac{n-1}{n}\right)^2 (D_2 - D_1 D_0^{-1} D_1) \right] D_0^{-1}
\end{aligned}$$

(3.1)

where  $D_k = \sum_{i=1}^n (1 - w_i)^{-k} x_i x_i'$ .

If we assume that  $w_i$  is of order  $n^{-1}$ , it can be shown that  $\{Var(\beta_J) - Var(\beta)\}$  is of order  $n^{-2}$  (Hinkley, 1977).

The jackknife estimate of the variance of the jackknife estimate  $\beta_J$  is the appropriately scaled sample variance of the pseudovalue; namely,

$$var_J(\beta_J) = \{n(n-1)\}^{-1} \sum_{i=1}^n (P_i - \beta_J)(P_i - \beta_J)'$$

Miller (1974b) and Hinkley (1978) argue that the jackknife variance estimate can be used to estimate the variance of the least squares estimator  $\beta$ . We will denote  $var_J(\beta_J)$  by  $var_J$ . Using simple algebra, it can easily be shown that the expectation of this jackknife variance estimate is (Hinkley, 1977)

$$E(var_J) = \sigma^2 \left( \frac{n-1}{n} \right) D_0^{-1} \left[ D_1 - \frac{1}{n} (D_2 - D_1 D_0^{-1} D_1) \right] D_0^{-1} .$$

Comparing this expectation to the true variance of the jackknife estimate  $\hat{\beta}_J$ , it can be shown that the jackknife variance estimator  $var_J$  is biased for both  $Var(\hat{\beta}) = \sigma^2 D_0^{-1}$ , and for  $Var(\hat{\beta}_J)$ ; see (3.1). This bias occurs because the jackknife variance estimator is designed for balanced problems, while the linear model is unbalanced. For general results concerning the bias of the jackknife estimator, see Efron and Stein (1981).

Because the jackknife variance estimator is biased and because it may not have the optimal variance, it is not advisable to use it in linear models to estimate the variance of the parameter estimate,  $\hat{\beta}$ . Weighted jackknife variance estimators attempt to account for this imbalance; they will be discussed in the next section.

### 3.3 Weighted Jackknife in Linear Models

In this section we will introduce two weighted versions of the jackknife estimators of  $Var(\hat{\beta})$ .

These estimators were first suggested by Hinkley (1977) and by Wu (1986); therein may be found proofs of results and conclusions mentioned in this section.

#### 3.3.1 Approach Using Influence Functions

The influence function of a functional  $\theta$  at the point  $x_i$  indicates the extent to which the value of  $\theta$  would change when an infinitesimal fraction of the population is moved to  $x_i$ . Alternately, the influence function gives "the suitably scaled differential influence of one additional observation at  $x$ , as the sample size approaches infinity" (paraphrased from Huber, 1977). Several authors have

pointed out the relation between pseudovalues and the influence function, in particular Jaeckel (1972) and Mallows (1975).

The influence function of a functional  $T$  at a distribution  $F$  and a point  $x$  is defined (Hampel, 1968, 1974) as:

$$IF_F(x;T) = \lim_{\epsilon \rightarrow 0} \frac{T[(1-\epsilon)F + \epsilon\delta_x] - T[F]}{\epsilon} ,$$

where  $\delta_x$  is a point mass at  $x$  .

Hinkley (1977) argues that the linear model is generally unbalanced; the imbalance being evident in the different squared "distances" of  $x_i$  from the origin, scaled by  $D_0^{-1}$ :  $x_i' D_0^{-1} x_i$ . Because the linear model is unbalanced, Hinkley argues the pseudovalues should be modified; for example, by weighting. He goes on to argue from an intuitive point of view what the weights should be. However, Hinkley's strongest argument is the suggestion that Quenouille (1956) essentially defined a pseudovalue as the estimate of a parameter plus the estimated influence function at the observation under consideration; that is,

$$P_i = \hat{\theta} + \hat{IF}(x_i; \hat{\theta}) .$$

Thus, Hinkley suggests that for unbalanced problems, particularly linear models, one should use as the  $i^{th}$  pseudovalue the parameter estimate plus the estimated influence function at the  $i^{th}$  observation. We shall denote this modified pseudovalue by  $Q_i$ ,

$$Q_i = \hat{\theta} + \hat{IF}(x; \hat{\theta}) .$$

The estimated influence function,  $\hat{IF}(x_i; \hat{\theta})$ , is simply an estimate of the influence function  $IF(x_i; \theta)$ . Mallows (1975) discusses six consistent estimators of the influence function and their usefulness in the theory of robustification. Jaeckel (1972) proposes a modification of the jackknife, known as the infinitesimal jackknife, based on one particular approach toward estimation of the influence function. (Also see Efron 1979, 1982).

Hinkley has shown that for linear models, the influence function of  $\beta$  is

$$IF_F(x_i, Y_i; \beta) = \Sigma^{-1} x_i (Y_i - x_i^t \beta) ,$$

where  $F$  is the joint distribution of  $x_i$  and  $Y_i$ ,

$$\Sigma = E_F(x_i x_i^t) , \text{ and}$$

$$\beta = \Sigma^{-1} E_F(x_i Y_i) .$$

One estimate of the influence function of  $\beta$  is (Hinkley, 1977)

$$\hat{IF}_F(x_i, y_i; \beta) = n D_0^{-1} x_i r_i ,$$

corresponding to which the pseudo-value is

$$\begin{aligned} Q_i &= \beta + \hat{IF}(x_i, y_i; \beta) \\ &= \beta + n D_0^{-1} x_i r_i , \end{aligned} \tag{3.2}$$

and by the Lemma in Section 2.2 ,

$$= \beta + n(1 - w_i)(\beta - \hat{\beta}_{-i}) .$$

Hinkley goes on to show that the weighted jackknife estimate is simply the least squares estimate,

$$\hat{\beta}_H = \frac{1}{n} \sum_{i=1}^n Q_i = \beta , \tag{3.3}$$



since the sum of  $x_i r_i$  is zero in ordinary least squares. The  $Q_i$  are referred to as weighted pseudovalues.

Similarly, Hinkley suggests the use of these weighted pseudovalues to estimate the variance of  $\beta$ .

$$var_H(\beta) = \{n(n-p)\}^{-1} \sum_{i=1}^n (Q_i - \bar{Q})(Q_i - \bar{Q})' , \quad (3.4)$$

which for linear models is (by (3.2) and (3.3)),

$$= \left( \frac{1}{1 - \frac{p}{n}} \right) \sum_{i=1}^n (1 - w_i)^2 (\beta_{-i} - \beta)(\beta_{-i} - \beta)' .$$

The denominator Hinkley proposes differs from the usual denominator  $\{n(n-1)\}$ . However, he suggests that  $\{n(n-p)\}^{-1}$  is preferable because it appropriately reflects the degrees of freedom, and makes  $var_H(\beta)$  unbiased in the case  $w_i = p n^{-1}$ , usually referred to as the balanced case.

The expectation of  $var_H(\beta)$  under the distribution function of the errors is (Miller, 1974b; Hinkley, 1977)

$$E(var_H(\beta)) = \left( \frac{n}{n-p} \right) D_0^{-1} (D_0 - D_1^*) D_0^{-1} \cdot \sigma^2 ,$$

$$\text{where } D_1^* = \sum_{i=1}^n w_i^h x_i x_i' .$$

In the balanced case,  $D_1^* = \frac{p}{n} D_0$  and therefore  $var_H(\beta)$  is unbiased for  $Var(\beta)$ . More generally, in

linear models with homoscedastic errors, Wu (1986) shows that  $var_H(\beta)$  has bias of order  $n^{-1}$

irrespective of balance or lack thereof. Because Hinkley's proposed variance estimator  $var_H(\beta)$  is

unbiased in the homoscedastic case, unlike the Quenouille-Tukey variance estimator  $var_J(\beta)$ , one would presumably prefer Hinkley's variance estimator over the Quenouille-Tukey estimator.

### 3.3.2 Approach Using the Fisher Information Matrix

We will now proceed to the jackknife variance estimator proposed by Wu (1986). Whereas Hinkley's version of the jackknife variance estimator is motivated by a particular definition of the pseudo-value, Wu does not propose a new version of the pseudo-values since, he says, "It's extension to non-iid situations lacks firm theoretical foundation." His version of the jackknife variance estimator was motivated by the representation of  $\beta$  as a weighted average of the  $\beta_{-i}$ 's, the LSE's based on the "delete-one" samples. For the general p-dimensional case, Wu (1986) shows that

$$\beta = \sum_{i=1}^n u_i \beta_{-i}$$

$$\text{where } u_i = \frac{|X_{-i}^t X_{-i}|}{\sum_{j=1}^n |X_{-j}^t X_{-j}|} . \quad (|\cdot| \text{ denotes the determinant})$$

An alternate way of viewing this representation comes from the recognition that in this representation the weights are the determinants of the Fisher Information matrices of the particular data subsets under consideration. Of course, the Fisher Information matrix depends upon distributional and model assumptions; the above weights are obtained under a homoscedastic linear model, with normally distributed errors, but this idea can be generalized under other distributional and model assumptions. Interpreting the weights  $u_i$  as the relative information content of the sample with the  $i^{\text{th}}$  observation removed, led Wu (1986) to propose the following weighted jackknife estimate of the variance of  $\beta$ :

$$var_w(\beta) = (n-1) \sum_{i=1}^n u_i (\beta_{-i} - \beta)(\beta_{-i} - \beta)' . \quad (3.5)$$

In fact, Wu's proposal is more general than the above method which we have described. He considers the more general version of the jackknife which removes an arbitrary number of points, not just one as in the "delete-one" version. However, we will only deal with the delete-one version of Wu's variance estimator. We do this for three reasons: (i) in order to compare  $var_w(\beta)$  with  $var_H(\beta)$ . Otherwise, the weights are difficult to compare because both  $u_i$  and  $(\beta_{-i} - \beta)(\beta_{-i} - \beta)'$  change; (ii) because the robustness of the delete-one version of  $var_w(\beta)$  can be established (Shao and Wu, 1987); and, (iii) because it has been established that the delete-one version of the jackknife is optimal for estimation of the variance of the classical ratio estimator (Chakrabarty and Rao, 1968). Wu (1986) shows that  $var_w$  is unbiased in the homoscedastic linear model, unlike  $var_H$  which was biased.

To summarize, the weighted versions of the jackknife variance estimators proposed by Hinkley (1977) and Wu (1986) eliminate or reduce the bias of the unweighted versions. Though our attention has been restricted to linear models, we may still conclude that by taking into account model imbalance, the standard jackknife can be improved in some situations. Though both these approaches have been presented only for linear models, the same approaches may be useful in more general unbalanced situations.

### 3.4 Jackknife Variance Estimates in a Heteroscedastic Linear Model

In this section, we will explicitly derive weighted versions of the jackknife variance estimator for the linear model for regression through the origin with variance structure specified by

$Var(e_i) = \sigma^2 x_i^\lambda$ . We have chosen this version of the linear model because in many practical applications of the ratio estimator, this form of variance does indeed occur; and furthermore, for the particular value of  $\lambda = 1$ , the weighted least squares estimate of  $\beta$  is the classical ratio estimator. From this section onwards, we will set  $p=1$ ; that is,  $x_i$  and  $\beta$  will be one-dimensional and not vectors.

The linear model with heteroscedasticity of the form  $Var(e_i) = \sigma^2 x_i^\lambda$  can be converted to a homoscedastic case. Thus, the model

$$Y_i = x_i \beta + e_i ,$$

*Model 1:*  $E(e_i) = 0$  ,  $Var(e_i) = \sigma^2 x_i^\lambda$  and  
 $e_i$  independent of  $e_j$  for  $i \neq j$  ,

may be transformed to the model

$$Y_i^* = x_i^* \beta + e_i^* ,$$

*Model 2:*  $E(e_i^*) = 0$  ,  $Var(e_i^*) = \sigma^2$  and  
 $e_i^*$  independent of  $e_j^*$  for  $i \neq j$  ,

where the transformation  $*$  simply denotes division by  $x_i^{\frac{\lambda}{2}}$ . When  $\lambda = 1$ , the least squares

estimator  $\hat{\beta}$  of  $\beta$  in model 2, which corresponds to MLE in model 1 had we assumed normal

errors, is easily seen to be the ratio estimate,  $\hat{R} = \frac{\bar{y}}{\bar{x}}$ . This suggests one way of obtaining

unbalanced jackknife estimates of the variance of  $\hat{R}$ : Using model 2 and  $\lambda = 1$ , Hinkley's and Wu's

weighted jackknife variance estimators will be obtained. These will then correspond to model-based estimators of the variance of  $\hat{R}$ .

### 3.4.1 Influence Function-Based Jackknife

From Section 3.3.1, Hinkley's estimate of the variance of  $\hat{\beta}$  is (3.4):

$$var_H = \frac{1}{n(n-1)} \sum_{i=1}^n (Q_i - Q.)^2$$

and since the linear model weighted pseudovalues for model 2 are given by  $Q_i = \hat{\beta} + n D_0^{-1} x_i^* r_i^*$ , we have

$$var_H = \frac{n}{n-1} \sum_{i=1}^n D_0^{-1} [x_i^{*2} r_i^{*2}] D_0^{-1} .$$

But for our transformed problem

$$D_0^{-1} = (X^{*t} X^*)^{-1} = \left( \sum_{i=1}^n x_i^{2-\lambda} \right)^{-1} .$$

Therefore,

$$\begin{aligned} var_H &= \frac{n}{n-1} \sum_{i=1}^n (y_i^* - x_i^* \hat{\beta})^2 x_i^{*2} \cdot \left( \sum_{j=1}^n x_j^{2-\lambda} \right)^{-2} , \\ &= \frac{n}{n-1} \sum_{i=1}^n [(y_i - x_i \hat{\beta})^2 x_i^{2-2\lambda}] \cdot \left( \sum_{j=1}^n x_j^{2-\lambda} \right)^{-2} ; \end{aligned}$$

For the case of  $\lambda = 1$ , this expression becomes

$$var_H(\hat{R}) = \frac{n}{n-1} \sum_{i=1}^n (y_i - x_i \hat{R})^2 \cdot \frac{1}{(n\bar{x})^2} .$$

Therefore,

$$\begin{aligned}
 var_H(\hat{Y}) &= var_H(N\bar{X}\hat{R}) , \\
 &= \frac{N^2}{n(n-1)} \cdot \sum_{i=1}^n (y_i - x_i \hat{R})^2 \cdot \left( \frac{\bar{X}}{\bar{x}} \right)^2 , \\
 &= var_2(\hat{Y}) / (1-f) .
 \end{aligned}$$

Thus when the finite population correction factor is included, the variance expression resulting from Hinkley's version of the unbalanced jackknife is the same as Cochran's second estimate of the variance,  $var_2(\hat{Y})$ .

If we re-express  $var_H$  as the sum of squared delete-one estimates of the parameter, as in (3.4), we obtain

$$\begin{aligned}
 var_H(\beta) &= \left( \frac{n}{n-1} \right) \sum_{i=1}^n (1 - x_i^{*t} (X^{*t} X^*)^{-1} x_i^*)^2 (\beta_{-i} - \beta)^2 , \\
 &= \left( \frac{n}{n-1} \right) \sum_{i=1}^n \left( 1 - \frac{x_i^{2-\lambda}}{\sum_{j=1}^n x_j^{2-\lambda}} \right)^2 (\beta_{-i} - \beta)^2 .
 \end{aligned}$$

For the case  $\lambda = 1$ , this becomes

$$\begin{aligned}
 var_H(\hat{R}) &= \left( \frac{n}{n-1} \right) \sum_{i=1}^n \left( \frac{(n-1)\bar{x}_{-i}}{n\bar{x}} \right)^2 (\hat{R}_{-i} - \hat{R})^2 , \\
 &= \left( \frac{n-1}{n} \right) \sum_{i=1}^n \left( \frac{\bar{x}_{-i}}{\bar{x}} \right)^2 (\hat{R}_{-i} - \hat{R})^2 .
 \end{aligned}$$

In Hinkley's version of the weighted jackknife, the weights are  $\left( \frac{\bar{x}_{-i}}{\bar{x}} \right)^2$ , where  $\bar{x}_{-i}$  is the mean of the

sub-sample obtained by deleting  $i^{th}$  observation on  $x$ 's. Thus the weights are a function of the sub-sample balance on the first moment. If the sub-sample mean is less than the sample mean, then the weights are less than one, implying that  $(R_{-i} - R)^2$  gets down-weighted relative to the same terms corresponding to a sub-sample with high first moment. At first glance, it is surprising that subsamples with a high first order moment receive more weight. Of course under the super-population model, the larger the  $x_i$ 's, the smaller is the variability of  $\hat{R}$ , and perhaps this explains why subsamples with high  $x_i$ 's are weighted more.

### 3.4.2 Information Matrix-Based Jackknife

We will now consider Wu's version of the weighted jackknife. From (3.5) in Section 3.3.2, Wu's weighted jackknife estimator of the variance of  $\hat{\beta}$  (for  $p=1$ ) is defined as

$$var_w(\hat{\beta}) = (n-1) \sum_1^n u_i (\hat{\beta}_{-i} - \hat{\beta})^2,$$

and the weights for model 2 are:

$$u_i = \frac{|X_{-i}^{*t} X_{-i}^*|}{\sum_{j=1}^n |X_{-j}^{*t} X_{-j}^*|} = \frac{\sum_{j \neq i} x_j^{2-\lambda}}{\sum_{j=1}^n \sum_{l \neq j} x_l^{2-\lambda}}.$$

When  $\lambda = 1$ , this becomes  $u_i = \frac{(n-1)\bar{x}_{-i}}{n(n-1)\bar{x}}$  which leads to

$$var_w(\beta) = \left( \frac{n-1}{n} \right) \sum_{i=1}^n \left( \frac{\bar{x}_{-i}}{\bar{x}} \right) (\hat{R}_{-i} - \hat{R})^2 .$$

Thus, it is seen that whereas the weights are  $\left( \frac{\bar{x}_{-i}}{\bar{x}} \right)^2$  in Hinkley's version, they are  $\left( \frac{\bar{x}_{-i}}{\bar{x}} \right)$  in Wu's

version. In both instances, the weights are increasing functions of the "information content" and hence are appealing and reasonable. Even when the Fisher Information of a "leave-one-out" sample is not  $\bar{x}_{-i}$ , these weights appear to be reasonable because many studies have indicated the importance of the first moment balance (Royall and Cumberland, 1981; Wu and Deng, 1983; Royall and Cumberland, 1985; Robinson, 1987).

Wu's weighted variance estimator can also be represented (using the Lemma in Chapter 2) as

$$var_w(\hat{R}) = \frac{1}{n\bar{x}^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - x_i\beta)^2 \left( \frac{\bar{x}}{\bar{x}_{-i}} \right) . \quad (3.6)$$

In the similar representation of  $var_H(\hat{R})$  in Section 3.4.1, it was found that  $var_H(\hat{R})$  was equal to  $var_2(\hat{R})$ . For Wu's variance estimator, we see that  $var_w(\hat{R})$  is not equivalent to any variance estimator which we have considered.

In this heteroscedastic linear model, we have obtained two weighted jackknife variance estimators. Wu (1986) and Shao and Wu (1987) derive asymptotic results for the heteroscedastic-robustness of the weighted jackknife methods presented here. Their results deal with the more general heteroscedastic models, where the variances of the errors are not a known function of the  $x_i$ 's. They show that both  $var_w(\beta)$  and  $var_H(\beta)$  have order  $n^{-1}$  bias in the general heteroscedastic case.



They further show that  $var_w(\hat{\beta})$  is unbiased in the specific heteroscedastic case where  $x_i' D_0^{-1} x_j$  is equal to zero for any  $i, j$  with  $\sigma_i^2$  not equal to  $\sigma_j^2$ .

In this section, weighted versions of the jackknife variance estimator were derived for a heteroscedastic linear model. These estimators were obtained from a model-based approach to ratio estimation. In the following section, a design-based approach will be taken in deriving weighted jackknife variance estimators.

### 3.5 Weighted Jackknife Variance Estimates in Ratio Estimation

In this section, we will explicitly derive weighted versions of the jackknife variance estimator of the classical ratio estimator. Following the approach of Hinkley (1977) and Wu (1986), we will show how Hinkley's version of the weighted jackknife variance estimator leads to the two variance estimators  $var_0$  and  $var_2$ . Thereafter we will show that Wu's approach leads to the variance estimator which was discussed in Section 3.4.2.

#### 3.5.1 Influence Function-Based Approach

Recall that Hinkley's approach in deriving weighted pseudovalues employed estimated influence functions. Therefore, we will first derive the estimated influence function of  $R$ , and thereafter derive the corresponding pseudovalues and estimators.

Let us define the functional  $R$  of the joint distribution  $G$  of  $X$  and  $Y$  as follows:

$$R(G) = \frac{\int y dF_y}{\int x dF_x} = \frac{\mu_y}{\mu_x},$$

where  $F_y$  and  $F_x$  are the respective marginal densities of  $Y$  and  $X$ .

If  $\delta_z$  represents a point mass at  $(X_z, Y_z)$  then the influence function of  $R$  is defined as follows:

$$IF(X_z, Y_z; R) = \lim_{\varepsilon \rightarrow 0} \left\{ \frac{R[(1-\varepsilon)G + \varepsilon\delta_z] - R(G)}{\varepsilon} \right\}.$$

Straightforward calculation yields

$$\begin{aligned} IF(X_z, Y_z; R) &= \lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \left\{ \frac{(1-\varepsilon)\mu_y + \varepsilon Y_z}{(1-\varepsilon)\mu_x + \varepsilon X_z} - \frac{\mu_y}{\mu_x} \right\} \\ &= \lim_{\varepsilon \rightarrow 0} \left\{ \frac{Y_z\mu_x - \mu_y X_z}{(1-\varepsilon)\mu_x^2 + \varepsilon X_z\mu_x} \right\} \\ &= \frac{1}{\mu_x} \left( Y_z - \frac{\mu_y}{\mu_x} X_z \right) \\ &= \frac{1}{\mu_x} (Y_z - RX_z). \end{aligned}$$

There are two sample analogues of this influence function, depending on whether we use the sample estimate of  $\mu_x$  or the actual value:

$$\begin{aligned} \hat{IF}_1(X_z, Y_z; \hat{R}) &= \frac{1}{\bar{x}} (Y_z - \hat{R}X_z) \quad \text{or} \\ \hat{IF}_2(X_z, Y_z; \hat{R}) &= \frac{1}{\bar{X}} (Y_z - \hat{R}X_z). \end{aligned}$$

The modified pseudovalues based on  $\hat{IF}_1$  are

$$\begin{aligned} \bar{Q}_i &= \hat{R} + \hat{IF}_1(x_i, y_i; \hat{R}) \\ &= \hat{R} + \frac{1}{\bar{x}} (y_i - \hat{R}x_i), \quad i=1, \dots, n, \end{aligned}$$

with the corresponding estimator

$$\bar{Q} = \frac{1}{n} \sum_{i=1}^n \bar{Q}_i = \hat{R} + \frac{1}{\bar{x}} \cdot \frac{1}{n} \sum_{i=1}^n (y_i - \hat{R}x_i) = \hat{R},$$

the usual estimator. The corresponding weighted variance estimator is

$$\begin{aligned}
 var_H(\hat{R}) &= \frac{1}{n(n-1)} \sum_{i=1}^n (\tilde{Q}_i - \tilde{Q})^2 , \\
 &= \frac{1}{n(n-1)} \sum_{i=1}^n \left( \frac{1}{\bar{x}} (y_i - \hat{R}x_i) \right)^2 , \\
 &= \frac{1}{n\bar{x}^2} \left( \frac{1}{n-1} \right) \sum_{i=1}^n (y_i - \hat{R}x_i)^2 , \\
 &= var_2(\hat{R}) / (1-f) .
 \end{aligned}$$

Thus, when the finite population correction factor is incorporated, the weighted estimator of the variance corresponding to  $\hat{IF}_1$  is  $var_2$ , the second estimator suggested in Cochran (1977).

Similarly, the use of  $\hat{IF}_2$  results in modified pseudovalues which lead to the variance estimator  $var_0$ . Thus, explicit derivation of Hinkley's version of the weighted jackknife results in two known estimators, in contrast to the results obtained in Section 3.4.1 where Hinkley's version results in just one known estimator,  $var_2$ . We will now proceed on to Wu's version of the weighted jackknife.

### 3.5.2 Information Matrix-Based Approach

Recall that Wu's approach can be motivated by the use of weights proportional to the determinant of the Fisher Information Matrix of subsamples obtained by systematically removing individual observations. The use of these weights results in the weighted jackknife estimate being equal to the original estimate.

If the pairs  $(x_i, Y_i)$  satisfy the following model,

$$Y_i = Rx_i + e_i, \text{ where the } e_i \text{ are independent with distribution } N(0, \sigma^2 x_i) ,$$

then the Fisher Information number of  $R$  given an observation  $(x_i, Y_i)$  is  $(x_i/\sigma^2)$ . Thus, the weights  $u_i$  in (3.5) are  $(\bar{x}_{-i}/n\bar{x})$ ,  $i=1,\dots,n$ . It is interesting to note that these weights satisfy the following:

$$\sum_{i=1}^n u_i \hat{R}_{-i} = \sum_{i=1}^n \frac{\bar{x}_{-i}}{n\bar{x}} \cdot \frac{\bar{y}_{-i}}{\bar{x}_{-i}} = \hat{R}.$$

From (3.5), Wu's jackknife variance estimator becomes

$$\begin{aligned} var_w^*(\hat{R}) &= (n-1) \sum_{i=1}^n u_i (\hat{R}_{-i} - \hat{R})^2 \\ &= \left( \frac{n-1}{n} \right) \sum_{i=1}^n \left( \frac{\bar{x}_{-i}}{\bar{x}} \right) (\hat{R}_{-i} - \hat{R})^2, \end{aligned}$$

which is the same as  $var_w(\hat{R})$  derived in Section 3.4.2.

### 3.6 Summary

In this chapter, we concentrated on deriving jackknife based estimators of the variance. Using linear models at first, we attempted to show the shortcomings of the jackknife in unbalanced problems. We then derived weighted jackknife variance estimators and showed they are preferable to the jackknife variance estimator because they have lower bias. Finally, in the preceding section, we derived weighted jackknife variance estimators for the classical ratio estimator. Of the three newly-derived estimators, two were shown to be equivalent to already well-known variance estimators, while the third is a new estimator.

## **Chapter 4**

### **Bootstrapping in Ratio Estimation**

We will now proceed to a sophisticated resampling technique known as the bootstrap. In Section 1, we introduce bootstrap estimates of standard error in a general statistical setting. In Section 2, we introduce and discuss a bootstrap estimate of the standard error of the ratio estimator. This bootstrap may be described as arising from a design-based view. In Section 3, using the approach of a bootstrap for linear models, we derive a second bootstrap estimate of standard error. In Section 4, we introduce and discuss a weighted bootstrap estimate of standard error. Finally in Section 5, we discuss some of the potential drawbacks and weaknesses of the three proposed estimators.

#### **4.1 Introduction**

The bootstrap was first discussed in Efron (1979). The theory was further expanded and refined in Bickel and Freedman (1981), Rubin (1981), Singh (1981) and Efron (1982). An excellent recent summary of the bootstrap is provided in Hinkley (1988) and DiCiccio and Romano (1988). The variance estimators which we will derive are based upon approaches first suggested in Efron (1979), Efron and Gong (1983) and Wu (1986) along with the accompanying discussion.

The bootstrap is a resampling technique which is primarily used to estimate the statistical error of estimates, providing non-parametric estimates of the bias, standard deviation and confidence intervals. In most problems, the bootstrap is simple to apply, requiring very little theoretical work before its application. It does, however, require a large amount of computing power and is not applicable in all problems. Schenker (1985) provides examples of situations where the bootstrap does not work.

The basic idea behind the bootstrap is as follows. Consider  $n$  independent and identically distributed observations  $x_1, \dots, x_n$  from a distribution  $F$ . Further consider that a statistic, which depends on  $F$ , is calculated from these observations. Then to estimate the sampling distribution of the statistic, substitute the unknown underlying distribution  $F$  with its empirical estimate  $\hat{F}_n$  which simply places mass  $\frac{1}{n}$  at each of  $n$  observations. Recall that the jackknife estimates

standard error by scaling the sample standard deviation of the jackknife estimates which are obtained by systematically "leaving-out" observations. Similarly, the bootstrap estimate of standard error is the sample standard deviation of the bootstrap estimates of the parameter. However, the bootstrap parameter estimates are obtained by via simple random sampling with replacement from the original, observed data. Thus, we are replacing the true distribution  $F$  by the observed empirical distribution  $\hat{F}_n$ . In the delete-one version of the jackknife, a total of  $n$  estimates of the parameter are obtained, whereas in the bootstrap,  $n^n$  estimates of the parameter are obtained.

More formally, the bootstrap estimate of standard error for the context of simple random sampling is as follows: Let  $\theta = \theta(F)$  be the parameter of interest and let  $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$  be its corresponding estimator. We wish to estimate the standard error of  $\hat{\theta}$ . Let the true standard error of  $\hat{\theta}$  be  $\sigma = \sigma(n, \hat{\theta}, F)$ . The bootstrap estimate of  $\sigma$  is  $\hat{\sigma}_B = \sigma(n, \hat{\theta}, \hat{F}_n)$ . Since we know  $\hat{F}_n$ , and we know how to calculate  $\hat{\theta}$  for any given sample of size  $n$  drawn from  $\hat{F}_n$ , we could, if we wished, proceed by calculating  $\hat{\theta}$  from each of the  $n^n$  possible bootstrap samples (i.e. the possible "re-samples"). This is the essence of the bootstrap. The estimate of the standard error of  $\hat{\theta}$  is calculated as the

standard deviation of the  $n^*$  estimates of  $\theta$ ; that is, of the estimates obtained from the bootstrap samples. However,  $n^*$  grows extremely quickly and the number of bootstrap estimates becomes unmanageable for even relatively small values of  $n$ . The solution is to approximate the standard deviation of the  $n^*$   $\hat{\theta}$ 's. Thus, a large number of simple random samples of size  $n$  are drawn with replacement from  $\hat{F}_n$ . This is repeated until a sufficiently accurate estimate  $\hat{\sigma}_B$  is obtained.

To illustrate the above, suppose we wish to estimate the variability of a sample correlation coefficient. We have observed  $n$  pairs of  $(s_i, t_i) = u_i$ ,  $i=1, \dots, n$ . We draw a simple random sample of size  $n$ , with replacement from  $\{u_1, u_2, \dots, u_n\}$  and calculate the correlation coefficient in this "bootstrap sample". We repeat this a large number of times, say  $B=1,000$ . The sample standard deviation of these 1,000 correlation coefficients is the bootstrap estimate of the standard error of the correlation coefficient from the original data.

In some instances, it is possible to calculate the standard error of the  $n^*$  possible bootstrap estimates of  $\theta$  exactly via direct, theoretical calculations. For example, let  $X_1, \dots, X_n$  be  $n$  independent and identically distributed observations, with distribution function  $F$ . Let

$\hat{\theta}(X_1, \dots, X_n) = \bar{X}$ . Then the true standard error of  $\bar{X}$  is  $\frac{\sigma}{\sqrt{n}}$ , where  $\sigma = \sqrt{E_F(X^2) - E_F(X)^2}$  is the

standard deviation of the underlying distribution  $F$ . The bootstrap estimate of the standard error

of  $\bar{X}$  is  $\frac{\hat{\sigma}_B}{\sqrt{n}}$  where  $\hat{\sigma}_B = \sqrt{E_{\hat{F}_n}(X^2) - E_{\hat{F}_n}(X)^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \left( \frac{1}{n} \sum_{i=1}^n x_i \right)^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$  is the

standard deviation of the empirical distribution  $\hat{F}_n$ . Thus  $E_{\hat{F}_n}(X^2)$  and  $E_{\hat{F}_n}(X)$  were found theoretically, without need for Monte Carlo. Note that,  $\hat{\sigma}_B$  corresponds to the usual estimate of  $\sigma$  (except for the term replacement of  $n - 1$  by  $n$ ).

The bootstrap procedure employing Monte Carlo simulation is as follows:

- (i) Construct the empirical distribution function,  $\hat{F}_n$ , by placing mass  $\frac{1}{n}$  at each observation.
- (ii) a. Draw a simple random sample of size  $n$ , with replacement, from  $\hat{F}_n$ , and compute  $\hat{\theta}$  from this bootstrap sample:  $\hat{\theta}^*$ .  
 b. Repeat (a)  $B$  times. Denote the  $B$  bootstrap realizations of  $\hat{\theta}$  by  $\hat{\theta}_i^*$ ,  $i=1,\dots,B$ .
- (iii) Approximate the exact value of  $\hat{\sigma}_B = \sigma(n, \hat{\theta}, \hat{F}_n)$  by  $\left[ \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_b^* - \hat{\theta}^*)^2 \right]^{1/2}$ .

Had we used the direct theoretical approach, we would skip step (ii) and instead evaluate  $\hat{\sigma}_B = \sigma(n, \hat{\theta}, \hat{F}_n)$  theoretically. Step (ii) replaces theoretical derivations by sheer computing power.

The method of Monte Carlo simulation is by far the easiest and most commonly used approach. If sufficient computing power were available, one could calculate  $\sigma(n, \hat{\theta}, \hat{F}_n)$  exactly by using all  $n^n$  possible bootstrap samples. However, in practice one approximates  $\sigma(n, \hat{\theta}, \hat{F}_n)$  by randomly selecting  $B$  samples and calculating the standard error of these. The value of  $B$  is often set between 100 and 1,000. One way to obtain more information for a given level of computing power is to select samples which are balanced in the sense that the samples selected cover the hyper-



dimensional lattice cube  $\{1, 2, \dots, n\}^n$  as evenly as possible (Graham, et al, 1990). The balanced design approach probably produces "a four-fold or five-fold reduction in B for any given level of simulation error" (Hinkley, 1988). Other general numerical techniques are discussed in Therneau (1983), Hesterberg (1988) and Efron (1990). (For a more theoretical approach, using saddlepoint approximation methods, see Davison and Hinkley (1988)).

The advantages and appeal of the bootstrap are immediately obvious: No distributional assumptions have been made, and the method may be applied (depending on the available computing power) to all statistics, regardless of their complexity. This leads to a pitfall of the bootstrap. Because it is so easy to apply, many users assume that the bootstrap works in all situations. This is obviously wrong. The bootstrap depends on certain assumptions; a fact many users forget. The primary assumption of the bootstrap (as described above) is that the observations are exchangeable. That is, all the  $X_i$  have the same distribution but need not be independent. We will discuss this in later sections.

We conjecture that the bootstrap may provide a statistically efficient and accurate estimate of variance of the classical ratio estimator in the context of forest weigh scaling. However, there are several ways of applying the bootstrap idea to our problem and we discuss these in the following sections.

## **4.2 Bootstrap Variance Estimates - A Design-Based Approach**

Our population consists of  $N$  pairs of  $(x_i, y_i)$  of which we have observed a random sample of  $n$  pairs. Thus, the pairs are exchangeable. This immediately suggests one method of applying the bootstrap to estimate  $Var(\hat{R})$ . We shall call this method Bootstrap I:

- (i) Randomly select, with replacement,  $n$  pairs of  $(x, y)$  from the original sample and calculate the ratio estimate for this bootstrap sample.
- (ii) Repeat this  $B$  times and denote the  $B$  resulting ratio estimates by  $\hat{R}_b^*$ ,  $b=1,2,\dots,B$ .
- (iii) Calculate the sample variance of  $\hat{R}_b^*$ ,  $b=1,2,\dots,B$  and let this be the estimate of  $Var(\hat{R})$ :

$$var_{BI}(\hat{R}) = \frac{1}{B-1} \sum_{b=1}^B (\hat{R}_b^* - \hat{R}^*)^2, \text{ where } \hat{R}^* = \frac{1}{B} \sum_{b=1}^B \hat{R}_b^*.$$

This version of the bootstrap is based upon two assumptions: one,  $\hat{F}_n$  adequately approximates  $F$  and two, the pairs  $(x_i, y_i)$  are exchangeable. The first assumption is the basis of the bootstrap. It is an appropriate assumption, provided that the sample pairs were selected randomly (though various schemes of selective sampling would also suffice) and enough observations were made. The second assumption is also valid, because each pair of the population has an equal chance of selection. This version of the bootstrap may particularly appeal to proponents of the design-based school, because no model assumptions such as model (2.1) were made.

Exact theoretical calculation of the bootstrap estimate of standard error is difficult since both covariates,  $x$  and  $y$ , are random. Hence, in the empirical studies of Chapter 5, we will evaluate Bootstrap I via Monte Carlo simulation.

### 4.3 Bootstrap Variance Estimates - A Model-Based Approach

In Chapter 3, we used the linear model approach to derive jackknife variance estimators. The justification for this approach to deriving estimators was that the data upon which ratio estimation was being applied often satisfied the model assumptions. Another reason for deriving both model-based and design based variance estimators is because estimators derived from the two approaches perform differently in many common problems.

The bootstrap has been applied to linear models and has been shown to perform reasonably well in empirical tests (Freedman and Peters, 1984; Carroll, Ruppert and Wu, 1986; and Wu, 1986).

Under the homoscedastic linear model, the exchangeable components are the errors. Consequently, in the linear model, bootstrap methods resample the residuals and not the observations  $(x_i, y_i)$ , though the latter can also be done.

The linear model bootstrap algorithm for the general case of  $p$  dimensions is as follows:

- (i) Estimate  $\beta$  using the  $n$  observations and evaluate the residuals:

$$\{(y_1 - \beta x_1), \dots, (y_n - \beta x_n)\} = \{\hat{e}_i, i=1, \dots, n\}.$$

In this instance the empirical distribution of interest is that of  $\{\hat{e}_i\}$ .

- (ii) a. Randomly select, with replacement,  $n$  elements from  $\{\hat{e}_i, i=1, \dots, n\}$ . Using these resampled

errors  $\hat{e}_i^*$ ,  $i=1, \dots, n$ , create the bootstrap realization of  $Y$  by  $\hat{Y}_i^* = x_i \beta + \hat{e}_i^*$ ,  $i=1, \dots, n$ . From

these  $\hat{Y}_i^*$ 's, obtain the bootstrap realization of  $\beta$ :  $\beta^*$ .

- b. Repeat (a)  $B$  times. Denote the  $B$  realizations of  $\beta$  by  $\beta_b^*$ ,  $b=1, \dots, B$ .

- (iii) The bootstrap estimate of the variance-covariance matrix of  $\beta$  is

$$\hat{\Sigma}_{B2} = \frac{1}{B-1} \sum_{b=1}^B (\beta_b^* - \beta^*)(\beta_b^* - \beta^*)^T$$

where  $\beta^* = \frac{1}{B} \sum_{b=1}^B \beta_b^*$ .

The reasons this version of the bootstrap is particularly appealing to proponents of the

model-based school is because it obtains conditional estimates of the variance of  $\beta$ :  $var(\beta | X = x)$ ,

and because for linear estimators  $\hat{\beta}$ ,  $Var_{\hat{F}_n}(\hat{\beta}^*)$  can be derived theoretically. This is appealing because the variance estimates take into account the covariates which are important as mentioned earlier in Section 3.4.2 and because there is no need to perform Monte Carlo. For  $\hat{\beta} = (X^T X)^{-1} X^T Y$ , then

$$\begin{aligned} Var_{\hat{F}_n}(\hat{\beta}) &= Var_{\hat{F}_n}[(X^T X)^{-1} X^T (X\hat{\beta} + \hat{e}^*)] , \\ &= Var_{\hat{F}_n}((X^T X)^{-1} X^T \hat{e}^*) , \\ &= (X^T X)^{-1} X^T \cdot Var_{\hat{F}_n}(\hat{e}^*) \cdot X (X^T X)^{-1} , \\ &= (X^T X)^{-1} X^T \cdot X (X^T X)^{-1} \cdot \sigma_{\hat{F}_n}^2 , \\ &= (X^T X)^{-1} \sigma_{\hat{F}_n}^2 , \end{aligned}$$

$$where \sigma_{\hat{F}_n}^2 = \frac{1}{n} \sum_1^n (\hat{e}_i - \bar{\hat{e}})^2 = \frac{1}{n} \sum \hat{e}_i^2 .$$

Efron and Gong (1983) suggest a modification to the bootstrap algorithm Bootstrap II, which

consists of resampling from  $\left\{ \left( \frac{n}{n-1} \right)^{1/2} \hat{e}_i \right\}$ . This minor modification makes the bootstrap variance

estimates unbiased in the case of  $p = 1$ . This version of the bootstrap can be calculated directly, without the use of Monte Carlo methods.

We will now derive the analogue of this version of the bootstrap variance estimator for the ratio estimator. Using the linear model for ratio estimation,

$$Y_i = R X_i + e_i, \quad E_F(e_i) = 0 \text{ and } Var_F(e_i) = \sigma^2 ,$$

we obtain a model-based version of the bootstrap variance estimator, which we shall name Bootstrap II:

(i) Estimate  $\hat{R}$  using the  $n$  observations. Let

$$\left\{ (y_1 - \hat{R}x_1) \sqrt{\frac{n}{n-1}}, \dots, (y_n - \hat{R}x_n) \sqrt{\frac{n}{n-1}} \right\} = \{\tilde{e}_i, i=1, \dots, n\} \text{ be the bootstrap residuals.}$$

(ii) a. Randomly select, with replacement, a simple random sample of size  $n$  from  $\{\tilde{e}_i\}$ . Using

these resampled errors  $\tilde{e}_i^*$ ,  $i=1, \dots, n$  create the bootstrap realization of  $Y$  by  $\tilde{Y}_i^* = x_i \hat{R} + \tilde{e}_i^*$ ,

$i=1, \dots, n$ . From these  $\tilde{Y}_i^*$ 's, obtain the bootstrap realization of  $\hat{R}$ :  $\hat{R}^*$ .

b. Repeat (a)  $B$  times. Denote the  $B$  realizations of  $\hat{R}$  by  $\hat{R}_b^*$ ,  $b=1, \dots, B$ .

(iii) Estimate the variance of  $\hat{R}$  by the sample variance of  $\hat{R}_b^*$ ,  $b=1, \dots, B$ .

As shown earlier, because the ratio estimator is linear in  $Y$ , there is no need to perform Monte Carlo simulation to estimate  $Var_{\hat{F}_*}(\hat{R}^*)$ . It can be derived directly.

$$\begin{aligned} Var_{\hat{F}_*}(\hat{R}^*) &= Var_{\hat{F}_*} \left( \frac{\sum_{j=1}^n \hat{Y}_j^*}{\sum_{j=1}^n x_j} \right), \\ &= Var_{\hat{F}_*} \left( \sum_{j=1}^n (\hat{R}x_j + \tilde{e}_j^*) \right) \cdot \frac{1}{(n\bar{x})^2}, \end{aligned}$$

$$\begin{aligned}
&= \text{Var}_{\hat{F}_*} \left( \sum_{j=1}^n \tilde{e}_j^* \right) \cdot \frac{1}{(n\bar{x})^2} , \\
&= \frac{n}{(n\bar{x})^2} \cdot \text{Var}_{\hat{F}_*}(\tilde{e}_j^*), \text{ for any } j , \\
&= \frac{n}{(n\bar{x})^2} \cdot \frac{n}{n-1} \cdot \frac{1}{n} \sum_1^n (\hat{e}_i - \hat{e}_.)^2 .
\end{aligned}$$

Therefore,

$$\begin{aligned}
\text{var}_{B2}(\hat{R}) &\equiv \text{var}_{\hat{F}_*}(\hat{R}) = \frac{n}{(n\bar{x})^2} \cdot \frac{1}{n-1} \cdot \sum_1^n \hat{e}_i^2 , \\
&= \frac{1}{n\bar{x}^2} \cdot \frac{1}{n-1} \cdot \sum_1^n (y_i - \hat{R}x_i)^2 , \\
&= \text{var}_2(\hat{R}) / (1-f) .
\end{aligned}$$

Once the finite population correction is taken into account, the model-based approach to bootstrapping leads to Cochran's second version of the variance estimator,  $\text{var}_2$ .

#### 4.4 Weighted Bootstrap Variance Estimates

In the previous two sections, we derived two estimators of the variance of the classical ratio estimator using the bootstrap. The two estimators,  $\text{var}_{B1}$  and  $\text{var}_{B2}$ , were motivated by two different approaches towards estimation of the variance of the ratio estimator:  $\text{var}_{B1}$  was derived from a design-based approach, whereas  $\text{var}_{B2}$  was derived using a model-based approach. In this section, we will derive an estimator based on a combination of both approaches.

Recall that the idea behind Wu's weighted jackknife was to weight a residual by the determinant of the expected Fisher Information matrix. Recall that in (3.5) of Chapter 3, similarly weighted jackknife residuals were used to obtain an estimate of variance. This idea can be extended to the

case of weighting bootstrap residuals. Thus, the squared deviation  $(\hat{R}_b^* - \hat{R}^*)^2$  will be weighted by the information content of the  $b^{th}$  resample. Assuming the linear model discussed in Section 2.1, the Fisher Information matrix of a bootstrap resample is simply the sum of the  $x$ 's in that resample.

We refer to this approach as a hybrid of the design and model-based approaches because the bootstrap samples of Bootstrap I, which was design-based, are averaged using weights derived from a particular model and distributional assumption. The model and distributional assumptions upon which the weights will be based, will be similar to the assumptions made in Section 3.4 where Wu's weighted jackknife variance estimator was discussed. The weighted bootstrap algorithm replaces steps (iii) of Bootstrap I with:

- (iii) Calculate the bootstrap estimate of the variance of  $\hat{R}$  by a weighted average of the squared deviations  $(\hat{R}_b^* - \hat{R}^*)^2$ . The variance estimator is

$$var_{BW}(\hat{R}) = \sum_{b=1}^B \bar{x}_b^* (\hat{R}_b^* - \hat{R}^*)^2 / \sum_{b=1}^B \bar{x}_b^*$$

where  $\bar{x}_b^*$  is the average value of  $x$  in the  $b^{th}$  resample.

For certain heteroscedastic models, both Hinkley's and Wu's weighted jackknife variance estimators correspond, to the first order, with  $var_{BW}$  (Beran, 1986). In an unpublished article, Mason and Newton (1990) discuss and prove the consistency of a weighted bootstrap using rank statistics. Thus, it seems plausible that our proposed weighted bootstrap variance estimator (which is due to Wu (1986) for regression models) may indeed be consistent for the homoscedastic

model. The  $var_{BW}$  is very difficult to approximate theoretically and consequently, Monte Carlo methods must be used.

#### 4.5 Some Comments on Proposed Estimators

For the heteroscedastic linear model, with  $var(e_i) = \sigma_i^2$ , all three bootstrap variance estimators are inconsistent. In fact, in this case consistent estimates of the variance do not exist (Wu, 1986; Beran, 1986). In heteroscedastic linear models, the residuals are not exchangeable and Wu (1986) suggests using Hadamard matrices to resample the residuals. The basic idea is to add  $k \cdot \hat{e}_i$  to  $X_i^T \beta$ , where  $k=+1$  or  $-1$ . This method assumes symmetric distributions for the errors.

Because we are estimating the variance of the ratio estimator when the data are obtained by simple random sampling from a finite population, we propose to downscale the variance estimates using the finite population correction factor:  $(1-f) = \left(1 - \frac{n}{N}\right)$ . Both Chao and Lo (1985) and

McCarthy and Snowden (1985) argue that the bootstrap needs to be modified in this way in order to apply it to estimators which use samples drawn without replacement from finite populations.

Chao and Lo (1985) argue that for a finite population sampled without replacement with  $n=N$ , the bootstrap distribution which is used to obtain  $\sigma(n, \hat{\theta}, \hat{F}_n)$  should be degenerate but is not. Thus, they argue, the (Fisher) consistency requirement of the bootstrap is not satisfied. They further argue that the bootstrap distribution which is used to obtain  $\sigma(n, \hat{\theta}, \hat{F}_n)$  estimates the true distribution of  $\hat{\theta}$  over many realizations of the data. But when  $n = N$ ,  $\hat{F}_n$  equals  $F$  and therefore, to maintain Fisher consistency,  $\hat{\sigma}_B = \sigma(N, \hat{\theta}, \hat{F}_N)$  should equal  $\sigma(N, \hat{\theta}, F)$ . However,  $\sigma(N, \hat{\theta}, \hat{F}_n)$  does



not equal  $\sigma(N, \hat{\theta}, F)$  because  $\hat{\sigma}_B$  is obtained by sampling with replacement whereas  $\sigma$  involves sampling without replacement, and is thus, degenerate. Therefore, Chao and Lo argue that the bootstrap should be modified to take into account this (Fisher) inconsistency.

On the other hand, McCarthy and Snowden (1985) argue that if sampling with replacement occurs from  $F$ , then all bootstrap samples drawn from  $\hat{F}_n$  might have been obtained as regular samples from  $F$ . However, if sampling without replacement occurs from  $F$ , then the only bootstrap sample from  $\hat{F}_n$  which might have been obtained from  $F$  is the original sample. Thus, McCarthy and Snowden argue that the bootstrap needs to be modified when used in a finite population situation.

Both Chao and Lo (1985), and McCarthy and Snowden (1985) suggest a similar alternative bootstrap resampling method to account for their respective discrepancies. The methods they use were first suggested in Gross (1980) and Bickel and Freedman (1984). (Chao and Lo claim precedence over Bickel and Freedman.)

We propose that for finite populations, when the sampling rate is small ( $f = \frac{n}{N}$  is small)

$\sigma(n, \hat{\theta}, \hat{F}_n) \cdot (1-f)$  should be used as the bootstrap estimator of  $\sigma(n, \hat{\theta}, F)$ . By standard design-based arguments, the  $(1-f)$  factor is appropriate because  $\hat{\sigma}_B$  estimates the variance of an infinite population.

## **4.6 Summary**

In this chapter, bootstrap variance estimators were introduced and applied in estimating the variance of a classical ratio estimator. Three variations of the bootstrap were derived, one of which was shown to be equivalent to a variance estimator which was considered in Chapter 2. These three estimators were obtained from model-based, design-based and combined approaches.

## **Chapter 5**

### **Sensitivity of Variance Estimates: An Empirical Study**

In this chapter, we will present the results of empirical studies on the performance of various estimators of the classical ratio estimator. In Sections 1 and 2, we will discuss the underlying motivation of the empirical study, and we will present an overview of the empirical study. In Section 3, we will describe the populations which will be used in the simulation study. Following these, in Sections 4 and 5 we will discuss the results of the empirical analysis. Finally, in Section 6, we will present a summary of the empirical results.

#### **5.1 Purpose of Empirical Study**

The primary purpose of this empirical study is to evaluate, under a variety of population characteristics, the performance of various estimators of the variance of the classical ratio estimator; thus, it is the robustness with respect to these characteristics which will be studied. In particular, we are interested in evaluating the performance of the variance estimators in the context of small sample sizes and finite populations. Recall from Chapter 1 that one of the goals of this study is to identify an efficient and accurate, small sample estimator. (This is in order to allow the Ministry of Forests to reduce the minimum sample size requirement from the present requirement of thirty).

The performance of the estimators will be evaluated via their sensitivity to various population characteristics. In particular, this study will investigate the effects of non-zero intercepts and heteroscedasticity. Recall that some of the estimators which were discussed in earlier chapters were derived from a model-based approach, whereas others were non-parametric. Consequently, it is of interest to ascertain the difference in performance between model-based and design-based

estimators. Furthermore, since these methods would presumably be applied to many data sets, with a variety of population characteristics, it is of interest to determine the robustness of the estimators under a variety of populations; i.e. populations which do and do not satisfy assumptions of the model-based approach. Thus, the "best performing" estimator will, in a sense, be the most robust with respect both to the presence of a non-zero intercept and to the variance structure.

Another important purpose is to study the performance of resampling-based estimators of the variance. Recall our earlier conjecture that because of their unique non-parametric approach, such estimators might perform better overall. In addition, we also proposed methods to accommodate the inherent imbalance of the ratio estimation problem. Thus, it is of interest to gauge the success of the weighted jackknife and the weighted bootstrap variance estimators in handling this imbalance.

## 5.2 Overview of Empirical Study

In this section, we will describe the variance estimators to be studied, along with the range of conditions under which their performance will be investigated. We will also discuss the assessment criteria which will be used to measure the performance or effectiveness of the estimators.

The estimators to be studied may be classified into three categories. First, we will use the two basic estimators  $var_0$  and  $var_2$ . We are not aware of empirical studies which considered the squared error of both  $var_0$  and  $var_2$  and the two estimators may perform quite differently in similar situations, so both are of interest. Furthermore,  $var_2$  can also be motivated using Hinkley's weighted jackknife approach. Another estimator which will be studied is that suggested by Fuller,  $var_{REG}$ . Recall that this estimator is derived using a very different approach that

attempts to correct for the deviation in the first sample moment (i.e.  $(\bar{x} - \bar{X})$ ). Thus, we will study three estimators which are not based on resampling ideas.

The second category of estimators which we will study are the jackknife based estimators. One of these,  $var_H$ , is essentially due to Hinkley and coincides with either  $var_2$  or  $var_0$  depending on how one chooses to estimate the influence function of  $\hat{R}$ . Thus, we will study the remaining two jackknife-based estimators:  $var_J$ , the classical Quenouille-Tukey variance estimator and  $var_w$ , the weighted jackknife variance estimator with weights proportional to the (approximate) information content of the subsamples. Recall that the latter estimator is obtained using the approach of Wu (1986).

The third category of estimators which we will study are bootstrap based estimators. Recall that in Chapter 4, we discussed three bootstrap estimators of the variance of the classical ratio estimator. The Bootstrap II variance estimator, based on a linear model approach can be calculated exactly, without recourse to Monte Carlo simulation, and is equivalent to  $var_2$ . However, the remaining two bootstrap estimators need to be calculated via Monte Carlo simulation, and these will be studied. The two estimators, which were derived and discussed in Sections 4.2 and 4.4, are  $var_{BI}$  and  $var_{BW}$ .

Of the seven estimators to be studied, three are model-based and four are design-based. The model-based estimators are  $var_{REG}$ ,  $var_w$  and  $var_{BW}$ . All seven estimators will be studied under a variety of conditions which reflect the concerns and issues which were discussed in Section 5.1. The two primary factors which will be studied are the sample sizes and the population characteristics.

The general format of the empirical study will be as follows. Random samples will be repeatedly drawn from a finite population consisting of  $N$  pairs  $(x_i, y_i)$ . This finite population will exhibit some of the selected characteristics of a real population. In other words, we will try to simulate a real population with selected features, and each simple random sample selected (without replacement) will represent a sample which might have been drawn from a real population. For each sample, the seven estimates of variance will be calculated and compared to an empirical estimate of the true variance; as mentioned earlier, the exact, theoretical variance of  $\hat{R}$  is difficult to calculate. The above simulations will be repeated for several different sample sizes.

The data which will form the populations will be either provided by the Ministry of Forests or will be simulated. The data from the Ministry of Forests will be large samples which have already been collected and were used by the Ministry during the calendar year 1989 to estimate actual volumes. Because these samples were drawn randomly (using block sampling) from much larger populations, they adequately approximate the population of annual truckloads from which they were drawn. And since the primary objective of this thesis is to identify effective small sample variance estimators for the types of populations faced by the Ministry, the data provided by the Ministry will be ideal for the purpose of empirical study.

The simulated data will be created using random number generators and will have characteristics resembling those of real populations. That is, the simulated data will consist of pairs  $(x_i, y_i)$  where the distribution of the simulated  $x_i$ 's will be very similar to the distribution of the real  $x_i$ 's from the Ministry's data; details will be discussed in the next section. The important point to note is that the artificial populations will have marginal and joint distributions similar to the marginal and joint distributions exhibited by the Ministry data.

The assessment criteria which we will use to determine the performance of the seven variance estimators must reflect, in some sense, the underlying problem. Thus, it seems appropriate to use the average squared error of the variance estimates as the primary performance yardstick. As a secondary criteria, the percentage bias of the variance estimators will also be calculated. We will present the details of the above criteria in the following sections. At this point it is sufficient to note that the performance criteria will be the mean squared error and the bias.

### **5.3 Description of Populations**

In this section we will describe in detail the populations upon which the simulations will be based. We will first describe the actual populations and later, the artificial populations.

The Ministry of Forests provided a total of 18 large samples which were collected during the calendar year 1989. These samples were drawn from populations of truckloads consisting of an average of 5,000 truckloads. The sample sizes range from 150 to 250 truckloads. For our purposes here, we will consider these 18 large samples as populations.

The data for these 18 populations are graphed in Figures 1a and 1b, where to facilitate visual comparison all populations are plotted on the same scales. A point to note is that the correlation appears high because the data are plotted on scales which are larger than the range of the data. Of the 18 populations, only populations 4 and 9 have significant outliers; outliers in the sense that the observations are grossly unlike other observations in that population. However, further study of the plots reveal that some of the populations are quite similar in structure; although maybe not in the scale or the range of the data. The data for some populations seem ideal for the linear model assumptions used in earlier chapters: examples are populations 5, 12, 15, 17 and 18. On the other hand, the data for some populations such as 2, 3 and 6, appear to exhibit non-zero intercepts, though it is not obvious what relationship exists between weight and volume in these

populations. Finally, some populations such as 15, 16 and 17 appear odd in that the distribution of weights seems to be multi-modal.

We have selected three populations from these 18 for the simulation study. The three particular populations were chosen because they represented, broadly speaking, three groups around which the other populations could be classified. In other words, using the intercept, distribution of  $x$ 's and variance structure, we identified three major population types and selected a population of each type. One type, with zero intercept and homoscedastic variance, is the most prevalent among the 18 populations. Population 5 was selected to represent this group and will be referred to as population B. Another group is comprised of those populations with zero intercept and either heteroscedastic variance or multi-modal distribution of weights. This group appears to be the least prevalent among the 18 populations. Population 16 was selected to represent this population and it will be referred to as population C. Population 3, referred to as population A, will represent the group of populations which exhibit both non-zero intercept and slight heteroscedasticity.

Population A, graphed in Figure 2, has truckload weights which range from 27 tonnes to 41 tonnes. These weights appear to be normally distributed with few, if any, outliers. The volumes of the truckloads range from just under 30 cubic meters to just over 50 cubic meters. This data set does not appear to satisfy the assumption of a zero intercept. A linear regression suggests that the intercept is about 30 cubic metres. However, the variance structure of the data appears to be roughly homoscedastic; that is, the variance of the volumes appears constant, regardless of the weights. There are 191 members in this population and the correlation between weight and volume is 0.29.

Population B, graphed in Figure 3, has the largest range of both weight and volume among all 18 populations. The weights of the truckloads range from 25 tonnes to almost 100 tonnes, whereas



the volumes range from just under 40 cubic meters to 100 cubic meters per truckload. This data set satisfies the assumption of approximately zero intercept and the variance structure appears to be homoscedastic. There were 213 members in this population and the correlation between weights and volumes was 0.87.

Population C, graphed in Figure 4, has weights which range from 26 tonnes to 40 tonnes. The volumes of the truckloads range from 35 cubic meters to 55 cubic meters. This data set satisfies the assumption of approximately zero intercept, however the distribution of weights is multi-modal. The variance appears to be constant, regardless of weights. There were 206 members in this population and the correlation between weight and volume is 0.78.

The distribution and parameters of the artificial data were selected based on the distributions of the weights and estimated errors from the Ministry data. In the majority of Ministry populations, the weights were normally distributed, with the mean approximately 35 tonnes and the standard deviation approximately 5 tonnes. The estimated errors, obtained from linear regressions, were also normally distributed, with the mean (obviously) 0 and the standard deviation approximately 2 cubic meters. The ratio was selected to be  $4/3$  because the ratio of the 18 Ministry populations ranged from 1 to  $3/2$ . Furthermore, because the Ministry draws samples from population which range in size from 100 to 5,000, the artificial populations were selected to have 1,000 members.

The artificial data consisted of six populations. The ideal population was created in the following manner: A vector of 1,000 normally distributed, independent random variables was created, with mean 35 and standard deviation 5. Another vector of 1,000 normally distributed, independent random variables with mean 0 and standard deviation 2 was also generated. The first vector simulated the  $x$ 's (i.e. weights) and the second simulated the errors. Using a slope of  $(4/3)$ , the "random"  $y_i$ 's for the first population were generated by multiplying the vector of  $x_i$ 's by  $(4/3)$  and

adding the vector of errors. Thus, the first population was "ideal" in the sense that the intercept was 0 and the errors were normally distributed with constant variance. We shall denote this population by Y00 - the first digit indicates the error structure and the second digit indicates the zero intercept. More specifically, the first digit indicates  $\lambda$  in the variance assumption:

$var(e_i) = \sigma^2 x_i^\lambda$ . The second digit is a categorical indicator. A zero means no intercept (i.e. zero intercept) and a one means a non-zero intercept.

The other five populations were generated as variations of the first population using the same vectors of weights and errors. Two additional populations were created with zero intercept. However, the variance structure of the populations differed from Y00. One of the populations satisfied the error structure  $\sigma^2 x_i$  and the other satisfied the error structure  $\sigma^2 x_i^2$ . These two populations were created by appropriate addition of the errors to the  $x_i$ 's. Let  $x_i$  represent the elements of the vector of weights and let  $e_i$  represent the elements of the vector of errors. Then the three populations Y00, Y10 and Y20 were generated by:

$$Y00: Y_i = (4/3) x_i + e_i, \quad i=1, \dots, 1000;$$

$$Y10: Y_i = (4/3) x_i + e_i \sqrt{x_i}, \quad i=1, \dots, 1000; \text{ and,}$$

$$Y20: Y_i = (4/3) x_i + e_i x_i, \quad i=1, \dots, 1000.$$

The other three artificial populations, Y01, Y11 and Y21, were constructed by the addition of a constant to the volumes from the three populations Y00, Y10 and Y20. It seemed appropriate to select a value of the intercept that in some way reflected the magnitude of the covariates  $x_i, y_i$  and the slope. The following relationship

$$\bar{Y} = \alpha + \frac{4}{3} \bar{X} + \bar{e}$$

suggested that the intercept be set at some proportion of  $(4/3)\bar{X}$ . The value of the intercept which we chose was set at 50% of  $(4/3)\bar{X}$  because it seemed that for a higher value, one would not normally use a ratio estimator; for a smaller value, the effect of the non-zero intercept would not be significant. The constant  $\alpha$  was calculated as follows:

$$\alpha = (0.50)\bar{X}R = (0.50)(4/3)\bar{X} = (0.50)(4/3)(35) = 23.3m^3.$$

To summarize, six artificial populations were created. They were created with parameter values and distributions similar to real populations. However, selected features of these artificial populations were controlled to allow the study of their effects on the performance of the variance estimators. The populations are denoted by  $Y_{ij}$  where  $i \in \{0,1,2\}$  indicates the exponent of  $x$  in the following variance structure:  $Var(e_i) = \sigma^2 x_i^\lambda$ . The presence of a non-zero intercept ( $= 23.3m^3$  in each case) is indicated by  $j = 1$ ; otherwise, for a zero intercept,  $j = 0$ . The six populations are  $Y_{00}, Y_{10}, Y_{20}, Y_{01}, Y_{11}, Y_{21}$ .

As was mentioned in earlier chapters, the estimate of interest is the estimate of total volume obtained using the classical ratio estimate. However, for any given population, the difference between these two estimates ( $\hat{R}$  and  $\hat{Y}$ ) is the constant  $(N\bar{X})$ . Thus, to avoid incorporation of this constant, the empirical study will concentrate upon  $\hat{R}$  and not  $\hat{Y}$ . The results and conclusions of the study will not differ in any way because of this change.

A brief description of the simulations will now be given. For each sample size to be investigated, 5,000 simple random samples will be drawn without replacement. All seven variance estimators

will be evaluated for each simulated sample; for the bootstrap variance estimators, 2,500 bootstrap samples will be drawn from each simulated sample. All seven estimators will be simulated at sample sizes of 5, 10, 20, 30, 50 and 100 for the real populations, and at sample sizes of 10, 20 and 50 for the artificial populations. Since all three real populations are approximately of size 200, the simulations with sample size 100 will be of particular interest in light of the arguments for a finite population modification to the bootstrap.

The assessment of the performance of the variance estimators will be in terms of the relative mean squared error and the percentage bias of these estimators as estimates of the standard error of  $\hat{R}$ ; the evaluation of these performance criteria will now be described. For each sample size, the mean squared error is approximated from the 5,000 estimates of  $\hat{R}$  as follows:

$$MSE(\hat{R}) = \frac{1}{5000} \sum_{i=1}^{5000} (\hat{R}_i - R)^2 ,$$

where  $\hat{R}_i$  is the estimate of  $R$  from the  $i^{th}$  sample, and  $R$  is the true population ratio.

The percentage bias of method  $t$  is calculated as

$$PB_t = \left\{ \frac{\frac{1}{5000} \sum_{i=1}^{5000} (\sqrt{v_{t,i}}) - \sqrt{MSE(\hat{R})}}{\sqrt{MSE(\hat{R})}} \right\} \cdot 100$$

where  $v_{t,i}$  is the variance estimate based on method  $t$  and calculated from the  $i^{th}$  simulated sample.

Similarly, the mean squared error of method  $t$  as an estimator of the standard error of  $\hat{R}$  is calculated as:

$$MSE_t = \frac{1}{5000} \sum_{i=1}^{5000} \left( \sqrt{v_{ti}} - \sqrt{MSE(\hat{R})} \right)^2 .$$

Finally, the relative mean squared error of method  $t$  is obtained by calculating the percentage difference between its mean squared error and the minimum mean squared error among all methods for a given sample size; that is,

$$RMSE_t = \left\{ \frac{MSE_t - \min_u [MSE_u]}{\min_u [MSE_u]} \right\} \cdot 100$$

The results obtained for all seven methods of variance estimation and all sample sizes for a given real or artificial population, will be presented in a single table. From these results, we will attempt to draw general conclusions regarding the relative performance of these methods of estimating the variance of  $\hat{R}$  in the context of populations similar to those faced by the Ministry of Forests.

## 5.4 Results from Real Populations

We will first discuss the results from each population and will thereafter summarize the overall results from the three Ministry data sets.

The results for population A, which was approximately heteroscedastic with zero intercept, summarized in Table 1, indicate that Cochran I and Fuller's method are the best estimators regardless of the sample size. For the smaller sample sizes ( $n=5, 10, 20$ ) these two methods outperform the other methods by approximately 5%. For the larger sample sizes, ( $n=30, 50, 100$ )

Cochran I and Fuller's method significantly out-perform the remaining methods, with a mean squared error approximately 10% lower than for the other methods. The bias of the bootstrap methods is significantly higher than that of the other methods; but all methods underestimate the true standard error for the smaller sample sizes. However, at the larger sample sizes, all methods are approximately unbiased. As the sample size increases, the bias of all methods is steadily reduced from -5% at  $n=5$  to -0.5% at  $n=50$ . Peculiarly, for  $n=100$  the bias increases to -1.3%. As for the weighted versions of the resampling methods, both the weighted jackknife and the weighted bootstrap have mean squared errors similar to mean squared errors of their unweighted counterparts. The bootstrap methods, though approximately unbiased, have much larger mean squared errors than the other methods, and their relative mean squared error rises with increasing sample sizes. This may be due to the need for a finite population modification to the bootstrap.

The results for population B, which was close to ideal, tabulated in Table 2, can be summarized as follows: All methods, except the two bootstrap-based methods at larger sample sizes, perform similarly. Once again, Cochran I and Fuller's method have the lowest mean squared error. However, the remaining methods do not show significantly larger mean squared errors. The average difference in the squared errors between the methods is approximately 2%. Once again, the bootstraps do not perform very well at the largest sample size, though they do perform well at sample sizes less than 50. In both this and the previous population, the bias of the bootstrap methods does not differ greatly from the other methods, however the relative mean squared error does differ significantly. The weighted resampling methods perform similar to their unweighted counterparts. Once again, the bias of all methods is steadily reduced as the sample size increases. However at  $n=50$ , the bias increases and becomes positive, before dropping to almost zero at  $n=100$ .

The results for population C, which exhibited a non-zero intercept and multi-modal distribution of weights, summarized in Table 3, are very similar to the results from population B. Once again, all methods perform similarly, both in terms of mean squared error and in terms of bias. However, for this population, the bootstraps do not show as high a relative squared error at the largest sample size as for population B. Once again, the bias has peculiar behaviour as the sample size increases.

The combined results from all three populations are as follows. For the larger sample sizes, all methods except for the bootstrap perform similarly. All methods have approximately the same mean squared error and are approximately unbiased, except for the bootstrap which has longer relative mean squared errors. For the smaller sample sizes the methods differ and, there is sometimes, significant underestimation of the standard error of  $\hat{R}$ . The weighted jackknife has a mean squared error which is slightly smaller than the mean squared error of the unweighed jackknife. The Cochran II method, which is also Hinkley's weighted jackknife, has a mean squared error and bias which is quite similar to that of the two jackknife-based methods. However, the Cochran II method has a consistently lower mean squared error than the unweighted bootstrap. In summary, it is the particular characteristics of the population that determine the quality of performance of a particular method for that population.

Based on the above, the following conclusions may be drawn. For populations with zero intercept and slight heteroscedasticity, Cochran I and Fuller's method perform best, regardless of sample size. For populations with non-zero intercept and weights distributed with multiple modes, all seven methods perform similarly. Thus, one may conclude that Cochran I and Fuller's method are the most robust with respect to these three populations. However, it is difficult to generalize these conclusions; it is still not known which estimates are robust with respect to the intercept, which are robust with respect to the variance structure and which are robust with respect to both. This

is because we have not isolated these population characteristics and studied their effects in a more controlled setting. Therefore, to study the sensitivity of the seven estimators, a similar simulation was performed on artificial data. However, these simulations will not involve sample sizes with large sampling fractions because we have already established that at these large sampling fractions the bootstrap methods have higher mean squared error.

## **5.5 Results from Artificial Populations**

The purpose of conducting empirical studies on artificial populations is to study the effects of non-zero intercept and heteroscedasticity on the relative performance of the various estimators. The six artificial populations described in Section 5.3 were created with this in mind. The six populations can, in a very general sense, be described as the results of a 3 by 2 factorial design; the factors being, respectively, heteroscedasticity and non-zero intercept.

The results of the empirical study on these artificial populations are summarized in Tables 4 - 9. For each of the six artificial populations, all seven methods were tested at sample sizes 10, 20 and 50. Because of the consistency of results across sample sizes, sample sizes 5, 30 and 100 were not used. Rather than present the results on a population by population basis, the results will be presented with respect to the two factors. The sensitivity of the estimators with respect to the intercept will be discussed first, followed by sensitivity with respect to the variance structure and finally, with respect to both factors.

All results are consistent across the sample sizes. Thus, if one method performs well at the smallest sample size, it will also perform well at larger sample sizes. Consequently, except for the following two observations, no further mention of change over sample sizes will be made. As sample sizes get larger, the discrepancies between the different methods get smaller, as one would expect. The second observation is that for the smaller sample sizes, the bias is consistently



negative; however, the bias reduces at the larger sample sizes resulting in approximately unbiased estimators.

For the ideal population Y00 (homoscedastic and zero intercept), Cochran I and Fuller's method are slightly better than the other methods in terms of mean squared error, the difference being approximately 2%. However, as we move towards homoscedastic population with a non-zero intercept (from Y00, Table 4 to Y01, Table 7), it is clear that Cochran I and Fuller's method have a substantially smaller mean squared error than the other methods, though all methods are approximately unbiased. Similarly, for populations with heteroscedastic variance structure, the presence of a non-zero intercept (from Y10, Table 5 to Y11, Table 8 and from Y20, Table 6 to Y21, Table 7) causes Cochran I and Fuller's method to either have smaller mean squared errors than the other methods or to reduce the difference in the mean squared errors between these two groups. Thus, it may be concluded that Cochran I and Fuller's method are more robust with respect to non-zero intercept than the other methods. In terms of bias, there are minor differences between the two groups in the presence of a non-zero intercept.

In the presence of heteroscedasticity, Cochran II and the resampling-based estimators have smaller mean squared errors than Cochran I and Fuller's method. For populations with zero intercept, the presence of heteroscedasticity causes an increase in the relative mean squared error of Cochran I and Fuller's estimator, though the bias remains unaffected. For populations with non-zero intercept, the presence of heteroscedasticity once again causes the relative mean squared errors of Cochran I and Fuller's method to increase. However, because these two methods were substantially better than the other methods for homoscedastic and non-zero intercept populations (Y01), it is not until substantial heteroscedasticity is present (Y21) that Cochran I and Fuller's method exhibit larger mean squared error. Based on the preceding results, we conclude that Cochran II and the resampling methods are more robust with respect to heteroscedasticity than

the other two methods. It is interesting to note that as we increase heteroscedasticity in the presence of non-zero intercept (from Y01, Table 7 to Y11, Table 8 to Y21, Table 9), resampling-based methods along with Cochran I, improve relative to the analytic methods. The relative improvement is possibly due to the analytic methods performing worse than the resampling-based methods in the presence of both heteroscedasticity and non-zero intercept.

In terms of both a non-zero intercept and non-constant variance, no estimator performs consistently well; that is, no estimator is robust with respect to both non-zero intercept and non-constant variance. In the factorial design analogy, there are no interaction effects. That is, some estimators are robust with respect to heteroscedasticity whereas others are robust with respect to non-zero intercept. However, no estimators are robust with respect to the presence of both factors.

The performance of the resampling estimators was very similar in terms of the mean squared errors; for all population sizes, the mean squared errors of the four estimators rarely differed by more than 2%. However, the bootstrap-based estimators were more biased than the jackknife-based estimators at the smaller sample sizes i.e.  $n = 10, 20$ . On average, the underestimation of the standard error of  $\hat{R}$  by the bootstrap-based estimators was approximately 5% less than the underestimation by the jackknife-based estimators. In terms of the weighted and unweighted versions, the weighted jackknife always outperformed the unweighted jackknife, though the average difference in mean squared errors was approximately 0.5%. For non-ideal populations, the bias of the weighted jackknife slightly exceeded the bias of the unweighted jackknife. The weighted and unweighted bootstraps showed different robustness; the weighted bootstrap was more robust with respect to heteroscedasticity, whereas the unweighted bootstrap was more robust with respect to non-zero intercept. For both the jackknife and the bootstrap, the differences between the weighted and unweighted versions were minimal. The mean squared errors rarely differed by more than 1%.

It is clear from the table that for a homoscedastic population, regardless of the intercept, Cochran I and Fuller's method perform better than the remaining five methods. In fact, when the intercept is non-zero and the error structure is homoscedastic, Cochran I and Fuller's method produce mean squared errors approximately 10% lower than those of the other methods. Of the five remaining methods, the performance is very similar.

On the other hand, for very heteroscedastic populations, that is Y20 and Y21, Cochran II and the resampling methods out-perform Cochran I and Fuller's method. However, in these instances the discrepancy between the two groups is half of what it was for the homoscedastic case; that is 5%.

For the populations Y10 and Y11, the results are quite interesting: The resampling methods, along with the Cochran II method, out-perform (by approximately 2%) the remaining two methods for small sample sizes when the intercept is zero. However when the intercept is non-zero, the two methods (Cochran I and Fuller's) out-perform the remaining methods by approximately 9%.

## **5.6 Summary**

The overall conclusions of the study on artificial and real populations are that for a homoscedastic population, with zero intercept, one should use either Cochran I or Fuller's method; since Cochran I is much simpler, it should be used instead of Fuller's method. If the population is highly heteroscedastic, then the resampling-based methods or the Cochran II method should be used. The presence of a non-zero intercept should be considered together with the variance structure. If the violation of the homoscedastic variance assumption exceeds, in a qualitative sense, the violation of the zero intercept assumption, then the resampling based methods should be preferred over the remaining two methods. Otherwise, Cochran I or Fuller's method should be used.

In terms of sensitivity to population characteristics, Cochran I and Fuller's method are robust with respect to a non-zero intercept, whereas the other five estimators are more robust with respect to heteroscedasticity. As to the robustness with respect to both characteristics, no estimators performed consistently well. However, due to the higher spread between the two groups for homoscedastic and non-zero intercept populations than for heteroscedastic and zero intercept populations, Cochran I and Fuller's method could be said to be more robust.

As for the smaller sample sizes, all estimators underestimate the true mean squared error, with average underestimation of approximately 4% at sample size 10 and approximately 2% at sample size 20. The jackknife-based methods had a consistently lower bias than the other methods, whereas the bootstrap-based methods had a consistently higher bias.

The resampling-based estimators performed very similarly - especially if we consider the Cochran II method as a weighted jackknife method. This seems reasonable not only because the Cochran II estimator can be derived from the influence function-based approach to weighting the jackknife estimator, but also because its performance in the empirical study was very similar to the resampling-based methods. Thus, it is apparent that the weighted versions of the jackknife and bootstrap do not perform significantly better than their unweighted counterparts within the context of estimation of the variance of the classical ratio estimator, and within the context of the Ministry of Forests' data.

## Chapter 6

### Conclusions and Recommendations

In this chapter, specific conclusions concerning weigh scaling at the Ministry of Forests will be made. In particular, the issues of variance estimation and modification of resampling-based estimators will be addressed. Recommendations for the choice of robust variance estimators will also be made.

The first, and possibly most significant, conclusion is that the current variance estimator, Cochran II, is effective at sample sizes well below the present minimum requirement of 30. Furthermore, empirical studies show that the seven methods studied all performed consistently well in terms of squared error at sample sizes as low as 10. However at smaller sample sizes, the estimators also underestimated the true standard error of  $\hat{Y}$ , the average percentage bias being approximately -4%. This implies that confidence intervals constructed at these small sample sizes are 5% shorter on average than they would be if the estimators were unbiased. Therefore, a 1% confidence interval estimated using 10 samples will likely have a lower than 1% probability of including the population total. The bias reduces to -2% at sample size 20, and the estimators are effectively unbiased at larger sample sizes. Thus, the issue of bias only arises at smaller sample sizes.

The second conclusion is that all seven methods have approximately similar mean squared errors. At the smaller sample sizes, the difference between the estimators range from 5-10%, the exact differences being dependent upon population characteristics. Cochran I and Fuller's method were found to be bias-robust, whereas Cochran II and the resampling-based estimators were found to be heteroscedastic-robust. No estimators were robust with respect to both bias and heteroscedasticity.

To determine the best estimators for use in weigh scaling, the Ministry would have to decide whether non-zero intercepts or heteroscedasticity is the major "violation" in their data, and then select an appropriately robust estimator.

The third and final conclusion is that in terms of squared error and bias, the unweighted resampling-based variance estimators performed almost as well as their weighted counterparts. Perhaps in different estimation problems, the weighted resampling-based methods would perform differently from their unweighted counterparts. However, for the variance of the classical ratio estimator, the two versions have approximately similar mean squared errors.

There are three major areas that require further research. The first is the empirical performance of the estimators at small sample sizes for non-normally distributed and finite populations. Recall that in the empirical study discussed in Chapter 5, the errors and the weights were simulated using the normal distribution. It is not clear whether our conclusions, particularly about the small sample effectiveness of the estimators, can be extended to non-normal populations. Furthermore, the finite population adjustment needs to be studied, particularly for samples with large sampling fractions. These issues were not dealt with in this thesis because the Ministry populations are approximately normal and because their samples are drawn with small sampling fractions. A second area which requires further study is the combination of variance and ratio estimators. Recall that in Chapter 2 we choose to restrict our attention to the classical ratio estimator, and thus these conclusions only apply to estimation of population total by the classical ratio estimator. Perhaps there exists a combination of ratio and variance estimators which is in some sense optimal for the estimation of confidence intervals in the context of the Ministry of Forests' applications. Finally, a third area which requires further research is resampling-based confidence intervals, particularly bootstrap-based confidence intervals. These appear to be particularly promising because of their non-parametric and non-symmetric approaches.

The specific recommendations arising from this research are as follows: The Ministry of Forests should select either Cochran I or Cochran II as the variance estimator of their estimates of population total volume. The particular choice of estimator should be based on the types of populations encountered by the Ministry. The minimum sample size restriction may be reduced from 30 samples to 10 samples.

## Bibliography

1. Beale, E.M.L. (1962). Some Use of Computers in Operational Research. *Industrielle Organisation*, 31, 27-28.
2. Beran, R. (1986). Comment on "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis" by C.F.J. Wu. *Annals of Statistics*, 14, 1295-1298.
3. Bickel, P.J. and Freedman, D.A. (1981). Some Asymptotic Theory for the Bootstrap. *Annals of Statistics*, 9, 1196-1217.
4. Bickel, P.J. and Freedman, D.A. (1984). Asymptotic Normality and the Bootstrap in Stratified Sampling. *The Annals of Statistics*, 12, 470-482.
5. Brewer, K.W.R. (1963). Ratio Estimation in Finite Populations: Some Results Deducible from the Assumption of an Underlying Stochastic Process. *Australian Journal of Statistics*, 5, 93-105.
6. Carroll, R.J., Ruppert, D. and Wu, C.F.J. (1986). Generalized Least Squares: Variance Expansions, the Bootstrap and the Number of Cycles. Unpublished.
7. Chakrabarty, R.P. and Rao, J.N.K. (1968). The Bias and Stability of the Jackknife Variance Estimator in Ratio Estimation. *Journal of the American Statistical Association*, 63, 748-749.
8. Chao, M-T. and Lo, S-H. (1985). A Bootstrap Method for Finite Population. *Sankhya*, 47, Series A, 399-405.
9. Cochran, W.G. (1977). **Sampling Techniques**. Third Edition. John Wiley and Sons, Inc., New York, New York.
10. Davison, A.C. and Hinkley, D.V. (1988). Saddlepoint Approximations in Resampling Methods. *Biometrika*, 75, 417-431.
11. DiCiccio, T.J. and Romano, J.P. (1988). A Review of Bootstrap Confidence Intervals. *Journal of the Royal Statistical Society, Series B*, 50, 338-354.
12. Durbin, J. (1959). A Note on the Application of Quenouille's Method of Bias Reduction to the Estimation of Ratios. *Biometrika*, 46, 477-480.
13. Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7, 1-26.
14. Efron, B. (1982). **The Jackknife, the Bootstrap and Other Resampling Plans**. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.
15. Efron, B. (1990). More Efficient Bootstrap Computations. *Journal of the American Statistical Association*, 85, 79-89.
16. Efron, B. and Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife and Cross-Validation. *The American Statistician*, 37, 36-48.



17. Efron, B. and Stein, C. (1981). The Jackknife Estimate of Variance. *The Annals of Statistics*, 9, 586-596.
18. Fieller, E.C. (1932). The Distribution of the Index in a Normal Bivariate Population. *Biometrika*, 24, 428-440.
19. Fox, T., Hinkley, D.V., and Larntz, K. (1980). Jackknifing in Non-Linear Regression. *Technometrics*, 22, 29-33.
20. Freedman, D.A. and Peters, S.C. (1984). Bootstrapping a Regression Equation: Some Empirical Results. *Journal of the American Statistical Association*, 79, 97-106.
21. Fuller, W.A. (1981). Comment on "An Empirical Study of the Ratio Estimator and Estimators of its Variance", by R.M. Royall and W.G. Cumberland. *Journal of the American Statistical Association*, 76, 78-80.
22. Graham, R.L., Hinkley, D.V., John, P.W.M. and Shi, S. (1990). Balanced Design of Bootstrap Simulations. *Journal of the Royal Statistical Society, Series B*, 52, 1, 185-202.
23. Gross, S.T. (1980). Median Estimation in Sample Surveys. *Proceedings of the American Statistical Association, Section on Survey Research Methods*.
24. Hampel, F.R. (1968). Contributions to the Theory of Robust Estimation. Ph.D. Thesis, Department of Statistics, University of California, Berkeley, California.
25. Hampel, F.R. (1974). The Influence Curve and its Role in Robust Estimation. *Journal of the American Statistical Association*, 69, 383-393.
26. Hartley, H.O. and Ross, A. (1954). Unbiased Ratio Estimators. *Nature*, 174, 270-271.
27. Hesterberg, T. (1988). Variance Reduction Techniques for Bootstrap and Other Monte Carlo Simulations. Ph.D. Thesis, Department of Statistics, Stanford University, Stanford, California.
28. Hinkley, D.V. (1977). Jackknifing in Unbalanced Situations. *Technometrics*, 19, 285-292.
29. Hinkley, D.V. (1978). Improving the Jackknife with Special Reference to Correlation Estimation. *Biometrika*, 65, 13-21.
30. Hinkley, D.V. (1988). Bootstrap Methods. *Journal of the Royal Statistical Society, Series B*, 50, 321-337.
31. Huber, P.J. (1977). **Robust Statistical Procedures**. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.
32. Hutchison, M.C. (1971). A Monte Carlo Comparison of Some Ratio Estimators. *Biometrika*, 58, 313-321.
33. Jaeckel, L.A. (1972). The Infinitesimal Jackknife. Unpublished.

34. Kish, L. and Frankel, M.R. (1974). Inference from Complex Samples. *Journal of the Royal Statistical Society, Series B*, 36, 1-37.
35. Krewski, D. and Chakrabarty, R.P. (1981). On the Stability of the Jackknife Variance Estimator in Ratio Estimation. *Journal of Statistical Planning and Inference*, 5, 71-78.
36. Mallows, C.L. (1975). On Some Topics in Robustness. Unpublished.
37. Mason, D.M. and Newton, M.A. (1990). A Rank Statistics Approach to the Consistency of a General Bootstrap. Unpublished.
38. McCarthy, P.J. and Snowden, C.B. (1985). The Bootstrap and Finite Population Sampling. *Vital and Health Statistics, Series 2*, No. 95.
39. Mickey, M.R. (1959). Some Finite Population Unbiased Ratio and Regression Estimators. *Journal of the American Statistical Association*, 54, 594-612.
40. Miller, R.G. Jr. (1974a). The Jackknife - A Review. *Biometrika*, 61, 1-15.
41. Miller, R.G. Jr. (1974b). An Unbalanced Jackknife. *The Annals of Statistics*, 2, 880-891.
42. Quenouille, M.H. (1949). Approximate Tests of Correlation in Time-Series. *Journal of the Royal Statistical Society, Series B*, 68-84.
43. Quenouille, M.H. (1956). Notes on Bias in Estimation. *Biometrika*, 43, 353-60.
44. Rao, J.N.K. (1968). Some Small Sample Results in Ratio and Regression Estimation. *Journal of the Indian Statistical Association*, 6, 160-168.
45. Rao, J.N.K. and Webster, J.T. (1966). On Two Methods of Bias Reduction in the Estimation of Ratios. *Biometrika*, 53, 571-577.
46. Rao, P.S.R.S. (1969). Comparison of Four Ratio-Type Estimates Under a Model. *Journal of American Statistical Association*, 64, 574-580.
47. Rao, P.S.R.S. (1979). On Applying the Jackknife Procedure to the Ratio Estimator. *Sankhya, Series C*, 115-126.
48. Rao, P.S.R.S. and Rao, J.N.K. (1971). Small Sample Results for Ratio Estimators. *Biometrika*, 58, 625-630.
49. Robinson, J. (1987). Conditioning Ratio Estimates Under Simple Random Sampling. *Journal of the American Statistical Association*, 82, 826-831.
50. Royall, R.M. (1970). On Finite Population Sampling Theory Under Certain Linear Regression Models. *Biometrika*, 57, 377-387.
51. Royall, R.M. (1971). *Linear Regression Models in Finite Population Sampling Theory. Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott (eds.). Holt, Rinehart & Winston, Toronto, Ontario, 259-279.

52. Royall, R.M. and Cumberland, W.G. (1978). Variance Estimation in Finite Population Sampling. *Journal of the American Statistical Association*, 73, 351-358.
53. Royall, R.M. and Cumberland, W.G. (1981). An Empirical Study of the Ratio Estimator and Estimators of its Variance. *Journal of the American Statistical Association*, 76, 66-77.
54. Royall, R.M. and Cumberland, W.G. (1985). Conditional Coverage Properties of Finite Population Confidence Intervals. *Journal of the American Statistical Association*, 80, 355-359.
55. Royall, R.M. and Eberhardt, K.R. (1975). Variance Estimates for the Ratio Estimator. *Sankhya*, Series C, 37, 43-52.
56. Rubin, D.B. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, 9, 130-134.
57. Schenker, N. (1985). Qualms about Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, 80, 360-361.
58. Shao, J. and Wu, C.F.J. (1987). Heteroscedasticity-Robustness of Jackknife Variance Estimators in Linear Models. *The Annals of Statistics*, 15, 1563-1579.
59. Simonoff, J.S. and Tsai, C-L. (1986). Jackknife-Based Estimators and Confidence Regions in Non-Linear Regression. *Technometrics*, 28, 103-112.
60. Singh, K. (1981). On the Asymptotic Accuracy of Efron's Bootstrap. *The Annals of Statistics*, 9, 1186-1195.
61. Smith, T.M.F. (1981). Comment on "An Empirical Study of the Ratio Estimator and Estimators of its Variance", by R.M. Royall and W.G. Cumberland. *Journal of the American Statistical Association*, 76, 78-80.
62. Thernau, T. (1983). Variance Reduction Techniques for the Bootstrap. Ph.D. Thesis, Department of Statistics, Stanford University, Stanford, California.
63. Tin, M. (1965). Comparison of Some Ratio Estimators. *Journal of the American Statistical Association*, 60, 294-307.
64. Tukey, J.W. (1958). Bias and Confidence in not Quite Large Samples, Abstract. *Annals of Mathematical Statistics*, 29, 614.
65. Wu, C.F.J. (1982). Estimation of Variance of the Ratio Estimator. *Biometrika*, 69, 183-189.
66. Wu, C.F.J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, 14, 1261-1295.
67. Wu, C.F.J. and Deng, L.Y. (1983). *Estimation of Variance of the Ratio Estimator: An Empirical Study. Scientific Influence, Data Analysis and Robustness*. Academic Press, Inc., New York, New York.

Table 1

**Population A**  
**Empirical Results of Variance Estimators**

RMSE=Relative MSE PB=Percentage Bias	Sample Size											
	5		10		20		30		50		100	
Method	RMSE	PB	RMSE	PB	RMSE	PB	RMSE	PB	RMSE	PB	RMSE	PB
Cochran I	0.0	-4.7	0.0	-2.8	0.2	-2.7	0.2	-0.5	0.6	-0.2	1.8	-1.3
Cochran II	5.1	-4.2	5.7	-2.6	7.0	-2.6	8.0	-0.5	7.8	-0.2	9.0	-1.3
Fuller	0.1	-4.6	0.0	-2.7	0.0	-2.6	0.0	-0.5	0.0	-0.2	0.0	-1.3
Jackknife	5.4	-4.3	6.1	-2.7	7.3	-2.7	8.2	-0.5	8.0	-0.2	9.3	-1.3
Wtd. Jackknife	5.3	-4.3	5.9	-2.6	7.2	-2.7	8.1	-0.5	7.9	-0.2	9.2	-1.3
Bootstrap I	3.7	-13.8	7.6	-7.3	12.2	-4.9	10.1	-2.0	11.6	-1.1	26.0	-1.8
Wtd. Bootstrap	3.7	-14.1	7.8	-7.4	12.6	-5.0	10.2	-2.1	11.7	-1.2	26.3	-1.8

Results are summarized from 5,000 simple random samples. Bootstrap estimates involved 2,500 resamples.

Table 2

**Population B**  
**Empirical Results of Variance Estimators**

RMSE=Relative MSE PB=Percentage Bias	Sample Size											
	5		10		20		30		50		100	
Method	RMSE	PB	RMSE	PB	RMSE	PB	RMSE	PB	RMSE	PB	RMSE	PB
Cochran I	0.0	-7.4	0.0	-3.7	0.0	-1.6	0.4	-0.3	0.7	1.2	0.6	-0.5
Cochran II	0.9	-7.0	1.9	-3.4	1.6	-1.4	0.5	-0.2	1.3	1.2	1.4	-0.5
Fuller	0.0	-7.3	0.0	-3.7	0.0	-1.5	0.0	-0.3	0.0	1.2	0.0	-0.5
Jackknife	2.9	-6.2	2.6	-2.9	1.8	-1.2	0.7	-0.1	1.7	1.3	1.3	-0.5
Wtd. Jackknife	1.8	-6.6	2.2	-3.2	1.7	-1.3	0.6	-0.2	1.5	1.2	1.4	-0.5
Bootstrap I	0.8	-16.0	3.6	-8.1	3.2	-3.8	1.1	-1.8	0.5	0.2	11.1	-1.0
Wtd. Bootstrap	1.2	-16.5	3.7	-8.3	3.3	-3.9	1.1	-1.9	0.3	0.2	11.2	-1.0

Results are summarized from 5,000 simple random samples. Bootstrap estimates involved 2,500 resamples.

Table 3

**Population C**  
**Empirical Results of Variance Estimators**

RMSE=Relative MSE PB=Percentage Bias	Sample Size											
	5		10		20		30		50		100	
Method	RMSE	PB	RMSE	PB	RMSE	PB	RMSE	PB	RMSE	PB	RMSE	PB
Cochran I	4.2	-3.4	0.1	-2.5	0.0	-2.1	0.0	-1.6	0.0	0.2	0.0	1.3
Cochran II	6.1	-3.2	1.1	-2.5	1.9	-2.1	1.4	-1.5	1.0	0.2	1.5	1.3
Fuller	4.1	-3.4	0.0	-2.5	0.2	-2.1	0.1	-1.6	0.3	0.2	1.5	1.3
Jackknife	6.9	-3.1	1.5	-2.4	2.2	-2.1	1.6	-1.5	1.1	0.2	1.6	1.3
Wtd. Jackknife	6.5	-3.1	1.2	-2.5	2.0	-2.1	1.5	-1.5	1.1	0.2	1.5	1.3
Bootstrap I	0.0	-13.1	1.4	-7.3	4.9	-4.5	5.0	-3.2	1.9	-0.7	3.1	0.8
Wtd. Bootstrap	0.1	-13.3	1.6	-7.4	5.1	-4.5	5.2	-3.2	1.9	-0.8	3.1	0.8

Results are summarized from 5,000 simple random samples. Bootstrap estimates involved 2,500 resamples.

Table 4

**Population Y00**  
**Empirical Results of Variance Estimators**

RMSE=Relative MSE PB=Percentage Bias	Sample Size					
	10		20		50	
Method	RMSE	PB	RMSE	PB	RMSE	PB
Cochran I	0.0	-3.7	0.0	-2.2	0.0	0.0
Cochran II	1.1	-3.5	2.2	-2.1	2.9	0.0
Fuller	0.0	-3.7	0.0	-2.2	0.0	0.0
Jackknife	1.6	-3.4	2.5	-2.1	3.1	0.0
Wtd. Jackknife	1.3	-3.5	2.3	-2.1	3.0	0.0
Bootstrap I	1.9	-8.2	4.0	-4.5	4.0	-0.8
Wtd. Bootstrap	2.3	-8.4	4.3	-4.6	4.0	-0.9

Results are summarized from 5,000 simple random samples.  
 Bootstrap estimates involved 2,500 resamples.

Table 5

**Population Y10**  
**Empirical Results of Variance Estimators**

RMSE=Relative MSE PB=Percentage Bias	Sample Size					
	10		20		50	
Method	RMSE	PB	RMSE	PB	RMSE	PB
Cochran I	2.2	-4.0	1.2	-2.2	0.5	0.0
Cochran II	0.0	-3.8	0.0	-2.1	0.0	0.0
Fuller	2.2	-4.0	1.1	-2.2	0.4	0.0
Jackknife	1.0	-3.6	0.5	-2.0	0.3	0.1
Wtd. Jackknife	0.5	-3.7	0.2	-2.1	0.1	0.0
Bootstrap I	1.0	-8.7	1.5	-4.6	0.6	-0.9
Wtd. Bootstrap	1.5	-8.8	1.8	-4.6	0.8	-0.9

Results are summarized from 5,000 simple random samples.  
 Bootstrap estimates involved 2,500 resamples.



Table 6

**Population Y20**  
**Empirical Results of Variance Estimators**

RMSE=Relative MSE PB=Percentage Bias	Sample Size					
	10		20		50	
Method	RMSE	PB	RMSE	PB	RMSE	PB
Cochran I	5.2	-4.3	4.2	-2.2	3.6	0.0
Cochran II	0.0	-4.3	0.0	-2.2	0.0	0.0
Fuller	5.1	-4.3	4.1	-2.2	3.2	0.0
Jackknife	1.4	-3.9	0.8	-2.0	0.5	0.1
Wtd. Jackknife	0.7	-4.1	0.4	-2.1	0.2	0.0
Bootstrap I	1.1	-9.2	1.1	-4.7	0.3	-1.0
Wtd. Bootstrap	1.6	-9.2	1.4	-4.7	0.4	-0.9

Results are summarized from 5,000 simple random samples.

Bootstrap estimates involved 2,500 resamples.

Table 7

**Population Y01**  
**Empirical Results of Variance Estimators**

RMSE=Relative MSE PB=Percentage Bias	Sample Size					
	10		20		50	
Method	RMSE	PB	RMSE	PB	RMSE	PB
Cochran I	0.0	-2.1	0.0	-0.8	0.0	1.7
Cochran II	11.6	-1.7	11.2	-0.6	11.5	1.8
Fuller	0.0	-2.1	0.0	-0.8	0.1	1.7
Jackknife	12.9	-1.5	11.5	-0.5	11.6	1.8
Wtd. Jackknife	12.2	-1.6	11.3	-0.6	11.5	1.8
Bootstrap I	11.4	-6.0	11.3	-2.8	9.7	0.9
Wtd. Bootstrap	10.2	-6.4	10.7	-2.9	9.0	0.8

Results are summarized from 5,000 simple random samples.  
 Bootstrap estimates involved 2,500 resamples.

Table 8

**Population Y11**  
**Empirical Results of Variance Estimators**

RMSE=Relative MSE PB=Percentage Bias	Sample Size					
	10		20		50	
Method	RMSE	PB	RMSE	PB	RMSE	PB
Cochran I	0.0	-2.3	0.0	-1.4	0.0	1.7
Cochran II	9.1	-1.9	8.9	-1.2	9.2	1.8
Fuller	0.0	-2.2	0.0	-1.3	0.1	1.7
Jackknife	10.9	-1.7	9.6	-1.1	9.6	1.8
Wtd. Jackknife	10.0	-1.8	9.2	-1.2	9.4	1.8
Bootstrap I	9.2	-6.3	9.7	-3.4	6.9	0.9
Wtd. Bootstrap	8.6	-6.6	9.4	-3.6	6.5	0.8

Results are summarized from 5,000 simple random samples.

Bootstrap estimates involved 2,500 resamples.

Table 9

**Population Y21**  
**Empirical Results of Variance Estimators**

RMSE=Relative MSE PB=Percentage Bias	Sample Size					
	10		20		50	
Method	RMSE	PB	RMSE	PB	RMSE	PB
Cochran I	3.9	-4.0	3.0	-2.4	3.0	0.4
Cochran II	0.0	-3.9	0.0	-2.4	0.4	0.5
Fuller	3.8	-4.0	2.9	-2.4	2.7	0.4
Jackknife	1.7	-3.5	0.9	-2.2	1.0	0.6
Wtd. Jackknife	0.8	-3.7	0.4	-2.3	0.7	0.5
Bootstrap I	0.9	-8.8	1.5	-4.9	0.0	-0.5
Wtd. Bootstrap	1.4	-8.8	1.8	-4.9	0.1	-0.5

Results are summarized from 5,000 simple random samples.  
 Bootstrap estimates involved 2,500 resamples.

# Populations 1-9 from B.C. Ministry of Forests

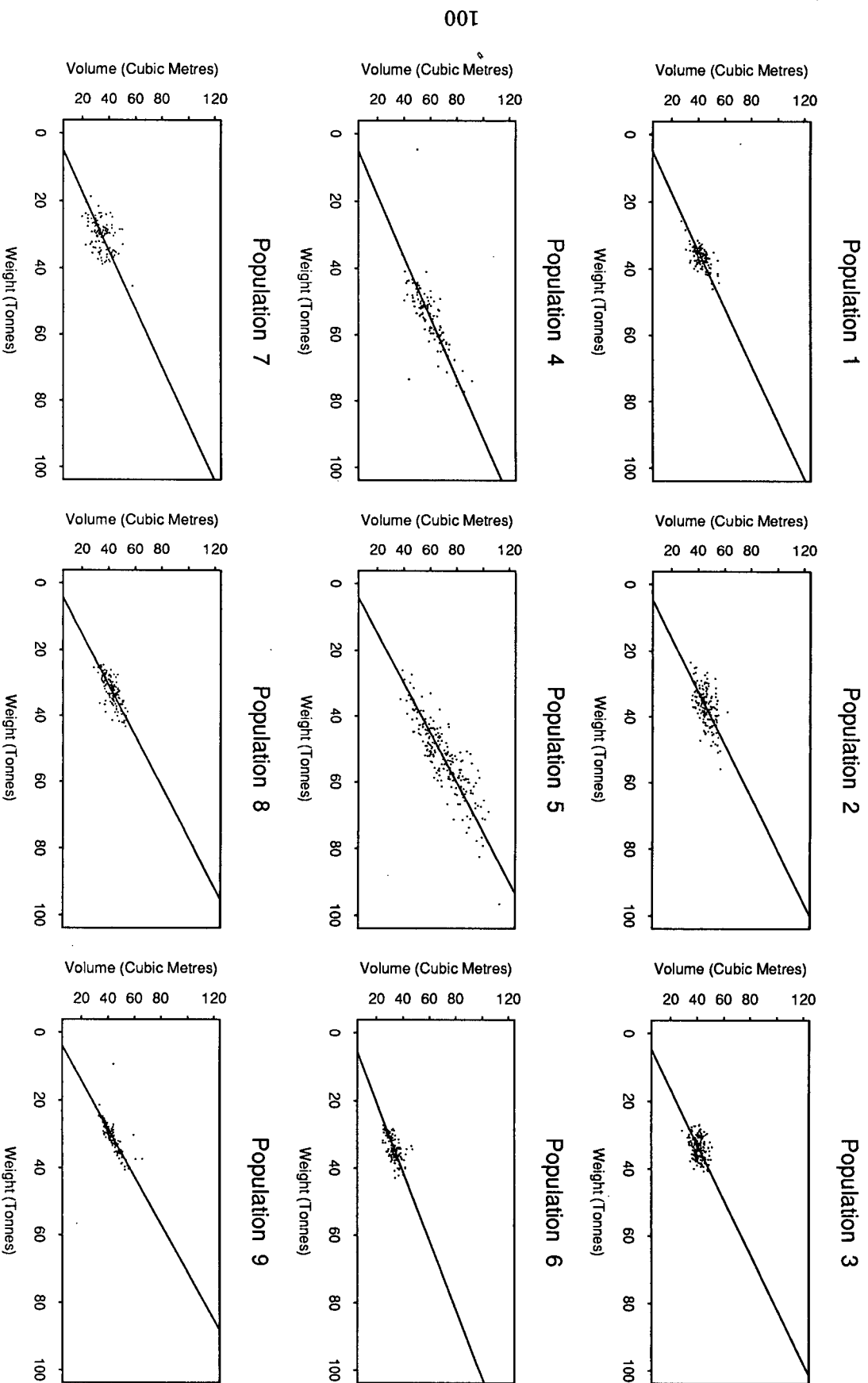


Figure 1a

Lines through Scatterplot are (Population Ratio \* Weight)

Populations 10-18 from B.C. Ministry of Forests

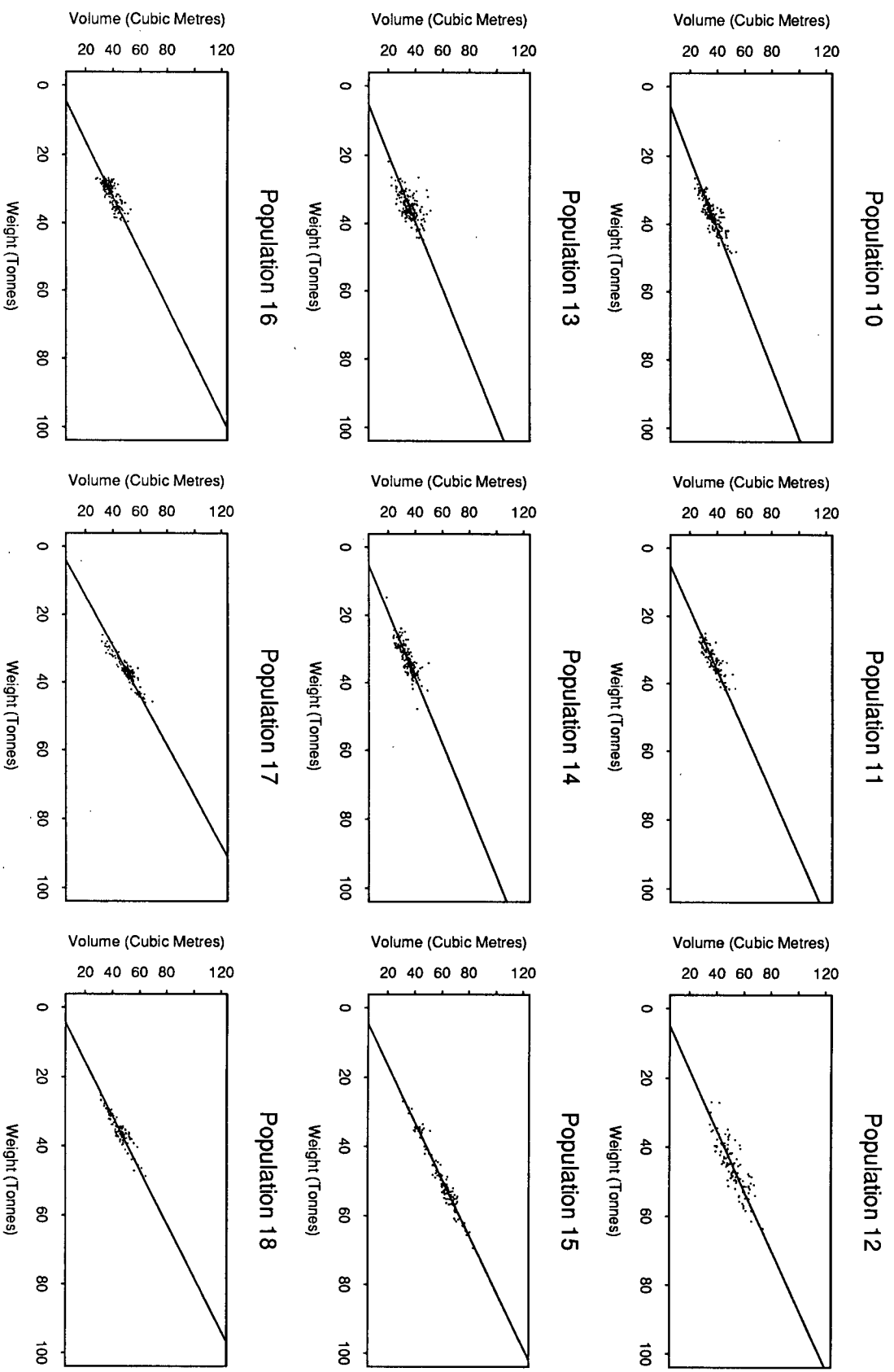
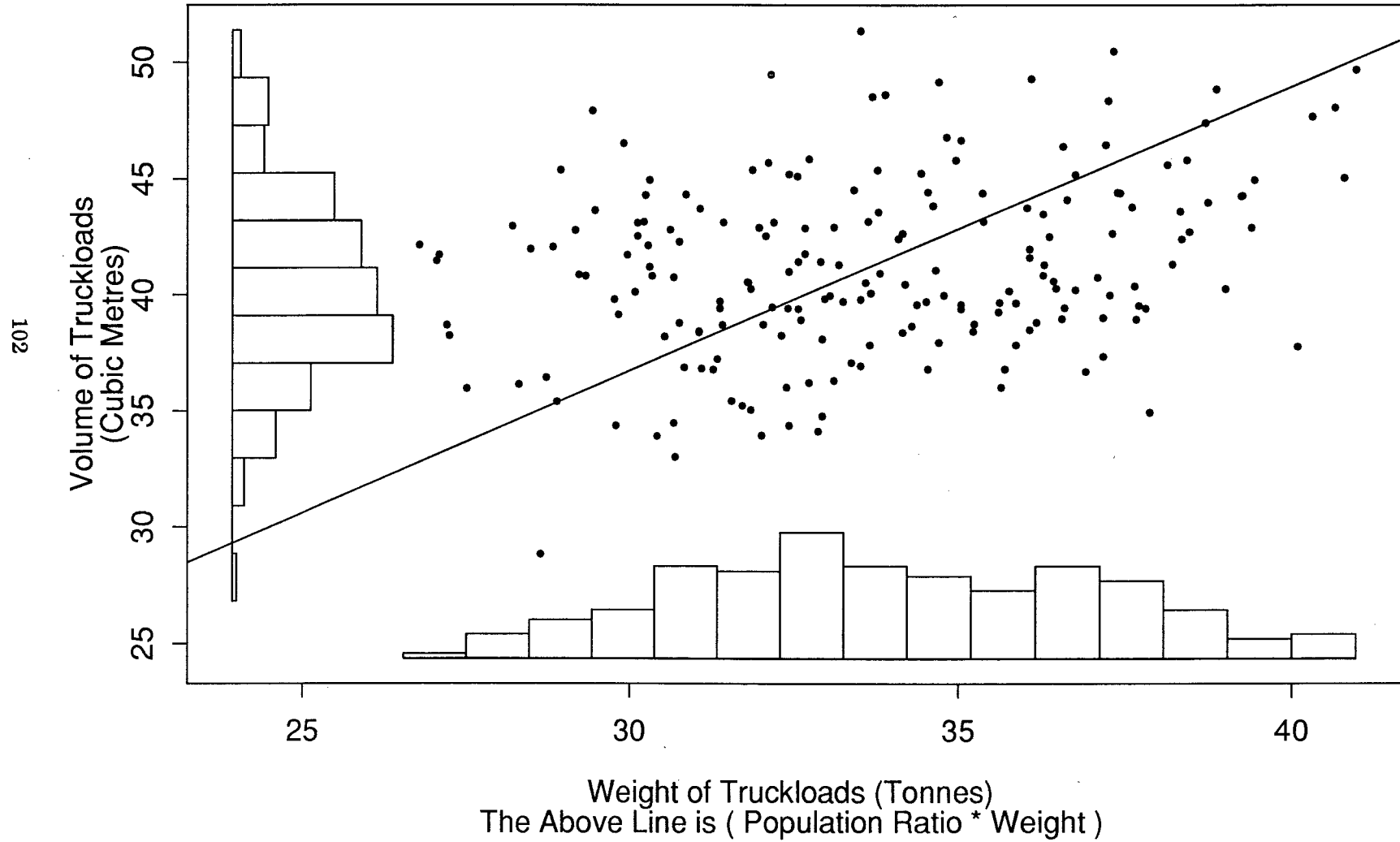


Figure 1b

Lines through Scatterplot are (Population Ratio \* Weight)

# Population A



## Population B



Figure 3



# Population C

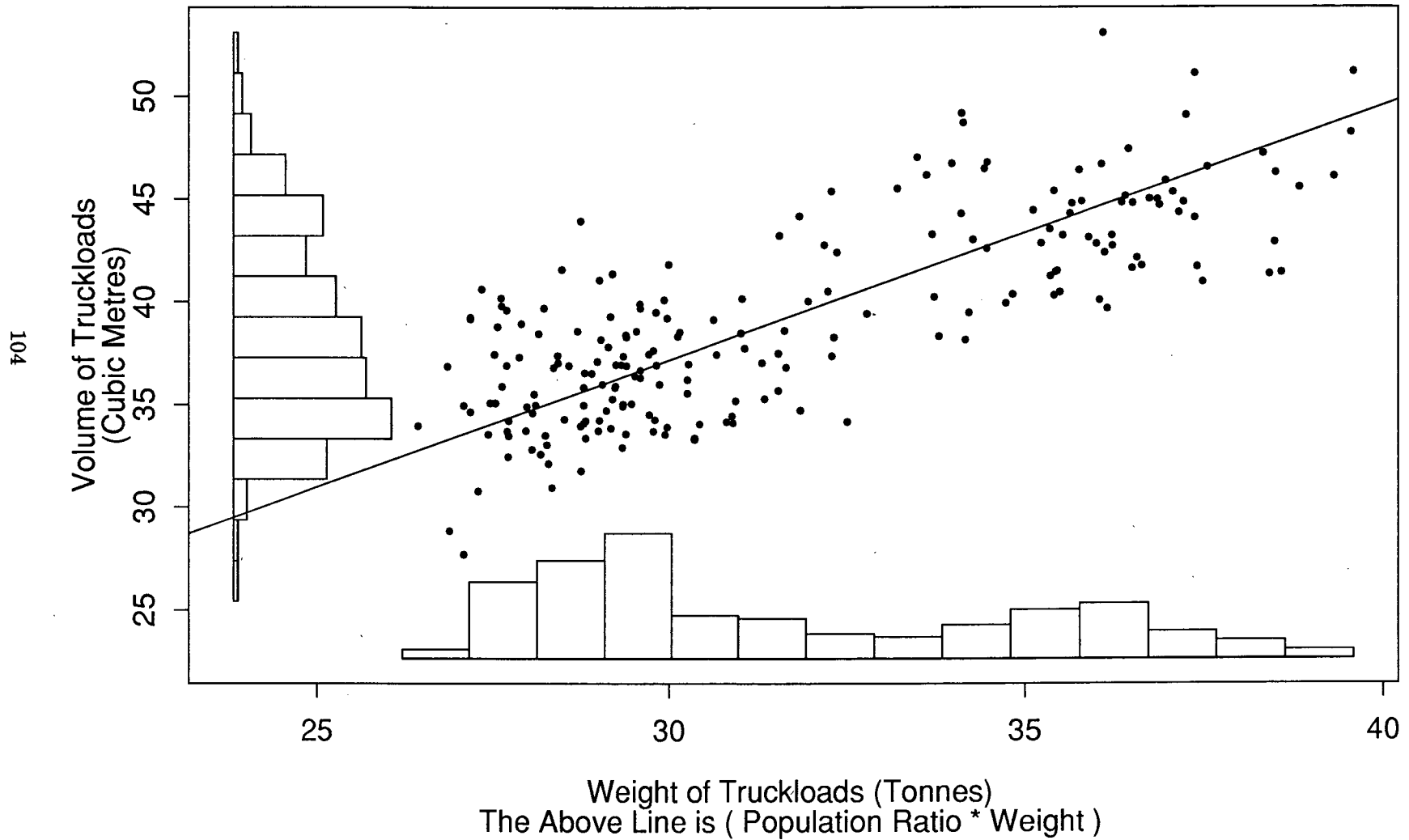


Figure 4