INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI[®]

Contributions to Imputation for Missing Survey

Data

by

David Haziza, B.Sc., M.Sc.

A thesis submitted to

the Faculty of Graduate Studies and Research

in partial fulfillment of

the requirements for the degree of Doctor of Philosophy

Ottawa-Carleton Institute of Mathematics and Statistics

Carleton University, Ottawa, Ontario

June 2005

Copyright © 2005 by David Haziza

Library and
Archives Canada

Bibliothèque et
Archives Canada

Published Heritage
Branch

Direction du
Patrimoine de l'édition

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

0-494-06736-5

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

Imputation is widely used in surveys in order to compensate for item nonresponse. In this dissertation, we study several aspects of statistical inference in the presence of imputed data. The dissertation consists of nine chapters. Chapter 1 is an introduction to imputation for missing survey data. It presents a review of the different approaches used in the literature to make inference (point and variance estimation) in the presence of imputed data.

In Chapter 2, the problem of unweighted imputation is considered. We show that, in the case of unequal probability designs, unweighted imputation generally leads to biased estimators for population totals means under the design-based approach. A robust bias-adjusted estimator is then proposed, and its correct variance estimator is derived.

In Chapter 3, we make inference on a vector of domain totals (and means). We show that the imputed estimator of a vector of domain totals is biased under the design-based approach. We propose a bias-adjusted estimator which is approximately unbiased under both the design-based and the model-based approaches. We also estimate the variance of the bias-adjusted estimator. The case of unweighted imputation is also considered. Finally, a second class of bias-adjusted estimators is proposed which are approximately unbiased under both the design-based and model-based approaches and a more general imputation model.

In Chapter 4, we propose a modified regression imputation method that leads to unbiased imputed estimators of population total under ignorable response variance

estimators are also derived.

In Chapter 5, we consider the problem of variance estimation in the case of two-stage cluster sampling. In this case, we show that the usual approaches to inference may not be tenable because the intracluster correlations are not taken into account.

In Chapter 6, we consider estimating bivariate parameters such as domain means, coefficients of correlation and regression coefficients. Under several versions of random hot-deck imputation, we show that the imputed estimator of bivariate parameters is asymptotically biased. Extending the idea of Skinner and Rao (2002) for simple random sampling, we propose bias-adjusted estimators which are robust in the sense of validity under both the design-based or the model-based approaches. Finally, we show that the joint imputation procedure proposed by Shao and Wang (2002) is valid under the design-based approach.

In Chapter 7, we consider the problem of constructing imputation classes. First, we give a conditional justification for the formation of the classes. An extensive simulation study is then performed in order to assess the performance of two methods frequently used in practice for the formation of the classes: the score method and the cross-classification method.

In Chapter 8, asymptotic consistency as well as asymptotic normality of some imputed estimators are established.

Finally, some suggestions for further research are presented in Chapter 9.

Acknowledgments

I am deeply grateful to my thesis supervisor, Professor J.N.K. Rao, without whose help and advice this thesis could not have been written. Throughout my graduate studies at Carleton University, Professor Rao has been an inspiring mentor. He gave me a tremendous support with his valuable suggestions, inspiring ideas and insightful guidance. By his side, not only I learned infinitely many aspects of survey sampling and statistics in general, but he communicated to me his passion for research, and taught me that research takes hard work and discipline. It has been a great privilege to work under his supervision.

I would also like to thank Professors P. Lahiri, M. Alvo, M. Mojirsheibani and D.R. Thomas for serving as members of my advisory committee.

I would like to thank Professor I.S. Alalouf who encouraged me tremendously to pursue my graduate studies when I was a Master's student at UQAM.

Aux miens, dont l'affection a nourri ce travail: me voici arrivé au bout de mon long voyage académique. Un voyage qui n'aurait jamais pu se faire sans l'appui inconditionnel de mes parents qui ont toujours veillé à ce que j'aie accès à la meilleure éducation et la meilleure formation académique. Je leur en serai éternellement reconnaissant.

Finally, I wish to express my sincere gratitude to my spouse, Mélanie, for her unconditional love and support and without whom this thesis may never have seen the light.

Dedication

À mon père

Contents

Abstract	ii
Acknowledgments	iv
Dedication	v
Contents	vi
List of Tables	xvi
List of Figures	xviii
1 Literature review	1
1.1 Introduction	1
1.2 Sampling Designs	4
1.3 An Imputed Estimator	6
1.4 Imputation Methods	7
1.4.1 Some deterministic imputation methods	8
1.4.2 Some random imputation methods	9
1.5 Nonresponse Versus Two-phase Sampling	10
1.6 Response Mechanism	11
1.7 Approaches to Inference	13

1.8	Properties of the Imputed Estimator	14
1.8.1	Properties of the imputed estimator when the assumptions are satisfied	15
1.8.2	Properties of the imputed estimator when the assumptions are not satisfied	16
1.9	Distortion of Distributions	18
1.10	Variance estimation	20
1.10.1	Two-Phase Approach under Assumption DB	22
1.10.2	Two-Phase Approach under Assumption MB	23
1.10.3	The Reverse Approach	25
1.10.4	Jackknife Method	28
1.10.5	Bootstrap Method	30
1.10.6	Multiple Imputation	31
1.10.7	Fractionally Weighted Imputation	35
1.11	Nearest-Neighbour Imputation	37
1.12	Estimation of relationships	38
1.13	Outline of the Thesis	40
2	Inference under unweighted imputation	43
2.1	Introduction	43
2.2	Framework and Assumptions	44
2.3	Ratio Imputation	46
2.3.1	Estimation of a Mean	46

2.3.2	Variance estimation	51
2.4	Random Hot-Deck Imputation	54
2.4.1	Estimation of a Mean	55
2.4.2	Variance Estimation	56
2.5	Simulation Study	59
2.6	Remarks	62
2.7	Appendix	64
2.7.1	Derivation of v_1	64
2.7.2	Derivation of v_2 for the estimator $\bar{y}_{IR}$ under unweighted imputation	66
2.7.3	Derivation of v_2 for the estimator $\bar{y}_{IR}^a$	67
2.7.4	Non-negativity of $V_\star(\bar{y}_{IH}^a - \bar{Y})$	69
3	Inference for domain means	70
3.1	Introduction	70
3.2	Design-Based Approach	73
3.2.1	Unadjusted estimators	74
3.2.2	Bias-adjusted estimators	76
3.2.3	Complete case estimators	78
3.2.4	Survey-weighted random regression imputation	79
3.3	Model-Based Approach	80
3.4	Variance Estimation	81
3.4.1	Survey-weighted regression imputation	82

3.4.2	Survey-weighted random regression imputation	86
3.4.3	Unadjusted imputed estimator: assumption MB	86
3.5	Simulation Study	88
3.6	Unweighted imputation	90
3.6.1	Bias-adjusted estimator	90
3.6.2	Variance estimation	93
3.7	Alternative Bias-Adjusted Estimators	95
3.7.1	Bias-Adjusted Estimators	95
3.7.2	Simulation Study	98
4	Nonresponse model approach to inference	104
4.1	Introduction	104
4.2	Estimation of a Total	106
4.2.1	Deterministic regression imputation	107
4.2.2	A bias-adjusted estimator	108
4.2.3	Modified deterministic regression imputation	110
4.2.4	Random regression imputation	113
4.3	Variance Estimation	114
4.3.1	Known p_i	115
4.3.2	Unknown p_i	118
4.4	Simulation Studies	119
4.4.1	Simulation study 1	120
4.4.2	Simulation study 2	123

4.5	Estimation of Domain Means	127
5	Inference in two-stage cluster sampling	130
5.1	Introduction	130
5.2	Approaches suitable for cluster sampling	133
5.3	Comparison of conditional variances under the two-phase approach .	134
5.3.1	Design-based approach	135
5.3.2	Model-based approach	136
5.4	Variance estimation under the model-based approach	138
5.4.1	Complete response	138
5.4.2	Särndal's method	139
5.4.3	Shao-Steel's method	141
5.5	Comparison of the variance under the reverse approach	143
5.5.1	Uniform response mechanism	143
5.5.2	"Beta-binomial" response mechanism	144
5.5.3	Choice of the model	145
5.6	Simulation Study	149
5.7	Variance estimation under random hot-deck imputation	153
5.8	Deterministic EBLUP imputation	154
6	Inference for Bivariate Statistics	162
6.1	Introduction	162
6.2	Framework and Assumptions	164
6.3	Random hot-deck procedures	167

6.3.1	Common Donor Random Hot-Deck Imputation	168
6.3.2	Marginal Random Hot-Deck Imputation	168
6.3.3	Hybrid Random Hot-Deck Imputation	169
6.4	Bias of the imputed estimator	169
6.4.1	Assumption DB	169
6.4.2	Assumption MB	170
6.5	Bias-Adjusted Estimators	171
6.6	Simulation study	172
6.7	Variance estimation	174
6.8	Validity of the Shao-Wang procedure under assumption DB	181
6.9	Appendix	186
6.9.1	Bias of $\hat{\theta}_I$ under assumption DB	186
6.9.2	Bias of $\hat{\theta}_I$ under assumption MB	187
6.9.3	Unbiasedness of $\hat{\theta}_I^a$ under assumption DB	188
6.9.4	Unbiasedness of $\hat{\theta}_I^a$ under assumption MB	193
7	On the construction of imputation classes	195
7.1	Introduction	195
7.2	Background and notation	198
7.3	Evaluating the quality of the imputed estimator	200
7.3.1	Nonresponse Model Approach	202
7.3.2	Imputation Model Approach	203
7.3.3	Random hot-deck imputation	205

7.3.4	Modeling issues	206
7.3.5	Variance estimation	207
7.4	Methods for constructing classes	208
7.4.1	The score method	208
7.4.2	Cross-classification method	211
7.5	Empirical determination of the number of classes	212
7.5.1	Coefficient of determination	213
7.5.2	Within-class bias	213
7.6	Multivariate case	214
7.7	Simulation study	215
7.7.1	The score method	217
7.7.2	The cross-classification method	217
7.8	Discussion of the results	218
7.8.1	The score method	218
7.8.2	The cross-classification method	221
7.8.3	Comparison of the methods	222
8	Some asymptotic results	231
8.1	Introduction	231
8.2	Consistency of some imputed estimators	233
8.3	Asymptotic normality under ratio imputation	240
8.3.1	Deterministic weighted ratio imputation	242
8.3.2	Asymptotic normality under random ratio imputation	244

CONTENTS

xiii

9 Future Research

248

Bibliography

251

List of Tables

1.1	Coverage probability and median length of the confidence intervals at a nominal rate of 95%	35
2.1	Bias ratio (%) and CV (%) of the imputed estimators	62
2.2	Relative bias (%) of the variance estimators and comparisons of the CV ratios of the variance estimators	63
3.1	Mean of the variable WE by age group in the population.	90
3.2	Percent absolute bias ratio of the unadjusted, adjusted and complete case estimators of domain totals and means under linear regression imputation	91
3.3	Efficiency of the adjusted estimator relative to the unadjusted estimator (E_{an}) and the complete case estimator (E_{ac}) of domain totals and means under linear regression imputation	92
3.4	Percent absolute bias ratio of the unadjusted, complete case, adjusted estimator ($\hat{Y}_{I(j)}^a$) and adjusted estimator ($\hat{Y}_{I(j)}^{a(2)}$) under the response Mechanism 1.	102

3.5	Efficiency of the adjusted estimator relative to the unadjusted estimator (E_{au}), the complete case estimator (E_{ac}) and the adjusted estimator, $\hat{Y}_{I(j)}^{a(2)}(E_{aa2})$ under the response Mechanism 1.	102
3.6	Percent absolute bias ratio of the unadjusted, complete case, adjusted estimator $\left(\hat{Y}_{I(j)}^a\right)$ and adjusted estimator $\hat{Y}_{I(j)}^{a(2)}$ (denoted E_{aa2}) under the response Mechanism 2.	102
3.7	Efficiency of the adjusted estimator relative to the unadjusted estimator (E_{au}), the complete case estimator (E_{ac}) and the adjusted estimator, $\hat{Y}_{I(j)}^{a(2)}(E_{aa2})$ under the response Mechanism 2.	103
4.1	Models used for imputation	123
4.2	Relative bias (%) and RRMSE (%) of the imputed estimators under Mechanism 1	124
4.3	Relative bias (%) and RRMSE (%) of the imputed estimators under Mechanism 2	125
4.4	Relative bias (%) and RRMSE (%) of the imputed estimators	127
5.1	Magnitude of R^* for different values of ρ_m and ρ_p and $M = 20$	145
5.2	Magnitude of $Q \times 100\%$ with $\frac{n}{N} = 0.1$	149
5.3	Magnitude of $Q \times 100\%$ with $\frac{n}{N} = 0.25$	150
5.4	Magnitude of $Q \times 100\%$ with $\frac{n}{N} = 0.5$	151
5.5	Magnitude of $Q' \times 100\%$ with $\frac{n}{N} = 0.1$	152
5.6	Magnitude of $Q' \times 100\%$ with $\frac{n}{N} = 0.25$	153
5.7	Magnitude of $Q' \times 100\%$ with $\frac{n}{N} = 0.5$	154

5.8	Relative bias of (%) the variance estimators with $n/N = 0.1$	155
5.9	Relative bias (%) of the variance estimators with $n/N = 0.5$	156
5.10	Percent relative efficiency (RE) for $r_o/m_o = 0.4$	159
5.11	Percent relative efficiency (RE) for $r_o/m_o = 0.6$	160
5.12	Percent relative efficiency (RE) for $r_o/m_o = 0.8$	161
6.1	Relative bias (%) of the estimators	174
6.2	Mean square of the estimators ($\times 10^5$)	175
7.1	Relative bias and RRMSE of the imputed estimator under the response mechanism 1	229
7.2	Relative bias and RRMSE of the imputed estimator under the response mechanism 2	230

List of Figures

7.1	Relative bias of the imputed estimator using the classification algorithm and $\hat{y}$ -criterion	224
7.2	RRMSE of the imputed estimator using the classification algorithm and $\hat{y}$ -criterion	224
7.3	Relative bias of the imputed estimator using the classification algorithm and $\hat{p}$ -criterion	225
7.4	RRMSE of the imputed estimator using the classification algorithm and $\hat{p}$ -criterion	225
7.5	Relative bias of the imputed estimator using the classification algorithm under both $\hat{y}$ -criterion and $\hat{p}$ -criterion	226
7.6	RRMSE of the imputed estimator using the classification algorithm under both $\hat{y}$ -criterion and $\hat{p}$ -criterion	226
7.7	Relative bias of the imputed estimator: equal quantile method vs classification algorithm under the $\hat{y}$ -criterion	227
7.8	R^2 of the imputed estimator for the first response mechanism: equal quantile method vs classification score	227

- 7.9 Relative bias of the imputed estimator using the classification algorithm 228
- 7.10 Relative bias of the imputed estimator using the classification algorithm 228

Chapter 1

Literature review

1.1 Introduction

Despite the best efforts made by the survey staff to maximize response, it is almost certain that some degree of nonresponse will occur in large scale surveys. Essentially, survey statisticians distinguish between two types of nonresponse, total or unit nonresponse (when no information is collected on a sampled unit) and partial or item nonresponse (when the absence of information is limited to some variables only). Unit nonresponse occurs, for example, when the sampled unit is not-at-home or refuses to participate in the survey. Item nonresponse occurs, for example, because the sampled unit may refuse to respond to sensitive items, may not know the answer to some items, or because of edit failures. Generally, weighting adjustment methods are used to compensate for unit nonresponse whereas imputation is used to compensate for item nonresponse. The main idea behind a weighting adjustment is to increase the sampling weights of the respondents in order to compensate for the nonrespondents,

while imputation is a process where one or more “plausible values” are produced to replace a missing value. It is customary in both weighting and imputation to first classify respondents and nonrespondents into classes, formed on the basis of information recorded for all units in the sample.

The word imputation comes from the Latin verb *imputare* that may be translated as “guesstimate inside”. The first use of the word imputation in its current meaning is due to Hansen, Hurwitz and Madow (1953) where the term is used to refer to the corrections made in the 1948 American Survey of Retail Shares. In the 70’s, the development of automated editing processes led to the proliferation of imputation as a technique to compensate for item nonresponse. For more details on the history of imputation, the reader is referred to Rancourt (2001).

We distinguish between single and multiple imputation. Single imputation consists of creating a single imputed value to “fill the hole” of a missing value which leads to the creation of a single complete data file. Multiple imputation, proposed by Rubin (1978), consists of creating $M \geq 2$ imputed values to “fill the hole” of a missing value which leads to the creation of M complete data files. Multiple imputation is discussed in section 1.10.6.

Single imputation is widely used for treating item nonresponse because it presents the following advantages: (1) Unlike weighting adjustment for each item, imputation allows for the use of a single sampling weight for all items. (2) Because of a single sampling weight for all items, imputation leads to the creation of a complete data file so that the results of different analyses are consistent with each other. (3) The information available on partial respondents may be used as auxiliary information to

improve the quality of imputed values.

It is nonetheless important to note that imputation also presents certain risks. The most significant risks are as follows: (1) Even though imputation leads to the creation of a complete data file, inferences are valid only if the underlying assumptions about the response mechanism and/or the imputation model are satisfied. (2) Some imputation methods tend to distort the distribution of the variables of interest (i.e., the variables being imputed). (3) Treating the imputed values as if they were observed may lead to a substantial underestimation of the variance of the estimator, especially if the item nonresponse rate is appreciable. (4) Marginal imputation for each item separately has the effect of distorting relationships between variables.

Imputation is essentially a modeling exercise. Hence, the quality of the estimates will depend on the availability of good auxiliary information that will be used in the construction of imputed values and/or the construction of imputation classes. We define *auxiliary information* as a set of variables that are available for all units in the sample or all units in the population.

A popular method used in presence of missing data known as complete-case analysis, consists of discarding units whose information is incomplete. However, complete-case analysis may lead to severely biased estimators unless the data is Missing Completely At Random (see section 1.6). Also, the sampling weights cannot be used unless they are first adjusted. As a result, the estimation of population totals cannot be performed. For these reasons, complete-case analysis will not be considered in this thesis.

1.2 Sampling Designs

Let $U = \{1, 2, \dots, N\}$ be a population of N identifiable elements. A sample s is any set of population elements such that $s \subseteq U$. Let S be the collection of all samples s . A sampling design is a probability function $p(\cdot)$ on S . Thus, $p(s) \geq 0$ for all $s \in S$ and $\sum_{s \in S} p(s) = 1$. Commonly used sampling designs include simple random sampling, probability proportional to size sampling, stratified random sampling and stratified multistage designs. Associated with the i^{th} element of the population is a vector of characteristics denoted by $\mathbf{y}_i$ and the population of vectors $\mathbf{y}_i$ is denoted by $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$ and that of auxiliary variables $\mathbf{z}_i$ by $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)$. Throughout the thesis, we assume the absence of sample selection bias (see Pfefferman, 1993). That is,

$$P(\delta_i = 1 | \mathbf{y}_i, \mathbf{z}_i) = P(\delta_i = 1 | \mathbf{z}_i),$$

where δ_i is the sample selection indicator for unit i , such that $\delta_i = 1$ if $i \in s$ and $\delta_i = 0$, otherwise. Let $\theta = \theta(\mathbf{y}_1, \dots, \mathbf{y}_N)$ be a parameter of interest and $\hat{\theta}$ be an estimator of θ based on the realized sample s . An estimator $\hat{\theta}$ is said to be design unbiased if

$$E_p(\hat{\theta}) = \theta,$$

where $E_p(\cdot)$ denotes the expectation operator with respect to the sampling design $p(\cdot)$. In practice, we are often interested in parameters such as population totals and population means given respectively by $Y = \sum_{i \in U} y_i$ and $\bar{Y} = \frac{1}{N} \sum_{i \in U} y_i$. Under a given sampling design $p(\cdot)$, an approximately unbiased estimator (that does not make

use of auxiliary information) of the mean $\bar{Y}$ is given by

$$\bar{y} = \frac{1}{\sum_{i \in s} w_i} \sum_{i \in s} w_i y_i, \quad (1.1)$$

where $w_i = 1/\pi_i$ is the design weight attached to unit i and $\pi_i = P(i \in s)$ is the first-order probability of inclusion of unit i in the sample. An unbiased estimator of the total is the Horvitz-Thompson estimator $\hat{Y} = \sum_{i \in s} w_i y_i$, assuming that $\pi_i > 0$ for all units $i \in U$.

We now describe the stratified multistage sampling design frequently used in practice. The population under consideration is stratified into L strata with N_h primary sampling units (PSU's) or clusters in the h^{th} stratum. Within each stratum, $n_h \geq 2$ clusters are selected from stratum h , independently across strata. The first-stage clusters are usually selected without replacement to avoid the selection of the same cluster more than once. Within the $(hi)^{th}$ sampled first-stage cluster, m_{hi} ultimate units (elements) are sampled according to some probability sampling method, $i = 1, \dots, n_h$; $h = 1, \dots, L$. Note that we do not need to specify the number of stages or the sampling methods beyond the first stage. We simply assume that subsampling within sampled clusters is performed to ensure unbiased estimation of cluster totals, Y_{hi} , $i = 1, \dots, n_h$; $h = 1, \dots, L$. Associated with the j^{th} sampled element in the i^{th} sampled cluster belonging to the h^{th} stratum is the variable of interest, y_{hij} , and the basic survey weights, w_{hij} . In this case, it is more convenient to write the estimator (1.1) as

$$\bar{y} = \frac{1}{\sum_{(hij) \in s} w_{hij}} \sum_{(hij) \in s} w_{hij} y_{hij}, \quad (1.2)$$

where s denotes the sample of elements. Similarly, $\hat{Y} = \sum_{(hij) \in s} w_{hij} y_{hij}$.

At the stage of variance estimation it is a common practice to treat the sample as if the first-stage clusters are drawn with replacement, to simplify the derivation of variance estimators. This approximation leads to overestimation of the true variance, but the bias will be small if the overall first-stage sampling fraction, $\sum_{h=1}^L n_h / \sum_{h=1}^L N_h$ is small (Shao, 2002). In this case, a linearization estimator of variance of $\bar{y}$ in (1.2) is given by

$$v_L(\bar{y}) = \sum_{h=1}^L \frac{1}{n_h(n_h - 1)} \sum_{i \in s_h} \left(\hat{Z}_{hi} - \hat{Z}_h \right)^2, \quad (1.3)$$

where $\hat{Z}_{hi} = \sum_j n_h w_{hij} z_{hij}$ with $z_{hij} = y_{hij} - \bar{y}$ and $\hat{Z}_h = \frac{1}{n_h} \sum_{i \in s_h} \hat{Z}_{hi}$. An estimator of variance of $\hat{Y}$ is obtained from (1.3) by changing z_{hij} to y_{hij} .

1.3 An Imputed Estimator

We focus on the estimation of mean $\bar{Y}$. In the presence of nonresponse to item y , it is not possible to compute the estimator (1.1) since some y -values are missing. In this case, we define an imputed estimator of the population mean $\bar{Y}$ as

$$\bar{y}_I = \frac{1}{\sum_{i \in s} w_i} \left[\sum_{i \in s_r} w_i y_i + \sum_{i \in s_m} w_i y_i^* \right], \quad (1.4)$$

where s_r denotes the set of respondents to item y , of size r , s_m denotes the set of nonrespondents to item y , of size m , and y_i^* denotes the imputed value for missing y_i . Note that the imputed estimator (1.4) is simply the weighted mean of observed and imputed values in the sample.

1.4 Imputation Methods

Imputation methods may be classified into two broad classes: deterministic and stochastic. Deterministic methods are those that yield a fixed imputed value given the sample if the imputation process is repeated. Random methods are those that have a random component; hence these methods do not necessarily yield the same imputed value given the sample if the imputation process is repeated. Many imputation methods used in practice can be represented as special cases of the following “imputation model” (Kalton and Kasprzyk, 1986):

$$\begin{aligned} y_i &= f(\mathbf{z}_i) + \epsilon_i, \\ E_m(\epsilon_i) &= 0, Cov_m(\epsilon_i, \epsilon_j) = 0 \text{ if } i \neq j, V_m(\epsilon_i) = \sigma_i^2 = \sigma^2 b(\mathbf{z}_i) \end{aligned} \quad (1.5)$$

where $f(\cdot)$ is an arbitrary function, $\mathbf{z}_i$ is a q -vector of auxiliary variables available for all $i \in s$, σ^2 is an unknown parameter, $b(\cdot)$ is a known function and E_m , V_m , and Cov_m denote respectively the expectation, the variance and the covariance operators with respect to the model (1.5). Note that the imputation model (1.5) is used to simply motivate the particular imputation used. In the case of deterministic imputation, the imputed value y_i^* is obtained by estimating $f(\mathbf{z})$ by $\hat{f}_r(\mathbf{z})$ based on the responding units, $i \in s_r$; that is, $y_i^* = \hat{f}_r(\mathbf{z}_i)$ for $i \in s_m$. Random imputation can be seen as a deterministic imputation plus a random noise ϵ^* ; that is,

$$y_i^* = \hat{f}_r(\mathbf{z}_i) + \hat{\sigma}_i \epsilon_i^*, \text{ for } i \in s_m, \quad (1.6)$$

where $\hat{\sigma}_i$ is an estimator of σ_i . For example, the random component ϵ_i^* may be drawn from a normal distribution with mean 0 and variance c for some $c > 0$. However,

it is more natural to select (with replacement) the random noise from the set, $A_r = \{e_j - \bar{e}_r; j \in s_r\}$, of standardized residuals observed from the responding units, where $e_j = \frac{1}{\hat{\sigma}_j} [y_j - \hat{f}_r(z_j)]$ and $\bar{e}_r = \sum_{j \in s_r} w_j e_j / \sum_{j \in s_r} w_j$, see section 1.4.2.

Note that any imputation method applied within classes (e.g., mean within classes or random hot-deck within classes) is a special case of model (1.5) because it suffices to include the appropriate indicator variables in (1.5).

1.4.1 Some deterministic imputation methods

Commonly used deterministic imputation methods include the following:

- i) (Weighted) regression imputation: In this case, $f(z_i) = z_i' \beta$ and $b(z_i) = \lambda' z_i$ for a specified vector of constants λ . We have

$$y_i^* = z_i' \hat{B}_r, i \in s_m, \quad (1.7)$$

where

$$\hat{B}_r = \left(\sum_{i \in s_r} w_i z_i z_i' / (\lambda' z_i) \right)^{-1} \sum_{i \in s_r} w_i z_i y_i / (\lambda' z_i) \quad (1.8)$$

is the weighted least square estimator of β based on the responding units.

- ii) (Weighted) ratio imputation: In this case, a single auxiliary variable is available for all the sampled units, $f(z_i) = \beta z_i$ and $b(z_i) = z_i$. We have

$$y_i^* = \frac{\bar{y}_r}{\bar{z}_r} z_i, i \in s_m, \quad (1.9)$$

where $(\bar{y}_r, \bar{z}_r) = \frac{1}{\sum_{i \in s_r} w_i} \sum_{i \in s_r} w_i (y_i, z_i)$ are the weighted means of the respondents for variables y and z respectively.

- iii) (Weighted) mean imputation: In this case, $z_i = 1$ for all $i \in s$, $f(z_i) = \beta$ and $b(z_i) = 1$. We have

$$y_i^* = \bar{y}_r, i \in s_m. \quad (1.10)$$

- iv) Nearest-neighbour imputation (NNI): NNI is a nonparametric imputation method.

Hence, we do not attempt to specify the form of $f(z_i)$ nor the variance structure σ_i^2 . We have

$$y_i^* = y_j \text{ such that } \text{dist}(z_i, z_j) \text{ is minimum, } j \in s_r, \quad (1.11)$$

where $\text{dist}(\cdot, \cdot)$ is a distance measure to be determined.

1.4.2 Some random imputation methods

Commonly used stochastic imputation methods include the following:

- (i) (Weighted) random hot-deck imputation: (Weighted) random hot-deck imputation consists in selecting donor values with replacement from the set of respondents s_r with probabilities $w_j / \sum_{l \in s_r} w_l$ to replace missing values. That is,

$$y_i^* = y_j \text{ with } P(y_i^* = y_j) = \frac{w_j}{\sum_{l \in s_r} w_l}, j \in s_r. \quad (1.12)$$

Random hot-deck imputation can be seen as weighted mean imputation plus a random noise; that is, $y_i^* = \bar{y}_r + \epsilon_i^*$, where $\epsilon_i^* = e_j^* - \bar{e}_r$ such that $P(e_i^* = e_j) = \frac{w_j}{\sum_{l \in s_r} w_l}$ with $e_j = (y_j - \bar{y}_r)$ and $\bar{e}_r = \sum_{j \in s_r} w_j e_j / \sum_{j \in s_r} w_j$. If the donors are selected by simple random sampling with replacement from s_r , then we have

unweighted random hot-deck imputation. Note that the imputed value under random hot-deck imputation is taken from a respondent item value, unlike in the case of mean imputation $y_i^* = \bar{y}_r$. For example, if y is binary (0, 1), imputed values y^* would take only the value 0 or 1 under random hot-deck imputation.

- (ii) (Weighted) random regression imputation: (Weighted) random regression imputation can be seen as deterministic regression imputation given in (1.7) plus a random noise. We have

$$y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r + (\boldsymbol{\lambda}' \mathbf{z}_i)^{1/2} \epsilon_i^*, \quad (1.13)$$

where $\epsilon_i^* = e_i^* - \bar{e}_r$ such that $P(e_i^* = e_j) = w_j / \sum_{l \in s_r} w_l$ with

$$e_j = (\boldsymbol{\lambda}' \mathbf{z}_j)^{-1/2} (y_j - \mathbf{z}_j' \hat{\mathbf{B}}_r) \text{ and } \bar{e}_r = \sum_{j \in s_r} w_j e_j / \sum_{j \in s_r} w_j.$$

Unweighted random regression imputation uses simple random sampling with replacement. Note that the imputed values under random regression imputation may not correspond to a respondent value unlike the value under random hot-deck imputation or nearest neighbour imputation.

1.5 Nonresponse Versus Two-phase Sampling

Nonresponse is often compared to two-phase sampling. Two-phase sampling is often used in surveys when the sampling frame contains little or no auxiliary information. In this case, a first sample s_1 , of usually large size, is selected from the population according to a sampling design $p_1(\cdot)$. This sample will allow the survey staff to collect inexpensive information about the population that will serve as auxiliary information

at the second phase. From s_1 , a sample s_2 is selected according to a sampling design $p_2(\cdot|s_1)$. An estimator (that does not make use of auxiliary information) of a population mean $\bar{Y}$ is given by

$$\bar{y}_{TP} = \frac{1}{\sum_{i \in s_2} w_{1i} w_{2i}} \sum_{i \in s_2} w_{1i} w_{2i} y_i, \quad (1.14)$$

where $w_{1i} = 1/\pi_{1i}$, $\pi_{1i} = P(i \in s_1)$, $w_{2i} = 1/\pi_{2i|s}$ and $\pi_{2i|s} = P(i \in s_2|s_1, i \in s_1)$. The estimator $\bar{y}_{TP}$ in (1.14) is approximately unbiased for $\bar{Y}$; that is $E(\bar{y}_{TP}) = E_{p_1} E_{p_2}(\bar{y}_{TP}) \approx \bar{Y}$, where $E_{p_1}(\cdot)$ and $E_{p_2}(\cdot)$ denote respectively the expectation operators with respect to the sampling designs $p_1(\cdot)$ and $p_2(\cdot|s_1)$. In two-phase sampling, the survey statistician knows both the inclusion probabilities π_{1i} and $\pi_{2i|s}$. In the presence of nonresponse to item y , the set of respondents s_r can be viewed as a second phase sample, except that the conditional inclusion probabilities for s_r are unknown. In this case, let $\pi_{2i|s} = P(i \in s_r|s, i \in s)$ denote the probability of item response for unit i , $i \in s$. We assume that $\pi_{2i|s}$ does not depend on s , i.e., $\pi_{2i|s} = p_i$. Since the p_i 's are unknown, we have to make some assumptions about the response mechanism.

1.6 Response Mechanism

Let a_i be an item response indicator for unit i defined by

$$a_i = \begin{cases} 1 & \text{if } i \in s_r, \\ 0 & \text{if } i \notin s_r. \end{cases}$$

The distribution of the response indicators, $P(a_i|s)$, is called the response mechanism. Since it is unknown, we do not have the choice but to make assumptions about

the response mechanism. We assume that the units respond independently of one another; that is, $p_{ij} = P(a_i = 1, a_j = 1 | s, i \in s, j \in s, i \neq j) = p_i p_j$. The assumption of independence is usually satisfied in practice, although it is easy to come up with situations where it is not satisfied. For example, in cluster sampling, the units within clusters (e.g., households) may not respond independently of one another (see, Chapter 5). We distinguish between three types of response mechanisms:

- (i) Uniform response mechanism: A response mechanism is said to be uniform if $p_i = p$ for all $i \in U$. In other words, a response mechanism is uniform when the probability of response is independent of all the variables available. When the response mechanism is uniform, we say that the data is *Missing Completely At Random* (MCAR). This mechanism is, in general, not realistic in practice. It is then customary to assume uniform response $p_i = p_\nu$ within imputation classes ν .
- (ii) Ignorable response mechanism: If $P(a_i = 1 | y, \mathbf{z}) = P(a_i = 1 | \mathbf{z})$, then the response mechanism is said to be ignorable. That is, the response mechanism is ignorable if the probability of response to item y depends on auxiliary variables $\mathbf{z}$ but not on the variable of interest (i.e., the variable being imputed). Note that uniform response is a particular case of ignorable response. When the response mechanism is ignorable, we say that the data is *Missing At Random* (MAR).
- (iii) Nonignorable response mechanism: A response mechanism is nonignorable if the probability of response depends on the variable of interest y . When the response mechanism is nonignorable, we say that the data is *Not Missing At*

Random (NMAR)

The previous terminology is due to Rubin (1976).

1.7 Approaches to Inference

To reduce or eliminate biases of the imputed estimators, imputation is often done by first dividing the population U into J nonoverlapping cells (or classes), called imputation cells, and then imputing nonrespondents in an imputation cell using respondents as donors within the same imputation cell, independently across the J imputation cells. In the literature, researchers have considered two distinct approaches: (i) design-based (Rao, 1990) and (ii) model-based (Särndal, 1990). Approach (i) may be called Nonresponse Model approach, and approach (ii) may be called as the Imputation Model approach. In this dissertation, we use this terminology interchangeably.

Under the design based approach, the following assumption holds:

Assumption DB: Within an imputation cell, the response mechanism is uniform, but varies across cells ν , i.e., $p_i = p_\nu$ for unit i in cell ν . For simplicity, we assume a single imputation cell, but the extension to multiple cells is straightforward.

Under the model-based approach, the following assumption holds:

Assumption MB: Within an imputation cell, the response mechanism is ignorable. Imputation is performed according to an imputation model. Under linear regression imputation, the imputation model is given by

$$E_m(y_i) = \mathbf{z}_i' \boldsymbol{\beta}, V_m(y_i) = \sigma_i^2 = \sigma^2 \boldsymbol{\lambda}' \mathbf{z}_i, Cov_m(y_i, y_j) = 0 \text{ if } i \neq j, \quad (1.15)$$

where β is q -vector of unknown parameters, $\mathbf{z}_i$ is a q -vector of auxiliary variables available for all $i \in s$, λ is q -vector of specified constants and σ^2 is an unknown parameter. The restriction $\sigma_i^2 = \sigma^2 \lambda' \mathbf{z}_i$ does not severely restrict the range of imputation models. Note that a particular case of (1.15) leads to the ratio imputation model given by

$$E_m(y_i) = \beta z_i, V_m(y_i) = \sigma^2 z_i, Cov_m(y_i, y_j) = 0 \text{ if } i \neq j, \quad (1.16)$$

where β and z_i are the scalar versions of β and $\mathbf{z}_i$ respectively.

Another particular case of (1.15) leads to the mean imputation model given by

$$E_m(y_i) = \beta, V_m(y_i) = \sigma^2, Cov_m(y_i, y_j) = 0 \text{ if } i \neq j, \quad (1.17)$$

Random hot-deck imputation is also motivated by (1.17).

Assumption DB consists in describing the response mechanism completely. The estimators will then be valid if the response mechanism holds. The response mechanism in Assumption MB is much weaker than that in Assumption DB, but the inference depends on the validity of the imputation model (1.15). Under assumption MB, only MAR is assumed without actually specifying the response mechanism. We attempt in this thesis to make robust inferences in the sense of validity under both approaches.

1.8 Properties of the Imputed Estimator

In the presence of nonresponse to item y , the imputed estimator (1.4) is approximately unbiased if the assumptions about the response mechanism or the imputation model

are satisfied.

1.8.1 Properties of the imputed estimator when the assumptions are satisfied

Proposition 1.1 *The imputed estimator (1.4) under random regression imputation (1.13) is approximately unbiased under assumption DB; that is,*

$$E(\bar{y}_I) = E_p E_r E_\star(\bar{y}_I) \approx \bar{Y},$$

where $E_r(\cdot)$ denotes the expectation operator with respect to the response mechanism and $E_\star(\cdot)$ denotes the expectation operator with respect to the imputation mechanism.

Proof. First, it is easily seen that $E_\star(\epsilon_i^\star) = 0$. Hence, we have

$$E_\star(\bar{y}_I) = \frac{1}{\sum_{i \in s} w_i} \left[\sum_{i \in s_r} w_i y_i + \sum_{i \in s_m} w_i \mathbf{z}_i' \hat{\mathbf{B}}_r \right]. \quad (1.18)$$

It follows that

$$E_r E_\star(\bar{y}_I) \approx \frac{1}{\sum_{i \in s} w_i} \left[p \sum_{i \in s} w_i y_i + (1-p) \sum_{i \in s} w_i \mathbf{z}_i' \hat{\mathbf{B}} \right], \quad (1.19)$$

where $E_r(\hat{\mathbf{B}}_r) \approx \hat{\mathbf{B}} = (\sum_{i \in s} w_i \mathbf{z}_i \mathbf{z}_i' / (\lambda \mathbf{z}_i'))^{-1} \sum_{i \in s} w_i \mathbf{z}_i y_i / (\lambda \mathbf{z}_i')$. Now, since $1 = \sigma_i^{-2} \sigma^2 \lambda' \mathbf{z}_i$, the second term on the right hand side of (1.19) can be written as

$$\begin{aligned} \sum_{i \in s} w_i \mathbf{z}_i' \hat{\mathbf{B}} &= \sum_{i \in s} w_i \sigma_i^{-2} \sigma^2 \lambda' \mathbf{z}_i \mathbf{z}_i' \hat{\mathbf{B}} \\ &= \sigma^2 \lambda' \sum_{i \in s} w_i \mathbf{z}_i y_i / \sigma_i^2 \\ &= \sum_{i \in s} w_i y_i. \end{aligned}$$

Hence, $E_r E_\star(\bar{y}_I) \approx \frac{1}{\sum_{i \in s} w_i} \sum_{i \in s} w_i y_i$. It now follows that $E(\bar{y}_I) = E_p E_r E_\star(\bar{y}_I) \approx \bar{Y}$. ■

Proposition 1.2 *The imputed estimator (1.4) under (weighted) random regression imputation (1.13) is approximately unbiased under assumption MB and imputation model (1.15) ; that is,*

$$E(\bar{y}_I) = E_m E_p E_r E_\star(\bar{y}_I) \approx \bar{Y}.$$

Proof. We first note that the order of the expectations E_m , E_p and E_r may be interchanged because of the MAR assumption and absence of sample selection bias so that $E_m E_p E_r E_\star(.)$ may be evaluated as $E_p E_r E_m E_\star(.)$; see Särndal (1992). Once again, $E_\star(\bar{y}_I)$ is given by (1.18). Hence,

$$\begin{aligned} E_m E_\star(\bar{y}_I) &= \frac{1}{\sum_{i \in s} w_i} \left[\sum_{i \in s_r} w_i \mathbf{z}_i' \beta + \sum_{i \in s_m} w_i \mathbf{z}_i' \beta \right] \\ &= \frac{1}{\sum_{i \in s} w_i} \sum_{i \in s} w_i \mathbf{z}_i' \beta, \end{aligned} \tag{1.20}$$

noting that $E_m(\hat{\mathbf{B}}_r) = \beta$. On the other hand, $E(\bar{Y}) = E_m(\bar{Y}) = \frac{1}{N} \sum_{i \in U} \mathbf{z}_i' \beta$.

Hence,

$$E(\bar{y}_I - \bar{Y}) \approx 0.$$

■

Note that the results in Proposition 1 and Proposition 2 are valid for the weighted mean, ratio, regression and random hot-deck imputation methods.

1.8.2 Properties of the imputed estimator when the assumptions are not satisfied

We now study the imputed estimator when the assumptions are not satisfied, and show that it is generally biased.

Bias under assumption DB

Consider the case of ratio imputation. Using the imputed values (1.5) in (1.4) leads to

$$\bar{y}_I = \frac{\bar{y}_r}{\bar{z}_r} \bar{z}, \quad (1.21)$$

where $\bar{z} = \sum_{i \in s} w_i z_i / \sum_{i \in s} w_i$. From Proposition 1, note that the imputed estimator (1.21) is approximately unbiased under assumption DB (uniform response) for $\bar{Y}$. However, we now show that if the response mechanism is not uniform, the imputed estimator is biased. To illustrate this point, consider the case of a response mechanism for which the probability of response for unit i is p_i , $i = 1, \dots, N$. In this case, the imputed estimator (1.21) is biased:

$$\text{Bias}(\bar{y}_I) = E_p E_r (\bar{y}_I - \bar{Y}) \approx \frac{\bar{Z}}{\bar{Z}_p \bar{P}} \frac{1}{N} \sum_{i \in U} (p_i - \bar{P}) (y_i - R z_i), \quad (1.22)$$

where $(\bar{Z}, \bar{P}) = \sum_{i \in U} (z_i, p_i) / N$, $\bar{Z}_p = \sum_{i \in U} p_i z_i / \sum_{i \in U} p_i$ and $R = \bar{Y} / \bar{Z}$. Note that the bias in (1.22) is 0 if the population covariance between the probabilities of response p_i and the residuals $y_i - R z_i$, is 0, which, holds, for example, in the case of uniform response, $p_i = p$. In general, however, the bias is not equal to zero.

Bias under assumption MB

Once again, we consider the case of ratio imputation. Ratio imputation suggests the use of model (1.16). From Proposition 1.2, the imputed estimator (1.21) is approximately unbiased for $\bar{Y}$, provided the assumed model is valid for all the units in the population. If the model (1.16) is not valid, the imputed estimator (1.21) is biased.

Indeed, suppose that the true model linking y and auxiliary variable z is not (1.16) but rather

$$y_i = \sum_{j=0}^J \delta_j \beta_j z_i^j + v_i^{1/2} \epsilon_i, \quad (1.23)$$

where the errors are $\epsilon_i \sim (0, \sigma^2)$ and uncorrelated, β_j 's are a set of unknown parameters, $j = 0, \dots, J$, δ_j 's are 0-1 variables indicating whether the j^{th} power term is in the model or not and v_i is a known positive function. Assume that the probability of response for unit i is p_i , $i = 1, \dots, N$. Then, it is easily seen that, according to model (1.23), the imputed estimator (1.21) is biased:

$$Bias(\bar{y}_I) = E_p E_r E_{m^*} (\bar{y}_I - \bar{Y}) \approx \bar{Z} \sum_{j=0}^J \delta_j \beta_j \left[\frac{\bar{Z}_p^{(j)}}{\bar{Z}_p} - \frac{\bar{Z}^{(j)}}{\bar{Z}} \right], \quad (1.24)$$

where $E_{m^*}(\cdot)$ is the expectation with respect to the true model (1.23),

$\bar{Z}_p^{(j)} = \frac{1}{\sum_{i \in U} p_i} \sum_{i \in U} p_i z_i^j$, $\bar{Z}_p = \frac{1}{\sum_U p_i} \sum_U p_i z_i$ and $\bar{Z}^{(j)} = \frac{1}{N} \sum_U z_i^j$. Note that the bias in (1.24) is 0 if $\frac{\bar{Z}_p^{(j)}}{\bar{Z}_p} = \frac{\bar{Z}^{(j)}}{\bar{Z}}$ which occurs, for example, under uniform response.

However, in general, the bias is not equal to zero.

1.9 Distortion of Distributions

It is well known that some deterministic imputation methods tend to distort the distribution of the variable of interest, whereas random imputation methods tend to preserve it. Consider the following example: we want to estimate the population variance $S_y^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \bar{Y})^2$. To this end, we select a simple random sample without replacement, s , of size n , from the population. An imputed estimator of S_y^2

is defined by

$$s_{yI}^2 = \frac{1}{n-1} \left[\sum_{i \in s_r} (y_i - \bar{y}_I)^2 + \sum_{i \in s_m} (y_i^* - \bar{y}_I)^2 \right]. \quad (1.25)$$

Note that (1.25) is the sample variance of observed and imputed values.

First, consider the case of simple linear regression imputation that uses the imputed values $y_i^* = \hat{\beta}_{0r} + \hat{\beta}_{1r}z_i$ for missing value y_i . In this case, s_{yI}^2 in (1.25) reduces to

$$s_{yI}^2 = \frac{1}{n-1} \left[(r-1) s_{yr}^2 + (m-1) \hat{\beta}_{1r}^2 s_{zm}^2 + \frac{mr}{n} (\bar{z}_m - \bar{z}_r)^2 \right], \quad (1.26)$$

where $s_{yr}^2 = \frac{1}{r-1} \sum_{i \in s_r} (y_i - \bar{y}_r)^2$, $s_{zm}^2 = \frac{1}{m-1} \sum_{i \in s_m} (z_i - \bar{z}_m)^2$ and $\bar{z}_m = \frac{1}{m} \sum_{i \in s_m} z_i$.

Under Assumption DB, it is easily seen that the relative bias of (1.26) is given by

$$RB(s_{yI}^2) = \frac{E_p E_r (s_{yI}^2) - S_y^2}{S_y^2} = - \left[1 - E_p \left(\frac{r}{n} \right) \right] (1 - \rho_{yz}^2) \leq 0, \quad (1.27)$$

where ρ_{yz} is the coefficient of correlation between variables y and z . The relative bias (1.27) is zero when the expected response rate $E_p \left(\frac{r}{n} \right)$ is equal to 1 or when $|\rho_{yz}| = 1$. Expression (1.27) shows that simple linear regression imputation does not preserve the population variance S_y^2 . In fact, regression imputation tends to underestimate the natural variability we would have observed under full response. The relative distortion depends on the correlation between the variable of interest y and the auxiliary variable z . If the correlation is strong, the relative distortion is not affected much by the expected response rate. Under mean imputation, it is interesting to note that the expression (1.27) reduces to

$$RB(s_{yI}^2) = - \left[1 - E_p \left(\frac{r}{n} \right) \right]. \quad (1.28)$$

In this case, the relative distortion depends only on the expected response rate.

Now, consider the case of weighted random hot-deck imputation that uses the imputed values (1.12). In this case, it is easily seen that the relative bias of s_{yI}^2 is approximately zero. Hence, unlike regression imputation, random hot-deck imputation preserves the variance S_y^2 . Note that preserving the mean and the variance of a distribution does not necessarily imply that the distribution function $F(y)$ is preserved. Chen, Rao and Sitter (2000) showed that, under some regularity conditions, $\hat{F}_I(y) \rightarrow F(y)$ in probability under weighted random hot-deck imputation, where $\hat{F}_I(y)$ is the imputed estimator of $F(y)$ given by

$$\hat{F}_I(y) = \frac{1}{\sum_{i \in s} w_i} \left[\sum_{i \in s_r} w_i I(y_i \leq y) + \sum_{i \in s_m} w_i I(y_i^* \leq y) \right],$$

where $I(\cdot)$ is the usual indicator function.

1.10 Variance estimation

Traditionally, researchers have used the following sample-response path (two-phase approach) for variance estimation:

Population $\rightarrow$ complete sample $\rightarrow$ sample with nonrespondents.

Under deterministic imputation and assumption DB, the variance of the imputed estimator (1.4) is given by

$$V(\bar{y}_I - \bar{Y}) = V_p E_r(\bar{y}_I - \bar{Y}|s) + E_p V_r(\bar{y}_I - \bar{Y}|s), \quad (1.29)$$

where $V_p(\cdot)$ denotes the variance with respect to the sampling design and $V_r(\cdot)$ denotes the variance with respect to the response mechanism. Under random imputation, one

has to take the variance due to the imputation mechanism also into account. In this case,

$$\begin{aligned} V(\bar{y}_I - \bar{Y}) &= V_p E_r E_* (\bar{y}_I - \bar{Y}|s) + E_p V_r E_* (\bar{y}_I - \bar{Y}|s) \\ &\quad + E_p E_r V_* (\bar{y}_I - \bar{Y}|s), \end{aligned} \quad (1.30)$$

where $V_*(.)$ denotes the variance operator with respect to the imputation mechanism.

Under the two-phase approach, Rao (1990) proposed to estimate the variance under assumption DB, and Särndal (1992) proposed to estimate the variance under assumption MB. Fay(1991) proposed to use a different path which reverses the order of sampling and response (we will call it the reverse approach) that can be depicted as:

Population $\longrightarrow$ census with nonrespondents $\longrightarrow$ sample with nonrespondents.

In this case, under assumption DB and deterministic imputation,

$$V(\bar{y}_I - \bar{Y}) = E_r V_p (\bar{y}_I - \bar{Y}|\mathbf{a}) + V_r E_p (\bar{y}_I - \bar{Y}|\mathbf{a}), \quad (1.31)$$

where $\mathbf{a} = (a_1, \dots, a_N)'$ is the vector of response indicators; see, Shao and Steel, (1999). Under assumption MB, $E_r(.)$ and $V_r(.)$ are replaced by $E_r E_m(.)$ and $E_r V_m(.) + V_r E_m(.)$, respectively.

The reverse approach, unlike the two-phase approach, can handle simultaneously both assumption DB and assumption MB. Under this approach, an estimator of the overall variance $V(\bar{y}_I - \bar{Y})$ is given by $v_t = v_1 + v_2$, where v_1 is an estimator of $V_p(\bar{y}_I - \bar{Y}|\mathbf{a})$ conditional on the response indicators, and v_2 is an estimator of $V_r E_p(\bar{y}_I - \bar{Y}|\mathbf{a})$. Finding v_1 does not depend on the response mechanism and/or the

imputation model, and thus v_1 is valid under either assumption DB or assumption MB. Since the component $V_p(\bar{y}_I - \bar{Y}|\mathbf{a})$ represents sampling variability, its estimation may be performed using any standard variance estimation method; for example, Taylor linearization, jackknife or bootstrap. Furthermore, the ratio $\frac{V_r E_p(\bar{y}_I - \bar{Y}|\mathbf{a})}{E_r V_p(\bar{y}_I - \bar{Y}|\mathbf{a})}$ is of order $O\left(\frac{n}{N}\right)$, so when the overall sampling fraction is negligible, the second component in (1.31) is negligible relative to the first component. We can then omit the derivation of v_2 which could be quite tedious (Shao and Steel, 1999). Finally, the variance estimators obtained under the reverse approach are consistent.

Under random imputation, the variance of the imputed estimator (1.4) is given by

$$V(\bar{y}_I - \bar{Y}) = E_r V_p E_\star(\bar{y}_I - \bar{Y}|\mathbf{a}) + E_r E_p V_\star(\bar{y}_I - \bar{Y}|\mathbf{a}) + V_r E_p E_\star(\bar{y}_I - \bar{Y}|\mathbf{a}). \quad (1.32)$$

In what follows, we describe several variance estimation methods that take nonresponse and imputation into account.

1.10.1 Two-Phase Approach under Assumption DB

Rao (1990) and Rao and Sitter (1995) proposed to estimate the variance under assumption DB. To illustrate the method, consider the case of ratio imputation and simple random sampling without replacement. Noting that given s and r , the set of respondents s_r is a simple random sample without replacement selected from s , the variance of the imputed estimator (1.21) under uniform response is given by

$$\begin{aligned} V(\bar{y}_I - \bar{Y}) &= V_p E_r(\bar{y}_I - \bar{Y}|s, r) + E_p V_r(\bar{y}_I - \bar{Y}|s, r) \\ &\approx \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{E_p(r)} - \frac{1}{n}\right) S_e^2, \end{aligned} \quad (1.33)$$

where $S_e^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - Rz_i)^2$ and $R = \bar{Y}/\bar{Z}$. Noting that $S_y^2 = S_e^2 + 2RS_{ez} + R^2S_z^2$ with $S_{ez} = \frac{1}{N-1} \sum_{i \in U} (z_i - \bar{Z})(y_i - Rz_i)$, expression (1.33) can be rewritten as

$$V(\bar{y}_I - \bar{Y}) \approx \left(\frac{1}{E_p(r)} - \frac{1}{N} \right) S_e^2 + 2 \left(\frac{1}{n} - \frac{1}{N} \right) RS_{ez} + \left(\frac{1}{n} - \frac{1}{N} \right) R^2 S_z^2. \quad (1.34)$$

Assuming $N \rightarrow \infty$, a correct variance estimator is obtained by substituting estimators for the unknown quantities in (1.34) which leads to

$$v_{cor}(\bar{y}_I - \bar{Y}) = \frac{A}{r} + 2\frac{B}{n} + \frac{C}{n}, \quad (1.35)$$

where

$$A = \frac{1}{r-1} \sum_{i \in s_r} (y_i - \hat{R}z_i)^2, \quad (1.36)$$

with $\hat{R} = \frac{\bar{y}_r}{\bar{z}_r}$,

$$B = \left(\frac{\bar{y}_r}{\bar{z}_r} \right) \frac{1}{r-1} \sum_{i \in s_r} (y_i - \hat{R}z_i)(z_i - \bar{z}_r), \quad (1.37)$$

and

$$C = \left(\frac{\bar{y}_r}{\bar{z}_r} \right)^2 \frac{1}{n-1} \sum_{i \in s} (z_i - \bar{z})^2. \quad (1.38)$$

1.10.2 Two-Phase Approach under Assumption MB

Särndal (1990, 1992) proposed a variance estimator under assumption MB in the case of simple random sampling. The method is based on the following decomposition

$$\bar{y}_I - \bar{Y} = (\bar{y} - \bar{Y}) + (\bar{y}_I - \bar{y}). \quad (1.39)$$

That is, the total error of $\bar{y}_I$ can be expressed as the sum of the sampling error and the nonresponse error. Assuming that the imputed estimator $\bar{y}_I$ is approximately

unbiased for $\bar{Y}$, the variance of $\bar{y}_I$ is then given by

$$V(\bar{y}_I) = E_m E_p E_r (\bar{y}_I - \bar{Y})^2 = V_{sam} + V_{imp} + V_{mix} \equiv V_{tot}, \quad (1.40)$$

where $V_{sam} = E_m V_p (\bar{y} - \bar{Y})$, $V_{imp} = E_p E_r E_m (\bar{y}_I - \bar{y}|s)$ and

$V_{mix} = 2E_m E_p [(\bar{y} - \bar{Y}) (E_r (\bar{y}_I - \bar{y}) | s)]$. In the case of self-weighting uni-stage designs and regression imputation, the component V_{mix} is exactly 0. Even when it is not exactly zero, Särndal argued that this component is typically much smaller than the terms V_{sam} and V_{imp} , so it may be omitted in the computation of the total variance. However, Brick, Kalton and Kim (2004) showed that in the case of unequal probability designs, the contribution (positive or negative) of V_{mix} may be important. As a result, the computation of this component should not be omitted. The estimation of V_{sam} and V_{imp} may be done as follows:

- (i) Let $\hat{V}_p$ be the variance estimator under full response and let $\hat{V}_I$ be the “naive” variance estimator of $\bar{y}_I$. That is, $\hat{V}_I$ denotes the naive variance estimator of the imputed estimator obtained by treating the imputed values as if they were observed. It is well known that for several imputation methods (in particular, deterministic methods), $\hat{V}_I$ underestimates V_{sam} . To compensate for this underestimation, evaluate the following expectation

$$E_m (\hat{V}_p - \hat{V}_I | s) \equiv V_{dif}.$$

Then, for given s and s_r , determine a model unbiased estimator, denoted $\hat{V}_{dif}$, of V_{dif} . This will usually require the estimation of certain parameters of the assumed imputation model. Then

$$\hat{V}_{sam} = \hat{V}_I + \hat{V}_{dif}$$

is unbiased for V_{sam} in the sense $E_p E_m (\hat{V}_{sam}|s) = V_{sam}$.

- (ii) Determine a model unbiased estimator, denoted $\hat{V}_{imp}$, of V_{imp} . Again, this may require the estimation of unknown parameters of the imputation model m . We have,

$$E_m (\hat{V}_{imp}|s) = V_{imp}.$$

Finally, an overall unbiased estimator of V_{tot} , denoted by $\hat{V}_{tot}$, is given by

$$\hat{V}_{tot} = \hat{V}_I + \hat{V}_{dif} + \hat{V}_{imp}.$$

Note that in the case of random imputation methods (e.g., random hot-deck imputation and random regression imputation), the component $\hat{V}_I$ is approximately unbiased for V_{sam} so that the derivation of $\hat{V}_{dif}$ may be omitted in this case. Deville and Särndal (1994) extended the method to the case of general sampling designs.

Under ratio imputation and simple random sampling without replacement and assuming $N \rightarrow \infty$, Särndal (1992) obtained the variance estimator

$$v_S (\bar{y}_I - \bar{Y}) = \left(\frac{\bar{z}}{\bar{z}_r} \right)^2 \frac{A}{r} + 2 \left(\frac{r}{n} \right) \frac{B}{n} + \frac{C}{n}, \quad (1.41)$$

where A, B and C are given respectively by (1.36), (1.37) and (1.38).

1.10.3 The Reverse Approach

To illustrate the reverse approach, consider ratio imputation and simple random sampling without replacement. To estimate the total variance (1.31), we estimate the two components separately.

- (i) Estimation of $V_1 = E_r V_p (\bar{y}_I - \bar{Y} | \mathbf{a})$: It suffices to estimate $V_p (\bar{y}_I - \bar{Y} | \mathbf{a})$. To this end, first write the imputed estimator (1.21) as a function of the response indicators a_i which gives

$$\bar{y}_I = \frac{\sum_{i \in s} a_i y_i}{\sum_{i \in s} a_i z_i} \bar{z} = \psi \left(\sum_{i \in s} a_i y_i, \sum_{i \in s} a_i z_i, \sum_{i \in s} z_i \right), \quad (1.42)$$

where $\psi(\cdot)$ is a smooth function of totals. The problem is now reduced to estimating the sampling variance of a smooth function of totals based on the full sample. Using Taylor linearization, an approximately unbiased and consistent estimator, v_1 , of V_1 is given by

$$v_1 \approx \left(1 - \frac{n}{N}\right) \left[\left(\frac{\bar{z}}{\bar{z}_r}\right)^2 \frac{A}{r} + 2 \left(\frac{\bar{z}}{\bar{z}_r}\right) \frac{B}{n} + \frac{C}{n} \right], \quad (1.43)$$

where A, B and C are given respectively by (1.36), (1.37) and (1.38).

- (ii) Estimation of $V_2 = V_r E_p (\bar{y}_I - \bar{Y} | \mathbf{a})$ under Assumption DB: Under this approach, we have $V_r(a_i) = p(1-p)$. Using Taylor linearization, V_2 is given by

$$V_2 \approx \left[\frac{\bar{Z}}{E_r(Z_r)} \right]^2 p(1-p) \frac{1}{N} \sum_{i \in U} (y_i - R z_i)^2, \quad (1.44)$$

where $Z_r = \sum_{i \in U} a_i z_i$. A consistent estimator v_{2DB} of V_2 is obtained by substituting estimators for the unknown quantities in (1.44), which leads to

$$v_2 = \left(\frac{\bar{z}}{\bar{z}_r}\right)^2 \left(\frac{n}{N} - \frac{r}{N}\right) \frac{A}{r}. \quad (1.45)$$

An estimator of the total variance under Assumption DB is given by

$$\begin{aligned} v_t = v_1 + v_{2DB} &= \left(1 - \frac{r}{N}\right) \left(\frac{\bar{z}}{\bar{z}_r}\right)^2 \frac{A}{r} + 2 \left(1 - \frac{n}{N}\right) \left(\frac{\bar{z}}{\bar{z}_r}\right) \frac{B}{n} \\ &\quad + \left(1 - \frac{n}{N}\right) \frac{C}{n}. \end{aligned} \quad (1.46)$$

(iii) Estimation of $V_2 = V_r E_p (\bar{y}_I - \bar{Y} | \mathbf{a})$ under Assumption MB: Here, inferences are based on the imputation model (1.16). We have,

$$V_2 = E_r V_m E_p (\bar{y}_I - \bar{Y} | \mathbf{a}) + V_r E_m E_p (\bar{y}_I - \bar{Y} | \mathbf{a}). \quad (1.47)$$

The second component in the right hand side of (1.47) is approximately equal to 0, so V_2 reduces to

$$V_2 \approx E_r V_m \left(\sum_{i \in U} c_i y_i \right), \quad (1.48)$$

where $c_i = a_i \left(\frac{\bar{z}}{\bar{z}_r} - \frac{1}{N} \right)$. It follows that

$$V_2 \approx \sigma^2 \sum_{i \in U} E_r (c_i) z_i. \quad (1.49)$$

Under Assumption MB, a consistent estimator, v_{2MB} of V_2 , is given by

$$v_{2MB} = \left[\frac{n}{r} \left(\frac{\bar{z}}{\bar{z}_r} \right)^2 - \left(\frac{\bar{z}}{\bar{z}_r} \right) \right] \frac{A}{N}. \quad (1.50)$$

Note that v_{2MB} can be expressed as

$$v_{2MB} = v_{2DB} + \frac{r}{N} \left(\frac{\bar{z}}{\bar{z}_r} \right) \left(\frac{\bar{z}}{\bar{z}_r} - 1 \right). \quad (1.51)$$

It follows that $v_{2MB} \approx v_{2DB}$ under uniform response, since in this case, $\frac{\bar{z}}{\bar{z}_r} \rightarrow 1$ in probability. Also, note that the ratio $\frac{v_1}{v_{2DB}}$ is equal to

$$\frac{v_1}{v_{2DB}} = \frac{N \left(1 - \frac{n}{N} \right)}{n \left(1 - \frac{r}{N} \right)} + \frac{N}{n} \times O(1) \rightarrow \infty, \text{ as } \frac{n}{N} \rightarrow 0.$$

Hence, the second component v_{2DB} is negligible with respect to the first component when the sampling fraction n/N is negligible.

1.10.4 Jackknife Method

The jackknife method was originally proposed by Quenouille (1949) as a method of reducing bias. Tukey (1958) used it for variance estimation and construction of confidence intervals. We first describe the jackknife in the case of complete response for stratified multistage designs. We first assume that the first-stage units are selected with replacement. The jackknife method proceeds by calculating a set of replicate estimates, derived from a subset of the sample data, and estimating the variance using the replicate estimates. Let θ be a parameter of interest and $\hat{\theta}$ be its estimator based of the complete sample s and weights w_{hij} . Let $\hat{\theta}_{(gl)}$ denote the estimator obtained from $\hat{\theta}$, using the jackknife weights $w_{hij}b_{(gl)}$ and

$$b_{(gl)} = \begin{cases} 1 & \text{if } h \neq g \\ \frac{n_g}{n_g - 1} & \text{if } h = g, i \neq l \\ 0 & \text{if } i = l. \end{cases}$$

Note that $\hat{\theta}_{(gl)}$ is obtained by deleting the (gl) -th cluster. A jackknife variance estimator of $\hat{\theta}$ is then given by

$$v_J(\hat{\theta}) = \sum_{g=1}^L \left(\frac{n_g - 1}{n_g} \right) \sum_{i=1}^{n_h} \left(\hat{\theta}_{(gl)} - \hat{\theta} \right)^2. \quad (1.52)$$

This method is called *delete-cluster* jackknife. In the presence of nonresponse to item y , we consider the case $\theta = \bar{Y}$. The use of (1.52), treating the imputed values as true values, may lead to serious underestimation of the variance of the estimator, especially if the nonresponse rate is appreciable. Rao and Shao (1992) proposed an *adjusted jackknife* method that may be applied in the case of weighted deterministic imputation (mean, ratio and regression) as well as for weighted random hot-deck

imputation. Under weighted random hot-deck imputation, the Rao-Shao adjusted jackknife is calculated in the usual way except that whenever the (gl) -th cluster is deleted, we adjust the imputed values, y_{hij}^* to

$$y_{hij}^{*(gl)} = y_{hij}^* + [E_{*(gl)}(y_{hij}^*) - E_*(y_{hij}^*)],$$

where $E_{*(gl)}(.)$ denotes the expectation with respect to the imputation mechanism when the $(gl)^{th}$ cluster has been deleted. The adjusted imputed values $y_{hij}^{*(gl)}$ reflect the fact that the donor set is changed when a cluster is deleted from the sample. In the case of deterministic imputation, $y_{hij}^{*(gl)}$ reduces to $E_{*(gl)}(y_{hij}^*)$. Therefore, the adjustment is the same as reimputing missing values in the pseudo-replicates obtained by deleting the $(gl)^{th}$ cluster, using the donors in that pseudo-replicate.

Using the adjusted imputed values, the Rao-Shao jackknife variance estimator is given by

$$v_{JRS}(\bar{y}_I) = \sum_{g=1}^L \left(\frac{n_g - 1}{n_g} \right) \sum_{i=1}^{n_h} (\bar{y}_{I(gl)}^a - \bar{y}_I)^2, \quad (1.53)$$

where $\bar{y}_{I(gl)}^a$ is computed the same way as $\bar{y}_I$ in (1.4) but with the adjusted imputed values $y_{hij}^{*(gl)}$ instead of the imputed values y_{hij}^* . Cohen (2002) proposes a method for implementing the method using extra records and standard jackknife software.

Note that the jackknife method assumes that the clusters are selected with replacement or that the overall first-stage sampling fraction $\sum n_h / \sum N_h$ is small. Also, note that the jackknife finds a natural justification under the reverse approach since $v_{JRS}(\bar{y}_I)$ is an estimator of the component V_1 . Hence, it is valid under both Assumption DB and Assumption MB.

1.10.5 Bootstrap Method

The bootstrap method proposed by Efron (1979) is a useful replication method for obtaining variance estimators for complex parameters. Rao and Wu (1988) extended the bootstrap to stratified multistage designs. Let θ be a smooth or non-smooth parameter to be estimated. We select a sample, s , of size n , according to a stratified multistage design. Let $\hat{\theta}_I$ be an imputed estimator of θ based on observed and imputed values. Applying the Rao-Wu bootstrap in the presence of missing responses, treating the missing values as true values, may lead to serious underestimation of the variance of the estimator. Under imputation for missing data, Shao and Sitter (1996) proposed a bootstrap for imputed survey data that may be described as follows:

- i) Draw a bootstrap sample (i.e., simple random sample with replacement) s^* of size $n^* = n - 1$ from s after imputation.
- ii) Let a_i^* be the response indicator for unit i in s^* . Let $s_r^* = \{i \in s^* : a_i^* = 1\}$ and let $s_m^* = \{i \in s^* : a_i^* = 0\}$. Apply the same imputation procedure used to obtain the imputed estimator $\hat{\theta}_I$ for reimputing the nonrespondent values in s^* (i.e., the missing values in s_m^* , using the bootstrap donor set s_r^*).
- iii) Compute the imputed estimator $\hat{\theta}_I^*$, from the imputed data in (ii).
- iv) Repeat (i)-(iii) B times to get $\hat{\theta}_{I1}^*, \dots, \hat{\theta}_{IB}^*$.

A bootstrap variance estimator is then given by

$$v_B(\hat{\theta}_I) = \sum_{b=1}^B \left(\hat{\theta}_{Ib}^* - \bar{\theta}_I^* \right)^2, \quad (1.54)$$

where $\bar{\theta}_I^* = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_{Ib}^*$. Note that the bootstrap, like the jackknife, assumes that the clusters are selected with replacement or that the first-stage overall sampling fractions $\sum n_h / \sum N_h$ is small. Finally, when the within-stratum cluster sample size, n_h is small (say, $n_h = 2$), the Shao-Sitter bootstrap may lead to heavily biased variance estimators. Saigo, Shao and Sitter (2001) modified the Shao-Sitter procedure to overcome this difficulty.

1.10.6 Multiple Imputation

Multiple imputation was originally proposed by Rubin (1978) to handle missing data. For each nonrespondent, M imputations are generated, resulting in M completed data sets, which allows the analyst to use standard techniques of analysis designed for complete data. Analyses are performed separately on each of the M completed data sets and the results are then combined to get point estimates as well as variance estimates. Let θ be a parameter of interest and $\hat{\theta}$ be an estimator of θ under full response and let U be the variance estimator associated with $\hat{\theta}$.

Let $\hat{\theta}_I^{(t)}$ and $U_I^{(t)}$ be the point estimator and variance estimator using the t -th imputed data set, treating the imputed values as true values, $t = 1, \dots, M$. The multiple imputed estimator of $\hat{\theta}$ is given by

$$\hat{\theta}_{I,M} = \frac{1}{M} \sum_{t=1}^M \hat{\theta}_I^{(t)}. \quad (1.55)$$

The variance estimator associated with $\hat{\theta}_{I,M}$ is given by

$$T_M = U_M + \frac{M+1}{M} B_M \quad (1.56)$$

where

$$U_M = \frac{1}{M} \sum_{t=1}^M U_I^{(t)} \quad (1.57)$$

and

$$B_M = \frac{1}{M-1} \sum_{t=1}^M \left(\hat{\theta}_I^{(t)} - \hat{\theta}_{I,M} \right)^2.$$

Inferences are then based on the approximation

$$T_M^{-1/2} \left(\hat{\theta} - \theta \right) \sim t_\nu,$$

where the degrees of freedom, ν , is given by

$$\nu = (M-1) \left[1 + \frac{1}{r} \right]^2$$

with $r = \frac{(M-1)B_M}{U_M}$. Rubin (1987) interprets r as *the relative increase in variance due to nonresponse*.

For inferences to be valid under the frequentist approach, the imputation method should be proper (Rubin 1987, p. 118). An imputation method is called proper under the assumed response mechanism if

$$E_r \left(\hat{\theta}_{I,\infty} | \mathbf{y} \right) = \hat{\theta} \quad (1.58)$$

$$E_r \left(\bar{U}_\infty | \mathbf{y} \right) = \bar{U} \quad (1.59)$$

$$E_r \left(B_\infty | \mathbf{y} \right) = V_r \left(\bar{\theta}_\infty | \mathbf{y} \right), \quad (1.60)$$

where $\mathbf{y} = (\mathbf{y}_{obs}, \mathbf{y}_{mis})$ with $\mathbf{y}_{obs}$ and $\mathbf{y}_{mis}$ denoting the observed part and missing part, respectively. The conditions (1.58)-(1.60) are not easy to verify and may not hold generally (Binder and Sun, 1996). Under the Bayesian approach, the imputed values are

generated from the conditional distribution $P(\mathbf{y}_{mis}|\mathbf{y}_{obs})$, using iterative procedures such as Gibbs sampling (Schafer, 1997) and assuming an imputation model.

There has been some controversy about the validity of multiple imputation. One problem is that the definition of a proper imputation depends upon the specific analysis of interest; what is proper for one analysis may not be proper for another (Fay, 1992). Also, Wang and Robins (1998) showed with a simple example that, even when the imputation is proper, multiple imputation can be inefficient when we turn our attention to interval estimation. We now give this example. Let $y_1, \dots, y_n$ be n independent and identically distributed random variables such that $y_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$. Wang and Robins generated nonresponse to item y according to an uniform response mechanism with probabilities 0.25 and 0.5. They considered two imputation methods:

- a) Type A: The imputed values are drawn from a normal distribution with parameters $\hat{\mu}_{MLE}$ and $\hat{\sigma}_{MLE}^2$, where $\hat{\mu}_{MLE}$ and $\hat{\sigma}_{MLE}^2$ denote respectively the maximum likelihood estimators of μ and σ^2 . This method is not proper in the sense of Rubin. Wang and Robins (1998) give asymptotically correct variance estimator of the imputed estimator.
- b) Type B: This method is proper in the sense of Rubin and may be described as follows:
 - i) For each of the imputations $k = 1, 2, \dots, M$, draw a realization from a χ_{r-1}^2 distribution and let

$$\sigma_{(k)}^{*2}|\mathbf{y}_{obs} \sim \hat{\sigma}_{MLE}^2(r-1)/\chi_{r-1}^2$$

ii) Draw a random variable of the distribution

$$\mu_{(k)}^* | (\mathbf{y}_{obs}, \sigma_{(k)}^*) \sim N(\hat{\mu}_{MLE}, \sigma_{(k)}^{*2}/r)$$

iii) For each of the missing values $i \in s_m$, draw

$$e_{i(k)}^{**} | (\mu_{(k)}^*, \sigma_{(k)}^*) \sim N(0, \sigma_{(k)}^{*2}), i = 1, \dots, M.$$

Let $y_{i(k)}^{**} = \mu_{(k)}^* + e_{i(k)}^{**}$ be the k -th imputed value for missing y_i , $k = 1, \dots, M$.

vi) Repeat (i)-(iii) M independent times.

Table 1 gives the coverage probability as well as the median length of the confidence intervals at a nominal rate of 95%, with $n = 50$, $M = 3$ and $p = 0.25; 0.5$.

It is clear from Table 1 that the median length of Type A normal theory interval can be much smaller than the median length of Rubin's t -intervals, for small M .

Table 1.1: Coverage probability and median length of the confidence intervals at a nominal rate of 95%

	Coverage probability	Length of C.I
Type A ($p = 0.25$)	97.8	1.00
Type B ($p = 0.25$)	98.0	1.87
Type A ($p = 0.0.5$)	96.5	1.00
Type B ($p = 0.5$)	94.3	1.52

1.10.7 Fractionally Weighted Imputation

Fay (1996) discussed fractionally weighted imputation (FWI), originally proposed by Kalton and Kish (1984), as an alternative to multiple imputation (MI). FWI is similar to MI in the sense that each missing value is replaced by $M \geq 2$ imputed values, but may be distinguished as follows: (a) Under MI, the imputation method should be proper for the estimators to be valid whereas FWI uses methods such as random hot-deck imputation that are not proper. (b) Under MI, the point estimator is given by (1.55) whereas under FWI, the point estimator is of the form (1.4) with y_i^* replaced by the average of the M imputed values for $i \in s_m$. (c) Under MI, the variance estimator is given by (1.56), whereas FWI uses the Rao-Shao type adjusted jackknife or a similar method.

FWI assigns M imputed values y_{il}^* , $l = 1, \dots, M$ for each $i \in s_m$. Each imputed value receives $1/M$ times the original weight. In this case, the imputed estimator

(1.4) becomes

$$\bar{y}_{FWI} = \frac{1}{\sum_{i \in s} w_i} \left[\sum_{i \in s_r} w_i y_i + \sum_{i \in s_m} w_i \left(\frac{1}{M} \sum_{l=1}^M y_{il}^* \right) \right]. \quad (1.61)$$

The advantage of FWI over single random hot-deck imputation is that FWI leads to a reduction of the imputation variance as we show now. Consider simple random sampling without replacement and assumption DB. Under the two-phase approach, the variance of an imputed estimator under random imputation is given by (1.30). First, note that under single random hot-deck imputation (with replacement), we have $E_*(y_i^*) = \bar{y}_r$ and $V_*(y_i^*) = \frac{r-1}{r} s_{yr}^2$. Hence, under single random hot-deck imputation, the variance of the imputed estimator (1.4) is given by

$$\begin{aligned} V(\bar{y}_I - \bar{Y}) &= V_p E_r(\bar{y}_r - \bar{Y} | s, r) + E_p V_r E_*(\bar{y}_r - \bar{Y} | s, r) \\ &\quad + \left[1 - E_p \left(\frac{r}{n} \right) \right] \frac{S_y^2}{n}, \end{aligned} \quad (1.62)$$

for large r .

If random hot-deck (with replacement) imputation is used to select the M imputed values, we have $E_*(\bar{y}_{FWI}) = \bar{y}_r$, so the variance of the imputed estimator (1.61) is given by

$$\begin{aligned} V(\bar{y}_{FWI} - \bar{Y}) &\approx V_p E_r(\bar{y}_r - \bar{Y} | s, r) + E_p V_r E_*(\bar{y}_r - \bar{Y} | s, r) \\ &\quad + \left[1 - E_p \left(\frac{r}{n} \right) \right] \frac{S_y^2}{Mn}, \end{aligned} \quad (1.63)$$

for large r . A comparison of (1.62) and (1.63) shows that the first two terms on the right hand side are identical. However, the third term on the right hand side of (1.62) is M times larger than the corresponding component in (1.63), showing the advantage of FWI over single random hot-deck imputation. Expression (1.63) shows that the

component $E_p E_r V_* (\bar{y}_r - \bar{Y} | s, r)$ goes to 0 as $M \rightarrow \infty$. Hence, FWI may considerably reduce the variance due random imputation when M is large. For variance estimation, Fay (1996) proposed a natural extension of the Rao-Shao adjusted jackknife (Rao and Shao, 1992).

1.11 Nearest-Neighbour Imputation

Although nearest-neighbour imputation (NNI) is frequently used, its theoretical properties were not fully known until recently in the survey sampling context. Chen and Shao (2000, 2001) studied the properties of the imputed estimator and proposed a method for variance estimation in the presence of imputed values. In the case of NNI, the imputed estimator (1.4) of the population mean $\bar{Y}$ becomes

$$\bar{y}_{NNI} = \frac{1}{\sum_{i \in s} w_i} \left[\sum_{i \in s_r} w_i y_i + \sum_{j \in s_m} w_j \left(\sum_{i \in s_r} d_{ij} y_i \right) \right], \quad (1.64)$$

where

$$d_{ij} = \begin{cases} 1 & \text{if } i \text{ is the nearest neighbour of } j \\ 0 & \text{otherwise.} \end{cases}$$

Chen and Shao (2000) assumed MAR and the following general imputation model

$$y_i = f(z_i) + \epsilon_i, \quad (1.65)$$

$$E_m(\epsilon_i) = 0, E_m(\epsilon_i \epsilon_j) = 0, i \neq j, V_m(\epsilon_i) = \sigma_i^2,$$

where $f(\cdot)$ is a smooth function. Under this set-up, the imputed estimator (1.64) is asymptotically unbiased for $\bar{Y}$; that is,

$$E(\bar{y}_{NNI} - \bar{Y}) = E_p E_m(\bar{y}_{NNI} - \bar{Y}) = O_p(n^{-1/2}).$$

Since $f(z_i)$ and $V_m(\epsilon_i)$ do not have parametric forms, Chen and Shao (2001) proposed to estimate the variance using a nonparametric variance estimation method. The simplest nonparametric resampling method for variance estimation is the jackknife. For imputation methods such as regression imputation or random regression imputation, Rao and Shao's (1992) adjusted jackknife leads to consistent variance estimators. However, in the case of NNI, Chen and Shao (2001) showed that the Rao-Shao adjusted jackknife overestimates the variance of $\bar{y}_{NNI}$. Therefore, they proposed to use a partially adjusted jackknife that works in the same way as the Rao-Shao jackknife except that, when a responding unit $i \in s_r$ is deleted, the imputed values are adjusted to

$$y_i^{*a} = y_i^* + t_i (E_{*(j)}(y_i) - y_i^*),$$

where $t_i = \frac{\sqrt{6(d_j)^2 + 6d_j + 4} - 2}{3d_j}$ and $d_j = \sum_{t \in s_m} \frac{w_t}{w_j} d_{jt}$. It is easily seen that $0 \leq t_i \leq 1$ so that the adjustment is a partial adjustment. Chen and Shao (2001) showed that, under some regularity conditions, the partially adjusted jackknife variance estimator is consistent.

1.12 Estimation of relationships

So far, we have discussed the problem of inference for population totals and population means. In practice, it is of interest to estimate more complex parameters such as domain means, regression coefficients and coefficients of correlation. Skinner and Rao (2002) have considered the problem of estimating bivariate parameters of the

form

$$\theta_N = \frac{1}{N} \sum_{i \in U} g(x_i, y_i) \quad (1.66)$$

which include common parameters as special cases. For example, the mean $\bar{Y}$ is a special case of (1.66) with $g(x_i, y_i) = y_i$. Both variables x and y are now potentially missing. Let s_{rr} be the set of units who responded to both x and y , of size n_{rr} , s_{rm} be the set of units who responded to x but not y , of size n_{rm} , s_{mr} be the set of units who responded to y but not x , of size n_{mr} , and s_{mm} be the set of units who did not respond to either x or y , of size n_{mm} . In the case of simple random sampling without replacement, an imputed estimator of θ_N is given by

$$\hat{\theta}_I = \frac{1}{n} \sum_{i \in s} g(\tilde{x}_i, \tilde{y}_i), \quad (1.67)$$

where $\tilde{x}_i = x_i$ if $i \in s_{rr} \cup s_{rm}$, $\tilde{x}_i = x_i^*$ if $i \in s_{mr} \cup s_{mm}$, $\tilde{y}_i = y_i$ if $i \in s_{rr} \cup s_{mr}$, $\tilde{y}_i = y_i^*$ if $i \in s_{rm} \cup s_{mm}$. Skinner and Rao (2002) showed that under uniform response and common donor imputation, the imputed estimator $\hat{\theta}_I$ is biased. They proposed a bias-adjusted estimator $\hat{\theta}_I^a$ which is approximately unbiased for θ_N . We do not present their results in more details here since we extend their results in Chapter 6 to the case of stratified multistage designs.

Shao and Wang (2002) considered the problem of estimating the finite population coefficient of correlation between x and y that may be expressed as

$$\rho_{xy} = \frac{\sum_{i \in U} x_i y_i - N \bar{Y} \bar{X}}{[(\sum_{i \in U} y_i^2 - N \bar{Y}^2)(\sum_{i \in U} x_i^2 - N \bar{X}^2)]^{\frac{1}{2}}}.$$

They showed that the usual random regression imputation does not preserve unbiasedness for coefficients of correlation. They propose a joint regression imputation

method that preserves unbiasedness for coefficients of correlation under assumption MB. Once again, we do not present their procedure in details here since we study its validity under the design-based approach in Chapter 6.

1.13 Outline of the Thesis

This dissertation is organized as follows:

In Chapter 2, the problem of unweighted imputation is considered. We show that, in the case of unequal probability designs, unweighted imputation generally leads to biased estimators for population means under the design-based approach. A robust bias-adjusted estimator is then proposed and its correct variance is derived. A simulation study is conducted to assess the performance of different estimators. A paper based on Chapter 2 is published in *Survey Methodology* (Haziza and Rao, 2003).

In Chapter 3, we make inference on a vector of domain totals (and means). We show that the imputed estimator of a vector of domain totals is biased under the design-based approach. We propose a bias-adjusted estimator which is approximately unbiased under both the design-based and the model-based approaches. Variance estimators are derived using the reverse approach of Fay (1991). A simulation study is performed to assess the performance of alternative estimators. The case of unweighted imputation is also considered. Finally, a second class of bias-adjusted estimators is proposed which are approximately unbiased under both the design-based and model-based approaches and a more general imputation model. A paper based on Chapter

3 is published in the *Canadian Journal of Statistics* (Haziza and Rao, 2005).

In Chapter 4, we propose a modified regression imputation method that leads to unbiased imputed estimators under ignorable response. The method is shown to be robust in the sense of validity under the usual approaches used for inference. Domain estimation is also considered. Under modified regression imputation, we obtain a bias-adjusted estimator of a vector of domain totals. Two simulation studies are performed in order to assess the performance of the modified imputation method. A paper based on Chapter 4 is tentatively accepted for publication in *Survey Methodology* (Haziza and Rao, 2005).

In Chapter 5, we consider the problem of variance estimation in the case of two-stage cluster sampling. In this case, we show that the usual approaches to inference may not be tenable because the intracluster correlations are not taken into account. We derive correct variance estimators under both Särndal's method (Särndal, 1990) and Shao-Steel's method (Shao and Steel, 1999). A simulation study is then performed to compare the performance of the two methods. Finally, imputation based on empirical best linear unbiased prediction (EBLUP), under a random effects imputation model, is studied.

In Chapter 6, we consider estimating bivariate parameters such as domain means, coefficients of correlation and regression coefficients. Under different versions of random hot-deck imputation, we show that the imputed estimator of bivariate parameters under multistage sampling is asymptotically biased. Following Skinner and Rao (2002), we propose a bias-adjusted estimator which is robust in the sense of validity under both the design-based or the model-based approaches. Comparisons of the

different imputation method are also provided. Finally, we show that the procedure proposed by Shao and Wang (2002) is valid under the design-based approach.

In Chapter 7, we consider the problem of constructing imputation classes. First, we give a conditional justification for the formation of the classes. An extensive simulation study is then performed in order to assess the performance of two methods frequently used in practice for the formation of the classes: the score method and the cross-classification method.

In Chapter 8, asymptotic consistency as well as asymptotic normality of some imputed estimators are established.

Finally, some suggestions for further research are presented in Chapter 9.

Chapter 2

Inference under unweighted imputation

2.1 Introduction

In the presence of item nonresponse, weighted or unweighted imputation may be used. Weighted (deterministic or stochastic) imputation uses the sampling weights induced by the sampling design to select donors. However, weighted imputation is not feasible in practice when the sampling weights are not available at the imputation stage. Note that unweighted and weighted imputation methods lead to identical results for self-weighting designs (i.e., designs with equal weights). Also, unweighted imputation methods are appealing to users.

Unweighted imputation generally leads to biased estimators under uniform response within imputation classes. In this chapter, following the approach of Skinner and Rao (2002), we propose bias-adjusted estimators of population means under

unweighted imputation. We then derive linearization variance estimators using the reverse approach of Fay (1991) and Shao and Steel (1999). As noted in Chapter 1, Fay's reverse approach facilitates variance estimation.

We set out our basic framework and assumptions in section 2.2. In section 2.3, we study both weighted and unweighted ratio imputation. We show that the imputed estimator under unweighted imputation is asymptotically biased, and propose a bias-adjusted estimator. The estimator under weighted imputation and the bias-adjusted estimator under unweighted imputation are shown to be robust in the sense of validity under both the design-based and model-based approaches. We also derive linearization variance estimators of the imputed estimators in section 2.3. We consider the case of random hot-deck imputation in section 2.4. A small simulation is conducted in section 2.5 to compare the performances of the imputed estimators in terms of bias and mean squared error. Relative bias and relative stability of the variance estimators are also studied.

2.2 Framework and Assumptions

Let U be a finite population of possibly unknown size N . The objective is to estimate the population mean $\bar{Y} = \frac{1}{N} \sum_{i \in U} y_i$ when imputation has been used to compensate for nonresponse. Suppose a probability sample, s , of size n is selected according to a specified design $p(s)$ from U and let π_i for $i \in U$ be the inclusion probabilities. Let s_r be the sample of respondents of size r and let s_m be the sample of nonrespondents of size m ; $r + m = n$. For simplicity, we assume the case of a single imputation cell

but the extension to multiple imputation cells is straightforward.

The usual imputed estimator of the population mean $\bar{Y}$ is given by

$$\bar{y}_I = \frac{1}{\sum_{i \in s} w_i} \left[\sum_{i \in s_r} w_i y_i + \sum_{i \in s_m} w_i y_i^* \right], \quad (2.1)$$

where w_i is the sampling (or design) weight attached to unit i and y_i^* denotes the value imputed for missing y_i . We use the Horvitz-Thompson weight $w_i = 1/\pi_i$, where π_i is the probability of including unit i in the sample.

We consider the two approaches presented in Chapter 1: (i) design-based and (ii) model-based. Under the design-based approach, we assume a uniform response mechanism within cells so that the following assumption holds:

Assumption DB: Within an imputation cell, the response probability for a given variable of interest is constant and the response statuses for different units are independent.

Under the model-based approach, the following assumption holds:

Assumption MB: Within an imputation cell the response mechanism is ignorable or unconfounded in the sense that the response status of a unit does not depend on the variable being imputed but may depend on covariates used for imputation. In this case, an imputation model is assumed.

Under ratio imputation, the imputation model used is the “ratio model” given by (1.16). Under random hot-deck imputation, the imputation model used is given by (1.17).

2.3 Ratio Imputation

In this section, we study the properties of the imputed estimator (2.1) under both weighted and unweighted ratio imputation. We also derive linearization variance estimators. We study point estimation in section 2.3.1 under weighted and unweighted ratio imputation, and corresponding variance estimation in section 2.3.2.

2.3.1 Estimation of a Mean

Weighted Ratio Imputation

Weighted ratio imputation uses $y_i^* = \hat{R}_r z_i$ for missing y_i , where $\hat{R}_r = \frac{\bar{y}_r}{\bar{z}_r}$ and $(\bar{y}_r, \bar{z}_r) = \frac{1}{\sum_{i \in s_r} w_i} \sum_{i \in s_r} w_i (y_i, z_i)$ are the weighted means of respondents for variables y and z respectively. Using the y_i^* 's, the imputed estimator (2.1) reduces to

$$\bar{y}_{IR} = \hat{R}_r \bar{z}, \quad (2.2)$$

where $\bar{z} = \frac{1}{\sum_{i \in s} w_i} \sum_{i \in s} w_i z_i$. It is easy to verify that $\bar{y}_{IR}$ is approximately unbiased for $\bar{Y}$ under assumption DB or assumption MB, (Särndal, 1992). Hence $\bar{y}_{IR}$ is robust in the sense of validity under both approaches.

Unweighted Ratio Imputation

Unweighted ratio imputation uses $y_i^* = \hat{R}_r^{un} z_i$ for missing y_i , where $\hat{R}_r^{un} = \frac{\bar{y}_r^{un}}{\bar{z}_r^{un}}$ and $(\bar{y}_r^{un}, \bar{z}_r^{un}) = \frac{1}{r} \sum_{i \in s_r} (y_i, z_i)$ are the unweighted means of respondents for variables y and z respectively. Using the y_i^* 's, the imputed estimator (2.1) reduces to

$$\bar{y}_{IR} = \frac{1}{\sum_{i \in s} w_i} \left[\sum_{i \in s_r} w_i y_i + \hat{R}_r^{un} \sum_{i \in s_m} w_i z_i \right]. \quad (2.3)$$

Under the ratio model (1.16) and assumption MB, the imputed estimator (2.3) is approximately unbiased for $\bar{Y}$, i.e., $E_p E_r E_m (\bar{y}_{IR}) \approx E_m (\bar{Y})$. However, it is biased under uniform response (assumption DB). We have $E_p E_r (\bar{y}_{IR}) \approx p\bar{Y} + (1-p) \frac{\bar{Y}_\pi}{\bar{Z}_\pi} \bar{Z}$, where $(\bar{Y}_\pi, \bar{Z}_\pi) = \frac{1}{\sum_{i \in U} \pi_i} \sum_{i \in U} \pi_i (y_i, z_i)$. Hence, the relative bias of $\bar{y}_{IR}$, $RB(\bar{y}_{IR}) = [E_p E_r (\bar{y}_{IR}) - \bar{Y}] / \bar{Y}$, is given by

$$RB(\bar{y}_{IR}) \approx (1-p) \left[\frac{\bar{Z}}{\bar{Z}_\pi} \frac{\bar{Y}_\pi}{\bar{Y}} - 1 \right] \quad (2.4)$$

$$= (1-p) \frac{\bar{Z}}{\bar{Z}_\pi} C_\pi [C_y \rho_{\pi y} - C_z \rho_{\pi z}], \quad (2.5)$$

where $\bar{Z} = \frac{1}{N} \sum_{i \in U} z_i$, $\rho_{\pi y}$ and $\rho_{\pi z}$ are the finite population correlation coefficients between the variables π and y and π and z respectively, C_π , C_z and C_y are respectively the coefficients of variation of π , z and y , and p is the probability of response to y . The bias is nonzero generally. It vanishes in the full response case (i.e., $p = 1$) or if

$$C_\pi [C_y \rho_{\pi y} - C_z \rho_{\pi z}] = 0, \quad (2.6)$$

which is satisfied when $C_\pi = 0$ (the case when the design is self-weighting) or when

$$\frac{\rho_{\pi y}}{\rho_{\pi z}} = \frac{C_z}{C_y}. \quad (2.7)$$

We further explore the relative bias (2.4) for three cases. First, we consider unweighted mean imputation, $y_i^* = \bar{y}_r^{un}$, which is a special case of unweighted ratio imputation with $z_i = 1$. Assume that a size variable x is available for all the units in the population and that the sample s is selected according to a probability proportional to size (pps) sampling without replacement design, using x as the size, such that

$\pi_i = nx_i/X$, where $X = \sum_{i \in U} x_i$. For example, one may use the well-known Rao-Sampford method (Rao, 1965; Sampford, 1967). Noting that $\rho_{\pi y} = \rho_{xy}$, $\bar{Z}/\bar{Z}_\pi = 1$ and $C_\pi = C_x$, expression (2.5) for the relative bias may be written as

$$RB(\bar{y}_{IR}) \approx (1-p) C_x C_y \rho_{xy}. \quad (2.8)$$

Two particular cases of (2.8) are of interest. First, if x and y are uncorrelated the bias of the imputed estimator vanishes. The case of weakly correlated x and y (i.e., $\rho_{xy} \approx 0$) may occur in surveys with multiple characteristics y (Rao, 1966). Second, if $y_i \propto x_i$, the relative bias (2.8) reduces to $(1-p) C_x^2$ which decreases with C_x . Note that, since $C_x = C_\pi$, the sampling design approaches a self-weighting design as C_x decreases.

Consider next the more general case of unweighted ratio imputation based on z_i , $i \in s$, and pps sampling based on x_i , $i \in U$. In this case, the relative bias (2.5) is zero if and only if

$$\frac{\rho_{xy}}{\rho_{xz}} = \frac{C_y}{C_z},$$

provided $p < 1$ and $C_\pi \neq 0$. If $C_y = C_z$, then the relative bias (2.5) is zero if and only if $\rho_{xy} = \rho_{xz}$.

Finally, we consider the case of stratified random sampling. In this case, the population U is partitioned into L strata U_h with N_h sampling units in the h -th stratum; $U = \bigcup_{h=1}^L U_h$, $N = \sum_{h=1}^L N_h$. We then independently select a simple random sample without replacement s_h of size n_h from each stratum; $s = \bigcup_{h=1}^L s_h$ and $n = \sum_{h=1}^L n_h$. Two situations may occur in practice: (1) Imputation is done independently in each stratum (i.e., the imputation cells coincide with the strata). In

this case, under unweighted ratio imputation, the imputed estimator is approximately unbiased under uniform response within strata. (2) The imputation is done across strata. In this case, we note from (2.5) that the imputed estimator is approximately unbiased if and only if $n_h = n \frac{N_h}{N}$ (proportional allocation) which leads to a self-weighting design.

A bias-adjusted estimator of $\bar{Y}$ under unweighted ratio imputation and assumption DB is given by

$$\bar{y}_{IR}^a = \hat{p}^{-1} \bar{y}_{IR} + (1 - \hat{p}^{-1}) \frac{\bar{z}}{\bar{z}^{un}} \bar{y}_{IR}^{un}, \quad (2.9)$$

where $\hat{p} = \frac{\sum_{i \in s_r} w_i}{\sum_{i \in s} w_i}$ is a consistent estimator of the response probability p , $\bar{z}^{un} = \frac{1}{n} \sum_{i \in s} z_i$ and $\bar{y}_{IR}^{un}$ is the unweighted mean of the observed values y_i and the imputed values $y_i^* = \hat{R}_r^{un} z_i$. This estimator may be derived from the method of moments, following Skinner and Rao (2002), by solving

$$E(\bar{y}_{IR}) = E_P E_r(\bar{y}_{IR}) = p \bar{Y} + (1 - p) \frac{\bar{Y}_\pi}{\bar{Z}_\pi} \bar{Z}$$

for $\bar{Y}$ and replacing $E(\bar{y}_{IR})$ by its estimator $\bar{y}_{IR}$, $\frac{\bar{Y}_\pi}{\bar{Z}_\pi} \bar{Z}$ by its estimator

$$\hat{R}_r^{un} \bar{z} = \left(\frac{\bar{z}}{\bar{z}^{un}} \right) \bar{y}_{IR}^{un}, \quad (2.10)$$

and p^{-1} by its estimator $\hat{p}^{-1}$. Note that the estimator $\bar{z}$ of $\bar{Z}$ makes use of the full sample z -values, unlike $\bar{z}_r$. If $\bar{z}_r$ is used to estimate $\bar{Z}$, then the bias-adjusted estimator requires response identifiers, unlike (2.9).

We now show that the bias-adjusted estimator (2.9) is approximately unbiased under both the design-based and the model-based approaches. Hence, unlike the unadjusted estimator (2.3), the adjusted estimator is robust in the sense of validity

under both approaches. First, noting that $\bar{y}_{IR}$ may be expressed as $\hat{p}\bar{y}_r + R_r^{un}(\bar{z} - \hat{p}\bar{z}_r)$ and using (2.10), the bias-adjusted estimator (2.9) reduces to

$$\bar{y}_{IR}^a = \bar{y}_r + \hat{R}_r^{un}(\bar{z} - \bar{z}_r). \quad (2.11)$$

Comparing (2.2) and (2.11), we see that $\bar{y}_{IR}$ under weighted ratio imputation is not equal to the bias-adjusted estimator $\bar{y}_{IR}^a$ under unweighted ratio imputation, unless $z_i = 1$ for all i . In the latter case, both estimators reduce to $\bar{y}_r$. Note that, the form (2.9) for $\bar{y}_{IR}^a$ does not require response identifiers, provided $\hat{p}$ is available, unlike (2.11).

Since $E_m(\bar{y}_{IR}^a) = \beta\bar{z}$ and $E_m(\bar{Y}) = \beta\bar{Z}$ under the ratio model (1.16), we have $E_p E_m(\bar{y}_{IR}^a - \bar{Y}) \approx 0$; that is, the adjusted estimator is approximately unbiased under the model-based approach. On the other hand, since $E_p E_r(\bar{y}_r) \approx \bar{Y}$ and $E_r(\bar{z} - \bar{z}_r) \approx 0$ under uniform response, it follows from (2.11) that $E_p E_r(\bar{y}_{IR}^a) \approx \bar{Y}$ so that the adjusted estimator is approximately unbiased under uniform response.

We note several points here: (1) The survey analyst can easily implement the adjusted estimator $\bar{y}_{IR}^a$, given by (2.9), from the imputed data file without response identifiers, i.e., $(w_i, \tilde{y}_i, z_i, i \in s)$, where $\tilde{y}_i = y_i$ if $i \in s_r$ and $\tilde{y}_i = y_i^*$ if $i \in s_m$. Note that the response identifiers are not needed on the data file, but the response rate $\hat{p}$ should be available to the analyst, which we assume to be the case here. In the case of multiple imputation classes, response rates within classes and imputation class identifiers need to be provided with the file. (2) The bias-adjusted estimator coincides with the unadjusted estimator $\bar{y}_{IR}$, given by (2.2), under a self-weighting design $w_i = w$. (3) The adjusted estimator $\bar{y}_{IR}^a$ in (2.11) has the form of a regression

estimator in two-phase sampling. (4) Under mean imputation, (2.11) reduces to the weighted mean of respondents $\bar{y}_r$, so the correction made to the unadjusted estimator eliminates the effect of using unweighted mean imputation.

Another approach to getting a bias-adjusted estimator, $\bar{y}_{IR}^a$, is to subtract an estimator, $b(\bar{y}_{IR})$, of the bias of $\bar{y}_{IR}$, from $\bar{y}_{IR}$, i.e.,

$$\bar{y}_{IR}^a = \bar{y}_{IR} - b(\bar{y}_{IR}). \quad (2.12)$$

It follows from (2.4) that an estimator of the bias of $\bar{y}_{IR}$ is given by

$$b^{(1)}(\bar{y}_{IR}) = (1 - \hat{p}) \left(\hat{R}_r^{un} \bar{z} - \bar{y}_r \right). \quad (2.13)$$

But the resulting bias-adjusted estimator is not identical to (2.9), and it depends on response identifiers, unlike (2.9). On the other hand, if one uses

$$b^{(2)}(\bar{y}_{IR}) = (1 - \hat{p}) \left(\hat{R}_r^{un} \bar{z}_r - \bar{y}_r \right), \quad (2.14)$$

it is easy to verify that the resulting bias-adjusted estimator is identical to (2.9).

2.3.2 Variance estimation

We study variance estimation under uniform response in this section. We assume that response identifiers are available with the variance estimation file. If imputation classes are used, their identifiers are also needed.

Variance Estimation under Weighted Ratio Imputation

In this subsection, we obtain a linearization variance estimator of the imputed estimator (2.2) based on weighted ratio imputation, using the reverse approach of Fay

(1991). First, express (2.2) as

$$\bar{y}_{IR} = \frac{\sum_{i \in s} w_i a_i y_i}{\sum_{i \in s} w_i a_i z_i} \bar{z},$$

where a_i is a response indicator to item y such that $a_i = 1$ if $i \in s_r$ and $a_i = 0$, otherwise. It follows from (1.31) that the variance $V(\bar{y}_{IR})$ of $\bar{y}_{IR}$ can be estimated by $v_t = v_1 + v_2$, where v_1 is an estimator of $V_p(\bar{y}_{IR} - \bar{Y} | \mathbf{a})$ conditional on the vector of response indicators $\mathbf{a}$, and v_2 is an estimator of $V_r E_p(\bar{y}_{IR} - \bar{Y} | \mathbf{a})$. Denote the estimator of the variance of the estimated total $\hat{Y} = \sum_{i \in s} w_i y_i$ based on the full sample as $v(y)$. Then, using the delta method, a linearization variance estimator, v_1 , in the operator notation $v(\cdot)$, is given by

$$v_1 = v(\hat{\xi}), \quad (2.15)$$

where the value of $\hat{\xi}$ for $i \in s$ is given by

$$\hat{\xi}_i = \frac{1}{\sum_{i \in s} w_i} [\hat{\xi}_{1i} - \bar{y}_{IR}],$$

with

$$\hat{\xi}_{1i} = a_i y_i + (1 - a_i) \hat{R}_r z_i + \hat{c} a_i (y_i - \hat{R}_r z_i),$$

where

$$\hat{c} = \frac{\sum_{i \in s} w_i (1 - a_i) z_i}{\sum_{i \in s} w_i a_i z_i}.$$

Note that v_1 is valid regardless of the response mechanism and the imputation model.

The derivation of (2.15) is given in Appendix 2.7.1. Shao and Steel (1999) derived a linearization variance estimator of the imputed estimator $\hat{Y} = \sum_s w_i a_i y_i + \sum_s w_i (1 - a_i) \hat{R}_r z_i$ of the total Y . They first expressed $\hat{Y}$ as

$$\hat{Y} = \sum_{i \in s} w_i [a_i y_i + (1 - a_i) R_r z_i] + \hat{c} \sum_{i \in s} w_i a_i (y_i - R_r z_i),$$

where $R_r = \frac{\bar{Y}_r}{\bar{Z}_r}$ with $(Y_r, Z_r) = \sum_{i \in U} a_i (y_i, z_i)$ and then replaced $\hat{c}$ by $\bar{c} = \frac{\sum_{i \in U} (1-a_i)z_i}{\sum_{i \in U} a_i z_i}$ to get a linear approximation for $\hat{Y} \approx \sum_{i \in s} w_i \eta_i$, where

$$\eta_i = a_i y_i + (1 - a_i) R_r z_i + \bar{c} a_i (y_i - R_r z_i).$$

Now replacing R_r by $\hat{R}_r$ and $\bar{c}$ by $\hat{c}$ in the above expression for η_i we get $\hat{\eta}_i = a_i y_i + (1 - a_i) \hat{R}_r z_i + \hat{c} a_i (y_i - \hat{R}_r z_i)$ which leads to the linearization variance estimator $v_1 = v(\hat{\eta})$. The delta method in Appendix 2.7.1 may be used to obtain this result in a straightforward manner.

Next, using the delta method,

$$V_r E_p (\bar{y}_{IR} - \bar{Y} | \mathbf{a}) \approx p(1-p) \left(\frac{Z}{E_r(Z_r)} \right)^2 \frac{S_e^2}{N}, \quad (2.16)$$

where $Z = \sum_{i \in U} z_i$, $Z_r = \sum_{i \in U} a_i z_i$ and $S_e^2 = \frac{1}{N} \sum_{i \in U} (y_i - E_r(R_r) z_i)^2$. The component v_2 is then obtained by substituting estimators for the unknown quantities in (2.16). We obtain

$$v_2 = \hat{p}(1-\hat{p}) \left(\frac{\hat{Z}}{\hat{Z}_r} \right)^2 \frac{s_{er}^2}{\hat{N}}, \quad (2.17)$$

where $\hat{Z} = \sum_{i \in s} w_i z_i$, $\hat{Z}_r = \sum_{i \in s} w_i a_i z_i$, $\hat{N} = \sum_{i \in s} w_i$ and

$$s_{er}^2 = \frac{1}{\sum_{i \in s} w_i a_i} \sum_{i \in s} w_i a_i (y_i - \hat{R}_r z_i)^2.$$

The sum of (2.15) and (2.17) gives v_t , the estimator of the overall variance of $\bar{y}_{IR}$.

Variance Estimation under Unweighted Ratio Imputation

We now give a linearization estimator of variance of the imputed estimator (2.3) based on unweighted ratio imputation. Using the delta method, see Appendix 2.7.1,

we obtain

$$v_1 = v \left(\hat{\xi} \right), \quad (2.18)$$

where

$$\hat{\xi}_i = \frac{1}{\sum_{i \in s} w_i} \left[\hat{\xi}_{1i} - \bar{y}_{IR} \right],$$

with

$$\hat{\xi}_{1i} = a_i y_i + (1 - a_i) \hat{R}_r^{un} z_i + \hat{d} \frac{a_i}{w_i} \left(y_i - \hat{R}_r^{un} z_i \right)$$

and $\hat{d} = \frac{\sum_{i \in s} w_i (1 - a_i) z_i}{\sum_{i \in s} a_i z_i}$. The component v_2 is given by (2.39) in Appendix 2.7.2.

Variance Estimation for the Bias-Adjusted Estimator

In this subsection, we give a linearization variance estimator of the bias-adjusted estimator (2.9). Using the delta method, we obtain

$$v_1 = v \left(\hat{\xi} \right), \quad (2.19)$$

where

$$\hat{\xi}_i = \frac{a_i}{\sum_{i \in s} w_i a_i} \left[(y_i - \bar{y}_r) + \hat{R}_r^{un} (z_i - \bar{z}_r) \right] + \frac{\hat{R}_r^{un}}{\hat{N}} (z_i - \bar{z}) + (\bar{z} - \bar{z}_r) \frac{1}{\sum_{i \in s} a_i z_i} \frac{a_i}{w_i} \left(y_i - \hat{R}_r^{un} z_i \right);$$

see Appendix 2.7.1. The component v_2 is given by (2.37) in Appendix 2.7.3.

2.4 Random Hot-Deck Imputation

In this section, we study the properties of the imputed estimator (2.1) under weighted and unweighted random hot-deck imputation. We also derive consistent variance estimators under uniform response.

2.4.1 Estimation of a Mean

In section 2.4.1 we study point estimators under weighted and unweighted random hot-deck imputation.

Weighted Random Hot-Deck Imputation

Under weighted random hot-deck imputation, we select the donors $j \in s_r$ with replacement with selection probabilities $\frac{w_j}{\sum_{i \in s_r} w_i}$ and use $y_i^* = y_j$, $i \in s_m$. The imputed estimator, $\bar{y}_{IH}$, is given by (2.1) with the above imputed values. It is approximately unbiased for the population mean $\bar{Y}$ under both assumption DB and assumption MB. The latter uses the mean model (1.17).

Unweighted Random Hot-Deck Imputation

Under unweighted random hot-deck imputation, we select the donors $j \in s_r$ with replacement with equal probabilities $1/r$ and use $y_i^* = y_j$, $i \in s_m$. The imputed estimator, $\bar{y}_{IH}$, is given by (2.1) with the above imputed values. It is approximately unbiased for $\bar{Y}$ under the mean model (1.17), but biased under uniform response. The bias of $\bar{y}_{IH}$ is given by

$$B(\bar{y}_{IH}) = (1 - p)(\bar{Y}_\pi - \bar{Y}). \quad (2.20)$$

A bias-adjusted estimator of $\bar{Y}$ under unweighted random hot-deck imputation is given by

$$\bar{y}_{IH}^a = \hat{p}^{-1} \bar{y}_{IH} + (1 - \hat{p}^{-1}) \bar{y}_{IH}^{un}, \quad (2.21)$$

where $\hat{p} = \frac{\sum_{i \in s_r} w_i}{\sum_{i \in s} w_i}$ is a consistent estimator of the response probability p and $\bar{y}_{IH}^{un}$ is the unweighted mean of the observed values y_i and the imputed values y_i^* . The estimator (2.21) may be derived from the method of moments, following Skinner and Rao (2002), by solving

$$E(\bar{y}_{IH}) = p\bar{Y} + (1 - p)\bar{Y}_\pi$$

for $\bar{Y}$, replacing $E(\bar{y}_{IH})$ by its estimator $\bar{y}_{IH}$, $\bar{Y}_\pi$ by its estimator $\bar{y}_{IH}^{un}$ and p^{-1} by its estimator $\hat{p}^{-1}$. The adjusted estimator is approximately unbiased for $\bar{Y}$ under both the design-based and the model-based approaches. As in section 2.3.1, note that the survey analyst can easily implement the adjusted estimator $\bar{y}_{IR}^a$ from the imputed data file without response identifiers, i.e., $(w_i, \tilde{y}_i, z_i, i \in s)$, where $\tilde{y}_i = y_i$ if $i \in s_r$ and $\tilde{y}_i = y_i^*$ if $i \in s_m$, provided the response rate, $\hat{p}$, is available.

Note that the method of subtracting an estimator of the bias of $\bar{y}_I$ from $\bar{y}_I$, using (2.20), will lead to a bias-adjusted estimator that depends on response identifiers, unlike (2.21). It is not possible to obtain the bias-adjusted estimator (2.21) by this approach, unlike in the case of deterministic ratio imputation studied in subsection 2.3.1.

2.4.2 Variance Estimation

We study variance estimation under uniform response in this section. We assume that response identifiers are available with the variance estimation file. If imputation classes are used, their identifiers are also needed.

Variance Estimation under Weighted Random Hot-Deck Imputation

We now obtain a linearization variance estimator of the imputed estimator $\bar{y}_{IH}$ under weighted random hot-deck imputation. First, note that under weighted random hot-deck imputation, $E_*(\bar{y}_{IH}) = \bar{y}_r$. This is a particular case of (2.2) with $z_i = 1$ for all i . Hence, using (2.15), v_1 is given by

$$v_1 = v\left(\hat{\xi}\right), \quad (2.22)$$

where

$$\begin{aligned} \hat{\xi}_i &= \frac{1}{\sum_{i \in s} w_i} \left[\hat{\xi}_{1i} - \bar{y}_r \right], \\ \hat{\xi}_{1i} &= a_i y_i + (1 - a_i) \bar{y}_r + \hat{c} a_i (y_i - \bar{y}_r), \end{aligned}$$

with $\hat{c} = \frac{\sum_{i \in s} w_i (1 - a_i)}{\sum_{i \in s} w_i a_i}$. Straightforward algebra shows that $\hat{\xi}_i$ simplifies to $\hat{\xi}_i = \frac{1}{\sum_{i \in s} w_i a_i} a_i (y_i - \bar{y}_r)$. Now, noting that $V_*(y_i^*) = \frac{1}{\sum_{i \in s} w_i a_i} \sum_{i \in s} w_i a_i (y_i - \bar{y}_r)^2 \equiv s_{yr}^2$, we have

$$v_* = \frac{\sum_{i \in s} w_i^2 (1 - a_i)}{\left(\sum_{i \in s} w_i\right)^2} s_{yr}^2. \quad (2.23)$$

The component v_2 is the same as for the deterministic case. Hence, under weighted random hot-deck imputation, v_2 is given by (2.17) with $z_i = 1$ for all i , which leads to

$$v_2 = \hat{p}(1 - \hat{p}) \frac{s_{yr}^2}{\hat{N}}. \quad (2.24)$$

The sum of (2.22), (2.23) and (2.24) gives v_i , the estimator of overall variance.

Variance Estimation under Unweighted Random Hot-Deck Imputation

We now obtain a linearization estimator of variance of the imputed estimator (2.1) under unweighted random hot-deck imputation. First, note that $E_*(\bar{y}_{IH})$ reduces to (2.3) with $z_i = 1$ for all i . Hence, v_1 is given by

$$v_1 = v(\hat{\xi}), \quad (2.25)$$

where

$$\begin{aligned} \hat{\xi}_i &= \frac{1}{\sum_s w_i} [\hat{\xi}_{1i} - E_*(\bar{y}_{IH})], \\ \hat{\xi}_{1i} &= a_i y_i + (1 - a_i) \bar{y}_r^{un} + \hat{d} \frac{a_i}{w_i} (y_i - \bar{y}_r^{un}), \end{aligned}$$

with $\hat{d} = \frac{\sum_{i \in s} w_i (1 - a_i)}{\sum_{i \in s} a_i}$. Now, noting that $V_*(y_i^*) = \frac{1}{\sum_{i \in s} a_i} \sum_{i \in s} a_i (y_i - \bar{y}_r^{un})^2 \equiv s_{yr}^{2un}$, the component $E_r E_p V_*(\bar{y}_{IH} - \bar{Y} | \mathbf{a})$ can be estimated by

$$v_* = \frac{\sum_{i \in s} w_i^2 (1 - a_i)}{(\sum_{i \in s} w_i)^2} s_{yr}^{2un}. \quad (2.26)$$

As noted in section 1, v_2 is the same as for the deterministic case. Hence, under unweighted random hot-deck imputation, v_2 is given by (2.17) with $z_i = 1$ for all i , which reduces to

$$v_2 = \hat{p}(1 - \hat{p}) \frac{s_{yr}^{2un}}{\hat{N}} \quad (2.27)$$

The sum of (2.25), (2.26) and (2.27) gives v_t .

Variance Estimation for the Bias-Adjusted Estimator

We now obtain a linearization variance estimator of the bias-adjusted estimator given by (2.21). First, note that, $E_*(\bar{y}_{IH}^a)$ reduces to $\bar{y}_r$, the mean of the y -values respondent. Hence, v_1 is given by (2.22) and v_2 is given by (2.24). Now, noting that

$V_*(y_i^*) = s_{yr}^{2un}$ and $Cov_*(y_i^*, y_j^*) = 0$ for $i \neq j$, one can show that $E_r E_p V_*(\bar{y}_{IH}^2 - \bar{Y}|\mathbf{a})$ can be estimated by

$$v_* = \left[\frac{\hat{p}^{-2}}{(\sum_{i \in s} w_i)^2} \sum_{i \in s} w_i^2 (1 - a_i) - (1 - \hat{p}^{-1})^2 \left(\frac{r + n}{n^2} \right) \right] s_{yr}^{2un}. \quad (2.28)$$

The sum of (2.25), (2.26) and (2.28) gives v_t . Note that even though v_* given by (2.28) is expressed as the difference between two terms, it is always nonnegative, as shown in Appendix 2.7.4.

2.5 Simulation Study

As a complement to the theory, we present some results from a limited simulation study. We generated a population of $N = 800$ item values (y_i, z_i) according to the ratio model $y = \beta z + \epsilon$, where z and ϵ were generated from a normal distribution such that the correlation, ρ_{yz} , between y and z equaled 0.05, 0.30, 0.70 and 0.90. The objective is to estimate the population total $Y = \sum_{i \in U} y_i$. We drew $R = 10000$ samples, each of size $n = 75$, according to Rao-Sampford's probability proportional to size (pps) sampling method, using z as the measure of size. Nonresponse to item y was then generated from each pps sample according to a uniform response mechanism with a response rate of 0.7; item z was observed for all units in the sample. We used random hot-deck imputation to compensate for nonresponse to item y .

The estimator of the first component in the variance formula (1.31) was computed using the well known Sen-Yates-Grundy estimator. Let $v(\xi)$ denote the variance

estimator of $\sum_s w_i \xi_i$. The Sen-Yates-Grundy estimator of variance is then given by

$$v(\xi) = -\frac{1}{2} \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{\xi_i}{\pi_i} - \frac{\xi_j}{\pi_j} \right)^2, \quad (2.29)$$

where $\pi_i = P(i \in s)$ is the probability of inclusion of unit i in the sample and $\pi_{ij} = P(i \in s \text{ and } j \in s)$ is the joint probability of inclusion of units i and j in the sample. Sampford's method ensures $\pi_{ij} - \pi_i \pi_j \leq 0$ for all i, j so that the variance estimator in (2.29) is always nonnegative.

As a measure of the bias of an imputed estimator $\hat{Y}_I$ of Y , we used the bias ratio $B_r(\hat{Y}_I) = \text{Bias}(\hat{Y}_I) / \text{s.e}(\hat{Y}_I)$, where $\text{s.e}(\hat{Y}_I)$ denotes the standard error of $\hat{Y}_I$. To compare the efficiencies, we used the coefficient of variation of $\hat{Y}_I$, denoted $CV(\hat{Y}_I)$ and given by $CV(\hat{Y}_I) = \frac{\sqrt{MSE}}{Y}$. The variance estimators were compared in terms of their relative bias and CV. The relative bias of a variance estimator, v_t , is measured by $B_{rel}(v_t) = \frac{E(v_t) - MSE(\hat{Y}_I)}{MSE(\hat{Y}_I)}$ and its CV is given by $CV(v_t) = \frac{\sqrt{MSE(v_t)}}{MSE(\hat{Y}_I)}$. Values of the above measures were calculated from the simulated pps samples.

Table 2.1 reports the simulation results on the bias ratio (B_r) of the three imputed estimators of Y , denoted $B_r(\text{weighted})$, $B_r(\text{unweighted})$ and $B_r(\text{adjusted})$ and the CVs of the estimators, denoted $CV(\text{weighted})$, $CV(\text{unweighted})$ and $CV(\text{adjusted})$. It is clear from Table 2.1 that the bias ratio of the estimator under unweighted imputation is large ($\geq 30\%$) if $\rho_{xy} \geq 0.5$, while the bias ratios of the estimator under weighted imputation and the adjusted estimator are small ($\leq 4\%$) for all values of ρ_{xy} . Due to large bias, the CV of the unweighted estimator is larger than the CV of the weighted estimator if $\rho_{xy} \geq 0.5$ and also larger than the CV of the adjusted estimator if $\rho_{xy} \geq 0.7$, but the increase in CV is not large. Also, $CV(\text{weighted})$ is

slightly smaller than CV (adjusted) for all values of ρ_{xy} .

Table 2.2 reports the relative bias (B_{rel}) and the CV ratios of the variance estimators. As expected, the variance estimator v_t (unweighted) leads to serious underestimation of MSE of the estimator for large ρ_{xy} (≥ 0.7), while the absolute relative bias of the variance estimators v_t (weighted) and v_t (adjusted) is small ($\leq 6\%$) for all values of ρ_{xy} . Turning to the CV ratios of the variance estimators, Table 2.2 shows that v_t (unweighted) has the smallest CV followed by v_t (weighted) and v_t (adjusted) for $\rho_{xy} \geq 0.3$.

Table 2.1: Bias ratio (%) and CV (%) of the imputed estimators

	$\rho_{xy} = 0.05$	$\rho_{xy} = 0.30$	$\rho_{xy} = 0.50$	$\rho_{xy} = 0.70$	$\rho_{xy} = 0.90$
B_r (weighted)	-0.78	1.99	-0.79	0.40	3.27
B_r (unweighted)	1.82	18.6	30.5	49.2	64.2
B_r (adjusted)	-1.12	1.47	0.01	0.61	2.94
CV (weighted)	18.8	15.3	11.6	5.87	4.69
CV (unweighted)	18.0	15.2	12.5	6.83	5.93
CV (adjusted)	20.9	16.8	13.5	6.10	4.78

2.6 Remarks

If the imputation flags are available in the data file and imputation is deterministic, the imputed values can be replaced by those under weighted imputation. For example, in the case of unweighted ratio imputation, $y_i^* = \frac{\bar{y}_r^{un}}{\bar{z}_r^{un}} z_i$, one could either multiply each imputed value by $\frac{\bar{z}_r^{un}}{\bar{y}_r^{un}} \frac{\bar{y}_r}{\bar{z}_r}$ to reproduce the values $\frac{\bar{y}_r}{\bar{z}_r} z_i$ under weighted ratio imputation, provided edits are not applied after imputation. Alternatively, one could reimpute values using the sampling weights w_i . In both cases, the adjusted estimator does not present advantages over the imputed estimator based on weighted imputation other than assuring that the imputed values in the data file are not changed. In the case of random hot-deck imputation, however, the only way to implement weighted random hot-deck imputation is to reimpute using a weighted hot-deck scheme. We believe that analysts do not like to change the imputed values on the data file produced by

Table 2.2: Relative bias (%) of the variance estimators and comparisons of the CV ratios of the variance estimators

	$\rho_{xy} = 0.05$	$\rho_{xy} = 0.30$	$\rho_{xy} = 0.50$	$\rho_{xy} = 0.70$	$\rho_{xy} = 0.90$
$B_{rel}(v_t)(\text{weighted})$	-2.43	-4.78	-4.28	3.96	-1.95
$B_{rel}(v_t)(\text{unweighted})$	-1.03	-3.47	-11.8	-18.5	-29.3
$B_{rel}(v_t)(\text{adjusted})$	-5.42	-1.06	-4.21	1.61	0.07
$\frac{CV(v_t)(\text{unweighted})}{CV(v_t)(\text{weighted})}$	1.016	0.9846	0.931	0.875	0.781
$\frac{CV(v_t)(\text{unweighted})}{CV(v_t)(\text{adjusted})}$	1.032	0.829	0.701	0.819	0.692
$\frac{CV(v_t)(\text{weighted})}{CV(v_t)(\text{adjusted})}$	1.016	0.843	0.751	0.935	0.886

the edit and imputation system.

The imputed estimator (2.1) can use poststratification (or calibration) weights, $\tilde{w}_i(s)$, based on known population auxiliary information, instead of design weights w_i . Note that the calibration weights, $\tilde{w}_i(s)$, depend on the whole sample s unlike the design weights w_i . If the calibration weights are used for ratio imputation, then we simply replace w_i by $\tilde{w}_i(s)$ in section 2.3.2 and the resulting linearization variance estimator, v_1 , uses $\hat{\xi}$ in (2.15) with w_i changed to $\tilde{w}_i(s)$. However, $v(\cdot)$ in (2.15) now refers to the linearization variance estimator of the full sample post-stratified estimator $\sum_{i \in s} \tilde{w}_i(s) y_i$.

Under unweighted imputation, linearization variance estimation becomes more complex because the bias-adjusted estimator based on the calibration weights will

involve both design weights and calibration weights. If the design weights, w_i , are available at the imputation stage but not the calibration weights, $\tilde{w}_i(s)$, the design weights can be used for imputation and the calibration weights for estimation. The resulting imputed estimator (2.1) based on calibration weights remains asymptotically unbiased under uniform response (within cells), but linearization variance estimation becomes more complex because both sets of weights are involved in the imputed estimator.

2.7 Appendix

2.7.1 Derivation of v_1

Suppose that an estimator $\hat{\theta}$ is expressed as

$$\hat{\theta} = \frac{1}{\hat{Y}_1} \left[\hat{Y}_2 + \frac{\hat{Y}_3}{\hat{Y}_4} (\hat{Y}_5 - \hat{Y}_6) \right] =: g(\hat{\mathbf{Y}}), \quad (2.30)$$

where $\hat{Y}_j = \sum_{i \in s} w_i y_{ji}$, $j = 1, \dots, 6$ and $\hat{\mathbf{Y}} = (\hat{Y}_1, \dots, \hat{Y}_6)'$. Letting $\theta = g(\mathbf{Y})$, $R_{34} = Y_3/Y_4$, $\hat{Y}_j = Y_j (1 + \delta\hat{Y}_j)$ with $\delta\hat{Y}_j = \frac{\hat{Y}_j - Y_j}{Y_j}$ and $Y_j = E_p(\hat{Y}_j)$, we have

$$\begin{aligned} \hat{\theta} - \theta &= \frac{1}{Y_1 (1 + \delta\hat{Y}_1)} \left\{ Y_2 (1 + \delta\hat{Y}_2) \right. \\ &\quad \left. + R_{34} \frac{(1 + \delta\hat{Y}_3)}{(1 + \delta\hat{Y}_4)} [Y_5 (1 + \delta\hat{Y}_5) - Y_6 (1 + \delta\hat{Y}_6)] \right\} - \theta \\ &\approx \frac{1}{Y_1} \left\{ (\delta\hat{Y}_2 - \delta\hat{Y}_1) Y_2 + R_{34} Y_5 (\delta\hat{Y}_3 - \delta\hat{Y}_4) \right\} - \delta\hat{Y}_1 \\ &\quad - R_{34} Y_6 (\delta\hat{Y}_3 - \delta\hat{Y}_4 + \delta\hat{Y}_6 - \delta\hat{Y}_1) \Big\}, \quad (2.31) \end{aligned}$$

neglecting higher order terms in $\delta\hat{Y}_j$'s. The expression (2.36) reduces to

$$\begin{aligned}\hat{\theta} - \theta &\approx \frac{1}{Y_1} \left\{ \hat{Y}_2 + R_{34} (\hat{Y}_5 - \hat{Y}_6) + \frac{Y_5 - Y_6}{Y_4} (\hat{Y}_3 - R_{34} \hat{Y}_4) - \theta \hat{Y}_1 \right\} \\ &= \sum_{i \in s} w_i \xi_i,\end{aligned}$$

where

$$\xi_i = \frac{1}{Y_1} (\xi_{1i} - \theta) \quad (2.32)$$

with

$$\xi_{1i} = y_{2i} + R_{34} (y_{5i} - y_{6i}) + \frac{Y_5 - Y_6}{Y_4} (y_{3i} - R_{34} y_{4i}).$$

Hence, the variance estimator of $\hat{\theta}$ from the delta method may be expressed as $v(\xi)$.

Now, replacing unknown quantities in (2.34) by their estimators, we get

$$\text{estvar}(\hat{\theta}) = v(\hat{\xi}),$$

where

$$\hat{\xi}_i = \frac{1}{\hat{Y}_1} (\hat{\xi}_{1i} - \hat{\theta}) \quad (2.33)$$

with

$$\hat{\xi}_{1i} = y_{2i} + \hat{R}_{34} (y_{5i} - y_{6i}) + \frac{\hat{Y}_5 - \hat{Y}_6}{\hat{Y}_4} (y_{3i} - \hat{R}_{34} y_{4i}).$$

Note that the delta method avoids evaluation of partial derivatives of $g(\hat{\mathbf{Y}})$ with respect to its components Y_j , unlike the usual Taylor linearization method.

Letting $\hat{Y}_1 = \sum_{i \in s} w_i$, $\hat{Y}_2 = \hat{Y}_3 = \sum_{i \in s} w_i a_i y_i$, $\hat{Y}_4 = \hat{Y}_6 = \sum_{i \in s} w_i a_i z_i$ and $\hat{Y}_5 = \sum_{i \in s} w_i z_i$ in (2.35), we get the variance estimator (2.15) of $\bar{y}_{IR}$ based on weighted ratio imputation. Also, letting $\hat{Y}_1 = \sum_{i \in s} w_i$, $\hat{Y}_2 = \sum_{i \in s} w_i a_i y_i$, $\hat{Y}_3 = \sum_{i \in s} w_i a_i (y_i/w_i)$,

$\hat{Y}_4 = \sum_{i \in s} w_i a_i (z_i/w_i)$, $\hat{Y}_5 = \sum_{i \in s} w_i z_i$ and $\hat{Y}_6 = \sum_{i \in s} w_i a_i z_i$ in (2.35), we get the variance estimator (2.18) of $\bar{y}_{IR}$ based on unweighted imputation. Finally, we note that the bias-adjusted estimator (2.9) written in the form (2.11) can be expressed as the sum of three components: $\bar{y}_r$, $\hat{R}_r^{un} \bar{z}$ and $-\hat{R}_r^{un} \bar{z}_r$. Each of these components is a special case of (2.35). Indeed, the component $\bar{y}_r$ is a special case of (2.35) with $\hat{Y}_1 = \sum_{i \in s} w_i a_i$, $\hat{Y}_2 = \sum_{i \in s} w_i a_i y_i$ with $\hat{Y}_5 = \hat{Y}_6$. The component $\hat{R}_r^{un} \bar{z}$ is a special case of (2.35) with $\hat{Y}_1 = \sum_{i \in s} w_i$, $\hat{Y}_2 = \hat{Y}_3 = \sum_{i \in s} w_i a_i (y_i/w_i)$, $\hat{Y}_4 = \hat{Y}_6 = \sum_{i \in s} w_i a_i (z_i/w_i)$ and $\hat{Y}_5 = \sum_{i \in s} w_i z_i$. The component $\hat{R}_r^{un} \bar{z}_r$ is a special case of (2.35) with $\hat{Y}_1 = \sum_{i \in s} w_i a_i$, $\hat{Y}_2 = \hat{Y}_3 = \sum_{i \in s} w_i a_i (y_i/w_i)$, $\hat{Y}_4 = \hat{Y}_6 = \sum_{i \in s} w_i a_i (z_i/w_i)$ and $\hat{Y}_5 = \sum_{i \in s} w_i a_i z_i$. We apply the delta method to each component separately to obtain $v_1 = v(\xi)$ given by (2.19).

2.7.2 Derivation of v_2 for the estimator $\bar{y}_{IR}$ under unweighted imputation

Using the delta method, it can be shown that $V_r E_p (\bar{y}_{IR} - \bar{Y})$ under unweighted ratio imputation is given by

$$\begin{aligned} V_r E_p (\bar{y}_{IR} - \bar{Y}) \approx & (1-p) \frac{1}{N} \left[S_{e(1)}^2 + \left(\frac{Z - E_r(Z_r)}{E_r(Z_{\pi r})} \right)^2 S_{e(2)}^2 \right. \\ & \left. + 2 \left(\frac{Z - E_r(Z_r)}{E_r(Z_{\pi r})} \right) S_{e(3)}^2 \right], \end{aligned} \quad (2.34)$$

where

$$\begin{aligned} S_{e(1)}^2 &= \frac{1}{N} \sum_{i \in U} (y_i - E_r(R_{\pi r}) z_i)^2, \\ S_{e(2)}^2 &= \frac{1}{N} \sum_{i \in U} \pi_i^2 (y_i - E_r(R_{\pi r}) z_i)^2, \\ S_{e(3)}^2 &= \frac{1}{N} \sum_{i \in U} \pi_i (y_i - E_r(R_{\pi r}) z_i)^2, \end{aligned}$$

with $R_{\pi r} = \frac{Y_{\pi r}}{Z_{\pi r}}$ and $(Y_{\pi r}, Z_{\pi r}) = \sum_{i \in U} \pi_i a_i (y_i, z_i)$. The component v_2 is obtained by estimating unknown quantities in (2.38). It is given by

$$v_2 \approx \hat{p}(1 - \hat{p}) \frac{1}{\hat{N}} \left[s_{er(1)}^2 + \left(\frac{\hat{Z} - \hat{Z}_r}{\hat{Z}_{\pi r}} \right)^2 s_{er(2)}^2 + 2 \left(\frac{\hat{Z} - \hat{Z}_r}{\hat{Z}_{\pi r}} \right) s_{er(3)}^2 \right], \quad (2.35)$$

where

$$\begin{aligned} s_{er(1)}^2 &= \frac{1}{\sum_{i \in s} w_i a_i} \sum_{i \in s} w_i a_i (y_i - \hat{R}_r^{un} z_i)^2, \\ s_{er(2)}^2 &= \frac{1}{\sum_{i \in s} w_i a_i} \sum_s w_i^{-1} a_i (y_i - \hat{R}_r^{un} z_i)^2, \\ s_{er(3)}^2 &= \frac{1}{\sum_{i \in s} w_i a_i} \sum_{i \in s} a_i (y_i - \hat{R}_r^{un} z_i)^2 \end{aligned}$$

and $\hat{Z}_{\pi r} = \sum_{i \in s} a_i z_i$.

2.7.3 Derivation of v_2 for the estimator $\bar{y}_{IR}^a$

Using the delta method, it can be shown that $V_r E_p (\bar{y}_{IR}^a - \bar{Y})$ for the bias-adjusted estimator is given by

$$\begin{aligned} V_r E_p (\bar{y}_{IR}^a - \bar{Y}) &\approx p(1 - p) \frac{N}{E_r \left[\left(\sum_{i \in U} a_i \right)^2 \right]} \left\{ S_{yr}^2 + E_r(R_{\pi r}^2) S_{zr}^2 \right. \\ &\quad \left. - 2E_r(R_{\pi r}) S_{y zr} + E_r \left[h \sum_{i \in U} a_i \right]^2 S_{e(2)}^2 \right. \\ &\quad \left. + 2E_r \left[h \sum_{i \in U} a_i \right] (S_{\pi ey} - E_r(R_{\pi r}) S_{\pi ez}) \right\}, \quad (2.36) \end{aligned}$$

where

$$\begin{aligned}
S_{yr}^2 &= \frac{1}{N} \sum_{i \in U} (y_i - E_r(\bar{Y}_r))^2, \\
S_{zr}^2 &= \frac{1}{N} \sum_{i \in U} (z_i - E_r(\bar{Z}_r))^2, \\
S_{yzt} &= \frac{1}{N} \sum_{i \in U} (y_i - E_r(\bar{Y}_r)) (z_i - E_r(\bar{Z}_r)), \\
S_{e(2)}^2 &= \frac{1}{N} \sum_{i \in U} \pi_i^2 (y_i - E_r(R_{\pi r}) z_i)^2, \\
S_{\pi ey} &= \frac{1}{N} \sum_{i \in U} \pi_i (y_i - E_r(\bar{Y}_r)) (y_i - E_r(R_{\pi r}) z_i), \\
S_{\pi ez} &= \frac{1}{N} \sum_{i \in U} \pi_i (z_i - E_r(\bar{Z}_r)) (y_i - E_r(R_{\pi r}) z_i),
\end{aligned}$$

$(\bar{Y}_r, \bar{Z}_r) = \frac{\sum_{i \in U} a_i(y_i, z_i)}{\sum_{i \in U} a_i}$ and $h = \frac{\bar{Z}_r - \bar{Z}_r}{\bar{Z}_{\pi r}}$. The component v_2 is obtained by estimating unknown quantities in (2.7.3). It is given by

$$\begin{aligned}
v_2 \approx & \hat{p}(1 - \hat{p}) \frac{\hat{N}}{\sum_{i \in s} w_i a_i} \left\{ s_{yr}^2 + (\hat{R}_r^{un})^2 s_{zr}^2 - 2\hat{R}_r^{un} s_{yzt} + \left(\hat{h} \sum_{i \in s} w_i a_i \right)^2 s_e^2 \right. \\
& \left. + 2 \left(\hat{h} \sum_{i \in s} w_i a_i \right) (s_{eyr} - \hat{R}_r^{un} s_{ezr}) \right\}, \tag{2.37}
\end{aligned}$$

where

$$\begin{aligned}
s_{yr}^2 &= \frac{1}{\sum_{i \in s} w_i a_i} \sum_{i \in s} w_i a_i (y_i - \bar{y}_r)^2, \\
s_{zr}^2 &= \frac{1}{\sum_{i \in s} w_i a_i} \sum_{i \in s} w_i a_i (z_i - \bar{z}_r)^2, \\
s_{yzt} &= \frac{1}{\sum_{i \in s} w_i a_i} \sum_{i \in s} w_i a_i (y_i - \bar{y}_r) (z_i - \bar{z}_r), \\
s_{er(2)}^2 &= \frac{1}{\sum_{i \in s} w_i a_i} \sum_{i \in s} w_i^{-1} a_i (y_i - \hat{R}_r^{un} z_i)^2, \\
s_{eyr} &= \frac{1}{\sum_{i \in s} w_i a_i} \sum_{i \in s} a_i (y_i - \bar{y}_r) (y_i - \hat{R}_r^{un} z_i), \\
s_{ezr} &= \frac{1}{\sum_{i \in s} w_i a_i} \sum_{i \in s} a_i (z_i - \bar{z}_r) (y_i - \hat{R}_r^{un} z_i),
\end{aligned}$$

and $\hat{h} = \frac{\bar{z} - \bar{z}_r}{\sum_{i \in s} a_i z_i}$.

2.7.4 Non-negativity of $V_*(\bar{y}_{IH}^a - \bar{Y})$

We show that the variance formula in (2.28) is always nonnegative. First, note that this expression can be expressed as

$$\begin{aligned} V_*(\bar{y}_{IH}^a - \bar{Y}) &= \frac{n^2 \sum_{i \in s_m} w_i^2 - (r+n) \left(\sum_{i \in s_m} w_i \right)^2}{n^2 \left(\sum_{i \in s_r} w_i \right)^2} \geq 0 \\ &\Leftrightarrow n^2 \sum_{i \in s_m} w_i^2 - (r+n) \left(\sum_{i \in s_m} w_i \right)^2 \geq 0 \\ &\Leftrightarrow n^2 \sum_{i \in s_m} w_i^2 - m(r+n) \frac{\left(\sum_{i \in s_m} w_i \right)^2}{m} \geq 0 \end{aligned}$$

On one hand, $n^2 \geq m(r+n) \Leftrightarrow n \geq m$ which is always true. On the other hand, using Cauchy-Schwarz inequality, it is easily seen that $\sum_{i \in s_m} w_i^2 \geq \frac{\left(\sum_{i \in s_m} w_i \right)^2}{m}$. The result follows.

Chapter 3

Inference for domain means

3.1 Introduction

In practice, estimates for various domains are often needed. For example, in the Canadian Labour Force Survey, estimates of unemployment are required by age-sex group and by industry at the provincial level. In practice, domains are generally not known at the imputation stage. As a result, the imputed estimators of domain means or domain totals are generally biased under regression imputation.

Suppose that the population U is partitioned into k domains of interest $U_1, U_2, \dots, U_k$. In this chapter, we are interested in making inference, under marginal imputation, on a k -vector of domain (or subpopulation) totals

$$\mathbf{Y}_{(d)} = \sum_{i=1}^N \mathbf{x}_i y_i \quad (3.1)$$

from a sample with missing responses y_i , where $\mathbf{x}_i$ is a k -vector of domain indicators known for all the units in the sample and N is the population size. Domains of

interest to analysts are typically not specified at the imputation stage, and in fact they can cut across imputation classes used by the data producer. The vector of domain means may be written as

$$\mathbf{Y}_{(d)} = \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^N \mathbf{x}_i y_i. \quad (3.2)$$

If $\mathbf{x}_i$ is an arbitrary vector of auxiliary variables related to y_i , then (3.2) may be regarded as a “census” regression vector $\mathbf{B}$. Little and Rubin (2002) studied the estimation of a regression vector in the case of missing y but not $\mathbf{x}$. In this chapter, however, we focus on the estimation of domain totals and means, but similar results can be obtained for the regression case if $\mathbf{x}$ is not missing.

Suppose that a probability sample, s , of size n , is selected according to a specified sampling design $p(s)$ from the population U . Let s_r be the set of respondents to item y of size r and s_m be the set of nonrespondents to item y of size m , $r + m = n$. We assume that the vector of domain indicators, $\mathbf{x}_i$, is known for all units i in the sample s .

Denoting the imputed value for a missing y_i as y_i^* , the completed (or imputed) data file reports $\tilde{y}_i = a_i y_i + (1 - a_i) y_i^*$ for $i \in s$, where a_i is the response indicator variable, and an imputed estimator of $\mathbf{Y}_{(d)}$ is given by

$$\hat{\mathbf{Y}}_{I(d)} = \sum_{i \in s} w_i \mathbf{x}_i \tilde{y}_i, \quad (3.3)$$

where w_i is the survey weight associated with unit i . It follows from (3.3) that $\hat{\mathbf{Y}}_{I(d)}$ can be computed without the knowledge of a_i , noting that the completed data file reports $\tilde{y}_i$ for $i \in s$. The estimator $\hat{\mathbf{Y}}_{I(d)}$ satisfies the additive property $\mathbf{1}' \hat{\mathbf{Y}}_{I(d)} = \sum_{i \in s} w_i \tilde{y}_i = \hat{Y}_I$, where $\hat{Y}_I$ is the imputed estimator of the population total

Y. Similarly, if the domains denote the cells of a two-way table, the estimators of cell counts add up to the estimators of the marginal counts. The additivity property is practically important.

As we have noted in Chapter 1, two distinct approaches for making inference on the domain totals $\mathbf{Y}_{(d)}$ or the domain means $\bar{\mathbf{Y}}_{(d)}$ may be used: (i) design-based and (ii) model-based. Under the design-based approach, the finite population values y_i are fixed and the usual assumption on the response mechanism is as follows:

Assumption DB: Within an imputation cell, the response probability for a given variable of interest, p , is a constant and the responses statuses for different units, i , are independent.

On the other hand, the model-based approach (Särndal, 1992) uses the following assumption:

Assumption MB: Within an imputation cell, the response mechanism is ignorable or unconfounded. Imputation is performed according to a model that relates the variable being imputed, y , to covariates, $\mathbf{z}$, used for imputation. Under assumption MB, we use the regression model (1.15) within an imputation class:

If $\mathbf{z}_i = (1, z_{i2}, \dots, z_{iq})'$, then linear regression with constant variance $\sigma_i^2 = \sigma^2$ is obtained by letting $\boldsymbol{\lambda} = (1, 0, \dots, 0)'$ in (1.15).

If $\mathbf{z}_i = 1$ and $\boldsymbol{\lambda} = 1$ then (1.15) corresponds to the mean model. Commonly used mean imputation and random “hot deck” imputation are motivated by the mean model. Ratio imputation is motivated by a ratio model which is a special case of (1.15) with $\mathbf{z}_i = z_i$, a scalar variable and $\boldsymbol{\lambda} = 1$.

In this chapter, we study survey-weighted regression imputation for missing y -

values in the sample under the two different approaches (i) and (ii). For simplicity, we assume a single imputation cell. We show that the imputed estimator $\hat{Y}_{I(d)}$ is generally biased under (i), while it is approximately unbiased under (ii); see sections 3.2 and 3.3. Following the method of Skinner and Rao (2002), we obtain a bias-adjusted estimator of $Y_{(d)}$ under (i) and show that it is robust in the sense of approximate unbiasedness under both (i) and (ii). Variance estimators are derived in Section 3.4, following the method of Fay (1991) and Shao and Steel (1999) in which the order of sampling and response is reversed. Section 3.5 reports the results of a simulation study based on some Canadian Labour Force Survey data. Efficiency of the estimators relative to a complete case estimator is also studied; the latter estimator requires the knowledge of response indicators, a_i . The case of unweighted imputation is considered in Section 3.6. Finally, in section 3.7, we consider a second class of bias-adjusted estimator which is approximately unbiased under assumption MB when the imputation model is not correctly specified as well as under assumption DB.

3.2 Design-Based Approach

In this section, we study the unadjusted estimator $\hat{Y}_{I(d)}$, given by (3.3), and a bias-adjusted estimator of $Y_{(d)}$ under assumption DB and survey-weighted regression imputation motivated by model (1.15). Survey-weighted regression imputation uses the predicted value $y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r$ for missing y_i , where $\hat{\mathbf{B}}_r$ is the weighted least squares

estimator of β given by

$$\hat{\mathbf{B}}_r = \left(\sum_{i \in s_r} w_i \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i) \right)^{-1} \left(\sum_{i \in s_r} w_i \mathbf{z}_i y_i / (\lambda' \mathbf{z}_i) \right).$$

For (survey-weighted) ratio imputation, $y_i^* = z_i (\sum_{i \in s_r} w_i y_i) / (\sum_{i \in s_r} w_i z_i)$. For (survey-weighted) mean imputation, $y_i^* = (\sum_{i \in s_r} w_i y_i) / (\sum_{i \in s_r} w_i)$.

3.2.1 Unadjusted estimators

The imputed estimator $\hat{Y}_I$ of the total $Y_N = \sum_{i \in U} y_i$ is asymptotically unbiased under assumption DB, but the domain total estimator $\hat{Y}_{I(d)}$ is asymptotically biased. We have

$$\text{Bias}(\hat{Y}_{I(d)}) \approx (1 - p) (\mathbf{U}\mathbf{B} - \mathbf{Y}_{(d)}), \quad (3.4)$$

where $\mathbf{U} = \sum_{i \in U} \mathbf{x}_i \mathbf{z}_i'$, $\mathbf{B} = (\sum_{i \in U} \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i))^{-1} (\sum_{i \in U} \mathbf{z}_i y_i / (\lambda' \mathbf{z}_i))$, and p is the constant probability of response. In a matrix form, expression (3.4) can be rewritten as

$$\text{Bias}(\hat{Y}_{I(d)}) \approx (1 - p) (\mathbf{X}\mathbf{X}')^{-1} \mathbf{X} \left[I_N - \mathbf{Z}' (\mathbf{Z}\Sigma^{-1}\mathbf{Z}')^{-1} \mathbf{Z}\Sigma^{-1} \right] \mathbf{y}, \quad (3.5)$$

where $\mathbf{X}$ is a $k \times N$ matrix whose columns are the k -dimensional vectors $\mathbf{x}_1, \dots, \mathbf{x}_N$, $\mathbf{Z}$ is a $q \times N$ matrix whose columns are the q -dimensional vectors $\mathbf{z}_1, \dots, \mathbf{z}_N$, Σ is a $N \times N$ diagonal matrix whose i^{th} component is σ_i^2 and $\mathbf{y} = (y_1, \dots, y_N)'$. From (3.5), it is clear that the bias is not zero for all $\mathbf{y}$ unless $p = 1$ (complete response case) or $\mathcal{R}(\mathbf{X}') \subseteq \mathcal{R}(\Sigma^{-1}\mathbf{Z}')$, where $\mathcal{R}(\mathbf{A})$ denotes the range of $\mathbf{A}$. In particular, if $\sigma_i^2 = \sigma^2$ then the latter condition reduces to $\mathcal{R}(\mathbf{X}') \subseteq \mathcal{R}(\mathbf{Z}')$, which means that the columns of $\mathbf{X}'$ can be expressed as linear combinations of the columns of $\mathbf{Z}'$. The

latter condition is satisfied if the domain indicators are included in the imputation model (1.15).

For ratio imputation, expression (3.5) reduces to

$$\frac{\rho_{x_j y}}{\rho_{x_j z}} = \frac{C_z}{C_y}, j = 1, \dots, k, \quad (3.6)$$

assuming $S_{x_j} > 0, j = 1, \dots, k$ and $\bar{Y} \neq 0$ and provided $\rho_{x_j z} \neq 0$, where $\rho_{x_j y}$ is the correlation between x_j and y , $\rho_{x_j z}$ is the correlation between x_j and z , C_y and C_z are the coefficients of variation of y and z respectively and S_{x_j} is the population standard deviation for variable $x_j, j = 1, \dots, k$. The condition (3.6) is satisfied when each domain ratio $R_{j(d)} = \sum_{i \in U} x_{ij} y_i / \sum_{i \in U} x_{ij} z_i, j = 1, \dots, k$ equals the overall ratio $R = \sum_{i \in U} y_i / \sum_{i \in U} z_i$.

For mean imputation, expression (3.5) reduces to

$$\rho_{x_j y} = 0, j = 1, \dots, k, \quad (3.7)$$

which is satisfied when each domain mean $\bar{Y}_{j(d)} = \sum_{i \in U} x_{ij} y_i / \sum_{i \in U} x_{ij}, j = 1, \dots, k$ equals the overall population mean $\bar{Y}$.

The imputed estimator of domain mean vector $\bar{\mathbf{Y}}_{(d)}$ is

$$\hat{\hat{\mathbf{Y}}}_{I(d)} = \hat{\mathbf{T}}^{-1} \hat{\mathbf{Y}}_{I(d)} \quad (3.8)$$

with

$$\text{Bias} \left(\hat{\hat{\mathbf{Y}}}_{I(d)} \right) \approx (1 - p) (\mathbf{T}^{-1} \mathbf{U} \mathbf{B} - \bar{\mathbf{Y}}_{(d)}), \quad (3.9)$$

where $\hat{\mathbf{T}} = \sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i'$ and $\mathbf{T} = \sum_{i \in U} \mathbf{x}_i \mathbf{x}_i'$. It follows from (3.9) that under mean imputation the asymptotic bias of $\hat{\hat{\mathbf{Y}}}_{I(d)}$ is not zero unless $p = 1$ or each domain mean

$\bar{Y}_{j(d)}$ equals the overall mean $\bar{Y}$. Similarly, under ratio imputation, the asymptotic bias is not zero unless $p = 1$ or each domain ratio $R_{j(d)}$ equals the overall ratio R .

3.2.2 Bias-adjusted estimators

A bias-adjusted estimator of $\mathbf{Y}_{(d)}$ is given by

$$\hat{\mathbf{Y}}_{I(d)}^a = \hat{p}^{-1} \hat{\mathbf{Y}}_{I(d)} + (1 - \hat{p}^{-1}) \hat{\mathbf{U}} \hat{\mathbf{B}}_I, \quad (3.10)$$

where $\hat{p} = \sum_{i \in s_r} w_i / \sum_{i \in s} w_i$ is an estimator of the response rate p , $\hat{\mathbf{U}} = \sum_{i \in s} w_i \mathbf{x}_i \mathbf{z}_i'$ and

$$\hat{\mathbf{B}}_I = \left(\sum_{i \in s} w_i \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i) \right)^{-1} \left[\sum_{i \in s} w_i \mathbf{z}_i \tilde{y}_i / (\lambda' \mathbf{z}_i) \right]. \quad (3.11)$$

The adjusted estimator (3.10) may be derived from the method of moments, following Skinner and Rao (2002), by solving

$$E(\hat{\mathbf{Y}}_{I(d)}) = p \mathbf{Y}_{(d)} + (1 - p) \mathbf{U} \mathbf{B} \quad (3.12)$$

for $\mathbf{Y}_{(d)}$ and then replacing $E(\hat{\mathbf{Y}}_{I(d)})$ by $\hat{\mathbf{Y}}_{I(d)}$, $\mathbf{B}$ by $\hat{\mathbf{B}}_I$ and $\mathbf{U}$ by $\hat{\mathbf{U}}$. The adjusted estimator (3.10) satisfies the additive property $\mathbf{1}' \hat{\mathbf{Y}}_{I(d)}^a = \hat{Y}_I$, and it can be implemented without the knowledge of response indicators, a_i , provided the response rate $\hat{p}$ is reported, but in general the completed data file should include the $\mathbf{z}_i$'s and identify them as imputation variables. However, for mean imputation and random hot deck imputation, $\hat{\mathbf{Y}}_{I(d)}^a$, reduces to

$$\hat{\mathbf{Y}}_{I(d)}^a = \hat{p}^{-1} \hat{\mathbf{Y}}_{I(d)} + (1 - \hat{p}^{-1}) \hat{\mathbf{N}}_{(d)} \hat{\hat{Y}}_I, \quad (3.13)$$

where $\hat{\mathbf{N}}_{(d)} = \sum_{i \in s} w_i \mathbf{x}_i$ is the vector of estimated domain sizes and

$\hat{\hat{Y}}_I = \sum_{i \in s} w_i \tilde{y}_i / \sum_{i \in s} w_i$ is the imputed estimator of the mean $\bar{Y}$. It follows from

(3.13) that $\hat{\mathbf{Y}}_{I(d)}^a$ can be implemented from the completed data file if the imputation classes are identified and the response rates within classes are reported.

In section 3.3, we show that the adjusted and unadjusted estimators are asymptotically unbiased under assumption MB and model (1.15). That is, the adjusted estimator is valid under both the design-based and model-based approaches, unlike the unadjusted estimator $\hat{\mathbf{Y}}_{I(d)}$. Hence, it is “robust” in the sense of validity under both approaches.

The adjusted estimator of the domain mean vector $\bar{\mathbf{Y}}_{(d)}$ is given by

$$\hat{\hat{\mathbf{Y}}}_{I(d)}^a = \hat{\mathbf{T}}^{-1} \hat{\mathbf{Y}}_{I(d)}^a. \quad (3.14)$$

It is also valid under both approaches to inference.

The following question now comes to mind: when is the bias-adjusted estimator (3.10) equal to the unadjusted estimator (3.3)?

Let $\mathbf{X}_r$ be a $k \times r$ matrix whose columns are the k -dimensional vectors $\mathbf{x}_1, \dots, \mathbf{x}_r$, $\mathbf{Z}_r$ is a $q \times r$ matrix whose columns are the q -dimensional vectors $\mathbf{z}_1, \dots, \mathbf{z}_r$, $\mathbf{W}_r$ is a $r \times r$ diagonal matrix whose i -th component is w_i , Σ_r is a $r \times r$ diagonal matrix whose i -th component is σ_i^2 and $\mathbf{y} = (y_1, \dots, y_r)'$.

Lemma 3.1 *The bias-adjusted estimator $\hat{\mathbf{Y}}_{I(d)}^a$ and the unadjusted estimate $\hat{\mathbf{Y}}_{I(d)}$ are equal if and only if*

$$\mathbf{W}_r \mathbf{X}_r \left[\mathbf{I}_r - \mathbf{Z}_r' (\mathbf{Z}_r \mathbf{W}_r^{1/2} \Sigma_r^{-1} \mathbf{W}_r^{1/2} \mathbf{Z}_r')^{-1} \mathbf{Z}_r' \Sigma_r^{-1} \mathbf{W}_r \right] \mathbf{y}_r = \mathbf{0}$$

which is satisfied, for all $\mathbf{y}_r$, if and only if $\mathcal{R}(\mathbf{X}_r') \subseteq \mathcal{R}(\Sigma_r^{-1} \mathbf{W}_r \mathbf{Z}_r')$. In the case of a self-weighting design, $w_i = w_0$, say and $\sigma_i^2 = \sigma^2$, the latter condition reduces $\mathcal{R}(\mathbf{X}_r') \subseteq$

$\mathcal{R}(\mathbf{Z}'_r)$, which is satisfied when the domain indicators are part of the imputation model (1.15).

3.2.3 Complete case estimators

If the response indicators, a_i , are provided with the completed data file, a complete case estimator of $\mathbf{Y}_{(d)}$ may be obtained from the respondent values only as

$$\hat{\mathbf{Y}}_{c(d)} = \hat{p}^{-1} \sum_{i \in s_r} w_i \mathbf{x}_i y_i. \quad (3.15)$$

We have $\mathbf{1}' \hat{\mathbf{Y}}_{c(d)} = \hat{p}^{-1} \sum_{i \in s_r} w_i y_i = \hat{Y}_c$. Hence, the complete case estimators $\hat{\mathbf{Y}}_{c(d)}$ add up to the complete case estimator, $\hat{Y}_c$, of the total Y , but not to the imputed estimator $\hat{Y}_I$. Note that $\hat{Y}_I$ uses the same survey weight for all items unlike $\hat{Y}_c$. The complete case estimator $\hat{\mathbf{Y}}_{c(d)}$ is asymptotically unbiased under assumption DB. The complete case estimator of the domain mean vector $\bar{\mathbf{Y}}_{(d)}$ is given by

$$\hat{\bar{\mathbf{Y}}}_{c(d)} = \left(\hat{p}^{-1} \sum_{i \in s_r} w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\hat{p}^{-1} \sum_{i \in s_r} w_i \mathbf{x}_i y_i \right). \quad (3.16)$$

It is also asymptotically unbiased under assumption DB. Note that $\hat{p}^{-1}$ cancels out in (3.16) for a single imputation class, but the response rates within classes do not cancel out in the case of multiple imputation classes:

$$\hat{\bar{\mathbf{Y}}}_{c(d)} = \left[\sum_{\nu} \hat{p}_{\nu}^{-1} \sum_{i \in s_{r\nu}} w_i \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \left[\sum_{\nu} \hat{p}_{\nu}^{-1} \sum_{i \in s_{r\nu}} w_i \mathbf{x}_i y_i \right],$$

where $s_{r\nu}$ and $\hat{p}_{\nu}$ denote the sample of respondents and estimated response rate in the ν^{th} imputation class.

It is interesting to note that the bias-adjusted estimator $\hat{\bar{\mathbf{Y}}}_{I(d)}^a$ reduces to the complete case estimator $\hat{\bar{\mathbf{Y}}}_{c(d)}$ if $y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r$ with $\mathbf{z}_i = \mathbf{x}_i$ and $\boldsymbol{\lambda} = \mathbf{1}$, that is, if the

imputation is done within domains. However, this is not possible unless the domains are specified in advance. Note that the bias-adjusted estimator $\hat{Y}_{I(d)}^a$ reduces to $(\sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i') (\sum_{i \in s_r} w_i \mathbf{x}_i \mathbf{x}_i')^{-1} (\sum_{i \in s_r} w_i \mathbf{x}_i y_i)$ which is not equal to $\hat{Y}_{c(d)}$.

3.2.4 Survey-weighted random regression imputation

Survey-weighted random imputation may be viewed as a weighted deterministic imputation plus random noise. In particular, survey-weighted random regression imputation, motivated by model (1.15), uses the value

$$y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r + (\boldsymbol{\lambda}' \mathbf{z}_i)^{1/2} \epsilon_i^* \quad (3.17)$$

for missing y_i , where $\epsilon_i^* = e_i^* - \bar{e}_{rw}$ with $\bar{e}_{rw} = \sum_{i \in s_r} w_i e_i / \sum_{i \in s_r} w_i$ and e_i^* is selected with replacement from the respondent standardized residuals $e_j = (\boldsymbol{\lambda}' \mathbf{z}_j)^{-1/2} (y_j - \mathbf{z}_j' \hat{\mathbf{B}}_r)$, $j \in s_r$, with probability $w_j / \sum_{l \in s_r} w_l$. If $\mathbf{z}_i = 1$, then y_i^* corresponds to weighted random hot-deck imputation: select donors $j \in s_r$ with replacement with selection probabilities $w_j / \sum_{l \in s_r} w_l$ and use $y_i^* = y_j$, $i \in s_m$.

Under survey-weighted random regression imputation, the unadjusted estimator $\hat{Y}_{I(d)}$ with y_i^* given by (3.17) is asymptotically biased, but the bias-adjusted estimator $\hat{Y}_{I(d)}^a$ is asymptotically unbiased under assumption DB, noting that $E_*(\epsilon_i^*) = 0$ and $E = E_p E_r E_*$. The bias-adjusted estimator $\hat{\mathbf{Y}}_{(d)}^a$, of the domain mean vector is also asymptotically unbiased.

3.3 Model-Based Approach

In this section, we study the unadjusted estimators and the adjusted estimators under assumption MB. Lemma 3.2 shows that the unadjusted estimator $\hat{\mathbf{Y}}_{I(d)}$ and the adjusted estimator $\hat{\mathbf{Y}}_{I(d)}^a$ are both valid under survey-weighted regression imputation and assumption MB. We first note that $E = E_m E_p E_r = E_p E_r E_m$ under assumption MB (Särndal, 1992).

Lemma 3.2 *Under survey-weighted regression imputation, $y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r$, the unadjusted estimator $\hat{\mathbf{Y}}_{I(d)}$ and the adjusted estimator $\hat{\mathbf{Y}}_{I(d)}^a$ are unbiased under the assumption MB with model (1.15). That is, $E[\hat{\mathbf{Y}}_{I(d)} - \mathbf{Y}_{(d)}] = 0$ and $E[\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)}] = 0$.*

Proof. First, we have $E(\mathbf{Y}_{(d)}) = E_m(\mathbf{Y}_{(d)}) = \sum_{i \in U} \mathbf{x}_i \mathbf{z}_i' \boldsymbol{\beta} = \mathbf{U}\boldsymbol{\beta}$. On the other hand,

$$\begin{aligned} E_m(\hat{\mathbf{Y}}_{I(d)}) &= \sum_{i \in s_r} w_i \mathbf{x}_i E_m(y_i) + \sum_{i \in s_m} w_i \mathbf{x}_i E_m(y_i^*) \\ &= \sum_{i \in s_r} w_i \mathbf{x}_i \mathbf{z}_i' \boldsymbol{\beta} + \sum_{i \in s_m} w_i \mathbf{x}_i \mathbf{z}_i' \boldsymbol{\beta} = \sum_{i \in s} w_i \mathbf{x}_i \mathbf{z}_i' \boldsymbol{\beta} = \hat{\mathbf{U}}\boldsymbol{\beta}. \end{aligned}$$

Hence, $E(\hat{\mathbf{Y}}_{I(d)}) = E_p(\sum_{i \in s} w_i \mathbf{x}_i \mathbf{z}_i' \boldsymbol{\beta}) = \sum_{i \in U} \mathbf{x}_i \mathbf{z}_i' \boldsymbol{\beta} = \mathbf{U}\boldsymbol{\beta} = E(\mathbf{Y}_{(d)})$. Similarly,

$$\begin{aligned} E_m(\hat{\mathbf{Y}}_{I(d)}^a) &= \hat{p}^{-1} E_m(\hat{\mathbf{Y}}_{I(d)}) + (1 - \hat{p}^{-1}) \hat{\mathbf{U}} E_m(\hat{\boldsymbol{\beta}}_I) \\ &= \hat{p}^{-1} \hat{\mathbf{U}} \boldsymbol{\eta} + (1 - \hat{p}^{-1}) \hat{\mathbf{U}} \boldsymbol{\beta} = \hat{\mathbf{U}} \boldsymbol{\beta}, \end{aligned}$$

noting that $E_m(\hat{\boldsymbol{\beta}}_I) = \boldsymbol{\beta}$ from (3.11). Hence,

$$E_m(\hat{\mathbf{Y}}_{I(d)}^a) = E(\hat{\mathbf{U}} \boldsymbol{\beta}) = \mathbf{U}\boldsymbol{\beta} = E(\mathbf{Y}_{(d)}).$$

■

Similarly, domain mean estimators $\hat{\mathbf{Y}}_{I(d)} = \hat{\mathbf{T}}^{-1} \hat{\mathbf{Y}}_{I(d)}$ and $\hat{\mathbf{Y}}_{I(d)}^a$ are asymptotically unbiased. Lemma 3.2 also holds for survey-weighted random regression imputation. The above results show that the adjusted estimators are robust in the sense of validity under both approaches, while the unadjusted estimators are valid only under the model-based approach (assumption MB).

The complete-case estimator $\hat{\mathbf{Y}}_{c(d)}$ given by (3.15) is biased under assumption MB.

We have

$$E_m(\hat{\mathbf{Y}}_{c(d)}) = \left(\sum_{i \in s} w_i / \sum_{i \in s} w_i a_i \right) \left(\sum_{i \in s} w_i a_i \mathbf{x}_i \mathbf{z}_i' \boldsymbol{\beta} \right)$$

and

$$E_r \left[E_m(\hat{\mathbf{Y}}_{c(d)}) \right] \approx \bar{p}_w^{-1} \left(\sum_{i \in s} w_i p_i \mathbf{x}_i \mathbf{z}_i' \boldsymbol{\beta} \right),$$

where p_i is the response probability for unit i and $\bar{p}_w = \sum_{i \in s} w_i p_i / \sum_{i \in s} w_i$.

Hence,

$$E(\hat{\mathbf{Y}}_{c(d)}) \approx \bar{P}_N^{-1} \left(\sum_{i \in U} p_i \mathbf{x}_i \mathbf{z}_i' \boldsymbol{\beta} \right), \quad (3.18)$$

where $\bar{P}_N = \bar{N}^{-1} \sum_{i \in U} p_i$. It follows from (3.18) that $E(\hat{\mathbf{Y}}_{c(d)}) \approx E_m(\mathbf{Y}_{(d)}) = \sum_{i \in U} \mathbf{x}_i \mathbf{z}_i' \boldsymbol{\beta}$ if $p_i = p$ for all $i \in U$; in general, it is asymptotically biased under assumption MB.

3.4 Variance Estimation

In this section, we obtain variance estimators of the bias-adjusted estimators under assumption DB and assumption MB, using the reverse approach of Fay (1991) and

Shao and Steel (1999). Under assumption DB and deterministic imputation, we have

$$V \left[\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)} \right] = E_r \left[V_p \left(\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)} \right) | \mathbf{a} \right] + V_r \left[E_p \left(\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)} \right) | \mathbf{a} \right], \quad (3.19)$$

It follows from (3.19) than an estimator of variance of $\hat{\mathbf{Y}}_{I(d)}^a$ is given by $\mathbf{v}_t = \mathbf{v}_1 + \mathbf{v}_2$, where $\mathbf{v}_1$ is an estimator of $V_p \left(\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)} | \mathbf{a} \right)$ and $\mathbf{v}_2$ is an estimator of $V_r \left[E_p \left(\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)} \right) | \mathbf{a} \right]$. Under assumption MB, E_r and V_r in (3.19) are changed to expectation and variance with respect to response mechanism and model, i.e., $E_r E_m$ and $E_r V_m + V_r E_m$. Note that $\mathbf{v}_1$ does not depend on the response mechanism or the model, and hence it is valid under either assumption DB or assumption MB.

In the case of random imputation, E_p and V_p in (3.19) are changed to expectation and variance with respect to imputation and sampling design, i.e., $\tilde{E} = E_p E_*$ and $\tilde{V} = V_p E_* + E_p V_*$, where V_* denotes variance with respect to imputation. Therefore, an estimator of $\tilde{V} \left(\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)} | \mathbf{a} \right)$ is given by $\tilde{\mathbf{v}}_1 = \mathbf{v}_1 + \mathbf{v}_*$, where $\mathbf{v}_1$ is an estimator of $V_p \left[E_* \left(\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)} \right) | \mathbf{a} \right]$ and $\mathbf{v}_*$ is an estimator of $E_p \left[V_* \left(\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)} \right) | \mathbf{a} \right]$. Further, noting that E_p is changed to $\tilde{E}$ in (3.19), the formula for $\mathbf{v}_2$ is the same as in the case of deterministic imputation, provided $E_* \left(\hat{\mathbf{Y}}_{I(d)}^a \right)$ agrees with the imputed estimator for the deterministic case. Hence, an estimator of the total variance of $\hat{\mathbf{Y}}_{I(d)}^a$ is given by $\mathbf{v}_t = \mathbf{v}_1 + \mathbf{v}_* + \mathbf{v}_2$.

3.4.1 Survey-weighted regression imputation

We first obtain a linearization variance estimator of the bias-adjusted estimator (3.10) of the total $\mathbf{Y}_{(d)}$ under assumption DB. We express (3.10) as a smooth function of

estimated totals $\hat{N}$, $\hat{R}$, $\hat{\mathbf{Y}}_{\mathbf{xr}}$, $\hat{\mathbf{U}}$, $\hat{\mathbf{U}}_r$, $\hat{\mathbf{S}}_r$, $\hat{\mathbf{t}}_r$ as follows, noting that $\hat{\mathbf{B}}_r = \hat{\mathbf{S}}_r^{-1}\hat{\mathbf{t}}_r$:

$$\hat{\mathbf{Y}}_{I(d)}^a = \frac{\hat{N}}{\hat{R}} \left[\hat{\mathbf{Y}}_{\mathbf{x}\mathbf{a}} + (\hat{\mathbf{U}} - \hat{\mathbf{U}}_a) \hat{\mathbf{S}}_a^{-1} \hat{\mathbf{t}}_a \right] + \left(1 - \frac{\hat{N}}{\hat{N}_r} \right) \hat{\mathbf{U}} \hat{\mathbf{S}}_r^{-1} \hat{\mathbf{t}}_r, \quad (3.20)$$

where $\hat{N} = \sum_{i \in s} w_i$, $\hat{R} = \sum_{i \in s} w_i a_i$, $\hat{\mathbf{Y}}_{\mathbf{xr}} = \sum_{i \in s} w_i a_i \mathbf{x}_i y_i$, $\hat{\mathbf{U}}_r = \sum_{i \in s} w_i a_i \mathbf{x}_i \mathbf{z}'_i$, $\hat{\mathbf{S}}_r = \sum_{i \in s} w_i a_i \mathbf{z}_i \mathbf{z}'_i / (\lambda' \mathbf{z}_i)$, $\hat{\mathbf{t}}_r = \sum_{i \in s} w_i a_i \mathbf{z}_i y_i / (\lambda' \mathbf{z}_i)$. Using the representation (3.20), a Taylor linearization variance estimator, $\mathbf{v}_1$, of $V_p \left(\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)} | \mathbf{a} \right)$ may be expressed, in an operator notation $\mathbf{v}(\cdot)$, as

$$\mathbf{v}_1 = \mathbf{v}(\hat{\xi}), \quad (3.21)$$

where $\mathbf{v}(\hat{\xi})$ denotes a standard variance estimator $\mathbf{v}(\mathbf{y})$ of a full sample total vector $\hat{\mathbf{Y}} = \sum_{i \in s} w_i \mathbf{y}_i$ with the vector $\mathbf{y}_i$ replaced by $\hat{\xi}_i$. Further,

$$\hat{\xi}_i = \frac{1}{\hat{R}} (1 - a_i \hat{p}^{-1}) \left[\hat{\mathbf{Y}}_{I(d)} - \hat{\mathbf{U}} \hat{\mathbf{B}}_r \right] + \hat{p}^{-1} \hat{\xi}_{1i} + (1 - \hat{p}^{-1}) \hat{\xi}_{3i} \quad (3.22)$$

with

$$\hat{\xi}_{1i} = a_i \mathbf{x}_i y_i + (1 - a_i) \mathbf{x}_i \mathbf{z}'_i \hat{\mathbf{B}}_r + (\hat{\mathbf{U}} - \hat{\mathbf{U}}_r) \hat{\mathbf{S}}_r^{-1} a_i \frac{\mathbf{z}_i}{(\lambda' \mathbf{z}_i)} (y_i - \mathbf{z}'_i \hat{\mathbf{B}}_r), \quad (3.23)$$

$$\hat{\xi}_{3i} = \mathbf{x}_i \mathbf{z}'_i \hat{\mathbf{B}}_r + \hat{\mathbf{U}} \hat{\xi}_{2i},$$

and

$$\hat{\xi}_{2i} = a_i \hat{\mathbf{S}}_r^{-1} \frac{\mathbf{z}_i}{(\lambda' \mathbf{z}_i)} (y_i - \mathbf{z}'_i \hat{\mathbf{B}}_r).$$

The linearization method of Demnati and Rao (2004) may be applied to obtain (3.22) in a straightforward manner. Using this method, the linearization variable (3.22) is simply obtained by expressing the estimator as a function of the weights $w_j, j \in s$ and then taking its partial derivative with respect to w_i . For the special case of mean

imputation (within classes), $\mathbf{z}_i = \mathbf{z}_i = 1$, (3.22) is somewhat simplified. Note that the response indicators, a_i , are required in calculating $\hat{\xi}_i$ for $i \in s$. If a resampling method of variance estimation, such as the jackknife or bootstrap, is applicable for the full sample total $\hat{\mathbf{Y}}$, then it can be applied in a straightforward manner to get a variance estimator, $\mathbf{v}_1$, of $V_p \left(\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)} | \mathbf{a} \right)$.

Denote the second component of $\mathbf{v}_t$ under assumption DB as $\mathbf{v}_{2D}$. After some calculation, we obtain

$$\mathbf{v}_{2D} = (1 - \hat{p}) \sum_{i \in s} w_i a_i \hat{\mathbf{c}}_{1i} \hat{\mathbf{c}}'_{1i}, \quad (3.24)$$

where

$$\begin{aligned} \hat{\mathbf{c}}_{1i} &= \hat{p}^{-1} \left(\hat{\xi}_{5i} - \hat{\mathbf{R}}^{-1} \hat{\mathbf{T}} \hat{\mathbf{Y}}_{I(d)} \right) + \hat{\mathbf{U}} \left[(1 - \hat{p}^{-1}) \hat{\xi}_{6i} + \hat{\mathbf{R}}^{-1} \hat{p}^{-1} \hat{\mathbf{B}}_r \right], \\ \hat{\xi}_{5i} &= \left[\mathbf{x}_i + (\hat{\mathbf{U}} - \hat{\mathbf{U}}_r) \hat{\mathbf{S}}_r^{-1} \frac{\mathbf{z}_i}{(\boldsymbol{\lambda}' \mathbf{z}_i)} \right] (y_i - \mathbf{z}'_i \hat{\mathbf{B}}_r) \end{aligned}$$

and

$$\hat{\xi}_{6i} = \hat{\mathbf{S}}_r^{-1} \mathbf{z}_i (y_i - \mathbf{z}'_i \hat{\mathbf{B}}_r) / (\boldsymbol{\lambda}' \mathbf{z}_i).$$

The total estimated variance is now given by $\mathbf{v}_t = \mathbf{v}_1 + \mathbf{v}_{2D}$. Note that a $\mathbf{v}_{2D}$ cannot be obtained directly using a resampling method, unlike $\mathbf{v}_1$. However, $\mathbf{v}_{2D}$ is negligible relative to $\mathbf{v}_1$ if the overall sampling fraction is small (Shao and Steel, 1999), so that $\mathbf{v}_t \approx \mathbf{v}_1$.

For the bias-adjusted estimator (3.14) of the domain mean $\bar{\mathbf{Y}}_{(d)}$, the first component, $\mathbf{v}_1$, of $\mathbf{v}_t$ under assumption DB is obtained from (3.21) by changing $\hat{\xi}$ to $\hat{\mathbf{T}}^{-1} \tilde{\xi}$, where $\tilde{\xi}_i$ is obtained from (3.22) with $\hat{\xi}_{1i}$ and $\hat{\xi}_{4i}$ changed to $\tilde{\xi}_{1i} = \hat{\xi}_{1i} - \mathbf{x}_i \mathbf{x}'_i \hat{\mathbf{Y}}_{I(d)}$ and $\tilde{\xi}_{4i} = \hat{\xi}_{3i} - \mathbf{x}_i \mathbf{x}'_i \hat{\mathbf{T}}^{-1} \hat{\mathbf{U}} \hat{\mathbf{B}}_r$, respectively. The second component, $\mathbf{v}_{2D}$, of $\mathbf{v}_t$ under assumption DB is obtained from (3.24) by simply changing $\hat{\mathbf{c}}_{1i}$ to $\tilde{\mathbf{c}}_{1i} = \hat{\mathbf{T}}^{-1} \hat{\mathbf{c}}_{1i}$.

We now turn to assumption MB and denote the second component of $\mathbf{v}_t$ as $\mathbf{v}_{2M}$.

Note that $\mathbf{v}_1$ is unchanged so that $\mathbf{v}_t = \mathbf{v}_1 + \mathbf{v}_{2M}$. To obtain $\mathbf{v}_{2M}$, we first write

$$E_p \left[\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)} \right] \approx \sum_{i \in U} \mathbf{c}_{ir} y_i \quad (3.25)$$

using (3.20), where

$$\mathbf{c}_{ir} = p_r^{-1} a_i \left[\mathbf{x}_i + (\mathbf{U} - \mathbf{U}_r) S_r^{-1} \mathbf{z}_i (\boldsymbol{\lambda}' \mathbf{z}_i)^{-1} \right] + (1 - p_r^{-1}) S_r^{-1} a_i \mathbf{z}_i (\boldsymbol{\lambda}' \mathbf{z}_i)^{-1} - \mathbf{x}_i, \quad (3.26)$$

with $p_r = R/N$ and $R = \sum_{i \in U} a_i \mathbf{U}_r = \sum_{i \in U} a_i \mathbf{x}_i \mathbf{z}_i'$ and $\mathbf{S}_r = \sum_{i \in U} a_i \mathbf{z}_i \mathbf{z}_i' (\boldsymbol{\lambda}' \mathbf{z}_i)^{-1}$.

The second component of the decomposition (3.19), under assumption MB, is changed to $E_r V_m \left[E_p \left(\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)} \right) | \mathbf{a} \right] + V_r E_m \left[E_p \left(\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)} \right) | \mathbf{a} \right]$. We have $E_m [E_p (.)] = E_p [E_m (.)]$ here and $E_p \left[E_m \left(\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)} \right) \right] = 0$. Hence, $V_r E_m [E_p (.)]$ is equal to zero. Next, using (3.25)

$$E_r V_m \left[E_p \left(\hat{\mathbf{Y}}_{I(d)}^a - \mathbf{Y}_{(d)} \right) \right] \approx \sigma^2 \sum_{i \in U} E_r (\mathbf{c}_{ir} \mathbf{c}_{ir}') (\boldsymbol{\lambda}' \mathbf{z}_i). \quad (3.27)$$

It now follows that the component $\mathbf{v}_{2M}$ of $\mathbf{v}_t$ may be obtained by substituting estimators for the unknown quantities in (3.27). We have

$$\mathbf{v}_{2M} = \hat{\sigma}^2 \sum_{i \in s} w_i \boldsymbol{\lambda}' \mathbf{z}_i \hat{\mathbf{c}}_{ir} \hat{\mathbf{c}}_{ir}', \quad (3.28)$$

where

$$\hat{\sigma}^2 = \frac{1}{\hat{R}} \sum_{i \in s} w_i a_i \left(y_i - \mathbf{z}_i \hat{\mathbf{B}}_r \right)^2$$

is an asymptotically unbiased estimator of σ^2 under the model (1.15), and

$$\hat{\mathbf{c}}_{ir} = \hat{p}^{-1} a_i \left[\mathbf{x}_i + \left(\hat{\mathbf{U}} - \hat{\mathbf{U}}_r \right) \hat{\mathbf{S}}_r^{-1} \mathbf{z}_i (\boldsymbol{\lambda}' \mathbf{z}_i)^{-1} \right] + (1 - \hat{p}^{-1}) \hat{\mathbf{S}}_r^{-1} a_i \mathbf{z}_i (\boldsymbol{\lambda}' \mathbf{z}_i)^{-1} - \mathbf{x}_i. \quad (3.29)$$

In the case of domain mean, $\mathbf{v}_1$ remains unchanged, i.e. $\mathbf{v}_1 = \mathbf{v}(\hat{\boldsymbol{\xi}})$ and $\mathbf{v}_{2M}$ is obtained from (3.28) by changing $\hat{\mathbf{c}}_{ir}$ to $\tilde{\mathbf{c}}_{ir} = \hat{\mathbf{T}}^{-1} \hat{\mathbf{c}}_{ir}$. Again, $\mathbf{v}_t$ is given by $\mathbf{v}_1 + \mathbf{v}_{2M}$.

3.4.2 Survey-weighted random regression imputation

We now obtain a linearization variance estimator of the bias-adjusted estimator $\hat{Y}_{I(d)}^a$ under survey-weighted random regression imputation. As noted before, we need only obtain the estimator $\mathbf{v}_*$ of $E_p V_*(\hat{Y}_{I(d)}^a - Y_{(d)})$. The estimator of total variance is then obtained as $\mathbf{v}_t = \mathbf{v}_1 + \mathbf{v}_* + \mathbf{v}_2$.

Under survey-weighted random regression imputation, we have from (3.17), $V_*(y_i^*) = (\lambda' \mathbf{z}_i) V_*(\epsilon_i^*)$,

$$V_*(\epsilon_i^*) = \frac{1}{\hat{R}} \sum_{i \in s} w_i a_i (e_i - \bar{e}_{aw})^2 \equiv s_{er}^2$$

noting that $E_*(\epsilon_i^*) = 0$, where $\bar{e}_{aw} = \sum_s w_i a_i e_i / \sum_s w_i a_i$. Hence, from (3.10), it follows that

$$\begin{aligned} \mathbf{v}_* &= V_*(\hat{Y}_{I(d)}^a) = \left\{ \hat{p}^{-2} \sum_{i \in s} w_i^2 (1 - a_i) \mathbf{x}_i \mathbf{x}_i' (\lambda' \mathbf{z}_i) \right. \\ &\quad \left. + \hat{p}^{-1} (1 - \hat{p}^{-1}) \left[\sum_{i \in s} w_i^2 (1 - a_i) (\hat{\mathbf{U}} \hat{\mathbf{S}}^{-1} \mathbf{z}_i \mathbf{x}_i' + \mathbf{x}_i \mathbf{z}_i' \hat{\mathbf{S}}^{-1} \hat{\mathbf{U}}) \right] \right. \\ &\quad \left. + (1 - \hat{p}^{-1})^2 \hat{\mathbf{U}} \hat{\mathbf{S}}^{-1} \left[\sum_{i \in s} w_i^2 (1 - a_i) \mathbf{z}_i \mathbf{z}_i' (\lambda' \mathbf{z}_i)^{-1} \right] \hat{\mathbf{S}}^{-1} \hat{\mathbf{U}}' \right\} s_{er}^2 \quad (3.30) \end{aligned}$$

noting that $Cov_*(y_i^*, y_j^*) = 0$ for $i \neq j$, where $\hat{\mathbf{S}} = \sum_{i \in s} w_i \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i)$. In the case of domain mean, $\mathbf{v}_*$, given by (3.30), is changed to $\tilde{\mathbf{v}}_* = \hat{\mathbf{T}}^{-1} \mathbf{v}_* \hat{\mathbf{T}}^{-1}$.

3.4.3 Unadjusted imputed estimator: assumption MB

In Lemma 3.1 we showed that the unadjusted estimator $\hat{Y}_{I(d)}$ is unbiased under assumption MB. Therefore, it would be of interest to provide a variance estimator of $\hat{Y}_{I(d)}$ under assumption MB. This will provide an extension of Särndal's (1992)

work on population totals to domain totals. However, we do not use the customary sample-response path adopted by Särndal (1992). Instead, we follow the reverse path approach which simplifies the derivation considerably. Also, the variance estimator will be robust to correlations in the model errors (as in cluster or two stage sampling) when $\mathbf{v}_1$ dominates $\mathbf{v}_t$, unlike the Särndal-type variance estimator (see Chapter 5).

In the case of survey-weighted regression imputation, we first express $\hat{\mathbf{Y}}_{I(d)}$ as a smooth function of estimated totals:

$$\hat{\mathbf{Y}}_{I(d)} = \hat{\mathbf{Y}}_{\mathbf{x}_r} + (\hat{\mathbf{U}} - \hat{\mathbf{U}}_r) \hat{\mathbf{S}}_r^{-1} \hat{\mathbf{t}}_r. \quad (3.31)$$

By Taylor linearization, it follows from (3.31) that

$$\mathbf{v}_1 = \mathbf{v} \left(\hat{\boldsymbol{\xi}}_1 \right), \quad (3.32)$$

where $\hat{\boldsymbol{\xi}}_{1i}$ is given by (3.23). Similarly,

$$\mathbf{v}_{2M} = \hat{\sigma}^2 \sum_{i \in s} w_i \boldsymbol{\lambda}' \mathbf{z}_i \hat{\mathbf{c}}_{1ir} \hat{\mathbf{c}}'_{1ir}, \quad (3.33)$$

where

$$\hat{\mathbf{c}}_{1ir} = a_i \left(\hat{\mathbf{U}} - \hat{\mathbf{U}}_r \right) \hat{\mathbf{S}}_r^{-1} \mathbf{z}_i (\boldsymbol{\lambda}' \mathbf{z}_i)^{-1} - (1 - a_i) \mathbf{x}_i.$$

In the case of survey-weighted random regression imputation, $\mathbf{v}_*$ is given by

$$\mathbf{v}_* = \sum_{i \in s} w_i^2 (1 - a_i) \mathbf{x}_i \mathbf{x}_i' s_{er}^2. \quad (3.34)$$

In the case of domain mean, $\hat{\boldsymbol{\xi}}_{1i}$ is changed to $\hat{\mathbf{T}}^{-1} \left(\hat{\boldsymbol{\xi}}_{1i} - \mathbf{x}_i \mathbf{x}_i' \hat{\mathbf{Y}}_{I(d)} \right)$ and $\hat{\mathbf{c}}_{1ir}$ to $\hat{\mathbf{T}}^{-1} \hat{\mathbf{c}}_{1ir}$.

3.5 Simulation Study

We conducted a limited simulation study on the performance of domain estimators, using a population of size $N = 11,270$ created from a Canadian Labour Force Survey (LFS) sample. The variable of interest, y , is taken as the weekly earnings (WE) and the domains of interest as the age groups. The overall mean, $\bar{Y}$, of the variable WE is 555, and Table 3.1 gives the domain means, $\bar{Y}_{j(d)}$, for $j = 1, \dots, 10$ domains. It is clear from Table 3.1 that the means for the age groups 15-19 and 19-24 are much smaller than $\bar{Y} = 555$, while the mean for the age group 60+ is close to $\bar{Y}$; middle age groups have larger means. For illustration, we selected the two extreme age groups (domains 1 and 10) and the middle age group (domain 5) to study the performance of domain estimators in the presence of nonresponse.

We selected $R = 2000$ independent simple random samples, each of size $n = 2000$ from the above population. Nonresponse was then generated in each sample according to uniform response with probability $p = 0.7$. Because of simple random sampling, the survey weights are all equal: $w_i = N/n$. We therefore used unweighted linear regression imputation with $\mathbf{z}_i = (1, z_{i1}, z_{i2}, z_{i3})$ and $\mathbf{X}'\mathbf{z}_i = 1$ or $\sigma_i^2 = \sigma^2$, where the categorical variables z_1, z_2 and z_3 respectively denote sex, province and full time or part time job.

Table 3.2 reports the simulated values of the absolute bias ratio (BR) of the unadjusted, adjusted and complete case estimators under linear regression imputation, where $\text{BR} = (\text{bias of estimator}) / (\text{standard error of estimator})$. It is clear from Table 3.2 that both the adjusted and the complete case estimators perform well with $|\text{BR}|$

less than 10%. On the other hand, the unadjusted estimator exhibits large $|BR|$ when the domain mean differs significantly from the overall mean, as in the case of age group 15-19.

Table 3.3 reports the simulated values of efficiency of the adjusted estimator relative to the unadjusted estimator (denoted E_{au}) and the complete case estimator (denoted E_{ac}), where $E_{au} = (\text{MSE of unadjusted estimator}) / (\text{MSE of adjusted estimator})$ and $E_{ac} = (\text{MSE of complete case estimator}) / (\text{MSE of adjusted estimator})$. It is clear from Table 3.3 that the adjusted estimator leads to large gain in efficiency over the unadjusted estimator for the age group 15-19 because of large bias ratio of the unadjusted estimator. For the age group 60+ with domain mean nearly equal to the overall mean, the complete case estimator for age group 15-19 is significantly more efficient than the adjusted estimator, but it is less efficient for the other two age groups. Note that the complete case estimator is biased under the model-based approach unlike the adjusted estimator. Also, it requires the response indicators and the domain estimators do not add up to the imputed estimator $\hat{Y}_I$ of the total Y . The complete case estimator for domain mean performs better than the complete case estimator for domain total because of the cancelation of $\hat{p}^{-1}$ in the numerator and denominator when a single imputation class is used.

Table 3.1: Mean of the variable WE by age group in the population.

Age group	15-19	19-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60+
Mean of WE	139.7	343.6	513.9	587.2	625.6	661.5	704.5	692.3	629.3	549.3

3.6 Unweighted imputation

As noted in Chapter 2, weighted or unweighted imputation may be used in the case of unequal probability designs. As we have seen, unweighted imputation generally leads to biased estimators of population totals and population means under assumption DB. In this section, we consider the problem of estimating a vector of domain means when unweighted imputation was used to compensate for nonresponse.

3.6.1 Bias-adjusted estimator

In this section, we consider the case of unweighted random hot-deck imputation. Under this imputation method, we select the donors $j \in s_r$ with replacement with equal probabilities $1/r$ and use $y_i^* = y_j$, $i \in s_m$. Using the y_i^* 's, the bias of the imputed estimator (3.8) under uniform response is given by

$$Bias\left(\hat{\bar{Y}}_{I(d)}\right) = (1-p)\left(\mathbf{T}^{-1}\mathbf{X}\bar{Y}_\pi - \bar{Y}_{(d)}\right) \quad (3.35)$$

where $\bar{Y}_\pi = \frac{\sum_{i \in U} \pi_i y_i}{\sum_{i \in U} \pi_i}$ and $\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$. Assuming $S_y > 0$ and $\bar{X}_j \neq 0$, $j = 1, \dots, k$, the bias vanishes if and only if $p = 1$ (complete response case) or

$$\frac{\rho_{x_j y}}{\rho_{\pi y}} = \frac{C_\pi}{C_{x_j}}, j = 1, \dots, k; \quad (3.36)$$

Table 3.2: Percent absolute bias ratio of the unadjusted, adjusted and complete case estimators of domain totals and means under linear regression imputation

	Domain Total		
Domain (age group)	Unadjusted estimator	Adjusted estimator	Complete Case estimator
15-19	300.0	8.94	3.17
35-39	40.6	5.50	4.47
60+	9.52	0.45	1.77
	Domain Main		
15-19	402.1	10.3	5.20
35-39	68.3	2.40	2.60
60+	18.4	3.65	1.37

provided $\rho_{\pi y} \neq 0$, where $\rho_{x_j y}$ is the correlation between x_j and y , $\rho_{\pi y}$ is the correlation between π and y , C_{x_j} and C_π are the coefficients of variation of x_j , $j = 1, \dots, k$ and π respectively and S_y is the population standard deviation for item y . Note that when $\rho_{\pi y} = 0$, the expression for the bias (3.36) reduces to (3.7).

As in the case of weighted imputation, we propose a simple bias-adjusted estimator of $\bar{Y}_{I(d)}$ under unweighted random hot-deck imputation. It is given by

$$\hat{\bar{Y}}_{I(d)}^a = \hat{p}^{-1} \hat{\bar{Y}}_{I(d)} + (1 - \hat{p}^{-1}) \hat{\mathbf{T}}^{-1} \hat{\mathbf{X}} \bar{y}_I^{un}, \quad (3.37)$$

where $\hat{p} = \frac{\sum_{s \in s} w_i}{\sum_s w_i}$ is a consistent estimator of p , $\hat{\mathbf{X}} = \sum_{i \in s} w_i \mathbf{x}_i$ and $\bar{y}_I^{un}$ is the unweighted mean of the observed values y_i and the imputed values y_i^* . This estimator

Table 3.3: Efficiency of the adjusted estimator relative to the unadjusted estimator (E_{au}) and the complete case estimator (E_{ac}) of domain totals and means under linear regression imputation

	Domain Total		Domain Mean	
Age group	(E_{au})	(E_{ac})	(E_{au})	(E_{ac})
15-19	11.2	0.85	14.8	0.50
35-39	1.01	1.47	1.00	1.12
60+	0.81	1.31	0.64	1.21

may be derived from the method of moments, following Skinner and Rao (2002), by solving

$$E\left(\hat{\bar{Y}}_{I(d)}\right) = p\bar{Y}_{(d)} + (1-p)\mathbf{T}\mathbf{X}\bar{Y}_{\pi}$$

for $\bar{Y}_{(d)}$ and replacing $E\left(\hat{\bar{Y}}_{I(d)}\right)$ by its estimator $\hat{\bar{Y}}_{I(d)}$, $\bar{Y}_{\pi}$ by its estimator $\bar{y}_I^{un}$, $\mathbf{T}$ by its estimator $\hat{\mathbf{T}}$ and $\mathbf{X}$ by its estimator $\hat{\mathbf{X}}$.

It is easily seen that the adjusted estimator $\hat{\bar{Y}}_{I(d)}^a$ is approximately unbiased for $\bar{Y}_{(d)}$ under both Assumption DB and Assumption MB. Once again, the survey analyst can easily implement the adjusted estimator $\hat{\bar{Y}}_{I(d)}^a$ from the estimation file, i.e., the imputed data file $(w_i, \tilde{y}_i, z_i, i \in s)$, where $\tilde{y}_i = y_i$ if $i \in s_r$ and $\tilde{y}_i = y_i^*$ if $i \in s_m$, provided the response rate $\hat{p}$ is available to the analyst. In the case of multiple imputation classes, response rates within classes and imputation class identifiers need to be provided with the estimation file.

3.6.2 Variance estimation

In this section, we obtain linearization variance estimators for the bias-adjusted estimator (3.37) under both the design-based and model-based approaches. Once again, we use the reverse approach of Fay (1991). First, note that $E_* \left(\hat{\mathbf{Y}}_{I(d)}^a \right)$ under unweighted random hot-deck imputation can be expressed as a function of estimated totals:

$$E_* \left(\hat{\mathbf{Y}}_{I(d)}^a \right) = \frac{\hat{N}}{\hat{R}} \hat{\mathbf{T}}^{-1} \left[\hat{\mathbf{Y}}_{\mathbf{x}\mathbf{r}} + \left(\hat{\mathbf{X}} - \hat{\mathbf{X}}_{\mathbf{r}} \right) \bar{y}_r^{un} \right] + \left(1 - \frac{\hat{N}}{\hat{R}} \right) \hat{\mathbf{T}} \hat{\mathbf{X}} \bar{y}_r^{un},$$

where $\hat{\mathbf{X}}_{\mathbf{r}} = \sum_{i \in s} w_i a_i \mathbf{x}_i$, $\bar{y}_r^{un} = \frac{\hat{Y}_{r\pi}}{\hat{R}_\pi}$ with $\hat{Y}_{r\pi} = \sum_{i \in s} a_i y_i$ and $\hat{R}_\pi = \sum_{i \in s} a_i$. Then, one can show, using Taylor linearization, that $\mathbf{v}_1$ reduces to

$$\mathbf{v}_1 \approx \mathbf{v}(\hat{\xi}), \quad (3.38)$$

where

$$\begin{aligned} \hat{\xi}_i &= \frac{1}{\hat{R}} (1 - a_i \hat{p}^{-1}) \left[\hat{\mathbf{Y}}_{I(d)} - \hat{\mathbf{T}}^{-1} \hat{\mathbf{X}} \bar{y}_r^{un} \right] + \hat{p}^{-1} \hat{\mathbf{T}}^{-1} \left[\hat{\xi}_{1i} - \mathbf{x}_i \mathbf{x}_i' \hat{\mathbf{Y}}_{I(d)} \right] \\ &\quad + (1 - \hat{p}^{-1}) \hat{\mathbf{T}}^{-1} \left[\hat{\xi}_{3i} - \mathbf{x}_i \mathbf{x}_i' \hat{\mathbf{T}}^{-1} \hat{\mathbf{X}} \bar{y}_r^{un} \right], \end{aligned}$$

with

$$\begin{aligned} \hat{\xi}_{1i} &= a_i \mathbf{x}_i y_i + (1 - a_i) \mathbf{x}_i \bar{y}_r^{un} + \left(\hat{\mathbf{X}} - \hat{\mathbf{X}}_{\mathbf{a}} \right) \frac{1}{\hat{R}_\pi} a_i w_i^{-1} (y_i - \bar{y}_r^{un}), \\ \hat{\xi}_{3i} &= \mathbf{x}_i \bar{y}_r^{un} + \hat{\mathbf{X}} \hat{\xi}_{2i}, \end{aligned}$$

and

$$\hat{\xi}_{2i} = \frac{1}{\hat{R}_\pi} a_i w_i^{-1} (y_i - \bar{y}_r^{un}).$$

We now obtain the component of the variance due to imputation. Noting that

$V_*(y_i^*) = \frac{1}{\hat{R}_\pi} \sum_{i \in s} a_i (y_i - \bar{y}_r^{un})^2 = s_{yr}^{2un}$ and $Cov_*(y_i^*, y_j^*) = 0$ for $i \neq j$, one can show

that

$$\begin{aligned} V_* \left(\hat{\bar{Y}}_{I(d)} - \bar{Y}_{(d)} \right) &= \left\{ \hat{p}^{-2} \hat{\mathbf{T}}^{-1} \left[\sum_{i \in s} w_i^2 (1 - a_i) \mathbf{x}_i \mathbf{x}_i' \right] \hat{\mathbf{T}}^{-1} \right. \\ &\quad + (1 - \hat{p}^{-1})^2 \frac{m}{n^2} \hat{\mathbf{T}}^{-1} \hat{\mathbf{X}} \hat{\mathbf{X}}' \hat{\mathbf{T}}^{-1} \\ &\quad \left. + \frac{2}{n} \hat{p}^{-1} (1 - \hat{p}^{-1}) \hat{\mathbf{T}}^{-1} (\hat{\mathbf{X}} - \hat{\mathbf{X}}_a) \hat{\mathbf{X}}' \hat{\mathbf{T}}^{-1} \right\} s_{yr}^{2un}. \end{aligned} \quad (3.39)$$

We next give the second component of the variance estimator under the design-based and the model-based approaches, namely $\mathbf{v}_{2DB}$ and $\mathbf{v}_{2MB}$. We obtain

$$\mathbf{v}_{2DB} \approx \hat{p} (1 - \hat{p}) \sum_{i \in s} w_i a_i \hat{\mathbf{c}}_{1i} \hat{\mathbf{c}}_{1i}', \quad (3.40)$$

where

$$\begin{aligned} \hat{\mathbf{c}}_{1i} &= \hat{p}^{-1} \hat{\mathbf{T}}^{-1} \left[\hat{\xi}_{4i} - \frac{\hat{\mathbf{T}}}{\hat{R}} \hat{\bar{Y}}_{I(d)} \right] + \hat{\mathbf{T}}^{-1} \hat{\mathbf{X}} \left[\hat{\xi}_{5i} - \hat{p}^{-1} (\hat{\xi}_{5i} - \bar{y}_r^{un}) \right], \\ \hat{\xi}_{4i} &= \left(\mathbf{x}_i + \frac{\hat{\mathbf{X}} - \hat{\mathbf{X}}_a}{\hat{R}_\pi} w_i^{-1} \right) (y_i - \bar{y}_r^{un}) \end{aligned}$$

and

$$\hat{\xi}_{5i} = \frac{1}{\hat{R}_\pi} w_i^{-1} (y_i - \bar{y}_r^{un}).$$

Similarly, we obtain

$$\mathbf{v}_{2MB} \approx \hat{\sigma}^2 \sum_{i \in s} w_i \hat{\mathbf{c}}_{2i} \hat{\mathbf{c}}_{2i}', \quad (3.41)$$

where

$$\hat{\mathbf{c}}_{2i} = \hat{p}^{-1} \hat{\mathbf{T}}^{-1} a_i \left[\mathbf{x}_i + w_i^{-1} \frac{\hat{\mathbf{X}} - \hat{\mathbf{X}}_r}{\hat{R}_\pi} \right] + (1 - \hat{p}^{-1}) \hat{\mathbf{T}}^{-1} \frac{\hat{\mathbf{X}}}{\hat{R}_\pi} a_i w_i^{-1} - \hat{\mathbf{T}}^{-1} \mathbf{x}_i,$$

and

$$\hat{\sigma}^2 = \frac{1}{\hat{R}} \sum_{i \in s} w_i a_i (y_i - \bar{y}_r)^2$$

is an asymptotically unbiased and consistent estimator of σ^2 under the mean imputation model (1.17). The total variance $\mathbf{v}_t$ is then given by $\mathbf{v}_1 + \mathbf{v}_* + \mathbf{v}_{2_{DB}}$ under the design-based approach, and $\mathbf{v}_1 + \mathbf{v}_* + \mathbf{v}_{2_{MB}}$ under the model-based approach.

3.7 Alternative Bias-Adjusted Estimators

We have showed that the bias adjusted estimator $\hat{\mathbf{Y}}_{I(d)}^a$ is approximately unbiased under assumption DB as well as under assumption MB and imputation model (1.15). However, one can argue that model (1.15) is not realistic since it does not allow for different β coefficients in different domains. Indeed, it seems more realistic to assume different slopes in different domains, as it is often the case in practice. We consider the case of ratio imputation in which case a more realistic imputation model is given by

$$y_i = \beta_d z_i + \epsilon_i, i \in U_d. \quad (3.42)$$

3.7.1 Bias-Adjusted Estimators

In this section, we consider estimating the j^{th} domain mean $\bar{Y}_j$ (i.e., the j^{th} component of the vector of domain means $\bar{\mathbf{Y}}_{(d)}$) under ratio imputation.

In this case, an imputed estimator of $\bar{Y}_j$ is given by

$$\hat{Y}_{Ij} = \frac{1}{\hat{N}_j} \left(\sum_{i \in s} w_i a_i x_{ji} y_i + \sum_{i \in s} w_i (1 - a_i) x_{ji} z_i \frac{\sum_{i \in s} w_i a_i y_i}{\sum_{i \in s} w_i a_i z_i} \right), \quad (3.43)$$

where $x_{ij} = 1$ if unit i belongs to domain j and $x_{ij} = 0$, otherwise and $\hat{N}_j = \sum_{i \in s} w_i x_{ji}$ denotes the estimated size for domain j . Note that $\hat{Y}_j$ in (3.43) can be written as

$$\hat{Y}_{Ij} = \frac{1}{\hat{N}_j} \left(\sum_{i \in s} w_i a_i x_{ji} y_i + \frac{\sum_{i \in s} w_i (1 - a_i) x_{ji} z_i}{\sum_{i \in s} w_i a_i z_i} \sum_{d=1}^k \sum_{i \in s_d} w_i a_i y_i \right), \quad (3.44)$$

where $s_d = s \cap U_d$ denotes the subset of s which falls into domain d . It is easily seen that the asymptotic bias of $\hat{Y}_{Ij}$ under model (3.42) is given by

$$\begin{aligned} \text{Bias}(\hat{Y}_{Ij}) &= E_r E_p E_m (\hat{Y}_{Ij} - \bar{Y}_j) \\ &\approx \bar{Z}_j \left[1 - \tilde{p}_j \frac{\tilde{Z}_{jr}}{\bar{Z}_j} \right] \left[\sum_{d=1}^k \frac{\tilde{Z}_{dr}}{\tilde{Z}_r} \beta_d - \beta_j \right], \end{aligned} \quad (3.45)$$

where $\bar{Z}_j = Z_j/N_j$ with $Z_j = \sum_{i \in U} x_{ji} z_i$ and $N_j = \sum_{i \in U} x_{ji}$, $\tilde{p}_j = \tilde{R}_j/N_j$ with $\tilde{R}_j = E_r (\sum_{i \in U} a_i x_{ji})$, $\tilde{Z}_{jr} = \tilde{Z}_j/\tilde{R}_j$ with $\tilde{Z}_{jr} = E_r (\sum_{i \in U} a_i x_{ji} z_i)$ and $\tilde{Z}_r = E_r (\sum_{i \in U} a_i z_i)$.

Note that the bias in (3.45) vanishes if $\beta_d = \beta$ for all $d = 1, \dots, k$ noting that $\sum_{d=1}^k \frac{\tilde{Z}_{dr}}{\tilde{Z}_r} = 1$. The latter condition is equivalent to using the ratio model with a common slope for all the domains, which is a particular case given by (1.16). The bias is, in general, not equal to zero. As a result, the imputed estimator is not appropriate under model (3.42).

Now, consider the bias-adjusted estimator $\hat{Y}_{Ij}^a$ for domain j (i.e., the j^{th} component of the vector of bias-adjusted estimators $\hat{\mathbf{Y}}_{I(d)}^a$). Under ratio imputation, $\hat{Y}_{Ij}^a$ is given by

$$\hat{Y}_{Ij}^a = \hat{p}^{-1} \hat{Y}_{Ij} + (1 - \hat{p}^{-1}) \frac{\bar{z}_j}{\bar{z}} \bar{y}_I, \quad (3.46)$$

where $\bar{z}_j = \hat{Z}_j/\hat{N}_j$ with $\hat{Z}_j = \sum_{i \in s} w_i x_{ji}$, $\bar{z} = \sum_{i \in s} w_i z_i / \sum_{i \in s} w_i$ and $\bar{y}_I$ denotes the weighted mean of the observed values, y_i , and imputed values, $y_i^* = \hat{R}_r z_i$, where

$\hat{R}_r = \sum_{i \in s} w_i a_i y_i / \sum_{i \in s} w_i a_i z_i$. It is easily seen that, under model (3.42), the bias-adjusted estimator $\hat{Y}_{Ij}^a$ in (3.46) is asymptotically biased. The bias is given by

$$\text{Bias} \left(\hat{Y}_{Ij}^a \right) = E_r E_p E_m \left(\hat{Y}_{Ij}^a - \bar{Y}_j \right) \approx \left(\beta_j - \sum_{d=1}^k \beta_d \frac{\bar{Z}_{dr}}{\bar{Z}_r} \right) \left(\bar{p}^{-1} \bar{Z}_{jr} - Z_j \right), \quad (3.47)$$

where $\bar{p} = \bar{R}/N$ with $\bar{R} = E_r \left(\sum_{i \in U} a_i \right)$. Once again, the bias in (3.47) vanishes if $\eta_d = \eta$ for all $d = 1, \dots, k$, noting that $\sum_{d=1}^k \frac{\bar{Z}_{dr}}{\bar{Z}_r} = 1$. The bias in (3.47) is not zero, in general.

A bias-adjusted estimator of $\bar{Y}_j$ under model (3.42) is given by

$$\hat{Y}_{Ij}^{a(2)} = \hat{p}_j^{-1} \frac{\bar{z}_j}{\bar{z}_{jr}} \hat{Y}_{Ij} + \left(1 - \hat{p}_j^{-1} \frac{\bar{z}_j}{\bar{z}_{jr}} \right) \frac{\bar{z}_j}{\bar{z}} \bar{y}_I, \quad (3.48)$$

where $\bar{z}_j = \hat{Z}_j / \hat{N}_j$, $\hat{p}_j = \hat{R}_j / \hat{N}_j$ with $\hat{R}_j = \sum_{i \in s} w_i a_i x_{ji}$, $\bar{z}_{jr} = \hat{Z}_{jr} / \hat{R}_j$ with $\hat{Z}_{jr} = \sum_{i \in s} w_i a_i x_{ji} z_i$. Noting that $\beta_j = E_m(\bar{Y}_j) / \bar{Z}_j$, the bias-adjusted estimator $\hat{Y}_{Ij}^{a(2)}$ in (3.48) may be derived from the method of moments, following Skinner and Rao (2002), by solving

$$E \left(\hat{Y}_{Ij} \right) = \bar{p}_j \frac{\bar{Z}_{jr}}{\bar{Z}_j} E_m \left(\bar{Y}_j \right) + \bar{Z}_j \left(1 - \bar{p}_j \frac{\bar{Z}_{jr}}{\bar{Z}_j} \right) \sum_{d=1}^k \frac{\bar{Z}_{dr}}{\bar{Z}_r} \beta_d$$

for $E_m(\bar{Y}_j)$ and then replacing $E(\hat{Y}_{Ij})$ by $\hat{Y}_{Ij}$, $\bar{p}_j$ by $\hat{p}_j$, $\bar{Z}_{jr}$ by $\bar{z}_{jr}$, $\bar{Z}_j$ by $\bar{z}_j$ and $\sum_{d=1}^k \frac{\bar{Z}_{dr}}{\bar{Z}_r} \beta_d$ by $\bar{y}_I / \bar{z}$. It can be shown that the bias-adjusted estimator in (3.48) is approximately unbiased under both assumption MB under the imputation model (3.42) as well as under assumption DB. The bias-adjusted estimator $\hat{Y}_{Ij}^{a(2)}$ is therefore robust in the sense of validity under both approaches. It is interesting to note that under mean imputation (i.e., $z_i = 1$ for all i), the bias-adjusted estimator $\hat{Y}_{Ij}^a$ in (3.46) reduces to

$$\hat{Y}_{Ij}^a = \hat{p}^{-1} \hat{Y}_{Ij} + (1 - \hat{p}^{-1}) \bar{y}_I, \quad (3.49)$$

whereas the bias-adjusted estimator $\hat{Y}_{Ij}^{a(2)}$ in (3.48) reduces to

$$\hat{Y}_{Ij}^{a(2)} = \hat{p}_j^{-1} \hat{Y}_{Ij}^a + (1 - \hat{p}_j^{-1}) \bar{y}_I. \quad (3.50)$$

Hence, the difference between $\hat{Y}_{Ij}^a$ and $\hat{Y}_{Ij}^{a(2)}$ lies in the way we estimate the response probability p . In the case of $\hat{Y}_{Ij}^a$, we use the overall response rate $\hat{p}$, whereas in the case of $\hat{Y}_{Ij}^{a(2)}$ we use a “domain-specific” response rate $\hat{p}_j$. Also, note that the bias-adjusted estimator $\hat{Y}_{Ij}^{a(2)}$ is the estimator one would have obtained if ratio imputation within domains had been performed, which explains the robustness property of $\hat{Y}_{Ij}^{a(2)}$. However, the bias-adjusted estimator has two drawbacks: firstly, response identifiers, a_i , are required in order to compute $\hat{p}_j$ as well as $\bar{z}_{jr}$ in (3.48). Secondly, unlike the bias-adjusted estimator $\hat{Y}_{Ij}^a$, the bias-adjusted estimator $\hat{Y}_{Ij}^{a(2)}$ do not possess the additivity property, i.e.,

$$\sum_{j=1}^k \hat{N}_j \hat{Y}_{Ij}^{a(2)} = \hat{Y}_{IR} \neq \hat{Y}_I,$$

where $\hat{Y}_{IR}$ denotes the overall imputed estimator one would have obtained by performing ratio imputation within domains.

3.7.2 Simulation Study

We conducted a limited simulation study on the performance of the alternative bias-adjusted estimator $\hat{Y}_{I(j)}^{a(2)}$. First, we generated a population of size $N = 1000$ containing two domains U_1 and U_2 , of size $N_1 = N_2 = 500$, respectively. The population created contains 3 variables: a variable of interest y , an auxiliary variable z , and a domain indicator x .

First, the variable z was generated from a Gamma distribution with parameters $a = 6$ and $b = 9$. The first 500 units in the population were assigned to the first domain U_1 whereas the last 500 units were assigned to domain U_2 . In domain U_1 , the variable y was generated according to the model

$$y_{1i} = 10z_i + \epsilon_{1i},$$

where the ϵ_i 's were generated from a Normal distribution with parameters $\mu_1 = 0$ and $\sigma_1^2 = 100$. In domain U_2 , the variable y was generated according to the model

$$y_{2i} = 0.5z_i + \epsilon_{2i},$$

where the ϵ_i 's were generated from a Normal distribution with parameters $\mu_2 = 0$ and $\sigma_2^2 = 25$. As a result, the overall mean $\bar{Y}$ is equal to 281 and the domain means for U_1 and U_2 are equal to $\bar{Y}_1 = 537.05$ and $\bar{Y}_2 = 26.08$. Also, the correlation between y and z in domain U_1 , is equal to 0.9 whereas the correlation between y and z in the union of domains U_1 and U_2 is approximately equal to 0.9.

From the population, we selected $R = 2000$ simple random samples of size $n = 200$. In each sample, nonresponse was generated according to two different response mechanisms:

Mechanism 1: Response probability p_i for unit i is given by the logistic model

$$\log \frac{p_i}{1 - p_i} = \lambda_0 + \lambda_1 Z_i,$$

where the values of λ_0 and λ_1 were chosen so the mean of the p_i 's is approximately equal to 70%.

Mechanism 2 (uniform response): Response probability p_i is constant for all the units, i.e., $p_i = p$, where p was set to 70%.

To compensate for nonresponse to item y , we used ratio imputation based on z . Table 3.4 reports the simulated values of the absolute bias ratio (BR) of the unadjusted, complete case, bias-adjusted, $\hat{Y}_{I(j)}^a$, and bias-adjusted, $\hat{Y}_{I(j)}^{a(2)}$ under the response mechanism 1. It is clear from Table 3.4 that the bias ratio of all the estimators, excepting the bias-adjusted estimator $\hat{Y}_{I(j)}^{a(2)}$, is substantially large with an absolute bias ratio larger than 100% for both domain 1 and domain 2. The bias-adjusted estimator, $\hat{Y}_{I(j)}^{a(2)}$, performs well with an absolute bias ratio less than 2%.

Table 3.6 reports the simulated values of efficiency of the adjusted estimator, $\hat{Y}_{I(j)}^a$, relative to the unadjusted estimator (denoted E_{au}), the complete case estimator (denoted E_{ac}) and the bias-adjusted estimator, $\hat{Y}_{I(j)}^{a(2)}$, (denoted E_{aa_2}).

It is clear from Table 3.5 that the bias-adjusted estimator $\hat{Y}_{I(j)}^a$ is more efficient than the unadjusted estimator and the complete case estimator for domain 1 but less efficient than the bias-adjusted estimator, $\hat{Y}_{I(j)}^{a(2)}$. For domain 2, the bias-adjusted estimator $\hat{Y}_{I(j)}^a$ is more efficient than the unadjusted estimator but it is substantially less efficient than the complete case and the bias-adjusted estimator, $\hat{Y}_{I(j)}^{a(2)}$. Table 3.6 and 3.7 present respectively the percent absolute bias ratio and the relative efficiency for all the estimators under uniform response. In this cases, it is not surprising that all the estimators, except the unadjusted estimator, perform well in terms of bias ratio, as showed in Table 3.6. From Table 3.7, it is clear that the bias-adjusted estimator, $\hat{Y}_{I(j)}^{a(2)}$, is substantially more efficient than all the other estimators. These results show that the bias-adjusted estimator, $\hat{Y}_{I(j)}^{a(2)}$, is not only robust in the sense of validity

under assumption DB or assumption MB, but it is also more efficient than all the other estimators considered in the simulation study.

Table 3.4: Percent absolute bias ratio of the unadjusted, complete case, adjusted estimator $\left(\hat{Y}_{I(j)}^a\right)$ and adjusted estimator $\left(\hat{Y}_{I(j)}^{a(2)}\right)$ under the response Mechanism 1.

	Unadjusted estimator	Complete Case estimator	Adjusted estimator $\left(\hat{Y}_{I(j)}^a\right)$	Adjusted estimator $\left(\hat{Y}_{I(j)}^{a(2)}\right)$
Domain 1	200.0	234.6	113.0	1.69
Domain 2	465.8	255.3	317.7	1.97

Table 3.5: Efficiency of the adjusted estimator relative to the unadjusted estimator (E_{au}) , the complete case estimator (E_{ac}) and the adjusted estimator, $\hat{Y}_{I(j)}^{a(2)} (E_{aa2})$ under the response Mechanism 1.

	(E_{au})	(E_{ac})	(E_{aa2})
Domain 1	0.52	0.35	3.41
Domain 2	0.33	76.8	849.2

Table 3.6: Percent absolute bias ratio of the unadjusted, complete case, adjusted estimator $\left(\hat{Y}_{I(j)}^a\right)$ and adjusted estimator $\hat{Y}_{I(j)}^{a(2)}$ (denoted E_{aa2}) under the response Mechanism 2.

	Unadjusted estimator	Complete Case estimator	Adjusted estimator $\left(\hat{Y}_{I(j)}^a\right)$	Adjusted estimator $\left(\hat{Y}_{I(j)}^{a(2)}\right)$
Domain 1	-278.0	-4.62	-4.22	-2.95
Domain 2	466.4	3.68	9.72	5.06

Table 3.7: Efficiency of the adjusted estimator relative to the unadjusted estimator (E_{au}), the complete case estimator (E_{ac}) and the adjusted estimator, $\hat{Y}_{I(j)}^{a(2)}(E_{aa2})$ under the response Mechanism 2.

	(E_{au})	(E_{ac})	(E_{aa2})
Domain 1	0.11	0.99	1.37
Domain 2	0.03	102.0	138.4

Chapter 4

Nonresponse model approach to inference

4.1 Introduction

As mentioned in Chapter 1, two different approaches are generally used for making inference on totals, means and other parameters of interest in the presence of imputed values: (i) Nonresponse Model (NM) approach; (ii) Imputation model (IM) approach. Approach (i) is also called design-based approach (Särndal, 1992) and approach (ii) model-based approach (Shao and Steel, 1999). As noted in Chapter 1, NM approach is based on the following assumption:

Assumption NM: Response probability for a given item of interest is constant within imputation cells, and the response statuses for different units are independent. Inference under assumption NM is with respect to repeated sampling and uniform response mechanism within cells.

IM approach is based on the following assumption:

Assumption IM: Item values are missing at random (MAR) in the sense that the response probability does not depend on the item value being imputed but may depend on auxiliary variables used for imputation. Further, an (imputation) model that generates the finite population is assumed.

Inference under assumption IM is with respect to repeated sampling and the assumed model that generates the finite population and nonrespondents. Underlying response mechanism is not specified, except for the MAR assumption, unlike the NM approach. Under regression imputation, the imputation model is given by (1.15)

In this chapter, we propose a third approach, called Generalized Nonresponse Model (GNM) approach. GNM is based on the following assumption:

Assumption GNM: Item values are missing at random (MAR) and response probability is specified as a function of auxiliary variables and unknown parameters. Inference under assumption GNM is with respect to repeated sampling and the assumed response mechanism. Assumption NM is a special case of assumption GNM. We assume that the probability of response, p_i , for unit i , is linked to a l -vector of auxiliary variables $\mathbf{u}_i$ according to a logistic regression model so that

$$p_i = f(\mathbf{u}_i' \boldsymbol{\eta}) = \exp(\mathbf{u}_i' \boldsymbol{\eta}) / \exp(1 + \mathbf{u}_i' \boldsymbol{\eta}), \quad (4.1)$$

where $\boldsymbol{\eta}$ is l -vector of unknown parameters. Model (4.1) is the assumed nonresponse model.

As noted in Chapter 1, survey design weights are generally used in regression imputation. The resulting imputed estimator of a population mean is approximately

unbiased under either assumption NM or assumption IM. It is, therefore, robust in the sense of validity under both approaches. However, the imputed estimator is generally biased under assumption GNM. In this chapter, we propose a new method of regression imputation that is robust in the sense of leading to approximately unbiased estimators under either GNM or IM.

Section 4.2 develops the new method for the case of deterministic imputation as well as for random imputation, and demonstrates the robustness property in estimating a population total. Variance estimators are derived in section 4.3, using the reverse approach of Fay (1991). Results of a simulation study on the finite-sample performance of the imputed estimator under the new method of imputation are reported in section 4.4. Finally, the case of domain means is investigated in section 4.5.

4.2 Estimation of a Total

Let U be a finite population of possibly unknown size N . The objective is to estimate the population total $Y = \sum_{i \in U} y_i$ when imputation has been used to compensate for nonresponse on the item values y_i . Suppose a probability sample, s , of size n is selected according to a specified design $p(s)$ from U . In the presence of nonresponse to item y , an imputed estimator $\hat{Y}_I$ given by

$$\hat{Y}_I = \sum_{i \in s} w_i a_i y_i + \sum_{i \in s} w_i (1 - a_i) y_i^* = \sum_{i \in s} w_i \tilde{y}_i, \quad (4.2)$$

with $\tilde{y}_i = a_i y_i + (1 - a_i) y_i^*$, where y_i^* denotes the value imputed for missing y_i and a_i denotes the response indicator equal to 1 if unit i responds to item y and 0 otherwise.

Recall from Chapter 1, that the total error $\hat{Y}_I - Y$ may be decomposed as

$$\hat{Y}_I - Y = (\hat{Y} - Y) + (\hat{Y}_I - \hat{Y}). \quad (4.3)$$

Since the sampling error does not depend on nonresponse and on the imputation method, we focus on the nonresponse/imputation error $\hat{Y}_I - \hat{Y}$ to evaluate the properties of the imputed estimator, $\hat{Y}_I$. The properties of the imputed estimator are evaluated conditionally on the sample s . For simplicity, we assume a single imputation cell, but the extension to multiple cells is straightforward.

4.2.1 Deterministic regression imputation

Deterministic survey-weighted regression imputation uses the imputed values

$$y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r \quad (4.4)$$

for missing y_i , where $\hat{\mathbf{B}}_r$ is the weighted least squares estimator of β based on the responding units:

$$\hat{\mathbf{B}}_r = \left(\sum_{i \in s} w_i a_i \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i) \right)^{-1} \sum_{i \in s} w_i a_i \mathbf{z}_i y_i / (\lambda' \mathbf{z}_i). \quad (4.5)$$

Using (4.4), the imputed estimator (4.2) can be written as

$$\hat{Y}_I = \hat{Y}_r + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r)' \hat{\mathbf{B}}_r, \quad (4.6)$$

where $\hat{Y}_r = \sum_{i \in s} w_i a_i y_i$, $\hat{\mathbf{Z}} = \sum_{i \in s} w_i \mathbf{z}_i$ and $\hat{\mathbf{Z}}_r = \sum_{i \in s} w_i a_i \mathbf{z}_i$. Note that the imputed estimator (4.6) is similar to a regression estimator in the case of two-phase sampling.

Under assumption NM, $E_r(a_i | s) = p$ and the conditional bias of $\hat{Y}_I$, $\text{Bias}(\hat{Y}_I | s) = E_r(\hat{Y}_I - \hat{Y} | s)$ is approximately equal to 0. Furthermore, it can be shown that under assumption IM and regression model (1.15), the conditional bias $\text{Bias}(\hat{Y}_I | s) =$

$E_r E_m (\hat{Y}_I - \hat{Y}|s)$ is equal to 0. However, the imputed estimator (4.6) is conditionally biased under assumption GNM. The bias is given by

$$\text{Bias}(\hat{Y}_I|s) = E_r(\hat{Y}_I - \hat{Y}|s) \approx - \sum_{i \in s} w_i (1 - p_i) (y_i - \mathbf{z}'_i \hat{\mathbf{B}}_p), \quad (4.7)$$

where

$$\hat{\mathbf{B}}_p = \left(\sum_{i \in s} w_i p_i \mathbf{z}_i \mathbf{z}'_i / (\boldsymbol{\lambda}' \mathbf{z}_i) \right)^{-1} \sum_{i \in s} w_i p_i \mathbf{z}_i y_i / (\boldsymbol{\lambda}' \mathbf{z}_i). \quad (4.8)$$

This result follows from the fact that under assumption GNM, $E_r(a_i|s) = p_i$. Note that for the particular case of uniform response, $p_i = p$ (assumption NM), the bias term in (4.7) vanishes. It follows from (4.7) that the choice of imputed values (4.4) is, in general, not suitable under assumption GNM.

4.2.2 A bias-adjusted estimator

We assume for now that the response probabilities p_i are known. A natural approach for eliminating the bias of $\hat{Y}_I$ under assumption GNM is to consider a bias-adjusted estimator of the form

$$\hat{Y}_I^a = \hat{Y}_I - \hat{B}(\hat{Y}_I|s), \quad (4.9)$$

where $\hat{B}(\hat{Y}_I|s)$ is an estimator of the conditional bias (4.7) and it is given by

$$\hat{B}(\hat{Y}_I|s) = - \sum_{i \in s} w_i a_i \frac{(1 - p_i)}{p_i} (y_i - \mathbf{z}'_i \hat{\mathbf{B}}_r). \quad (4.10)$$

Note that $E_r[\hat{B}(\hat{Y}_I|s)|s] \approx \text{Bias}(\hat{Y}_I|s)$ under assumption GNM. The bias-adjusted estimator $\hat{Y}_I^a$ is then given by

$$\hat{Y}_I^a = \sum_{i \in s} \frac{w_i}{p_i} a_i (y_i - \mathbf{z}'_i \hat{\mathbf{B}}_r) + \sum_{i \in s} w_i \mathbf{z}'_i \hat{\mathbf{B}}_r. \quad (4.11)$$

In practice, the response probabilities p_i are unknown. Suppose we can obtain consistent estimators $\hat{p}_i$ of p_i by modeling p_i according to nonresponse model (4.1). Then, a bias-adjusted estimator is obtained by replacing p_i in (4.11) with $\hat{p}_i$. This estimator is also approximately conditionally unbiased for Y under assumption IM. Hence, the bias-adjusted estimator is robust in the sense of validity under both approaches, GNM and IM. However, the bias-adjusted estimator has two drawbacks. First, response identifiers, a_i , must be provided with the imputed data file, which may not always be the case in practice. Secondly, the estimated response probabilities $\hat{p}_i$ must also be provided with the data file. Both drawbacks can be eliminated by using the new imputation method, given in section 2.3, that leads to an approximately unbiased estimator under either assumptions GNM or IM.

It is interesting to note that the bias-adjusted estimator (4.11) with p_i replaced by $\hat{p}_i$ is identical to the estimator obtained using calibrated imputation (Beaumont, 2005). The idea behind calibrated imputation is to determine a set of final imputed values $\tilde{y}_i^*$, for $i \in s_m$, so as to satisfy the constraint $\hat{Y}_I = \hat{Y}_R$, where $\hat{Y}_I$ is given by (4.1) and $\hat{Y}_R$ is given by

$$\hat{Y}_R = \sum_{r \in s} w_r y_r^* + \sum_{i \in s} \frac{w_i}{\hat{p}_i} a_i (y_i - y_i^*).$$

Calibration imputation consists of finding final imputed values y_i^* , for $i \in s_m$, that minimize the distance function

$$\sum_{i \in s} q_i (\tilde{y}_i^* - y_i^*)^2.$$

subject to the constrain $\hat{Y}_I = \hat{Y}_R$, where q_i is a weight than can be calculated for every sample unit.

4.2.3 Modified deterministic regression imputation

We assume for now that the response probabilities p_i are known. We then use the imputed values y_i^* given by

$$y_i^* = \mathbf{z}_i' \boldsymbol{\beta} \quad (4.12)$$

for missing y_i and obtain the form of $\boldsymbol{\beta}$ that leads to an approximately unbiased estimator under assumption GNM.

Approximately unbiased estimator

The following lemma gives the form of $\boldsymbol{\beta}$ that leads to an approximately unbiased estimator under assumption GNM.

Lemma 4.1 : *Under assumption GNM, the choice of $\boldsymbol{\beta}$ that leads to an approximately unbiased estimator for the population total Y is given by*

$$\tilde{\mathbf{B}} = \left[\sum_{i \in s} w_i (1 - p_i) \mathbf{z}_i \mathbf{z}_i' / (\boldsymbol{\lambda}' \mathbf{z}_i) \right]^{-1} \sum_{i \in s} w_i (1 - p_i) \mathbf{z}_i y_i / (\boldsymbol{\lambda}' \mathbf{z}_i). \quad (4.13)$$

Proof. The conditional bias of $\hat{Y}_I$ with $y_i^* = \mathbf{z}_i' \boldsymbol{\beta}$ is given by

$$\text{Bias}(\hat{Y}_I | s) = - \sum_{i \in s} w_i (1 - p_i) (y_i - \mathbf{z}_i' \boldsymbol{\beta}).$$

Now, noting that $1 = \sigma^2 \boldsymbol{\lambda}' \mathbf{z}_i / \sigma_i^2$, the conditional bias of $\hat{Y}_I$ may be written as

$$\text{Bias}(\hat{Y}_I | s) = \sigma^2 \boldsymbol{\lambda}' \left[- \sum_{i \in s} w_i (1 - p_i) \mathbf{z}_i y_i / \sigma_i^2 + \sum_{i \in s} w_i (1 - p_i) \mathbf{z}_i \mathbf{z}_i' \boldsymbol{\beta} / \sigma_i^2 \right]. \quad (4.14)$$

It is clear from (4.14) that the conditional bias is zero if $\boldsymbol{\beta} = \tilde{\mathbf{B}}$ given by (4.13). ■

Note that $\tilde{\mathbf{B}}$ is unknown since the y -values are only observed for $i \in s_r$ and the response probabilities p_i are unknown. An estimator of $\tilde{\mathbf{B}}$, based on the responding units, is given by

$$\tilde{\mathbf{B}}_r = \left[\sum_{i \in s} w_i a_i \frac{(1 - \hat{p}_i)}{\hat{p}_i} \mathbf{z}_i \mathbf{z}_i' / (\boldsymbol{\lambda}' \mathbf{z}_i) \right]^{-1} \sum_{i \in s} w_i a_i \frac{(1 - \hat{p}_i)}{\hat{p}_i} \mathbf{z}_i y_i / (\boldsymbol{\lambda}' \mathbf{z}_i). \quad (4.15)$$

Given that $\hat{p}_i \approx p_i$, we have $E_r(\tilde{\mathbf{B}}_r | s) \approx \tilde{\mathbf{B}}_r$ so that $\tilde{\mathbf{B}}_r$ is conditionally approximately unbiased for $\tilde{\mathbf{B}}_r$ under assumption GNM. Hence, using the imputed values

$$y_i^* = \mathbf{z}_i' \tilde{\mathbf{B}}_r \quad (4.16)$$

with $\tilde{\mathbf{B}}_r$ given by (4.15) leads to an approximately unbiased estimator of the total Y . Note that $\tilde{\mathbf{B}}_r$ is a weighted least square estimator of $\boldsymbol{\beta}$ with respect to a new set of weights, $\tilde{w}_i = w_i \frac{(1 - \hat{p}_i)}{\hat{p}_i}$. This is equivalent to increasing the sampling weights for those units with a small probability of response and decreasing the sampling weights for those units with a large probability of response. The imputed estimator (4.2) can be implemented from the imputed data file containing the survey weights w_i and the $\tilde{y}_i$ -values. Response identifiers a_i and the estimated response probabilities, $\hat{p}_i$, are not required. The choice of imputed values (4.16) also leads to an approximately unbiased estimator of Y under assumption IM. Hence, this choice of imputed values leads to a robust imputed estimator in the sense of validity under both approaches.

Two particular cases of regression imputation are of interest: (i) ratio imputation ($\boldsymbol{\lambda} = 1, \mathbf{z}_i = z_i$), in which case the imputed values (4.16) reduce to

$$y_i^* = \frac{\sum_{i \in s} \tilde{w}_i a_i y_i}{\sum_{i \in s} \tilde{w}_i a_i z_i} z_i \quad (4.17)$$

and (ii) mean imputation ($\lambda = 1, \mathbf{z}_i = 1$) in which case the imputed values (4.17) reduce to

$$y_i^* = \frac{\sum_{i \in s} \tilde{w}_i a_i y_i}{\sum_{i \in s} \tilde{w}_i a_i}. \quad (4.18)$$

Under uniform response $p_i = p$, the imputed values (4.17) and (4.18) reduce to $(\sum_{i \in s_r} w_i y_i / \sum_{i \in s_r} w_i z_i) z_i$ and $\sum_{i \in s} w_i a_i y_i / \sum_{i \in s} w_i a_i$ respectively, which are the usual values that survey practitioners use for ratio imputation and mean imputation (Rao and Sitter, 1995).

Optimal choice of β

We now turn to the optimal choice of β by minimizing the conditional mean square error of the imputed estimator $\hat{Y}_I$ with $y_i^* = \mathbf{z}_i' \beta$. The conditional mean square error of the imputed estimator $\hat{Y}_I$ is given by

$$\begin{aligned} MSE_r(\hat{Y}_I|s) &= V_r(\hat{Y}_I|s) + [\text{Bias}(\hat{Y}_I|s)]^2 \\ &= \sum_{i \in s} w_i^2 p_i (1 - p_i) (y_i - \mathbf{z}_i' \beta)^2 \\ &\quad + \left[\sum_{i \in s} w_i (1 - p_i) (y_i - \mathbf{z}_i' \beta) \right]^2. \end{aligned} \quad (4.19)$$

We search for $\tilde{\beta}$ for which $MSE_r(\hat{Y}_I|s)$ is minimum.

Let $\mathbf{W}_s$ be a $n \times n$ diagonal matrix whose i^{th} component is w_i , $\mathbf{P}_s$ be a $n \times n$ diagonal matrix whose i^{th} component is p_i , $\mathbf{Z}_s$ be a $k \times n$ matrix whose columns are the k -dimensional vectors $\mathbf{z}_1, \dots, \mathbf{z}_n$, $\mathbf{y}_s = (y_1, \dots, y_n)'$ and $\mathbf{q}_s = (1 - p_1, \dots, 1 - p_n)'$.

Lemma 4.2 : The choice of $\tilde{\mathbf{B}}$ that minimizes $MSE_r(\hat{Y}_I|s)$ is given by

$$\begin{aligned} \tilde{\mathbf{B}}_{\text{opt}} = \{ \mathbf{Z}_s \mathbf{W}_s [\mathbf{P}_s (\mathbf{I}_s - \mathbf{P}_s) + \mathbf{q}_s \mathbf{q}_s'] \mathbf{W}_s \mathbf{Z}_s' \}^{-1} [\mathbf{Z}_s \mathbf{W}_s \mathbf{P}_s (\mathbf{I} - \mathbf{P}_s) \mathbf{W}_s \mathbf{y}_s \\ + \mathbf{q}_s' \mathbf{W}_s \mathbf{y}_s \mathbf{Z}_s \mathbf{W}_s \mathbf{q}_s], \end{aligned} \quad (4.20)$$

where $\mathbf{I}_s$ denotes the identity matrix of order n .

The optimal choice, $\tilde{\mathbf{B}}_{\text{opt}}$, of β is complex in general, but in the special case of ratio imputation $\tilde{\mathbf{B}}_{\text{opt}}$ reduces to

$$\tilde{\mathbf{B}}_{\text{opt}} = \frac{\sum_{i \in s} w_i (1 - p_i) y_i \sum_{i \in s} w_i (1 - p_i) z_i + \sum_{i \in s} w_i^2 p_i (1 - p_i) y_i z_i}{[\sum_{i \in s} w_i (1 - p_i) z_i]^2 + \sum_{i \in s} w_i^2 p_i (1 - p_i) z_i^2}. \quad (4.21)$$

Assume that the sampling weights w_i satisfy $\max(\frac{n}{N} w_i) = O(1)$ and that there exists a positive constant C such that $C < p_i$. Then,

$$\begin{aligned} \tilde{\mathbf{B}}_{\text{opt}} &= \frac{\sum_{i \in s} w_i (1 - p_i) y_i}{\sum_{i \in s} w_i (1 - p_i) z_i} + O\left(\frac{1}{n}\right) \\ &= \tilde{\mathbf{B}} + O\left(\frac{1}{n}\right). \end{aligned}$$

Hence, for large sample sizes, the choice $\tilde{\mathbf{B}}$ is nearly optimal for ratio imputation.

Similarly, $\tilde{\mathbf{B}}$ is nearly optimal for mean imputation which is a special case of ratio imputation.

4.2.4 Random regression imputation

Random imputation can be viewed as deterministic imputation plus a random noise.

Let s_r and s_m denote the sets of sample respondents and nonrespondents respectively,

and let $e_j = (y_j - \mathbf{z}_j' \hat{\mathbf{B}}_r) / (\lambda' \mathbf{z}_j)^{1/2}$ be the standardized residuals for the respondents

$j \in s_r$ under deterministic regression imputation. Further, $e_i^* = e_j$ with $P(e_i^* = e_j) =$

$w_j / \sum_{i \in s} w_i a_i$ independently for each $i \in s_m$. Then, random regression imputation uses the imputed values $y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r + \epsilon_i^*$, $i \in s_m$, where $\epsilon_i^* = (\boldsymbol{\lambda}' \mathbf{z}_i)^{1/2} (e_i^* - \bar{e}_r)$ with $\bar{e}_r = \sum_{i \in s} w_j a_j e_j / \sum_{i \in s} w_j a_j$. We have $E_*(\epsilon_i^*) = 0$ and $E_*(\hat{Y}_I)$ equals (4.6). Hence, the imputed estimator $\hat{Y}_I$ is approximately unbiased under either NM or IM.

The imputed estimator based on random regression imputation is asymptotically biased under assumption GNM. To obtain an approximately unbiased estimator for Y , we propose modified random regression imputation. Let $\tilde{e}_j = (y_j - \mathbf{z}_j' \tilde{\mathbf{B}}_r) / (\boldsymbol{\lambda}' \mathbf{z}_j)^{1/2}$ and $\tilde{e}_i^* = \tilde{e}_j$ with $P(\tilde{e}_i^* = \tilde{e}_j) = \tilde{w}_j / \sum_{i \in s} \tilde{w}_i a_i$ independently for each $i \in s_m$, where $\tilde{\mathbf{B}}_r$ is given by (4.15) and $\tilde{w}_i = w_i (1 - \hat{p}_i) / \hat{p}_i$. Then, modified random regression imputation uses the imputed values $y_i^* = \mathbf{z}_i' \tilde{\mathbf{B}}_r + \tilde{\epsilon}_i^*$, where $\tilde{\epsilon}_i^* = (\boldsymbol{\lambda}' \mathbf{z}_i)^{1/2} (\tilde{e}_i^* - \tilde{e}_{r\bullet})$ with $\tilde{e}_{r\bullet} = \sum_{i \in s} \tilde{w}_j a_j \tilde{e}_j / \sum_{i \in s} \tilde{w}_j a_j$. We have $E_*(\tilde{\epsilon}_i^*) = 0$ and $E_*(\hat{Y}_I)$ equals the imputed estimator under modified deterministic regression imputation. Hence, the imputed estimator $\hat{Y}_I$ is approximately unbiased under either GNM or IM. For the special case of modified random hot-deck imputation, we have $\tilde{\mathbf{B}}_r = \sum_{i \in s} \tilde{w}_i a_i y_i / \sum_{i \in s} \tilde{w}_i a_i$ and $y_i^* = y_j$ corresponds to the respondent value y_j drawn at random with probability $\tilde{w}_j / \sum_{i \in s} \tilde{w}_i a_i$.

4.3 Variance Estimation

In this section, we derive a variance estimator of the imputed estimator, $\hat{Y}_I$, using the reverse approach of Fay (1991). The total variance of $\hat{Y}_I$ is given by

$$V(\hat{Y}_I - Y) = E_r V_p(\hat{Y}_I - Y | \mathbf{a}) + V_r E_p(\hat{Y}_I - Y | \mathbf{a}), \quad (4.22)$$

where $\hat{Y}_I$ is given by (4.3) with $y_i^* = \mathbf{z}_i' \tilde{\mathbf{B}}_r$ and $\mathbf{a} = (a_1, \dots, a_N)'$ is the vector of response indicators (Shao and Steel, 1999). An estimator of the overall variance $V(\hat{Y}_I - Y)$ in (4.22) is given by $v_t = v_1 + v_2$, where v_1 is an estimator of $V_p(\hat{Y}_I - Y|\mathbf{a})$ conditional on the response indicators a_i , and v_2 is an estimator of $V_r[E_p(\hat{Y}_I - Y|\mathbf{a})]$.

Under modified random regression imputation, $y_i^* = \mathbf{z}_i' \tilde{\mathbf{B}}_r + \tilde{\epsilon}_i^*$, the variance of the imputed estimator $\hat{Y}_I$ is given by

$$\begin{aligned} V(\hat{Y}_I - Y) &= E_r \left[V_p E_\star (\hat{Y}_I - Y|\mathbf{a}) + E_r E_p V_\star (\hat{Y}_I - Y|\mathbf{a}) \right] \\ &\quad + V_r E_p E_\star (\hat{Y}_I - Y|\mathbf{a}). \end{aligned} \quad (4.23)$$

We assume that $E_\star(\hat{Y}_I|\mathbf{a})$ agrees with the imputed estimator for the deterministic case. Hence, $E_r V_p E_\star(\hat{Y}_I - Y|\mathbf{a})$ is estimated by v_1 above for the deterministic case. Similarly, $V_r E_p E_\star(\hat{Y}_I - Y|\mathbf{a})$ is estimated by v_2 above for the deterministic case. The additional contribution to variance due to random imputation comes from the component $E_r E_p V_\star(\hat{Y}_I - Y|\mathbf{a})$, which is estimated by $v_\star = V_\star(\hat{Y}_I - Y|\mathbf{a})$. Hence, it follows from (4.23) that the overall variance $V(\hat{Y}_I - Y)$ is estimated by $v_t = v_1 + v_\star + v_2$. The term $v_\star$ is absent for deterministic imputation.

4.3.1 Known p_i

In this section, we assume that the response probabilities p_i are known. We first consider the case of modified deterministic regression imputation in section 4.2.3. The case of modified random regression imputation will be considered in section 4.2.4.

Modified deterministic regression imputation

Under modified deterministic regression imputation, the imputed estimator $\hat{Y}_I$ may be written as

$$\hat{Y}_{Ip} = \sum_{i \in s} w_i a_i y_i + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r)' \tilde{\mathbf{B}}_{rp}, \quad (4.24)$$

where

$$\tilde{\mathbf{B}}_{rp} = \left[\sum_{i \in s} w_i a_i \frac{(1-p_i)}{p_i} \mathbf{z}_i \mathbf{z}_i' / (\boldsymbol{\lambda}' \mathbf{z}_i) \right]^{-1} \left[\sum_{i \in s} w_i a_i \frac{(1-p_i)}{p_i} \mathbf{z}_i y_i / (\boldsymbol{\lambda}' \mathbf{z}_i) \right]. \quad (4.25)$$

To obtain v_1 , we use Taylor linearization which leads to

$$\hat{Y}_{Ip} - Y \approx \sum_{i \in s} w_i \hat{\xi}_i, \quad (4.26)$$

where

$$\tilde{\xi}_{ip} = a_i y_i + (1 - a_i) \mathbf{z}_i' \tilde{\mathbf{B}}_{rp} + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r)' \tilde{\mathbf{T}}_p^{-1} a_i \frac{(1-p_i)}{p_i} \frac{1}{(\boldsymbol{\lambda}' \mathbf{z}_i)} \mathbf{z}_i (y_i - \mathbf{z}_i' \tilde{\mathbf{B}}_{rp})$$

with $\tilde{\mathbf{T}}_p = \sum_{i \in s} w_i a_i \frac{(1-p_i)}{p_i} \mathbf{z}_i \mathbf{z}_i' / (\boldsymbol{\lambda}' \mathbf{z}_i)$. Denoting the variance estimator of the full sample estimator $\hat{Y} = \sum_{i \in s} w_i y_i$ as $v(y)$, it follows from (4.26) that an estimator of $V_p(\hat{Y}_I - Y | \mathbf{a})$ is given by

$$v_1 = v(\tilde{\xi}_p), \quad (4.27)$$

which is obtained by replacing y_i by $\tilde{\xi}_{ip}$ in the formula for $v(y)$.

To obtain the second component v_2 , first note that

$$E_p(\hat{Y}_{Ip} - Y | \mathbf{a}) \approx \sum_{i \in U} a_i y_i + \sum_{i \in U} (1 - a_i) \mathbf{B}_p - Y,$$

where $\mathbf{B}_p = \left[\sum_{i \in U} a_i \frac{(1-p_i)}{p_i} \mathbf{z}_i \mathbf{z}_i' / (\boldsymbol{\lambda}' \mathbf{z}_i) \right]^{-1} \sum_{i \in U} a_i \frac{(1-p_i)}{p_i} \mathbf{z}_i y_i / (\boldsymbol{\lambda}' \mathbf{z}_i)$. Using Taylor linearization, it can be shown that

$$V_r \left[E_p \left(\hat{Y}_{Ip} - Y | \mathbf{a} \right) \right] \approx \sum_{i \in U} \zeta_i^2 p_i (1 - p_i), \quad (4.28)$$

where

$$\zeta_i = \left[1 + \frac{(1-p_i)}{p_i} \frac{1}{(\boldsymbol{\lambda}' \mathbf{z}_i)} (\mathbf{Z} - \mathbf{Z}_r)' \mathbf{T}_p^{-1} \mathbf{z}_i (y_i - \mathbf{z}_i' \mathbf{B}_p) \right],$$

where $\mathbf{Z} = \sum_{i \in U} \mathbf{z}_i$, $\mathbf{Z}_r = \sum_{i \in U} a_i \mathbf{z}_i$ and $\mathbf{T}_p = \sum_{i \in U} a_i \frac{(1-p_i)}{p_i} \mathbf{z}_i \mathbf{z}_i' / (\boldsymbol{\lambda}' \mathbf{z}_i)$. The component v_2 is then obtained by estimating the unknown quantities in (4.28), which leads to

$$v_2 = \sum_{i \in s} w_i (1 - p_i) a_i \tilde{\zeta}_i^2, \quad (4.29)$$

where

$$\tilde{\zeta}_i = \left[1 + \frac{(1-p_i)}{p_i} \frac{1}{(\boldsymbol{\lambda}' \mathbf{z}_i)} (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r)' \tilde{\mathbf{T}}_p^{-1} \mathbf{z}_i (y_i - \mathbf{z}_i' \tilde{\mathbf{B}}_{rp}) \right].$$

An estimator of the total variance v_t is the sum of (4.27) and (4.29).

Modified random regression imputation

We first note that

$$V_*(y_i^*) = (\boldsymbol{\lambda}' \mathbf{z}_j) \sum_{i \in s} w_j \frac{(1-p_j)}{p_j} a_i (\tilde{e}_j - \tilde{e}_r)^2 / \sum_{i \in s} w_j \frac{(1-p_j)}{p_j} a_j \equiv \tilde{s}_e^2$$

and $Cov_*(y_i^*, y_j^*) = 0, i \neq j$. Hence, from (4.2) the component v_* , due to random imputation, is given by

$$v_* = \sum_{i \in s} w_i^2 (1 - a_i) V_*(y_i^*) = \sum_{i \in s} w_i^2 (1 - a_i) \tilde{s}_e^2. \quad (4.30)$$

An estimator of the total variance is the sum of (4.27), (4.29) and (4.30).

4.3.2 Unknown p_i

We use the linearization method of Binder (1983) to derive the component v_1 when the response probabilities are estimated. We assume that $p_i = f(\mathbf{u}_i' \boldsymbol{\eta})$, where $f(\mathbf{u}_i' \boldsymbol{\eta})$ is given by the logistic regression model (2). The estimated response probabilities are given by $\hat{p}_i = f(\mathbf{u}_i' \hat{\boldsymbol{\eta}})$, where $\hat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$. Let $\boldsymbol{\theta} = (\boldsymbol{\eta}'_N, \mathbf{B}', Y)'$, where $\boldsymbol{\eta}_N$ and $\boldsymbol{\gamma}_N$ are the “census” parameters corresponding to $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ respectively. An estimator of $\boldsymbol{\theta}$ given by $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\eta}}', \bar{\mathbf{B}}'_r, \hat{Y}_I)'$ can be expressed as a solution of the sample estimating equations

$$\hat{\mathbf{S}}(\boldsymbol{\theta}) = \mathbf{0},$$

where $\hat{\mathbf{S}}(\boldsymbol{\theta}) = (\hat{\mathbf{S}}_1(\boldsymbol{\theta}), \hat{\mathbf{S}}_2(\boldsymbol{\theta}), \hat{\mathbf{S}}_3(\boldsymbol{\theta}))'$ with

$$\begin{aligned}\hat{\mathbf{S}}_1(\boldsymbol{\theta}) &= \sum_{i \in s} w_i \mathbf{u}_i [a_i - f(\mathbf{u}_i' \boldsymbol{\eta}_N)] = 0, \\ \hat{\mathbf{S}}_2(\boldsymbol{\theta}) &= \sum_{i \in s} w_i a_i \mathbf{z}_i \frac{(1 - f(\mathbf{u}_i' \boldsymbol{\eta}))}{f(\mathbf{u}_i' \boldsymbol{\eta})} (y_i - \mathbf{z}_i' \mathbf{B}) / (\boldsymbol{\lambda}' \mathbf{z}_i) = 0\end{aligned}$$

and

$$\hat{\mathbf{S}}_3(\boldsymbol{\theta}) = Y - \sum_{i \in s} w_i \mathbf{z}_i' \mathbf{B} - \sum_{i \in s} w_i a_i (y_i - \mathbf{z}_i' \mathbf{B}) = 0.$$

Let $\hat{\mathbf{J}}(\boldsymbol{\theta}) = \partial \hat{\mathbf{S}}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ be the $(k + l + 1) \times (k + l + 1)$ matrix of partial derivative.

We have

$$\mathbf{V}(\hat{\boldsymbol{\theta}} | \mathbf{a}) = [\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})] \boldsymbol{\Sigma}(\boldsymbol{\theta}) [\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})]',$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ denotes the $(k + l + 1) \times (k + l + 1)$ covariance matrix. If $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is replaced by a consistent estimator $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$, say, we obtain a consistent variance estimator

$\mathbf{v}(\hat{\boldsymbol{\theta}})$ given by

$$\mathbf{v}(\hat{\boldsymbol{\theta}}) = [\hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}})] \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}) [\hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}})]'.$$

Since we are interested in the variance estimator, v_1 , of $\hat{Y}_I$, we need the final row $\hat{\mathbf{b}}$, say, of $\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. It follows that

$$v_1 = \hat{\mathbf{b}} \hat{\boldsymbol{\Sigma}} \hat{\mathbf{b}}'.$$

To obtain the component v_2 , we assume that the sampling weights w_i satisfy $\max(\frac{n}{N}w_i) = O(1)$ and that there exists a positive constant C such that $C < p_i$. Furthermore, we assume that $\hat{\boldsymbol{\eta}} - \boldsymbol{\eta} = O_p(n^{-1/2})$. By Taylor linearization, we have

$$\hat{Y}_I = \hat{Y}_{Ip} + (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \sum_{i \in U} p_i^{-1} (y_i - \bar{\mathbf{B}}_{\mathbf{a}}) \frac{\partial f(\mathbf{u}_i' \boldsymbol{\eta})}{\partial \boldsymbol{\eta}} + O_p(Nn^{-1}),$$

where

$$\bar{\mathbf{B}}_{\mathbf{a}} = \left[\sum_{i \in U} (1 - a_i) \mathbf{z}_i \mathbf{z}_i' / (\boldsymbol{\lambda}' \mathbf{z}_i) \right]^{-1} \left[\sum_{i \in U} (1 - a_i) \mathbf{z}_i y_i / (\boldsymbol{\lambda}' \mathbf{z}_i) \right].$$

Assuming that $\frac{\partial f(\mathbf{u}_i' \boldsymbol{\eta})}{\partial \boldsymbol{\eta}}$ is uniformly bounded, we have

$$E_p(\hat{Y}_I) = E_p(\hat{Y}_{Ip}) + O_p(Nn^{-1/2}).$$

Hence, the component $V_r \left[E_p(\hat{Y}_{Ip} - Y | \mathbf{a}) \right]$ is approximately given by (4.28) and v_2 is given by (4.29).

4.4 Simulation Studies

We performed two simulation studies to investigate the finite sample performance of the proposed imputation method. The first simulation study compares the performance of the traditional deterministic regression imputation and the proposed

modified deterministic regression imputation when the imputation model and/or the nonresponse model are not correctly specified. In practice, imputation classes homogeneous with respect to estimated response probabilities are often formed. It is then customary to assume uniform response within classes and use weighted random hot-deck or mean imputation in each class and independently across classes. Alternatively, the proposed modified regression imputation method may be used. The second simulation study compares the performance of the imputed estimator obtained by using imputation classes based on the estimated response probabilities with the imputed estimator obtained by using the proposed modified regression imputation method.

4.4.1 Simulation study 1

We generated a finite population of size $N = 1,000$ containing 3 variables: a variable of interest y and two auxiliary variables z_1 and z_2 . To do so, we first generated z_1 and z_2 independently from an exponential distribution with mean 4 and 30 respectively. Then the y -values were generated according to the regression model

$$y_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \epsilon_i,$$

where the ϵ_i 's are generated from a normal distribution with mean 0 and variance σ^2 . The values of the parameters γ_0 , γ_1 and γ_2 were respectively set to 20, 2 and 0.1 and the variance σ^2 was chosen to lead to a model R^2 -value approximately equal to 0.75. The objective is to estimate the population total $Y = \sum_{i \in U} y_i$.

We generated $R = 5,000$ simple random samples without replacement of size

$n = 100$ from the finite population. In each sample, nonresponse to item y was generated according to the following response mechanisms:

Mechanism 1: Response probability p_{1i} for unit i is given by the logistic model

$$\log \frac{p_{1i}}{1 - p_{1i}} = \lambda_0 + \lambda_1 z_{1i}.$$

Mechanism 2: Response probability p_{2i} for unit i is given by the logistic model

$$\log \frac{p_{2i}}{1 - p_{2i}} = \lambda_0 + \lambda_1 y_i.$$

The values of λ_0 and λ_1 were chosen so the mean of the p_i 's is approximately equal to 70%. The response indicators a_{1i} and a_{2i} were then generated independently from a Bernoulli distribution with parameters p_{1i} and p_{2i} , respectively.

To compensate for nonresponse to item y , we used the traditional deterministic regression imputation for which the imputed values are given by (4.4) and the modified deterministic regression imputation for which the imputed values are given by (4.16). Imputations were based on the models for y and for p listed in Table 4.1. Note that in the case of the nonresponse mechanism 2, the probability of response depends on the variable of interest y . Hence, we performed stepwise regression analyses for each simulated sample to obtain the best model for p given the variables z_1 and z_2 . The resulting model for the response probability is the one containing an intercept as well as the variable z_1 , denoted by $p_{(1)}$ in Table 4.1.

From each simulated sample, we calculated the imputed $\hat{Y}_I$ given by (4.2) with the imputed values (4.4) and (4.16) respectively. As a measure of the bias of an estimator $\hat{\theta}$ of θ , we used the percent relative bias given by

$$B_r(\hat{\theta}) = \frac{\text{Bias}(\hat{\theta})}{\theta} \times 100. \quad (4.31)$$

As a measure of variability of, we used the percent relative root mean square error (RRMSE) given by

$$RRMSE(\hat{\theta}) = \frac{\sqrt{MSE(\hat{\theta})}}{\theta} \times 100. \quad (4.32)$$

The results on relative bias and RRMSE are shown in Table 4.2 for the response mechanism 1, and in Table 4.3 for the response mechanism 2. From Table 4.2, it is clear that when the imputation is performed according to the correct model (i.e., $y_{(1)}$), the traditional deterministic regression imputation leads to an approximately unbiased estimator and is slightly more efficient than the modified deterministic regression imputation. However, when the imputation model is incorrectly specified (e.g., $y_{(2)}$ and $y_{(4)}$), the traditional deterministic imputation leads to biased estimators whereas the bias of the modified deterministic imputation is small to negligible provided the nonresponse model is correctly specified (e.g., $p_{(1)}$). As a result, the RRMSE for the traditional deterministic imputation is larger than that for the modified regression imputation. When both the imputation and nonresponse models are not correctly specified (e.g., $y_{(4)} - p_{(2)}$), all the estimators are biased. From Table 4.2, it is clear that the imputed estimator obtained under modified regression imputation performs equally or better than the imputed estimator obtained under traditional regression imputation in all the scenarios.

Table 4.1: Models used for imputation

Models for y	Intercept	z_1	z_2
$y_{(1)}$	yes	yes	yes
$y_{(2)}$	yes	no	yes
$y_{(3)}$	yes	yes	no
$y_{(4)}$	no	yes	yes
Models for p_i	Intercept	z_1	z_2
$p_{(1)}$	yes	yes	no
$p_{(2)}$	yes	no	yes
$p_{(3)}$	no	yes	no

4.4.2 Simulation study 2

We first generated a finite population of size $N = 1,000$ containing 4 variables: a variable of interest y and 3 auxiliary variables z_1 , z_2 and z_3 . To do so, we first generated z_1 , z_2 and z_3 independently from an exponential distribution with mean 100. Then the y -values were generated according to the regression model

$$y_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{1i}^2 + \epsilon_i,$$

where the ϵ_i 's are generated from a normal distribution with mean 0 and variance σ^2 .

The values of the parameters γ_0 , γ_1 , γ_2 and γ_3 were respectively fixed to 20, 10, 0.5 and 10. The variance σ^2 has been chosen to lead to a model R^2 approximately equal to 0.66. The objective is to estimate the population mean $\bar{Y} = \sum_{i \in U} y_i / N$. In order

Table 4.2: Relative bias (%) and RRMSE (%) of the imputed estimators under Mechanism 1

Model	Bias (traditional)	Bias (modified)	RRMSE (traditional)	RRMSE (modified)
$y_{(1)} - p_{(1)}$	0.19	-0.01	1.85	2.33
$y_{(2)} - p_{(1)}$	5.2	0.16	5.60	2.66
$y_{(3)} - p_{(1)}$	0.17	-0.04	1.87	2.37
$y_{(4)} - p_{(1)}$	-14.8	-3.5	15.0	6.7
$y_{(1)} - p_{(2)}$	0.19	0.12	1.85	1.86
$y_{(4)} - p_{(2)}$	-14.8	-14.8	15.0	14.6
$y_{(1)} - p_{(3)}$	0.19	0.05	1.85	1.88

to focus on the nonresponse/imputation error, we considered the case of a census, i.e., $n = N = 1000$. From the population, nonresponse to item y was generated according to the following response mechanisms:

Mechanism 1: the response probability p_{1i} for unit i is given by the logistic model

$$\log \frac{p_{1i}}{1 - p_{1i}} = \lambda_0 + \lambda_1 z_{1i} + \lambda_2 z_{3i}.$$

Mechanism 2: the response probability p_{2i} for unit i is given by the logistic model

$$\log \frac{p_{2i}}{1 - p_{2i}} = \lambda_0 + \lambda_1 y_i + \lambda_2 z_{3i}.$$

The values of λ_0 , λ_1 and λ_2 have been chosen so the mean of the response probabilities is approximately equal to 70%. The response indicators a_{1i} and a_{2i} were then gener-

Table 4.3: Relative bias (%) and RRMSE (%) of the imputed estimators under Mechanism 2

Model	Bias	Bias	RRMSE	RRMSE
Model	(traditional)	(modified)	(traditional)	(modified)
$y_{(1)} - p_{(1)}$	1.84	1.83	2.55	2.54
$y_{(2)} - p_{(1)}$	4.46	1.84	4.89	2.65
$y_{(3)} - p_{(1)}$	2.03	2.02	2.70	2.70
$y_{(4)} - p_{(1)}$	-4.58	-3.04	5.07	3.81
$y_{(1)} - p_{(2)}$	1.84	1.84	2.55	2.55
$y_{(4)} - p_{(2)}$	-4.58	-1.70	5.07	2.88
$y_{(1)} - p_{(3)}$	1.84	1.84	2.55	2.55

ated independently $R = 1,000$ times from a Bernoulli distribution with parameters p_{1i} and p_{2i} , respectively. To compensate for nonresponse, two strategies were used: the first used the modified regression imputation method based on the auxiliary variables. The second consists in dividing the sample, s , into imputation classes $s_1, s_2, \dots, s_c$ based on the auxiliary variables. To form the classes, we used the score method which may be described as follows: Using the auxiliary information, first estimate the response probabilities, p_i , to obtain $\hat{p}_i$ for both the respondents and the nonrespondents using logistic regression. Using the $\hat{p}_i$'s, partition the population using a classification algorithm FASTCLUS of SAS (that uses the "k-means" classification algorithm). The score method leads to a partition of the population in such a way

that, within classes, units (respondents and nonrespondents) are homogeneous with respect to $\hat{p}_i$. The goal of the simulation study is thus to compare the performance of two imputed estimators: (a) the first, denoted $\bar{y}_I$, is given by

$$\bar{y}_I = \frac{1}{\hat{N}} \left[\sum_{i \in s} w_i a_i y_i + \sum_{i \in s} w_i (1 - a_i) y_i^* \right], \quad (4.33)$$

where the y_i^* are the imputed values by (4.16) and $\hat{N} = \sum_{i \in s} w_i$; (b) the imputed estimator based on C imputation classes given by

$$\bar{y}_I^C = \sum_{c=1}^C \frac{\hat{N}_c}{\hat{N}} \bar{y}_{Ic},$$

where

$$\bar{y}_{Ic} = \frac{1}{\hat{N}_c} \left[\sum_{i \in s_c} w_i a_i y_i + \sum_{i \in s_c} w_i (1 - a_i) y_i^* \right], \quad (4.34)$$

and $\hat{N}_c = \sum_{i \in s_c} w_i$. We used mean imputation within classes; i.e.,

$y_i^* = \sum_{i \in s_c} w_i a_i y_i / \sum_{i \in s_c} w_i a_i$. Note that, in this simulation study, $w_i = 1$ for all $i \in U$. Also, note that for Mechanism 1, the response probabilities p_i were estimated using the variable z_1 and z_3 whereas we used the variables z_1 , z_2 and z_3 for Mechanism 2. Finally, we compared these estimators in terms of their relative bias given by (4.31) and their RRMSE given by (4.32). From Table 4.4, it is clear that the estimator given by (4.33) performs considerably better than the estimator (4.34) in terms of RRMSE for both Mechanism 1 and Mechanism 2.

Table 4.4: Relative bias (%) and RRMSE (%) of the imputed estimators

Nonresponse mechanism	Number of classes	$B_r(\bar{y}_I^C)$	$RRMSE(\bar{y}_I^C)$
Mechanism 1	1	14.4	14.5
	5	-0.02	4.26
	10	-0.85	7.33
	20	-0.20	8.61
Imputed estimator using modified regression imputation	—	1.11	1.90
Mechanism 2	1	29.0	29.1
	5	21.4	21.4
	10	21.0	21.1
	20	20.9	21.0
Imputed estimator using modified regression imputation	—	10.9	10.9

4.5 Estimation of Domain Means

As noted in Chapter 3, domains are generally not specified at the imputation stage and domain estimates are obtained from the imputed data. For this reason, we propose a bias-adjusted estimator along the lines of section 4.2.2. The bias-adjusted estimator can be obtained at the estimation stage and does not require the specification of the domains at the imputation stage.

A vector of domain means may be expressed as

$$\bar{\mathbf{Y}}_{(d)} = \left(\sum_{i \in U} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i \in U} \mathbf{x}_i y_i \quad (4.35)$$

where $\mathbf{x} = (x_{1i}, \dots, x_{di}, \dots, x_{ki})'$ is the vector of domain indicators, x_{di} , such that $x_{di} = 1$ if $i \in \text{domain } d$ and $x_{di} = 0$, otherwise. We assume that $\mathbf{x}$ is known for all the units $i \in s$. In other words, only item y may be missing.

In the presence of nonresponse to item y , an imputed estimator of $\bar{\mathbf{Y}}_{(d)}$ is given by (3.3)

In Chapter 3, we showed that the imputed estimator $\hat{\mathbf{Y}}_{I(d)}$ is biased under assumption NM. We proposed a bias-adjusted estimator which is approximately unbiased under either NM or IM. In this section, we propose an extension of the bias-adjusted estimator which is approximately unbiased under either GNM or IM.

It is easily seen that the asymptotic conditional bias of the imputed estimator (3.3) that uses the modified deterministic regression imputation (4.16) is given by

$$\text{Bias} \left(\hat{\mathbf{Y}}_{I(d)} | s \right) = E_r \left(\hat{\mathbf{Y}}_{I(d)} - \bar{\mathbf{Y}}_{(d)} | s \right) \approx -\hat{\mathbf{T}}^{-1} \left[\sum_{i \in s} w_i (1 - p_i) \mathbf{x}_i \left(y_i - \mathbf{z}_i' \tilde{\mathbf{B}}_r \right) \right] \quad (4.36)$$

where $\tilde{\mathbf{B}}_r$ is given by (4.13). An approximately conditionally unbiased estimator of the bias in (4.36) is given by

$$\hat{B} \left(\hat{\mathbf{Y}}_{I(d)} | s \right) \approx -\hat{\mathbf{T}}^{-1} \left[\sum_{i \in s} \tilde{w}_i a_i \mathbf{x}_i \left(y_i - \mathbf{z}_i' \tilde{\mathbf{B}}_r \right) \right], \quad (4.37)$$

where $\tilde{\mathbf{B}}_r$ is given by (4.15). A bias-adjusted estimator, $\hat{\mathbf{Y}}_{I(d)}^{*a}$, is then obtained as $\hat{\mathbf{Y}}_{I(d)} - \hat{B} \left(\hat{\mathbf{Y}}_{I(d)} | s \right)$, which leads to

$$\hat{\mathbf{Y}}_{I(d)}^{*a} = \hat{\mathbf{T}}^{-1} \left[\sum_{i \in s} \frac{w_i}{\hat{p}_i} a_i \mathbf{x}_i \left(y_i - \mathbf{z}_i' \tilde{\mathbf{B}}_r \right) + \sum_{i \in s} w_i \mathbf{x}_i \mathbf{z}_i' \tilde{\mathbf{B}}_r \right]. \quad (4.38)$$

The bias-adjusted estimator (4.38) is approximately unbiased under either IM or GNM. Hence, it is robust in the sense of validity under both IM and GNM. However, it requires both the imputation flags and the estimated response probabilities $\hat{p}_i$, unlike in the case of estimating the population total (section 4.2.3). It is possible to obtain a bias-adjusted estimator of the form (4.38) if we use the traditional deterministic regression imputation instead. It is interesting to note that in this case the bias-adjusted estimator is identical to the estimator obtained using calibrated imputation (Beaumont, 2005). In the case of calibrated imputation, however, the domains must be specified at the imputation stage, which is not generally the case in practice.

If the nonresponse model (4.1) contains only the intercept, we have $\hat{p}_i = \hat{p}$, where $\hat{p}$ is the overall response rate. In this case, the bias-adjusted estimator (4.38) reduces to

$$\hat{\mathbf{Y}}_{\mathbf{I}(\mathbf{d})}^{\mathbf{a}} = \hat{p}^{-1} \hat{\mathbf{Y}}_{\mathbf{I}(\mathbf{d})} + (1 - \hat{p}^{-1}) \hat{\mathbf{T}}^{-1} \sum_{i \in s} w_i \mathbf{x}_i \mathbf{z}_i' \hat{\mathbf{B}}_{\mathbf{I}}, \quad (4.39)$$

noting that under deterministic regression, imputation,

$$\hat{\mathbf{B}}_{\mathbf{I}} = \left(\sum_{i \in s} w_i \mathbf{z}_i \mathbf{z}_i' / (\boldsymbol{\lambda}' \mathbf{z}_i) \right)^{-1} \left[\sum_{i \in s} w_i a_i \mathbf{z}_i y_i / (\boldsymbol{\lambda}' \mathbf{z}_i) + \sum_{i \in s} w_i (1 - a_i) \mathbf{z}_i y_i^* / (\boldsymbol{\lambda}' \mathbf{z}_i) \right] = \hat{\mathbf{B}}_{\mathbf{r}},$$

which is identical to (3.10).

Chapter 5

Inference in two-stage cluster sampling

5.1 Introduction

Multi-stage sampling is often used in surveys, especially when direct element sampling is impractical (or impossible). In this chapter, we confine to the case of two-stage cluster sampling. We adopt the following notation: Let a finite population consist of N nonoverlapping clusters, U_i of sizes M_i , $i = 1, \dots, N$. Let $K = \sum_{i=1}^N M_i$ denote the total number of ultimate units (elements) in the population. Further, let y_{ij} be the value of the variable of interest y for the j^{th} element in the i^{th} cluster, $i = 1, \dots, N$; $j = 1, \dots, M_i$ and Y_i be the i -th cluster total. The objective is to estimate the population total $Y = \sum_{i=1}^N Y_i = \sum_{i=1}^N \sum_{j=1}^{M_i} y_{ij}$, by selecting a sample according to a two-stage design: at the first stage, a random sample of clusters, s , of size n , is selected according to a given design $p(s)$ from the population of clusters. At the second stage, a random

sample of element, s_i , of size m_i ($i = 1, \dots, n$) is selected according to a given design $p_i(s_i)$ if the i -th cluster is sampled. Under complete response to item y , an estimator of Y , denoted $\hat{Y}$, is given by

$$\hat{Y} = \sum_{i \in s} \sum_{j \in s_i} w_{ij} y_{ij}, \quad (5.1)$$

with $w_{ij} = w_i w_{j|i}$, where $w_i = 1/\pi_i$, $w_{j|i} = 1/\pi_{j|i}$, $\pi_i = P(i \in s)$ is the inclusion probability of cluster i in s , and $\pi_{j|i}$ is the conditional probability of inclusion of unit j belonging to cluster i in s_i , $i = 1, \dots, N$; $j = 1, \dots, M_i$. The estimator $\hat{Y}$ given in (5.1) is design-unbiased for Y ; that is, $E_p(\hat{Y}) = Y$, where $E_p(\cdot)$ denotes the expectation with respect to the sampling design. In the case of two-stage sampling, note that $E_p(\cdot) = E_1 E_2(\cdot|s)$, where $E_1(\cdot)$ and $E_2(\cdot)$ denote respectively the expectation with respect to the first and second stages.

In the case of nonresponse to item y , an imputed estimator of Y , denoted $\hat{Y}_I$, is given by

$$\hat{Y}_I = \sum_{i \in s} \left[\sum_{j \in s_i} w_{ij} a_{ij} y_{ij} + \sum_{j \in s_i} w_{ij} (1 - a_{ij}) y_{ij}^* \right], \quad (5.2)$$

where y_{ij}^* denotes the imputed value for missing y_{ij} and a_{ij} is a response indicator such that $a_{ij} = 1$ if element j in cluster i responds to item y and $a_{ij} = 0$ otherwise.

Let s_{r_i} be the set of respondents in cluster i , of size r_i and let $s_r = \bigcup_{i=1}^n s_{r_i}$ be the total set of respondents of size $r = \sum_{i \in s} r_i$.

For simplicity, we assume a single imputation class, but the results extend to multiple classes. In this chapter, we first consider the case of mean imputation that

uses the imputed values

$$y_{ij}^* = \bar{y}_r = \frac{\sum_{i \in s} \sum_{j \in s_i} w_{ij} a_{ij} y_{ij}}{\sum_{i \in s} \sum_{j \in s_i} w_{ij} a_{ij}}. \quad (5.3)$$

Using the imputed values (5.3) in (5.2), the imputed estimator $\hat{Y}_I$ reduces to

$$\hat{Y}_I = \frac{\sum_{i \in s} \sum_{j \in s_i} w_{ij}}{\sum_{i \in s} \sum_{j \in s_i} w_{ij} a_{ij}} \sum_{i \in s} \sum_{j \in s_i} w_{ij} a_{ij} y_{ij}. \quad (5.4)$$

As noted in Chapter 1, two distinct approaches have been used in the literature to study the properties of the imputed estimator: (i) Design-based approach (ii) Model-based approach. The customary design-based approach is based on

Assumption DB: $E_r(a_{ij}) = p_1$ (uniform response) and $E_r(a_{ij}a_{i'j'}) = E_r(a_{ij})E_r(a_{i'j'}) = p_1^2$ except for $i = i'$ and $j = j'$ (independence of the response statuses).

The customary model-based approach is based on

Assumption MB: The response mechanism is ignorable in the sense that whether or not a unit responds does not depend on the variable being imputed but may depend on the covariates in the assumed imputation model. For mean imputation, the model is given by (1.18).

As noted in Chapter 1, the imputed estimator (5.4) is robust in the sense that it is approximately unbiased under assumption DB as well as assumption MB.

Assumptions DB and MB may not be tenable in the case of two-stage sampling because the within cluster correlations are not taken into account. Therefore, in section 5.2, we propose more realistic assumptions that reflect dependence within a cluster and show that the imputed estimator (5.4) remains valid. Section 5.3 compares the

conditional variances, given the sample, to study the effect of within-cluster correlations. In section 5.4, we derive variance estimators under the model-based approach, using Särndal's method and Shao-Steel's method (based on the reverse approach of Fay, 1991). In section 5.5, we compare the variances of the imputed estimator in the case of the reverse approach. In section 5.6, a simulation study is conducted to compare the two variance estimation methods when the imputation model is misspecified. Variance estimation under random hot-deck imputation is discussed in section 5.7. Finally, in section 5.8, we study the properties EBLUP imputation.

5.2 Approaches suitable for cluster sampling

In this section we propose weaker assumptions, corresponding to the two approaches, that reflect within-cluster correlations. The design-based approach is now based on

Assumption DBC: $E_r(a_{ij}) = p_1$ and $E_r(a_{ij}a_{ij'}) = p_2$, for $j \neq j'$. Note that, in general, $E_r(a_{ij}a_{ij'}) \neq E_r(a_{ij})E_r(a_{ij'}) = p_1^2$; that is, the response statuses between two units in the same cluster are not independent. In what follows, this response mechanism will be referred as "beta-binominal" response mechanism.

The model-based approach is now based on

Assumption MBC: The response mechanism is ignorable. For mean imputation, the model is the well known one-way ANOVA model with random effects given by

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad (5.5)$$

where μ is the general mean, α_i is i -th cluster random effect and ϵ_{ij} is the residual error. We assume that

- (i) $E_m(\alpha_i) = E_m(\epsilon_{ij}) = 0$,
- (ii) $Cov_m(\epsilon_{ij}, \epsilon_{i'j'}) = 0$ except for $i = i'$ and $j = j'$, $Cov_m(\alpha_i, \alpha_{i'}) = 0 \forall i \neq i'$,
 $Cov_m(\alpha_i, \epsilon_{i'j'}) = 0 \forall i, i'$ and j' ,
- (iii) $V_m(\alpha_i) = \sigma_\alpha^2 \forall i$, $V_m(\epsilon_{ij}) = \sigma_\epsilon^2 \forall i, j$.

From (i)-(iii), we get

$$Cov_m(y_{ij}, y_{i'j'}) = \begin{cases} \sigma_\alpha^2 & \text{if } i = i' \text{ and } j \neq j' \\ \sigma_\alpha^2 + \sigma_\epsilon^2 & \text{if } i = i' \text{ and } j = j' \\ 0 & \text{if } i \neq i' \end{cases}$$

Once again, the imputed estimator (5.4) is robust in the sense that it is approximately unbiased under assumption DBC or assumption MBC. Note that model (5.5) is also used for random hot-deck imputation (see section 5.7).

5.3 Comparison of conditional variances under the two-phase approach

Recall from Chapter 1 that the total error, $\hat{Y}_I - Y$, may be decomposed as

$$\hat{Y}_I - Y = (\hat{Y} - Y) + (\hat{Y}_I - \hat{Y}). \quad (5.6)$$

Since the sampling error does not depend on nonresponse and on the imputation method, we focus on the nonresponse component $\hat{Y}_I - \hat{Y}$ to study the conditional variance $\hat{Y}_I - \hat{Y}$, under the design-based and the model-based approaches, given the full sample of elements $\tilde{s} = \bigcup_{i \in s} s_i$.

5.3.1 Design-based approach

In this section, we study the conditional variance $V_r \left[\left(\hat{Y}_I - \hat{Y} | \tilde{s} \right) \right]$, under assumption DB and assumption DBC. First, using Taylor linearization, it is easily seen that under assumption DB,

$$V_r^{DB} \left(\hat{Y}_I - \hat{Y} | \tilde{s} \right) \approx \frac{p_1 (1 - p_1)}{p_1^2} \sum_{i \in s} \sum_{j \in s_i} w_{ij}^2 (y_{ij} - \bar{y})^2, \quad (5.7)$$

where $\bar{y} = \sum_{i \in s} \sum_{j \in s_i} w_{ij} y_{ij} / \sum_{i \in s} \sum_{j \in s_i} w_{ij}$.

Similarly, under assumption DBC,

$$\begin{aligned} V_r^{DBC} \left(\hat{Y}_I - \hat{Y} | \tilde{s} \right) &\approx \frac{p_1 (1 - p_1)}{p_1^2} \sum_{i \in s} \sum_{j \in s_i} w_{ij}^2 (y_{ij} - \bar{y})^2 \\ &+ \frac{p_2 - p_1^2}{p_1^2} \sum_{i \in s} \sum_{j \in s_i} \sum_{\substack{j' \in s_i \\ j' \neq j}} w_{ij} w_{ij'} (y_{ij} - \bar{y}) (y_{ij'} - \bar{y}). \end{aligned} \quad (5.8)$$

Let $\tilde{R}_{DB} = V_r^{DBC} \left(\hat{Y}_I - \hat{Y} | \tilde{s} \right) / V_r^{DB} \left(\hat{Y}_I - \hat{Y} | \tilde{s} \right)$. Then, we have

$$\tilde{R}_{DB} = 1 + \rho_p (\hat{c}_{DB} - 1), \quad (5.9)$$

where

$$\rho_p = \frac{Cov_r(a_{ij}, a_{ij'})}{\sqrt{V_r(a_{ij}) V_r(a_{ij'})}} = \frac{p_2 - p_1^2}{p_1 (1 - p_1)}$$

is the intraclass correlation of the response indicators under assumption DBC and

$$\hat{c}_{DB} = \frac{\sum_{i \in s} \left(\sum_{j \in s_i} w_{ij} (y_{ij} - \bar{y}) \right)^2}{\sum_{i \in s} \sum_{j \in s_i} w_{ij}^2 (y_{ij} - \bar{y})^2}.$$

Several points may be noted from (5.9). First, if $\rho_p = 0$ then $\tilde{R}_{DB} = 1$, as expected.

If ρ_p is large, then $\tilde{R}_{DB}$ may be substantial. Hence, in this case, using assumption

DB instead of assumption DBC may lead to severe underestimation of the variance

of the imputed estimator $\hat{Y}_I$. Using the Cauchy-Schwarz inequality, $\left(\sum_{j \in s_i} b_{ij} c_{ij}\right)^2 \leq \sum_{j \in s_i} b_{ij}^2 \sum_{j \in s_i} c_{ij}^2$, with $b_{ij} = 1$ and $c_{ij} = w_{ij} (y_{ij} - \bar{y})$, we have

$$\left(\sum_{j \in s_i} w_{ij} (y_{ij} - \bar{y})\right)^2 \leq m_i \sum_{j \in s_i} w_{ij}^2 (y_{ij} - \bar{y})^2,$$

which implies that

$$\tilde{c}_{DB} \leq \frac{\sum_{i \in s} \sum_{j \in s_i} m_i w_{ij}^2 (y_{ij} - \bar{y})^2}{\sum_{i \in s} \sum_{j \in s_i} w_{ij}^2 (y_{ij} - \bar{y})^2} \equiv \tilde{d}_{DB}.$$

It follows that an upper bound for $\tilde{R}_{DB}$ is given by

$$\tilde{R}_{DB} \leq 1 + \rho_p (\tilde{d}_{DB} - 1) \quad (5.10)$$

if $\rho_p \geq 0$. In the particular case of equal subsample sizes, $m_i = m$, (5.10) becomes

$$\tilde{R}_{DB} \leq 1 + \rho_p (m - 1). \quad (5.11)$$

Expression (5.11) suggests that the ratio $\tilde{R}_{DB}$ increases, for fixed ρ_p , as the number of elements selected in each cluster, m , increases or as ρ_p increases for fixed m .

5.3.2 Model-based approach

In this section, we study the conditional variance $V_m(\hat{Y}_I - \hat{Y}|\bar{s})$, under assumption MB and assumption MBC. First, noting that

$$\hat{Y}_I - \hat{Y} = \sum_{i \in s} \sum_{j \in s_i} w_{ij} d_{ij} y_{ij}$$

with

$$d_{ij} = \frac{\sum_{i \in s} \sum_{j \in s_i} w_{ij} a_{ij}}{\sum_{i \in s} \sum_{j \in s_i} w_{ij}} a_{ij} - 1,$$

it is easily seen that under assumption MB

$$V_m^{MB} \left(\hat{Y}_I - \hat{Y} | \bar{s} \right) = \sigma^2 \sum_{i \in s} \sum_{j \in s_i} w_{ij}^2 d_{ij}^2. \quad (5.12)$$

Similarly, it is easily seen that under assumption MBC

$$\begin{aligned} V_m^{MBC} \left(\hat{Y}_I - \hat{Y} | \bar{s} \right) &= (\sigma_\alpha^2 + \sigma_\epsilon^2) \sum_{i \in s} \sum_{j \in s_i} w_{ij}^2 d_{ij}^2 \\ &+ \sigma_\alpha^2 \sum_{i \in s} \sum_{j \in s_i} \sum_{\substack{j' \in s_i \\ j \neq j'}} w_{ij} w_{ij'} d_{ij} d_{ij'}. \end{aligned} \quad (5.13)$$

Noting that $\sigma^2 = \sigma_\alpha^2 + \sigma_\epsilon^2$, the ratio $\tilde{R}_{MB} = V_m^{MBC} \left(\hat{Y}_I - \hat{Y} | \bar{s} \right) / V_m^{MB} \left(\hat{Y}_I - \hat{Y} | \bar{s} \right)$ is equal to

$$\tilde{R}_{MB} = 1 + \rho_m (\tilde{c}_{MB} - 1), \quad (5.14)$$

where

$$\rho_m = \frac{\text{Cov}_m(y_{ij}, y_{ij'})}{\sqrt{V_m(y_{ij}) V_m(y_{ij'})}} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}$$

is the intraclass correlation of the y -values under assumption MBC and

$$\tilde{c}_{MB} = \frac{\sum_{i \in s} \left(\sum_{j \in s_i} w_{ij} d_{ij} \right)^2}{\sum_{i \in s} \sum_{j \in s_i} w_{ij}^2 d_{ij}^2}.$$

Once again, using Cauchy-Schwarz inequality, we have

$$\left(\sum_{j \in s_i} w_{ij} d_{ij} \right)^2 \leq m_i \sum_{j \in s_i} w_{ij}^2 d_{ij}^2,$$

which implies that

$$\tilde{c}_{MB} \leq \frac{\sum_{i \in s} \sum_{j \in s_i} m_i w_{ij}^2 d_{ij}^2}{\sum_{i \in s} \sum_{j \in s_i} w_{ij}^2 d_{ij}^2} \equiv \tilde{d}_{MB}.$$

It follows that an upper bound for $\tilde{R}_{MB}$ is given by

$$\tilde{R}_{MB} \leq 1 + \rho_m (\tilde{d}_{MB} - 1) \quad (5.15)$$

if $\rho_m \geq 0$. In the particular case $m_i = m$, the expression (5.15) becomes

$$\tilde{R}_{MB} \leq 1 + \rho_m (m - 1). \quad (5.16)$$

It follows from (5.16) that the ratio $\tilde{R}_{MB}$ increases as m increases for fixed ρ_m or as ρ_m increases for fixed m .

5.4 Variance estimation under the model-based approach

In this section we derive variance estimators under both Särndal's method (Särndal, 1992) and the Shao-Steel method (Shao and Steel, 1999).

5.4.1 Complete response

In the case of complete response, the variance of $\hat{Y}$ in (5.1) is given by

$$V(\hat{Y}) = V_1 E_2(\hat{Y}|s) + E_1 V_2(\hat{Y}|s),$$

where $V_1(\cdot)$ and $V_2(\cdot)$ denote the variance with respect to first-stage and second-stage sampling, respectively. An estimator of the variance of $\hat{Y}$ is given by

$$v(\hat{Y}) = \sum_{i \in s} \sum_{k \in s} \Omega_{ik} \hat{Y}_i \hat{Y}_k + \sum_{i \in s} \frac{\hat{V}_i}{\pi_i}, \quad (5.17)$$

where $\Omega_{ik} = \frac{1}{\pi_i \pi_k} - \frac{1}{\pi_{ik}}$, with π_{ik} denoting the joint inclusion probability of clusters i and k in the sample s ,

$$\hat{Y}_i = \sum_{j \in s_i} \frac{y_{ij}}{\pi_{j|i}}, \quad (5.18)$$

and $\hat{V}_i$ is an estimator of $V_2(\hat{Y}_i|s)$ given by

$$\hat{V}_i = \sum_{j \in s_i} \sum_{l \in s_i} \Lambda_{jl|i} y_{ij} y_{il}, \quad (5.19)$$

where $\Lambda_{jl|i} = \frac{1}{\pi_{j|i}\pi_{l|i}} - \frac{1}{\pi_{jl|i}}$ with $\pi_{jl|i}$ denoting the joint inclusion probability of elements j and l in the subsample s_i . Note that the variance estimator $v(\hat{Y})$ in (5.17) is unbiased for $V(\hat{Y})$; i.e., $E_1 E_2 [v(\hat{Y}) | s] = V(\hat{Y})$.

5.4.2 Särndal's method

In this section, we derive an estimator of the variance under mean imputation, using Särndal's method and assumption MBC. Recall from Chapter 1, that an estimator of the variance of $\hat{Y}_I$ is given by

$$\hat{V}_{tot} = \hat{V}_{sam} + \hat{V}_{imp} + 2\hat{V}_{mix}, \quad (5.20)$$

where $\hat{V}_{sam} = \hat{V}_I + \hat{V}_{dif}$ denotes an estimator of the sampling variance.

We now obtain expressions for the different components in (5.20). First, note that the component $\hat{V}_I$ in $\hat{V}_{sam}$ is obtained by replacing $\hat{Y}_i$ by $\hat{Y}_{Ii}$ and $\hat{V}_i$ by $\hat{V}_{Ii}$ in (5.17), where $\hat{Y}_{Ii}$ and $\hat{V}_{Ii}$ are computed the same way as $\hat{Y}_i$ and $\hat{V}_i$ by treating the imputed values as if they were true values. The expression for $\hat{V}_{dif}$ involves tedious but somewhat straightforward algebra. We obtain

$$\hat{V}_{dif} = \hat{V}_{dif}^{(1)} + \hat{V}_{dif}^{(2)}, \quad (5.21)$$

where

$$\begin{aligned}\hat{V}_{dif}^{(1)} &= \sum_{i \in s} \Omega_{ii} \left[\hat{\sigma}_\epsilon^2 (\hat{Q}_i - \hat{U}_i) + \hat{\sigma}_\alpha^2 (\hat{M}_i^2 - \hat{R}_i^2) \right] \\ &\quad - \frac{1}{\hat{R}^2} \left[\hat{\sigma}_\epsilon^2 \sum_{i \in s} w_i^2 \hat{U}_i + \hat{\sigma}_\alpha^2 \sum_{i \in s} w_i^2 \hat{R}_i^2 \right] \sum_{i \in s} \sum_{k \in s} \Omega_{ik} (\hat{M}_i - \hat{R}_i) (\hat{M}_k - \hat{R}_k) \\ &\quad - \frac{2}{\hat{R}} \sum_{i \in s} \sum_{k \in s} \Omega_{ik} w_i (\hat{M}_i - \hat{R}_i) (\hat{\sigma}_\epsilon^2 \hat{V}_i + \hat{\sigma}_\alpha^2 \hat{R}_i),\end{aligned}$$

with

$$\hat{Q}_i = \sum_{j \in s_i} w_{j|i}^2, \hat{U}_i = \sum_{j \in s_i} w_{j|i}^2 a_{ij}, \hat{M}_i = \sum_{j \in s_i} w_{j|i} \text{ and } \hat{R}_i = \sum_{j \in s_i} w_{j|i} a_{ij},$$

and

$$\hat{V}_{dif}^{(2)} = \sum_{i \in s} w_i \hat{V}_i^{(dif)},$$

where

$$\begin{aligned}\hat{V}_i^{(dif)} &= \hat{\sigma}_\epsilon^2 \sum_{j \in s_i} \Lambda_{jj|i} (1 - a_{ij}) + 2\hat{\sigma}_\alpha^2 \sum_{j \in s_i} \sum_{l \in s_i} \Lambda_{jl|i} a_{ij} (1 - a_{il}) \\ &\quad + \hat{\sigma}_\alpha^2 \sum_{j \in s_i} \sum_{l \in s_i} \Lambda_{jl|i} (1 - a_{ij}) (1 - a_{il}) \\ &\quad - 2\hat{\sigma}_\alpha^2 \frac{\hat{R}_i}{\hat{R}} w_i \sum_{j \in s_i} \sum_{l \in s_i} \Lambda_{jl|i} a_{ij} (1 - a_{il}) \\ &\quad - 2\frac{\hat{\sigma}_\epsilon^2}{\hat{R}} w_i \sum_{j \in s_i} \sum_{l \in s_i} \Lambda_{jl|i} a_{ij} (1 - a_{il}) w_{j|i} \\ &\quad - \frac{1}{\hat{R}^2} \left(\hat{\sigma}_\epsilon^2 \sum_{i \in s} w_i^2 \hat{Q}_i + \hat{\sigma}_\alpha^2 \sum_{i \in s} w_i^2 \hat{R}_i^2 \right) \sum_{j \in s_i} \sum_{l \in s_i} \Lambda_{jl|i} (1 - a_{ij}) (1 - a_{il}),\end{aligned}$$

and $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_\alpha^2$ are estimators of σ_ϵ^2 and σ_α^2 , respectively.

Also, we obtain

$$\hat{V}_{imp} = \hat{\sigma}_\epsilon^2 \sum_{i \in s} \sum_{j \in s_i} w_{ij}^2 (\hat{p}^{-1} a_{ij} - 1)^2 + \hat{\sigma}_\alpha^2 \sum_{i \in s} w_i^2 (\hat{p}^{-1} \hat{R}_i - \hat{M}_i)^2, \quad (5.22)$$

where $\hat{p} = \sum_{i \in s} \sum_{j \in s_i} w_{ij} a_{ij} / \sum_{i \in s} \sum_{j \in s_i} w_{ij}$.

Finally, we obtain

$$\begin{aligned} \hat{V}_{mix} = & \hat{\sigma}_\epsilon^2 \sum_{i \in s} \sum_{j \in s_i} w_{ij}^2 (\hat{p}^{-1} a_{ij} - 1) + \hat{\sigma}_\alpha^2 \sum_{i \in s} w_i^2 \hat{M}_i (\hat{p}^{-1} \hat{R}_i - \hat{M}_i) \\ & - \sum_{i \in s} (\hat{p}^{-1} \hat{R}_i - \hat{M}_i) (\hat{M}_i \hat{\sigma}_\alpha^2 + \hat{\sigma}_\epsilon^2). \end{aligned} \quad (5.23)$$

Note that $\hat{V}_{mix} = 0$ in the case of a self-weighting design, $w_{ij} = w_o$, noting that $\hat{p}^{-1} \hat{R}_i - \hat{M}_i = 0$.

The estimators $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\epsilon^2$ can be obtained using available methods such as maximum likelihood (ML), restricted maximum likelihood (REML) or minimum norm quadratic unbiased estimation (MINQUE), (e.g., Searle, Casella and McCulloch, 1992). Finally, the variance estimator of $\hat{Y}_I$ under assumption MB is obtained by setting $\hat{\sigma}_\alpha^2 = 0$ in (5.21) - (5.23).

5.4.3 Shao-Steel's method

To apply the Shao-Steel method, we first express $\hat{Y}_I$ as $\hat{Y}_I = \hat{K} \hat{Y}_r / \hat{R}$, where $\hat{K} = \sum_{i \in s} \sum_{j \in s_i} w_{ij}$, $\hat{Y}_r = \sum_{i \in s} \sum_{j \in s_i} w_{ij} a_{ij} y_{ij}$, $\hat{R} = \sum_{i \in s} \sum_{j \in s_i} w_{ij} a_{ij}$. Now, denote the estimator of the variance of $\hat{Y} = \sum_{i \in s} \sum_{j \in s_i} w_{ij} y_{ij}$ based on the full sample as $v(y)$. Then, one can show, using Taylor linearization, that v_1 reduces to

$$v_1 = v(\hat{\xi}), \quad (5.24)$$

where

$$\hat{\xi}_{ij} = a_{ij} y_{ij} + (1 - a_{ij}) \bar{y}_r + \frac{(\hat{K} - \hat{R})}{\hat{R}} a_{ij} (y_{ij} - \bar{y}_r). \quad (5.25)$$

Hence, the component v_1 is obtained from (5.17) by replacing y_{ij} by $\hat{\xi}_{ij}$, where $\hat{\xi}_{ij}$ is given by (5.25). To obtain the second component, v_{2MBC} , under assumption MBC, note that $E_p(\hat{Y}_I - Y|\mathbf{a}) = \sum_{i \in U} \sum_{j \in U_i} b_{ij} y_{ij}$, where $b_{ij} = (K/R) a_{ij} - 1$ with $R = \sum_{i \in U} R_i$ and $R_i = \sum_{j \in U_i} a_{ij}$. Also, it is easy to show that

$$\begin{aligned} E_r V_m E_p(\hat{Y}_I - Y|\mathbf{a}) &\approx \sigma_\epsilon^2 K \left(\frac{K}{E_r(R)} - 1 \right) + \sigma_\alpha^2 \left[\frac{K^2}{E_r(R^2)} \sum_{i \in U} E_r(R_i^2) \right. \\ &\quad \left. - 2 \frac{K}{E_r(R)} \sum_{i \in U} M_i E_r(R_i) + \sum_{i \in U} M_i^2 \right]. \end{aligned} \quad (5.26)$$

The component v_{2MBC} is then obtained by substituting estimators for the unknown quantities in (5.26), which leads to

$$\begin{aligned} v_{2MBC} &= \hat{\sigma}_\epsilon^2 \hat{K} \left(\frac{\hat{K}}{\hat{R}} - 1 \right) + \hat{\sigma}_\alpha^2 \left\{ \frac{\hat{K}^2}{\hat{R}^2} \sum_{i \in s} w_i \left[\hat{R}_i^2 - \hat{V}_2(\hat{R}_i) \right] \right. \\ &\quad \left. - 2 \frac{\hat{K}}{\hat{R}} \sum_{i \in s} w_i M_i \hat{R}_i + \sum_{i \in s} w_i M_i^2 \right\}, \end{aligned} \quad (5.27)$$

where $\hat{R}_i = \sum_{j \in s_i} w_{j|i} a_{ij}$.

Note that since $V_2(\hat{R}_i) = E_2(\hat{R}_i^2) - E_2(\hat{R}_i)^2$, an estimate of R_i^2 is given by $\hat{R}_i^2 - \hat{V}_2(\hat{R}_i)$, where $\hat{V}_2(\hat{R}_i)$ is an estimate of $V_2(\hat{R}_i)$. The sum of (5.24) and (5.27) gives v_{LMBC} . The variance estimator under assumption MB is given by $v_1 + v_{2MB}$, where v_{2MB} is obtained by setting $\hat{\sigma}_\alpha^2 = 0$ in (5.27). Note that the ratio v_1/v_{2MBC} is $O(n/N)$. Hence, the component v_{2MBC} may be negligible relative to v_1 when the first-stage sampling fraction is negligible. As a result, when n/N is negligible, the computation of v_{2MBC} may be omitted.

5.5 Comparison of the variance under the reverse approach

In this section, we compare the magnitude of the second component $E_r V_m E_p (\hat{Y}_I - Y | \mathbf{a})$ under assumption MBC and assumption MB. For simplicity, we consider the particular case of single-stage cluster sampling with equal cluster sizes (i.e., $m_i = M_i = M$).

Denote the component $E_r V_m E_p (\hat{Y}_I - Y | \mathbf{a})$ under assumption MBC by $\tilde{V}_m^{MBC} E_p (\hat{Y}_I - Y | \mathbf{a})$ and by $\tilde{V}_m^{MB} E_p (\hat{Y}_I - Y | \mathbf{a})$ under assumption MB. Note that $\tilde{V}_m^{MBC} E_p (\hat{Y}_I - Y | \mathbf{a})$ is given by (5.26), whereas $\tilde{V}_m^{MB} E_p (\hat{Y}_I - Y | \mathbf{a})$ is obtained by setting $\sigma_\alpha^2 = 0$ in (5.26) which leads to

$$\tilde{V}_m^{MBC} \approx \sigma^2 K \left(\frac{K}{E_r(R)} - 1 \right). \quad (5.28)$$

We are interested in the ratio

$$R_{MB}^* = \frac{\tilde{V}_m^{MBC} E_p (\hat{Y}_I - Y | \mathbf{a})}{\tilde{V}_m^{MB} E_p (\hat{Y}_I - Y | \mathbf{a})}.$$

To determine the ratio R_{MB}^* , we need to specify a response mechanism. We consider the uniform response mechanism as specified in assumption DB as well as the “beta-binomial” response mechanism as specified in assumption DBC.

5.5.1 Uniform response mechanism

Under uniform response, $p_i = p_1$, it is easily seen that

$$\tilde{V}_m^{MBC} E_p (\hat{Y}_I - Y | \mathbf{a}) = p_1^{-1} (1 - p_1) K (\sigma_\epsilon^2 + \sigma_\alpha^2), \quad (5.29)$$

and

$$\tilde{V}_m^{MB} E_p \left(\hat{Y}_I - Y | \mathbf{a} \right) \approx p_1^{-1} (1 - p_1) K \sigma^2. \quad (5.30)$$

Noting that $\sigma^2 = \sigma_\epsilon^2 + \sigma_\alpha^2$, it follows that $R_{MB}^* = 1$. This suggests that the uniform response assumption “washes off” the effect of using an incorrect model. Note that the order for both (5.29) and (5.30) is $O(K)$. Also, note that if there is no nonresponse (i.e., $p_1 = 1$), both (5.29) and (5.30) are equal to 0, as expected.

5.5.2 “Beta-binomial” response mechanism

Under the “beta-binomial” response mechanism specified in assumption DBC, it is easily seen that

$$\tilde{V}_m^{MB} E_p \left(\hat{Y}_I - Y | \mathbf{a} \right) \approx -K (M \sigma_\alpha^2 + \sigma_\epsilon^2) + K (M - 1) \sigma_\alpha^2 \frac{p_2}{p_1^2} + \frac{K}{p_1} (\sigma_\alpha^2 + \sigma_\epsilon^2). \quad (5.31)$$

Also, the component $\tilde{V}_m^{MB} E_p \left(\hat{Y}_I - Y | \mathbf{a} \right)$ is given by (5.30). It follows that

$$R_{MB}^* = 1 + (M - 1) \rho_m \rho_p. \quad (5.32)$$

First, note that under uniform response, $p_2 = p_1^2$ (or $\rho_p = 0$), (5.32) reduces to 1.

Also, note that the order of (5.31) is $O(KM)$ as opposed to (5.30) which is only $O(K)$. Table 5.1 shows the ratio R_{MB}^* with $M = 20$ for different values of ρ_m and ρ_p .

Table 5.1: Magnitude of R^* for different values of ρ_m and ρ_p and $M = 20$

$\rho_m \backslash \rho_p$	0.1	0.3	0.5	0.7	0.9
0.1	1.19	1.57	1.95	2.33	2.71
0.3	1.57	2.71	3.85	4.99	6.13
0.5	1.95	3.85	5.75	7.65	9.55
0.7	2.33	4.99	7.65	10.31	12.97
0.9	2.71	6.13	9.55	12.97	16.39

5.5.3 Choice of the model

Results in Table 5.1 show that choosing the wrong model may lead to appreciable underestimation of the second component $\tilde{V}_m E_p (\hat{Y}_I - Y | \mathbf{a})$, especially for large M , which might suggest that, in the case of single-stage or two-stage cluster sampling, one should use model (5.5) and not model (1.18). It should be noted, however, that as M increases, ρ_m and ρ_p typically decrease.

We now compare the total variance under uniform response and the “beta-binomial” response mechanism. First, under uniform response, the first component $\tilde{E}_r E_m V_p (\hat{Y}_I - Y | \mathbf{a})$ is given by

$$E_r E_m V_p (\hat{Y}_I - Y | \mathbf{a}) \approx NK \left(\frac{1}{n} - \frac{1}{N} \right) \times [p_1^{-1} (\sigma_\alpha^2 + \sigma_\epsilon^2) + (M - 1) \sigma_\alpha^2]. \quad (5.33)$$

The order of the first component in (5.33) is thus $O\left(\frac{K^2}{n}\right)$. We have noted in section 5.5.1 that, under uniform response, the order of the second component $\tilde{V}_m^{MBC} E_p (\hat{Y}_I - Y | \mathbf{a})$ is $O(K)$ so the order of $\frac{\tilde{V}_m^{MBC} (\hat{Y}_I - Y | \mathbf{a})}{\tilde{E}_r E_m V_p (\hat{Y}_I - Y | \mathbf{a})}$ is $O\left(\frac{n}{K}\right)$. Typically, $n \ll K$ in cluster sam-

pling so the second component is negligible relative to the first component. In this case, the computation of v_2 may be omitted. This may, however, be not true for the beta-binomial response mechanism. Under this response mechanism, one can show that the first component $E_r E_m V_p (\hat{Y}_I - Y | \mathbf{a})$ is given by

$$\begin{aligned} E_r E_m V_p (\hat{Y}_I - Y | \mathbf{a}) &\approx \frac{K^2}{A} N \left(\frac{1}{n} - \frac{1}{N} \right) \\ &\times \left(\left[(\sigma_\alpha^2 + \sigma_\epsilon^2) p_1 + (M-1) \sigma_\alpha^2 p_2 \left(1 + \frac{p_1 - p_2}{A} \right) \right] \right. \\ &+ \left[\frac{K-M}{N-1} (p_2 - p_1^2) \right. \\ &\left. \left. \times \left(\frac{\sigma_\alpha^2 + \sigma_\epsilon^2}{K p_1} + \frac{(M-1) \sigma_\alpha^2 p_2}{A} \right) \right] \right), \end{aligned} \quad (5.34)$$

where $A = M(p_2 - p_1^2) + K p_1^2$. First, note that under uniform response $p_2 = p_1^2$, (5.34) reduces to (5.33), assuming $K p_1 + (1 - p_1) \approx K p_1$. Second, note that the order (5.34) is $O\left(\frac{K^2}{n}\right)$. Third, note that under full response $p_1 = p_2 = 1$, (5.34) reduces to $E_m E_r V_p (\hat{Y}_I - Y | \mathbf{a}) \approx N K \left(\frac{1}{n} - \frac{1}{N}\right) (M \sigma_\alpha^2 + \sigma_\epsilon^2)$, as expected. Finally, in section 5.2, we have noted that under the beta-binomial response mechanism, the second component $\tilde{V}_m^{MBC} E_p (\hat{Y}_I - Y | \mathbf{a})$ was of order $O(KM)$ and therefore the order of $\frac{\tilde{V}_m^{MBC} E_p (\hat{Y}_I - Y | \mathbf{a})}{E_r E_m V_p (\hat{Y}_I - Y | \mathbf{a})}$ is $O\left(\frac{n}{N}\right)$. The second component is hence negligible only if the first-stage sampling fraction is negligible which might not be the case in practice. Indeed, it not unusual, in cluster sampling, to have first-stage sampling fractions as high as 0.5. In this case, computation of the second component must be performed and the choice of the imputation model may then become relevant.

We now investigate the magnitude of the underestimation of the total variance occurring when using the wrong model. A measure of underestimation, denoted by Q ,

is defined as

$$Q = 1 - \frac{E_m E_r V_p (\hat{Y}_I - Y | \mathbf{a}) + \tilde{V}_m^{MB} E_p (\hat{Y}_I - Y | \mathbf{a})}{E_m E_r V_p (\hat{Y}_I - Y | \mathbf{a}) + \tilde{V}_m^{MBC} E_p (\hat{Y}_I - Y | \mathbf{a})},$$

where $E_r E_m V_p (\hat{Y}_I - Y | \mathbf{a})$ is given by (5.34). Note that the first component of the total variance is computed under the correct model since its estimator is valid regardless of the model and/or the response mechanism. Tables 5.2, 5.3 and 5.4 show the magnitude of the underestimation for different values of ρ_m and ρ_p , with $M = 20$, $p_1 = 0.7$ and $\frac{n}{N} = 0.1, 0.25$ and 0.5 . The following conclusions may then be drawn from these tables: for modest and small first-stage sampling fractions, the underestimation is small to negligible; in this case, the choice of the model may not be relevant. As the first-stage sampling fraction $\frac{n}{N}$ increases, the underestimation increases for a given ρ_m and may be quite substantial (e.g., $\frac{n}{N} = 0.1$, $\rho_m = 0.3 \implies 0.36\% \leq Q \leq 2.54\%$ and for $\frac{n}{N} = 0.5$, $\rho_m = 0.3 \implies 3.03\% \leq Q \leq 18.40\%$). For a given $\frac{n}{N}$ and ρ_m , Q increases as ρ_p increases; (e.g., $\frac{n}{N} = 0.5$, $\rho_m = 0.5 \implies 3.34\% \leq Q \leq 19.61\%$). This indicates that departure from uniform response leads to underestimation of the total variance. For a given $\frac{n}{N}$ and ρ_p , Q increases as ρ_m increases; (e.g., $\frac{n}{N} = 0.5$, $\rho_p = 0.5 \implies 8.90\% \leq Q \leq 13.93\%$). Hence, correct specification of the model becomes critical when $\frac{n}{N}$, ρ_m and ρ_p increase.

Finally, we investigate the relative magnitude of underestimation, Q' , occurring when the computation of the second component is not performed. It is given by

$$Q' = 1 - \frac{E_m E_p V_p (\hat{Y}_I - Y | \mathbf{a})}{E_r E_m V_p (\hat{Y}_I - Y | \mathbf{a}) + \tilde{V}_m^{MBC} E_p (\hat{Y}_I - Y | \mathbf{a})}.$$

Tables 5.5, 5.6 and 5.7 show the magnitude of Q' for different values of ρ_m , ρ_p , with $M = 20$, $p_1 = 0.7$ and $\frac{n}{N} = 0.1, 0.25$ and 0.5 . These tables show that for a given ρ_m

and ρ_p , Q' increases as $\frac{n}{N}$ increases. Also, for large $\frac{n}{N}$ and given ρ_p , Q' decreases as ρ_m increases because, as ρ_m increases, the numerator of Q' increases faster than the second component in the denominator of Q' (Tables 5.6 and 5.7).

Table 5.2: Magnitude of $Q \times 100\%$ with $\frac{n}{N} = 0.1$

$\rho_m \backslash \rho_p$	0.1	0.3	0.5	0.7	0.9
0.1	0.26	0.74	1.18	1.57	1.94
0.3	0.36	1.01	1.58	2.09	2.54
0.5	0.39	1.09	1.70	2.23	2.70
0.7	0.41	1.13	1.76	2.30	2.78
0.9	0.42	1.16	1.79	2.34	2.82

5.6 Simulation Study

We conducted a limited simulation study on the relative performance of the variance estimators obtained by Särndal's method and the Shao-Steel's method. We generated several populations of $N = 120$ clusters with $M_i = M = 5$ or 20 elements in each cluster i such that the intraclass correlation $\rho_m = 0.05, 0.1, 0.2$. To do this, we first generated i.i.d. random variables

$$y_i \sim N(\mu, \sigma_\alpha^2), i = 1, \dots, N,$$

with $\mu = 200$ and $\sigma_\alpha^2 = 100$. Then, we generated i.i.d. random variables

$$\epsilon_{ij} \sim N(0, \sigma_\epsilon^2), j = 1, \dots, M_i,$$

with $\sigma_\epsilon^2 = \frac{1-\rho_m}{\rho_m} \sigma_\alpha^2$ for $\rho_m \neq 0$. The y -values are then generated by letting

$$y_{ij} = y_i + \epsilon_{ij}.$$

Table 5.3: Magnitude of $Q \times 100\%$ with $\frac{n}{N} = 0.25$

$\rho_m \backslash \rho_p$	0.1	0.3	0.5	0.7	0.9
0.1	0.75	2.14	3.37	4.48	5.49
0.3	1.07	2.95	4.57	5.96	7.18
0.5	1.16	3.20	4.92	6.38	7.65
0.7	1.21	3.32	5.08	6.58	7.87
0.9	1.24	3.40	5.18	6.70	8.00

From each generated population, we selected $R = 5000$ samples of clusters of size n according to simple random sampling without replacement, where $n/N = 0.1, 0.5$.

To simplify the discussion, we consider single-stage cluster sampling, i.e., $m_i = M_i$. In each sample of clusters, we assigned a probability p_i to cluster $i, i = 1, \dots, n$, where p_i was generated from a beta distribution. Then, response indicators a_{ij} were generated from a Bernoulli distribution with parameter p_i . The overall response rate was set to 0.65. From each sample, the imputed estimator $\hat{Y}_I$, given by (5.2), was computed and its variance was estimated using both Särndal's method and the Shao-Steel's method. For each method, the variance estimators were calculated under assumption MBC and assumption MB. The imputed estimator $\hat{Y}_I$ was approximately unbiased for all the scenarios with a Monte Carlo relative bias less than 0.3%.

As a measure of the bias of a variance estimator, $v(\hat{Y}_I)$, we used the relative bias

$$RB[v(\hat{Y}_I)] = \frac{E[v(\hat{Y}_I)] - MSE(\hat{Y}_I)}{MSE(\hat{Y}_I)} \times 100,$$

Table 5.4: Magnitude of $Q \times 100\%$ with $\frac{n}{N} = 0.5$

$\rho_m \backslash \rho_p$	0.1	0.3	0.5	0.7	0.9
0.1	2.07	5.75	8.90	11.63	14.03
0.3	3.03	8.11	12.21	15.57	18.40
0.5	3.34	8.84	13.20	16.70	19.61
0.7	3.49	9.19	13.66	17.24	20.18
0.9	3.58	9.40	13.93	17.55	20.52

where $MSE(\hat{Y}_I)$ denotes the mean square error of $\hat{Y}_I$. Table 5.1 and 5.2 report the relative bias values for $n/N = 0.1$ and 0.5 , respectively.

From Table 5.8 and Table 5.9, we note that both Särndal's and Shao-Steel's method perform well when the intracluster correlation is taken into account (i.e., the variance estimators are derived under assumption MBC) since the absolute value of the relative bias of the variance estimator is less than 7% in all cases. Note that the component $\hat{V}_{mix}$ in Särndal's method is equal to 0 under this particular design since it is self-weighting. When the intracluster correlation is not taken into account (i.e., the variance estimators are derived under assumption MB), Särndal's method leads to underestimation of the true variance. The underestimation increases as the intracluster correlation coefficient ρ_m and/or the size M increases. For example, when $n/N = 0.1$ and $\rho_m = 0.1$ the relative bias is -9.2% for $M_i = 5$ and -30.1% for $M = 20$. Also, $n/N = 0.1$ and $M = 20$ the relative bias is -17.6% for $\rho_m = 0.05$ and -39.8% if $\rho_m = 0.2$. Note that the sampling fraction n/N does not seem to have

Table 5.5: Magnitude of $Q' \times 100\%$ with $\frac{n}{N} = 0.1$

$\rho_m \backslash \rho_p$	0.1	0.3	0.5	0.7	0.9
0.1	1.64	2.05	2.43	2.77	3.09
0.3	1.01	1.72	2.15	2.62	3.04
0.5	1.82	1.49	2.07	2.58	3.03
0.7	1.72	1.43	2.03	2.56	3.02
0.9	1.64	1.39	2.01	2.55	3.02

a significant impact on the relative bias of the variance estimators. Also, it is interesting to note that Särndal's method performs poorly when the imputation model is incorrectly specified mainly because the component $\hat{V}_{dif}$ tends to be severely biased in those cases. The Shao-Steel's method, on the other hand, performs fairly well when the imputation model is incorrectly specified as the absolute value of the relative bias is less than 5% in all cases. This result is not surprising since the first component, v_1 , of the total variance is robust in the sense that it does not depend on the assumptions made about the imputation model and since the second component v_{2MBC} , is negligible with respect to v_1 when n/N is negligible. In summary, the Shao-Steel's method seems to be robust unlike Särndal's method when the assumptions about the model are violated.

Table 5.6: Magnitude of $Q' \times 100\%$ with $\frac{n}{N} = 0.25$

$\rho_m \backslash \rho_p$	0.1	0.3	0.5	0.7	0.9
0.1	4.74	5.90	6.93	7.86	8.70
0.3	2.95	4.69	6.17	7.46	8.58
0.5	2.39	4.32	5.95	7.34	8.55
0.7	2.12	4.15	5.85	7.29	8.53
0.9	1.96	4.05	5.78	7.26	8.52

5.7 Variance estimation under random hot-deck imputation

Random hot-deck imputation uses the imputed values (1.12). Note that $E_*(y_{ij}^*) = \bar{y}_r$, where $\bar{y}_r$ is given by (5.3). It follows that $E_*(\hat{Y}_I)$ is given by (5.4). We chose the Shao-Steel's method to estimate the variance under assumption MBC random hot deck imputation because of its simplicity. Indeed, Särndal's method involves very tedious algebra for obtaining the component $\hat{V}_{dif}$. Note that the components $\hat{V}_{imp}$ and $\hat{V}_{mix}$ are identical as for the case of mean imputation. Under assumption MBC, a variance estimator of $\hat{Y}_I$ under random hot-deck imputation is then given by $v_t = v_1 + v_* + v_{2MBC}$, where v_1 is given by (5.24), v_{2MBC} is given by (5.27) and v_* is given by

$$v_* = \sum_{i \in s} \sum_{j \in s_i} w_{ij}^2 (1 - a_{ij}) s_{yr}^2,$$

Table 5.7: Magnitude of $Q' \times 100\%$ with $\frac{n}{N} = 0.5$

$\rho_m \backslash \rho_p$	0.1	0.3	0.5	0.7	0.9
0.1	13.00	15.84	18.27	20.38	22.24
0.3	8.35	12.86	16.49	19.48	21.98
0.5	6.86	11.95	15.96	19.21	21.90
0.7	6.12	11.50	15.71	19.09	21.87
0.9	5.68	11.24	15.56	19.02	21.85

where $s_{yr}^2 = \frac{1}{\sum_{i \in s} \sum_{j \in s_i} w_{ij} a_{ij}} \sum_{i \in s} \sum_{j \in s_i} w_{ij} a_{ij} (y_{ij} - \bar{y}_r)^2$.

5.8 Deterministic EBLUP imputation

In previous sections, we have studied the case of mean and random hot-deck imputation. However, under assumption MBC, a better alternative would be to impute the empirical best linear predictor (EBLUP) of $\mu + \alpha_i$ based on the respondents. To illustrate this idea, we consider the following self-weighting design: simple random sampling without replacement at the first stage and simple random sampling without replacement at the second stage such that m_i/M_i is constant. It follows that $w_{ij} = w_o$, say. For missing y_{ij} , we impute $y_{ij}^* = EBLUP(\mu + \alpha_i)$, where $EBLUP(\mu + \alpha_i)$ is given by

$$EBLUP(\mu + \alpha_i) = \hat{\lambda}_i \bar{y}_{r_i} + (1 - \hat{\lambda}_i) \hat{\mu}, \quad (5.35)$$

Table 5.8: Relative bias of (%) the variance estimators with $n/N = 0.1$

	$\rho_m = 0.05$		$\rho_m = 0.1$		$\rho_m = 0.2$	
	$M_i = 5$	$M_i = 20$	$M_i = 5$	$M_i = 20$	$M_i = 5$	$M_i = 20$
Särndal's method (under MBC)	6.5	6.7	6.3	3.3	1.6	2.3
Shao-Steel method (under MBC)	-4.0	0.2	-1.2	-1.6	-3.2	-1.6
Särndal's method (under MB)	-6.2	-17.6	-9.2	-30.1	-19.6	-39.8
Shao-Steel method (under MB)	-4.7	-0.2	-1.5	-1.8	-3.6	-2.1

where $\hat{\mu}$ is the generalized least square estimator of μ given by $\hat{\mu} = \frac{\sum_s h_i \bar{y}_{r_i}}{\sum_s h_i}$ with $h_i = (\hat{\sigma}_\alpha^2 + \hat{\sigma}_\epsilon^2/r_i)^{-1}$, $\bar{y}_{r_i} = \frac{1}{r_i} \sum_{j \in s_i} a_{ij} y_{ij}$, r_i denotes the number of respondents in cluster i and $\hat{\lambda}_i = \hat{\sigma}_\alpha^2 h_i$.

Note that the imputed estimator $\hat{Y}_I$ is approximately unbiased for Y under assumption MBC and EBLUP imputation; i.e., $E_m(\hat{Y}_I - \hat{Y}|s) = 0$.

In order to compare mean imputation and EBLUP imputation in terms of efficiency, we assume that the variance components σ_α^2 and σ_ϵ^2 are known. Also, for simplicity, we assume equal $m_i (= m_o, \text{ say})$ and equal $r_i (= r_o, \text{ say})$.

Table 5.9: Relative bias (%) of the variance estimators with $n/N = 0.5$

	$\rho_m = 0.05$		$\rho_m = 0.1$		$\rho_m = 0.2$	
	$M_i = 5$	$M_i = 20$	$M_i = 5$	$M_i = 20$	$M_i = 5$	$M_i = 20$
Särndal's method (under MBC)	1.3	1.4	2.6	3.3	2.4	3.2
Shao-Steel method (under MBC)	0.2	0.3	1.2	0.2	0.6	1.5
Särndal's method (under MB)	-4.9	-18.0	-8.8	-27.1	-15.7	-38.1
Shao-Steel method (under MB)	-0.4	-1.6	-1.5	-2.1	-2.2	-2.6

In this case,

$$\text{EBLUP} = (\mu + \alpha_i) = \lambda \bar{y}_{ri} + (1 - \lambda) \bar{y}_r, \quad (5.36)$$

where $\lambda = \sigma_\alpha^2 h$ with $h = (\sigma_\alpha^2 + \sigma_\epsilon^2 / r_o)^{-1}$. It can be shown that the conditional variance under EBLUP imputation, denoted as $V_m^{\text{EBLUP}}(\hat{Y}_I - \hat{Y}|_s)$ is given by

$$\begin{aligned} V_m^{\text{EBLUP}}(\hat{Y}_I - \hat{Y}|_s) &= w_o^2 (n - 1) (m_o - r_o)^2 (1 - \lambda)^2 h^{-1} \\ &\quad + w_o^2 n (m_o - r_o) \left(2 - \frac{m_o}{r_o} \right) \sigma_\epsilon^2 \end{aligned} \quad (5.37)$$

Also, it can be shown that the conditional variance under mean imputation, denoted

by $V_m^{mean}(\hat{Y}_I - \hat{Y}|s)$, is given by

$$\begin{aligned} V_m^{mean}(\hat{Y}_I - \hat{Y}|s) &= w_o^2(n-1)(m_o - r_o)^2 h^{-1} \\ &\quad + w_o^2 n(m_o - r_o) \left(2 - \frac{m_o}{r_o}\right) \sigma_\epsilon^2. \end{aligned} \quad (5.38)$$

To compare mean imputation and EBLUP imputation, we use the following measure of percent relative efficiency, RE, given by

$$RE = \frac{V_m^{EBLUP}(\hat{Y}_I - \hat{Y}|s) - V_m^{mean}(\hat{Y}_I - \hat{Y}|s)}{V_m^{mean}(\hat{Y}_I - \hat{Y}|s)} \times 100. \quad (5.39)$$

Combining (5.37) and (5.38) as in (5.39) leads to

$$RE = \frac{\left[1 + \left(\frac{\rho_m}{1-\rho_m}\right) r_o\right]^{-2} - 1}{1 + \frac{n}{(n-1)} \left[\left(\frac{\rho_m}{1-\rho_m}\right) + r_o^{-1}\right] [(m_o - r_o)^{-1} - r_o^{-1}]}. \quad (5.40)$$

Note that $RE = 0$ when $\rho_m = 0$, as expected. Also, note that $RE \leq 0$ for any value of n, r_o, m_o and $\rho_m \geq 0$, showing that EBLUP imputation is always more efficient than mean imputation if $\rho_m > 0$.

Tables 5.10, 5.11 and 5.12 give values of RE for $n = 20, \rho_m = 0.05, 0.1, 0.3, 0.5; m_o = 5, 10, 20$ and $r_o/m_o = 0.4, 0.6$ and 0.8 .

It is clear from Tables 5.10, 5.11 and 5.12 that EBLUP imputation is increasingly more efficient than mean imputation as ρ_m increases for fixed m_o and r_o/m_o . For example, for $r_o/m_o = 0.6$ and $m_o = 5$, $RE = -17\%$ for $\rho_m = 0.05$ and $RE = -82\%$ for $\rho_m = 0.5$. Also, EBLUP imputation becomes increasingly more efficient than mean imputation as m_o increases for a fixed r_o/m_o and ρ_m . For example, for $r_o/m_o = 0.6$ and $\rho_m = 0.1$, $RE = -31\%$ for $m_o = 0.5$ and $RE = -66\%$ for $m_o = 20$. Finally, it is clear that EBLUP imputation is increasingly more efficient than mean imputation as

r_o/m_o decreases for fixed ρ_m and m_o . For example, for $\rho_m = 0.1$ and $m_o = 10$, we have $\text{RE} = -68\%$, whereas $\text{RE} = -26\%$ for $\rho_m = 0.1, m_o = 10$ and $r_o/m_o = 0.8$.

Note that the imputed estimator $\hat{Y}_I$ that uses the imputed values (5.36) is biased under assumption DBC. The bias is given by

$$\text{Bias}(\hat{Y}_I) = E_p E_r (\hat{Y}_I - Y|s) \approx (1-p) \lambda \left[\sum_{i \in U} \bar{Y}_i - Y \right], \quad (5.41)$$

where $\bar{Y}_i = Y_i/M_i$. The bias in (5.41) is equal to zero if $p = 1$ or if $\sum_{i \in U} \bar{Y}_i = Y$, which occurs, for example, when $\bar{Y}_i = \frac{Y}{N} \equiv \bar{Y}_0$, say for all $i \in U$. In general, the bias is not zero. Hence, EBLUP imputation can be justified under assumption MBC but not under assumption DBC. As a result, the imputed estimator $\hat{Y}_I$ under EBLUP imputation is not robust in the sense of validity under both approaches.

Table 5.10: Percent relative efficiency (RE) for $r_o/m_o = 0.4$

ρ_m	m_o	RE
0.05	5	-26
0.1	5	-46
0.3	5	-87
0.5	5	-100
0.05	10	-44
0.1	10	-68
0.3	10	-99
0.5	10	-100
0.05	20	-67
0.1	20	-88
0.3	20	-100
0.5	20	-100

Table 5.11: Percent relative efficiency (RE) for $r_o/m_o = 0.6$

ρ_m	m_o	RE
0.05	5	-17
0.1	5	-31
0.3	5	-65
0.5	5	-82
0.05	10	-30
0.1	10	-48
0.3	10	-80
0.5	10	-91
0.05	20	-47
0.1	20	-66
0.3	20	-89
0.5	20	-95

Table 5.12: Percent relative efficiency (RE) for $r_o/m_o = 0.8$

ρ_m	m_o	RE
0.05	5	-8
0.1	5	-16
0.3	5	-39
0.5	5	-58
0.05	10	-15
0.1	10	-26
0.3	10	-55
0.5	10	-73
0.05	20	-25
0.1	20	-40
0.3	20	-70
0.5	20	-84

Chapter 6

Inference for Bivariate Statistics

6.1 Introduction

As we have seen in Chapter 1, imputation generally leads to unbiased estimators of totals (or means) under uniform response (assumption DB) or an imputation model and the weaker assumption of ignorable response (assumption MB). In Chapter 3, we considered the problem of estimating the mean of a domain, d

$$\bar{Y}_d = \frac{\sum_{i \in U} x_i y_i}{\sum_{i \in U} x_i}. \quad (6.1)$$

We assumed that the domain indicator x_i was known for all the units in the sample. We showed that the imputed estimator of the mean $\bar{Y}_d$ was biased under assumption DB, which motivated us to consider a bias-adjusted estimator.

In this chapter, we consider the general case for which both items x and y may be missing. Two additional parameters of practical interest are the census regression

coefficient given by

$$B = \frac{\sum_{i \in U} x_i y_i}{\sum_{i \in U} x_i^2}, \quad (6.2)$$

and the finite population coefficient of correlation between items x and y given by

$$\rho_{xy} = \frac{1}{K-1} \left[\frac{\sum_{i \in U} x_i y_i - N \bar{X} \bar{Y}}{S_x S_y} \right], \quad (6.3)$$

where K denotes the number of elements in the population and S_x and S_y denote the finite population standard deviation for item x and item y , respectively. For now, consider the case of the correlation coefficient (6.3). Obtaining an imputed estimator that is approximately unbiased for (6.3) depends on the possibility of obtaining an approximately unbiased estimator for each component, $\sum_{i \in U} x_i y_i$, $\bar{X}$, $\bar{Y}$, S_x and S_y separately, which is not difficult for $\bar{X}$, $\bar{Y}$, S_x and S_y (provided we use a stochastic imputation method for S_x and S_y as we have shown in Chapter 1, see section 1.9). The difficulty lies in trying to obtain an approximately unbiased estimator for the cross-product term, $\sum_{i \in U} x_i y_i$. This component can be seen as a measure of the relationship between items x and y . However, imputation generally distorts the relationship between variables, which explains the difficulty. Two strategies may be used to obtain an approximately unbiased estimator for the cross-product term, $\sum_{i \in U} x_i y_i$: (i) Use a tailor-made imputation method and use a simple imputed estimator (e.g., Shao and Wang, 2002). (ii) Use a simple imputation method (such as random hot-deck imputation) and use a tailor-made estimator (e.g., Skinner and Rao, 2002). Skinner and Rao (2002) proposed a bias-adjusted estimator under simple random sampling without replacement and assumption DB in the context of common donor imputation.

In this chapter, we extend the results obtained by Skinner and Rao to the case of stratified multistage sampling and under design-based assumption DB or model-based assumption MB, in the context of several versions of random hot-deck imputation. We set out our basic framework and assumptions in section 6.2. In section 6.3 we present three versions of random hot-deck imputation. We study the bias of a bivariate parameter and obtain a bias-adjusted estimator that is robust in the sense of approximate unbiasedness under either assumption DB or assumption MB in section 6.5. In section 6.6, we conduct a simulation study to compare the performance of the estimators considered in sections 6.4 and 6.5 under three versions of random hot-deck imputation. In section 6.7, we derive a variance estimator of the bias-adjusted estimator, using the jackknife method under the reverse approach of Fay (1991). In section 6.8, we show that the imputation procedure proposed by Shao and Wang (2002) is valid under assumption DB. Their procedure is thus robust in the sense of validity under both approaches.

6.2 Framework and Assumptions

Let U be a finite population of size K which may be unknown. The objective is to estimate a parameter of the form

$$\theta = \frac{1}{K} \sum_{i \in U} g(x_i, y_i),$$

where $g(\cdot, \cdot)$ is a given function, or some smooth function of θ , say $f(\theta)$. We assume that the function $g(\cdot, \cdot)$ is separable, i.e., $g(x, y)$ can be expressed as the product $g_1(x) \times g_2(y)$, for some smooth functions $g_1(\cdot)$ and $g_2(\cdot)$. Note that the cross-

product term $\sum_{i \in U} x_i y_i$ is a particular case with $g_1(x) = x$ and $g_2(y) = y$. Suppose a probability sample, s , of size n is selected according to a stratified multistage design $p(s)$ from U . That is, the population under consideration is stratified into L strata with N_h primary sampling units (PSU's) or clusters in the h^{th} stratum. Within each stratum, $n_h \geq 2$ clusters are selected from stratum h , independently across strata. The first-stage clusters are usually selected without replacement to avoid the selection of the same cluster more than once. Within the $(hi)^{th}$ sampled first-stage cluster, m_{hi} ultimate units (elements) are sampled according to some probability sampling method, $i = 1, \dots, n_h$; $h = 1, \dots, L$. Note that we do not need to specify the number of stages or the sampling methods beyond the first stage. We simply assume that subsampling within sampled clusters is performed to ensure unbiased estimation of cluster totals, Y_{hi} , $i = 1, \dots, n_h$; $h = 1, \dots, L$ under complete response. Associated with the j^{th} sampled element in the i^{th} sampled cluster belonging to the h^{th} stratum is the variable of interest, y_{hij} , and the basic survey weights, w_{hij} . In the case of full response, an approximately unbiased estimator of θ is given by

$$\hat{\theta} = \frac{1}{\hat{K}} \sum_{(hij) \in s} w_{hij} g(x_{hij}, y_{hij}) \quad (6.4)$$

where s denotes the sample of elements, and $\hat{K} = \sum_{(hij) \in s} w_{hij}$.

We now consider the case where items x and y may be missing. Let s_{rr} be the set of units that responded to both items x and y , s_{rm} be the set of units that responded to item x but not to item y , s_{mr} be the set of units that responded to item y but not to item x and s_{mm} be the set of units that did not respond either to items x or y . Also, let $s_{r_x} = s_{rr} \cup s_{rm}$ be the set of units that responded to

item x and $s_{r_y} = s_{rr} \cup s_{mr}$ be the set of units that responded to item y . Further, let $p_{rr} = P[(hij) \in s_{rr}]$, $p_{rm} = P[(hij) \in s_{rm}]$, $p_{mr} = P[(hij) \in s_{mr}]$ and $p_{mm} = P[(hij) \in s_{mm}]$. Also, let $p_x = P[(hij) \in s_{r_x}] = p_{rr} + p_{rm}$ and $p_y = P[(hij) \in s_{r_y}] = p_{rr} + p_{mr}$. Note that $p_{rr} + p_{rm} + p_{mr} + p_{mm} = 1$.

An imputed estimator of θ is given by

$$\hat{\theta}_I = \frac{1}{\hat{K}} \sum_{(hij) \in s} w_{hij} g(\tilde{x}_{hij}, \tilde{y}_{hij}), \quad (6.5)$$

where

$$\tilde{x}_{hij} = \begin{cases} x_{hij} & \text{if } (hij) \in s_{r_x} \\ x_{hij}^* & \text{if } (hij) \in s_{m_x} \end{cases}$$

$$\tilde{y}_{hij} = \begin{cases} y_{hij} & \text{if } (hij) \in s_{r_y} \\ y_{hij}^* & \text{if } (hij) \in s_{m_y}, \end{cases}$$

and x_{hij}^* and y_{hij}^* denote the imputed values for missing x_{hij} and y_{hij} respectively.

Note that the imputed estimator (6.5) can be alternatively written as

$$\hat{\theta}_I = \frac{1}{\hat{K}} \left[\sum_{(hij) \in s} w_{hij} a_{hij} b_{hij} g(x_{hij}, y_{hij}) + \sum_{(hij) \in s} w_{hij} a_{hij} (1 - b_{hij}) g(x_{hij}, y_{hij}^*) + \sum_{(hij) \in s} w_{hij} (1 - a_{hij}) b_{hij} g(x_{hij}^*, y_{hij}) + \sum_{(hij) \in s} w_{hij} (1 - a_{hij}) (1 - b_{hij}) g(x_{hij}^*, y_{hij}^*) \right] \quad (6.6)$$

where

$$a_{hij} = \begin{cases} 1 & \text{if } (hij) \in s_{r_x} \\ 0 & \text{if } (hij) \in s_{m_x} \end{cases}$$

and

$$b_{hij} = \begin{cases} 1 & \text{if } (hij) \in s_{r_y} \\ 0 & \text{if } (hij) \in s_{m_y} \end{cases}$$

We consider two different approaches for inference: (i) design-based and (ii) model-based. Under the design-based approach, we assume a uniform response mechanism within cells so that the following assumption holds:

Assumption DB: Within an imputation cell, the probability of response to item x is p_x , the probability of response to item y is p_y and the probability of response to both items x and y is p_{rr} . Also, the response statuses for different units are independent. Note that, in general, $p_{rr} \neq p_x p_y$.

Under the model-based approach, the assumption is given by:

Assumption MB: Within an imputation cell the response mechanism is ignorable.

The imputation model used is given by

$$\begin{aligned} y_{hik} &= \beta + \epsilon_{hij}, E_m(\epsilon_{hij}) = 0, V_m(\epsilon_{hij}) = \sigma_\epsilon^2, Cov_m(\epsilon_{hij}, \epsilon_{glk}) = 0 \text{ if } (hij) \neq (glk), \\ x_{hik} &= \gamma + \eta_{hij}, E_m(\eta_{hij}) = 0, V_m(\eta_{hij}) = \sigma_\eta^2, Cov_m(\eta_{hij}, \eta_{glk}) = 0 \text{ if } (hij) \neq (glk), \\ \sigma_{\epsilon\eta} &= Cov_m(\epsilon_{hij}, \eta_{hij}), \end{aligned} \tag{6.7}$$

where β , γ , σ_ϵ^2 , σ_η^2 and $\sigma_{\epsilon\eta}$ are unknown parameters.

The response mechanism under assumption MB is weaker than the response mechanism under assumption DB, but the imputation model (6.7) is assumed to hold.

6.3 Random hot-deck procedures

In this section, we describe the following three random hot-deck procedures: common donor imputation (CDI), marginal imputation (MI) and a hybrid imputation (HI).

6.3.1 Common Donor Random Hot-Deck Imputation

The CDI scheme may be described as follows:

For $(hij) \in s_{mr}$, missing x_{hij} is imputed by $x_{hij}^* = x_{glk}$, $(glk) \in s_{rr}$, such that

$$P(x_{hij}^* = x_{glk}) = \frac{w_{glk}}{\sum_{(glk) \in s} w_{glk} a_{glk} b_{glk}}, \quad (glk) \in s_{rr};$$

For $(hij) \in s_{rm}$, missing y_{hij} is imputed by $y_{hij}^* = y_{glk}$, $(glk) \in s_{rr}$ such that

$$P(y_{hij}^* = y_{glk}) = \frac{w_{glk}}{\sum_{(glk) \in s} w_{glk} a_{glk} b_{glk}}, \quad (glk) \in s_{rr};$$

For $(hij) \in s_{mm}$, missing (x_{hij}, y_{hij}) is imputed by $(x_{hij}^*, y_{hij}^*) = (x_{glk}, y_{glk})$, $(glk) \in s_{rr}$ such that

$$P[(x_{hij}^*, y_{hij}^*) = (x_{glk}, y_{glk})] = \frac{w_{glk}}{\sum_{(glk) \in s} w_{glk} a_{glk} b_{glk}}, \quad (glk) \in s_{rr}.$$

6.3.2 Marginal Random Hot-Deck Imputation

The MI scheme may be described as follows:

For $(hij) \in s_{mr}$, missing x_{hij} is imputed by $x_{hij}^* = x_{glk}$, $(glk) \in s_{rx}$ such that

$$P(x_{hij}^* = x_{glk}) = \frac{w_{glk}}{\sum_{(glk) \in s_{rx}} w_{glk} a_{glk}}, \quad (glk) \in s_{rx};$$

For $(hij) \in s_{rm}$, missing y_{hij} is imputed by $y_{hij}^* = y_{glk}$, $(glk) \in s_{ry}$ such that

$$P(y_{hij}^* = y_{glk}) = \frac{w_{glk}}{\sum_{(glk) \in s_{ry}} w_{glk} b_{glk}}, \quad (glk) \in s_{ry};$$

For $(hij) \in s_{mm}$, missing (x_{hij}, y_{hij}) is imputed by $(x_{hij}^*, y_{hij}^*) = (x_{glk}, y_{efg})$, $(glk) \in s_{rx}$ and $(efg) \in s_{ry}$ such that

$$P[(x_{hij}^*, y_{efg}^*) = (x_{glk}, y_{glk})] = \frac{w_{glk}}{\sum_{(glk) \in s_{rx}} w_{glk} a_{glk}} \frac{w_{efg}}{\sum_{(efg) \in s_{ry}} w_{efg} b_{efg}}, \quad (glk) \in s_{rx}$$

and $(efg) \in s_{ry}$.

6.3.3 Hybrid Random Hot-Deck Imputation

The HI scheme may be described as follows:

For $(hij) \in s_{mr}$, missing x_{hij} is imputed by $x_{hij}^* = x_{glk}$, $(glk) \in s_{rx}$ such that

$$P(x_{hij}^* = x_{glk}) = \frac{w_{glk}}{\sum_{(glk) \in s_{rx}} w_{glk} a_{glk}}, \quad (glk) \in s_{rx};$$

For $(hij) \in s_{rm}$, missing y_{hij} is imputed by $y_{hij}^* = y_{glk}$, $(glk) \in s_{ry}$ such that

$$P(y_{hij}^* = y_{glk}) = \frac{w_{glk}}{\sum_{(glk) \in s_{ry}} w_{glk} b_{glk}}, \quad (glk) \in s_{ry};$$

For $(hij) \in s_{mm}$, missing (x_{hij}, y_{hij}) is imputed by $(x_{hij}^*, y_{hij}^*) = (x_{glk}, y_{glk})$, $(glk) \in s_{rr}$ such that

$$P[(x_{hij}^*, y_{hij}^*) = (x_{glk}, y_{glk})] = \frac{w_{glk}}{\sum_{(glk) \in s} w_{glk} a_{glk} b_{glk}}, \quad (glk) \in s_{rr}.$$

Note that HI may be seen as an hybrid method between CDI and HI.

6.4 Bias of the imputed estimator

6.4.1 Assumption DB

We show in Appendix 6.9.1 that, under CDI, MI and HI, the imputed estimator $\hat{\theta}$ in (6.5) is biased under assumption DB and that the bias is given by

$$\text{Bias}(\hat{\theta}_I) = E_p E_r E_* (\hat{\theta}_I - \theta) \approx (1 - c) \left[\frac{T}{K^2} - \theta \right], \quad (6.8)$$

where

$$c = \begin{cases} p_{rr} + p_{mm} & \text{for CDI and HI} \\ p_{rr} & \text{for MI} \end{cases}$$

and

$$T = \sum_{(hij) \in U} \sum_{\substack{(glk) \in U \\ (hij) \neq (glk)}} g(x_{hij}, y_{glk}).$$

Note that under CDI and HI, the bias of $\hat{\theta}_I$ vanishes if $p_{rm} = p_{mr} = 0$. In this case, common donor and hybrid imputation preserve the relationships between items x and y . Also, note that the bias does not vanish under MI, even if $p_{rm} = p_{mr} = 0$. Hence, MI always distorts the relationship between items x and y , unlike CDI or HI.

Denote the imputed estimator $\hat{\theta}_I$ under CDI or HI as $\hat{\theta}_{I(CDI-HI)}$ and the imputed estimator $\hat{\theta}_I$ under MI as $\hat{\theta}_{I(MI)}$. From (6.8), we obtain

$$\left| \frac{\text{Bias} [\hat{\theta}_{I(CDI-HI)}]}{\text{Bias} [\hat{\theta}_{I(MI)}]} \right| = 1 - \frac{p_{mm}}{1 - p_{rr}} \leq 1, \quad (6.9)$$

Expression (6.9) shows that the bias occurring under MI is always greater or equal to the bias occurring under CDI or HI. Hence, MI distorts the relationship between x and y to a greater extent than CDI or HI. Finally, note that in the particular case $g(x, y) = xy$, expression (6.8) reduces to

$$\text{Bias} (\hat{\theta}_I) \approx - (1 - c) S_x S_y \rho_{xy}. \quad (6.10)$$

It is clear from (6.10) that the bias of $\hat{\theta}_I$ is equal to 0 if and only if $\rho_{xy} = 0$, provided $c \neq 1, S_x > 0$ and $S_y > 0$. Also, note that the bias of $\hat{\theta}_I$ is always negative if $\rho_{xy} > 0$.

6.4.2 Assumption MB

Under the assumption MB approach we consider the particular case $g(x, y) = xy$.

We show in Appendix 6.9.2 that the bias of the imputed estimator $\hat{\theta}_I$ is given by

$$\text{Bias} [\hat{\theta}_I] = E_m E_p E_r E_* (\hat{\theta}_I - \theta) \approx -\frac{\sigma_{\epsilon\eta}}{K} \sum_{(hij) \in U} E_r (c_{hij}), \quad (6.11)$$

where

$$c_{hij} = \begin{cases} a_{hij}(1 - b_{hij}) + (1 - a_{hij})b_{hij} & \text{for CDI and HI} \\ 1 - a_{hij}b_{hij} & \text{for MI} \end{cases}$$

Note that

$$\frac{1}{K} \sum_{(hij) \in U} E_r(c_{hij}) = \begin{cases} 1 - \left(\frac{N_{rr} + N_{mm}}{K}\right) & \text{for CDI and HI} \\ 1 - \frac{N_{rr}}{K} & \text{for MI} \end{cases}$$

where $N_{rr} = E_r \left(\sum_{(hij) \in U} a_{hij}b_{hij} \right)$ and $N_{mm} = E_r \left[\sum_{(hij) \in U} (1 - a_{hij})(1 - b_{hij}) \right]$.

From (6.11), it is clear that the bias of $\hat{\theta}_I$ is equal to zero if $\sigma_{\epsilon\eta} = 0$ or if $(N_{rr} + N_{mm})/K = 1$ under CDI or HI and $N_{rr}/K = 1$ under MI. Once again, note that the bias is always negative provided $\sigma_{\epsilon\eta} > 0$, which is frequently the case in practice.

6.5 Bias-Adjusted Estimators

Following Skinner and Rao (2002), we obtain bias-adjusted estimators under CDI, HI

and MI. Let $\hat{p}_{rr} = \frac{\sum_{(hij) \in s} w_{hij} a_{hij} b_{hij}}{\sum_{(hij) \in s} w_{hij}}$, $\hat{p}_{rm} = \frac{\sum_{(hij) \in s} w_{hij} a_{hij} (1 - b_{hij})}{\sum_{(hij) \in s} w_{hij}}$, $\hat{p}_{mr} = \frac{\sum_{(hij) \in s} w_{hij} (1 - a_{hij}) b_{hij}}{\sum_{(hij) \in s} w_{hij}}$ and $\hat{p}_{mm} = \frac{\sum_{(hij) \in s} w_{hij} (1 - a_{hij}) (1 - b_{hij})}{\sum_{(hij) \in s} w_{hij}}$ be estimators of p_{rr} , p_{rm} , p_{mr} and p_{mm} , respectively.

Under CDI, HI and MI, a bias-adjusted estimator of θ , denoted by $\hat{\theta}_I^a$, is given by

$$\hat{\theta}_I^a = \hat{c}^{-1} \hat{\theta}_I + (1 - \hat{c}^{-1}) \frac{\hat{T}_I}{\hat{K}^2}, \quad (6.12)$$

where

$$\hat{c} = \begin{cases} \hat{p}_{rr} + \hat{p}_{mm} & \text{for CDI and HI} \\ \hat{p}_{rr} & \text{for MI} \end{cases}$$

and

$$\hat{T}_I = \sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (glk)}} w_{hij} w_{glk} g(\tilde{x}_{hij}, \tilde{y}_{glk}). \quad (6.13)$$

It is shown in Appendix 6.9.3 and 6.9.4 that $\hat{\theta}_I^a$ is approximately unbiased for θ under approach DB and approach MB, respectively. Since the function $g(x, y)$ is separable, it is not necessary to calculate the $n(n-1)$ values $g(\tilde{x}_{hij}, \tilde{y}_{glk})$ to obtain the bias-adjusted estimator (6.12).

6.6 Simulation study

We performed a simulation study in order to assess the performance of the estimators considered in this chapter under the different versions of random hot-deck imputation (CDI, HI and MI). We generated three populations each of size $K = 5,000$ consisting of two variables x and y such that the coefficient of correlation between x and y , ρ_{xy} , equal to 0.05, 0.5 and 0.9. The goal is to estimate the cross-product moment $\theta = \frac{1}{K} \sum_{i \in U} x_i y_i$. From these populations, we selected $R = 5,000$ simple random samples each of size $n = 200$. In each sample, we generated nonresponse to item x and y such that $p_{rr} = 0.4, p_{rm} = 0.2, p_{mr} = 0.2$ and $p_{mm} = 0.2$. To compensate for nonresponse to items x and y , we used CDI, HI and MI. As a measure of the bias of an estimator $\hat{\theta}$ of θ , we used the percent relative bias

$$B_r(\hat{\theta}) = \frac{\text{Bias}(\hat{\theta})}{\theta} \times 100.$$

As a measure of variability of $\hat{\theta}$, we used the mean square error (MSE)

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2.$$

Simulation results are reported in Tables 6.1 and 6.2. Table 6.1 clearly shows that the unadjusted estimator is negatively biased. As expected, the absolute bias increases as the correlation between item x and y increases. Also, MI leads to a larger relative bias than CDI and HI, for a given ρ_{xy} . The bias-adjusted estimators are approximately unbiased in all the scenarios with absolute relative bias less than 1%. Table 6.2 shows that the MSE of the unadjusted estimators increases as ρ_{xy} increases, which is not surprising since the bias increases with ρ_{xy} . For $\rho_{xy} = 0.05$, the bias is small and in this case MI has the smallest MSE (2.7×10^5). This result can be explained by the fact that MI uses all the possible donors for item x and y , unlike CDI (which only uses respondents in s_{rr} in all the cases) and HI (which only uses respondents in s_{rr} when both x and y are missing). Turning to the bias-adjusted estimators, we note that for $\rho_{xy} = 0.5$ and $\rho_{xy} = 0.9$, MI leads to the smallest MSE, which can be explained, once again, by the fact that MI uses all the possible respondents to item x and y , which leads to more stable estimators.

Table 6.1: Relative bias (%) of the estimators

	$\rho_{xy} = 0.05$	$\rho_{xy} = 0.5$	$\rho_{xy} = 0.9$
CDI (unadjusted)	-2.8	-15.4	-19.3
HI (unadjusted)	-2.5	-14.5	-18.5
MI (unadjusted)	-3.6	-22.7	-28.8
CDI (adjusted)	0.2	0.3	0.1
HI (adjusted)	0.6	0.9	0.9
MI (adjusted)	0.9	0.3	0.4

6.7 Variance estimation

To estimate the variance of the bias-adjusted estimator $\hat{\theta}_I^a$ in (6.12), we use the jackknife method. We do not use Taylor linearization which would be too involved due to the complexity of $\hat{\theta}_I^a$. Skinner and Rao (2002) extended the idea of Rao and Shao's (1992) adjusted jackknife estimator (see section 1.10.4). In this Section, we propose to use the jackknife under the reverse approach of Fay (1991) because of its simplicity. Recall from Chapter 1 (see section 1.10) that, under random imputation, the total variance of

$$V\left(\hat{\theta}_I^a - \theta\right) = E_r V_p E_* \left(\hat{\theta}_I^a - \theta | \mathbf{a}\right) + E_r E_p V_* \left(\hat{\theta}_I^a - \theta | \mathbf{a}\right) + V_r E_p E_* \left(\hat{\theta}_I^a - \theta | \mathbf{a}\right). \quad (6.14)$$

Recall that the third component on the right-hand side of (6.14) is negligible with respect to the sum of the first two components when n/K is negligible, where $n = \sum_{h=1} n_h$ denotes the number of clusters selected at the first stage. In practice, we

Table 6.2: Mean square of the estimators ($\times 10^5$)

	$\rho_{xy} = 0.05$	$\rho_{xy} = 0.5$	$\rho_{xy} = 0.9$
CDI (unadjusted)	3.2	13.8	27.9
HI (unadjusted)	2.9	12.7	26.4
MI (unadjusted)	2.7	16.3	39.7
CDI (adjusted)	5.0	15.2	24.6
HI (adjusted)	4.8	14.5	24.6
MI (adjusted)	6.4	13.5	21.7

often have $n \ll K$ so we omit the derivation of an estimator of $V_{\tau} E_p E_{\star} (\hat{\theta}_I^a - \theta | \mathbf{a})$.

We derive the variance estimator for MI but the cases of CDI and JI can be handled in a similar fashion. The jackknife may be used in a straightforward manner in order to estimate the first component on the right-hand side of (6.14).

First, note that in the case of MI,

$$\tilde{\theta}_I^a = E_{\star} (\hat{\theta}_I^a) = \hat{c}^{-1} E_{\star} (\hat{\theta}_I) + (1 - \hat{c}^{-1}) \frac{E_{\star} (\hat{T}_I)}{\hat{K}^2},$$

where $E_{\star} (\hat{\theta}_I)$ is given by (6.31) and $E_{\star} (\hat{T}_I)$ is obtained by adding (6.43) to (6.46)

and $\hat{c} = \hat{p}_{rr} + \hat{p}_{mm}$.

Let $w_{hij}^{(gl)} = w_{hij} b_{(gl)}$ be the so-called jackknife weights, where

$$b_{(gl)} = \begin{cases} 1 & \text{if } h \neq g \\ \frac{n_g}{n_g - 1} & \text{if } h = g, i \neq l \\ 0 & \text{if } i = l \end{cases}$$

Let $\tilde{\theta}_{I(gl)}^a$ be the bias-adjusted estimator obtained by deleting the $(gl)^{th}$ cluster. Note that $\tilde{\theta}_{I(gl)}^a$ is computed the same way as $\hat{\theta}_I^a$ but by using the jackknife weights $w_{hij}^{(gl)}$ instead of the original weights w_{hij} . An estimator of the component $E_r V_p E_* \left(\hat{\theta}_I^a - \theta | \mathbf{a} \right)$ is given by

$$v_J \left(\tilde{\theta}_I^a \right) = \sum_{g=1}^L \left(\frac{n_g - 1}{n_g} \right) \sum_{i=1}^{n_h} \left(\tilde{\theta}_{I(gl)}^a - \tilde{\theta}_I^a \right)^2. \quad (6.15)$$

Next, we derive an estimator of the component $E_r E_p V_* \left(\hat{\theta}_I^a - \theta | \mathbf{a} \right)$ due to random imputation. For simplicity, we consider the case $g(x, y) = xy$. We have

$$\begin{aligned} V_* \left(\hat{\theta}_I^a \right) &= \hat{c}^{-2} V_* \left(\hat{\theta}_I \right) + (1 - \hat{c}^{-1})^2 \hat{K}^{-4} V_* \left(\hat{T}_I \right) \\ &\quad + 2\hat{c}^{-1} (1 - \hat{c}^{-1}) \frac{1}{\hat{K}^2} Cov_* \left(\hat{\theta}_I, \hat{T}_I \right). \end{aligned} \quad (6.16)$$

First, note that under MI, $E_* (x_i^*) = \sum_{(hij) \in s} w_{hij} a_{hij} x_{hij} / \sum_{(hij) \in s} w_{hij} a_{hij} \equiv \bar{x}_r$

and that $E_* (y_i^*) = \sum_{(hij) \in s} w_{hij} b_{hij} y_{hij} / \sum_{(hij) \in s} w_{hij} b_{hij} \equiv \bar{y}_r$.

Also, we have

$$\begin{aligned} V_* (x_i^*) &= \frac{1}{\sum_{(hij) \in s} w_{hij} a_{hij}} \sum_{(hij) \in s} w_{hij} a_{hij} (x_{hij} - \bar{x}_r)^2 \equiv s_{xr}^2, \\ V_* (y_i^*) &= \frac{1}{\sum_{(hij) \in s} w_{hij} b_{hij}} \sum_{(hij) \in s} w_{hij} b_{hij} (y_{hij} - \bar{y}_r)^2 \equiv s_{yr}^2, \end{aligned}$$

and

$$\begin{aligned} V_* (x_{hij}^* y_{hij}^*) &= E_* (x_{hij}^{*2} y_{hij}^{*2}) - E_* (x_{hij}^*)^2 E_* (y_{hij}^*)^2 \\ &= \left[V_* (x_{hij}^*) + E_* (x_{hij}^*)^2 \right] \left[V_* (y_{hij}^*) + E_* (y_{hij}^*)^2 \right] - E_* (x_{hij}^*)^2 E_* (y_{hij}^*)^2 \\ &= [s_{xr}^2 + \bar{x}_r^2] [s_{yr}^2 + \bar{y}_r^2] - \bar{x}_r^2 \bar{y}_r^2 \equiv C_{xyr}. \end{aligned}$$

Finally, note that under MI, $Cov_*(x_i^*, y_j^*) = 0 \forall i, j$.

Hence,

$$\begin{aligned} V_*(\hat{\theta}_I) &= \hat{K}^{-2} \left\{ \sum_{(hij) \in s} w_{hij}^2 a_{hij} (1 - b_{hij}) x_{hij}^2 s_{yr}^2 + \sum_{(hij) \in s} w_{hij}^2 (1 - a_{hij}) b_{hij} y_{hij}^2 s_{xr}^2 \right. \\ &\quad \left. + \sum_{(hij) \in s} w_{hij}^2 (1 - a_{hij}) (1 - b_{hij}) C_{xyr} \right\}. \end{aligned} \quad (6.17)$$

We now determine $V_*(\hat{T}_I)$. We first write $\hat{T}_I$ as

$$\hat{T}_I = \hat{X}_I \hat{Y}_I - \hat{T}_{Ixy},$$

where $(\hat{X}_I, \hat{Y}_I) = \sum_{(hij) \in s} w_{hij} (\tilde{x}_{hij} \tilde{y}_{hij})$ and $\hat{T}_{Ixy} = \sum_{(hij) \in s} w_{hij}^2 \tilde{x}_{hij} \tilde{y}_{hij}$.

Note that under MI,

$$E_*(\hat{X}_I, \hat{Y}_I) = \hat{K}(\bar{x}_r, \bar{y}_r), \quad (6.18)$$

$$\begin{aligned} V_*(\hat{X}_I) &= \sum_{(hij) \in s} w_{hij}^2 (1 - a_{hij}) s_{xr}^2, \\ V_*(\hat{Y}_I) &= \sum_{(hij) \in s} w_{hij}^2 (1 - b_{hij}) s_{yr}^2. \end{aligned}$$

It follows that

$$\begin{aligned} V_*(\hat{X}_I \hat{Y}_I) &= E_*(\hat{X}_I^2 \hat{Y}_I^2) - E_*(\hat{X}_I)^2 E_*(\hat{Y}_I)^2 \\ &= \left[V_*(\hat{X}_I) + E_*(\hat{X}_I)^2 \right] \left[V_*(\hat{Y}_I) + E_*(\hat{Y}_I)^2 \right] - E_*(\hat{X}_I)^2 E_*(\hat{Y}_I)^2 \\ &= \left[\sum_{(hij) \in s} w_{hij}^2 (1 - a_{hij}) s_{xr}^2 + \hat{K}^2 \bar{x}_r^2 \right] \left[\sum_{(hij) \in s} w_{hij}^2 (1 - b_{hij}) s_{yr}^2 + \hat{K}^2 \bar{y}_r^2 \right] - \hat{K}^2 \bar{x}_r^2 \bar{y}_r^2 \end{aligned}$$

and

$$\begin{aligned} V_*(\hat{T}_{Ixy}) &= \sum_{(hij) \in s} w_{hij}^4 a_{hij} (1 - b_{hij}) s_{yr}^2 \\ &\quad + \sum_{(hij) \in s} w_{hij}^4 (1 - a_{hij}) b_{hij} s_{xr}^2 + \sum_{(hij) \in s} w_{hij}^4 (1 - a_{hij}) (1 - b_{hij}) C_{xyr}. \end{aligned}$$

Finally, note that

$$Cov_* (\hat{X}_I \hat{Y}_I, \hat{T}_{Ixy}) = E_* (\hat{X}_I \hat{Y}_I \hat{T}_{Ixy}) - E_* (\hat{X}_I \hat{Y}_I) E_* (\hat{T}_{Ixy}). \quad (6.19)$$

After tedious but somewhat straightforward algebra, we obtain

$$\begin{aligned} E_* (\hat{X}_I \hat{Y}_I \hat{T}_{Ixy}) &= \hat{K}^2 \bar{x}_r \bar{y}_r \sum_{(hij) \in s} w_{hij}^2 a_{hij} b_{hij} x_{hij} y_{hij} \\ &+ \hat{K} \bar{y}_r \left[\sum_{(hij) \in s} w_{hij}^3 (1 - a_{hij}) b_{hij} y_{hij} s_{xr}^2 + \hat{K} \bar{x}_r^2 \sum_{(hij) \in s} w_{hij}^2 (1 - a_{hij}) b_{hij} y_{hij} \right] \\ &+ \hat{K} \bar{x}_r \left[\sum_{(hij) \in s} w_{hij}^3 (1 - b_{hij}) a_{hij} x_{hij} s_{yr}^2 + \hat{K} \bar{y}_r^2 \sum_{(hij) \in s} w_{hij}^2 a_{hij} (1 - b_{hij}) \right] \\ &+ \hat{K}^2 \bar{x}_r^2 \bar{y}_r^2 \sum_{(hij) \in s} w_{hij}^2 (1 - a_{hij}) (1 - b_{hij}) \\ &+ \hat{K} \bar{y}_r \left[\sum_{(hij) \in s} w_{hij}^3 (1 - a_{hij}) (1 - b_{hij}) (s_{xr}^2 + \bar{x}_r^2) \bar{y}_r \right. \\ &\left. + \bar{x}_r^2 \bar{y}_r \sum_{(hij) \in s} w_{hij} (1 - a_{hij}) \sum_{(hij) \in s} w_{hij} (1 - a_{hij}) b_{hij} \right] \\ &+ \hat{K} \bar{x}_r \left[\sum_{(hij) \in s} w_{hij}^3 (1 - a_{hij}) (1 - b_{hij}) (s_{yr}^2 + \bar{y}_r^2) \bar{x}_r \right. \\ &\left. + \bar{x}_r \bar{y}_r^2 \sum_{(hij) \in s} w_{hij} (1 - b_{hij}) \sum_{(hij) \in s} w_{hij} a_{hij} (1 - b_{hij}) \right] \\ &+ \sum_{(hij) \in s} w_{hij}^4 (1 - a_{hij}) (1 - b_{hij}) C_{xr} \\ &+ \sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (glk)}} w_{hij}^3 w_{glk} (1 - a_{hij}) (1 - b_{hij}) (1 - b_{glk}) (s_{xr}^2 + \bar{x}_r^2) \bar{y}_r^2 \\ &+ \sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (glk)}} w_{hij}^3 w_{glk} (1 - a_{hij}) (1 - b_{hij}) (1 - a_{glk}) (s_{yr}^2 + \bar{y}_r^2) \bar{x}_r^2. \quad (6.20) \end{aligned}$$

Also, we have

$$\begin{aligned}
 E_* \left(\hat{T}_{Ixy} \right) &= \sum_s w_{hij}^2 a_{hij} b_{hij} x_{hij} y_{hij} \\
 &+ \sum_{(hij) \in s} w_{hij}^2 (1 - a_{hij}) b_{hij} y_{hij} \bar{x}_r + \sum_{(hij) \in s} w_{hij}^2 (1 - b_{hij}) a_{hij} x_{hij} \bar{y}_r \\
 &+ \sum_{(hij) \in s} w_{hij}^2 (1 - a_{hij}) (1 - b_{hij}) \bar{x}_r \bar{y}_r.
 \end{aligned} \tag{6.21}$$

Combining (6.18), (6.19), (6.20) and (6.21) we obtain $Cov_* \left(\hat{X}_I \hat{Y}_I, \hat{T}_{Ixy} \right)$.

Hence,

$$\begin{aligned}
 V_* \left(\hat{T}_I \right) &= \left[V_* \left(\hat{X}_I \right) + E_* \left(\hat{X}_I \right)^2 \right] \left[V_* \left(\hat{Y}_I \right) + E_* \left(\hat{Y}_I \right)^2 \right] \\
 &+ V_* \left(\hat{T}_{Ixy} \right) - 2Cov_* \left(\hat{X}_I \hat{Y}_I, \hat{T}_{Ixy} \right).
 \end{aligned} \tag{6.22}$$

Next, we determine $Cov_* \left(\hat{\theta}_I, \hat{T}_I \right)$. We have

$$\begin{aligned}
 Cov_* \left(\hat{\theta}_I, \hat{T}_I \right) &= Cov_* \left(\hat{\theta}_I, \hat{X}_I \hat{Y}_I - \hat{T}_{xyI} \right) \\
 &= Cov_* \left(\hat{\theta}_I, \hat{X}_I \hat{Y}_I \right) - Cov_* \left(\hat{\theta}_I, \hat{T}_{xyI} \right).
 \end{aligned} \tag{6.23}$$

Note that

$$\begin{aligned}
 Cov_* \left(\hat{\theta}_I, \hat{T}_{xyI} \right) &= \frac{1}{\hat{K}} \left[\sum_{(hij) \in s} w_{hij}^3 (1 - a_{hij}) b_{hij} y_{hij}^2 s_{xr}^2 \right. \\
 &+ \sum_{(hij) \in s} w_{hij}^3 a_{hij} (1 - b_{hij}) x_{hij}^2 s_{yr}^2 \\
 &\left. + \sum_{(hij) \in s} w_{hij}^3 (1 - a_{hij}) (1 - b_{hij}) C_{xyr} \right].
 \end{aligned} \tag{6.24}$$

Also, note that

$$Cov_* \left(\hat{\theta}_I, \hat{X}_I \hat{Y}_I \right) = E_* \left(\hat{\theta}_I \hat{X}_I \hat{Y}_I \right) - E_* \left(\hat{\theta}_I \right) E_* \left(\hat{X}_I \hat{Y}_I \right). \tag{6.25}$$

After tedious but somewhat straightforward algebra, we obtain

$$\begin{aligned}
E_* \left(\hat{\theta}_I \hat{X}_I \hat{Y}_I \right) &= \hat{K}^2 \bar{x}_r \bar{y}_r \sum_{(hij) \in s} w_{hij} a_{hij} b_{hij} x_{hij} y_{hij} \\
&+ \hat{K} \bar{y}_r \left[\sum_{(hij) \in s} w_{hij}^2 (1 - a_{hij}) b_{hij} y_{hij} s_{xr}^2 + \hat{K} \bar{x}_r^2 \sum_{(hij) \in s} w_{hij} (1 - a_{hij}) b_{hij} y_{hij} \right] \\
&+ \hat{K} \bar{x}_r \left[\sum_{(hij) \in s} w_{hij}^2 (1 - b_{hij}) a_{hij} x_{hij} s_{yr}^2 + \hat{K} \bar{y}_r^2 \sum_{(hij) \in s} w_{hij} a_{hij} (1 - b_{hij}) \right] \\
&+ \hat{K}^2 \bar{x}_r^2 \bar{y}_r^2 \sum_{(hij) \in s} w_{hij} (1 - a_{hij}) (1 - b_{hij}) \\
&+ \hat{K} \bar{y}_r \left[\sum_{(hij) \in s} w_{hij}^2 (1 - a_{hij}) (1 - b_{hij}) (s_{xr}^2 + \bar{x}_r^2) \bar{y}_r \right. \\
&+ \left. \bar{x}_r^2 \bar{y}_r \sum_{(hij) \in s} w_{hij} (1 - a_{hij}) \sum_{(hij) \in s} w_{hij} (1 - a_{hij}) b_{hij} \right] \\
&+ \hat{K} \bar{x}_r \left[\sum_{(hij) \in s} w_{hij}^2 (1 - a_{hij}) (1 - b_{hij}) (s_{yr}^2 + \bar{y}_r^2) \bar{x}_r \right. \\
&+ \left. \bar{x}_r \bar{y}_r^2 \sum_{(hij) \in s} w_{hij} (1 - b_{hij}) \sum_{(hij) \in s} w_{hij} a_{hij} (1 - b_{hij}) \right] \\
&+ \sum_{(hij) \in s} w_{hij}^2 (1 - a_{hij}) (1 - b_{hij}) C_{xr} \\
&+ \sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (glk)}} w_{hij}^2 w_{glk} (1 - a_{hij}) (1 - b_{hij}) (1 - b_{glk}) (s_{xr}^2 + \bar{x}_r^2) \bar{y}_r^2 \\
&+ \sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (glk)}} w_{hij}^2 w_{glk} (1 - a_{hij}) (1 - b_{hij}) (1 - a_{glk}) (s_{yr}^2 + \bar{y}_r^2) \bar{y}_r^2. \quad (6.26)
\end{aligned}$$

Combining (6.18), (6.31) and (6.26) in (6.25), we obtain $Cov_* \left(\hat{\theta}_I, \hat{X}_I \hat{Y}_I \right)$. Finally, combining (6.17), (6.22) and (6.25) in (6.16), we obtain $V_* \left(\hat{\theta}_I^a \right)$.

6.8 Validity of the Shao-Wang procedure under assumption DB

Shao and Wang (2002) proposed an imputation method that preserves unbiasedness for marginal totals, second moments and cross-product moments. Hence, their procedure also preserves coefficients of correlation. They demonstrated the validity of their procedure under assumption MB, but they did not study its properties under assumption DB. In this section, we show that the Shao-Wang procedure preserves marginal totals, second moment and coefficients of correlation under assumption DB as well. Hence, their imputation procedure is robust in the sense of validity under both approaches. We start by describing their imputation procedure in the case of random hot-deck imputation. Imputation is performed according to the following scheme:

- (i) For $a_{hij} = 0$ and $b_{hij} = 1$ (x_{hij} missing, y_{hij} observed), we use the imputed values

$$x_{hij}^* = \hat{\gamma}_r + \eta_{hij}^*,$$

where

$$\eta_{hij}^* = \frac{\hat{\sigma}_{\epsilon\eta}}{\hat{\sigma}_{\epsilon}^2} (y_{hij} - \hat{B}_r) + \tilde{\eta}_{hij}^*,$$

where $\hat{B}_r = \sum_{(hij) \in s} w_{hij} b_{hij} y_{hij} / \sum_{(hij) \in s} w_{hij} b_{hij} = \bar{y}_r$ and, given the observed data, the $\tilde{\eta}_{hij}^*$'s are independent random variables with mean 0 and variance

$\hat{\sigma}_\eta^2 - \hat{\sigma}_{\epsilon\eta}^2 / \hat{\sigma}_\epsilon^2$ with

$$\hat{\sigma}_\epsilon^2 = \frac{1}{\sum_{(hij) \in s} w_{hij} b_{hij}} \sum_{(hij) \in s} w_{hij} b_{hij} (y_{hij} - \hat{B}_r)^2,$$

$$\hat{\sigma}_\eta^2 = \frac{1}{\sum_{(hij) \in s} w_{hij} a_{hij}} \sum_{(hij) \in s} w_{hij} a_{hij} (x_{hij} - \hat{\gamma}_r)^2,$$

with $\hat{\gamma}_r = \sum_{(hij) \in s} w_{hij} a_{hij} y_{hij} / \sum_{(hij) \in s} w_{hij} a_{hij} = \bar{x}_r$ and

$$\hat{\sigma}_{\eta\epsilon} = \frac{1}{\sum_{(hij) \in s} w_{hij} a_{hij} b_{hij}} \sum_{(hij) \in s} w_{hij} a_{hij} b_{hij} (x_{hij} - \hat{\gamma}_r) (y_{hij} - \hat{B}_r).$$

(ii) For $a_{hij} = 1$ and $b_{hij} = 0$, we use the imputed values

$$y_{hij}^* = \hat{B}_r + \epsilon_{hij}^*,$$

where

$$\epsilon_{hij}^* = \frac{\hat{\sigma}_{\epsilon\eta}}{\hat{\sigma}_\epsilon^2} (x_{hij} - \hat{\gamma}_r) + \tilde{\epsilon}_{hij}^*,$$

and, given the observed data, the $\tilde{\epsilon}_{hij}^*$'s are independent random variables with mean 0 and variance $\hat{\sigma}_\epsilon^2 - \hat{\sigma}_{\epsilon\eta}^2 / \hat{\sigma}_\eta^2$.

(iii) For $a_{hij} = 0$ and $b_{hij} = 0$, $(\epsilon_{hij}^*, \eta_{hij}^*)$'s are independently distributed with mean 0 and covariance matrix

$$\begin{pmatrix} \hat{\sigma}_\epsilon^2, \hat{\sigma}_{\epsilon\eta} \\ \hat{\sigma}_{\epsilon\eta}, \hat{\sigma}_\eta^2 \end{pmatrix}$$

Under assumption DB, note that

$$E_r (\hat{\sigma}_\epsilon^2) \approx \frac{1}{\hat{K}} \sum_{(hij) \in s} w_{hij} (y_{hij} - \hat{B})^2 \equiv \tilde{\sigma}_\epsilon^2,$$

where $\hat{B} = \bar{y} = \sum_{(hij) \in s} w_{hij} y_{hij} / \sum_{(hij) \in s} w_{hij}$.

Also, note that

$$E_r(\hat{\sigma}_\eta^2) \approx \frac{1}{\hat{K}} \sum_{(hij) \in s} w_{hij} (x_{hij} - \hat{\gamma})^2 \equiv \tilde{\sigma}_\eta^2,$$

where $\hat{\gamma} = \bar{x} = \sum_{(hij) \in s} w_{hij} x_{hij} / \sum_{(hij) \in s} w_{hij}$.

Finally, note that

$$E_r(\hat{\sigma}_{\epsilon\eta}^2) \approx \frac{1}{\hat{K}} \sum_{(hij) \in s} w_{hij} (x_{hij} - \hat{\gamma}) (x_{hij} - \hat{\beta}) \equiv \tilde{\sigma}_{\epsilon\eta}^2.$$

We first show that the imputed estimator, $\hat{Y}_I$, of the total Y , is approximately unbiased under the Shao-Wang procedure and assumption DB. We have

$$\hat{Y}_I = \sum_{(hij) \in s} w_{hij} b_{hij} y_{hij} + \sum_{(hij) \in s} w_{hij} (1 - b_{hij}) \left[\hat{B}_r + \frac{\hat{\sigma}_{\epsilon\eta}}{\hat{\sigma}_\eta^2} (x_{hij} - \hat{\gamma}_r) + \tilde{\epsilon}_{hij}^* \right].$$

Noting that $E_*(\tilde{\epsilon}_{hij}^*) = 0$, we obtain

$$E_*(\hat{Y}_I) = \sum_{(hij) \in s} w_{hij} b_{hij} y_{hij} + \sum_{(hij) \in s} w_{hij} (1 - b_{hij}) \left[\hat{B}_r + \frac{\hat{\sigma}_{\epsilon\eta}}{\hat{\sigma}_\eta^2} (x_{hij} - \hat{\gamma}_r) \right].$$

Now,

$$E_r E_*(\hat{Y}_I) \approx p_y \hat{Y} + (1 - p_y) \left[\hat{Y} + \frac{\tilde{\sigma}_{\epsilon\eta}}{\tilde{\sigma}_\eta^2} (\hat{X} - \hat{X}) \right] = p_y \hat{Y} + (1 - p_y) \hat{Y} = \hat{Y},$$

where $(\hat{X}, \hat{Y}) = \sum_{i \in s} w_{hij} (x_{hij}, y_{hij})$. It follows that $E(\hat{Y}_I) = E_p E_r E_*(\hat{Y}_I) \approx Y$.

Secondly, we show that the imputed estimator of the second moment $Y^{(2)} = \sum_{(hij) \in U} y_{hij}^2$, denoted as $\hat{Y}_I^{(2)}$, is approximately unbiased under the Shao-Wang procedure and assumption DB.

We have

$$\hat{Y}_I^{(2)} = \sum_{(hij) \in s} w_{hij} b_{hij} y_{hij}^2 + \sum_{(hij) \in s} w_{hij} (1 - b_{hij}) \left[\hat{B}_r + \epsilon_{hij}^* \right]^2.$$

Note that

$$E_{\star}(\epsilon_{hij}^{*2}) = \frac{\hat{\sigma}_{\epsilon\eta}^2}{\hat{\sigma}_{\eta}^4} (x_{hij} - \hat{\gamma}_r)^2 + \hat{\sigma}_{\epsilon}^2 - \frac{\hat{\sigma}_{\epsilon\eta}^2}{\hat{\sigma}_{\eta}^2}.$$

Then,

$$\begin{aligned} E_{\star}(\hat{Y}_I^{(2)}) &= \sum_{(hij) \in s} w_{hij} b_{hij} y_{hij}^2 + \sum_{(hij) \in s} w_{hij} (1 - b_{hij}) \left[\hat{B}_r + \sigma_{\epsilon}^2 \right] \\ &\quad + \sum_{(hij) \in s} w_{hij} (1 - b_{hij}) \left[\frac{\hat{\sigma}_{\epsilon\eta}^2}{\hat{\sigma}_{\eta}^4} (x_{hij} - \hat{\gamma}_r)^2 - \frac{\hat{\sigma}_{\epsilon\eta}^2}{\hat{\sigma}_{\eta}^2} \right]. \end{aligned}$$

It follows that

$$\begin{aligned} E_r E_{\star}(\hat{Y}_I^{(2)}) &\approx p_y \sum_{(hij) \in s} w_{hij} y_{hij}^2 + (1 - p_y) \sum_{(hij) \in s} w_{hij} \left\{ V_r(\hat{B}_r) + E_r(\hat{B}_r)^2 \right. \\ &\quad \left. + \frac{\sum_{(hij) \in s} w_{hij} y_{hij}^2}{\sum_{(hij) \in s} w_{hij}} - \left[V_r(\hat{B}_r) + E_r(\hat{B}_r)^2 \right] \right\} \\ &\quad + (1 - p_y) \frac{\hat{\sigma}_{\epsilon\eta}^2}{\hat{\sigma}_{\eta}^2} \left[\frac{1}{\hat{\sigma}_{\eta}^2} \sum_{(hij) \in s} w_{hij} (x_{hij} - \hat{\gamma})^2 - \hat{K} \right] \\ &\approx p_y \sum_{(hij) \in s} w_{hij} y_{hij}^2 + (1 - p_y) \sum_{(hij) \in s} w_{hij} y_{hij}^2 \\ &= \sum_{(hij) \in s} w_{hij} y_{hij}^2, \end{aligned}$$

noting that $\sum_{(hij) \in s} w_{hij} (x_{hij} - \hat{\gamma})^2 / \hat{\sigma}_{\eta}^2 \approx \hat{K}$. Hence, $E(\hat{Y}_I^{(2)}) = Y^{(2)}$ and the Shao-Wang procedure preserves the second moment $Y^{(2)}$.

Next, we show that the imputed estimator (6.5) with $g(x, y) = xy$ is approximately unbiased for $\theta = \frac{1}{K} \sum_{(hij) \in U} x_{hij} y_{hij}$ under Shao-Wang procedure.

Clearly,

$$E_r E_{\star} \left(\sum_{(hij) \in s} w_{hij} a_{hij} b_{hij} x_{hij} y_{hij} \right) = p_{rr} \sum_{(hij) \in s} w_{hij} x_{hij} y_{hij}. \quad (6.27)$$

Now,

$$E_{\star} \left(\sum_{(hij) \in s} w_{hij} a_{hij} (1 - b_{hij}) x_{hij} y_{hij}^* \right) = \sum_{(hij) \in s} w_{hij} a_{hij} (1 - b_{hij}) x_{hij} \frac{\hat{\sigma}_{\epsilon\eta}}{\hat{\sigma}_{\eta}^2} (x_{hij} - \hat{\gamma}_r).$$

It follows that

$$\begin{aligned}
 E_r E_* \left(\sum_{(hij) \in s} w_{hij} a_{hij} (1 - b_{hij}) x_{hij} y_{hij}^* \right) &\approx p_{rm} \hat{K} \hat{\gamma} \hat{B} \\
 &+ p_{rm} E_r \left\{ \frac{\tilde{\sigma}_{\epsilon\eta}}{\tilde{\sigma}_\eta^2} \sum_{(hij) \in s} w_{hij} (x_{hij} - \hat{\gamma}_r) x_{hij} \right\} \\
 &\approx p_{rm} \hat{K} \hat{\gamma} \hat{B} + p_{rm} \hat{K} \tilde{\sigma}_{\epsilon\eta}, \tag{6.28}
 \end{aligned}$$

where $(\hat{\gamma}, \hat{B}) = \hat{K}^{-1} (\hat{X}, \hat{Y})$ and noting that $E_r \left[\sum_{(hij) \in s} w_{hij} (x_{hij} - \tilde{\gamma}_r) x_{hij} \right] \approx \hat{K} \tilde{\sigma}_\eta^2$.

Similarly, we have

$$E_r E_* \left(\sum_{(hij) \in s} w_{hij} (1 - a_{hij}) b_{hij} x_{hij}^* y_{hij} \right) \approx p_{mr} \hat{K} \hat{\gamma} \hat{B} + p_{mr} \hat{K} \tilde{\sigma}_{\epsilon\eta}. \tag{6.29}$$

Finally,

$$E_* \left(\sum_{(hij) \in s} w_{hij} (1 - a_{hij}) (1 - b_{hij}) x_{hij}^* y_{hij}^* \right) = \sum_{(hij) \in s} w_{hij} (1 - a_{hij}) (1 - b_{hij}) \left[\hat{B}_r \hat{\gamma}_r + \tilde{\sigma}_{\epsilon\eta} \right],$$

Noting that $E_* (x_{hij}^* y_{hij}^*) = E_* (\hat{\gamma}_r \hat{B}_r + \hat{B}_r \eta_i^* + \hat{\gamma}_r \epsilon_i^* + \epsilon_i^* \eta_i^*) = \hat{\gamma}_r \hat{B}_r + Cov_*(\epsilon_i^* \eta_i^*)$.

It follows that

$$E_r E_* \left(\sum_{(hij) \in s} w_{hij} (1 - a_{hij}) (1 - b_{hij}) x_{hij}^* y_{hij}^* \right) \approx p_{mm} \hat{K} \hat{\gamma} \hat{B} + p_{mm} \hat{K} \tilde{\sigma}_{\epsilon\eta}. \tag{6.30}$$

Combining (6.27), (6.28), (6.29) and (6.30), we obtain

$$\begin{aligned}
 E_r E_* (\hat{\theta}_I) &\approx \frac{1}{\hat{K}} \left\{ p_{rr} \sum_{(hij) \in s} w_{hij} x_{hij} y_{hij} + p_{rm} \hat{K} \hat{\gamma} \hat{B} + p_{rm} \hat{K} \tilde{\sigma}_{\epsilon\eta} \right. \\
 &\quad \left. + p_{mr} \hat{K} \hat{\gamma} \hat{B} + p_{mr} \hat{K} \tilde{\sigma}_{\epsilon\eta} + p_{mm} \hat{K} \hat{\gamma} \hat{B} + p_{mm} \hat{K} \tilde{\sigma}_{\epsilon\eta} \right\} \\
 &\approx \frac{1}{\hat{K}} \left\{ (p_{rr} + p_{rm} + p_{mr} + p_{mm}) \sum_{(hij) \in s} w_{hij} x_{hij} y_{hij} \right\} \\
 &= \frac{1}{\hat{K}} \sum_{(hij) \in s} w_{hij} x_{hij} y_{hij},
 \end{aligned}$$

noting that $\tilde{\sigma}_{\epsilon\eta} = \frac{1}{\hat{K}} \left[\sum_{hij \in s} w_{hij} x_{hij} y_{hij} - \hat{K} \hat{\gamma} \hat{B} \right]$. It follows that $E (\hat{\theta}_I) \approx \theta$.

6.9 Appendix

6.9.1 Bias of $\hat{\theta}_I$ under assumption DB

We derive the bias expression (6.8) for MI. The proofs for CDI and HI are very similar.

First,

$$\begin{aligned}
 E_{\star}(\hat{\theta}_I) &= \frac{1}{\hat{K}} \left[\sum_{(hij) \in s} w_{hij} a_{hij} b_{hij} g(x_{hij}, y_{hij}) \right. \\
 &\quad + \frac{1}{\sum_{(gk) \in s} w_{gk} b_{gk}} \sum_{(hij) \in s} \sum_{(gk) \in s} w_{hij} w_{gk} a_{hij} (1 - b_{hij}) b_{gk} g(x_{hij}, y_{gk}) \\
 &\quad + \frac{1}{\sum_{(gk) \in s} w_{gk} a_{gk}} \sum_{(hij) \in s} \sum_{(gk) \in s} w_{hij} w_{gk} (1 - a_{hij}) b_{hij} a_{gk} z(x_{gk}, y_{hij}) \\
 &\quad + \frac{1}{\sum_{(gk) \in s} w_{gk} a_{gk} \sum_{(efg) \in s} w_{efg} b_{efg}} \times \\
 &\quad \left. \sum_{(hij) \in s} \sum_{(gk) \in s} \sum_{(efg) \in s} w_{hij} w_{gk} w_{efg} (1 - a_{hij}) (1 - b_{hij}) a_{gk} b_{efg} g(x_{gk}, y_{efg}) \right] \quad (6.31)
 \end{aligned}$$

Hence,

$$\begin{aligned}
 E_r E_{\star}(\hat{\theta}_I) &= \frac{1}{\hat{K}} \left\{ p_{rr} \sum_{(hij) \in s} w_{hij} g(x_{hij}, y_{hij}) \right. \\
 &\quad \left. + \frac{1}{\hat{K}} \left[(p_{rm} + p_{mr} + p_{mm}) \sum_{(hij) \in s} \sum_{\substack{(gk) \in s \\ (hij) \neq (gk)}} w_{hij} w_{gk} g(x_{hij}, y_{gk}) \right] \right\} \\
 &= p_{rr} \hat{\theta} + (1 - p_{rr}) \frac{\hat{T}}{\hat{K}^2}. \quad (6.32)
 \end{aligned}$$

Finally, we have

$$E_p E_r E_{\star}(\hat{\theta}_I) \approx p_{rr} \theta + (1 - p_{rr}) \frac{T}{K^2} \quad (6.33)$$

Hence,

$$Bias(\hat{\theta}_I) = (1 - p_{rr}) \left[\frac{T}{K^2} - \theta \right]. \quad (6.34)$$

6.9.2 Bias of $\hat{\theta}_I$ under assumption MB

We derive the bias expression (6.11) under assumption MB in the case of MI. The proofs for CDI and MI are very similar.

First, note that $E_\star(\hat{\theta}_I)$ is obtained by replacing $g(x, y)$ by xy in (6.33), which leads to

$$\begin{aligned} E_\star(\hat{\theta}_I) = & \frac{1}{\hat{K}} \left[\sum_{(hij) \in s} w_{hij} a_{hij} b_{hij} x_{hij} y_{hij} \right. \\ & + \sum_{(hij) \in s} w_{hij} a_{hij} (1 - b_{hij}) x_{hij} \left(\frac{\sum_{(glk) \in s} w_{glk} b_{glk} y_{glk}}{\sum_{(glk) \in s} w_{glk} b_{glk}} \right) \\ & + \sum_{(hij) \in s} w_{hij} (1 - a_{hij}) b_{hij} y_{hij} \left(\frac{\sum_{(glk) \in s} w_{glk} a_{glk} x_{glk}}{\sum_{(glk) \in s} w_{glk} a_{glk}} \right) \\ & \left. + \sum_{(hij) \in s} w_{hij} (1 - a_{hij}) (1 - b_{hij}) \left(\frac{\sum_{(glk) \in s} w_{glk} a_{glk} x_{glk}}{\sum_{(glk) \in s} w_{glk} a_{glk}} \right) \left(\frac{\sum_{(glk) \in s} w_{glk} b_{glk} y_{glk}}{\sum_{(glk) \in s} w_{glk} b_{glk}} \right) \right] \end{aligned}$$

Now, we determine $E_m E_p E_\star(\hat{\theta}_I)$. First, we have

$$E_m E_p \left(\sum_{(hij) \in s} w_{hij} a_{hij} b_{hij} x_{hij} y_{hij} \right) = \sum_{(hij) \in U} a_{hij} b_{hij} [\sigma_{\epsilon\eta} + \gamma\beta]. \quad (6.35)$$

Secondly, we have

$$\begin{aligned} E_m E_p \left[\sum_{(hij) \in s} w_{hij} a_{hij} (1 - b_{hij}) x_{hij} \left(\frac{\sum_{(hij) \in s} w_{hij} b_{hij} y_{hij}}{\sum_{(hij) \in s} w_{hij} b_{hij}} \right) \right] \\ \approx \gamma\beta \sum_{(hij) \in U} a_{hij} (1 - b_{hij}), \end{aligned} \quad (6.36)$$

noting that $Cov_m \left(\sum_{(hij) \in U} a_{hij} (1 - b_{hij}) x_{hij}, \sum_{(hij) \in U} b_{hij} y_{hij} \right) = 0$.

Similarly,

$$\begin{aligned} E_m E_p \left[\sum_{(hij) \in s} w_{hij} (1 - a_{hij}) b_{hij} y_{hij} \left(\frac{\sum_{(glk) \in s} w_{glk} a_{glk} x_{glk}}{\sum_{(glk) \in s} w_{glk} a_{glk}} \right) \right] \\ \approx \gamma\beta \sum_{(hij) \in U} (1 - a_{hij}) b_{hij}. \end{aligned} \quad (6.37)$$

and

$$E_m E_p \left[\sum_{(hij) \in s} w_{hij} (1 - a_{hij}) (1 - b_{hij}) \left(\frac{\sum_{(glk) \in s} w_{glk} a_{glk} x_{glk}}{\sum_{(glk) \in s} w_{glk} a_{glk}} \right) \left(\frac{\sum_{(glk) \in s} w_{glk} b_{glk} y_{glk}}{\sum_{(glk) \in s} w_{glk} b_{glk}} \right) \right] \\ \approx \sum_{(hij) \in U} (1 - a_{hij}) (1 - b_{hij}) \left[\frac{\sum_{(glk) \in U} a_{glk} b_{glk}}{\sum_{(glk) \in U} a_{glk} \sum_{(glk) \in U} b_{glk}} \sigma_{\epsilon\eta} + \gamma\beta \right] \quad (6.38)$$

Combining (6.35), (6.36), (6.37) and (6.38), we obtain

$$E(\hat{\theta}_I) = \gamma\beta + \frac{\sigma_{\epsilon\eta}}{K} E_r \left[\sum_{(hij) \in U} a_{hij} b_{hij} + \sum_{(hij) \in U} (1 - a_{hij}) (1 - b_{hij}) \left(\frac{\sum_{(hij) \in U} a_{hij} b_{hij}}{\sum_{(hij) \in U} a_{hij} \sum_{(hij) \in U} b_{hij}} \right) \right]$$

It now follows that

$$\begin{aligned} \text{Bias}(\hat{\theta}_I) &= -\frac{\sigma_{\epsilon\eta}}{K} E_r \left[\sum_{(hij) \in U} (1 - a_{hij} b_{hij}) \right. \\ &\quad \left. + \sum_{(hij) \in U} (1 - a_{hij}) (1 - b_{hij}) \frac{\sum_{(hij) \in U} a_{hij} b_{hij}}{\sum_{(hij) \in U} a_{hij} \sum_{(hij) \in U} b_{hij}} \right] \\ &= -\frac{\sigma_{\epsilon\eta}}{K} \sum_{(hij) \in U} E_r (1 - a_{hij} b_{hij}) + O\left(\frac{1}{K}\right) \end{aligned}$$

6.9.3 Unbiasedness of $\hat{\theta}_I^a$ under assumption DB

We show that the bias-adjusted estimator $\hat{\theta}_I^a$ in (6.12) for MI is approximately unbiased under assumption DB. The proofs for CDI and HI are very similar. We first show that $E(\hat{T}_I) \approx T$. To that end, we decompose $\hat{T}_I$ as

$$\hat{T}_I = \hat{T}_{I1} + \hat{T}_{I2} + \hat{T}_{I3} + \hat{T}_{I4},$$

where

$$\begin{aligned}
 \hat{T}_{I1} = & \sum_{(hij) \in s_{rr}} \sum_{\substack{(glk) \in s_{rr} \\ (hij) \neq (glk)}} w_{hij} w_{glk} g(x_{hij}, y_{glk}) \\
 & + \sum_{(hij) \in s_{rr}} \sum_{(glk) \in s_{mr}} w_{hij} w_{glk} g(x_{hij}, y_{glk}) \\
 & + \sum_{(hij) \in s_{rm}} \sum_{(glk) \in s_{mr}} w_{hij} w_{glk} g(x_{hij}, y_{glk}) \\
 & + \sum_{(hij) \in s_{rm}} \sum_{(glk) \in s_{rr}} w_{hij} w_{glk} g(x_{hij}, y_{glk}), \tag{6.39}
 \end{aligned}$$

$$\begin{aligned}
 \hat{T}_{I2} = & \sum_{(hij) \in s_{mm}} \sum_{\substack{(glk) \in s_{rr} \\ (hij) \neq (glk)}} w_{hij} w_{glk} g(x_{hij}^*, y_{glk}) \\
 & + \sum_{(hij) \in s_{mm}} \sum_{(glk) \in s_{mr}} w_{hij} w_{glk} g(x_{hij}^*, y_{glk}) \\
 & + \sum_{(hij) \in s_{mr}} \sum_{\substack{(glk) \in s_{mr} \\ (hij) \neq (glk)}} w_{hij} w_{glk} g(x_{hij}^*, y_{glk}) \\
 & + \sum_{(hij) \in s_{mr}} \sum_{(glk) \in s_{rr}} w_{hij} w_{glk} g(x_{hij}^*, y_{glk}), \tag{6.40}
 \end{aligned}$$

$$\begin{aligned}
 \hat{T}_{I3} = & \sum_{(hij) \in s_{rr}} \sum_{\substack{(glk) \in s_{rm} \\ (hij) \neq (glk)}} w_{hij} w_{glk} g(x_{hij}, y_{glk}^*) \\
 & + \sum_{(hij) \in s_{rr}} \sum_{(glk) \in s_{mm}} w_{hij} w_{glk} g(x_{hij}, y_{glk}^*) \\
 & + \sum_{(hij) \in s_{rm}} \sum_{\substack{(glk) \in s_{rm} \\ (hij) \neq (glk)}} w_{hij} w_{glk} g(x_{hij}, y_{glk}^*) \\
 & + \sum_{(hij) \in s_{rm}} \sum_{(glk) \in s_{mm}} w_{hij} w_{glk} g(x_{hij}, y_{glk}^*), \tag{6.41}
 \end{aligned}$$

and

$$\begin{aligned}
\hat{T}_{I4} = & \sum_{(hij) \in s_{mm}} \sum_{\substack{(glk) \in s_{mm} \\ (hij) \neq (gkl)}} w_{hij} w_{glk} g(x_{hij}^*, y_{glk}^*) \\
& + \sum_{(hij) \in s_{mm}} \sum_{(glk) \in s_{rm}} w_{hij} w_{glk} g(x_{hij}^*, y_{glk}^*) \\
& + \sum_{(hij) \in s_{rm}} \sum_{(glk) \in s_{mm}} w_{hij} w_{glk} g(x_{hij}^*, y_{glk}^*) \\
& + \sum_{(hij) \in s_{mr}} \sum_{(glk) \in s_{rm}} w_{hij} w_{glk} g(x_{hij}^*, y_{glk}^*). \tag{6.42}
\end{aligned}$$

From (6.39), we obtain

$$E_* \left[\sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (gkl)}} w_{hij} w_{glk} g(x_{hij}, y_{glk}) \right] = \sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (gkl)}} w_{hij} w_{glk} g(x_{hij}, y_{glk}) \tag{6.43}$$

From (6.40), we obtain

$$\begin{aligned}
& E_* \left[\sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (gkl)}} w_{hij} w_{glk} g(x_{hij}^*, y_{glk}) \right] \\
& = \sum_{(hij) \in s_{mm}} \sum_{(glk) \in s_{rr}} w_{hij} w_{glk} \left[\frac{\sum_{(abc) \in sr_x} w_{abc} g(x_{abc}, y_{glk})}{\sum_{(abc) \in Ar_x} w_{abc}} \right] \\
& + \sum_{(hij) \in s_{mm}} \sum_{(glk) \in s_{mr}} w_{hij} w_{glk} \left[\frac{\sum_{(abc) \in sr_x} w_{abc} g(x_{abc}, y_{glk})}{\sum_{(abc) \in Ar_x} w_{abc}} \right] \\
& + \sum_{(hij) \in s_{mr}} \sum_{\substack{(glk) \in s_{mr} \\ (hij) \neq (gkl)}} w_{hij} w_{glk} \left[\frac{\sum_{(abc) \in sr_x} w_{abc} g(x_{abc}, y_{glk})}{\sum_{(abc) \in Ar_x} w_{abc}} \right] \\
& + \sum_{(hij) \in s_{mr}} \sum_{(glk) \in s_{rr}} w_{hij} w_{glk} \left[\frac{\sum_{(abc) \in sr_x} w_{abc} g(x_{abc}, y_{glk})}{\sum_{(abc) \in Ar_x} w_{abc}} \right]. \tag{6.44}
\end{aligned}$$

From (6.41), we obtain

$$\begin{aligned}
& E_* \left[\sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (glk)}} w_{hij} w_{glk} g(x_{hij}, y_{glk}^*) \right] \\
&= \sum_{(hij) \in s_{rr}} \sum_{(glk) \in s_{rm}} w_{hij} w_{glk} \left[\frac{\sum_{(abc) \in s_{ry}} w_{abc} g(x_{hij}, y_{abc})}{\sum_{(abc) \in A_{ry}} w_{abc}} \right] \\
&+ \sum_{(hij) \in s_{rr}} \sum_{(glk) \in s_{mm}} w_{hij} w_{glk} \left[\frac{\sum_{(abc) \in s_{ry}} w_{abc} g(x_{hij}, y_{abc})}{\sum_{(abc) \in A_{ry}} w_{abc}} \right] \\
&+ \sum_{(hij) \in s_{rr}} \sum_{(glk) \in s_{mm}} w_{hij} w_{glk} \left[\frac{\sum_{(abc) \in s_{ry}} w_{abc} g(x_{hij}, y_{abc})}{\sum_{(abc) \in A_{ryx}} w_{abc}} \right] \\
&+ \sum_{(hij) \in s_{mr}} \sum_{(glk) \in s_{rr}} w_{hij} w_{glk} \left[\frac{\sum_{(abc) \in s_{ry}} w_{abc} g(x_{hij}, y_{abc})}{\sum_{(abc) \in A_{ry}} w_{abc}} \right]. \tag{6.45}
\end{aligned}$$

From (6.42), we obtain

$$\begin{aligned}
& E_* \left[\sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (glk)}} w_{hij} w_{glk} g(x_{hij}^*, y_{glk}^*) \right] = \\
& \sum_{(hij) \in s_{mm}} \sum_{\substack{(glk) \in s_{mm} \\ (hij) \neq (glk)}} w_{hij} w_{glk} \left[\frac{\sum_{(abc) \in s_{rx}} \sum_{(def) \in s_{ry}} w_{abc} w_{def} g(x_{abc}, y_{def})}{\sum_{(abc) \in s_{rx}} w_{abc} \sum_{(def) \in s_{ry}} w_{def}} \right] \\
&+ \sum_{(hij) \in s_{mm}} \sum_{(glk) \in s_{rm}} w_{hij} w_{glk} \left[\frac{\sum_{(abc) \in s_{rx}} \sum_{(def) \in s_{ry}} w_{abc} w_{def} g(x_{abc}, y_{def})}{\sum_{(abc) \in s_{rx}} w_{abc} \sum_{(def) \in s_{ry}} w_{def}} \right] \\
&+ \sum_{(hij) \in s_{mr}} \sum_{(glk) \in s_{mm}} w_{hij} w_{glk} \left[\frac{\sum_{(abc) \in s_{rx}} \sum_{(def) \in s_{ry}} w_{abc} w_{def} g(x_{abc}, y_{def})}{\sum_{(abc) \in s_{rx}} w_{abc} \sum_{(def) \in s_{ry}} w_{def}} \right] \\
&+ \sum_{(hij) \in s_{mr}} \sum_{(glk) \in s_{rm}} w_{hij} w_{glk} \left[\frac{\sum_{(abc) \in s_{rx}} \sum_{(def) \in s_{ry}} w_{abc} w_{def} g(x_{abc}, y_{def})}{\sum_{(abc) \in s_{rx}} w_{abc} \sum_{(def) \in s_{ry}} w_{def}} \right] \tag{6.46}
\end{aligned}$$

From (6.43), we obtain

$$\begin{aligned}
& E_r E_* \left[\sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (glk)}} w_{hij} w_{glk} g(x_{hij}, y_{glk}) \right] \\
&= (p_{rr}^2 + p_{rr}p_{mr} + p_{rm}p_{mr} + p_{rm}p_{rr}) \hat{T}, \tag{6.47}
\end{aligned}$$

where $\hat{T} = \sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (glk)}} w_{hij} w_{glk} g(x_{hij}, y_{glk})$.

From (6.44), we obtain

$$\begin{aligned} E_r E_* \left[\sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (glk)}} w_{hij} w_{glk} g(x_{hij}^*, y_{glk}) \right] \\ \approx (p_{mm} p_{rr} + p_{mm} p_{mr} + p_{mr}^2 + p_{rm} p_{rr}) \hat{T}. \end{aligned} \quad (6.48)$$

From (6.45), we obtain

$$\begin{aligned} E_r E_* \left[\sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (glk)}} w_{hij} w_{glk} g(x_{hij}, y_{glk}^*) \right] \\ \approx (p_{rr} p_{rm} + p_{rr} p_{mm} + p_{rm}^2 + p_{rm} p_{mm}) \hat{T}. \end{aligned} \quad (6.49)$$

From (6.46), we obtain

$$\begin{aligned} E_r E_* \left[\sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (glk)}} w_{hij} w_{glk} g(x_{hij}^*, y_{glk}^*) \right] \\ \approx (p_{mm}^2 + p_{mm} p_{rm} + p_{mr} p_{rm} + p_{mr} p_{rm}) \hat{T}. \end{aligned} \quad (6.50)$$

Adding (6.43), (6.44), (6.45) and (6.46), we get

$$E_r E_* (\hat{T}_I) \approx (p_{rr} + p_{rm} + p_{mr} + p_{mm})^2 \hat{T} = \hat{T},$$

noting that $p_{rr} + p_{rm} + p_{mr} + p_m = 1$. Finally,

$$E(\hat{T}_I) \approx E_p E_r E_* (\hat{T}_I) = E_p (\hat{T}) = T.$$

Hence, we have

$$E(\hat{\theta}_I^a) \approx p_{rr}^{-1} \left[p_{rr} \theta + (1 - p_{rr}) \frac{T}{K^2} \right] + (1 - p_{rr}^{-1}) \frac{T}{K^2} = \theta$$

6.9.4 Unbiasedness of $\hat{\theta}_I^a$ under assumption MB

We show that the bias-adjusted estimator (6.12) for MI is approximately unbiased under assumption MB in the particular case $g(x, y) = xy$. The proofs for CDI and JI are very similar. We first derive $E(\hat{T}_I)$. To that end, note that

$$\begin{aligned} \frac{\hat{T}_I}{\hat{K}^2} &= \frac{1}{\hat{K}^2} \sum_{(hij) \in s} \sum_{\substack{(glk) \in s \\ (hij) \neq (glk)}} w_{hij} w_{glk} \tilde{x}_{hij} \tilde{y}_{glk} \\ &= \bar{x}_I \bar{y}_I - \frac{1}{\hat{K}^2} \sum_{(hij) \in s} w_{hij}^2 \tilde{x}_{hij} \tilde{y}_{hij}, \end{aligned}$$

where

$$(\bar{x}_I, \bar{y}_I) = \frac{1}{\hat{K}} \sum_{hij \in s} w_{hij} (\tilde{x}_{hij}, \tilde{y}_{hij}).$$

Next, we determine $E_*(\hat{T}_I/\hat{K}^2)$. We have $E_*(\bar{x}_I \bar{y}_I) = \text{Cov}_*(\bar{x}_I, \bar{y}_I) + E_*(\bar{x}_I) E_*(\bar{y}_I)$.

Noting that $\text{Cov}_*(x_{hij}^*, y_{hij}^*) = 0$ under MI, and that

$$E_*(\bar{x}_I, \bar{y}_I) = \frac{\sum_{(hij) \in s} w_{hij} (a_{hij} x_{hij}, b_{hij} y_{hij})}{\sum_{(hij) \in s} w_{hij} (a_{hij}, b_{hij})},$$

we obtain

$$\begin{aligned} E_*(\bar{x}_I, \bar{y}_I) &= E_*(\bar{x}_I) E_*(\bar{y}_I) \\ &= \left(\frac{\sum_{(hij) \in s} w_{hij} a_{hij} x_{hij}}{\sum_{(hij) \in s} w_{hij} a_{hij}} \right) \left(\frac{\sum_{(hij) \in s} w_{hij} b_{hij} y_{hij}}{\sum_{(hij) \in s} w_{hij} b_{hij}} \right). \end{aligned} \quad (6.51)$$

Similarly,

$$\begin{aligned}
E_* \left[\frac{1}{\hat{K}^2} \sum_{(hij) \in s} w_{hij}^2 x_{hij}^2 y_{hij}^2 \right] &= \sum_{(hij) \in s} w_{hij}^2 a_{hij} b_{hij} x_{hij} y_{hij} \\
&+ \sum_{(hij) \in s} w_{hij}^2 a_{hij} (1 - b_{hij}) x_{hij} \frac{\sum_{(glk) \in s} w_{glk} b_{glk} y_{glk}}{\sum_{(glk) \in s} w_{glk} b_{glk}} \\
&+ \sum_{(hij) \in s} w_{hij}^2 (1 - a_{hij}) b_{hij} y_{hij} \frac{\sum_{(glk) \in s} w_{glk} a_{glk} x_{glk}}{\sum_{(glk) \in s} w_{glk} a_{glk}} \\
&+ \sum_{(hij) \in s} w_{hij}^2 (1 - a_{hij}) (1 - b_{hij}) \\
&\frac{\sum_{(glk) \in s} w_{glk} a_{glk} x_{glk}}{\sum_{(glk) \in s} w_{glk} a_{glk}} \frac{\sum_{(glk) \in s} w_{glk} b_{glk} y_{glk}}{\sum_{(glk) \in s} w_{glk} b_{glk}} \quad (6.52)
\end{aligned}$$

Now, from (6.51) we obtain

$$E_m E_p E_* (\bar{x}_I \bar{y}_I) \approx \frac{\sum_{(hij) \in U} a_{hij} b_{hij}}{\left(\sum_{(hij) \in U} a_{hij} \right) \left(\sum_{(hij) \in U} b_{hij} \right)} \sigma_{\epsilon\eta} + \gamma\beta. \quad (6.53)$$

Similarly, from (6.52), we obtain

$$\begin{aligned}
E_m E_p E_* \left(\frac{1}{\hat{K}^2} \sum_{(hij) \in s} w_{hij}^2 x_{hij}^2 y_{hij}^2 \right) &\approx \left(\sum_{(hij) \in U} w_{hij} / K^2 \right) \gamma\beta \\
&+ \frac{\sigma_{\epsilon\eta}}{K^2} \left[\sum_{(hij) \in U} w_{hij} a_{hij} b_{hij} + \frac{\sum_{(hij) \in U} w_{hij} (1 - a_{hij}) (1 - b_{hij})}{\sum_{(hij) \in U} a_{hij} \sum_{(hij) \in U} b_{hij}} \sum_{(hij) \in U} a_{hij} b_{hij} \right]. \quad (6.54)
\end{aligned}$$

Noting that $E_m(\theta) = \gamma\beta + \sigma_{\epsilon\eta}$, it follows from (6.53) and (6.54) that $E(\hat{\theta}_I^a - \theta) \approx 0$.

Chapter 7

On the construction of imputation classes

7.1 Introduction

In this chapter, we focus on random hot-deck imputation, where missing values of a given nonrespondent are imputed by the values of a randomly chosen respondent. Random hot-deck imputation is one of the most popular imputation methods, especially for categorical variables, and is very often used in household surveys. We also consider mean imputation since it has statistical properties similar to random hot-deck imputation, apart from the fact that it does not suffer from an additional component of variance due to the use of a random imputation mechanism.

As noted in Chapter 1, it is customary to first divide respondents and nonrespondents into classes before imputing missing values. These imputation classes are formed on the basis of auxiliary information recorded for all the units in the sample.

The objective in forming the classes is first to reduce the nonresponse bias but also to reduce the nonresponse variance as well as the imputation variance in the case of random imputation. It is also very common to form weighting classes to compensate for unit nonresponse before using weighting adjustment methods (see, for example, Eltinge and Yansaneh, 1997, Little, 1986, Oh and Scheuren, 1983, and Thomsen, 1973). The formation of imputation classes is related to stratification in survey sampling. Stratification techniques can thus be borrowed to form imputation classes, although the objectives of stratification are different.

There are at least four main advantages motivating the use of random hot-deck imputation within classes, instead of imputing directly the predicted value from a regression model: (i) it is more robust to model misspecification; (ii) it better preserves the variability of the original data; (iii) it is convenient when it is desired to impute more than one variable at a time since a unique donor can be used to impute all missing values of a given nonrespondent; and (iv) it leads to plausible values. The latter is particularly important in the case of categorical variables of interest. Random hot-deck imputation has also two main disadvantages: (i) its robustness property is usually accompanied by a loss of efficiency when the regression model holds; and (ii) it has an additional component of variance due to the use of a random imputation mechanism.

In practice, there are several methods used for the formation of imputation classes. Common methods include: (1) using strata or groups of strata; (2) cross-classifying simple categorical variables such as province, age, sex, etc.; (3) using available decision tree algorithms such as CHAID (Kass, 1980) and CART (Steinberg and Colla,

1995), or using classification algorithms; and (4) using the score method (see, for example, Thomsen, 1973; Little, 1986, and Eltinge and Yansaneh, 1997). Method (1) is common in business surveys and might be the only possible choice when there is no other auxiliary information than stratification. Method (2) is very popular, especially in household surveys. However, it can often lead to a large number of classes and hence has a tendency to be less stable than other methods. When some imputation classes are too small, some form of collapsing is usually performed with this method. The idea behind methods (3) and (4) is similar; that is, constructing classes homogeneous with respect to the propensity to respond or to the variable of interest. However, the criterion used to achieve homogeneity is different. In this paper, we focus mainly on method (4) but also consider method (2) in a simulation study for comparison purposes.

We begin by introducing some notation in section 7.2. We focus on the case of a single variable requiring imputation. In section 7.3, we describe two approaches that can be used to justify the formation of classes. In the first approach, the bias and variance of the imputed estimator are evaluated with respect to the nonresponse model (or nonresponse mechanism). In the second approach, these quantities are evaluated with respect to the imputation model. Unlike Thomsen (1973), Little (1986), and Eltinge and Yansaneh (1997), we condition on the realized sample in both approaches. This allows us to obtain a theoretical justification for the use of classes formed on the basis of auxiliary variables available for the sampled units only. Some modeling issues are also raised. Methods for constructing classes are described in section 7.4. Section 7.5 provides some methods used to determine the number

of classes. The extension to the multivariate case, where more than one variable of interest is subject to nonresponse, is discussed in section 7.6. A simulation study was conducted to evaluate and compare the different methods. Results of the simulation study are presented in section 7.7. In section 7.8, we discuss the results obtained in the simulation study.

7.2 Background and notation

Let U be a finite population of size N . The objective is to estimate the population mean $\bar{Y} = \frac{1}{N} \sum_{i \in U} y_i$ of a variable of interest y . Suppose that a random sample s of size n is selected from U according to a given design $p(s)$. We assume that a vector of auxiliary variables $\mathbf{x}_i$ is available for all units in the sample and that the vector of population totals, $\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$, is known. In the case of complete response, an approximately design-unbiased estimator of $\bar{Y}$ is the Generalized REGression (GREG) estimator given by

$$\bar{y}^G = \frac{1}{\hat{N}} \left[\sum_{i \in s} \tilde{w}_i y_i \right], \quad (7.1)$$

where $\hat{N} = \sum_{i \in s} \tilde{w}_i$, $\tilde{w}_i = w_i g_i$, $w_i = \frac{1}{\pi_i}$ denotes the design weight for unit i , $\pi_i = P(i \in s)$ is the first-order probability of inclusion in the sample for unit i , and g_i denotes the so-called g -weight (or calibrator factor) given by

$$g_i = 1 + (\mathbf{X} - \hat{\mathbf{X}})' \hat{\mathbf{T}}^{-1} \mathbf{x}_i / v_i,$$

where $\hat{\mathbf{X}} = \sum_{i \in s} w_i \mathbf{x}_i$, $\hat{\mathbf{T}} = \sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i' / v_i$ and v_i is a known constant proportional to the error variance of the working model underlying the GREG estimator. The calibra-

tion weights $\tilde{w}_i$ ensure consistency with the known $\mathbf{X}$ in the sense that $\sum_{i \in s} \tilde{w}_i \mathbf{x}_i = \mathbf{X}$.

In the presence of nonresponse to a variable y , it is customary to first partition the sample s into C imputation classes $s_1, \dots, s_c$ of sizes $n_1, \dots, n_c$. We define an imputed estimator of the population mean $\bar{Y}$ based on the C classes by

$$\bar{y}_I^G = \sum_{c=1}^C \frac{\hat{N}_c}{\hat{N}} \bar{y}_{Ic}, \quad (7.2)$$

where

$$\bar{y}_{Ic} = \frac{1}{\hat{N}_c} \left[\sum_{i \in s_{rc}} \tilde{w}_{ci} y_{ci} + \sum_{i \in s_{mc}} \tilde{w}_{ci} y_{ci}^* \right] \quad (7.3)$$

is the imputed estimator of the mean in class c , $\hat{N}_c = \sum_{i \in s_c} \tilde{w}_{ci}$, s_{rc} denotes the random set of respondents to variable y in class c , s_{mc} denotes the random set of nonrespondents to variable y in class c and y_{ci}^* denotes the imputed value for missing y_{ci} .

In this chapter, we consider two imputation methods: mean and random hot-deck imputation within classes. In the case of mean imputation, we have

$$y_{ci}^* = \bar{y}_{rc} \equiv \sum_{i \in s_{rc}} \tilde{w}_{ci} y_{ci} / \sum_{i \in s_{rc}} \tilde{w}_{ci} \quad (7.4)$$

The imputed estimator (7.2) reduces to $\bar{y}_{HI}^G = \sum_{c=1}^C \frac{\hat{N}_c}{\hat{N}} \bar{y}_{rc}$ under mean imputation.

In the case of random hot-deck imputation, a missing y -value in class c is replaced by the y -value of a donor (respondent) selected randomly from the set of respondent s_{rc} with replacement and with probability proportional to its calibration weight $\tilde{w}_{ci}$; that is,

$$y_{ci}^* = y_{cj} \text{ for some } j \in s_{rc} \text{ such that } P(y_{ci}^* = y_{cj}) = \tilde{w}_{cj} / \sum_{i \in s_{rc}} \tilde{w}_{ci}. \quad (7.5)$$

7.3 Evaluating the quality of the imputed estimator

Different approaches may be used for evaluating the quality of the imputed estimator. To understand the nature of these approaches, we first identify three sources of randomness: (i) the imputation model m , which generates the vector of population values $\mathbf{y} = (y_1, \dots, y_N)'$; (ii) the sampling design $p(s)$, which generates the vector of sample indicators $\boldsymbol{\delta} = (\delta_1, \dots, \delta_N)'$ such that $\delta_i = 1$ if unit i is selected in the sample and $\delta_i = 0$ otherwise; and (iii) the nonresponse mechanism which generates the vector of response indicators $\mathbf{a} = (a_1, \dots, a_N)'$ such that $a_i = 1$ if unit i belongs to the set of respondents s_r and $a_i = 0$ otherwise. Different combinations of these distributions may be used to evaluate the properties of the imputed estimator (7.2). For example, Thomsen (1973) and Little (1986) determine the bias and the variance of the estimators with $\mathbf{y}$ and $\mathbf{a}$ held fixed. In this case, the only source of randomness left is the vector of sample indicators $\boldsymbol{\delta}$. In their approach, expectations and variances are thus evaluated with respect to the sampling design. It also implicitly assumes that the nonresponse mechanism does not depend on s . Eltinge and Yansaneh (1997) determine the bias and the variance of the estimators with only $\mathbf{y}$ held fixed. Expectations and variances are thus evaluated with respect to both the sampling design and the nonresponse mechanism. In all these approaches, the imputation classes are defined at the population level.

In practice, the imputation classes are generally formed after the sample s is selected. For this reason, we prefer to study the bias and the variance conditional on

the realized sample s . Under this conditional framework, we consider two approaches: (i) a Generalized Nonresponse Model Approach (GNM) and (ii) an Imputation Model Approach (IM). Under an GNM, we calculate the bias and variance of the imputed estimator (2.2) with $\mathbf{y}$ and δ held fixed. Under an IM, we calculate the bias and variance of the imputed estimator (2.2) with δ and $\mathbf{a}$ held fixed. In the case of random hot-deck imputation, there is a fourth source of randomness: the imputation mechanism $I(s_m^*|s, s_r)$, where s_m^* is the random set of donors (respondents) used to replace the missing values in the set of nonrespondents s_m . We do not condition on s_m^* neither under GNM nor under IM.

Similar to the decomposition (1.39), the total error, $\bar{y}_I^G - \bar{Y}$, can be decomposed as follows:

$$\bar{y}_I^G - \bar{Y} = (\bar{y}^G - \bar{Y}) + (\bar{y}_I^G - \bar{y}^G). \quad (7.6)$$

Since the sampling error, $\bar{y}_I^G - \bar{Y}$, does not depend on nonresponse and on the imputation method chosen, we focus on the nonresponse/imputation error, $(\bar{y}_I^G - \bar{y}^G)$, to evaluate the properties of the imputed estimator.

In sections 7.3.1 and 7.3.2, we give the conditional nonresponse bias and the conditional nonresponse variance of the imputed estimator (7.2) under GNM and IM respectively. We only consider mean imputation within classes. The extension to random hot-deck imputation within classes is described in section 7.3.3. In section 7.3.4, some modeling issues are raised. Variance estimation is briefly discussed in section 7.3.5.

7.3.1 Nonresponse Model Approach

This approach consists in describing the nonresponse mechanism by means of a nonresponse model. First, we assume that the nonresponse model parameters are known. Let $p_i = P(i \in s_r \mid s, i \in s)$ denote the probability of response for unit i . We assume that units respond independently of one another; that is, $p_{ij} = P(i \in s_r, j \in s_r \mid s, i \in s, j \in s) = p_i p_j$ for all $i \neq j$. Under this nonresponse model, it is easily seen that for mean imputation within classes, the conditional nonresponse bias of the imputed estimator (7.2) is approximated by

$$E_r(\bar{y}_{MI}^G - \bar{y}^G \mid s) \approx \frac{1}{N} \sum_{c=1}^C \bar{p}_c^{-1} \sum_{i \in s_c} \tilde{w}_{ci} (p_{ci} - \bar{p}_c) (y_{ci} - \bar{y}_c), \quad (7.7)$$

where $\bar{p}_c = \sum_{i \in s_c} \tilde{w}_{ci} p_{ci} / \hat{N}_c$,

$$\bar{y}_c = \sum_{i \in s_c} \tilde{w}_{ci} y_{ci} / \hat{N}_c. \quad (7.8)$$

It follows that the conditional bias (7.7) is approximately equal to zero if the covariance between the response probability and the variable of interest is approximately equal to zero in each class. This is the case, for example, when the imputation classes are homogeneous with respect to the response probabilities and/or the variable of interest. Forming classes homogeneous with respect to the response probabilities can be essentially achieved by forming classes homogeneous with respect to estimated response probabilities. However, it is not possible to form imputation classes based on the variable y itself since the y -values are not observed for the nonrespondents and cannot be estimated under this approach. One possible approximate solution to this problem can be obtained by finding a proxy variable closely related to y and observed

for every sample units. Classes can then be made homogeneous with respect to this proxy variable. In section 7.3.2, we see that a proxy variable comes out naturally under IM.

It is also easily seen that, under mean imputation within classes, the conditional nonresponse variance of the imputed estimator (7.2) is given by

$$V_r(\bar{y}_{MI}^G - \bar{y}^G | s) \approx \frac{1}{\hat{N}^2} \sum_{c=1}^C \frac{1}{\bar{p}_c^2} \sum_{i \in s_c} \tilde{w}_{ci}^2 p_{ci} (1 - p_{ci}) (y_{ci} - \bar{y}_c^{(p)})^2, \quad (7.9)$$

where $\bar{y}_c^{(p)} = \sum_{i \in s_c} \tilde{w}_{ci} p_{ci} y_{ci} / \sum_{i \in s_c} \tilde{w}_{ci} p_{ci}$. It follows from (7.9) that the variance of the imputed estimator $\bar{y}_{MI}^G$ can be reduced by forming imputation classes that are homogeneous with respect to the variable of interest. Again, a proxy variable closely related to y and observed for every sample units is needed since the variable y is not observed for the nonrespondents. From expression (7.9), it is not clear, however, that forming imputation classes homogeneous with respect to estimated response probabilities would reduce the variance of the imputed estimator $\bar{y}_{MI}^G$ unless the response probability is correlated with the variable of interest. This point was also noted by Little (1986).

7.3.2 Imputation Model Approach

This approach consists in describing the relationship between the variable of interest and a vector of auxiliary variables $\mathbf{z}$, available for all units in the sample, used at the imputation stage to construct imputation classes. Note that $\mathbf{z}_i$ is not necessarily equal to $\mathbf{x}_i$. We assume that the sampling design and the nonresponse mechanism

are ignorable (Rubin, 1976) with respect to the following imputation model

$$\begin{aligned} m : E_m(y_i | \mathbf{z}_1, \dots, \mathbf{z}_N) &= \mu_i; Cov_m(y_i, y_j | \mathbf{z}_1, \dots, \mathbf{z}_N) = 0, i \neq j; \\ V_m(y_i | \mathbf{z}_1, \dots, \mathbf{z}_N) &= \sigma_i^2, \end{aligned} \quad (7.10)$$

A particular case of (7.10) frequently used in practice is the linear regression imputation model in which $\mu_i = \mathbf{z}_i' \boldsymbol{\beta}$ and $\sigma_i^2 = \boldsymbol{\lambda}' \mathbf{z}_i$, where $\boldsymbol{\beta}$ is a vector of unknown parameters and $\boldsymbol{\lambda}$ is a vector of specified constants.

Under the imputation model (7.10), it is easily seen that the conditional nonresponse bias of the imputed estimator (7.2) is given by

$$E_m(\bar{y}_{MI}^G - \bar{y}^G | s, s_r) = -\frac{1}{\hat{N}} \sum_{c=1}^C \sum_{i \in s_c} \tilde{w}_{ci} (\mu_{ci} - \bar{\mu}_{r_c}), \quad (7.11)$$

where $\bar{\mu}_{r_c} = \sum_{i \in s_{r_c}} \tilde{w}_{ci} \mu_{ci} / \sum_{i \in s_{r_c}} \tilde{w}_{ci}$. The bias in (7.11) vanishes if μ_{ci} is constant within an imputation class; that is,

$$\mu_{ci} = \mu_c, \text{ for all } i \text{ in class } c \text{ and for } c = 1, \dots, C. \quad (7.12)$$

Expression (7.12) corresponds in fact to the model underlying mean imputation within classes. Hence, the objective will be to form classes in order for (7.12) to hold. This can be essentially achieved by forming classes homogeneous with respect to estimates $\hat{\mu}_i$ of the conditional means μ_i .

Also, it is easily seen that, under model (7.10), the conditional nonresponse variance of the imputed estimator (7.2) is given by

$$V_m(\bar{y}_{MI}^G - \bar{y}^G | s, s_r) = \frac{1}{\hat{N}^2} \sum_{c=1}^C \frac{1}{\hat{p}_c^2} \left[\hat{p}_c^2 \sum_{i \in s_{m_c}} \tilde{w}_{ci}^2 \sigma_{ci}^2 + (1 - \hat{p}_c)^2 \sum_{i \in s_{r_c}} \tilde{w}_{ci}^2 \sigma_{ci}^2 \right], \quad (7.13)$$

where $\hat{p}_c = \sum_{i \in s_{rc}} \tilde{w}_{ci} / \sum_{i \in s_c} \tilde{w}_{ci}$ is the estimated response rate in class c . The variance (7.13) will be small if the model (7.10) has good predictive power or, in other words, if the model variances σ_{ci}^2 are small. This is the case when the vector of auxiliary variables $\mathbf{z}$ explains well the variability of the variable y .

7.3.3 Random hot-deck imputation

The conditional nonresponse bias can still be expressed as (7.7) and (7.11) under random hot-deck imputation within classes. This can be shown by noting that $E_\star (\bar{y}_I^G - \bar{y}_{MI}^G | s, s_r) = 0$. As a result, $E_r E_\star (\bar{y}_I^G - \bar{y}^G | s) = E_r (\bar{y}_{MI}^G - \bar{y}^G | s)$ and $E_m E_\star (\bar{y}_I^G - \bar{y}^G | s, s_r) = E_m (\bar{y}_{MI}^G - \bar{y}^G | s, s_r)$. Also, it is straightforward to show that

$$V_{r\star} (\bar{y}_I^G - \bar{y}^G | s) = V_r (\bar{y}_{MI}^G - \bar{y}^G | s) + E_r V_\star (\bar{y}_I^G | s, s_r) \quad (7.14)$$

and that

$$V_{m\star} (\bar{y}_I^G - \bar{y}^G | s, s_r) = V_m (\bar{y}_{MI}^G - \bar{y}^G | s, s_r) + E_m V_\star (\bar{y}_I^G | s, s_r) \quad (7.15)$$

Therefore, the conditional nonresponse variance of the imputed estimator under random hot-deck imputation is equal to the conditional nonresponse variance under mean imputation plus an extra term representing the variability due to the imputation mechanism. Under either an GNM (7.14) or an IM (7.15), this extra term can be estimated by $V_\star (\bar{y}_I^G | s, s_r)$, where

$$V_\star (\bar{y}_I^G - \bar{y}^G | s, s_r) = \frac{1}{\hat{N}^2} \sum_{c=1}^C \sum_{i \in s_{mc}} \tilde{w}_{ci}^2 s_{rc}^2, \quad (7.16)$$

and $s_{r_c}^2 = \frac{1}{\sum_{i \in s_{r_c}} \tilde{w}_i} \sum_{i \in s_{r_c}} \tilde{w}_{ci} (y_{ci} - \bar{y}_{r_c})^2$ is the variability of the y -values for the respondents in class c , $c = 1, \dots, C$. The imputation variance (7.16) will be small if the imputation classes are homogeneous with respect to the respondent y -values. We might thus expect that the imputation variance (7.11) will be small when the imputation classes are made homogeneous with respect to estimates $\hat{\mu}_i$ of μ_i , especially if these estimates are well correlated with the variable of interest.

7.3.4 Modeling issues

Results from sections 7.3.1 - 7.3.3 show that forming classes homogeneous with respect to response probabilities p_i and/or with respect to conditional means μ_i will help in reducing the nonresponse bias and might also reduce the nonresponse/imputation variance. In practice, however, p_i and μ_i are unknown. This problem can be addressed by forming classes homogeneous with respect to estimated response probabilities $\hat{p}_i$ and/or with respect to estimated conditional means $\hat{\mu}_i$. A nonresponse model (for example, a logistic regression model) is needed to obtain $\hat{p}_i$ while an imputation model is needed to obtain $\hat{\mu}_i$. Therefore, it is important that the nonresponse model and the imputation model be adequate descriptions of the unknown nonresponse mechanism and of the unknown relationship between y and $\mathbf{z}$, respectively.

The appropriate selection of auxiliary variables to include in these models is also very important. All the variables used to construct the estimator should be considered as potential auxiliary variables. The types of population parameters to be estimated should also be considered when selecting auxiliary variables. For example, it is always

wise to verify whether the indicators of main domains of interest should be included in the models or not so as to make sure that the models hold within these domains. This will help in reducing the nonresponse bias and the nonresponse variance when the population parameter to be estimated is a domain mean.

7.3.5 Variance estimation

There are a variety of variance estimation methods that have been developed under the assumption that the nonresponse mechanism is uniform within classes (see Chapter 1). These methods cover many imputation methods including mean and random hot-deck imputation within classes. All these methods can be used to estimate the variance of the imputed estimator (7.2) provided the imputation classes are homogeneous with respect to response probabilities so that the uniform nonresponse assumption holds reasonably well.

The uniform nonresponse assumption will not always hold if classes are made homogeneous with respect to $\hat{\mu}_i$. In such a case, it may be preferable to assume that (7.10) holds; that is, assuming an imputation model where the conditional means are constant within a class. Särndal (1992) and Shao and Steel (1999) provide two different methods for estimating the variance of the imputed estimator under the assumption that an imputation model holds.

7.4 Methods for constructing classes

In this section, we describe two methods for constructing imputation classes: the score method, and the cross-classification method. In section 7.7, we study the performance of these methods in a simulation study.

7.4.1 The score method

This method leads to a partition of the sample in such a way that, within classes, units (respondents and nonrespondents) are homogeneous with respect to scores, usually $\hat{p}$ or $\hat{\mu}$. We define a score to be any variable for which the values can be computed for all sampled units.

Using one score

The steps for the formation of classes may be described as follows:

1. Estimate the response probabilities p_i by $\hat{p}_i$, $i \in s$, using a nonresponse model, or estimate the conditional means μ_i by $\hat{\mu}_i$, $i \in s$, using an imputation model. Note that auxiliary variables, available for all sampled units, are used to build the models. Choose one of the two scores $\hat{p}$ or $\hat{\mu}$. Alternately, any other variable observed for all sampled units could be used as a score.
2. Using the chosen score, partition the sample using either (i) an equal-quantile method (see, for example, Eltinge and Yansaneh, 1997) or (ii) a classification algorithm (see, for example, Seber, 1984, Chapter 7). For the equal-quantile method, the class boundaries are determined by the j/C quantiles of a given

score, for $j = 1, \dots, C - 1$, where C is the desired number of classes. In the simulation study, we used the “ k -means” (or in our notation “ C -means”) classification algorithm to evaluate the properties of method (ii) using the procedure FASTCLUS of SAS. We chose this algorithm for its efficiency in terms of computer time. Efficiency is important in the context of a simulation study but is also important in practice for large datasets.

It might sometimes occur that some classes contain only nonrespondents, especially when a classification algorithm is used. A possible solution to this problem is to collapse some classes together. Another solution consists in first partitioning the set of respondents into C classes. Then, each nonrespondent is assigned to the class that minimizes a distance function (for example, the Euclidian distance) between the score of the nonrespondent and the class center (for example, the mean of the score in a class). This second solution has been used in the simulation study for method (ii). It ensures that each class will contain at least one respondent. Using the procedure FASTCLUS, it is also possible to ensure that each class will contain at least a specified number of respondents (greater than one).

Using more than one score

It is possible to use simultaneously more than one score to form the classes (for example, $\hat{p}$ and $\hat{\mu}$). The simultaneous use of $\hat{p}$ and $\hat{\mu}$ may give some robustness to a misspecification of one model or the other. The use of both scores $\hat{p}$ and $\hat{\mu}$ may come in different versions such as:

1. A classification algorithm using both scores may be used.
2. A second method consists in first forming C_1 classes with respect to one score, $\hat{p}$ (say) using an equal-quantile method. Within each $\hat{p}$ -based class, we form C_2 classes with respect to $\hat{\mu}$ using, once again, an equal-quantile method. This method leads to the creation of $C_1 \times C_2$ classes. It has been used by Smith et al. (2004) in the context of reweighting in the U.S. National Immunization Survey.
3. A third method consists in forming separately C_1 classes with respect to one score $\hat{p}$ (say) and C_2 classes with respect to $\hat{\mu}$ and cross-classify to obtain $C_1 \times C_2$ classes, (Vartivarian and Little, 2002).

Methods (i), (ii), and (iii) above can be extended straightforwardly to handle more than two scores, not necessarily including $\hat{p}$ and $\hat{\mu}$. For example, classes could be made homogeneous with respect to the auxiliary variables used in the nonresponse and imputation models provided that these auxiliary variables are numeric and can be ordered. This would give some robustness against a model misspecification. However, as the number of scores increases, the number of classes required to obtain homogeneous classes is likely to increase as well. Therefore, only the scores correlated with the response probability and/or with the variable of interest should be used. Also, note that methods (ii) and (iii) can lead to a very large number of classes when the number of scores is large. Method (i) allows the user to better control the number of classes in such a case.

Nearest-Score Imputation

A particular case of the score method is the Nearest-Score Imputation (NSI). Indeed, NSI imputes the respondent value, y_j , for missing y_i , such that the distance (for example, the Euclidian distance) between the scores of units i and j is minimum among all respondents j . NSI can thus be obtained by first partitioning the set of respondents into n_r classes, where n_r is the number of respondents. Then, each nonrespondent is assigned to the class that minimizes a distance function between the scores of the nonrespondent and the class center. Therefore, NSI can be viewed as a limiting case of the score method when $C = n_r$. That is, it leads to the finest partition possible of the sample into classes, where each class contains one respondent. It is possible that some classes contain more than one respondent in the presence of ties. In this case, a respondent is usually selected at random within these classes.

7.4.2 Cross-classification method

To apply this method, we first select a set of auxiliary variables that explain the response probability or the variable of interest well. Then, we rank the selected variables from the most significant to the least significant variable. An equivalent procedure frequently used in practice consists in assigning a weight to each variable; the larger the weight, the more significant is the variable. If some auxiliary variables are continuous, they must be categorized.

The classes are then formed by cross-classifying these categorical variables. To ensure stability of the point estimator, it is customary to specify the following two

constraints:

1. The minimum number of donors in a given class is fixed, say n_{min} .
2. In a given class, the proportion of donors must be greater than p_{min} (for example, $p_{min} = 1/2$).

Random Hot-deck imputation is then performed in all the classes satisfying the above constraints. However, since the initial number of classes is potentially very large, it is likely that a large number of classes will not satisfy the constraints. Hence some form of collapsing is needed to impute the nonrespondents in these classes. Here, the least significant variable is dropped from the list and the classes are defined by cross-classifying the remaining variables. Once again, random hot-deck imputation is performed in all classes that satisfy the constraints. If, there are still some classes that do not satisfy the constraints, the least significant variable is dropped from the list and so on. This process continues until each nonrespondent has found a donor. Note that unlike the score method, the classes are not disjoint; therefore, the same donor may be used in more than one class.

7.5 Empirical determination of the number of classes

In this section, we propose two measures useful in determining the number of classes: the coefficient of determination and the intraclass bias.

7.5.1 Coefficient of determination

Let $\hat{\lambda}_i$ denote either the score $\hat{p}_i$ or the score $\hat{y}_i$. The coefficient of determination R^2 is the correlation between the dependent variable $\hat{\lambda}_i$ and the class identifier variable. As the number of classes increases, R^2 increases as the classes become increasingly homogeneous with respect to $\hat{\lambda}_i$. We seek for a large value of R^2 . However, this might lead to a substantial number of classes which in turns leads to small classes. In practice, a compromise is then preferable. The R^2 possesses the attractive property that it does not depend on the parameter to be estimated.

7.5.2 Within-class bias

Another possible diagnostic tool is the assessment of the within-class bias. Consider the following estimators

$$\bar{y}_{wc}^{(1)} = \frac{1}{\sum_{i \in s_c} \tilde{w}_{ci} \hat{p}_{ci}^{-1}} \sum_{i \in s_c} \tilde{w}_{ci} \hat{p}_{ci}^{-1} y_{ci} \quad (7.17)$$

and

$$\bar{y}_{wc}^{(2)} = \frac{1}{\sum_{i \in s_c} \tilde{w}_{ci}} \sum_{i \in s_c} \tilde{w}_{ci} \hat{y}_{ci} \quad (7.18)$$

It is easily seen that the within-class estimator (7.17) is approximately unbiased for $\bar{Y}_c$ under the sampling design and the response mechanism provided that $\hat{p}_i = p_i$ for all $i \in s_c$. Similarly, the within-cell estimator (7.18) is approximately unbiased for $\bar{Y}_c$ provided that $E_m(\hat{y}_i - y_i) = 0$. These two estimators may then be used to measure the within-cell bias $\hat{B}_c = \bar{y}_c - \bar{y}_{wc}^{(j)}$, $j = 1, 2$, where $\bar{y}_c$ is given by (7.8). Large values of $\hat{B}_c$ may indicate a large within-cell bias in $\bar{y}_c$ but as Eltinge and

Yansaneh (1997) pointed out, it may also indicate that the errors $\hat{p}_i - p_i$ (or $\hat{y}_i - y_i$) are large. Hence, caution should be applied when interpreting large values of $\hat{B}_c$. If however, one considers both diagnostics (7.17) and (7.18) simultaneously, large $\hat{B}_c$ in only the probability of response p_i or the variable of interest y is less indicative of bias than if both values are large. Regardless of the cause, a large value of $\hat{B}_c$ should be investigated.

7.6 Multivariate case

So far, we have considered the case of a single variable, y , requiring imputation. In practice, this is rarely the case and there are often many variables of interest with missing values. It is quite impractical to derive different imputation classes for different variables of interest. Therefore, a compromise is needed. To obtain such a compromise, we can first find appropriate scores separately for each variable and then take the union of the scores to form classes using one of the methods described in section 4.1.2. With such a compromise, there will be many scores if there are many variables of interest. Therefore, to obtain homogeneity within classes with respect to all these scores and to remove (or at least reduce) nonresponse bias, a large number of classes is likely to be needed. However, this might have the effect of increasing the nonresponse variance. In the limit case, the NSI method might be required to reduce or eliminate the nonresponse bias for all variables of interest at the price of increasing the nonresponse variance. Therefore, this method offers a bias-variance compromise when there are many variables. An alternate compromise can be obtained by finding

scores for a few main variables only. In such a case, the imputation classes would be efficient for these main variables but might be inefficient for other variables not taken into account in the construction of classes.

How to form imputation classes for more complex parameters involving more than one variable has not been considered in this chapter. We expect that proper specifications of the nonresponse and imputation models will help reduce the nonresponse bias and variance for many parameters. The choice of auxiliary variables is crucial. For example, if it is desired to estimate population regression parameters, it is wise to consider including in the models, if possible, all the variables used to compute these parameters.

7.7 Simulation study

We conducted a simulation study to test the performance of both the score method and the cross-classification method. We first generated a finite population of size $N = 1,000$ containing 7 variables: a variable of interest y and 6 auxiliary variables $z_1, \dots, z_6$. The model used to generate y , z_1 and z_2 is given by

$$y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 z_{1i}^2 + \epsilon_i, \quad (7.19)$$

where z_1 and z_2 were generated from an exponential distribution with mean 100. The ϵ_i 's were generated according to a normal distribution with mean 0 and variance σ^2 . The remaining variables z_3, z_4, z_5 and z_6 were generated from an exponential distribution with means 100, 40, 60 and 80 respectively. The parameters $\beta_0, \beta_1, \beta_2$ and β_3 were fixed respectively to 20, 10, 0.5 and 10. The variance σ^2 was chosen

to give a model R^2 (coefficient of determination) approximately equal to 0.66. The variables $z_1, \dots, z_6$ were also been categorized into 5 classes to fit the needs of the cross-classification method (that uses categorical variables only). This led to the creation of 6 new variables $z_{1c}, \dots, z_{6c}$.

We considered the case of a census ($n = N = 1000$) in order to focus on the non-response error only. From the population, we generated nonresponse to variable y according to the following two response mechanisms:

Mechanism 1: The response probability p_{1i} for unit i is given by the logistic model

$$\log \frac{p_{1i}}{1 - p_{1i}} = \lambda_0 + \lambda_1 z_{1i} + \lambda_2 z_{3i},$$

where λ_0, λ_1 and λ_2 have been chosen so the mean of the p_{1i} 's is approximately equal to 70%.

Mechanism 2: The response probability p_{2i} for unit i is given by the logistic model

$$\log \frac{p_{2i}}{1 - p_{2i}} = \lambda_0 + \lambda_1 y_i + \lambda_2 z_{3i},$$

where λ_0, λ_1 and λ_2 have been chosen so the mean of the p_{2i} 's is approximately equal to 70%.

The response indicators a_{1i} and a_{2i} , for $i \in U$, were then generated independently 1000 times from a Bernoulli distribution with parameter p_{1i} and p_{2i} . This led to 1000 sets of respondents for each nonresponse mechanism. Note that the auxiliary variables $z_1, \dots, z_6$ are observed for all the sampled units. Then, for each set of respondent, we formed imputation classes using the score method and the cross-classification method and imputed missing values using random hot-deck imputation within classes.

From the simulation, we estimated the Percent Relative Bias (RB) of the imputed estimator given by

$$RB(\bar{y}_I^G) = \frac{E(\bar{y}_I^G) - \bar{Y}}{\bar{Y}} \times 100\%.$$

We also estimated the Percent Relative Root Mean Square Error (RRMSE) of the imputed estimator given by

$$RRMSE(\bar{y}_I^G) = \frac{\sqrt{MSE(\bar{y}_I^G)}}{\bar{Y}} \times 100\%.$$

7.7.1 The score method

For the score method, we wish to investigate the following aspects:

1. The impact of the response mechanism.
2. The impact of the choice between the two scores $\hat{y}$ and $\hat{p}$.
3. The quality of modeling the variable of interest and the response probability.
4. The impact of the choice between the equal-quantile method and a classification algorithm for forming the classes

7.7.2 The cross-classification method

For the cross-classification method, we wish to investigate the following aspects:

1. The impact of the response mechanism.
2. The quality of modeling the variables of interest.

3. The order of the variables to be cross-classified. We considered several ordering: (correct: $z_{1c}, z_{2c}, z_{3c}, z_{4c}, z_{5c}, z_{6c}$, incorrect: $z_{6c}, z_{5c}, z_{4c}, z_{3c}, z_{2c}, z_{1c}$ and $z_{2c}, z_{3c}, z_{1c}, z_{4c}, z_{5c}, z_{6c}$)
4. The impact of the constraints (see section 7.4.2): the minimum number of donors per class, n_{min} . We used $n_{min} = 1$ and $n_{min} = 9$. Also, the presence of the constraint “more donors than recipients” in a given class.

7.8 Discussion of the results

In this section, we discuss the simulation results for both the score method and the cross-classification method in sections 7.8.1 and 7.8.2 respectively. In section 7.8.3 we attempt to compare the two methods by stating the advantages and drawback for each of them.

7.8.1 The score method

In the light of the simulation results, we can make the following remarks for the score method:

When the imputation model (that leads to $\hat{y}_i$) or the nonresponse model (that leads to $\hat{p}_i$) are correctly specified, we note from Figure 7.1 ($\hat{y}$ -criterion) and Figure 7.3 ($\hat{p}$ -criterion) that the bias of the imputed estimator goes to 0 as the number of classes increase in the case of Mechanism 1 (ignorable). For Mechanism 2 (non ignorable), although the imputation classes help in reducing the bias, it cannot be eliminated (around 20% under $\hat{y}$ -criterion and 21% under $\hat{p}$ -criterion). Also, we note that in

the case of $\hat{y}$ -criterion (see Figure 7.1) the bias of the imputed estimator stabilizes somewhat rapidly (around 20 classes for Mechanism 1). This is not consistent with the suggestion that $C = 5$ classes may provide most of the effective bias reduction (see Eltinge and Yansaneh, 1997 and Rosenbaun and Rubin, 1984). This result can be explained by the presence of the quadratic term in the model (7.19) that generated y -values. In this case, it takes more classes for the model (7.12) to be satisfied. In the case of $\hat{p}$ -criterion (see Figure 7.3), the bias stabilizes very rapidly (around 5 classes for both Mechanism 1 and Mechanism 2). This is consistent with the suggestion that $C = 5$ classes may provide most of the effective bias reduction. In both cases ($\hat{y}$ -criterion and $\hat{p}$ -criterion), it is interesting to note that a moderate number of classes ($C < 25$) provide an effective bias reduction, which in turn may lead to more stable estimates since the number of respondents in each class is likely to be fairly large.

Turning to the RRMSE of the estimator, we note that in the case of the $\hat{y}$ -criterion and Mechanism 1 (see Figure 7.2), RRMSE rapidly decreases (because the bias rapidly decreases) and quickly stabilizes (around 1.5%) starting at $C = 20$ classes. For the Mechanism 2, the behaviour of RRMSE is very close to the behaviour of relative bias (RB) in Figure 7.1 because of the presence of substantial bias. In the case of the $\hat{p}$ -criterion and Mechanism 1 (see Figure 7.4), we note that RRMSE decreases rapidly (between 1 and 4 classes) but increases for $C \geq 5$ and becomes stable for $C \geq 15$ with a value of 9.5% approximately. Hence, although the use of the $\hat{p}$ -criterion is effective in reducing the bias, there is no guarantee that it will be effective in reducing the variance of the imputed estimator. On the other hand, the use of the $\hat{y}$ -criterion is effective in reducing both the bias and the variance of the imputed estimator. This

is consistent with the variance expressions (7.9) and (7.13).

When both the $\hat{y}$ -criterion and the $\hat{p}$ -criterion are simultaneously used (using the classification algorithm) in forming the classes (see Figure 7.5), the relative bias of the imputed estimator stabilizes rapidly (around 10 classes for Mechanism 1). From Figure 7.6, we note that RRMSE of the imputed estimator stabilizes around 4%. Using both scores in the formation of the classes is thus a compromise between the $\hat{y}$ -criterion and the $\hat{p}$ -criterion used separately in terms of both relative bias and RRMSE. However, as we note below, using both scores ensures some degree of robustness when either the model for the variable of interest or the model for the response probabilities is incorrectly specified.

Turning to the method of sample partitioning (equal quantile method vs classification algorithm), we note in Figure 7.7 that although, the classification algorithm presents a slight advantage over the equal quantile method for $C \leq 20$, the two methods give very similar results when the number of classes is large. This result is not surprising since the classification algorithm is specifically designed to construct homogeneous classes even if the number of classes is small, whereas the equal quantile method may lead to somewhat heterogeneous classes in this case.

As a measure of homogeneity we used the coefficient of determination (R^2). From Figure 7.8, we can once again note that the classification algorithm does better than the equal quantile method when the number of classes is small. In practice, the coefficient may be useful for determining the number of classes to be used.

When the model for the variable of interest is incorrectly specified for obtaining the score $\hat{y}_i$ (the variables z_1 and z_1^2 were omitted from the imputation model), then

the imputed estimator is heavily biased for both mechanisms as shown in Figure 7.9 (around 14.5 % for Mechanism 1 and 29 % for Mechanism 2) when the $\hat{y}$ -criterion is used. However, when both scores are used and the nonresponse model is correctly specified, Figure 7.10 shows that as the number of classes increase, the relative bias of the imputed estimator decreases. However, in this case, a large number of classes is needed to obtain an effective bias reduction (more than 50 classes) unlike in the case of one score.

Finally, when all the variables are included in the model $(z_1, z_2, z_1^2, z_3, z_4, z_5, z_6)$, the results (not shown here) were very similar to those of Figure 7.1-7.4. In other words, the inclusion of the redundant variables (z_3, z_4, z_5, z_6) in the imputation model had no significant impact on both the relative bias of the imputed estimator and its RRMSE.

7.8.2 The cross-classification method

Results for the cross-classification are reported in Tables 7.1 and 7.2. First, let us discuss the case of the response mechanism 1. From Table 7.1, note that when the ordering of the variable is correctly specified, the relative bias estimator is equal to 1.30% for $n_{min} = 1$ and 1.18% for $n_{min} = 9$. Also, the RRMSE of the imputed estimator is equal to 1.71% for $n_{min} = 1$ and 1.60% for $n_{min} = 9$. When the ordering of the variables is incorrectly specified, the imputed estimator is biased. Indeed, for the incorrect ordering $(z_{6c}, z_{5c}, z_{4c}, z_{3c}, z_{2c}, z_{1c})$, relative bias is approximately equal to 14.8% for $n_{min} = 1$. For the incorrect ordering $(z_{2c}, z_{3c}, z_{1c}, z_{4c}, z_{5c}, z_{6c})$ the relative

bias is also, approximately equal to 14.8% for $n_{min} = 1$. The bias can easily be explained by the fact that, when the ordering of the variables is incorrectly specified, the important variables (z_{1c} and z_{2c}) are dropped at an early stage of the process. For the response mechanism 2 (see Table 7.2), we note that the imputed estimator is heavily biased in all scenarios (20% - 30%). The bias is larger when the ordering is incorrectly specified (around 29% for $n_{min} = 1$).

7.8.3 Comparison of the methods

In this section, we give a point by point comparison between the score method and the cross-classification method:

- The score method requires a relatively small number of classes ($C \leq 25$) for reducing the bias to a maximum extent, whereas the cross-classification method usually requires a much larger number of classes. As a result, the number of donors per class under the cross-classification method may be very small, potentially leading to unstable estimators.
- The cross-classification method requires a proper ordering of the variables to be cross-classified whereas the score method does not require such an ordering.
- In terms of RRMSE, results not reported here have shown that the score method leads to a smaller RRMSE than the cross-classification method.
- The score method facilitates variance estimation because it leads to disjoint imputation classes. As a result, any variance estimation method (see section

7.3.5) can be used in a somewhat straightforward manner. In the case of the cross-classification method, application of the available methods is not straightforward because of the collapsing of the classes at each stage of the process. In this case, the bootstrap method may be used but in each bootstrap sample, the whole process (formation of classes and imputation) needs to be performed.

Overall, the score method presents several advantages over the cross-classification method. In practice, the score method allows for more flexibility (in determining the number of classes, for example).

Figure 7.1: Relative bias of the imputed estimator using the classification algorithm and $\hat{y}$ -criterion

Figure 7.2: RRMSE of the imputed estimator using the classification algorithm and $\hat{y}$ -criterion

Figure 7.3: Relative bias of the imputed estimator using the classification algorithm and $\hat{p}$ -criterion

Figure 7.4: RRMSE of the imputed estimator using the classification algorithm and $\hat{p}$ -criterion

Figure 7.5: Relative bias of the imputed estimator using the classification algorithm under both $\hat{y}$ -criterion and $\hat{p}$ -criterion

Figure 7.6: RRMSE of the imputed estimator using the classification algorithm under both $\hat{y}$ -criterion and $\hat{p}$ -criterion

Figure 7.7: Relative bias of the imputed estimator: equal quantile method vs classification algorithm under the $\hat{y}$ -criterion

Figure 7.8: R^2 of the imputed estimator for the first response mechanism: equal quantile method vs classification score

Figure 7.9: Relative bias of the imputed estimator using the classification algorithm

Figure 7.10: Relative bias of the imputed estimator using the classification algorithm

Table 7.1: Relative bias and RRMSE of the imputed estimator under the response mechanism 1

Ordering of the variables	More donors than recipients	Minimum number of donors per class	Relative bias of the imputed estimator	RRMSE of the imputed estimator
$z_{1c}z_{2c}z_{3c}z_{4c}z_{5c}z_{6c}$	yes	1	1.31	1.71
	yes	9	1.19	1.61
	no	1	1.41	1.80
	no	9	1.05	1.41
$z_{6c}z_{5c}z_{4c}z_{3c}z_{2c}z_{1c}$	yes	1	14.81	14.89
	yes	9	14.37	14.45
	no	1	13.99	14.05
	no	9	14.52	14.59
$z_{2c}z_{3c}z_{1c}z_{4c}z_{5c}z_{6c}$	yes	1	15.03	15.11
	yes	9	16.24	16.32
	no	1	14.81	14.90
	no	9	19.35	19.43

Table 7.2: Relative bias and RRMSE of the imputed estimator under the response mechanism 2

Ordering of the variables	More donors than recipients	Minimum number of donors per class	Relative bias of the imputed estimator	RRMSE of the imputed estimator
$z_{1c}z_{2c}z_{3c}z_{4c}z_{5c}z_{6c}$	yes	1	20.57	20.58
	yes	9	20.95	20.96
	no	1	21.17	21.18
	no	9	19.81	19.82
$z_{6c}z_{5c}z_{4c}z_{3c}z_{2c}z_{1c}$	yes	1	29.73	29.76
	yes	9	29.09	29.12
	no	1	29.08	29.11
	no	9	29.20	29.23
$z_{2c}z_{3c}z_{1c}z_{4c}z_{5c}z_{6c}$	yes	1	26.12	26.15
	yes	9	29.25	29.28
	no	1	22.90	22.92
	no	9	30.40	30.43

Chapter 8

Some asymptotic results

8.1 Introduction

In this chapter, we investigate the asymptotic properties of some estimators studied in this dissertation. To this end, we define a sequence of population and an associated sequence of samples. Let $\{u_j\}$ denote a sequence of elements. Let $\{U_t\}$ denote a sequence of finite populations of size $\{N_t\}$, where $0 < N_1 < N_2 < N_3 < \dots$. That is, U_1 is composed of the first N_1 elements of $\{u_j\}$, $U_2 \supset U_1$ is composed of the first elements of $\{u_j\}$, and so on. Let a sequence of samples $\{s_t\}$ of size $\{n_t\}$ be created from the sequence of populations by a sequence of designs such that s_1 is composed of n_1 distinct elements selected from U_1 , s_2 is composed of n_2 distinct elements selected from U_2 , and so on, where $n_1 < n_2 < n_3, \dots$, and $n_t < N_t$ for all t . For simplicity, the index t will be suppressed in what follows and all limiting processes will be understood to be as $N \rightarrow \infty, n \rightarrow \infty$ and $N - n \rightarrow \infty$. In the complete response case, Isaki and Fuller (1982) gave sufficient conditions for the Horvitz-Thompson estimator $\bar{y}_{HT}$ of

the mean $\bar{Y}$ to be design consistent.

Lemma 8.1 *Let a sequence of populations and samples be as described. Let*

$$\bar{y}_{HT} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i}$$

and

$$\Delta_{ij} = \begin{cases} \pi_i \pi_j - \pi_{ij} & \text{if } \pi_i \pi_j - \pi_{ij} \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Assume that $\frac{1}{N} \sum_{i \in U} \left[\frac{N}{n} (y_i - \bar{Y}) \right]^{2k} < M < \infty$, for some constant M and that the inclusion probabilities satisfy

$$\frac{1}{N^2} \sum_{i \in U} \sum_{i \neq j, j \in U} \Delta_{ij}^r = O(n^{-2r\delta}),$$

for $\delta > 0$, where $r^{-1} + k^{-1} = 1$. Then, for $\delta = 1/2$,

$$\bar{y}_{HT} - \bar{Y} = O_p(n^{-1/2}).$$

We now state a useful lemma.

Lemma 8.2 (Fuller, 1996) *Let $\{X_n\}$ be a sequence of random variables and let $\{h_n\}$ a sequence of real numbers such that*

$$E([X_n - E(X_n)]^2) = O(h_n)$$

and $E(X_n) = O(\sqrt{h_n})$

Then,

$$X_n = O_p(\sqrt{h_n}).$$

8.2 Consistency of some imputed estimators

In this section, we establish the asymptotic consistency of some imputed estimators under weighted ratio imputation. We assume the following regularity conditions:

C1: $\max_i \left(\frac{n}{N} w_i \right) = O(1)$.

C2: $\frac{n}{N^2} V_p(\hat{Y}) = O(1)$, where $\hat{Y} = \sum_{i \in s} w_i y_i$ denotes the estimator of Y under complete response with $w_i = \frac{1}{\pi_i}$.

C3: The response probabilities satisfy

$$K_1 < p_i,$$

for some nonnegative constant K_1 .

C4: $P(a_i = 1, a_j = 1) = P(a_i = 1) P(a_j = 1), \forall i \neq j$.

Condition C1 says that no unit contribution, w_i , is of disproportionate size. Condition C2 says that the variance of the estimator under complete response is bounded when multiplied by the factor $\frac{n}{N^2}$. Condition C3 says that the response probabilities are not “too” small. Finally, condition C4 says that the units respond independently of one another.

Under weighted ratio imputation, the imputed estimator of a total $Y = \sum_{i \in U} y_i$ is given by

$$\hat{Y}_I = \hat{B}_r \hat{Z}, \quad (8.1)$$

where $\hat{B}_r = \frac{\tilde{Y}_r}{\tilde{Z}_r}$ with $(\tilde{Y}_r, \tilde{Z}_r) = p^{-1} \sum_{i \in s} w_i a_i (y_i, z_i)$ and $\hat{Z} = \sum_{i \in s} w_i z_i$.

Note that $\hat{Y}_I$ in (8.1) can be alternatively written as

$$\hat{Y}_I = \tilde{Y}_r + \left(\hat{Z} - \tilde{Z}_r \right) \hat{B}_r. \quad (8.2)$$

It follows that the total error of $\hat{Y}_I, \tilde{Y}_I - Y$, can be written as

$$\hat{Y}_I - Y = (\tilde{Y}_r - Y) + (\hat{Z} - \tilde{Z}_r) B + (\hat{Z} - \tilde{Z}_r) (\hat{B}_r - B), \quad (8.3)$$

where $B = \sum_{i \in U} y_i / \sum_{i \in U} z_i$. Note that $E(\tilde{Y}_r, \tilde{Z}_r) = E_p E_r[(\tilde{Y}_r, \tilde{Z}_r) | s] = (Y, Z)$.

First, using condition C4, we have

$$\begin{aligned} E(\tilde{Y}_r - Y)^2 &= E_p E_r \left\{ (\tilde{Y}_r - Y)^2 | s \right\} \\ &= V_p E_r(\tilde{Y}_r | s) + E_p V_r(\tilde{Y}_r | s) \\ &= V_p(\hat{Y}) + \frac{(1-p)}{p} E_p \left(\sum_{i \in s} w_i^2 y_i^2 \right) \\ &\quad V_p(\hat{Y}) + \frac{(1-p)}{p} \sum_{i \in U} w_i y_i^2 \\ &= O\left(\frac{N^2}{n}\right), \end{aligned}$$

where the last equality follows from C1, C2 and C3. Hence, by Lemma 8.2, we have

$$\frac{1}{N} (\tilde{Y}_r - Y) = O_p(n^{-1/2}). \quad (8.4)$$

Now, decompose the term $(\hat{Z} - \tilde{Z}_r)$ as

$$\hat{Z} - \tilde{Z}_r = (\hat{Z} - Z) - (\tilde{Z}_r - Z).$$

Following the same argument used for $(\tilde{Y}_r - Y)$, we have

$$\frac{1}{N} (\tilde{Z}_r - Z) = O_p(n^{-1/2}).$$

Also, it follows from Lemma 8.1 that

$$\frac{1}{N} (\hat{Z} - Z) = O_p(n^{-1/2}).$$

Hence,

$$\frac{1}{N} \left(\hat{Z} - \tilde{Z}_r \right) = O_p \left(n^{-1/2} \right). \quad (8.5)$$

Finally, by (8.4) and (8.5), we obtain, using Taylor linearization,

$$\hat{B}_r - B = O_p \left(n^{-1/2} \right). \quad (8.6)$$

Inserting (8.4), (8.5) and (8.6) in (8.3), it follows that

$$\frac{1}{N} \left(\hat{Y}_I - Y \right) = O_p \left(n^{-1/2} \right).$$

That is, $\hat{Y}_I$ is a consistent estimator for Y .

In Chapter 3, we proposed a bias-adjusted estimator, $\hat{Y}_{I(d)}^a$, which is approximately unbiased for the vector of domain totals $\mathbf{Y}_{(d)}$ under both assumption DB and assumption MB. We now show the consistency of the elements of $\hat{Y}_{I(d)}^a$ under weighted ratio imputation. First, note that under weighted ratio imputation, the j -th component of $\hat{Y}_{I(d)}^a$ is given by

$$\hat{Y}_{Ij}^a = \hat{p}^{-1} \hat{Y}_{Ij} + (1 - \hat{p}^{-1}) \frac{\hat{Z}_j}{\hat{Z}} \hat{Y}_I, \quad (8.7)$$

where

$$\hat{Y}_{Ij} = \sum_{i \in s} w_i x_{ji} \tilde{y}_i$$

with $x_{ji} = 1$ if unit i is in domain j and $x_{ji} = 0$, otherwise and $\tilde{y}_i = y_i$ if unit i responded to item y and $\tilde{y}_i = y_i^*$ otherwise,

$$\hat{Y}_I = \sum_{i \in s} w_i \tilde{y}_i$$

and $\hat{Z}_j = \sum_{i \in s} w_i x_{ji} z_i$. First, note that $\hat{Y}_{Ij}^a$ in (8.7) can be written as

$$\hat{Y}_{Ij}^a = \hat{Y}_{jr} + (\hat{Z}_j - \hat{Z}_{jr}) \hat{B}_r, \quad (8.8)$$

where $(\hat{Y}_{jr}, \hat{Z}_{jr}) = \hat{p}^{-1} \sum_{i \in s} w_i a_i x_{ji} (y_i, z_i)$. We can write the total error of $\hat{Y}_{Ij}^a, \hat{Y}_{Ij}^a - Y_j$, as

$$\hat{Y}_{Ij}^a - Y_j = (\hat{Y}_{jr} - Y_j) + (\hat{Z}_j - \hat{Z}_{jr}) B + (\hat{Z}_j - \hat{Z}_{jr}) (\hat{B}_r - B). \quad (8.9)$$

First, note that

$$\hat{Y}_{jr} - Y_j = (\hat{Y}_{jr} - \tilde{Y}_{jr}) + (\tilde{Y}_{jr} - Y_j),$$

where $\tilde{Y}_{jr} = p^{-1} \sum_{i \in s} w_i a_i x_{ji} y_i$. Further, using condition C4, we have

$$\begin{aligned} E(\tilde{Y}_{jr} - Y_j)^2 &= V_p E_r(\tilde{Y}_{jr} | s) + E_p V_r(\tilde{Y}_{jr} | s) \\ &= V_p(\hat{Y}_j) + \frac{(1-p)}{p} E_p \left(\sum_{i \in s} w_i^2 x_{ji} y_i^2 \right) \\ &= V_p(\hat{Y}_j) + \frac{(1-p)}{p} \sum_{i \in s} w_i x_{ji} y_i^2 \\ &= O(n^{-1} N^2), \end{aligned}$$

where $\hat{Y}_j = \sum_{i \in s} w_i x_{ji} y_i$ and the last equality follows from C1, C2 (with y_i replaced by $x_{ji} y_i$) and C3. By Lemma 8.2, we have

$$\frac{1}{N} (\tilde{Y}_{jr} - Y_j) = O_p(n^{-1/2}). \quad (8.10)$$

Now, note that

$$\hat{Y}_{jr} - \tilde{Y}_{jr} = (\hat{p}^{-1} - p^{-1}) \sum_{i \in s} w_i a_i x_{ji} y_i. \quad (8.11)$$

From Lemma 8.1, we have

$$\frac{1}{N} (\hat{N} - N) = O_p(n^{-1/2}), \quad (8.12)$$

where $\hat{N} = \sum_{i \in s} w_i$. Also, let $\hat{R} = \sum_{i \in s} w_i a_i$. We have, noting that $E(\hat{R}) = Np$,

$$\begin{aligned} E(\hat{R} - Np)^2 &= V_p E_r(\hat{R}|s) + E_p V_r(\hat{R}|s) \\ &= p^2 V_p(\hat{N}) + p(1-p) E_p \left(\sum_{i \in s} w_i^2 \right) \\ &= p^2 V_p(\hat{N}) + p(1-p) \hat{N} \\ &= O\left(\frac{N^2}{n}\right), \end{aligned}$$

where the last equality follows from conditions C1 and C2 (with y_i replaced by 1).

Hence, applying Lemma 8.2, we obtain

$$\frac{1}{N} (\hat{R} - Np) = O_p(n^{-1/2}). \quad (8.13)$$

Using (8.12) and (8.13), we can apply Taylor linearization to obtain

$$\hat{p}^{-1} - p^{-1} = O_p(n^{-1/2}). \quad (8.14)$$

Assuming that $\frac{1}{N} \sum_{i \in U} w_i y_i = O(1)$, we have under condition C1, $\frac{1}{N} \sum_{i \in s} w_i a_i x_{ji} y_i = O_p(1)$. It follows that

$$\frac{1}{N} (\hat{Y}_{jr} - \tilde{Y}_{jr}) = O_p(n^{-1/2}). \quad (8.15)$$

Hence, from (8.10), (8.11), (8.14) and (8.15), we obtain

$$\frac{1}{N} (\hat{Y}_{jr} - Y_j) = O_p(n^{-1/2}). \quad (8.16)$$

Now, consider the term $(\hat{Z}_j - \hat{Z}_{jr})$ that can be decomposed as

$$\hat{Z}_j - \hat{Z}_{jr} = (\hat{Z}_j - Z_j) - (\hat{Z}_{jr} - Z_j).$$

It follows from Lemma 8.1 (with y_i replaced by $x_{ji}y_i$) that

$$\frac{1}{N} \left(\hat{Z}_j - Z_j \right) = O_p \left(n^{-1/2} \right). \quad (8.17)$$

Using the same arguments as for $\hat{Y}_{jr} - Y_j$, we have

$$\frac{1}{N} \left(\hat{Z}_{jr} - Z_j \right) = O_p \left(n^{-1/2} \right). \quad (8.18)$$

Hence, from (8.17) and (8.18), we obtain

$$\frac{1}{N} \left(\hat{Z}_j - \hat{Z}_{jr} \right) = O_p \left(n^{-1/2} \right). \quad (8.19)$$

Inserting (8.16), (8.19) and (8.6) in (8.9), we obtain

$$\frac{1}{N} \left(\hat{Y}_{Ij}^a - Y_j \right) = O_p \left(n^{-1/2} \right).$$

In Chapter 4, we proposed a modified imputation method that leads to robust imputed estimators under IM or GNM. In the case of modified ratio imputation, assuming that the response probabilities p_i are known for simplicity, we use the imputed values

$$y_i^* = \frac{\tilde{Y}_{rp}}{\tilde{Z}_{rp}} z_i,$$

where $(\tilde{Y}_{rp}, \tilde{Z}_{rp}) = \sum_{i \in s} w_i a_i \frac{(1-p_i)}{p_i} (y_i, z_i)$. The resulting imputed estimator is given by

$$\tilde{Y}_{Ip} = \sum_{i \in s} w_i a_i y_i + \sum_{i \in s} w_i (1 - a_i) z_i \frac{\tilde{Y}_{rp}}{\tilde{Z}_{rp}}. \quad (8.20)$$

The estimator $\tilde{Y}_{Ip}$ in (8.20) can be expressed as

$$\tilde{Y}_{Ip} = \tilde{Y}_p + \frac{\tilde{Y}_{rp}}{\tilde{Z}_{rp}} \left(\hat{Z} - \tilde{Z}_p \right),$$

where $(\tilde{Y}_p, \tilde{Z}_p) = \sum_{i \in s} \frac{w_i}{p_i} a_i(y_i, z_i)$. The total error of $\tilde{Y}_{Ip}, \tilde{Y}_{Ip} - Y$, is thus given by

$$\tilde{Y}_{Ip} - Y = (\tilde{Y}_p - Y) + (\hat{Z} - \tilde{Z}_p) \frac{Y_p}{Z_p} + (\hat{Z} - \tilde{Z}_p) \left(\frac{\tilde{Y}_{rp}}{\tilde{Z}_{rp}} - \frac{Y_p}{Z_p} \right), \quad (8.21)$$

where $(Y_p, Z_p) = \sum_{i \in U} (1 - p_i) (y_i, z_i)$.

First, using condition C4, we have

$$\begin{aligned} E \left(\tilde{Y}_p - Y \right)^2 &= V_p E_r \left(\tilde{Y}_p | s \right) + E_p V_r \left(\tilde{Y}_p | s \right) \\ &= V_p \left(\hat{Y} \right) + E_p \left[\sum_{i \in s} w_i^2 \frac{(1 - p_i)}{p_i} y_i^2 \right] \\ &= V_p \left(\hat{Y} \right) + \sum_{i \in U} w_i \frac{(1 - p_i)}{p_i} y_i^2 \\ &= O \left(\frac{N^2}{n} \right), \end{aligned}$$

where the last equality follows from conditions C1, C2 and C3. Hence, from Lemma 8.2, it follows that

$$\frac{1}{N} \left(\tilde{Y}_p - Y \right) = O_p \left(n^{-1/2} \right). \quad (8.22)$$

Now, the term $\hat{Z} - \tilde{Z}_p$ may be written as

$$\hat{Z} - \tilde{Z}_p = (\hat{Z} - Z) - (\tilde{Z}_p - Z).$$

From Lemma 8.1, we have

$$\frac{1}{N} \left(\hat{Z} - Z \right) = O_p \left(n^{-1/2} \right).$$

Also, following the proof of (8.22), we obtain

$$\frac{1}{N} \left(\tilde{Z}_p - Z \right) = O_p \left(n^{-1/2} \right). \quad (8.23)$$

Hence,

$$\frac{1}{N} \left(\hat{Z} - \tilde{Z}_p \right) = O_p \left(n^{-1/2} \right). \quad (8.24)$$

Now, note that $(\tilde{Y}_{rp}, \tilde{Z}_{rp})$ may be expressed as

$$(\tilde{Y}_{rp}, \tilde{Z}_{rp}) = \left((\tilde{Y}_p - \hat{Y}_a), (\tilde{Z}_p - \hat{Z}_a) \right),$$

where $(\hat{Y}_a, \hat{Z}_a) = \sum_{i \in s} w_i a_i z_i$. Let $(Y_a, Z_a) = \sum_{i \in U} a_i (y_i, z_i)$. Noting that

$$\frac{1}{N} \left(\hat{Y}_a - Y_a, \hat{Z}_a - Z_a \right) = O_p \left(n^{-1/2} \right),$$

it follows from (8.22), (8.23) that, applying Taylor linearization,

$$\frac{\tilde{Y}_{rp}}{\tilde{Z}_{rp}} - \frac{Y_p}{Z_p} = O_p \left(n^{-1/2} \right). \quad (8.25)$$

Hence, the last term of (8.21) is of order $O_p(n^{-1})$. It now follows from (8.22) (8.24)

and (8.25) that

$$\frac{1}{N} \left(\tilde{Y}_{rp} - Y \right) = O_p \left(n^{-1/2} \right).$$

8.3 Asymptotic normality under ratio imputation

In this section we establish the asymptotic normality in the context of stratified multistage designs of an imputed estimator under both deterministic weighted ratio imputation and random weighted ratio imputation. We will use the following useful lemmas.

Lemma 8.3 (*Central Limit Theorem*) Suppose $\{X_t\}_{t=1}^T$ are independent nonidentically distributed random variables with $E(X_t) = \mu_t$ and $V(X_t) = \sigma_t^2$. If $\frac{1}{T} \sum_{t=1}^T \sigma_t^2 \rightarrow$

σ^2 and $\frac{1}{T} \sum_{t=1}^T E |X_t - \mu_t|^{2+\delta} = O(1)$ as $T \rightarrow \infty$ for some $\delta > 0$

then

$$\sqrt{T} (\bar{X} - \bar{\mu}) \rightarrow_d N(0, \sigma^2)$$

where $\bar{X} = \frac{1}{T} \sum_{t=1}^T X_t$ and $\bar{\mu} = \frac{1}{T} \sum_{t=1}^T \mu_t$.

Lemma 8.4 (*Slutsky's Theorem*) Suppose that $\{X_n\}, \{Y_n\}$ and X are random variables and c is a constant. If $X_n \rightarrow_d X$ and $Y_n \rightarrow_p c$, then

$$(i) \quad X_n + Y_n \rightarrow_d X + c$$

$$(ii) \quad X_n Y_n \rightarrow_d cX.$$

Lemma 8.5 (*Chen and Rao, 2004*) Let $\{U_n\}$ and $\{V_n\}$ two sequences of random variables and $\mathfrak{S}_n$ be a σ -finite algebra. Assume that

1. there exists $\sigma_{1n} > 0$ such that

$$\sigma_{1n}^{-1} V_n \rightarrow_d N(0, 1)$$

as $n \rightarrow \infty$, and V_n is $\mathfrak{S}_n$ measurable.

2. $E(U_n | \mathfrak{S}_n) = 0$ and $V(U_n | \mathfrak{S}_n) = \sigma_{2n}^2$ such that

$$\sup_t |P(\sigma_{2n}^{-1} U_n \leq t | \mathfrak{S}_n) - \Phi(t)| = o_p(1),$$

where $\Phi(\cdot)$ is the distribution function of the standard normal random variable.

3. $\gamma_n^2 = \sigma_{1n}^2 / \sigma_{2n}^2 = \gamma^2 + o_p(1)$

Then, as $n \rightarrow \infty$

$$\frac{U_n + V_n}{\sqrt{\sigma_{1n}^2 + \sigma_{2n}^2}} \rightarrow_d N(0, 1).$$

8.3.1 Deterministic weighted ratio imputation

In what follows, recall that K denotes the population size in the case of stratified multistage sampling. Under deterministic weighted ratio imputation, we have

$$\bar{y}_I = \hat{B}_r \bar{z},$$

where $\hat{B}_r = \frac{\hat{Y}_a}{\hat{Z}_a}$ with $(\hat{Y}_a, \hat{Z}_a) = \sum_{(hij) \in s} \tilde{w}_{hij} a_{hij} (y_{hij}, z_{hij})$ and

$\bar{z} = \sum_{(hij) \in s} \tilde{w}_{hij} z_{hij} / \sum_{(hij) \in s} \tilde{w}_{hij}$ and $\tilde{w}_{hij} = \frac{w_{hij}}{K}$ denotes the standardized weights for unit (hij) . We can decompose the total error of $\bar{y}_I, \bar{y}_I - \bar{Y}$, as

$$\begin{aligned} \bar{y}_I - \bar{Y} &= (\bar{y}_I - B_r \bar{Z}) + (B_r \bar{Z} - \bar{Y}), \\ &\equiv U_n + V_n, \end{aligned}$$

where $B_r = \bar{Y}_a / \bar{Z}_a$ with $(\bar{Y}_a, \bar{Z}_a) = \frac{1}{K} \sum_{(hij) \in U} a_{hij} (y_{hij}, z_{hij})$ and $\bar{Z} = \sum_{(hij) \in U} z_{hij} / K$.

We use Fay's reverse approach and first show that, given the response indicators a_{hij} , $(\bar{y}_I - B_r \bar{Z})$ is asymptotically normally distributed. Let $\mathfrak{S}_K = \sigma(a_{hij}, (hij) \in U)$.

First, using Taylor linearization, we have

$$U_n = \bar{y}_I - B_r \bar{Z} = \sum_{(hij) \in s} \tilde{w}_{hij} e_{hij} + O_p(n^{-1}), \quad (8.26)$$

where $e_{hij} = a_{hij} \frac{\bar{Z}}{\bar{Z}_a} (y_{hij} - B_r z_{hij}) + B_r (z_{hij} - \bar{Z})$. Note that given the response indicators a_{hij} , we have

$$E_p(\bar{e}) = 0,$$

where $\bar{e} = \sum_{(hij) \in s} \tilde{w}_{hij} e_{hij}$. We assume the following regularity conditions:

C1: $n^{1+\delta} \sum_h \sum_i E_p |e_{hi} - E_p(e_{hi})|^{2+\delta} = O(1)$ for some $\delta > 0$ and $e_{hi} = \sum_j \tilde{w}_{hij} e_{hij}$.

C2: $nV_p(\bar{e}) \rightarrow \sigma_e^2$, say.

Condition C1 is a standard Liapunov-type condition on the $2 + \delta$ moments used in establishing a central limit theorem for independent nonidentically distributed random variables. Condition C2 assumes that the limit of the variance of $\bar{e}$ exists when multiplied by the normalized factor n . Let

$$X_{hi} = n(e_{hi} - E_p(e_{hi})),$$

where $e_{hi} = \sum_j \tilde{w}_{hij} e_{hij}$. Under condition C2, we have

$$\frac{1}{n} \sum_h \sum_i V_p(X_{hi}) = n \sum_h \sum_i V_p(e_{hi}) = nV_p(\bar{e}) \rightarrow \sigma_e^2.$$

Also, under condition C1, we have

$$\frac{1}{n} \sum_h \sum_i E_p |X_{hi}|^{2+\delta} = n^{1+\delta} \sum_h \sum_i E_p |e_{hi} - E_p(e_{hi})|^{2+\delta} = O(1),$$

satisfying the conditions of the Central Limit Theorem in Lemma 8.3. Therefore, applying Lemma 8.4 to (8.26), we obtain, conditionally given $\mathfrak{S}_K$,

$$\frac{1}{\sigma_{2n}} U_n \rightarrow_d N(0, 1),$$

where $\sigma_{2n}^2 = V_p(\bar{y}_I | \mathfrak{S}_K)$.

We now show that the component $B_r \bar{Z} - \bar{Y}$ satisfies the second condition of Lemma 8.5. Using Taylor linearization, we obtain

$$V_n = \frac{\bar{Y}_r}{\bar{Z}_r} \bar{Z} - \bar{Y} = \frac{1}{pK} \sum_{(hij) \in U} a_{hij} (y_{hij} - Bz_{hij}) + o_p(N^{-1/2}).$$

Noting that $X_{hij} = a_{hij} (y_{hij} - Bz_{hij})$ are independent random variables, it follows from Lemma 8.3 that

$$\frac{V_n}{\sigma_{1n}} \rightarrow_d N(0, 1),$$

where $\sigma_{1n}^2 = V_r(B_r \bar{Z})$. Hence, by Lemma 8.5,

$$\frac{U_n + V_n}{\sqrt{\sigma_{1n}^2 + \sigma_{2n}^2}} \rightarrow_d N(0, 1)$$

or $\bar{y}_I$ is asymptotically normal.

8.3.2 Asymptotic normality under random ratio imputation

Under random weighted ratio imputation, the imputed estimator of the mean $\bar{Y}$ is given by

$$\bar{y}_I = \sum_{(hij) \in s} \tilde{w}_{hij} a_{hij} y_{hij} + \sum_{(hij) \in s} \tilde{w}_{hij} (1 - a_{hij}) \left(\hat{B}_r z_{hij} + \sqrt{z_{hij}} \epsilon_{hij}^* \right),$$

where $\epsilon_{hij}^* = e_{hij}^* - \bar{e}_r$ such that $P(e_{hij}^* = e_{glk}) = \frac{w_{glk}}{\sum_{(hij) \in s} w_{hij} a_{hij}}$, $(glk) \in s_r$ with $e_{glk} = \frac{1}{\sqrt{z_{glk}}} (y_{glk} - \hat{B}_r z_{glk})$ and $\bar{e}_r = \sum_{(hij) \in s} w_{hij} a_{hij} e_{hij} / \sum_{(hij) \in s} w_{hij} a_{hij}$. In this case, the total error of $\bar{y}_I, \bar{y}_I - \bar{Y}$, can be decomposed as

$$\bar{y}_I - \bar{Y} = [\bar{y}_I - E_*(\bar{y}_I)] + [E_*(\bar{y}_I) - B_r \bar{Z}] + [B_r \bar{Z} - \bar{Y}]. \quad (8.27)$$

Let $\mathfrak{S}_n = \sigma((\delta_{hij}, a_{hij}), (hij) \in s)$, where δ_{hij} denotes the sample selection indicator for unit (hij) . Note that since $E_*(\epsilon_{hij}^*) = 0$, we have $E_*(\bar{y}_I) = \hat{B}_r \bar{z}$ which coincides with the imputed estimators one would obtain by performing deterministic weighted ratio imputation. Hence, we can use the result in section 8.3.1, to conclude that the sum of the last two terms on the right hand side of (8.27) is asymptotically normally distributed.

Consider now the term $\bar{y}_I - E_*(\bar{y}_I)$ on the right hand side of (8.27). First, note

that

$$\begin{aligned}\bar{y}_I - E_*(\bar{y}_I) &= \sum_{(hij) \in s} \tilde{w}_{hij} (1 - a_{hij}) \sqrt{z_{hij}} \epsilon_{hij}^* \\ &= \sum_{(hij) \in s} \xi_{hij}^*,\end{aligned}$$

where $\xi_{hij}^* = \tilde{w}_{hij} (1 - a_{hij}) \sqrt{z_{hij}} \epsilon_{hij}^*$, $(hij) \in s$. In addition to C1 and C2, we assume the following regularity conditions:

$$\text{C3: } n \max_{h,i} \sum_j \tilde{w}_{hij} \sqrt{z_{hij}} = O_p(1).$$

$$\text{C4: } \sum_{(hij) \in s} \tilde{w}_{hij} |e_{hij} - \bar{E}|^{2+\delta} = O_p(1),$$

where $\bar{E} = \frac{1}{K} \sum_{(hij) \in U} E_{hij}$ with $E_{hij} = \frac{1}{\sqrt{z_{hij}}} (y_{hij} - B z_{hij})$ and $B = Y/Z$. Let

$$t_{hi}^* = \sum_j \xi_{hij}^* = \sum_j \tilde{w}_{hij} (1 - a_{hij}) \sqrt{z_{hij}} \epsilon_{hij}^*.$$

We have

$$\begin{aligned}E_* |t_{hi}^*|^{2+\delta} &= \left[\sum_{(glk) \in s} \tilde{w}_{glk} a_{glk} \left| \sum_j \tilde{w}_{hij} (1 - a_{hij}) \sqrt{z_{hij}} (e_{glk} - \bar{e}_r) \right|^{2+\delta} \right] \left[\sum_{(glk) \in s} \tilde{w}_{glk} a_{glk} \right]^{-1} \\ &= \left[\sum_{(glk) \in s} \tilde{w}_{glk} a_{glk} |e_{glk} - \bar{e}_r|^{2+\delta} \left| \sum_j \tilde{w}_{hij} (1 - a_{hij}) \sqrt{z_{hij}} \right|^{2+\delta} \right] \left[\sum_{(glk) \in s} \tilde{w}_{glk} a_{glk} \right]^{-1} \\ &\leq \left[\sum_{(glk) \in s} \tilde{w}_{hij} |e_{glk} - \bar{e}_r|^{2+\delta} \left| \sum_j \tilde{w}_{hij} \sqrt{z_{hij}} \right|^{2+\delta} \right] \left[\sum_{(glk) \in s} \tilde{w}_{glk} a_{glk} \right]^{-1} \quad (8.28)\end{aligned}$$

since $a_{hij} = 0$ or 1 and $\tilde{w}_{hij} \geq 0$.

Now, we have

$$\begin{aligned}\sum_{(glk) \in s} \tilde{w}_{glk} |e_{glk} - \bar{e}_r|^{2+\delta} &= \sum_{(glk) \in s} \tilde{w}_{glk} |e_{glk} - \bar{E} + \bar{E} - \bar{e}_r|^{2+\delta} \\ &\leq 2^{1+\delta} \left[\sum_{(glk) \in s} \tilde{w}_{glk} |e_{glk} - \bar{E}|^{2+\delta} \right. \\ &\quad \left. + \sum_{(glk) \in s} \tilde{w}_{glk} |\bar{E} - \bar{e}_r|^{2+\delta} \right], \quad (8.29)\end{aligned}$$

where the last inequality follows from the so-called C_r inequality (c.f., Sen and Singer, 1983, p.21), namely

for every $r > 0$, a and b real numbers,

$$|a + b|^r \leq C_r (|a|^r + |b|^r),$$

where

$$C_r = \begin{cases} 1 & \text{if } 0 \leq r \leq 1 \\ 2^{r-1} & \text{if } r > 1. \end{cases}$$

Now, noting that

$$\hat{B}_r - B \rightarrow_p 0, \text{ as } n \rightarrow \infty$$

we have,

$$\bar{e}_r - \bar{E} \rightarrow_p 0 \text{ as } n \rightarrow \infty,$$

and

$$\sum_{(g|k) \in s} \tilde{w}_{g|k} a_{g|k} \rightarrow_p p,$$

we obtain, using the conditions C4 in (8.28),

$$E_* |t_{hi}^*|^{2+\delta} \leq \left(\sum_j \tilde{w}_{hij} \sqrt{z_{hij}} \right)^{2+\delta} O_p(1).$$

Hence, using condition C3, we have

$$\begin{aligned} \frac{1}{n} \sum_h \sum_i E_* |nt_{hi}^*|^{2+\delta} &\leq n^{1+\delta} \sum_h \sum_i \left(\sum_j \tilde{w}_{hij} \sqrt{z_{hij}} \right)^{2+\delta} O_p(1) \\ &\leq n^{2+\delta} \max_{h,i} \left(\sum_j \tilde{w}_{hij} \sqrt{z_{hij}} \right)^{2+\delta} O_p(1) \\ &= \left(n \max_{h,i} \sum_j \tilde{w}_{hij} \sqrt{z_{hij}} \right)^{2+\delta} O_p(1) \\ &= O_p(1). \end{aligned}$$

Also, $\frac{1}{n} \sum_h \sum_i V_*(nt_{hi}^*) = nV_* \left(\sum_h \sum_i \sum_j \xi_{hij}^* \right) \rightarrow \sigma_\xi^2$, say, as $n \rightarrow \infty$,

and $\frac{1}{n} \sum_h \sum_i E_*(nt_{hi}^*) = \sum_h \sum_i \sum_j E_*(\xi_{hij}^*) = 0$. Applying Lemma 8.3, we obtain

$$\sqrt{n} \left(\frac{1}{n} \sum_h \sum_i nt_{hi}^* \right) \rightarrow_d N(0, \sigma_\xi^2).$$

Finally, applying Lemma 8.5 with $V_n = [E_*(\bar{y}_I) - B_r \bar{Z}] + [B_r \bar{Z} - \bar{Y}]$

and $U_n = [\bar{y}_I - E_*(\bar{y}_I)]$, we obtain

$$\frac{1}{\sqrt{V(\bar{y}_I)}} (\bar{y}_I - \bar{Y}) \rightarrow_d N(0, 1),$$

where $V(\bar{y}_I) = \sigma_{1n}^2 + \sigma_{2n}^2$ with $\sigma_{1n}^2 = V(E_*(\bar{y}_I))$ and $\sigma_{2n}^2 = V_*(\bar{y}_I)$.

Chapter 9

Future Research

In this thesis, we have studied several aspects of the problem of inference in the presence of imputation for missing survey data. For future research, we propose the following directions.

In Chapter 2, we studied the effect of unweighted imputation on the bias of the imputed estimators. We have proposed a bias-adjusted estimator which can be justified under both the design-based and the model-based approaches. Our “prototype” estimator was the Horvitz-Thompson estimator. It would be of interest to study the effect of unweighted imputation on the GENeralized Regression estimator (GREG) that include the ratio estimator as well as the post-stratified estimator as special cases and derive the correct variance estimators of the corresponding imputed estimators.

In Chapter 4, we considered a modified regression imputation that leads to valid imputed estimators under assumption GNM. We intend to consider the problem of

estimating bivariate parameters of the form

$$\theta = \frac{1}{N} \sum_{i \in U} g(x_i, y_i), \quad (9.1)$$

where both x and y may be missing. The goal will be to determine a modified regression imputation procedure that will lead to a valid imputed estimator of bivariate parameters under assumption GNM.

In Chapter 5, we investigated the case of mean imputation in two-stage sampling. The results obtained may be extended to the case of two-stage sampling and regression imputation. For example, suppose that a vector of q auxiliary variables $\mathbf{z} = (z_1, \dots, z_q)$ is observed on all the sampled elements (respondents and nonrespondents). In the case of assumption MBC, the imputation model is given by

$$y_{ij} = \mathbf{z}'_{ij}\boldsymbol{\beta} + \alpha_i\gamma + \epsilon_{ij},$$

where $\boldsymbol{\beta}$ is an unknown q -vector of fixed effects parameters, and α_i and ϵ_{ij} are as before. Regression imputation uses $y_{ij}^* = \mathbf{z}'_{ij}\hat{\mathbf{B}}_r$ or $y_{ij}^* = \mathbf{z}'_{ij}\hat{\mathbf{B}}_r + \hat{\alpha}_i$, where $\hat{\mathbf{B}}_r$ is the ordinary or weighted least squares estimators of $\boldsymbol{\beta}$ and $\hat{\alpha}_i$ is the predictor of the random effect.

In Chapter 6, we studied the problem of estimating bivariate parameters under random hot-deck imputation. In practice, nearest-neighbour imputation is frequently used to compensate for item nonresponse. Thus, it would be of practical interest to study the effect of nearest-neighbour imputation on bivariate parameters. We can, for example, extend the results obtained by Chen and Shao (2000) to the case of bivariate parameters of the form (9.1). Under MAR, we would assume the following

bivariate model:

$$y_i = f_1(z_{1i}) + \epsilon_i, E_m(\epsilon_i) = 0, E_m(\epsilon_i \epsilon_j) = 0, i \neq j, V_m(\epsilon_i) = \sigma_{\epsilon_i}^2$$

$$x_i = f_2(z_{2i}) + \eta_i, E_m(\eta_i) = 0, E_m(\eta_i \eta_j) = 0, i \neq j, V_m(\eta_i) = \sigma_{\eta_i}^2$$

$$Cov_m(\epsilon_i, \eta_i) = \sigma_{\epsilon\eta},$$

where z_1 and z_2 are the auxiliary (or matching) variables used for imputing variables y and x , respectively and $f_1(\cdot)$ and $f_2(\cdot)$ are two smooth functions.

In Chapter 7, we considered the problem of constructing imputation classes. For estimating a population total (or a population mean), we showed that the key for reducing the bias is to form classes that are homogeneous with respect to the estimated response probabilities and/or to the predicted item values. We intend to consider the problem of constructing classes when the parameter of interest is a domain total (or a domain mean). In this case, several scenarios need to be considered:

- (a) The domain indicators are strongly/weakly related to the variables of interest.
- (b) The domain indicators are used/not used in the imputation model for forming the classes.

Also, we have considered the case of only one variable requiring imputation. In practice, many variables of interest require imputation. Thus, we intend to extend the score method to two or more variables so the classes would be simultaneously homogeneous with respect to these variables.

Bibliography

- [1] Beaumont, J. F. (2005). Calibrated Imputation in Surveys under a Quasi-Model-Assisted Approach. *To appear in the Journal of the Royal Statistical Society B.*
- [2] Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 15, 279-292.
- [3] Binder, D. A. and Sun, W. (1996). Frequency valid multiple imputation for surveys with a complex design. *ASA Proceedings of the Section on Survey Research Methods*, 281-286.
- [4] Brick, J. M., Kalton, G. and Kim, J. K. (2004). Variance Estimation with Hot-Deck Imputation Using a Model. *Survey Methodology*, 30, 57-66.
- [5] Chen, H. L., Rao, J. N. K. and Sitter, R. R. (2000). Efficient Random Imputation for Missing Survey Data in Complex Survey. *Statistica Sinica*, 10, 1153-1169.
- [6] Chen, H. L. and Shao, J. (2000). Nearest-Neighbour Imputation for Survey data. *Journal of Official Statistics*, 16, 583-599.

- [7] Chen, H. L. and Shao, J. (2001). Jackknife Variance Estimation for Nearest-Neighbour Imputation. *Journal of the American Statistical Association*, 96, 260-269.
- [8] Chen, J. and Rao, J. N. K. (2004). Asymptotic normality under two-phase sampling designs. *Unpublished manuscript*.
- [9] Cohen, M. P. (2002). Implementing Rao-Shao Type Variance Estimation with Replicate Weights. *Survey Methodology*, 28, ???.
- [10] Demnati, A. and Rao, J. N. K. (2004). Linearization Variance Estimators for Survey Data. *Survey Methodology*, 30, 17-26.
- [11] Deville, J. C. and Särndal, C. E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Survey Methodology*, 23, 33-40.
- [12] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7, 1-26.
- [13] Eltinge, J. L. and Yansaneh, I. S. (1997). Diagnostics for formation of Nonresponse Adjustment Cells, With an Application to Income Nonresponse in the U. S. Consumer Expenditure Survey. *Survey Methodology*, 23, 33-40.
- [14] Fay, R. E. (1991). A Design-Based Perspective on Missing Data Variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the Census*, 429-440.

- [15] Fay, R. E. (1992). When are inferences from multiple imputation valid?. *ASA Proceedings of the Section on Survey Research Methods*, 227-232.
- [16] Fay, R. E. (1996). Alternative Paradigms for the Analysis of Imputed Survey Data. *Journal of the American Statistical Association*, 91, 490-498.
- [17] Fuller, W. A. (1996). Introduction to Statistical Time Series (2nd edition). *Introduction to Statistical Time Series*, 2nd edition. Wiley, New York.
- [18] Hansen, H., Hurwitz, W. and Madow, W. (1953). *Sample Survey Methods and Theory*. John Wiley & Sons, New York.
- [19] Haziza, D. and Rao, J. N. K. (2003). Inference for population means under unweighted imputation for missing survey data. *Survey Methodology*, 29, 81-90.
- [20] Haziza, D. and Rao, J. N. K. (2005). Inference for domain means and totals under imputation for missing data. *To appear in the Canadian Journal of Statistics*.
- [21] Haziza, D. and Rao, J. N. K. (2005). A nonresponse model approach to inference under imputation for missing survey data. *In review for Survey Methodology*.
- [22] Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- [23] Kalton, G. and Kasprzyk, D. (1986). The treatment of survey missing data. *Survey Methodology*, 12, 1-16.
- [24] Kalton, G. and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics, Part A- Theory and Methods*, 13, 1919-1939.

- [25] Kass, G. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29, 119-127.
- [26] Little, R. J. A. (1986). Survey Nonresponse Adjustments. *International Statistical Review*, 54, 139-157.
- [27] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. Wiley, New York.
- [28] Oh, H. L. and Scheuren, F. (1983). Weighted adjustment for nonresponse. *Incomplete Data in Sample Survey*, vol. 2 (Madow, W.G., Nisselson, H. and Olkin, I. Eds.), New York: Academic Press, 143-184.
- [29] Pfefferman, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.
- [30] Quenouille, M. H. (1949). Approximate tests of Correlation in time-series. *Journal of the Royal Statistical Society, series B*, 11, 68-84.
- [31] Rancourt, E. (2001). Edit and Imputation: From suspicious to scientific techniques. *Proceedings of the 2001 ISI conference*.
- [32] Rao, J. N. K. (1965). On two simple schemes of unequal probability sampling with replacement. *Journal of the Indian Statistical Association*, 3, 173-180.
- [33] Rao, J. N. K. (1966). Alternative Estimators in PPS sampling for Multiple Characteristics. *Sankhya, Series A*, vol. 28, Part 1, 47-59.

- [34] Rao, J. N. K. (1990). Variance estimation under imputation for missing data. Technical report, Statistics Canada, Ottawa.
- [35] Rao, J. N. K. and Shao, J. (1992). On variance estimation under imputation for missing data. *Biometrika*, 79, 811-822.
- [36] 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.
- [37] Rao, J. N. K. and Wu, C. F. J. (1988). Resampling Inference with Complex Survey Data. *Journal of the American Statistical Association*, 83, 231-241.
- [38] Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.
- [39] Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63, 581-590.
- [40] Rubin, D. B. (1978). Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse. *ASA Proceedings of the Section on Survey Research Methods*, 20-28.
- [41] Rubin, D. B. (1987). *Multiple imputations for nonresponse in surveys*. Wiley, New York.
- [42] Saigo, H., Shao, J. and Sitter, R. R. (2001). A repeated Half-Sample Bootstrap and Balanced Repeated Replications for Randomly Imputed Data. *Survey Methodology*, 27, 189-196.

- [43] Sampford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.
- [44] Särndal, C. E. (1990). Method for estimating the precision of survey estimates when imputation has been used. *Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality*. Statistics Canada, 337-347.
- [45] Särndal, C. E. (1992). Method for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.
- [46] Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components*. Wiley, New York.
- [47] Seber, G. A. F. (1984). *Multivariate Observations*. Wiley, New York.
- [48] Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman & Hall, New York.
- [49] Schafer, D. B. (1997). *Analysis of incomplete multivariate data*. Chapman-Hall.
- [50] Shao, J. and Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 93, 819-831.
- [51] Shao, J. and Steel, P. (1999). Variance Estimation for Survey Data With Composite Imputation and Nonnegligible Sampling Fractions. *Journal of the American Statistical Association*, 94, 254-265.

- [52] Shao, J. and Wang, H. (2002). Sample Correlation Coefficients Based on Survey Data under Regression Imputation. *Journal of the American Statistical Association*, 97, 522-544.
- [53] Shao, J. (2002). Replication Methods for Variance Estimation in Complex Surveys with Imputed Data. In *Survey Nonresponse edited by R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little*. Wiley, New York.
- [54] Skinner, C. J., and Rao, J. N. K. (2002). Jackknife Variance for Multivariate Statistics under Hot Deck Imputation from Common Donors. *Journal of Statistical Planning and Inference*, 102, 149-167.
- [55] Smith, P. J., Hoaglin, D. C., Rao, J. N. K., Battaglia, M. P. and Daniels, D. (2004). Evaluation of Adjustment for Partial Non-response Bias in the US National Immunization Survey. *Journal of the Royal Statistical Society A*, 167, 141-156.
- [56] Steinberg, D., and Colla, P. (1995). Tree-Structured Non-Parametric Data Analysis. *San Diego, CA: Salford Systems*.
- [57] Thomsen, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistik Tidskrift*, 16, 191-196.
- [58] Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29, 614.

- [59] Vartivarian, S. and Little, R. J. A. (2002). On the Formation of Weighting Class Adjustments for Unit Nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

- [60] Wang, N. and Robins, J. M. (1998). Large sample inference in parametric multiple imputation. *Biometrika*, 85, 935-948.