Variance Estimation for Sampling on Two Occasions

by

Heidi Kalbfleisch
BSc(Honours), Queen's University, 1995

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
in the Department
of
Mathematics and Statistics

SIMON FRASER UNIVERSITY
September 1997

APPROVAL

Name: Heidi Kalbfleisch

Degree: Master of Science

Title of thesis: Variance Estimation for Sampling on Two Occasions

Examining Committee: Dr. A. Lachlan
Chair

Dr. Randy Sitter
Senior Supervisor

Dr. C. Dean

Dr. T. Swartz

Dr. R. Routledge, Simon Fraser University, Burnaby
External Examiner

Date Approved: September 4, 1997

ABSTRACT

Sampling on two occasions is a valuable sampling scheme when observations are taken across two time points, and the correlation between these observations is high. Under this framework, estimates are based on a weighted average of a two-phase sampling estimator and an estimator from an independent random sample. There has been recent work in the literature on resampling methods such as the bootstrap and the jackknife in developing variance estimators under a two-phase sampling scheme. Resampling variance estimators have operational advantages over linearisation variance estimators. We extend these developments to sampling on two occasions. An attempt to use new linearisation variance estimators which make more complete use of the data available is made. Through simulation, a study of the unconditional properties of the different estimators is performed.

ACKNOWLEDGMENTS

I would like to thank Dr. Randy Sitter for his supervision and guidance in the creation and writing of this thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Sampling on two occasions is a valuable sampling scheme when observations are taken across two time points, and the correlation between these observations is high. Sen (1973) uses the methodology to estimate the kill of waterfowl per hunter in Ontario in 1968-1969. He noticed that the kill of waterfowl per hunter and the number of days hunted in the previous year, 1967-1968, were correlated with the desired estimate, kill of waterfowl in 1968-1969. Sampling on two occasions is also commonly used in the forest industry, often to estimate volume and growth. Before this sampling scheme was introduced in the forestry industry, *permanent* sample plots were frequently used. The term *permanent* indicates the same plots were sampled across all time points. It was eventually realized, however, when measurements over time are highly correlated, it is better to remeasure only a fraction of plots, and then with the remaining resources, establish an independent random sample of new plots. This is the general framework of sampling on two occasions, often termed sampling with partial replacement within the forest industry. Typically one characteristic is being observed, and the *successive occasions* are over time.

The general procedure involves taking an initial sample s_1 at time one of size n_1 and observing some characteristic denoted $x_1, \ldots, x_{n_1}$. It should be noted that these may be vector-valued measurements, though we focus on the scalar case. At time 2, a sub-sample $s_2 \subset s_1$ of size n_2 is taken and the same characteristic is again measured and denoted

$y_1, \ldots, y_{n_2}$. Also at time two, an independent random sample s_3 of size n_3 is taken from the entire population and the same characteristic is again observed to obtain $y_{n_2+1}, \ldots, y_{n_2+n_3}$. Note that the units in s_2 are matched with the same units in s_1 while units in s_3 are unmatched, and x and y denote the value of the measured characteristic at times 1 and 2 respectively. That is, s_1 is the set of units measured only at time 1, s_2 is the set of units measured at both time 1 and time 2, and s_3 is the set of units measured only at time 2. Thus, s_1 and s_2 form a two-phase sample over times 1 and 2, while s_3 is an independent sample from the population at time 2. Of interest is to estimate $\bar{Y}$, the population mean of the characteristic at time 2. There are two natural estimates available:

$$\bar{y}_{lr} = \bar{y}_2 + b(\bar{x}_1 - \bar{x}_2) \qquad \text{and} \qquad \bar{y}_3 = \textstyle\sum_{i \epsilon s_3} y_i / n_3,$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of the x_i on the first phase sample s_1 and second phase sample s_2, $\bar{y}_2$ and $\bar{y}_3$ are the means of the y units in s_2 and independent random sample s_3, and the regression coefficient b is calculated based on observations in s_2. Since these estimates give independent information, typically a weighted average of the estimates is used,

$$\bar{y}_w^{lr} = w_1 \bar{y}_{lr} + w_2 \bar{y}_3 \tag{1.1}$$

where $w_1 + w_2 = 1$. One could alternatively use the ratio estimator, for example, in replace of the regression estimator in the above formula as follows:

$$\bar{y}_w^{r} = w_1 \bar{y}_r + w_2 \bar{y}_3 \tag{1.2}$$

where $\bar{y}_r = (\bar{y}_2/\bar{x}_2)\bar{x}_1$. For simplicity of notation we will often restrict our comment to $\bar{y}_w^{lr}$. However, a similar development is available for $\bar{y}_w^{r}$ in each case.

Cochran (1977, p.346) obtains the optimal estimate of $\bar{Y}$ by weighting the two independent estimates inversely as their variances. That is, use

$$w_1 = \frac{V(\bar{y}_{lr})^{-1}}{V(\bar{y}_{lr})^{-1} + V(\bar{y}_3)^{-1}} \qquad \text{and} \qquad w_2 = \frac{V(\bar{y}_3)^{-1}}{V(\bar{y}_{lr})^{-1} + V(\bar{y}_3)^{-1}}$$

in (1.1). We could then use the estimated weights

$$\hat{w}_1 = \frac{v(\bar{y}_{lr})^{-1}}{v(\bar{y}_{lr})^{-1} + v(\bar{y}_3)^{-1}} \qquad \text{and} \qquad \hat{w}_2 = \frac{v(\bar{y}_3)^{-1}}{v(\bar{y}_{lr})^{-1} + v(\bar{y}_3)^{-1}}, \tag{1.3}$$

where $v(\bar{y}_{lr})$ and $v(\bar{y}_3)$ are consistent estimates of $V(\bar{y}_{lr})$ and $V(\bar{y}_3)$. This is the approach which is often used in the forest industry (see Schreuder, 1987; Ware and Cunia, 1962).

In large surveys with many measured characteristics, fixed weights may be operationally more desirable to avoid different weighting for different characteristics. Cochran (1977, p. 351) suggests using fixed weights when sampling on more than two occasions to avoid updating the weights w_1 and w_2. These fixed weights may be simple fractions based on sample sizes.

The main purpose of this thesis is to consider variance estimation in the context of sampling on two occasions. We will consider separately the two cases of fixed weights and weights estimated as in (1.3). In the fixed weight case the variance is

$$V(\bar{y}_w^{lr}) = w_1^2 V(\bar{y}_{lr}) + w_2^2 V(\bar{y}_3)$$

and the question becomes which estimators of $V(\bar{y}_{lr})$ and $V(\bar{y}_3)$ should be used. Rao and Sitter (1995) and Sitter (1997) derive new linearisation variance estimators and jackknife variance estimators for both the ratio and the regression estimator in the context of two-phase sampling. Resampling variance estimators have operational advantages over linearisation variances estimators. There is evidence that they may also have better conditional properties. We consider these as well as the usual estimate of variance for $\bar{y}_{lr}$ as given in Sukhatme & Sukhatme (1970, p.212) to estimate $V(\bar{y}_{lr})$. The usual estimate of variance for a simple mean, refer to Cochran (1977, p.23), based on s_3 is typically used to estimate $V(\bar{y}_3)$. We consider an alternative choice.

In the case where weights are estimated as in (1.3), it can be easily shown that treating the weights as fixed will yield consistent variance estimates. Scott (1984), Ware and Cunia (1962), and Bickford (1963) review sampling with partial replacement when estimating volume and growth in the forestry setting, and all suggest using an estimator suggested by Meier (1953) in a paper on variance estimation of weighted means. Both of these variance estimators can be written as a function of $\hat{w}_1$ given in (1.3) and thus their performance depends on the choice of $v(\bar{y}_{lr})$ and $v(\bar{y}_3)$. Schreuder (1987) considers various estimates $v(\bar{y}_{lr})$ including a grouped jackknife and a bootstrap, and uses these in Meier's variance estimator. He concludes that though the jackknife variance estimator or the bootstrap

variance estimator may be preferred for skewed populations and small sample sizes, Meier's original estimator is generally preferable to the jackknife and the bootstrap when considering bias and efficiency.

We consider various estimators for $V(\bar{y}_{lr})$ and $V(\bar{y}_3)$, and compare the one-term linearisation variance estimator to Meier's variance estimator. An attempt to use new linearisation estimators which make more complete use of the data available will be made. Through simulation, a study of the unconditional properties of the different estimators will be performed.

Chapter 2

Two-phase sampling

In this chapter we review results in two-phase sampling. This will facilitate both the notational and conceptual understanding of sampling on two occasions. Rao and Sitter (1995) and Sitter (1997) develop variance estimators for the ratio and regression estimators for two-phase sampling, also see Dorfman (1994). Simulation studies showed the newly developed variance estimators to have better conditional and unconditional properties than the simple linearisation variance estimator given in Sukhatme & Sukhatme (1970, p.212).

2.1 Two-phase sampling: a review

Under a single-phase sampling framework, the usual ratio and regression estimators require knowledge of the population parameter $\bar{X}$. This information may not be available, and thus two-phase sampling may be employed. In this type of sampling, we first take a large first-phase sample s_1 of size n_1 and observe some auxiliary variable x, which is cheaper or easier to obtain. Note that the variable of interest y is not measured in this preliminary sample. We can now calculate a simple arithmetic mean and get a good estimate, $\bar{x}_1$, of $\bar{X}$, the population mean. Next a subsample s_2 of size n_2 is taken without replacement from s_1 and both x and the variable of interest y are measured. With the information observed in both phases of the sampling procedure, we can now estimate $\bar{Y}$, the population mean for the y variables, using either a ratio or a regression estimator. The ratio estimator for $\bar{Y}$ is

given by $\bar{y}_r = (\bar{y}_2/\bar{x}_2)\bar{x}_1$ where $\bar{x}_1 = \sum_{i\epsilon s_1} x_i/n_1$, $\bar{y}_2 = \sum_{i\epsilon s_2} y_i/n_2$ and $\bar{x}_2 = \sum_{i\epsilon s_2} x_i/n_2$. The regression estimator is given by $\bar{y}_{lr} = \bar{y}_2 + b(\bar{x}_1 - \bar{x}_2)$ where b is the least squares regression coefficient of y_i on x_i computed from the second-phase sample. Typically this type of sampling is used when the cost efficiency is improved by taking this large preliminary sample of a correlated variable x, as opposed to simply taking a larger sample of variate y.

2.2 Variance estimators for the ratio estimator

The standard formula for a design-consistent linearisation variance estimator as given in Sukhatme & Sukhatme (1970, p.170) for the two-phase sampling ratio estimator, $\bar{y}_r = (\bar{y}_2/\bar{x}_2)\bar{x}_1$, is

$$v_0(\bar{y}_r) = \left(\frac{1}{n_2} - \frac{1}{n_1}\right) s_d^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right) s_y^2 \tag{2.1}$$

where $s_d^2 = \sum_{i\epsilon s_2} d_i^2/(n_2 - 1)$, $s_y^2 = \sum_{i\epsilon s_2}(y_i - \bar{y}_2)^2/(n_2 - 1)$, $d_i = y_i - \hat{R}x_i$ and $\hat{R} = \bar{y}/\bar{x}$. In this formulation, s_y^2 is used to estimate $S_y^2 = \sum_{i\epsilon S}(y_i - \bar{Y})^2/(N-1)$. Rao and Sitter (1995) exploit the identity

$$S_y^2 = S_D^2 + 2RS_{Dx} + R^2 S_x^2, \tag{2.2}$$

where S_{Dx} is the population covariance of d_i and x_i and $R = \bar{Y}/\bar{X}$, and propose a new linearisation variance estimator which makes fuller use of the data than the formula given in equation (2.1). To do this, they note that using s_y^2 to estimate S_y^2 is equivalent to estimating (2.2) term-by-term using s_d^2, $s_{dx} = \sum_{i\epsilon s_2} d_i(x_i - \bar{x}_2)/(n_2 - 1)$, $s_{x_2}^2 = \sum_{i\epsilon s_2}(x_i - \bar{x}_2)^2/(n_2 - 1)$ and $\hat{R}$ as estimates of S_d^2, S_{Dx}, S_x^2 and R. That is (2.1) can be rewritten as

$$v_0(\bar{y}_r) = \left(\frac{1}{n_2} - \frac{1}{N}\right) s_d^2 + 2\left(\frac{1}{n_1} - \frac{1}{N}\right) \hat{R}s_{dx} + \left(\frac{1}{n_1} - \frac{1}{N}\right) \hat{R}^2 s_{x_2}^2. \tag{2.3}$$

If instead $s_{x_1}^2 = \sum_{i\epsilon s_1}(x_i - \bar{x}_1)^2/(n_1 - 1)$ is used to estimate S_x^2, all x measurements observed in the sampling procedure are used. Noting the identity given above and using (2.1) we find

$$\begin{aligned} v_1(\bar{y}_r) &= \left(\frac{1}{n_2} - \frac{1}{n_1}\right) s_d^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right)(s_d^2 + 2\hat{R}s_{dx} + \hat{R}^2 s_{x_1}^2) \\ &= \left(\frac{1}{n_2} - \frac{1}{N}\right) s_d^2 + 2\left(\frac{1}{n_1} - \frac{1}{N}\right) \hat{R}s_{dx} + \left(\frac{1}{n_1} - \frac{1}{N}\right) \hat{R}^2 s_{x_1}^2 \end{aligned} \tag{2.4}$$

which is the proposed linearisation estimator given by Rao and Sitter (1995).

A jackknife variance estimator was also developed by Rao and Sitter (1995). In their delete-one-unit approach, $\bar{y}_r$ is recalculated with the j^{th} element removed for each $j \epsilon s_1$ giving the jackknife estimate $\bar{y}_r(j)$. Note that $\bar{y}_r(j) = [\bar{y}_2(j)/\bar{x}_2(j)]\bar{x}_1(j)$ where $\bar{y}_2(j)$, $\bar{x}_2(j)$ and $\bar{x}_1(j)$ are simple means calculated with the j^{th} element removed. In order to calculate these means, however, we must know if the j^{th} element is in both the first phase and second phase sample, or the first phase sample exclusively. Thus,

$$\bar{x}_2(j) = \begin{cases} \frac{n\bar{x}_2 - x_j}{n_2 - 1} & j \epsilon s_1 \\ \bar{x}_2 & j \epsilon s_1 - s_2, \end{cases}$$

$$\bar{y}_2(j) = \begin{cases} \frac{n\bar{y}_2 - y_j}{n_2 - 1} & j \epsilon s_1 \\ \bar{y}_2 & j \epsilon s_1 - s_2, \end{cases}$$

and $\bar{x}_1(j) = (n_1\bar{x}_1 - x_j)/(n_1 - 1)$ for all $j \epsilon s_1$. By applying the usual jackknife formula

$$v_J(\bar{y}_r) = \frac{n_1 - 1}{n_1} \sum_{j \epsilon s_1} (\bar{y}_r(j) - \bar{y}_r)^2 \tag{2.5}$$

the variance of the n_1 jackknife estimates is calculated. Note that the above formula ignores finite population corrections.

Rao and Sitter (1995) also give a linearised version of this jackknife variance estimator which they derive by approximating (2.5) for large n_2. The linearised jackknife variance estimator was found to be

$$v_{LJ}(\bar{y}_r) \doteq \left(\frac{\bar{x}_1}{\bar{x}_2}\right)^2 \frac{s_d^2}{n_2} + 2\left(\frac{\bar{x}_1}{\bar{x}_2}\right) \hat{R} \frac{s_{dx}}{n_1} + \hat{R}^2 \frac{s_{x_1}^2}{n_1}, \tag{2.6}$$

again ignoring finite population corrections. It is interesting to note that this estimator uses $s_{x_1}^2$ as in Rao and Sitter's new linearisation variance estimator given in (2.4). They also combine the above estimator with the appropriate finite population corrections to give a linearised jackknife variance estimator

$$v_2(\bar{y}_r) = \left(\frac{\bar{x}_1}{\bar{x}_2}\right)^2 \left(\frac{1}{n_2} - \frac{1}{N}\right) s_d^2 + 2\left(\frac{\bar{x}_1}{\bar{x}_2}\right)\left(\frac{1}{n_1} - \frac{1}{N}\right) \hat{R} s_{dx} + \left(\frac{1}{n_1} - \frac{1}{N}\right) \hat{R}^2 s_{x_1}^2. \tag{2.7}$$

A simulation study was done to investigate the performance of the proposed estimators relative to the standard linearisation estimator in (2.1). A finite population was created using various simple models. The conditional and unconditional properties of the new estimators were studied varying ρ, the correlation between the auxiliary variable x and the variable of interest y, and $C_x = \sigma_x/\mu_x$, the coefficient of variation of the x's. Rao and Sitter found that both v_1 and v_2 are more efficient than v_0 for $\rho \geq 0.8$ and large C_x. Further, conditional on $\bar{x}_1/\bar{x}_2$, v_2 and v_J both performed better in tracking the conditional MSE than v_0 and v_1. They argue that conditioning on $\bar{x}_1/\bar{x}_2$ is defensible since $\bar{x}_1$ is based on a large sample and thus is close to $\bar{X}$, which makes $\bar{x}_1/\bar{x}_2$ approximately ancillary.

2.3 Variance estimators for the regression estimator

Sitter (1997) extends the work discussed above to the regression estimator. This estimator is also commonly used in estimating the population mean, $\bar{Y}$, under a two-phase sampling framework. Sitter develops both a new linearisation variance estimator and a jackknife variance estimator and investigates their relative efficiency and conditional properties through a simulation study.

The simple linear regression estimator for two-phase sampling is

$$\bar{y}_{lr} = \bar{y}_2 + b(\bar{x}_1 - \bar{x}_2) \tag{2.8}$$

where $\bar{y}_2, \bar{x}_2$ and $\bar{x}_1$ are as defined previously, and $b = s_{xy_2}/s_{x_2}^2$ where $s_{xy_2} = \sum_{i \epsilon s_2}(x_i - \bar{x}_2)(y_i - \bar{y}_2)/(n_2 - 1)$ and $s_{x_2}^2 = \sum_{i \epsilon s_2}(x_i - \bar{x}_2)^2/(n_2 - 1)$. Cochran (1977, p.343) gives the standard linearisation variance estimators of $\bar{y}_{lr}$,

$$v_0(\bar{y}_{lr}) = \left(\frac{1}{n_2} - \frac{1}{n_1}\right) s_d^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right) s_{y_2}^2, \tag{2.9}$$

where $d_i = y_i - \bar{y}_2 - b(x_i - \bar{x}_2)$, $s_d^2 = \sum_{i \epsilon s_2} d_i^2/(n_2 - 1)$ and $s_{y_2}^2 = \sum_{i \epsilon s_2}(y_i - \bar{y}_2)^2/(n_2 - 1)$. Recall in developing the new linearisation variance estimator for the ratio estimator, Rao and Sitter (1995) used the identity given in (2.2). Sitter (1997) notes a similar relationship

$$S_y^2 = S_D^2 + B^2 S_x^2, \tag{2.10}$$

where $B = S_{xy}/S_x^2$, S_D^2 and S_x^2 are the population variances of d_i and x_i. As explained in the previous section, we have two possible estimates of S_x^2, $s_{x_2}^2$ or $s_{x_1}^2$. The latter makes more complete use of the data, and thus Sitter suggests using

$$s_{y_2}^2 = s_d^2 + b^2 s_{x_1}^2$$

in (2.9). Thus Sitter's proposed linearisation variance estimator is

$$\begin{aligned} v_1(\bar{y}_{lr}) &= \left(\frac{1}{n_2} - \frac{1}{n_1}\right) s_d^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right)(s_d^2 + b^2 s_{x_1}^2) \\ &= \left(\frac{1}{n_2} - \frac{1}{N}\right) s_d^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right) b^2 s_{x_1}^2, \end{aligned} \tag{2.11}$$

refer also to Dorfman (1994). Sitter then goes on to develop the jackknife variance estimator, again using a delete-one-unit approach. We calculate $\bar{y}_{lr}(j) = \bar{y}_2(j) + b(j)(\bar{x}_1(j) - \bar{x}_2(j))$ for each $j \epsilon s_1$, where $\bar{x}_1(j)$, $\bar{x}_2(j)$, and $\bar{y}_2(j)$ are as defined in Section 2.2 and

$$b(j) = \begin{cases} b - \frac{(x_j - \bar{x}_2)d_j}{(n_2-1)s_{x_2}^2(1-k_j)} & j \epsilon s_1 \\ b & j \epsilon s_1 - s_2, \end{cases}$$

where $k_j = 1/n_2 + (x_j - \bar{x}_2)^2/\{(n_2-1)s_{x_2}^2\}$. We can then apply the general jackknife formula

$$v_J(\bar{y}_{lr}) = \frac{n_1 - 1}{n_1} \sum_{j \epsilon s_1} (\bar{y}_{lr}(j) - \bar{y}_{lr})^2. \tag{2.12}$$

Sitter also finds the linearised version of his jackknife variance estimator. He notes for large n_2,

$$\bar{y}_{lr}(j) - \bar{y}_{lr} \doteq \begin{cases} -b\left(\frac{x_j - \bar{x}_1}{n_1 - 1}\right) - \frac{d_j}{n_2}\left(1 + \frac{a_j}{(1-k_j)}\right) & j \epsilon s_2 \\ -b\left(\frac{x_j - \bar{x}_1}{n_1 - 1}\right) & j \epsilon s_1 - s_2, \end{cases}$$

where $a_j = \{n_2(x_j - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)\}/\{(n_2 - 1)s_{x_2}^2\}$. The linearised version of $v_J(\bar{y}_{lr})$ with finite population corrections is then

$$v_{LJ}(\bar{y}_{lr}) = \left(\frac{1}{n_2} - \frac{1}{N}\right) s_d^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right) b^2 s_{x_1}^2 + \left\{\frac{\bar{x}_1 - \bar{x}_2}{(n_2 - 1)s_{x_2}^2}\right\}^2 \sum_{j \epsilon s_2} \frac{d_j^2 (x_j - \bar{x}_2)^2}{(1 - k_j)^2} + R \tag{2.13}$$

where

$$R = \frac{2}{n_2}\left\{\frac{1}{n_2} \sum_{j \epsilon s_2} \frac{d_j^2 a_j}{(1 - k_j)} + \frac{b}{n_1 - 1} \sum_{j \epsilon s_2} \frac{d_j a_j (x_j - \bar{x}_1)}{(1 - k_j)}\right\}. \tag{2.14}$$

A simulation study was used to examine the conditional and unconditional properties of the new variance estimators versus the simple linearisation estimator given in (2.9). As in the previous simulation study discussed, a finite population was created from simple models, and values of ρ, the correlation of x and y, and C_x, the coefficient of variation of the x's, were varied. Results from the simulation study were similar to those found for the ratio simulation study. That is, $v_1(\bar{y}_{lr})$ and the linearised jackknife variance estimator were found to be considerably more efficient than $v_0(\bar{y}_{lr})$ for $\rho \geq 0.8$ and large C_x. Unconditionally, v_1 had the smallest MSE. Conditional properties were studied by conditioning on the size of $\bar{x}_1 - \bar{x}_2$. Sitter found that the jackknife variance estimator and its linearised version performed better in tracking the conditional MSE than v_0 or v_1 when $\bar{x}_1 - \bar{x}_2$ was small or large.

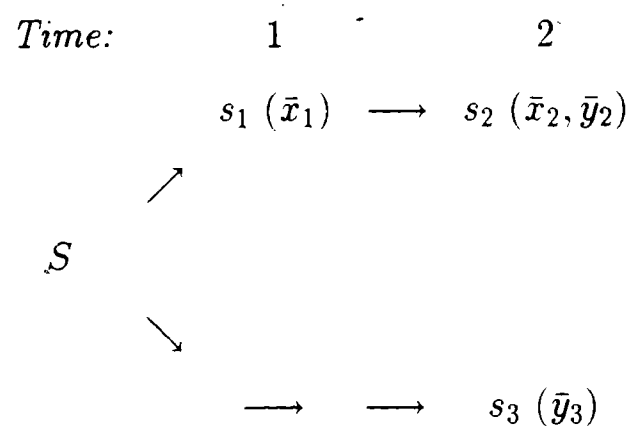
Chapter 3

Sampling on two occasions

Sampling on two occasions, also called successive sampling or sampling with partial replacement, is considered to be an efficient sampling scheme particularly by those in the forest industry. In the past, the forest industry sampled the same plots, termed *permanent* sample plots, to evaluate volume and growth. It was eventually realized, however, there is a gain in precision when the measurements taken over time are highly correlated if only a portion of permanent plots are remeasured and an additional random sample from the population is taken at the second sampling occasion. Scott (1984) explains that based on his own experience in the northeastern United States, the correlation between measurements is strong enough to warrant the use of sampling with partial replacement even with 25 years between surveys.

Since the estimator we will be primarily discussing is a linear combination of a regression (or ratio) estimator from a two-phase sample and a mean from an independent simple random sample, we expect that by extending the work that was discussed in Chapter 2, we can develop more efficient variance estimators than those presently in use. Typically the variance estimator used in the forestry literature under sampling with partial replacement is based on that suggested by Meier (1953). Schreuder (1987) has done some work in variance estimation when sampling with partial replacement, looking in particular at resampling methods to get improved variance estimators.

Figure 3.1: Means generated from sampling on two occasions



3.1 Sampling on two occasions: the procedure and notation

Typically in this sampling scheme the same characteristic is being measured at each occasion. Note that more than one characteristic could potentially be observed at each time point giving rise to vector-valued observations, though we focus on the scalar case. The successive sampling occasions are typically over time. Consider, for example, measurements being taken at two time points. We will use x for first occasion (or first time point) measurements, and y for second occasion measurements. The procedure is then as follows. At time one, a sample s_1 of size n_1 is taken without replacement and $x_1, \ldots, x_{n_1}$ are measured. At the second occasion, a subsample s_2 of size n_2 from the n_1 units is taken without replacement and $x_1, \ldots, x_{n_2}$ are noted and $y_1, \ldots, y_{n_2}$ are measured for the n_2 matched units. Also at the second occasion an independent random sample s_3 of size n_3 is taken without replacement from the entire population generating the sample $y_1, \ldots, y_{n_3}$. Figure 3.1 gives a pictorial representation of the sampling scheme. S represents the finite population of units. The means in parentheses are those obtained from the indicated sample, s_1, s_2 or s_3: $\bar{x}_1 = \sum_{i \epsilon s_1} x_i/n_1$, $\bar{x}_2 = \sum_{i \epsilon s_2} x_i/n_2$, $\bar{y}_2 = \sum_{i \epsilon s_2} y_i/n_2$ and $\bar{y}_3 = \sum_{i \epsilon s_3} y_i/n_3$.

We are interested in estimating $\bar{Y}$, the population mean at time 2. Note from the above picture we essentially have a two-phase sample, the top arm of the picture, and an independent random sample at the second time point, the bottom arm. We can find two

independent estimates of $\bar{Y}$. From the two-phase sample we could use the ratio estimator, $\bar{y}_r = (\bar{y}_2/\bar{x}_2)\bar{x}_1$, or the regression estimator,

$$\bar{y}_{lr} = \bar{y}_2 + b(\bar{x}_1 - \bar{x}_2), \tag{3.1}$$

where $b = s_{xy_2}/s_{x_2}^2$ with $s_{xy_2} = \sum_{i\epsilon s_2}(x_i - \bar{x}_2)(y_i - \bar{y}_2)/(n_2 - 1)$ and $s_{x_2}^2 = \sum_{i\epsilon s_2}(x_i - \bar{x}_2)^2/(n_2 - 1)$. From the independent random sample we have the mean $\bar{y}_3$ as defined above. To estimate $\bar{Y}$ consider a weighted average of the latter two estimates:

$$\bar{y}_w^{lr} = w_1\bar{y}_{lr} + w_2\bar{y}_3, \tag{3.2}$$

with $w_1 + w_2 = 1$. There are essentially two cases to consider here: (i) we could consider these weights to be fixed constants based, for instance, on the relative sample sizes of s_1, s_2 and s_3; or (ii) we could consider the optimal weights

$$w_1 = \frac{V(\bar{y}_{lr})^{-1}}{V(\bar{y}_{lr})^{-1} + V(\bar{y}_3)^{-1}} \quad \text{and} \quad w_2 = \frac{V(\bar{y}_3)^{-1}}{V(\bar{y}_{lr})^{-1} + V(\bar{y}_3)^{-1}}, \tag{3.3}$$

given in Cochran (p.346, 1977) and estimate them to get $\hat{w}_1$ and $\hat{w}_2$ by replacing $V(\bar{y}_{lr})$ and $V(\bar{y}_3)$ by sample estimates $v(\bar{y}_{lr})$ and $v(\bar{y}_3)$. That is, we could use

$$\hat{w}_1 = \frac{v(\bar{y}_{lr})^{-1}}{v(\bar{y}_{lr})^{-1} + v(\bar{y}_3)^{-1}} \quad \text{and} \quad \hat{w}_2 = \frac{v(\bar{y}_3)^{-1}}{v(\bar{y}_{lr})^{-1} + v(\bar{y}_3)^{-1}}. \tag{3.4}$$

Meier (1953) explains that although $\hat{w}_1$ and $\hat{w}_2$ are not the maximum likelihood weights, they provide an asymptotically efficient estimate of $\bar{Y}$ when both n_2 and n_3 are large, the estimates of $\bar{Y}$ are independently and normally distributed with mean $\bar{Y}$ and uncommon variances, and sample variance estimates are unbiased.

Note: One could alternatively use $\bar{y}_r$ in place of $\bar{y}_{lr}$ in (3.2) to give

$$\bar{y}_w^r = w_1\bar{y}_r + w_2\bar{y}_3. \tag{3.5}$$

Then use the weights specified above in (3.3) replacing $V(\bar{y}_{lr})$ by $V(\bar{y}_r)$.

3.2 Previous work in variance estimation

Consider the estimate given in (3.2). Since the weighted estimates in this expression are independent, in the case of fixed, or constant, weights we can write

$$V(\bar{y}_w^{lr}) = w_1^2 V(\bar{y}_{lr}) + w_2^2 V(\bar{y}_3). \tag{3.6}$$

This can be estimated by $v(\bar{y}_w^{lr}) = w_1^2 v(\bar{y}_{lr}) + w_2^2 v(\bar{y}_3)$, where $v(\bar{y}_{lr})$ and $v(\bar{y}_3)$ are sample variance estimates. Typically practitioners use $v_0(\bar{y}_{lr})$ given in equation (2.9) for the first term (or $v_0(\bar{y}_r)$ from equation (2.1) if $\bar{y}_r$ is being used), and $v_3(\bar{y}_3) = (1/n_3 - 1/N)s_{y_3}^2$ for the second term.

If we use estimated weights as in (3.4), but treat them as if they were fixed as in (3.6), we get

$$V(\bar{y}_w^{lr}) = \left(\frac{v(\bar{y}_{lr})^{-1}}{v(\bar{y}_{lr})^{-1} + v(\bar{y}_3)^{-1}}\right)^2 V(\bar{y}_{lr}) + \left(\frac{v(\bar{y}_3)^{-1}}{v(\bar{y}_{lr})^{-1} + v(\bar{y}_3)^{-1}}\right)^2 V(\bar{y}_3) \tag{3.7}$$

and its sample estimate

$$\begin{aligned} v(\bar{y}_w^{lr}) &= \left(\frac{v(\bar{y}_{lr})^{-1}}{v(\bar{y}_{lr})^{-1} + v(\bar{y}_3)^{-1}}\right)^2 v(\bar{y}_{lr}) + \left(\frac{v(\bar{y}_3)^{-1}}{v(\bar{y}_{lr})^{-1} + v(\bar{y}_3)^{-1}}\right)^2 v(\bar{y}_3) \\ &= \frac{1}{v(\bar{y}_{lr})^{-1} + v(\bar{y}_3)^{-1}}. \end{aligned} \tag{3.8}$$

One can show by taking a one-term Taylor Series expansion (see Appendix A) that even though (3.8) is obtained by treating the weights as fixed, it is in fact a consistent estimator of $V(\bar{y}_w^{lr})$ for estimated weights.

Meier (1953) and Cochran and Carroll (1953) discuss variance estimation of a weighted mean under weighting inversely as the estimated variance, such as $\hat{w}_1$ and $\hat{w}_2$ given above. The variance formula developed by Meier is commonly used in the forestry literature. Meier (1953) uses a second order Taylor series expansion in an attempt to get a better variance estimator. It should be noted Meier's estimator is not a standard linearisation variance estimator due to the assumptions made in its development. Applying Meier's technique to the estimator in (3.2), the following assumptions are made. The estimates $\bar{y}_{lr}$ and $\bar{y}_3$ are independent and normally distributed with mean $\bar{Y}$, and variances $\sigma_1^2 = V(\bar{y}_{lr})$ and

$\sigma_2^2 = V(\bar{y}_3)$. The estimates $s_1^2 = v(\bar{y}_{lr})$ and $s_2^2 = v(\bar{y}_3)$ are unbiased and independent of each other and of $\bar{y}_{lr}$ and $\bar{y}_3$. Moreover, Meier assumes $(n_1 - 1)s_1^2/\sigma_1^2 \sim \chi^2_{(n_1-1)}$ and $(n_3 - 1)s_2^2/\sigma_2^2 \sim \chi^2_{(n_3-1)}$. Using these assumptions and a second order Taylor expansion Meier gives an approximately unbiased estimate of $V(\bar{y}_w^{lr})$,

$$v_M(\bar{y}_w^{lr}) = \frac{1}{v(\bar{y}_{lr})^{-1} + v(\bar{y}_3)^{-1}} \left[1 + 4\hat{w}_1\hat{w}_2\left(\frac{1}{m_1} + \frac{1}{m_2}\right)\right] \tag{3.9}$$

where $m_1 = n_2 - 1$, $m_2 = n_3 - 1$, and $\hat{w}_1$ and $\hat{w}_2$ are as given in (3.4). Assuming the assumptions made hold, this estimate of $V(\bar{y}_w^{lr})$ contains bias of order $O(1/n_2^2 + 1/n_3^2)$.

Schreuder (1987) was interested in finding improved variance estimators for the estimator given in (3.2) with w_1 and w_2 as defined in (3.3). To estimate $V(\bar{y}_{lr})$ Schreuder uses the standard linearisation variance estimator given in (2.9). Since $\bar{y}_3$ is a linear estimator, $V(\bar{y}_3)$ is estimated using the simple formula $v_3(\bar{y}_3) = s_{y_3}^2/n_3$, assuming the finite population correction is negligible, where $s_{y_3}^2 = \sum_{i\epsilon s_3}(y_i - \bar{y}_3)^2/(n_3 - 1)$. He uses these in Meier's variance estimator given in (3.9) and calls it the classical method. Schreuder then considers alternate variance estimators of $V(\bar{y}_{lr})$ to use instead in (3.9).

He obtains a jackknife variance estimator using a grouped jackknife procedure. Instead of deleting a single observation, a group of n_1/n_2 units (using integer values) was deleted from random groupings of the sample of n_1 units in the set s_1, and one sample unit was deleted from the set s_2 which corresponds to one of the first phase sample units deleted in the random group of size n_1/n_2. There will be n_2 jackknife estimates in total. Using these jackknife estimates, Schreuder applies the general jackknife formula

$$v_J(\bar{y}_{lr}) = \frac{n_2 - 1}{n_2}\sum_{j=1}^{n_2}(\bar{y}_{lr}(j) - \bar{y}_{lr})^2. \tag{3.10}$$

He then defines

$$\hat{w}_1^J = \frac{v_J(\bar{y}_{lr})^{-1}}{v_J(\bar{y}_{lr})^{-1} + v_3(\bar{y}_3)^{-1}} \quad \text{and} \quad \hat{w}_2^J = \frac{v_3(\bar{y}_3)^{-1}}{v_J(\bar{y}_{lr})^{-1} + v_3(\bar{y}_3)^{-1}}.$$

and uses these weights in Meier's variance estimator given in equation (3.9). This gives Schreuder's jackknife variance estimator,

$$v_{J_{sch}}(\bar{y}_w^{lr}) = \frac{1}{v_J(\bar{y}_{lr})^{-1} + v_3(\bar{y}_3)^{-1}} \left[1 + 4\hat{w}_1^J\hat{w}_2^J\left(\frac{1}{m_1} + \frac{1}{m_2}\right)\right]. \tag{3.11}$$

A bootstrap variance estimator was also developed by Schreuder (1987). A similar approach was used to that used in developing the jackknife variance estimator; new weights were calculated now with $v_B(\bar{y}_{lr})$ and then substituted back into the variance formula given in (3.9). A simulation study was used taking samples from forest plots from the northeastern United States. Efficiency, estimation bias and confidence limit coverage were investigated in order to determine which estimators are preferable. They found their bootstrap and jackknife variance estimators to be more efficient in terms of mean squared error than the classical variance estimator given by Meier (1953) only for highly skewed populations. The classical estimator was found to be the most stable, although the jackknife coverage rates were found to be the best among all variance estimators.

Chapter 4

Development of variance estimators

4.1 Fixed weights: w_1 and w_2

4.1.1 Linearisation variance estimator

With fixed weights the linearisation estimator is easily obtained. Consider the estimator given in (3.2). Since the two-phase sample and the random sample on the second occasion are independent, we find for fixed w_1 and w_2:

$$V(\bar{y}_w^{lr}) = w_1^2 V(\bar{y}_{lr}) + w_2^2 V(\bar{y}_3).$$

We could use the linearisation estimator given in Chapter 2 equation (2.9) to estimate $V(\bar{y}_{lr})$. The variance component resulting from the independent random sample on the second occasion, $V(\bar{y}_3)$, may be estimated using the variance formula for a mean from a simple random sample. Thus one possible estimator is:

$$v_0(\bar{y}_w^{lr}) = w_1^2 \left[\left(\frac{1}{n_2} - \frac{1}{n_1} \right) s_d^2 + \left(\frac{1}{n_1} - \frac{1}{N} \right) s_{y_2}^2 \right] + w_2^2 \left[\left(\frac{1}{n_3} - \frac{1}{N} \right) s_{y_3}^2 \right] \qquad (4.1)$$

where $s_d^2 = \sum_{i \epsilon s_2} d_i^2/(n_2-1)$, $s_{y_2}^2 = \sum_{i \epsilon s_2}(y_i-\bar{y}_2)^2/(n_2-1)$, and $s_{y_3}^2 = \sum_{i \epsilon s_3}(y_i-\bar{y}_3)^2/(n_3-1)$. Note that $d_i = y_i - \bar{y}_2 - b(x_i - \bar{x}_2)$ and b is calculated based on observations from the second phase of the two-phase sample. We can use the identity $S_y^2 = S_D^2 + B^2 S_x^2$ and (4.1) simplifies to

$$v_0(\bar{y}_w^{lr}) = w_1^2 \left[\left(\frac{1}{n_2} - \frac{1}{N} \right) s_d^2 + \left(\frac{1}{n_1} - \frac{1}{N} \right) b^2 s_{x_2}^2 \right] + w_2^2 \left[\left(\frac{1}{n_3} - \frac{1}{N} \right) s_{y_3}^2 \right] \qquad (4.2)$$

where $s_{x_2}^2 = \sum_{i \epsilon s_2}(x_i - \bar{x}_2)^2/(n_2 - 1)$.

As discussed in Chapter 2, other variance estimators are possible which make fuller use of the data. Recall Sitter (1997) used the identity mentioned above to obtain an alternative linearisation variance estimator which makes fuller use of the data, see (2.11). We substitute this estimate in for $v(\bar{y}_{lr})$, and find

$$v_1(\bar{y}_w^{lr}) = w_1^2 \left[\left(\frac{1}{n_2} - \frac{1}{N} \right) s_d^2 + \left(\frac{1}{n_1} - \frac{1}{N} \right) b^2 s_{x_1}^2 \right] + w_2^2 \left[\left(\frac{1}{n_3} - \frac{1}{N} \right) s_{y_3}^2 \right] \tag{4.3}$$

where $s_{x_1}^2 = \sum_{i \epsilon s_1}(x_i - \bar{x}_1)^2/(n_1 - 1)$. Based on Sitter's simulation results, we expect that $v_1(\bar{y}_w^{lr})$ will be more efficient than $v_0(\bar{y}_w^{lr})$.

We may also consider various estimators for S_y^2 in $V(\bar{y}_3) = (1/n_3 - 1/N)S_y^2$ which make fuller use of the y-values. For example, we could use all y-values measured in both s_2 and s_3 to estimate S_y^2 to give

$$v_2(\bar{y}_w^{lr}) = w_1^2 \left[\left(\frac{1}{n_2} - \frac{1}{N} \right) s_d^2 + \left(\frac{1}{n_1} - \frac{1}{N} \right) b^2 s_{x_1}^2 \right] + w_2^2 \left[\left(\frac{1}{n_3} - \frac{1}{N} \right) s_{y_{23}}^2 \right] \tag{4.4}$$

where $s_{y_{23}}^2 = \sum_{s_2 \cup s_3}(y_i - \bar{y}_{23})^2/(n_2 + n_3 - 1)$ and $\bar{y}_{23} = \sum_{s2 \cup s_3} y_i/(n_2 + n_3)$. However, as Rao and Sitter (1997) discussed and showed in a different context through a simulation study, positive covariances between terms in variance estimators may be introduced in attempts to make fuller use of the data in this way. This can result in inflating the mean squared error of the variance estimator. For example in (4.4) above, if $s_{y_{23}}^2$ is positively correlated with the first term in the square brackets, the MSE of this estimator could be larger than that for $v_1(\bar{y}_w^{lr})$.

If we instead consider the estimate for $\bar{Y}$ which uses the ratio estimator instead of the regression estimator, given in (3.5), and again assume w_1 and w_2 are fixed, we can follow a parallel argument to that given above using results discussed in Chapter 2. We may simply use the linearisation estimator given in (2.3) to estimate $V(\bar{y}_r)$ to get

$$\begin{aligned} v_0(\bar{y}_w^r) &= w_1^2 \left[\left(\frac{1}{n_2} - \frac{1}{N} \right) s_d^2 + 2 \left(\frac{1}{n_1} - \frac{1}{N} \right) \hat{R} s_{dx} + \left(\frac{1}{n_1} - \frac{1}{N} \right) \hat{R}^2 s_{x_2}^2 \right] \\ &+ w_2^2 \left[\left(\frac{1}{n_3} - \frac{1}{N} \right) s_{y_3}^2 \right], \end{aligned} \tag{4.5}$$

where $s_d^2 = \sum_{i \epsilon s_2} d_i^2/(n_2 - 1)$ with $d_i = y_i - \hat{R}x_i$, $s_{dx_2} = \sum_{i \epsilon s_2} d_i(x_i - \bar{x}_2)/(n_2 - 1)$ and $\hat{R} = \bar{y}_2/\bar{x}_2$.

Recall Rao and Sitter (1995) suggested an alternative linearisation variance estimator for $V(\bar{y}_r)$ given in (2.4) which makes more complete use of the data. Using this estimator we find

$$\begin{aligned} v_1(\bar{y}_w^r) &= w_1^2\left[\left(\frac{1}{n_2}-\frac{1}{N}\right)s_d^2+2\left(\frac{1}{n_1}-\frac{1}{N}\right)\hat{R}s_{dx}+\left(\frac{1}{n_1}-\frac{1}{N}\right)\hat{R}^2s_{x_1}^2\right] \\ &+ w_2^2\left[\left(\frac{1}{n_3}-\frac{1}{N}\right)s_{y_2}^2\right]. \end{aligned} \tag{4.6}$$

Also $s_{y_3}^2$ could be replaced by $s_{y_{23}}^2$.

4.1.2 Jackknife variance estimator

A jackknife variance estimator may also be found for the fixed weight case. Deleting the j^{th} unit will effect the estimator differently depending on which set s_1, s_2 or s_3 the unit is in. We first rewrite equation (3.2):

$$\bar{y}_w^{lr} = w_1[\bar{y}_2+b(\bar{x}_1-\bar{x}_2)]+w_2\bar{y}_3. \tag{4.7}$$

We now develop the following notation:

$$\bar{x}_1(j)=\begin{cases}\frac{n_1\bar{x}_1-x_j}{n_1-1} & j\epsilon s_1\\ \bar{x}_1 & j\epsilon s_3,\end{cases}$$

$$\bar{x}_2(j)=\begin{cases}\frac{n_2\bar{x}_2-x_j}{n_2-1} & j\epsilon s_2\\ \bar{x}_2 & j\epsilon(s_1\cap s_2^c)\cup s_3,\end{cases}$$

$$\bar{y}_2(j)=\begin{cases}\frac{n_2\bar{y}_2-y_j}{n_2-1} & j\epsilon s_2\\ \bar{y}_2 & j\epsilon(s_1\cap s_2^c)\cup s_3,\end{cases}$$

$$\bar{y}_3(j)=\begin{cases}\frac{n_3\bar{y}_3-y_j}{n_3-1} & j\epsilon s_3\\ \bar{y}_3 & j\epsilon s_1\cup s_2,\end{cases}$$

and

$$b(j) = \begin{cases} b - \frac{(x_j - \bar{x}_2)d_j}{(n_2-1)s^2_{x_2}(1-k_j)} & j \epsilon s_2 \\ b & j \epsilon (s_1 \cap s_2^c) \cup s_3 \end{cases}$$

where $d_j = y_j - \bar{y}_2 - b(x_j - \bar{x}_2)$ and $k_j = 1/n_2 + (x_j - \bar{x}_2)^2/\{(n_2-1)s^2_{x_2}\}$.

Let

$$\begin{aligned} \bar{y}_w^{lr}(j) &= w_1 \bar{y}_{lr}(j) + w_2 \bar{y}_3(j) \\ &= w_1[\bar{y}_2(j) + b(j)(\bar{x}_1(j) - \bar{x}_2(j))] + w_2 \bar{y}_3(j). \end{aligned}$$

Ignoring finite population corrections, we may apply the usual jackknife formula

$$v_J(\bar{y}_w^{lr}) = \sum_{j \epsilon s_1 \cup s_3} (\bar{y}_w^{lr}(j) - \bar{y}_w^{lr})^2, \tag{4.8}$$

and $v_J(\bar{y}_w^{lr})$ is the jackknife variance estimator. Note that no bias correction factor is included in (4.8). Since we are jackknifing over two sets, s_1 and s_3, the usual correction factor cannot be applied.

We note for large n_2,

$$\bar{y}_w^{lr}(j) - \bar{y}_w^{lr} \doteq \begin{cases} -w_1 b\left(\frac{x_j - \bar{x}_1}{n_1 - 1}\right) & j \epsilon s_1 \\ -w_1\left[b\left(\frac{x_j - \bar{x}_1}{n_1-1}\right) + \frac{d_j}{n_2}\left(1 + \frac{a_j}{(1-k_j)}\right)\right] & j \epsilon s_2 \\ -w_2\left(\frac{y_j - \bar{y}_3}{n_3 - 1}\right) & j \epsilon s_3 \end{cases}$$

where $a_j = \{n_2(x_j - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)\}/\{(n_2-1)s^2_{x_2}\}$.

We use the above and (4.8) to obtain

$$v_J(\bar{y}_w^{lr}) \doteq w_1^2\left[\frac{s_d^2}{n_2} + \frac{b^2 s^2_{x_1}}{n_1}\right] + w_2^2 \frac{s^2_{y_3}}{n_3} + w_1^2\left(\frac{\bar{x}_1 - \bar{x}_2}{(n_2-1)s^2_{x_2}}\right)\sum_{j \epsilon s_2} \frac{d_j^2(x_j - \bar{x}_2)}{(1-k_j)^2} + R \tag{4.9}$$

where

$$R = \frac{2w_1^2}{n_2}\left[\frac{1}{n_2}\sum_{j \epsilon s_2} \frac{d_j^2 a_j}{(1-k_j)} + \frac{b}{n_1 - 1}\sum_{j \epsilon s_2} \frac{d_j a_j (x_j - \bar{x}_1)}{(1-k_j)}\right].$$

If we include the appropriate finite population corrections we find

$$v_{LJ}(\bar{y}_w^{lr}) = w_1^2\left[\left(\frac{1}{n_2} - \frac{1}{N}\right)s_d^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right)b^2 s^2_{x_1}\right] + w_2^2\left(\frac{1}{n_3} - \frac{1}{N}\right)s^2_{y_3}$$

$$+ \quad w_1^2 \left(\frac{\bar{x}_1 - \bar{x}_2}{(n_2 - 1)s_{x_2}^2} \right) \sum_{j \epsilon s_2} \frac{d_j^2 (x_j - \bar{x}_2)}{(1 - k_j)^2} + R \tag{4.10}$$

$$= \quad v_1(\bar{y}_w^{lr}) + w_1^2 \left(\frac{\bar{x}_1 - \bar{x}_2}{(n_2 - 1)s_{x_2}^2} \right) \sum_{j \epsilon s_2} \frac{d_j^2 (x_j - \bar{x}_2)}{(1 - k_j)^2} + R$$

where v_{LJ} indicates that this is a linearised jackknife variance estimator.

We can follow a parallel argument to develop a jackknife variance estimator for the estimator given in equation (3.5). Firstly we can write the jackknife estimate as

$$\bar{y}_w^r(j) = w_1 \frac{\bar{y}_2(j)}{\bar{x}_2(j)} \bar{x}_1(j) + w_2 \bar{y}_3(j), \tag{4.11}$$

where $\bar{x}_1(j)$, $\bar{x}_2(j)$, $\bar{y}_2(j)$ and $\bar{y}_3(j)$ are as defined previously. The general jackknife formula gives

$$v_J(\bar{y}_w^r) \quad = \quad \sum_{j \epsilon s_1 \cup s_3} (\bar{y}_w^r(j) - \bar{y}_w^r)^2. \tag{4.12}$$

For large n_2, we find

$$\bar{y}_w^r(j) - \bar{y}_w^r \doteq \begin{cases} -w_1 \hat{R} \left(\frac{x_j - \bar{x}_1}{n_1 - 1} \right) & j \epsilon s_1 \\ -w_1 \left[\hat{R} \left(\frac{x_j - \bar{x}_1}{n_1 - 1} \right) + \frac{\bar{x}_1(j)}{\bar{x}_2(j)} \left(\frac{y_j - \hat{R} x_j}{n_2 - 1} \right) \right] & j \epsilon s_2 \\ -w_2 \left(\frac{y_j - \bar{y}_3}{n_3 - 1} \right) & j \epsilon s_3. \end{cases}$$

We assume $\bar{x}_1(j)/\bar{x}_2(j) \doteq \bar{x}_1/\bar{x}_2$, use the above result and equation (4.12) to find

$$v_J(\bar{y}_w^r) \doteq w_1^2 \left[\frac{\hat{R}^2 s_{x_1}^2}{n_1} + 2 \left(\frac{\bar{x}_1}{\bar{x}_2} \right) \frac{\hat{R} s_{dx_2}}{n_1} + \left(\frac{\bar{x}_1}{\bar{x}_2} \right)^2 \frac{s_d^2}{n_2} \right] + w_2^2 \frac{s_{y_3}^2}{n_3}. \tag{4.13}$$

If we include the appropriate finite population corrections we obtain the linearised jackknife estimator,

$$v_{LJ}(\bar{y}_w^r) \quad = \quad w_1^2 \left(\frac{1}{n_1} - \frac{1}{N} \right) \hat{R}^2 s_{x_1}^2 + 2 w_1^2 \left(\frac{\bar{x}_1}{\bar{x}_2} \right) \left(\frac{1}{n_1} - \frac{1}{N} \right) \hat{R} s_{dx_2} + w_1^2 \left(\frac{\bar{x}_1}{\bar{x}_2} \right)^2 \left(\frac{1}{n_2} - \frac{1}{N} \right) s_d^2$$
$$+ \quad w_2^2 \left(\frac{1}{n_3} - \frac{1}{N} \right) s_{y_3}^2. \tag{4.14}$$

We note that the jackknife uses $s_{y_3}^2$ to estimate the variance of $\bar{y}_3$ in both $\bar{y}_w^{lr}$ and $\bar{y}_w^r$.

4.2 Estimated weights: w_1 and w_2

4.2.1 Linearisation variance estimator

For the case in which weights are estimated using (3.4), a linearisation variance estimator may again be obtained. We will develop a general formula for a linearisation variance estimator which will cover estimators of this form. The notation used in the following development is taken from Rao and Sitter (1997). These estimators can be written as a function of means, ie. $\hat{\theta} = g(\bar{\boldsymbol{z}}_3, \bar{\boldsymbol{w}}_2, \bar{\boldsymbol{v}}_1)$, estimating the population parameter $\theta = g(\bar{\boldsymbol{Z}}, \bar{\boldsymbol{W}}, \bar{\boldsymbol{V}})$, where $\bar{\boldsymbol{Z}}$ is the population mean of $\boldsymbol{z}$, $\bar{\boldsymbol{W}}$ is the population mean of $\boldsymbol{w} = (\boldsymbol{u}^T, \boldsymbol{v}^T)^T$ and $\bar{\boldsymbol{V}}$ is the population mean of $\boldsymbol{v}$ where $\boldsymbol{v}$ is observed for the entire first-phase sample, $\boldsymbol{u}$ is observed only on the second-phase sample and $\boldsymbol{z}$ is observed only on the independent random sample at time 2. For example, consider the estimator $\bar{y}_w^r = w_1\bar{y}_r + w_2\bar{y}_3$ with w_1 and w_2 fixed. Then using the notation defined above $\bar{\boldsymbol{v}}_1 = \bar{x}_1$, $\bar{\boldsymbol{w}}_2 = (\bar{u}_2, \bar{v}_2) = (\bar{y}_2, \bar{x}_2)$ and $\bar{\boldsymbol{z}}_3 = \bar{y}_3$, and the estimator can now be written as $\bar{y}_w^r = w_1\frac{\bar{y}_2}{\bar{v}_2}\bar{v}_1 + w_2\bar{z}_3$.

We can now write $\hat{\theta} = g(\bar{\boldsymbol{Z}} + \triangle\bar{\boldsymbol{z}}_3, \bar{\boldsymbol{W}} + \triangle\bar{\boldsymbol{w}}_2, \bar{\boldsymbol{V}} + \triangle\bar{\boldsymbol{v}}_1) = h(\triangle\bar{\boldsymbol{z}}_3, \triangle\bar{\boldsymbol{w}}_2, \triangle\bar{\boldsymbol{v}}_1)$ where $\triangle\bar{\boldsymbol{z}}_3 = \bar{\boldsymbol{z}}_3 - \bar{\boldsymbol{Z}}$, $\triangle\bar{\boldsymbol{w}}_2 = \bar{\boldsymbol{w}}_2 - \bar{\boldsymbol{W}}$, and $\triangle\bar{\boldsymbol{v}}_1 = \bar{\boldsymbol{v}}_1 - \bar{\boldsymbol{V}}$. By a first order Taylor expansion of $\hat{\theta} = h(\triangle\bar{\boldsymbol{z}}_3, \triangle\bar{\boldsymbol{w}}_2, \triangle\bar{\boldsymbol{v}}_1)$ around the point $(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0})$, we find

$$\hat{\theta} - \theta \doteq (\triangle\bar{\boldsymbol{z}}_3)^T\boldsymbol{h}^{(z)} + (\triangle\bar{\boldsymbol{w}}_2)^T\boldsymbol{h}^{(w)} + (\triangle\bar{\boldsymbol{v}}_1)^T\boldsymbol{h}^{(v)}$$

where $\boldsymbol{h}^{(z)} = \boldsymbol{h}^{(z)}(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0})$, $\boldsymbol{h}^{(w)} = \boldsymbol{h}^{(w)}(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0})$ and $\boldsymbol{h}^{(v)} = \boldsymbol{h}^{(v)}(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0})$ denote the vectors of derivatives with respect to $\triangle\bar{\boldsymbol{z}}_3$, $\triangle\bar{\boldsymbol{w}}_2$ and $\triangle\bar{\boldsymbol{v}}_1$ with lengths m, $k+l$ and l respectively. We can now get a general formula for the mean squared error of $\hat{\theta}$.

$$\begin{aligned} MSE(\hat{\theta}) \doteq & \sum_{\alpha=1}^{k+l}\sum_{\alpha'=1}^{k+l} h_\alpha^{(w)}h_{\alpha'}^{(w)}E(\triangle\bar{w}_{2\alpha}\triangle\bar{w}_{2\alpha'}) + \sum_{\beta=1}^{l}\sum_{\beta'=1}^{l} h_\beta^{(v)}h_{\beta'}^{(v)}E(\triangle\bar{v}_{1\beta}\triangle\bar{v}_{1\beta'}) \\ & + \sum_{\gamma=1}^{m}\sum_{\gamma'=1}^{m} h_\gamma^{(z)}h_{\gamma'}^{(z)}E(\triangle\bar{z}_{3\gamma}\triangle\bar{z}_{3\gamma'}) + 2\sum_{\alpha=1}^{k+l}\sum_{\beta=1}^{l} h_\alpha^{(w)}h_\beta^{(v)}E(\triangle\bar{w}_{2\alpha}\triangle\bar{v}_{1\beta}) \\ & + 2\sum_{\alpha=1}^{k+l}\sum_{\gamma=1}^{m} h_\alpha^{(w)}h_\gamma^{(z)}E(\triangle\bar{w}_{2\alpha}\triangle\bar{z}_{3\gamma}) + 2\sum_{\beta=1}^{l}\sum_{\gamma=1}^{m} h_\beta^{(v)}h_\gamma^{(z)}E(\triangle\bar{v}_{1\beta}\triangle\bar{z}_{3\gamma}) \end{aligned}$$

Since the two-phase sample and the independent random sample on the second phase are independent, we may rewrite the above as follows:

$$\begin{aligned} MSE(\hat{\theta}) &\doteq \sum_{\alpha=1}^{k+l}\sum_{\alpha'=1}^{k+l} h_{\alpha}^{(w)}h_{\alpha'}^{(w)}E(\Delta\bar{w}_{2\alpha}\Delta\bar{w}_{2\alpha'}) + \sum_{\beta=1}^{l}\sum_{\beta'=1}^{l} h_{\beta}^{(v)}h_{\beta'}^{(v)}E(\Delta\bar{v}_{1\beta}\Delta\bar{v}_{1\beta'}) \\ &+ \sum_{\gamma=1}^{m}\sum_{\gamma'=1}^{m} h_{\gamma}^{(z)}h_{\gamma'}^{(z)}E(\Delta\bar{z}_{3\gamma}\Delta\bar{z}_{3\gamma'}) + 2\sum_{\alpha=1}^{k+l}\sum_{\beta=1}^{l} h_{\alpha}^{(w)}h_{\beta}^{(v)}E(\Delta\bar{w}_{2\alpha}\Delta\bar{v}_{1\beta}) \\ &+ 2\sum_{\alpha=1}^{k+l}\sum_{\gamma=1}^{m} h_{\alpha}^{(w)}h_{\gamma}^{(z)}E(\Delta\bar{w}_{2\alpha})E(\Delta\bar{z}_{3\gamma}) + 2\sum_{\beta=1}^{l}\sum_{\gamma=1}^{m} h_{\beta}^{(v)}h_{\gamma}^{(z)}E(\Delta\bar{v}_{1\beta})E(\Delta\bar{z}_{3\gamma}). \end{aligned}$$

Note that $E(\Delta\bar{w}_{2\alpha}) = E(\Delta\bar{v}_{1\beta}) = E(\Delta\bar{z}_{3\gamma}) = 0$ and the above simplifies to

$$\begin{aligned} MSE(\hat{\theta}) &= \sum_{\alpha=1}^{k+l}\sum_{\alpha'=1}^{k+l} h_{\alpha}^{(w)}h_{\alpha'}^{(w)}S_{w\alpha\alpha'} + \sum_{\beta=1}^{l}\sum_{\beta'=1}^{l} h_{\beta}^{(v)}h_{\beta'}^{(v)}S_{v\beta\beta'} \\ &+ \sum_{\gamma=1}^{m}\sum_{\gamma'=1}^{m} h_{\gamma}^{(z)}h_{\gamma'}^{(z)}S_{z\gamma\gamma'} + 2\sum_{\alpha=1}^{k+l}\sum_{\beta=1}^{l} h_{\alpha}^{(w)}h_{\beta}^{(v)}S_{w\alpha,v\beta} \end{aligned} \tag{4.15}$$

where $S_{w\alpha\alpha'}, S_{v\beta\beta'}, S_{z\gamma\gamma'}$ and $S_{w\alpha,v\beta}$ are covariances of characteristics w_α and $w_{\alpha'}$, v_β and $v_{\beta'}$, z_γ and $z_{\gamma'}$, and w_α and v_β respectively. If we replace the covariance terms in equation (4.15) by their sample covariances, we obtain the following linearisation variance estimator

$$\begin{aligned} v_L(\hat{\theta}) &= \sum_{\alpha=1}^{k+l}\sum_{\alpha'=1}^{k+l} \hat{h}_{\alpha}^{(w)}\hat{h}_{\alpha'}^{(w)}s_{w\alpha\alpha'} + \sum_{\beta=1}^{l}\sum_{\beta'=1}^{l} \hat{h}_{\beta}^{(v)}\hat{h}_{\beta'}^{(v)}s_{v\beta\beta'} \\ &+ \sum_{\gamma=1}^{m}\sum_{\gamma'=1}^{m} \hat{h}_{\gamma}^{(z)}\hat{h}_{\gamma'}^{(z)}s_{z\gamma\gamma'} + 2\sum_{\alpha=1}^{k+l}\sum_{\beta=1}^{l} \hat{h}_{\alpha}^{(w)}\hat{h}_{\beta}^{(v)}s_{w\alpha,v\beta} \end{aligned} \tag{4.16}$$

Note that $s_{w\alpha,v\beta}$ must be estimated based on the second-phase sample only.

To illustrate, let us return to the example previously discussed. For the estimator $\bar{y}_w^r$ we have worked out the vector notation necessary to apply the above general formula. Recall

$$\begin{aligned} \hat{\theta} &= g(\bar{\boldsymbol{z}}_3, \bar{\boldsymbol{w}}_2, \bar{\boldsymbol{v}}_1) = g(\bar{\boldsymbol{Z}} + \Delta\bar{\boldsymbol{z}}_3, \bar{\boldsymbol{W}} + \Delta\bar{\boldsymbol{w}}_2, \bar{\boldsymbol{V}} + \Delta\bar{\boldsymbol{v}}_1) \\ &= w_1\frac{\bar{u}_2}{\bar{v}_2}\bar{v}_1 + w_2\bar{z}_3 \\ &= w_1\frac{\bar{U} + \Delta\bar{u}_2}{\bar{V} + \Delta\bar{v}_2}(\bar{V} + \Delta\bar{v}_1) + w_2(\bar{Z} + \Delta\bar{z}_3) \\ &= h(\Delta\bar{\boldsymbol{z}}_3, \Delta\bar{\boldsymbol{w}}_2, \Delta\bar{\boldsymbol{v}}_1) \end{aligned}$$

We calculate the partial derivatives necessary to apply the general MSE formula given in equation (4.15) and find $h_1^{(z)} = w_2, h_1^{(w)} = w_1, h_2^{(w)} = -w_1\frac{\bar{U}}{\bar{V}}$ and $h_1^{(v)} = w_1\frac{\bar{U}}{\bar{V}}$. Applying the formula we obtain

$$\begin{aligned} MSE(\hat{\theta}) &\doteq w_1^2 E(\triangle\bar{u}_2\triangle\bar{u}_2) - 2w_1^2\frac{\bar{U}}{\bar{V}}E(\triangle\bar{u}_2\triangle\bar{v}_2) + w_1^2\frac{\bar{U}^2}{\bar{V}^2}E(\triangle\bar{v}_2\triangle\bar{v}_2) \\ &+ w_1^2\frac{\bar{U}^2}{\bar{V}^2}E(\triangle\bar{v}_1\triangle\bar{v}_1) + w_2^2 E(\triangle\bar{z}_3\triangle\bar{z}_3) + 2w_1^2\frac{\bar{U}}{\bar{V}}E(\triangle\bar{u}_2\triangle\bar{v}_1) \\ &- 2w_1^2\frac{\bar{U}^2}{\bar{V}^2}E(\triangle\bar{v}_2\triangle\bar{v}_1). \end{aligned}$$

We use $R = \frac{\bar{U}}{\bar{V}} = \frac{\bar{Y}}{\bar{X}}$, ignore finite population corrections and evaluate the expectations to find

$$\begin{aligned} V_L(\bar{y}_w^r) &\doteq w_1^2\left[\frac{S_y^2}{n_2} - 2R\frac{S_{xy}}{n_2} + R^2\frac{S_x^2}{n_2} + R^2\frac{S_x^2}{n_1} + 2R\frac{S_{xy}}{n_1} - 2R^2\frac{S_x^2}{n_1}\right] + w_2^2\frac{S_y^2}{n_3} \\ &= w_1^2\left[\frac{S_d^2}{n_2} + R^2\frac{S_x^2}{n_1} + 2R\left(\frac{S_{xy} - RS_x^2}{n_1}\right)\right] + w_2^2\frac{S_y^2}{n_3} \\ &= w_1^2\left[\frac{S_d^2}{n_2} + 2R\frac{S_{dx}}{n_1} + R^2\frac{S_x^2}{n_1}\right] + w_2^2\frac{S_y^2}{n_3}. \end{aligned}$$

Though we have applied the general formula to an estimator with fixed weights, (4.15) would also apply to the estimator $\bar{y}_w^{lr}$ or $\bar{y}_w^r$ with weights w_1 and w_2 estimated as in (3.4), provided the estimator $\hat{w}_1$ of w_1 can be expressed as a function of means. To illustrate consider $\bar{y}_w^{lr} = \hat{w}_1\bar{y}_{lr} + \hat{w}_2\bar{y}_3$ where

$$\hat{w}_1 = \frac{v_0(\bar{y}_{lr})^{-1}}{v_0(\bar{y}_{lr})^{-1} + v_3(\bar{y}_3)^{-1}} \qquad \text{and} \qquad \hat{w}_2 = 1 - \hat{w}_1,$$

where $v_0(\bar{y}_{lr})$ is as given in (2.9) and $v_3(\bar{y}_3) = (1/n_3 - 1/N)s_{y_3}^2$. We note that $\bar{y}_{lr} = \bar{y}_2 + b(\bar{x}_1 - \bar{x}_2)$ can be written as a function of means if the regression coefficient, b, can be written as a function of means. We use some simple algebra to find

$$\begin{aligned} b &= \frac{s_{xy_2}}{s_{x_2}^2} \\ &= \frac{\sum_{i\epsilon s_2}(x_i - \bar{x}_2)(y_i - \bar{y}_2)}{\sum_{i\epsilon s_2}(x_i - \bar{x}_2)^2} \\ &= \frac{\frac{1}{n_2}\sum_{i\epsilon s_2}(x_i - \bar{x}_2)y_i}{\frac{1}{n_2}\sum_{i\epsilon s_2}(x_i - \bar{x}_2)^2} \\ &= \frac{\bar{u}_2 - \bar{x}_2\bar{y}_2}{\bar{v}_2 - \bar{x}_2^2}. \end{aligned}$$

Thus b is in fact a function of means, where $\bar{u}_2 = \sum_{i \epsilon s_2} x_i y_i / n_2$ and $\bar{v}_2 = \sum_{i \epsilon s_2} x_i^2 / n_2$. The variance estimators $v_0(\bar{y}_{lr})$ and $v_3(\bar{y}_3)$ can be written as functions of means using similar arguments and algebraic manipulation as shown above. This implies that $\bar{y}_w^{lr}$ can be written as a function of means, and is therefore in the class of estimators to which the general formula given in (4.15) may be applied. Note that all of the discussed estimators for $V(\bar{y}_{lr})$ and $V(\bar{y}_3)$ can be written as functions of means. However, $\bar{y}_w^{lr}$ with estimated weights will be a function of many means, and thus one would be hesitant to use this procedure. That said, the fact that this general method does apply allows for theoretical developments with the jackknife.

To use the jackknife procedure we need only be aware that the estimator of interest $\hat{\theta}$ can be written as a function of means. That is, the jackknife methodology is correct asymptotically for parameters $\hat{\theta}$ that may be written in this form. It is unnecessary to write it as such and expand. We may simply implement the delete-one-unit jackknife method to obtain the jackknife estimates, $\hat{\theta}(j)$, and apply the general formula

$$v_J(\hat{\theta}) = \sum_{j \epsilon s_1 \cup s_3} (\hat{\theta}(j) - \hat{\theta})^2. \tag{4.17}$$

4.2.2 Jackknife variance estimator

We now develop a general formula for a jackknife variance estimator for the case of estimated weights. Consider the vector notation given in the previous section where we claimed we can write the estimator $\hat{\theta} = g(\bar{\boldsymbol{z}}, \bar{\boldsymbol{w}}, \bar{\boldsymbol{v}})$. As in the fixed weight case, the effect of deleting a point will effect these means differently depending on which set, s_1, s_2, or s_3 the point was originally in. We can write

$$\bar{\boldsymbol{v}}_1\ (j) = \begin{cases} \frac{n_1 \bar{\boldsymbol{v}}_1 - \boldsymbol{v}_j}{n_1 - 1} & j \epsilon s_1 \\ \bar{\boldsymbol{v}}_1 & j \epsilon s_3, \end{cases}$$

$$\bar{\boldsymbol{w}}_2\ (j) = \begin{cases} \frac{n_2 \bar{\boldsymbol{w}}_2 - \boldsymbol{w}_j}{n_2 - 1} & j \epsilon s_2 \\ \bar{\boldsymbol{w}}_2 & j \epsilon (s_1 \cap s_2^c) \cup s_3, \end{cases}$$

and

$$\bar{z}_3(j) = \begin{cases} \frac{n_3\bar{z}_3 - z_j}{n_3-1} & j\epsilon s_3 \\ \bar{z}_3 & j\epsilon s_1. \end{cases}$$

We can now write the jackknife estimate $\hat{\theta}(j) = g(\bar{z}(j), \bar{w}(j), \bar{v}(j))$. To get the general formula we also note

$$\Delta\bar{v}_1(j) = \bar{v}_1(j) - \bar{v}_1 = \begin{cases} -\frac{(v_j - \bar{v}_1)}{n_1-1} & j\epsilon s_1 \\ 0 & j\epsilon s_3, \end{cases}$$

$$\Delta\bar{w}_2(j) = \bar{w}_2(j) - \bar{w}_2 = \begin{cases} -\frac{(w_j - \bar{w}_2)}{n_2-1} & j\epsilon s_2 \\ 0 & j\epsilon(s_1 \cap s_2^c) \cup s_3, \end{cases}$$

and

$$\Delta\bar{z}_3(j) = \bar{z}_3(j) - \bar{z}_3 = \begin{cases} -\frac{(z_j - \bar{z}_3)}{n_3-1} & j\epsilon s_3 \\ 0 & j\epsilon s_1. \end{cases}$$

Thus

$$\begin{aligned} \hat{\theta}(j) &= g(\bar{z}_3 + \Delta\bar{z}_3(j), \bar{w}_2 + \Delta\bar{w}_2(j), \bar{v}_1 + \Delta\bar{v}_1(j)) \\ &= \hat{h}(\Delta\bar{z}_3(j), \Delta\bar{w}_2(j), \Delta\bar{v}_1(j)) \end{aligned}$$

By a Taylor expansion of $\hat{\theta}(j) = \hat{h}(\Delta\bar{z}_3(j), \Delta\bar{w}_2(j), \Delta\bar{v}_1(j))$ around $(0,0,0)$ we find

$$\hat{\theta}(j) - \hat{\theta} \doteq (\Delta\bar{z}_3(j))^T\hat{h}^{(z)} + (\Delta\bar{w}_2(j))^T\hat{h}^{(w)} + (\Delta\bar{v}_1(j))^T\hat{h}^{(v)} \qquad (4.18)$$

and we may now use the jackknife formula

$$v_J(\hat{\theta}) = \sum_{j\epsilon s_1 \cup s_3} (\hat{\theta}(j) - \hat{\theta})^2. \qquad (4.19)$$

Let $s_{13} = s_1 \cup s_3$. Then using the above formula and (4.18) we obtain

$$\begin{aligned} v_J(\hat{\theta}) \doteq & \sum_{\alpha=1}^{k+l}\sum_{\alpha'=1}^{k+l} \hat{h}_\alpha^{(w)}\hat{h}_{\alpha'}^{(w)} \sum_{j\epsilon s_{13}} \Delta\bar{w}_{2\alpha}(j)\Delta\bar{w}_{2\alpha'}(j) + \sum_{\beta=1}^{l}\sum_{\beta'=1}^{l} \hat{h}_\beta^{(v)}\hat{h}_{\beta'}^{(w)} \sum_{j\epsilon s_{13}} \Delta\bar{v}_{1\beta}(j)\Delta\bar{v}_{1\beta'}(j) \\ + & \sum_{\gamma=1}^{m}\sum_{\gamma'=1}^{m} \hat{h}_\gamma^{(z)}\hat{h}_{\gamma'}^{(z)} \sum_{j\epsilon s_{13}} \Delta\bar{z}_{3\gamma}(j)\Delta\bar{z}_{3\gamma'}(j) + 2\sum_{\alpha=1}^{k+l}\sum_{\beta=1}^{l} \hat{h}_\alpha^{(w)}\hat{h}_\beta^{(v)} \sum_{j\epsilon s_{13}} \Delta\bar{w}_{2\alpha}(j)\Delta\bar{v}_{1\beta}(j) \\ + & 2\sum_{\alpha=1}^{k+l}\sum_{\gamma=1}^{m} \hat{h}_\alpha^{(w)}\hat{h}_\gamma^{(z)} \sum_{j\epsilon s_{13}} \Delta\bar{w}_{2\alpha}(j)\Delta\bar{z}_{3\gamma}(j) + 2\sum_{\beta=1}^{l}\sum_{\gamma=1}^{m} \hat{h}_\beta^{(v)}\hat{h}_\gamma^{(z)} \sum_{j\epsilon s_{13}} \Delta\bar{v}_{1\beta}(j)\Delta\bar{z}_{3\gamma}(j) \end{aligned}$$

where $\hat{h}_\alpha^{(w)}$ is the derivative of $\hat{h}$ with respect to the components of $\Delta\bar{\boldsymbol{w}}_2\,(j)$ evaluated at $(\mathbf{0},\mathbf{0},\mathbf{0})$, similarly for $\hat{h}_\beta^{(v)}$ and $\hat{h}_\gamma^{(z)}$. As in the linearisation variance estimator the covariance terms between the independent random sample and the double sample are zero and the above simplifies to

$$\begin{aligned} v_J(\hat{\theta}) \doteq & \sum_{\alpha=1}^{k+l}\sum_{\alpha'=1}^{k+l}\hat{h}_\alpha^{(w)}\hat{h}_{\alpha'}^{(w)}\sum_{j\epsilon s_{13}}\Delta\bar{w}_{2\alpha}(j)\Delta\bar{w}_{2\alpha'}(j) \\ & + \sum_{\beta=1}^{l}\sum_{\beta'=1}^{l}\hat{h}_\beta^{(v)}\hat{h}_{\beta'}^{(w)}\sum_{j\epsilon s_{13}}\Delta\bar{v}_{1\beta}(j)\Delta\bar{v}_{1\beta'}(j) \\ & + \sum_{\gamma=1}^{m}\sum_{\gamma'=1}^{m}\hat{h}_\gamma^{(z)}\hat{h}_{\gamma'}^{(z)}\sum_{j\epsilon s_{13}}\Delta\bar{z}_{3\gamma}(j)\Delta\bar{z}_{3\gamma'}(j) \\ & + 2\sum_{\alpha=1}^{k+l}\sum_{\beta=1}^{l}\hat{h}_\alpha^{(w)}\hat{h}_\beta^{(v)}\sum_{j\epsilon s_{13}}\Delta\bar{w}_{2\alpha}(j)\Delta\bar{v}_{1\beta}(j). \end{aligned} \tag{4.20}$$

We now turn to the example we have been following, and will use the above to reproduce the variance estimator given in equation (4.13). That is, we will find the jackknife variance estimator for the estimate given in equation (3.5). We use the same vector notation specified previously and we find

$$\hat{\theta}(j) \doteq w_1\left(\frac{\bar{u}_2+\Delta\bar{u}_2(j)}{\bar{v}_2+\Delta\bar{v}_2(j)}\right)(\bar{v}_1+\Delta\bar{v}_1(j))+w_2(\bar{z}_3+\Delta\bar{z}_3(j)).$$

We calculate partial derivatives and find $\hat{h}_1^{(v)} = w_1\hat{R}$, $\hat{h}_1^{(w)} = w_1\frac{\bar{x}_1}{\bar{x}_2}$, $\hat{h}_2^{(w)} = -w_1\hat{R}\frac{\bar{x}_1}{\bar{x}_2}$ and $\hat{h}_1^{(z)} = w_2$ where $\hat{R} = \frac{\bar{y}_2}{\bar{x}_2}$. Applying the formula given in equation (4.20) we find

$$\begin{aligned} v_J(\hat{\theta}) \doteq & w_1^2\left(\frac{\bar{x}_1}{\bar{x}_2}\right)^2\sum_{j\epsilon s_{13}}\Delta\bar{u}_2(j)\Delta\bar{u}_2(j) - 2w_1^2\hat{R}\sum_{j\epsilon s_{13}}\Delta\bar{u}_2(j)\Delta\bar{v}_2(j) \\ & + w_1^2\hat{R}^2\left(\frac{\bar{x}_1}{\bar{x}_2}\right)^2\sum_{j\epsilon s_{13}}\Delta\bar{v}_2(j)\Delta\bar{v}_2(j) + w_1^2\hat{R}^2\sum_{j\epsilon s_{13}}\Delta\bar{v}_1(j)\Delta\bar{v}_1(j) \\ & + w_2^2\sum_{j\epsilon s_{13}}\Delta\bar{z}_3(j)\Delta\bar{z}_3(j) + 2w_1^2\hat{R}\left(\frac{\bar{x}_1}{\bar{x}_2}\right)\sum_{j\epsilon s_{13}}\Delta\bar{u}_2(j)\Delta\bar{v}_1(j) \\ & - 2w_1^2\hat{R}^2\left(\frac{\bar{x}_1}{\bar{x}_2}\right)\sum_{j\epsilon s_{13}}\Delta\bar{v}_2(j)\Delta\bar{v}_1(j) \end{aligned}$$

This simplifies to

$$v_J(\bar{y}_w^r) \doteq w_1^2\left(\frac{\bar{x}_1}{\bar{x}_2}\right)^2\frac{s_y^2}{n_2} - 2w_1^2\hat{R}\left(\frac{\bar{x}_1}{\bar{x}_2}\right)\frac{s_{xy}}{n_2} + w_1^2\hat{R}^2\left(\frac{\bar{x}_1}{\bar{x}_2}\right)^2\frac{s_x^2}{n_2}$$

$$
\begin{aligned}
&+ \quad w_1^2 \hat{R}^2 \frac{s_{x_1}^2}{n_1} + 2w_1^2 \hat{R} \left(\frac{\bar{x}_1}{\bar{x}_2} \right) \frac{s_{xy}}{n_1} - 2w_1^2 \hat{R}^2 \left(\frac{\bar{x}_1}{\bar{x}_2} \right) \frac{s_x^2}{n_1} + w_2^2 \frac{s_y^2}{n_3} \\
&= \quad w_1^2 \left(\frac{\bar{x}_1}{\bar{x}_2} \right)^2 \frac{s_d^2}{n_2} + 2w_1^2 \hat{R}^2 \left(\frac{\bar{x}_1}{\bar{x}_2} \right) \frac{s_{dx}}{n_1} + w_1^2 \hat{R}^2 \frac{s_{x_1}^2}{n_1} + w_2^2 \frac{s_{y_3}^2}{n_3}
\end{aligned}
$$

which is identical to the estimator given in equation (4.13).

Chapter 5

Simulation study

This chapter describes simulation studies to compare the efficiencies of the discussed variance estimators, relative to the estimator typically used. Two simulation studies will be performed: one for the fixed weight case, and one for the case in which the weights are estimated.

5.1 Models and parameter settings

We need to generate a set of x and y characteristics, where x represents the characteristic at time 1 and y at time 2. We follow a similar design for the study as used in Rao and Sitter (1995) and Sitter (1997). We will create a finite population of size N=16,000. We use models of the following general form

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + x_i^a \epsilon_i, \tag{5.1}$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ are independent of $x_i \sim \text{gamma}(g, h)$, and vary our choice of $\alpha, \beta, g, h, \gamma$, and a. Since the $x_i \sim \text{gamma}(g, h)$, we know $\mu_x = gh$ and $\sigma_x^2 = gh^2$. The coefficient of variation of the x's is then C_x=$\sigma_x/\mu_x = 1/\sqrt{g}$. We consider two values of C_x, 1.0 and $\sqrt{2}$. Recall that the method of sampling on two occasions relies on the fact that x and y, the measurements of the characteristic over two time-points, are highly correlated. Therefore, we wish to vary our choice in ρ=corr(x_i, y_i). We use ρ=0.7 and ρ=0.85.

Table 5.1: 2^3 possible models

No.	β	γ	a	Model	
1)	0	0	0	$y_i = \epsilon_i$	
2)	1	0	0	$y_i = x_i + \epsilon_i$	*
3)	0	0.1	0	$y_i = 0.1x_i^2 + \epsilon_i$	*
4)	1	0.1	0	$y_i = x_i + 0.1x_i^2 + \epsilon_i$	*
5)	0	0	0.5	$y_i = \sqrt{x_i}\epsilon_i$	
6)	1	0	0.5	$y_i = x_i + \sqrt{x_i}\epsilon_i$	*
7)	0	0.1	0.5	$y_i = 0.1x_i^2 + \sqrt{x_i}\epsilon_i$	
8)	1	0.1	0.5	$y_i = x_i + 0.1x_i^2 + \sqrt{x_i}\epsilon_i$	

We need to discuss the choices for α, β, γ, and a. Since we will be investigating estimators which use the linear regression estimator, as opposed to the ratio estimator, the choice of α will not affect the results and so we set this parameter to 0 in all cases. For simplicity, consider two levels for each remaining parameter; β is 0 or 1, γ is 0 or 0.1 and a is 0 or 0.5. There are $2^3 = 8$ models to consider based on all combinations of these parameter settings, see Table 5.1.

Note that models 1) and 5) are of no interest since this will produce 0 correlation between x_i and y_i. Though models 7) and 8) are of some interest, we exclude them and study simpler models. Models 2) and 4) present the basic linear model and the linear model with a moderate size quadratic effect. Model 6) is again the basic linear model, but the variance now depends on x. Model 3) represents a departure from linearity with only the quadratic term plus error. Those models which we will consider have been marked with a "*" in Table 5.1, and are reproduced in Table 5.2.

For each of the 4 models chosen, we wish to run the simulation at C_x=1 and $C_x = \sqrt{2}$, and $\rho = 0.85$ and $\rho = 0.7$. Thus there will be $2 \times 2 = 4$ simulation runs for each model. C_x dictates the setting of g, and ρ dictates the setting of σ_ϵ^2. To determine h, we set μ_x=100 for models i) and ii) and μ_x=10 for models iii) and iv). See Appendix B for the development of

Table 5.2: Models to be used in simulation

No.	Model
i)	$y_i = x_i + \epsilon_i$
ii)	$y_i = x_i + \sqrt{x_i}\epsilon_i$
iii)	$y_i = 0.1x_i^2 + \epsilon_i$
iv)	$y_i = x_i + 0.1x_i^2 + \epsilon_i$

$Cov(X,Y)$ and σ_Y^2, and the parameter values used in the 4 simulations for each model. We give plots of 4 populations, one for each model with $\rho = 0.85$ and $C_x = \sqrt{2}$, see Figure 5.1.

For each simulation, we create a finite population of size $N = 16,000$ and take $B = 10,000$ independent samples using sampling on two occasions with $n_1 = 200, n_2 = 80$ and $n_3 = 100$. Scott (1984) discusses a forestry example applying sampling on two occasions with the above sample sizes. In each iteration we calculate our estimate of $\bar{Y}$ using $\bar{y}_w^{lr}$, refer to (3.2). We can obtain the "true" mean square error of $\bar{y}_w^{lr}$ through simulation using

$$MSE = \frac{1}{B}\sum_{b=1}^{B}(\bar{y}_{w(b)}^{lr} - \bar{Y})^2, \tag{5.2}$$

where $\bar{y}_{w(b)}^{lr}$ is the estimate $\bar{y}_w^{lr}$ obtained on the b^{th} simulation run and $\bar{Y}$ we calculate from the finite population. For each of the variance estimators v, we find its simulated mean square error using

$$MSE(v) = \frac{1}{B}\sum_{b=1}^{B}(v^{(b)} - MSE)^2, \tag{5.3}$$

where $v^{(b)}$ is the variance estimate v from the b^{th} simulation run. Since the finite population is large relative to the sample sizes under investigation, we remove the finite population corrections from all variance estimators. We report relative efficiencies of the variance estimators, using $v_0(\bar{y}_w^{lr})$ as the standard. That is, we report $MSE(v)/MSE(v_0)$ for each of the variance estimates v calculated.

We also report the percent relative bias for each variance estimator, v. To calculate this

Figure 5.1: Example populations for models (i)-(iv)

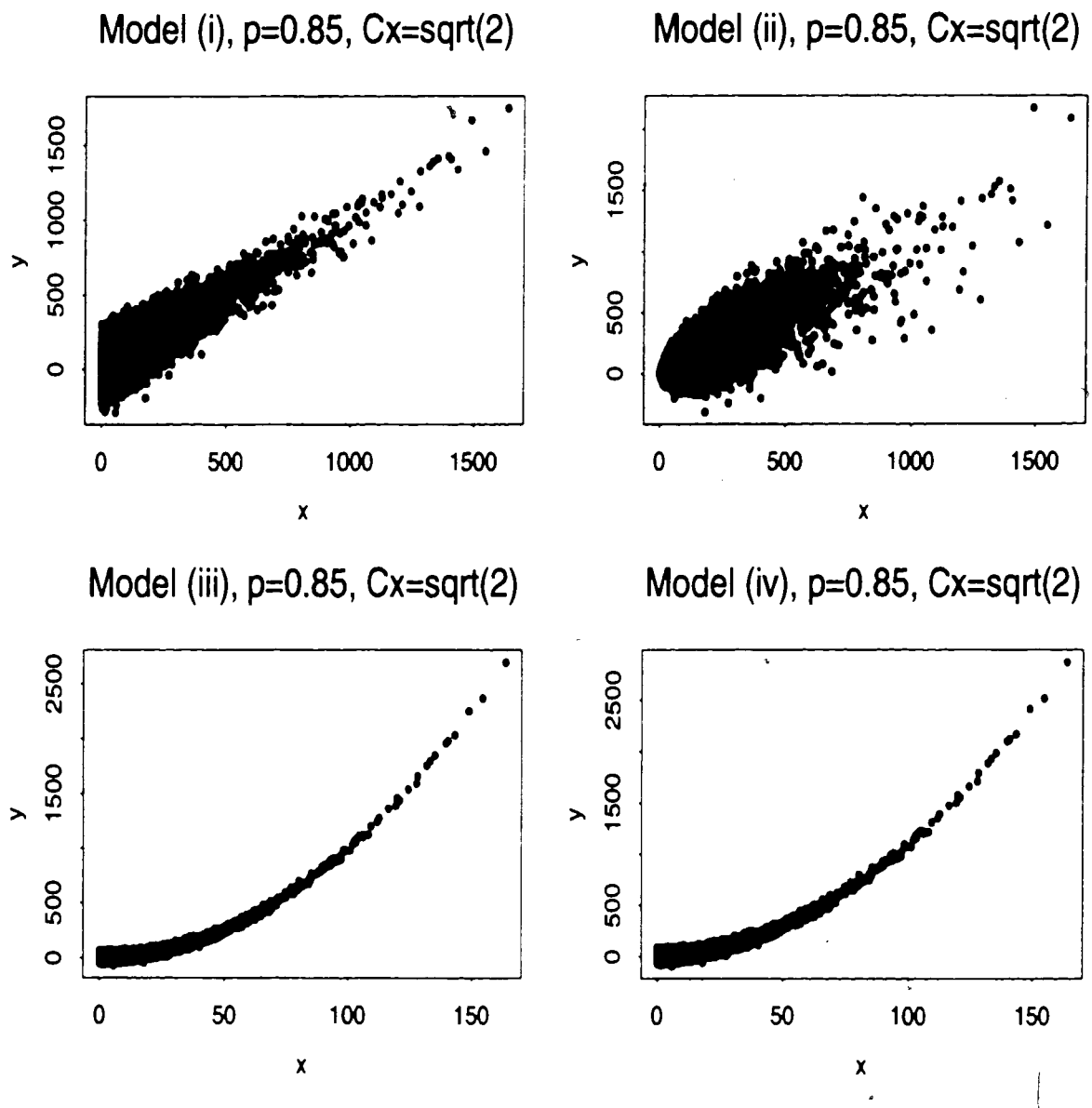


Table 5.3: Percent relative bias: $w_1 = 0.2, w_2 = 0.8$

	rbias(v_0)		rbias(v_1)		rbias(v_2)		rbias(v_{LJ})		rbias(v_J)	
$C_x =$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$
$\rho =$	(i) $y = x + \epsilon$									
0.7	2.36	0.98	2.37	1.18	2.66	1.18	2.41	1.04	3.50	2.13
0.85	2.96	0.73	2.97	0.75	3.40	1.11	2.30	0.77	4.06	1.18
$\rho =$	(ii) $y = x + \sqrt{x}\epsilon$									
0.7	1.47	0.26	1.48	0.27	1.96	0.61	1.59	0.43	2.78	1.69
0.85	2.24	0.24	2.26	0.26	2.80	0.74	2.32	0.35	3.43	1.49
$\rho =$	(iii) $y = 0.1x^2 + \epsilon$									
0.7	1.37	0.86	1.25	0.64	1.56	1.10	1.20	0.51	2.63	2.17
0.85	1.29	0.27	1.10	-0.05	1.58	0.67	0.99	-0.29	2.56	1.62
$\rho =$	(iv) $y = x + 0.1x^2 + \epsilon$									
0.7	1.68	0.90	1.57	0.71	.92	1.14	1.54	0.60	2.89	2.15
0.85	1.66	0.33	1.48	0.05	1.99	0.74	1.38	-0.16	2.85	1.59

we use

$$rbias(v) = \frac{\sum_{b=1}^{B} v^{(b)}/B - MSE}{MSE} \times 100. \tag{5.4}$$

5.2 Fixed weight simulation

For the fixed weight simulation we use three different weighting combinations: $w_1 = 0.2$ and $w_2 = 0.8$, $w_1 = w_2 = 0.5$, and $w_1 = 0.8$ and $w_2 = 0.2$. These are chosen to represent the three general cases of large weight on the double sample, equal weight on the double sample and independent random sample, and small weight on the double sample. For these weights we calculate $v_0(\bar{y}_u^{lr})$, $v_1(\bar{y}_u^{lr})$, $v_2(\bar{y}_u^{lr})$, $v_{LJ}(\bar{y}_u^{lr})$ and $v_J(\bar{y}_u^{lr})$ as presented in Chapter 4, Section 1. We report percent relative bias and relative efficiency of the variance estimators.

We first discuss percent relative bias of the estimates, see Tables 5.3 - 5.5. Note v_0, v_1, v_2, v_{LJ} and v_J are $v_0(\bar{y}_u^{lr})$, $v_1(\bar{y}_u^{lr})$, $v_2(\bar{y}_u^{lr})$, $v_{LJ}(\bar{y}_u^{lr})$ and $v_J(\bar{y}_u^{lr})$, respectively. We note relative bias of v_J is consistently the highest for weighting combinations $w_1 = 0.2, w_2 = 0.8$ and $w_1 = 0.5, w_2 = 0.5$. In Tables 5.3 and 5.4 all other estimators have approximately equal

Table 5.4: Percent relative bias: $w_1 = 0.5, w_2 = 0.5$

	rbias(v_0)		rbias(v_1)		rbias(v_2)		rbias(v_{LJ})		rbias(v_J)	
$C_x =$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$
$\rho =$	(i) $y = x + \epsilon$									
0.7	3.13	1.21	3.21	1.31	3.37	1.42	3.57	1.73	5.18	3.32
0.85	3.46	1.05	3.57	1.18	3.83	1.40	3.79	1.44	5.10	2.72
$\rho =$	(ii) $y = x + \sqrt{x}\epsilon$									
0.7	1.62	-0.55	1.76	-0.41	2.03	-0.22	2.77	1.04	5.22	4.25
0.85	2.43	0.02	2.60	0.19	2.93	0.49	3.23	1.08	5.02	3.32
$\rho =$	(iii) $y = 0.1x^2 + \epsilon$									
0.7	1.80	1.83	0.68	-0.15	0.86	0.11	0.23	-1.39	4.98	5.58
0.85	1.14	0.96	-0.68	-2.29	-0.39	-1.84	-1.81	-4.74	4.67	5.37
$\rho =$	(iv) $y = x + 0.1x^2 + \epsilon$									
0.7	2.25	1.96	1.26	0.22	1.46	0.47	0.91	-0.76	4.94	5.16
0.85	1.46	1.31	-0.32	-1.63	0.00	-1.18	-1.43	-3.72	4.17	4.85

Table 5.5: Percent relative bias: $w_1 = 0.8, w_2 = 0.2$

	rbias(v_0)		rbias(v_1)		rbias(v_2)		rbias(v_{LJ})		rbias(v_J)	
$C_x =$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$
$\rho =$	(i) $y = x + \epsilon$									
0.7	2.54	1.02	2.70	1.23	2.72	1.24	3.41	2.07	5.57	4.21
0.85	2.53	0.95	2.76	1.24	2.80	1.27	3.26	1.81	4.87	3.40
$\rho =$	(ii) $y = x + \sqrt{x}\epsilon$									
0.7	0.31	-3.10	0.58	-2.82	0.61	-2.79	2.56	0.14	6.37	5.30
0.85	1.09	-1.86	1.45	-1.49	1.49	-1.45	2.83	0.44	5.50	4.11
$\rho =$	(iii) $y = 0.1x^2 + \epsilon$									
0.7	1.11	1.56	-1.12	-2.46	-1.09	-2.42	-2.02	-4.98	6.45	8.11
0.85	0.09	0.96	-4.00	-6.53	-3.96	-6.46	-6.51	-12.15	6.67	9.80
$\rho =$	(iv) $y = x + 0.1x^2 + \epsilon$									
0.7	1.64	1.90	-0.38	-1.63	-0.35	-1.60	-1.08	-3.63	6.04	7.36
0.85	0.41	1.76	-3.84	-5.13	-3.79	-5.07	-6.48	-10.05	5.45	8.68

Table 5.6: Relative efficiency: $w_1 = 0.2, w_2 = 0.8$

	$\frac{MSE(v_1)}{MSE(v_0)}$		$\frac{MSE(v_2)}{MSE(v_0)}$		$\frac{MSE(v_{LJ})}{MSE(v_0)}$		$\frac{MSE(v_J)}{MSE(v_0)}$	
$C_x =$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$
$\rho =$	(i) $y = x + \epsilon$							
0.7	0.999	0.999	0.606	0.589	1.000	0.999	1.040	1.027
0.85	0.999	0.999	0.592	0.577	1.000	0.999	1.033	1.022
$\rho =$	(ii) $y = x + \sqrt{x}\epsilon$							
0.7	0.999	0.999	0.612	0.599	1.000	0.999	1.029	1.022
0.85	0.999	0.999	0.599	0.588	0.999	0.999	1.029	1.021
$\rho =$	(iii) $y = 0.1x^2 + \epsilon$							
0.7	0.999	0.999	0.587	0.593	1.000	0.998	1.023	1.021
0.85	0.999	0.999	0.583	0.589	1.000	0.999	1.022	1.020
$\rho =$	(iv) $y = x + 0.1x^2 + \epsilon$							
0.7	0.999	0.999	0.585	0.592	1.000	0.998	1.023	1.021
0.85	0.999	0.999	0.580	0.589	1.000	0.999	1.022	1.020

relative bias. Table 5.5 indicates relative bias is larger for v_1, v_2 and v_{LJ} as the models depart from linearity for $w_1 = 0.8, w_2 = 0.2$ for both large ρ and large C_x. Note that Table 5.5 reports that when the population is given by model (iii) or (iv) where $\rho = 0.85$ and $C_x = \sqrt{2}$, v_{LJ} is underestimating the true MSE by 12.15% and 10.05% respectively. For these models, v_0 performs best in terms of bias. For such non-linear populations, larger sample sizes would be needed to reduce the bias. We ran simulations doubling the size of n_1, n_2 and n_3 for the non-linear models, and found the relative percent bias decrease to approximately two-thirds of that given in Table 5.5.

Turning to relative efficiency, given in Tables 5.6 - 5.8, we see that for the first weighting scheme, $w_1 = 0.2$ and $w_2 = 0.8$, v_2 is considerably more efficient than the other variance estimates in all models considered while v_0, v_1, v_{LJ} and v_J all perform similarly, refer to Table 5.6. This may be due to the large weight on the independent random sample, s_3. It appears that if a covariance term has been introduced in v_2, it does not outweigh the benefits of reduced variance in using s^2_{y23} as opposed to s^2_{y3} as in v_1.

As weight is removed from s_3 and put on the double sample, refer to Table 5.7 where $w_1 =$

Table 5.7: Relative efficiency: $w_1 = 0.5, w_2 = 0.5$

	$\frac{MSE(v_1)}{MSE(v_0)}$		$\frac{MSE(v_2)}{MSE(v_0)}$		$\frac{MSE(v_{LJ})}{MSE(v_0)}$		$\frac{MSE(v_J)}{MSE(v_0)}$	
$C_x =$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$
$\rho =$	(i) $y = x + \epsilon$							
0.7	0.942	0.915	0.983	0.899	0.976	0.938	1.084	1.002
0.85	0.893	0.873	0.804	0.752	0.906	0.879	0.965	0.910
$\rho =$	(ii) $y = x + \sqrt{x}\epsilon$							
0.7	0.931	0.913	1.058	1.001	1.017	1.002	1.155	1.151
0.85	0.894	0.880	0.891	0.861	0.929	0.908	1.004	0.975
$\rho =$	(iii) $y = 0.1x^2 + \epsilon$							
0.7	0.888	0.869	0.966	0.945	0.899	0.842	1.173	1.171
0.85	0.885	0.870	0.959	0.939	0.893	0.844	1.160	1.160
$\rho =$	(iv) $y = x + 0.1x^2 + \epsilon$							
0.7	0.880	0.865	0.932	0.926	0.887	0.839	1.107	1.125
0.85	0.877	0.865	0.922	0.918	0.879	0.839	1.089	1.112

$w_2 = 0.5$, we find less obvious patterns in the relative efficiencies of the variance estimates. We see v_1 slightly outperforms v_2 in all models except model (i). We note v_{LJ} appears to be most efficient for model (iii). However Table 5.5 indicates v_{LJ} is underestimating the true variance for models (iii) and (iv). The efficiency of v_J is similar to that of v_0.

Table 5.8 presents the results with $w_1 = 0.8$ and $w_2 = 0.2$. Here, we put large weight on the double sample, and small weight on the independent random sample. We see v_1 is again more efficient that v_2 in all models considered. In models (i) and (ii), v_J performs well for $\rho = 0.85$. Moreover, v_{LJ} appears to have good efficiency, though, as previously noted, its relative bias is high for models (iii) and (iv) with $\rho = 0.85$. Tables 5.7 and 5.8 present results which concur with those found by Sitter (1997); v_1 and v_{LJ} perform better for large ρ and large C_x.

The most striking observation from this simulation is the relative efficiencies of v_1 and v_2 observed above for the different weighting combinations: when more (less) weight is placed on the double sample, v_1 (v_2) is better. This observation warrants further investigation.

Table 5.8: Relative efficiency: $w_1 = 0.8, w_2 = 0.2$

	$\frac{MSE(v_1)}{MSE(v_0)}$		$\frac{MSE(v_2)}{MSE(v_0)}$		$\frac{MSE(v_{L,I})}{MSE(v_0)}$		$\frac{MSE(v_J)}{MSE(v_0)}$	
$C_x =$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$
$\rho =$	(i) $y = x + \epsilon$							
0.7	0.843	0.753	0.893	0.801	0.907	0.816	1.025	0.904
0.85	0.603	0.531	0.649	0.575	0.631	0.552	0.679	0.582
$\rho =$	(ii) $y = x + \sqrt{x}\epsilon$							
0.7	0.840	0.784	0.895	0.839	1.041	1.021	1.270	1.329
0.85	0.667	0.619	0.722	0.674	0.768	0.722	0.882	0.869
$\rho =$	(iii) $y = 0.1x^2 + \epsilon$							
0.7	0.677	0.623	0.741	0.687	0.670	0.562	1.426	1.477
0.85	0.664	0.619	0.729	0.683	0.676	0.559	1.409	1.458
$\rho =$	(iv) $y = x + 0.1x^2 + \epsilon$							
0.7	0.640	0.601	0.704	0.664	0.646	0.539	1.243	1.350
0.85	0.620	0.592	0.685	0.657	0.616	0.529	1.216	1.325

5.2.1 Relative efficiency of $v_1(\bar{y}_w^{lr})$ and $v_2(\bar{y}_w^{lr})$

Recall $v_1(\bar{y}_w^{lr})$ and $v_2(\bar{y}_w^{lr})$ as given in Chapter 4. Ignoring finite population corrections, we have

$$v_1(\bar{y}_w^{lr}) = w_1^2 v_1(\bar{y}_{lr}) + w_2^2 v_3(\bar{y}_3) \tag{5.5}$$

and

$$v_2(\bar{y}_w^{lr}) = w_1^2 v_1(\bar{y}_{lr}) + w_2^2 v_{23}(\bar{y}_3), \tag{5.6}$$

where $v_1(\bar{y}_{lr}) = s_d^2/n_2 + b^2 s_{x_1}^2/n_1$, $v_3(\bar{y}_3) = s_{y_3}^2/n_3$, and $v_{23}(\bar{y}_3) = s_{y_{23}}^2/n_3$. The only difference being that in v_2 all of the y_i values are used in the second term.

We can decompose the MSE of the variance estimates v_1 and v_2 as follows,

$$\begin{aligned} MSE(v_1(\bar{y}_w^{lr})) &= E[w_1^2 v_1(\bar{y}_{lr}) + w_2^2 v_3(\bar{y}_3) - w_1^2 U_1 - w_2^2 U_2]^2 \\ &= w_1^4 MSE(v_1(\bar{y}_{lr})) + w_2^4 MSE(v_3(\bar{y}_3)) \\ &+ 2w_1^2 w_2^2 E[(v_1(\bar{y}_{lr}) - U_1)(v_3(\bar{y}_3) - U_2)], \end{aligned} \tag{5.7}$$

and

$$\begin{aligned} MSE(v_2(\bar{y}_w^{lr})) &= E[w_1^2 v_1(\bar{y}_{lr}) + w_2^2 v_{23}(\bar{y}_3) - w_1^2 U_1 - w_2^2 U_2]^2 \\ &= w_1^4 MSE(v_1(\bar{y}_{lr})) + w_2^4 MSE(v_{23}(\bar{y}_3)) \\ &+ 2w_1^2 w_2^2 E[(v_1(\bar{y}_{lr}) - U_1)(v_{23}(\bar{y}_3) - U_2)], \end{aligned} \tag{5.8}$$

where $U_1 = MSE(\bar{y}_{lr}) = E(\bar{y}_{lr} - \bar{Y})^2$ and $U_2 = MSE(\bar{y}_3) = E(\bar{y}_3 - \bar{Y})^2$.

We treat the set s_3 as if it is a sample from the entire set S, but in the simulation we do not allow units in s_1 to be sampled again in s_3. Subsequently, $\bar{y}_{lr}$ and $\bar{y}_3$ are not exactly independent. However, since the sampling fraction in our simulation is approximately zero, they are approximately independent. Thus, we would anticipate the last term in (5.7) to be nearly zero. One cannot expect the same of (5.8), since s_2 is used in both $v_1(\bar{y}_{lr})$ and $v_{23}(\bar{y}_{lr})$. We would expect $MSE(v_3(\bar{y}_3))$ in (5.7) to be larger than $MSE(v_{23}(\bar{y}_3))$ in (5.8) due to the increased sample size used in the latter. The question then is whether the sum of the second and third terms in (5.8) is larger or smaller than the second term of (5.7).

Recall to find the "true" MSE of $\bar{y}_w^{lr}$ via simulation we used

$$MSE = \frac{1}{B}\sum_{b=1}^{B}(\bar{y}_{w(b)}^{lr} - \bar{Y})^2,$$

where $\bar{y}_{w(b)}^{lr}$ is the estimate $\bar{y}_w^{lr}$ obtained on the b^{th} simulation run. Note that this "true" MSE does in fact capture any covariance between $\bar{y}_{lr}$ and $\bar{y}_3$. We then use this simulation estimate of MSE to calculate $MSE(v_1(\bar{y}_w^{lr}))$ and $MSE(v_2(\bar{y}_w^{lr}))$, see (5.3). However, when we calculate $MSE(v_1(\bar{y}_w^{lr}))$ using the expansion given in (5.7), the covariance between $\bar{y}_{lr}$ and $\bar{y}_3$ is not included in the MSE estimate. Therefore, the two different derivations of $MSE(v_1(\bar{y}_w^{lr}))$, using (5.3) and (5.7), will yield slightly different results. The same is true for $MSE(v_2(\bar{y}_w^{lr}))$. The difference, however, is negligible since $\bar{y}_{lr}$ and $\bar{y}_3$ are approximately independent as explained above.

We use simulation results and (5.3) to calculate $MSE(v_1)$ and $MSE(v_2)$ by (5.3) as explained in section 5.1. We can also empirically calculate the three terms on the right-hand of expressions (5.7) and (5.8) using simulation values. We use

$$\hat{U}_1 = \frac{1}{B}\sum_{i=1}^{B}(\bar{y}_{lr(b)} - \bar{Y})^2, \qquad \hat{U}_2 = \frac{1}{B}\sum_{i=1}^{B}(\bar{y}_{3(b)} - \bar{Y})^2$$

$$\hat{MSE}(v_1(\bar{y}_{lr})) = \frac{1}{B}\sum_{i=1}^{B}(v_1(\bar{y}_{lr})_{(b)} - \hat{U}_1)^2,$$

$$\hat{MSE}(v_3(\bar{y}_3)) = \frac{1}{B}\sum_{i=1}^{B}(v_3(\bar{y}_3)_{(b)} - \hat{U}_2)^2,$$

$$\hat{MSE}(v_{23}(\bar{y}_3)) = \frac{1}{B}\sum_{i=1}^{B}(v_{23}(\bar{y}_3)_{(b)} - \hat{U}_2)^2.$$

$$\hat{E}[(v_1(\bar{y}_{lr}) - U_1)(v_3(\bar{y}_3) - U_2)] = \frac{1}{B}\sum_{i=1}^{B}(v_1(\bar{y}_{lr})_{(b)} - \hat{U}_1)(v_3(\bar{y}_3)_{(b)} - \hat{U}_2),$$

and

$$\hat{E}[(v_1(\bar{y}_{lr}) - U_1)(v_{23}(\bar{y}_3) - U_2)] = \frac{1}{B}\sum_{i=1}^{B}(v_1(\bar{y}_{lr})_{(b)} - \hat{U}_1)(v_{23}(\bar{y}_3)_{(b)} - \hat{U}_2)$$

where (b) indicates the value was obtained from the b^{th} simulation run. Since $w_2 = 1 - w_1$, we can use simulation results and write $MSE(v_1(\bar{y}_w^{lr}))$ and $MSE(v_2(\bar{y}_w^{lr}))$ as functions of w_1. We may then plot $MSE(v_1(\bar{y}_w^{lr}))$ and $MSE(v_2(\bar{y}_w^{lr}))$ over $0 < w_1 < 1$. Figures 5.1-5.4 show the results for each model given in Table 5.2 at each combination of ρ and C_x used in the simulation.

Figures 5.2 - 5.5 reveal that $v_2(\bar{y}_w^{lr})$ appears to be more efficient than $v_1(\bar{y}_w^{lr})$ for w_1 less than 0.6, and nearly as efficient for w_1 greater than 0.6. This indicates that the gain in precision by using $v_{23}(\bar{y}_3)$ to estimate S_y^2 in $v_2(\bar{y}_w^{lr})$ outweighs the ***penalty*** introduced by the covariance term between $v_1(\bar{y}_{lr})$ and $v_{23}(\bar{y}_3)$. We note that both $v_1(\bar{y}_w^{lr})$ and $v_2(\bar{y}_w^{lr})$ obtain their minimum MSE in all models in the neighbourhood of $w_1 = 0.6$. This suggests the optimal weighting for the sample size under investigation is $w_1 = 0.6, w_2 = 0.4$.

Figure 5.2: Model i): $MSE(v_1)$ and $MSE(v_2)$ as a function of w_1

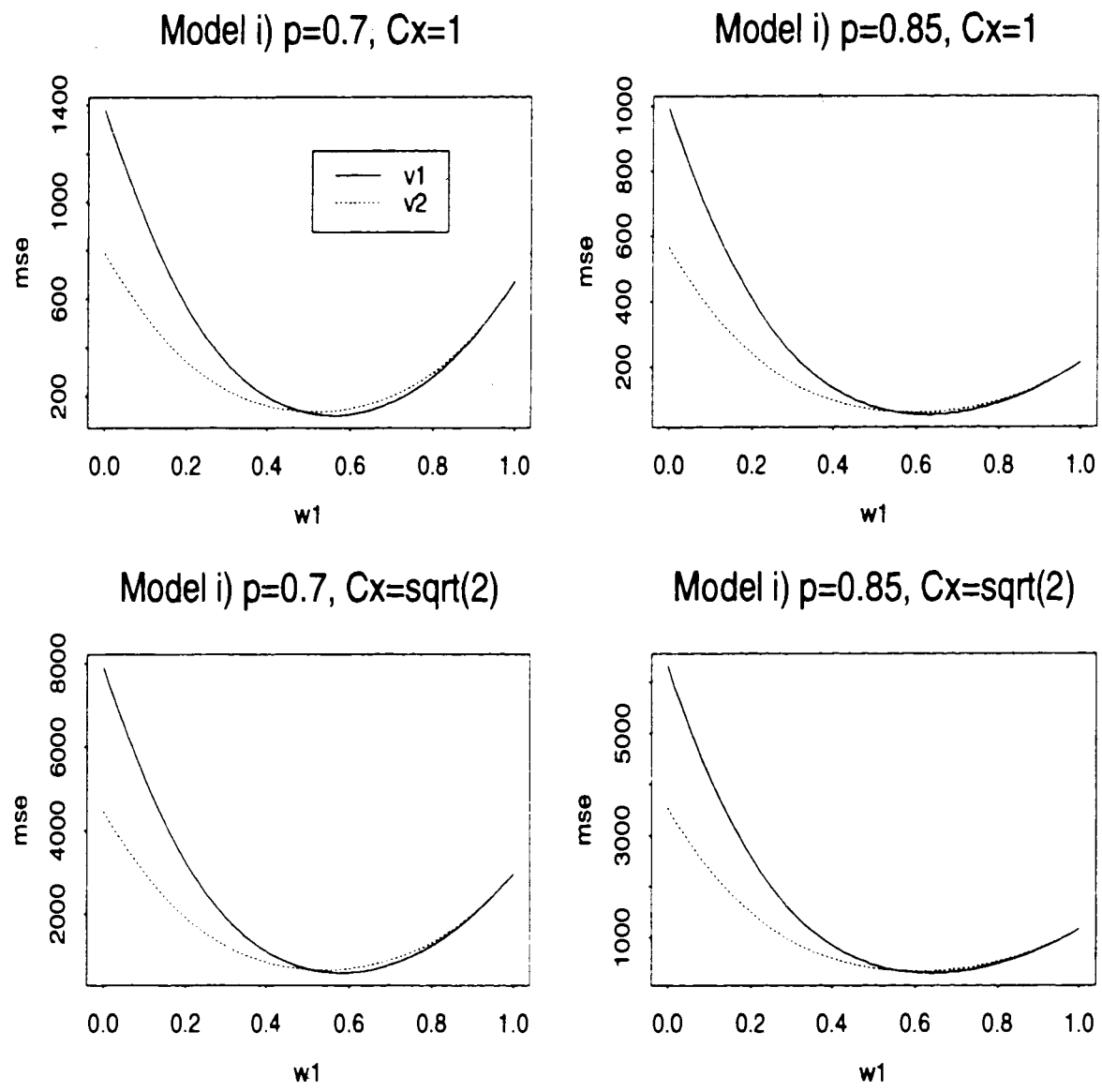


Figure 5.3: Model ii): $MSE(v_1)$ and $MSE(v_2)$ as a function of w_1

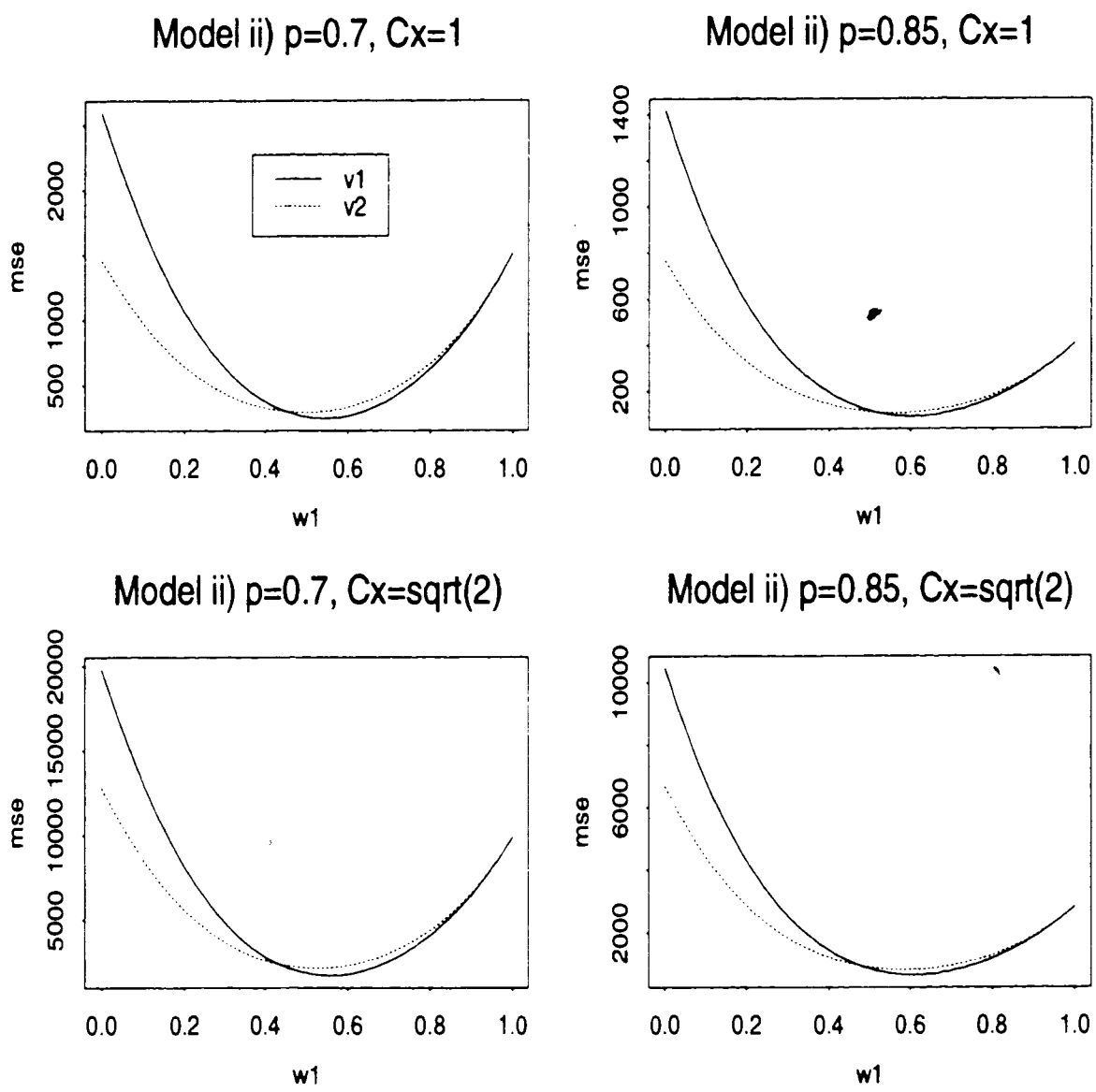


Figure 5.4: Model iii): $MSE(v_1)$ and $MSE(v_2)$ as a function of w_1

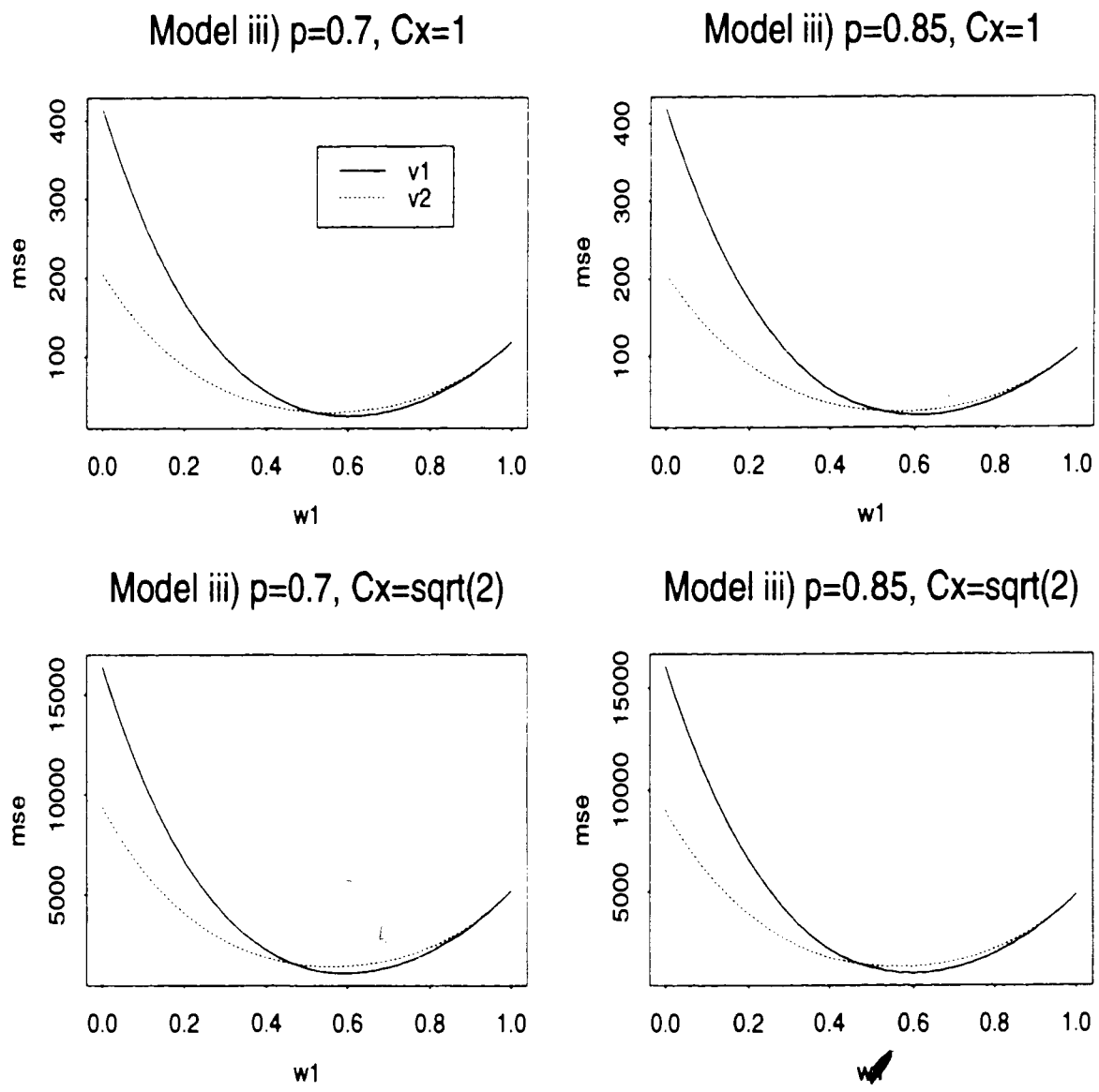
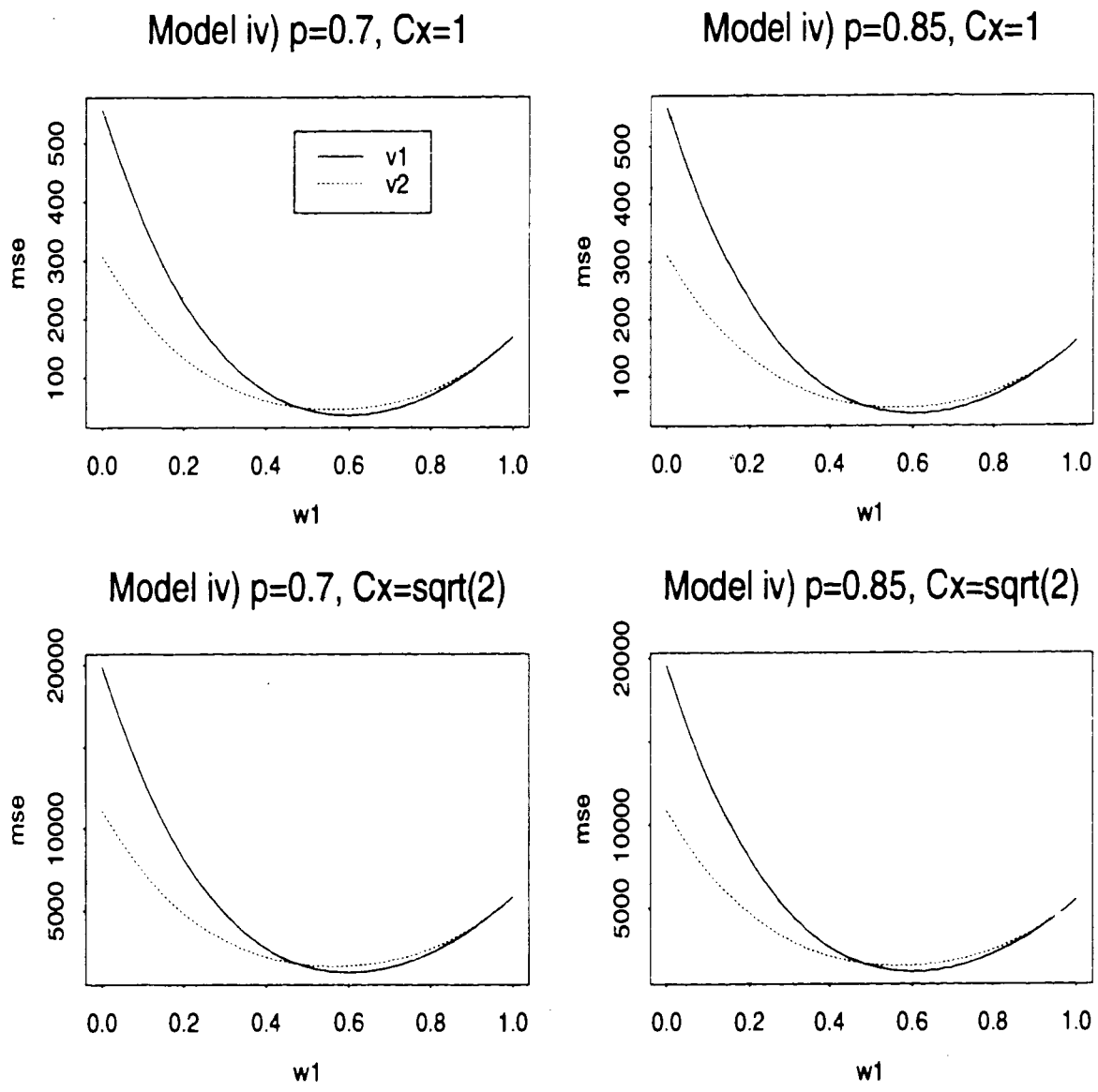


Figure 5.5: Model iv): $MSE(v_1)$ and $MSE(v_2)$ as a function of w_1



5.3 Estimated weight simulation

5.3.1 The point estimator $\bar{y}_w^{lr}$

To investigate variance estimation for $\bar{y}_w^{lr}$ with estimated weights, we first need to discuss the point estimator itself. Consider the weights given in (3.4). We have a variety of choices to estimate $V(\bar{y}_{lr})$ and two possible estimates of $V(\bar{y}_3)$. In Chapter 2 we introduced four estimates of $V(\bar{y}_{lr})$: $v_0(\bar{y}_{lr})$, $v_1(\bar{y}_{lr})$, $v_{LJ}(\bar{y}_{lr})$ and $v_J(\bar{y}_{lr})$, given in equations (2.9), (2.11), (2.12) and (2.13) respectively. Possible estimates for $V(\bar{y}_3)$ are $v_3(\bar{y}_3) = s_{y_3}^2/n_3$ and $v_{23}(\bar{y}_3) = s_{y_{23}}^2/n_3$. We wish to determine which combination of estimates used to calculate $\hat{w}_1$ and $\hat{w}_2$ yields the best estimate $\bar{y}_w^{lr}$. We have 8 such combinations to consider, and thus 8 possible point estimates, see Table 5.9.

Table 5.9: Estimates of $V(\bar{y}_{lr})$ and $V(\bar{y}_3)$

No.	pair	$v(\bar{y}_{lr})$	$v(\bar{y}_3)$
1	$(0,3)$	$v_0(\bar{y}_{lr})$	$v_3(\bar{y}_3)$
2	$(0,23)$	$v_0(\bar{y}_{lr})$	$v_{23}(\bar{y}_3)$
3	$(1,3)$	$v_1(\bar{y}_{lr})$	$v_3(\bar{y}_3)$
4	$(1,23)$	$v_1(\bar{y}_{lr})$	$v_{23}(\bar{y}_3)$
5	$(LJ,3)$	$v_{LJ}(\bar{y}_{lr})$	$v_3(\bar{y}_3)$
6	$(LJ,23)$	$v_{LJ}(\bar{y}_{lr})$	$v_{23}(\bar{y}_3)$
7	$(J,3)$	$v_J(\bar{y}_{lr})$	$v_3(\bar{y}_3)$
8	$(J,23)$	$v_J(\bar{y}_{lr})$	$v_{23}(\bar{y}_3)$

We employ the same models, and ρ and C_x values as used in the fixed weight simulation. To investigate the *best* point estimate, we use (5.2) to calculate the simulated mean square error for each estimate $\bar{y}_w^{lr}$ for the 8 possible weighting combinations. We report the percent relative bias of the point estimates, see Table 5.10. We estimate this quantity using,

$$rbias(\bar{y}_w^{lr}) = 100 \times \frac{1}{B}\sum_{b=1}^{B}(\bar{y}_w^{lr(b)} - \bar{Y})/\bar{Y}, \tag{5.9}$$

where (b) indicates the value obtained from the b^{th} simulation run. We also report relative mean squared error in Table 5.11 using combination (1) in Table 5.9 as the standard. We indicate the choice of $v(\bar{y}_{lr})$ and $v(\bar{y}_3)$ used to estimate $\hat{w}_1$ and $\hat{w}_2$ in Tables 5.10 and 5.11 by the pairs in brackets as given in Table 5.9.

Table 5.10 indicates that the bias increases with both ρ and C_x. For models (i) and (ii) all point estimates considered have similar and small percent relative bias. Results for the non-linear models, (iii) and (iv), display larger bias. We note that for these two models bias appears to be reduced by using $v_{23}(\bar{y}_3)$ to estimate $V(\bar{y}_3)$. In general, our results suggest the point estimate $\bar{y}_w^{lr(0,23)}$ performs best with respect to bias.

Turning to relative efficiency of the point estimates, see Table 5.11, we find for models (i) and (ii) estimates perform very similarly in terms of efficiency. For models (iii) and (iv) estimates which use $v_{23}(\bar{y}_3)$ as opposed to $v_3(\bar{y}_3)$ are slightly more efficient.

5.3.2 Variance estimator results

Recall using a one term Taylor series expansion with estimated weights, see Appendix A, we found the linearisation estimator

$$v(\bar{y}_w^{lr}) = \frac{1}{v(\bar{y}_{lr})^{-1} + v(\bar{y}_3)^{-1}}. \tag{5.10}$$

We also discussed Meier's variance estimator in Section 3.2 for estimated weights. He used a two term Taylor series expansion and made a number of assumptions in developing,

$$v_M(\bar{y}_w^{lr}) = \frac{1}{v(\bar{y}_{lr})^{-1} + v(\bar{y}_3)^{-1}} \left[1 + 4\hat{w}_1\hat{w}_2\left(\frac{1}{m_1} + \frac{1}{m_2}\right)\right] \tag{5.11}$$

where $m_1 = n_2 - 1$, $m_2 = n_3 - 1$. We consider both of the above variance estimators and explore all 8 possible estimates for $v(\bar{y}_{lr})$ and $v(\bar{y}_3)$ given in Table 5.9. That is, we consider 16 variance estimators in total. When computing the efficiency of these variance estimators using (5.3), we first need an estimate of the true MSE of $\bar{y}_w^{lr}$ using (5.2). We find the MSE of $\bar{y}_w^{lr}$ which uses the same estimates of $v(\bar{y}_{lr})$ and $v(\bar{y}_3)$ in $\hat{w}_1$ and $\hat{w}_2$ as does the variance estimate in question, $v(\bar{y}_w^{lr})$ or $v_M(\bar{y}_w^{lr})$. It is this MSE that is used to calculate the mean squared error of the variance estimator using equation (5.3). For example, to compute the

		(i) $y = x + \epsilon$		(ii) $y = x + \sqrt{x}\epsilon$		(iii) $y = 0.1x^2 + \epsilon$		(iv) $y = x + 0.1x^2 + \epsilon$	
	$\rho =$	0.7	0.85	0.7	0.85	0.7	0.85	0.7	0.85
$rbias((\bar{y}_w^{lr})^{(0,3)})$	1.0	-0.14	-0.33	-0.73	-0.67	-3.50	-5.10	-2.41	-3.83
	$\sqrt{2}$	-0.52	-0.86	-1.60	-1.49	-7.03	-10.2	-5.28	-7.72
$rbias((\bar{y}_w^{lr})^{(0,23)})$	1.0	0.00	0.00	0.00	0.00	-0.02	0.00	-0.01	-0.02
	$\sqrt{2}$	0.00	0.00	-0.01	-0.01	-0.05	-0.07	0.00	-0.05
$rbias((\bar{y}_w^{lr})^{(1,3)})$	1.0	-0.15	-0.34	-0.75	-0.69	-3.41	-5.00	-2.36	-3.75
	$\sqrt{2}$	-0.53	-0.88	-1.65	-1.54	-6.92	-10.1	-5.19	-7.64
$rbias((\bar{y}_w^{lr})^{(1,23)})$	1.0	0.00	-0.11	-0.29	-0.30	-2.06	-3.10	-1.39	-2.32
	$\sqrt{2}$	-0.22	-0.43	-0.75	-0.76	-4.43	-6.57	-3.28	-4.90
$rbias((\bar{y}_w^{lr})^{(LJ,3)})$	1.0	-0.14	-0.34	-0.74	-0.69	-3.40	-5.02	-2.35	-3.76
	$\sqrt{2}$	-0.54	-0.88	-1.65	-1.54	-6.88	-10.1	-5.15	-7.64
$rbias((\bar{y}_w^{lr})^{(LJ,23)})$	1.0	0.00	-0.11	-0.29	-0.30	-2.05	-3.10	-1.38	-2.32
	$\sqrt{2}$	-0.22	-0.43	-0.75	-0.77	-4.35	-6.51	-3.22	-4.85
$rbias((\bar{y}_w^{lr})^{(J,3)})$	1.0	-0.14	-0.34	-0.74	-0.69	-3.49	-5.13	-2.41	-3.85
	$\sqrt{2}$	-0.54	-0.88	-1.64	-1.54	-7.08	-10.3	-5.31	-7.80
$rbias((\bar{y}_w^{lr})^{(J,23)})$	1.0	0.00	-0.11	-0.29	-0.29	-2.16	-3.26	-1.45	-2.43
	$\sqrt{2}$	-0.22	-0.43	-0.73	-0.76	-4.65	-6.89	-3.43	-5.14

Table 5.10: Percent relative bias of point estimates $\bar{y}_w^{lr}$

Table 5.11: Relative efficiency of point estimates $\bar{y}_w^{lr}$

	$\frac{MSE(\bar{y}_w^{lr})^{(0,23)}}{MSE(\bar{y}_w^{lr})^{(0,3)}}$		$\frac{MSE(\bar{y}_w^{lr})^{(1,3)}}{MSE(\bar{y}_w^{lr})^{(0,3)}}$		$\frac{MSE(\bar{y}_w^{lr})^{(1,23)}}{MSE(\bar{y}_w^{lr})^{(0,3)}}$		$\frac{MSE(\bar{y}_w^{lr})^{(LJ,3)}}{MSE(\bar{y}_w^{lr})^{(0,3)}}$		$\frac{MSE(\bar{y}_w^{lr})^{(LJ,23)}}{MSE(\bar{y}_w^{lr})^{(0,3)}}$		$\frac{MSE(\bar{y}_w^{lr})^{(J,3)}}{MSE(\bar{y}_w^{lr})^{(0,3)}}$		$\frac{MSE(\bar{y}_w^{lr})^{(J,23)}}{MSE(\bar{y}_w^{lr})^{(0,3)}}$	
$C_x =$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	$C_x =$	1.0	$C_x =$	1.0	$C_x =$	1.0
						(i) $y = x + \epsilon$								
$\rho =$														
0.7	0.991	0.988	0.999	0.997	0.991	0.987	0.999	0.997	0.991	0.988	0.999	0.997	0.992	0.988
0.85	0.984	0.973	0.996	0.994	0.984	0.974	0.996	0.994	0.984	0.974	0.996	0.994	0.984	0.974
						(ii) $y = x + \sqrt{x}\epsilon$								
$\rho =$														
0.7	0.975	0.958	0.996	0.993	0.972	0.953	0.991	0.986	0.970	0.952	0.992	0.988	0.970	0.952
0.85	0.968	0.946	0.993	0.990	0.965	0.942	0.991	0.988	0.964	0.943	0.991	0.990	0.965	0.944
						(iii) $y = 0.1x^2 + \epsilon$								
$\rho =$														
0.7	0.965	0.946	0.996	0.993	0.967	0.949	0.991	0.984	0.971	0.953	0.997	0.995	0.968	0.947
0.85	0.907	0.866	0.989	0.990	0.906	0.865	0.980	0.979	0.910	0.871	0.992	0.997	0.905	0.864
						(iv) $y = x + 0.1x^2 + \epsilon$								
$\rho =$														
0.7	0.967	0.952	0.995	0.993	0.969	0.955	0.992	0.985	0.972	0.959	0.997	0.994	0.971	0.954
0.85	0.900	0.875	0.985	0.988	0.895	0.874	0.978	0.978	0.899	0.880	0.989	0.995	0.896	0.875

efficiency of $v(\bar{y}_w^{lr})$ which uses combination (1) in Table 5.9, we would first calculate the MSE of the point estimator $\bar{y}_w^{lr} = \hat{w}_1\bar{y}_{lr} + \hat{w}_2\bar{y}_3$ where $\hat{w}_1 = v_0(\bar{y}_{lr})^{-1}/(v_0(\bar{y}_{lr})^{-1} + v_3(\bar{y}_3)^{-1})$ and $\hat{w}_2 = v_3(\bar{y}_3)^{-1}/(v_0(\bar{y}_{lr})^{-1} + v_3(\bar{y}_3)^{-1})$. We would then use this value in equation (5.3) to find $MSE(v(\bar{y}_w^{lr}))^{(0,3)}$.

We report percent relative bias, calculated using equation (5.4), in Tables 5.12 and 5.13. The first table contains results using the linearisation estimator, equation (5.10), and the second contains results using Meier's variance estimator, equation (5.11). Again, we indicate the estimates used for $v(\bar{y}_{lr})$ and $v(\bar{y}_3)$ by pairs in brackets.

We first note that Meier's variance estimators have similar bias to that of the linearisation variance estimators, see Tables 5.12 and 5.13. Meier's correction factor typically increases the percent relative bias by approximately +2%. Variance estimators appear to have similar and small bias for model (i). For models (ii), (iii) and (iv) we find variance estimators which use $v_{23}(\bar{y}_3)$ as opposed to $v_3(\bar{y}_3)$ have smaller relative bias. For non-linear models (iii) and (iv), we find larger negative biases for all variance estimators. The problem worsens as both ρ and C_x increase. For example, for models (iii) and (iv) variance estimators are seriously underestimating the true variance for $\rho = 0.85$ and $C_x = \sqrt{2}$. As we noted in the fixed weight simulation, bias is reduced by increasing the sample sizes, n_1, n_2 and n_3. We found doubling the sample sizes decreased the biases of the variance estimates by approximately one third in models (iii) and (iv). Generally, our simulation results suggest that $v^{(0,23)}$ performs best in terms of bias for the models under investigation.

Tables 5.14 and 5.15 present relative efficiencies of the linearisation and Meier's variance estimators respectively. In both tables, relative efficiencies are reported with respect to the linearisation variance estimator which estimates $V(\bar{y}_{lr})$ and $V(\bar{y}_3)$ with $v_0(\bar{y}_{lr})$ and $v_3(\bar{y}_3)$ respectively. We note the same patterns exist within each table. However, comparing one table to the other, we see the linearisation estimators are slightly more efficient than Meier's estimators. For these reasons, we limit our discussion to the results contained in Table 5.14. Comparing $v^{(1,3)}$ and $v^{(1,23)}$, $v^{(LJ,3)}$ and $v^{(LJ,23)}$, and $v^{(J,3)}$ and $v^{(J,23)}$, we see that estimators which use $v_3(\bar{y}_3)$ are more efficient than those which use $v_{23}(\bar{y}_3)$ for all models especially for (iii) and (iv). This suggests a covariance term has been introduced by using $v_{23}(\bar{y}_3)$ in these

estimators. For models (i) through (iv), $v^{(1,3)}$ and $v^{(LJ,3)}$ are most efficient. The jackknife variance estimator of $V(\bar{y}_{lr})$ also appears to give stable estimates.

		(i) $y = x + \epsilon$		(ii) $y = x + \sqrt{x}\epsilon$		(iii) $y = 0.1x^2 + \epsilon$		(iv) $y = x + 0.1x^2 + \epsilon$	
	$\rho =$	0.7	0.85	0.7	0.85	0.7	0.85	0.7	0.85
$rbias(v^{(0,3)})$	1.0	0.71	-0.45	-4.39	-4.87	-8.60	-26.33	-6.97	-27.90
	$\sqrt{2}$	-1.83	-4.78	-10.31	-12.06	-15.34	-43.21	-12.40	-37.44
$rbias(v^{(0,23)})$	1.0	2.53	2.73	-0.04	0.50	1.46	-4.59	1.83	-6.39
	$\sqrt{2}$	0.58	0.11	-3.23	-3.37	1.12	-11.27	1.49	-8.32
$rbias(v^{(1,3)})$	1.0	1.02	0.45	-3.66	-3.54	-8.48	-25.47	-6.73	-26.35
	$\sqrt{2}$	-1.27	-3.46	-9.15	-10.22	-15.22	-42.68	-12.18	-36.59
$rbias(v^{(1,23)})$	1.0	2.50	2.73	0.45	1.01	0.18	-6.04	0.66	-4.45
	$\sqrt{2}$	0.57	0.02	-2.62	-2.75	-1.10	-14.06	-0.50	-10.92
$rbias(v^{(LJ,3)})$	1.0	1.39	0.74	-2.28	-2.56	-8.31	-25.44	-6.64	-25.59
	$\sqrt{2}$	-0.85	-3.16	-7.33	-9.08	-15.14	-43.37	-12.12	-37.25
$rbias(v^{(LJ,23)})$	1.0	2.86	3.01	1.30	1.58	-1.44	-9.48	-0.66	-10.70
	$\sqrt{2}$	0.95	0.30	-1.84	-2.43	-4.47	-20.72	-3.20	-16.82
$rbias(v^{(J,3)})$	1.0	2.25	1.38	-0.70	-1.42	-5.93	-22.5	-4.44	-23.79
	$\sqrt{2}$	0.00	-2.54	-5.23	-7.60	-12.28	-39.55	-9.49	-33.77
$rbias(v^{(J,23)})$	1.0	3.72	3.65	2.93	2.79	2.75	-2.72	2.95	-4.36
	$\sqrt{2}$	1.81	0.94	0.49	-0.70	2.28	-10.13	2.53	-7.40

Table 5.12: Percent relative bias of linearisation variance estimates

Table 5.13: Percent relative bias of Meier's variance estimates

		(i) $y = x + \epsilon$		(ii) $y = x + \sqrt{x}\epsilon$		(iii) $y = 0.1x^2 + \epsilon$		(iv) $y = x + 0.1x^2 + \epsilon$	
	$\rho =$	0.7	0.85	0.7	0.85	0.7	0.85	0.7	0.85
$rbias(v_M^{(0,3)})$	1.0	2.96	1.71	-2.29	-2.82	-6.68	-24.93	-5.00	-26.55
	$\sqrt{2}$	0.35	-2.73	-8.37	-10.20	-13.64	-42.22	-10.61	-36.32
$rbias(v_M^{(0,23)})$	1.0	4.85	5.00	2.29	2.70	3.70	-2.57	4.09	-4.45
	$\sqrt{2}$	2.85	2.32	-1.07	-1.26	3.34	-9.43	3.72	-6.41
$rbias(v_M^{(1,3)})$	1.0	3.29	2.64	-1.53	-1.46	-6.55	-24.03	-4.74	-24.96
	$\sqrt{2}$	0.93	-1.37	-7.17	-8.31	-13.50	-41.67	-10.37	-35.44
$rbias(v_M^{(1,23)})$	1.0	4.82	4.99	2.71	3.23	2.40	-4.05	2.89	-5.38
	$\sqrt{2}$	2.83	2.21	-0.44	-0.63	1.06	-12.28	1.69	-9.08
$rbias(v_M^{(LJ,3)})$	1.0	3.66	2.94	-0.12	-0.45	-6.37	-24.01	-4.65	-25.20
	$\sqrt{2}$	1.36	-1.05	-5.31	-7.15	-13.41	-43.37	-10.30	-36.10
$rbias(v_M^{(LJ,23)})$	1.0	5.18	5.28	3.57	3.81	0.72	-7.60	1.53	-8.87
	$\sqrt{2}$	3.23	0.30	0.34	-0.70	-2.41	-19.15	-1.10	-15.16
$rbias(v_M^{(J,3)})$	1.0	4.54	3.60	1.46	0.72	-3.95	-20.97	-2.41	-22.34
	$\sqrt{2}$	2.23	-0.42	-3.17	-5.62	-10.51	-38.48	-7.63	-32.57
$rbias(v_M^{(J,23)})$	1.0	6.07	5.94	5.23	5.05	5.01	-0.694	5.23	-2.39
	$\sqrt{2}$	4.11	3.16	2.72	1.47	4.48	-8.32	4.76	-5.52

Table 5.14: Relative efficiency of linearisation variance estimates

	$\frac{MSE(v^{(0,2)})}{MSE(v^{(0,3)})}$			$\frac{MSE(v^{(1,3)})}{MSE(v^{(0,3)})}$			$\frac{MSE(v^{(1,2)})}{MSE(v^{(0,3)})}$			$\frac{MSE(v^{(LJ,3)})}{MSE(v^{(0,3)})}$			$\frac{MSE(v^{(LJ,2)})}{MSE(v^{(0,3)})}$			$\frac{MSE(v^{(J,3)})}{MSE(v^{(0,3)})}$			$\frac{MSE(v^{(J,2)})}{MSE(v^{(0,3)})}$		
$C_x =$	1.0	$\sqrt{2}$		1.0	$\sqrt{2}$		1.0	$\sqrt{2}$		1.0	$\sqrt{2}$		$C_x =$	1.0		$C_x =$	1.0		$C_x =$	1.0	
$\rho =$	(i) $y = x + \epsilon$																				
0.7	1.282	1.268		0.921	0.883		1.059	0.956		0.954	0.903		1.094	0.970		0.993	0.919		1.138	0.993	
0.85	1.390	1.341		0.767	0.729		0.883	0.776		0.785	0.734		0.904	0.780		0.798	0.732		0.922	0.785	
$\rho =$	(ii) $y = x + \sqrt{x}\epsilon$																				
0.7	1.210	1.120		0.898	0.874		1.043	0.919		0.919	0.859		0.977	0.808		0.960	0.879		1.049	0.880	
0.85	1.272	1.151		0.773	0.766		0.897	0.776		0.793	0.758		0.854	0.701		0.815	0.762		0.899	0.747	
$\rho =$	(iii) $y = 0.1x^2 + \epsilon$																				
0.7	1.999	2.389		0.879	0.864		1.623	1.870		0.837	0.764		1.201	1.312		0.982	0.945		1.786	2.031	
0.85	1.272	1.737		0.773	0.872		0.897	1.352		0.792	0.796		0.854	0.959		0.815	0.906		0.899	1.429	
$\rho =$	(iv) $y = x + 0.1x^2 + \epsilon$																				
0.7	1.893	2.281		0.852	0.849		1.477	1.743		0.818	0.759		1.210	1.253		0.951	0.928		1.630	1.904	
0.85	1.613	1.766		0.800	0.841		1.208	1.326		0.748	0.766		0.951	0.950		0.837	0.879		1.292	1.416	

Table 5.15: Relative efficiency of Meier's variance estimates

	$\frac{MSE(v_M^{(0,23)})}{MSE(v^{(0,3)})}$		$\frac{MSE(v_M^{(1,3)})}{MSE(v(0,3))}$		$\frac{MSE(v_M^{(1,23)})}{MSE(v(0,3))}$		$\frac{MSE(v_M^{(LJ,3)})}{MSE(v(0,3))}$		$\frac{MSE(v_M^{(LJ,23)})}{MSE(v(0,3))}$		$\frac{MSE(v_M^{(J,3)})}{MSE(v(0,3))}$		$\frac{MSE(v_M^{(J,23)})}{MSE(v(0,3))}$	
$C_x =$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	1.0	$\sqrt{2}$	$C_x =$	1.0	$C_x =$	1.0	$C_x =$	1.0
$\rho=$	(i) $y = x + \epsilon$													
0.7	1.452	1.367	1.028	0.918	1.218	1.040	1.073	0.949	1.266	1.065	1.141	0.986	1.338	1.109
0.85	1.53	1.419	0.830	0.725	0.998	0.825	0.854	0.734	1.027	0.832	0.882	0.740	1.058	0.846
$\rho=$	(ii) $y = x + \sqrt{x}\epsilon$													
0.7	1.283	1.155	0.900	0.845	1.114	0.949	0.943	0.845	1.057	0.839	1.009	0.883	1.155	0.931
0.85	1.352	1.189	0.775	0.735	0.964	0.800	0.808	0.734	0.926	0.724	0.846	0.749	0.988	0.783
$\rho=$	(iii) $y = 0.1x^2 + \epsilon$													
0.7	2.103	2.502	0.872	0.854	1.703	1.954	0.829	0.750	1.354	1.357	0.992	0.945	1.881	2.124
0.85	1.352	1.809	0.775	0.868	0.964	1.404	0.808	0.789	0.927	0.984	0.846	0.906	0.988	1.482
$\rho=$	(iv) $y = x + 0.1x^2 + \epsilon$													
0.7	1.997	2.393	0.848	0.838	1.554	1.823	0.812	0.745	1.264	1.298	0.965	0.930	1.725	1.997
0.85	1.680	1.840	0.790	0.835	1.254	1.377	0.736	0.757	0.979	0.975	0.832	0.877	1.344	1.470

Chapter 6

Discussion

We have explored variance estimation for the typical point estimator generated under sampling on two occasions. Since this estimator is a weighted linear combination of two estimators, we consider two cases: fixed weights and estimated weights.

Under the fixed weight supposition, we explored three weighting combinations. As we increased the weight on the linear regression estimator from the double sample, we saw changes in the percent relative biases and relative efficiencies of the variance estimators in question. We found for the non-linear populations with $w_1 = 0.8$ and $w_2 = 0.2$ percent relative bias increased with both ρ and C_x. An increase in size of the sets s_1, s_2 and s_3 is necessary in order to reduce the bias of the variance estimates for these non-linear populations. In terms of efficiency, we discovered when more weight is placed on the double sample, v_1 performs better, whereas when less weight is placed on the double sample, v_2 performs better. By breaking down the MSE of these two estimates we were able to plot the simulated MSE of v_1 and v_2 over $0 < w_1 < 1$. We found $v_2(\bar{y}_w^{lr})$ to be more efficient than $v_1(\bar{y}_w^{lr})$ for $w_1 < 0.6$ and nearly as efficient for $w_1 > 0.6$. This finding suggests v_2's gain in efficiency by using $v_{23}(\bar{y}_3)$, as opposed to $v_3(\bar{y}_3)$, to estimate $V(\bar{y}_3)$ outweighs the *penalty* introduced by the covariance term between $v_1(\bar{y}_{lr})$ and $v_{23}(\bar{y}_3)$.

Turning to variance estimation where the point estimator has estimated weights, we first discuss the *best* choice of variance estimates to use in $\hat{w}_1$ and $\hat{w}_2$. We investigated 8 possible combinations of $\hat{w}_1$ and $\hat{w}_2$ to use in $\bar{y}_w^{lr}$. For models (i) and (ii) we found in terms of

both bias and efficiency all point estimates to be virtually equivalent, with $\bar{y}_w^{lr(0,23)}$ being a slight winner. For the non-linear models, (iii) and (iv), we found bias was smaller for those estimators which used $v_{23}(\bar{y}_3)$. Overall, the estimator $\bar{y}_w^{lr(0,23)}$ had smallest bias for models (i) through (iv).

Turning to estimators of the variance of $\bar{y}_w^{lr}$, we considered two general forms: the one-term Taylor series expansion linearisation variance estimator, and Meier's variance estimator. It is the latter estimator which is typically used by those in the forest industry, and claims to have a correction factor to reduce the order of the bias. We tried the same 8 combinations of estimates for $V(\bar{y}_{lr})$ and $V(\bar{y}_3)$ as used in the point estimator simulation in both the linearisation and Meier's variance estimator. We found for all models considered, Meier's correction factor affected the bias very little. Moreover, we found the linearisation variance estimator to be more efficient than Meier's estimator. On the basis of this study we would not recommend using Meier's variance estimator. It is based on parametric assumptions, is more complicated and does not seem to enhance performance.

As in the fixed weight simulation, we saw problems with percent relative bias for all variance estimates for non-linear populations. In such populations we found the bias increased with both ρ and C_x. Again, we found increasing n_1, n_2 and n_3 reduced the bias of our estimates. This suggests as populations become *more* non-linear, larger sample sizes are required in order to give unbiasedness (or near unbiasedness). In general, it appears bias is reduced in the linearisation estimators (and Meier's estimators) by using $v_{23}(\bar{y}_3)$ as opposed to $v_3(\bar{y}_3)$. However, estimators which use $v_{23}(\bar{y}_3)$ are less efficient. We found the linearisation variance estimator with $v_1(\bar{y}_{lr})$ or $v_{LJ}(\bar{y}_{lr})$ estimating $V(\bar{y}_{lr})$ and $v_3(\bar{y}_3)$ estimating $V(\bar{y}_3)$ yield the best variance estimates in terms of efficiency. However, taking bias into consideration $v_{23}(\bar{y}_3)$ is the preferable estimator of $V(\bar{y}_3)$.

We must consider the results from both the point estimator and the variance estimator simulation studies when recommending which estimates of $V(\bar{y}_{lr})$ and $V(\bar{y}_3)$ to use. Most important is the need to minimize the bias and MSE of the point estimator. Recall we found the point estimate $\bar{y}_w^{lr(0,23)}$ to give the best results in terms of bias. It also performed well in terms of efficiency. Next we consider the variance estimator results. We found estimating

$V(\bar{y}_3)$ with $v_{23}(\bar{y}_3)$ reduced bias in all models, at the price of lower efficiency. Based on these results, we suggest using $\bar{y}_w^{lr(0,23)}$ and $v^{(0,23)}$.

Bibliography

1. Cochran, W.G. (1977). *Sampling Techniques.* Third edition. John Wiley & Sons, New York.

2. Cochran, W.G. and Carroll, W.P. (1953). "A sampling investigation of the efficiency of weighting inversely as the estimated variance." *Biometrics*, 10, 447-459.

3. Dorfman, A.H. (1994). "A note on variance estimation for the regression estimator in double sampling." *Journal of the American Statistical Association*, 89, 137-140.

4. Meier, P. (1953). "Variance of a weighted mean." *Biometrics*, 9, 59-73.

5. Rao, J.N.K. and Sitter, R.R. (1997). "Variance estimation under stratified two-phase sampling with applications to measurement bias" in Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trein, D. (eds.), *Survey Measurement and Process Quality.* John Wiley & Sons, New York, Ch. 33.

6. Rao, J.N.K. and Sitter, R.R. (1995). "Variance estimation under two-phase sampling with application to imputation for missing data." *Biometrika*, 82, 453-460.

7. Schreuder, H.T., Li, H.G. and Scott, C.T. (1987). "Jackknife and bootstrap estimation for sampling with partial replacement." *Forest Science*, 33, 676-689.

8. Scott, C.T. (1984). "A new look at sampling with partial replacement." *Forest Science*, 30, 157-166.

9. Sen, A.R. (1973). "Theory and application of sampling on repeated occasions with several auxiliary variables." *Biometrics*, 29,381-385.

10. Sitter, R.R. (1997). "Variance estimation for the regression estimator in two-phase sampling." *Journal of the American Statistical Association*, 92, 780-787.

11. Sitter, R.R. and Rao, J.N.K. (1997). "Imputation for missing values and corresponding variance estimation." *The Canadian Journal of Statistics*, 25, 61-73.

12. Sukhatme, P.V. and Sukhatme, B.V. (1970). *Sampling Theory of Surveys with Applications*. Second Edition. Asia Publishing House, London.

13. Ware, K.D. and Cunia, T. (1952). "Continuous forest inventory with partial replacement of samples." *Forest Science Monograph*, 3, 40 p.

Appendix A

A.1 First order Taylor series expansion to get (3.8)

We have two estimates of $\bar{Y}$, $\bar{y}_{lr}$ and $\bar{y}_3$, with corresponding variances $\sigma_1^2 = V(\bar{y}_{lr})$ and $\sigma_2^2 = V(\bar{y}_3)$ respectively. We have consistent estimates of these variances $s_1^2 = v(\bar{y}_{lr})$ and $s_2^2 = v(\bar{y}_3)$. Then in this general notation the estimator of interest is

$$\bar{y}_w^{lr} = \frac{(s_1^2)^{-1}}{(s_1^2)^{-1} + (s_2^2)^{-1}} \bar{y}_{lr} + \frac{(s_2^2)^{-1}}{(s_1^2)^{-1} + (s_2^2)^{-1}} \bar{y}_3$$

Let $\hat{\boldsymbol{\theta}} = (\bar{y}_{lr}, \bar{y}_3, s_1^2, s_2^2)$ and $\boldsymbol{\theta} = (\bar{Y}, \bar{Y}, \sigma_1^2, \sigma_2^2)$. Then we may write the first order Taylor expansion

$$\bar{y}_w^{lr} = g(\hat{\boldsymbol{\theta}}) \doteq g(\boldsymbol{\theta}) + g'(\boldsymbol{\theta})\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

where $g'(\boldsymbol{\theta})$ is the vector of first order partial derivatives evaluated at $\boldsymbol{\theta}$. We calculate derivatives and note $\frac{\partial}{\partial \sigma_1^2} g(\boldsymbol{\theta}) = \frac{\partial}{\partial \sigma_2^2} g(\boldsymbol{\theta}) = 0$. We then simplify to find

$$g(\hat{\boldsymbol{\theta}}) \doteq \bar{Y} + \gamma_1(\bar{y}_{lr} - \bar{Y}) + \gamma_2(\bar{y}_3 - \bar{Y})$$

where $\gamma_1 = (\sigma_1^2)^{-1}/((\sigma_1^2)^{-1} + (\sigma_2^2)^{-1})$ and $\gamma_2 = (\sigma_2^2)^{-1}/((\sigma_1^2)^{-1} + (\sigma_2^2)^{-1})$. Therefore,

$$\begin{aligned} E(\bar{y}_w^{lr} - \bar{Y})^2 &\doteq E[\{(\gamma_1(\bar{y}_{lr} - \bar{Y}) + \gamma_2(\bar{y}_3 - \bar{Y})\}^2] \\ &= E[\gamma_1^2(\bar{y}_{lr} - \bar{Y})^2 + \gamma_2^2(\bar{y}_3 - \bar{Y})^2 + \gamma_1\gamma_2(\bar{y}_{lr} - \bar{Y})(\bar{y}_3 - \bar{Y})] \\ &= \gamma_1^2\sigma_1^2 + \gamma_2^2\sigma_2^2 \\ &= \frac{1}{(\sigma_1^2)^{-1} + (\sigma_2^2)^{-1}}. \end{aligned}$$

If one now replaces σ_1^2 and σ_2^2 with their consistent estimators s_1^2 and s_2^2, we get the same estimator as given in (3.8).

Appendix B

B.1 Simulation parameters

Recall $x_i \sim \text{gamma}(g,h)$. Then $E(X) = gh$ and $\sigma_x^2 = gh^2$. We find $Cov(X,Y)$ for the general model formula given in (5.1). Note that $E(Y) = \alpha + \beta\mu_x + \gamma(\sigma_x^2 + \mu_x^2)$.

$$\begin{aligned} Cov(X,Y) &= E(XY) - E(X)E(Y) \\ &= E\{X(\alpha + \beta X + \gamma X^2 + X^a\epsilon)\} - \mu_X\mu_Y \\ &= \alpha E(X) + \beta E(X^2) + \gamma E(X^3) - \mu_X\mu_Y \\ &= \alpha gh + \beta gh^2(g+1) + \gamma gh^3(g+1)(g+2) - gh\{\alpha + \beta gh + \gamma gh^3(g+1)(g+2)\} \\ &= \beta gh^2 + 2\gamma gh^3(g+1) \end{aligned}$$

We also require σ_Y^2 for each model:

$\underline{y_i = x_i + \epsilon_i}$

$$\begin{aligned} \sigma_Y^2 &= E\{(X - \epsilon - \mu_x)\}^2 \\ &= E\{(X - \mu_x) + \epsilon)\}^2 \\ &= gh^2 + \sigma_\epsilon^2 \end{aligned}$$

$y_i = x_i + \sqrt{x_i}\epsilon_i$

$$\begin{aligned}
\sigma_Y^2 &= E\{(X + \sqrt{X}\epsilon - \mu_x)\}^2 \\
&= E\{(X - \mu_x)^2 + 2(X - \mu_x)(\sqrt{X}\epsilon) + (\sqrt{X}\epsilon)^2\} \\
&= \sigma_x^2 + \mu_x\sigma_\epsilon^2 \\
&= gh^2\{h + \sigma_\epsilon^2\}
\end{aligned}$$

$y_i = \gamma x_i^2 + \epsilon_i$

$$\begin{aligned}
\sigma_Y^2 &= Cov(\gamma X^2 + \epsilon, \gamma X^2 + \epsilon) \\
&= \gamma^2\{E(X^4) - E(X^2)E(X^2)\} + \sigma_\epsilon^2 \\
&= 2\gamma^2 gh^4(g+1)(2g+3) + \sigma_\epsilon^2
\end{aligned}$$

$y_i = x_i + \gamma_i x_i^2 + \epsilon_i$

$$\begin{aligned}
\sigma_Y^2 &= Cov(X + \gamma X^2 + \epsilon, X + \gamma X^2 + \epsilon) \\
&= \sigma_x^2 + 2\gamma\{E(X^3 - E(X)E(X^2)\} + \gamma^2\{E(X^4) - E(X^2)E(X^2)\} + \sigma_\epsilon^2 \\
&= gh^2\{1 + 2\gamma[h(g+1)(g+2) - gh] + \gamma^2[h^2(g+1)(g+2)(g+3) - gh^2]\} + \sigma_\epsilon^2
\end{aligned}$$

Table B.1: Parameters for models (i)-(iv)

i) $y_i = x_i + \epsilon_i$

ρ	C_x	g	h	σ_ϵ
0.7	1	1	100	102.02
0.7	$\sqrt{2}$	0.5	200	144.28
0.85	1	1	100	61.97
0.85	$\sqrt{2}$	0.5	200	87.65

ii) $y_i = x_i + \sqrt{x_i}\epsilon_i$

0.7	1	1	100	10.202
0.7	$\sqrt{2}$	0.5	200	14.428
0.85	1	1	100	6.197
0.85	$\sqrt{2}$	0.5	200	8.765

iii) $y_i = 0.1x_i^2 + \epsilon_i$

0.7	1	1	10	35.57
0.7	$\sqrt{2}$	0.5	20	71.34
0.85	1	1	10	14.64
0.85	$\sqrt{2}$	0.5	20	19.12

iv) $y_i = x_i + 0.1x_i^2 + \epsilon_i$

0.7	1	1	10	41.26
0.7	$\sqrt{2}$	0.5	20	84.26
0.85	1	1	10	7.76
0.85	$\sqrt{2}$	0.5	20	25.77