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

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

Inference from Survey Data with Imputed Values

by

Steve Matthews

A thesis submitted to

the Faculty of Graduate Studies and Research

in partial fulfillment of the requirements for the degree of

Master of Science

School of Mathematics and Statistics

Ottawa-Carleton Institute for Mathematics and Statistics

Carleton University

Ottawa, Ontario, Canada

August, 2003

© Copyright 2003

Steve Matthews

National Library
of Canada

Bibliothèque nationale
du Canada

Acquisitions and
Bibliographic Services

Acquisisitons et
services bibliographiques

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 0-612-88778-2

Our file Notre référence

ISBN: 0-612-88778-2

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

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

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

Canada

The undersigned hereby recommend to the Faculty of
Graduate Studies and Research acceptance of the thesis

Inference from Survey Data with Imputed Values

Submitted by

Steve Matthews, B.Sc.

in partial fulfillment of the requirements
for the degree of Master of Science.

Director, School of Mathematics and Statistics

Dr. J.N.K. Rao, Thesis Supervisor

Carleton University

August, 2003

Abstract

This thesis considers inference from survey data with imputed values. In particular, we consider a stratified simple random sample design where non-response is treated by imputation methods that can be expressed as a linear regression model (includes mean, ratio and linear regression imputations). The thesis considers the bias and variance of the estimators under this class of imputation methods, applies procedures to derive variance estimators through the reverse approach framework elaborated by Shao and Steel (1999), and evaluates the derived estimators through Monte-Carlo simulations. We explore the quality of the point estimates and variance estimates when the model underlying the imputation method does not hold in the population (either the model used to describe non-response, the model underlying the imputation method, or both). Finally, the procedures are applied to data from a large-scale business survey at Statistics Canada.

Acknowledgements

I would like to express my sincere gratitude to my M.Sc. thesis supervisor, Professor J. N. K. Rao, for his supervision and encouragement during the preparation of my thesis, for the knowledge and skills he shared with me throughout my studies, and also his kindness and support. I would also like to thank my M.Sc. thesis committee members, Drs. R. Thomas, C. G. Park, and M. Csörgö of Carleton University and Dr. M. Alvo of the University of Ottawa for their time and effort in the examination of my thesis. I would also like to thank my wife Julie for her continuous support and encouragement throughout my studies. My thanks are also extended to David Haziza of Statistics Canada for his helpful discussions and suggestions. This research was partially supported by a research assistantship from Carleton University.

Contents

Acceptance Sheet	ii
Abstract	iii
Acknowledgements	iv
Contents	v
List of Tables	vii
1 INTRODUCTION	1
1.1 Notation	2
1.2 Non-Response	3
1.2.1 Non-response mechanisms	4
1.2.2 Treatment of Non-Response	7
2 SURVEY DATA WITH IMPUTED VALUES	10
2.1 Common Practice	10
2.1.1 Point Estimation from Imputed Survey Data	11
2.1.2 Variance Estimation from Imputed Survey Data	13

2.2	Review of Literature	16
2.2.1	Point Estimation from Imputed Survey Data	16
2.2.2	Variance Estimation from Imputed Survey Data	19
2.3	Scope of the Presented Work	26
2.3.1	Motivation	26
2.3.2	Development Required	27
3	REGRESSION IMPUTATION	30
3.1	Introduction	30
3.1.1	Design of Simulation Study	30
3.2	The Regression-Imputed Estimator	34
3.2.1	Bias of the estimator	36
3.2.2	Variance Estimation	44
3.3	Summary of Findings	57
4	APPLICATION OF RESULTS	61
4.1	Composite Imputation	62
4.1.1	Bias of the Estimator	62
4.1.2	Variance Estimation	64
4.2	Application to Survey Data	65
4.2.1	The Monthly Wholesale Retail Trade Survey	65
4.2.2	Imputation Strategy	67
4.2.3	Bias of the Estimator	70
4.2.4	Variance Estimation	71

4.2.5	Population Characteristics	76
4.3	Summary of Findings	90
5	CONCLUSIONS	93
5.1	Summary of Findings	93
5.2	Future Work	98
5.2.1	Model-assisted variance estimation under composite imputation	98
5.2.2	Collapsing of Imputation Cells	98
5.2.3	Stratification, Imputation Groups and Domain Estimation . .	99
	Bibliography	103
	Appendices	104
A	REGRESSION IMPUTATION	104
B	COMPOSITE IMPUTATION FOR THE MWRTS	109

List of Tables

3.1	Regression Imputation*: Properties of Point Estimator under MCAR	37
3.2	Regression Imputation*: Properties of Point Estimator under INR . .	38
3.3	Regression Imputation*: Properties of Point Estimator under NINR .	39
3.4	Regression Imputation**: Properties of Point Estimator under MCAR	41
3.5	Regression Imputation**: Properties of Point Estimator under INR .	42
3.6	Regression Imputation**: Properties of Point Estimator under NINR	43
3.7	Regression Imputation*: Relative Bias (%) of Variance Estimators under MCAR	46
3.8	Regression Imputation*: Relative Bias (%) of Variance Estimators under INR	47
3.9	Regression Imputation*: Relative Bias (%) of Variance Estimators under NINR	48
3.10	Regression Imputation*: Coverage Rates of 95% Confidence Intervals of Variance Estimators under MCAR	49
3.11	Regression Imputation*: Coverage Rates of 95% Confidence Intervals of Variance Estimators under INR	50

3.12	Regression Imputation*: Coverage Rates of 95% Confidence Intervals of Variance Estimators under NINR	51
3.13	Regression Imputation**: Relative Bias (%) of Variance Estimators under MCAR	52
3.14	Regression Imputation**: Relative Bias (%) of Variance Estimators under INR	53
3.15	Regression Imputation**: Relative Bias (%) of Variance Estimators under NINR	54
3.16	Regression Imputation**: Coverage Rates of 95% Confidence Inter- vals of Variance Estimators under MCAR	55
3.17	Regression Imputation**: Coverage Rates of 95% Confidence Inter- vals of Variance Estimators under INR	56
3.18	Regression Imputation**: Coverage Rates of 95% Confidence Inter- vals of Variance Estimators under NINR	57
4.1	MWRTS data - Distribution of response rates by cell	77
4.2	MWRTS data - $\hat{\rho}(y, \hat{y}_m^I)$ under each imputation model m	77
4.3	MWRTS data - Number of months with significant β_1 in logistic re- gression	80
4.4	MWRTS Estimates - Average Ratios of v_0 to v_D and v_1 to v_D	83
4.5	MWRTS Variance Estimates: Example 1	84
4.6	MWRTS Variance Estimates: Example 2	86
4.7	MWRTS Variance Estimates: Example 3	87
4.8	MWRTS Variance Estimates: Example 4	88

4.9	MWRTS Variance Estimation: Example 5	89
4.10	MWRTS Variance Estimates: Average Ratios of v_{2M} to v_{2D}	91

Chapter 1

INTRODUCTION

In virtually all sample surveys, a challenge results from an inability to obtain data for each sampled unit. The situation where a data item could not be obtained is referred to as non-response, and special procedures must be applied to use the incomplete data for the intended purposes. These special procedures introduce further complications when using survey data to draw conclusions on the population under study.

In the case of large-scale surveys, non-response can occur for a variety of reasons, and the way that the non-response is treated introduces an added complexity to the data that must be considered for valid inference. In particular, the treatment of non-response through imputation can lead to bias in point estimates, and complicates the derivation of meaningful measures of data quality such as the estimated variance. Among the many sources of errors in sample surveys, such as sampling error, response error, and processing errors, it is typically only the sampling error that is expressed through measures of data quality, and considered for drawing in-

ference via confidence intervals and hypothesis testing. One error that is typically not represented is that arising from imputation. Recently, this has been a common research topic, and developments have been made on procedures which better incorporate other sources of error in the overall quality assessment of statistical data from sample surveys.

This thesis considers the estimation of population totals, and variance estimation for these estimates in the context of sample surveys with nonresponse treated by imputation. Chapter 2 provides an introduction to the problem and summarizes research to date in the area. In Chapter 3, existing theory is extended to derive basic results under regression imputation, and several variance estimators are compared through an empirical study. The data used for the empirical study were created using models that in some cases meet the assumptions inherent in the estimation procedures, and in some cases do not. In Chapter 4 we develop variance estimators for application to data from a large-scale survey conducted by Statistics Canada, and generate variance estimators for recent survey data. In Chapter 5, our findings are summarized, and areas requiring further research are identified.

1.1 Notation

We use the following notation throughout this thesis:

A population U is divided into H strata $U_1, \dots, U_H$. From each strata h , a sample s_h is drawn by simple random sampling without replacement according to the stratum sampling fraction f_h . Let $s = s_1 \cup s_2 \cup \dots \cup s_H$ denote the entire sample. The index

(hi) will be used to identify the i th unit in the h th stratum of the population. For each unit hi in the population, a column-vector of auxiliary variables x_{hi} is available and we attempt to collect the value, y_{hi} , of a variable of interest, y , for each unit in the sample. The aim is to produce an estimate of the population total $Y = \sum_U y_{hi}$ using the collected data, where $\sum_U$ denotes summation over the units in the population U . The indicator variable a_{hi} is used to identify units that respond to the variable of interest; that is, $a_{hi} = 1$ if (hi) is a respondent, and $a_{hi} = 0$ if (hi) is a non-respondent to the variable of interest.

1.2 Non-Response

Non-response in sample surveys and censuses can occur due to poor survey planning, inability to contact a person able to provide a response, refusal to participate or some other reason. In large-scale surveys by statistical agencies, many measures are taken to reduce the amount of non-response, and its impacts on the quality of the data. These measures include providing incentives to survey respondents, follow-up strategies to obtain data from units that do not respond, and in some cases, federal legislation requiring the participation in a survey. Despite these efforts, it is virtually inevitable that some non-response will occur.

An important distinction is often made between item and unit non-response as in Kalton and Kasprzyk (1986) when different data items are collected from each unit of observation. Unit non-response occurs when data are not obtained from a particular unit for any of the data items requested. Item non-response occurs when

data are not obtained from a particular unit for a specific data item requested. The reasons for item and unit non-response are often quite different. In this thesis, we consider item nonresponse, but confine ourselves to a single item, y , and estimating its total under imputation for missing values.

1.2.1 Non-response mechanisms

If we consider response to a data item as a random event, with probability p_{hi} for each unit (hi), completely describing the phenomenon of non-response is equivalent to specifying the probabilities of response, p_{hi} , for each unit in the population. Of course, these probabilities are unknown, and a model is often constructed to describe their form and the probabilities are estimated via this model. A number of models have been proposed to describe the phenomenon of non-response and three classes of such models are described below.

Missing Completely at Random (MCAR)

In this non-response model each unit in the population has the same probability of response:

$$p_{hi} = p.$$

According to this response pattern, the set of respondents can be seen as a sub-sample drawn by Poisson sampling with constant probability p ; that is, each sample unit has a probability, p , of response to the item and the responses of each unit are independent of each other.

Ignorable Non-Response (INR)

In this non-response model, also referred to as Missing at Random (MAR), the probability of response for a unit (hi), is completely described through a function of auxiliary variables, $f(\mathbf{x}_{hi})$. This function is independent of the variable of interest, y_{hi} :

$$p_{hi} = f(\mathbf{x}_{hi})$$

For example, if $\mathbf{x}_{hi}$ consists of only one variable, x_{hi} , and $f(x_{hi})$ is a logit function, the relationship between p_{hi} and x_{hi} is monotonic, and bounded between zero and one. If the monotonic function is approximately a linear function in x , the set of respondents will be similar to what would have been drawn by Probability Proportional to Size sampling with replacement from the complete sample. For the purposes of this thesis, we can distinguish between two separate types of auxiliary variables; those used in imputation, and those not used in imputation. For example, some auxiliary variables may be available for all units in the population, and these could be used in processes such as stratification. Other auxiliary variables may be available only for those units in the sample, and these could be used for imputation or other purposes. In considering Ignorable Non-Response, we will consider the response to be ignorable if the probability of response is described through only those auxiliary variables used for imputation.

An example of ignorable non-response is the situation where units with high values of $\mathbf{x}_{hi}$ are more likely to respond than units with small values of $\mathbf{x}_{hi}$. This is common

in business surveys which collect financial data from companies. Typically, the units with smaller revenues are less likely to respond to financial data items. Reasons for this include: limitations inherent in the operating structures and monitoring activities of smaller businesses (i.e. reporting difficulties), and follow-up strategies which allocate more resources to obtain data from larger companies (due to their relative importance to the survey estimates). In these strategies, the priorities for follow-up are often determined directly from auxiliary data, which leads to differing probabilities of response conditional on the auxiliary data. If the response mechanism prior to follow-up is MCAR, and the same auxiliary data are used in imputation and in the follow-up strategy, overall the response mechanism will be INR.

Non-Ignorable Non-Response (NINR)

In this model the probability of response for each unit is described through a function of auxiliary variables, $\mathbf{x}_{hi}$, and the variable of interest, y_{hi} .

$$p_{hi} = f(y_{hi}, \mathbf{x}_{hi})$$

This is the most difficult type of non-response to treat in inference. In this case, the values y_{hi} contain information on the probability of response, and since we do not know the values y_{hi} for non-sampled units and non-responding sampled units, we are not able to make adjustments to use the set of respondents to represent the population. To the extent that the correlation between the variable of interest and the auxiliary variable used in imputation is high throughout the population, the individual response probabilities will be similar under INR and NINR.

1.2.2 Treatment of Non-Response

In order to draw inference from survey data with non-response, it is important that the sample adequately represents the entire population (through the sampling weights). With non-response, standard methods to draw population inferences are not applicable so a special process is required for inference to be valid for the population for which the survey was designed. Essentially, two different methods are available: imputation and re-weighting. In many cases, an imputation strategy can be translated to an analogous re-weighting method by re-expressing the resulting estimator.

Imputation

Commonly in large-scale surveys, imputation is used to correct for non-response. Imputation consists of estimating a response value according to some model for each non-responding unit, and using this estimate as if it were a reported value. Imputation has the following desirable properties.

- It results in a complete data file,
- performing analysis on different subsets of data will give consistent results,
- the process can be highly automated, and
- where available, auxiliary data can be exploited to yield imputed values with good properties.

We can make use of auxiliary data in estimating the value to be used from a model, and in forming groups of units called Imputation Cells. These groups are typically defined to produce homogeneous sets of units, and the imputation is done indepen-

dently within each group. Drawbacks of imputation are that bias can result if the model implied by the imputation method is not accurate, and the effects of treating imputed values improperly can lead to erroneous inferences (as discussed in later chapters).

In some cases, stochastic imputation, where a random component is included in imputed values, is used to complete non-respondent data. In deterministic imputation, repeated imputations for the same sample, and same set of respondents will yield identical results. Our focus is on business surveys, which typically use deterministic methods for operational reasons, so stochastic imputation will not be considered in this thesis.

Re-weighting

Another option to treat non-response is to modify the weights for responding units. The new weight is derived in such a way that the product of the non-response weight and the sampling weight can be used to represent the population when applied to the set of respondents. As in imputation, sub-groups of units can be formed using auxiliary data, within each of which the re-weighting is done independently.

In surveys where many different variables are collected, and estimates are produced for each variable, the use of re-weighting can lead to complications in analysis. For example, different sets of non-responding units for each variable lead to different sets of unit weights associated with each variable. This complicates the estimation,

and can lead to inconsistent results from the analysis of subsets of the population using the same survey data.

With these advantages and disadvantages, a common approach in social surveys is to treat unit non-response with re-weighting, and to fill in item non-response via imputation. In business surveys, it is more common to impute for both types of non-response, as the auxiliary information available typically leads to imputed values of high quality. As will be shown in the following chapters, care must be given in drawing inference from survey data with imputed values.

Chapter 2

SURVEY DATA WITH IMPUTED VALUES

Non-response is not a new phenomenon, and imputation has been used for quite some time to correct for missing values. A review of common imputation methods used at Statistics Canada is given by Kovar and Whitridge (1995). Section 2.1 gives an overview of the common practice used to draw inference from business survey data with imputed values at Statistics Canada. We focus on stratified simple random sampling with n_h units sampled without replacement from N_h population units in stratum h ($= 1, \dots, H$). Sampling is done independently across strata h .

2.1 Common Practice

We first describe the method that is typically used to draw inference from business survey data with imputed values at Statistics Canada.

2.1.1 Point Estimation from Imputed Survey Data

In the case where the item value is reported for each unit selected in the sample, the traditional Horvitz-Thompson estimator introduced in Horvitz and Thompson (1952) for the population total $Y = \sum_U y_{hi}$ takes the following form:

$$\hat{Y} = \sum_s w_{hi} y_{hi} \quad (2.1)$$

where $\sum_s$ denotes summation over units in the sample, s , $w_{hi} = \frac{1}{f_h}$ is the sampling weight of unit i in stratum h , and $f_h = \frac{n_h}{N_h}$ is the sampling fraction. This estimator is unbiased for Y , and an unbiased estimator of its variance $V(\hat{Y})$ is given by:

$$v = \sum_s (1 - f_h) \frac{N_h^2}{n_h} \frac{(y_{hi} - \bar{y}_h)^2}{(n_h - 1)}, \quad (2.2)$$

where $\bar{y}_h$ is the sample mean of the variable y for stratum h .

Imputation is designed in such a way that it is transparent for point estimation. Once a dataset has been completed through imputation, the imputed Horvitz-Thompson estimator is expressed as follows:

$$\hat{Y}^I = \sum_s w_{hi} y_{hi}^I, \quad (2.3)$$

where y_{hi}^I is equal to y_{hi} if sample unit (hi) responds to the variable y , and equal to the imputed value, y_{hi}^* if sample unit (hi) does not respond to the variable y .

In determining the importance of the bias on inference, we will follow the guideline suggested by Cochran (1977). Cochran suggested as a working rule that inferences on Y can be erroneous when the ratio of the bias to the standard error is above 10

%.

Typically, the data collected for a survey are limited in its information on characteristics of non-responding units. Any analysis of variables of interest is limited to the set of respondents since the variable of interest is not available for non-respondents. To avoid this limitation in studying non-response, the "Standards and Guidelines for Reporting of Non-response Rates" published by Statistics Canada (2000) recommends data analysis be based on the following practices:

- (i) *Obtain further information on non-respondents* : Through a non-response follow-up study one could select a sub-sample of non-respondents and collect a limited set of variables using more extensive data collection procedures. With this information, one could test the validity of the assumptions made on the response mechanism in the treatment of non-response.
- (ii) *Use existing information on non-respondents* : If related information is available for non-respondents, such as data from previous surveys, auxiliary data, or data directly from the survey, it can be used to suggest plausible adjustments to reduce non-response bias.
- (iii) *Use information from design of survey* : Comparisons of weighted and unweighted non-response rates can often serve as indicators on the type of non-response mechanism, and the potential risk of the assumptions made. The weighted non-response rates ($\sum_s w_{hi} a_{hi} / N$) indicate the portion of the estimate that is at risk of a potential bias due to the non-response adjustment procedure, and the unweighted rates ($\sum_s a_{hi} / n$) indicate the fraction of units that did not respond. If these rates are very different this may indicate that the weights themselves contain information

on the non-response pattern, and the results should be considered with caution.

In practice, through the use of auxiliary data that are highly correlated with the variable of interest in imputation, the effective difference between ignorable non-response (INR) and non-ignorable non-response (NINR) should not be significant, and methods that lead to unbiased estimators under INR will lead to approximately unbiased estimators under NINR. In Chapter 3 properties of the point estimators will be studied to identify the circumstances that lead to biased estimators.

2.1.2 Variance Estimation from Imputed Survey Data

Estimation of variance is more complicated when imputation has been used to treat non-response. In many cases, to estimate the variance of the imputed estimator given in (2.3) the following estimator is used:

$$v_0 = \sum_s (1 - f_h) \frac{N_h^2 (y_{hi}^I - \bar{y}_h^I)^2}{n_h (n_h - 1)}, \quad (2.4)$$

where $\bar{y}_h^I$ is the sample mean of the variable y^I (both imputed and reported values) for stratum h . This variance estimator is often referred to as the "naive" variance estimator. The term naive is used to indicate that the imputed values are treated as if they were actually reported values. It is well-known that this can lead to a serious underestimation of the variance. This underestimation occurs for two reasons: (i) under certain imputation methods the resulting dataset contains less than natural variability, and (ii) the imputed values are subject to an additional source of variability that is not estimated using traditional methods. This underestimation could

result in mis-use of the data due to over-confidence in the data. To avoid this, we would like to produce variance estimates that better reflect the total variability of the estimator.

Statistics Canada has developed a set of guidelines to be followed in publishing statistical data, and their indicators of data quality. At Statistics Canada, the guidelines are aimed towards providing data quality through the notion of "fitness for use." Since the uses and requirements of statistical data produced by survey programs vary so widely within the organization, the notion of fitness for use is meant to reflect the degree to which the data are of a quality suitable for its intended use. Thus, it includes characteristics such as bias and variance. In many surveys the sample design is complex, for example, stratified multiple stage sampling or two-phase sampling. This complexity is often introduced for operational reasons, to improve the efficiency of resulting estimates, or both. For such surveys, methods that incorporate the variance associated with imputation have not been developed. Instead, some surveys simply publish response rates in their data quality report to give the data user an idea of the amount of the data that corresponds to imputed values. These rates give a general idea of the amount of imputation, but cannot capture all aspects related to its impact. For example, different imputation methods lead to estimates with different amounts of variability and bias, given the same response fraction. In this way, although the response rates are general indicators, they do not express the data quality as directly as possible. It would be preferable to express the quality of the data in terms of estimated bias and an improved estimated variance

as neither one, on its own, can summarize the quality. Some alternatives currently used are:

- Report a data quality rating which is a function of the estimated variance and the response rate; for example, the rating is "A" if the Co-efficient of Variation is less than 5% and the Response Rate is above 95% , etc.
- In the case of simpler sample designs, calculate a variance estimate designed to include this source of variability, by some method such as those described in section 2.2.
- Apply an adjustment factor to the reported variance estimates in order to approximately reflect the increase in variance due to imputation.

All of the above alternatives, if correctly interpreted, can be used to obtain a general notion of "fitness for use". It would be most convenient to data users if one source of data quality information could be used on its own to accurately convey the uncertainty associated with a statistical estimate. Since the coefficient of variation is the most commonly published data quality indicator, it is the most viable candidate for uniform use across statistical programs. Although the methodology used to calculate the variance estimates would differ from survey to survey, depending on the sample design, and treatment of non-response, the end result of an estimated variance or coefficient of variation would help to standardize data quality reporting across survey programs. For these reasons, we seek to develop methods suitable for incorporating the uncertainty associated with non-response treatments into the estimated variance. In section 2.2 we summarize the progress in this regard.

2.2 Review of Literature

The traditional viewpoint taken in developing the methods described below has been as follows: From the population, a sample is drawn, and among the sampled units, some do not respond. In this way, for valid inference, we make use of an analogy between non-response, and a second-phase of sampling (although we have no control over the second-phase) since the theory behind two-phase sampling is well established. For this reason, this response path is often referred to as the "two-phase approach".

Population $\Rightarrow$ Sample $\Rightarrow$ Sample with non-respondents

2.2.1 Point Estimation from Imputed Survey Data

Depending on the non-response mechanism and the method used to impute the missing values, use of the imputed Horvitz-Thompson estimator can lead to bias in the estimates of population characteristics. Ideally, a desirable imputation strategy yields approximately unbiased estimators with low variance. Producing approximately unbiased estimators is typically the first priority in developing an imputation strategy.

One should not jump ahead to the assumption that the point estimates are unbiased without some validation. As stated in Rancourt, Lee and Sarndal (1994), "Imputation is sometimes applied without critical questioning, and, although widely used, imputation does not solve the critical problem of bias caused by non-response."

Many factors can cause imputed estimators to be biased, such as the presence of outliers and the way their values are used in imputation, or special procedures such as collapsing of imputation groups. Generally, the data required to estimate the bias in an estimator are unavailable due to non-response, so models are often employed to describe the response probabilities, and the bias is considered with respect to this model.

Bias correction procedures

Haziza and Rao (2003) demonstrate that when the design is not self-weighting within each imputation class, a bias is introduced by imputation using common methods if the design weights are not used in generating the imputed values for non-respondents. They develop a post-imputation adjustment which corrects for this bias under INR.

Little and Rubin (1987) propose a method to correct non-response bias resulting from confounded or non-ignorable non-response through use of auxiliary data and a Bayesian framework. Separate regression models for non-respondents and respondents are assumed for the variable of interest, with auxiliary data used as predictors, and the regression co-efficients for the non-respondents are estimated through a normal prior distribution with mean based on the coefficients estimated for the respondents. Once the coefficients of the model for the respondents are estimated, the imputation is carried out conditionally on the predicted response probabilities. There are a number of model assumptions required for the validity of this approach;

and it may be difficult or impossible to verify the models.

Alternatively, Greenless, Reece and Zieschang (1982) developed a maximum likelihood estimator to address the problem of non-ignorable non-response. Regression model parameters are estimated to predict the variable of interest and the probability of response. Imputed values are then generated from the predicted model, conditional on the auxiliary data. Again, the issue of validating assumptions is non-trivial as the data itself cannot be used, and another source of data is required. In addition, this method is computer intensive.

Rancourt, Lee and Sarndal (1994) studied the performance of ratio estimators under INR and NINR response mechanisms, and demonstrated conditions under which a bias can be introduced. The paper focused on the development of simple bias correction factors applied to imputed values. The authors avoid explicit modeling of the non-response mechanism. Essentially, the bias corrections estimate the ratio of the value that should have been imputed (had the response mechanism been ignorable) and the imputed value. The factors are estimated using ratios of totals of auxiliary variables for respondents, non-respondents and the entire sample (or ranked transformations to limit the impact of influential observations). Intuitively, the performance of the method depends on the validity of the assumption made on the response mechanism, and the distribution of the variable of interest conditional on the auxiliary variables. Variance estimators are also elaborated for the estimator using the correction factor. An empirical study showed that the cor-

rection performed well, especially when the underlying model used in imputation described the population well. The performance does depend on the model implied by the transformation, and cannot be expected to perform well if the respondents differ from the non-respondents in some way other than that explained by the model.

2.2.2 Variance Estimation from Imputed Survey Data

A number of different methods have been proposed to develop variance estimators with good properties when imputation has been used. This section gives an overview of methods that are suitable for use with mean, ratio or regression imputations.

Multiple Imputation

A computer-intensive method that can be used to estimate the variance of an estimator is derived through Multiple Imputation. In this method, described in Rubin (1986), a number of completed datasets are created. Each value to be imputed is replaced independently in each dataset, using the same imputation method and respondent data. Rubin (1986) expresses the motivation behind the method as follows: "One imputed value cannot itself represent any uncertainty about which value to impute: If one value were really adequate, then that value was never missing." In order to estimate the variability of the imputed estimator, two components are used: the average within-imputation variance associated with the estimate and the between-imputation variance of the estimates. In order to implement this method in practice, the structure of the data storage facility must be more elaborate. As well, the method was designed for stochastic imputation: for deterministic imputa-

tion methods the same value would be imputed each time imputation is carried out using the same respondent data. In addition to this disadvantage, this method has been subjected to other criticisms, as in Binder and Sun (1996). An important condition for the validity of these results is that the imputation is "proper". Imputation is termed "proper" if adequate sampling variability is introduced among repeated imputations. Rubin (1986) states that a relatively small number of replications is adequate to produce efficient estimates when proper imputation methods are used.

Adjusted Jack-knife

Variance estimation through jackknife procedures was first studied by Rao and Shao (1992) under random hot-deck imputation. In this method, a typical jackknife procedure is used, but an adjustment is made to the jackknife values. Typically, in jackknife variance estimation under simple random sampling and complete response, n estimates are each calculated by deleting one sampled unit in turn (i.e. each estimate is based on $(n - 1)$ sampled units). In order to obtain consistent estimators of variance under imputation, Rao and Shao (1992) developed the following modification:

- (i) If the unit being dropped from the sample is an imputed value, no adjustment is required.
- (ii) If the unit being dropped from the sample is a reported value, the imputed values are re-calculated based on the responding units in the jackknife sample (excluding the dropped unit). This method is easily extended to ratio imputation by re-calculating the imputed values using the ratio imputation method.

Rancourt, Lee and Sarndal (1994) and Rao and Sitter (1995) extend the method to the case where different imputation methods are used depending on available auxiliary information on non-respondents (sometimes referred to as composite imputation). This involves a further adjustment to the jackknife values that depend on the unit omitted from the sample. Essentially, all values imputed by a method making use of the data from the dropped record are re-imputed based on the responding values in the jack-knife sample.

Model-Assisted Approach

Sarndal (1992) developed a methodology to calculate variance estimators that include terms representing the variance due to imputation for non-response. A model-assisted approach is adopted, which relies on a number of assumptions:

- (i) the pattern of non-response is INR or MCAR, and
- (ii) the reported values follow the imputation model. The variance is decomposed as

$$V_{tot} = V_{sam} + V_{imp} + 2V_{mix}$$

and each term is estimated individually:

$$\hat{V}_{tot} = \hat{V}_{sam} + \hat{V}_{imp} + 2\hat{V}_{mix}$$

where $\hat{V}_{sam}$ is the variance estimator calculated using the full data-set (imputed and reported values), and $\hat{V}_{imp}$ and $\hat{V}_{mix}$ are adjustment terms to reflect the fact that the imputed dataset may contain less-than-natural variability. For example, mean

imputation yields equal imputed values for all non-respondents yielding an estimate of $\hat{V}_{sam}$ that underestimates the sampling variance, $\hat{V}_{imp}$ is the term that reflects the variability of imputed values themselves, and $\hat{V}_{mix}$ is a covariance term between the sampling and imputation (which is zero for many imputation strategies).

In Lee and Rancourt (1994) the method is applied under a variety of conditions, and the results indicate that the procedure is robust to model mis-specification. This method has also been considered for implementation with generalized estimation software packages in Gagnon, Lee, Rancourt and Sarndal (1996). Drawbacks to the method are that it relies heavily on the assumptions surrounding the imputation model, and in practice, these model assumptions can be difficult or impossible to validate.

Reverse Approach

Essentially, the methods described above treat non-response as another random process which occurs after the selection of the sample. The variance estimates are then developed by treating non-response as a 2nd phase of sampling. A modification to this response path was considered by Fay (1991), and further developed in Shao and Steel (1999). According to this response path units exist in the population that will not respond if included in the sample. The realized sample will thus include a number of responding units and a number of non-responding units. This response path essentially reverses the order of non-response and sampling in the progression

from the population to the sample of respondents.

Population $\Rightarrow$ Census with non-respondents $\Rightarrow$ Sample with non-respondents

Through this response path the variance can be decomposed into two components:

$$V(\hat{Y}^I) = E_r V_s(\hat{Y}^I) + V_r E_s(\hat{Y}^I) \equiv V_1 + V_2,$$

where V_r and E_r are the variance and expectation with respect to the response mechanism, and V_s and E_s are the variance and expectation with respect to sampling, conditional on the response set. Shao and Steel (1999) propose the following generalized method to develop variance estimators using this response path.

Step 1: Obtain an expression for the imputed estimator as a function of estimated totals.

Step 2: Obtain an expression for the variance with respect to sampling, conditional on the response set, for example, using a Taylor linearization technique or some other appropriate method such as the jackknife. In this step, to estimate $E_r V_s(\hat{Y}^I)$, when the formulas are derived through a Taylor Linearization, the quantity can be estimated using the standard formulae for the variance of the estimated total under the sample design. Since the variance we are estimating is conditional on the response mechanism, the response indicator variables in the expression are considered constant.

Step 3: Obtain an expression for the variance with respect to the response set of the expected value under sampling, for example using a Taylor linearization technique or some other appropriate method. A plug-in estimator, v_2 , is then used to estimate this quantity from the sample.

Each step can be carried out under a design-based or a model-assisted approach. The estimator of $E_r V_s(\hat{Y}^I)$ is the same whether the design-based or model-assisted approach is used. The difference between these approaches will only be in the estimation of $V_r E_s(\hat{Y}^I)$. For the design-based variance estimator, we note that the only components of the expression for $E_s(\hat{Y}^I)$ that are treated as random variables are the response indicator variables. Under the assumption of missing completely at random (MCAR), the response indicator variables are independent Bernoulli variables, although the probabilities are unknown and must be estimated from the sample. Using this approach, we develop v_{2D} , a design-based estimator of $V_r E_s(\hat{Y}^I)$ which relies on the assumption of MCAR. An estimator of the total variance is given by $v_D = v_1 + v_{2D}$.

In the model-assisted we estimate $V_r E_s(\hat{Y}^I)$ by estimating the variance with respect to the model rather than the non-response mechanism. The model used in imputation is used to estimate the variance of y conditional on the auxiliary data. We need only to assume that the response mechanism is ignorable (INR), unlike in the design-based approach, but need to assume that the imputation model is correct. The imputation model is used to get an estimator, v_{2M} , of the model variance. An

estimate of the overall variance is given by $v_M = v_1 + v_{2M}$.

Shao and Steel (1999) identify the following important features of this method.

- The derivation of the formula in Step 1 can be generalized to many imputation methods, used either individually or in combination, although the formula becomes complicated when several imputation methods are used. Since v_1 is conditional on the realized set of respondents, the estimator is valid regardless of the response mechanism.
- In Step 2 under the design-based approach, the Taylor linearization technique can lead to tedious derivations if many imputation methods are used. An alternative to the Taylor linearization is given in the paper, to simplify the derivation.
- Since linearization or some other approximate method is used in Steps 2 and 3, the resulting variance estimators are asymptotically unbiased and consistent, but we could observe biases in the variance estimates if sample sizes are small.
- The method can handle non-negligible sampling fractions by including the term derived in Step 3 (v_2). If the overall sampling fraction is negligible, v_1 will serve as a good estimator of the total variance.

Shao and Steel (1999) note the results from this method under ratio imputation with negligible sampling fractions are equivalent to the results derived under the two-phase approach developed by Rao and Sitter (1995). Also interesting is the result that the design-based and model-assisted methods lead to asymptotically equivalent estimators under MCAR and ratio imputation. Shao and Steel (1999)

applied their method to a survey conducted by the U. S. Bureau of the Census which uses composite imputation, to demonstrate the derivation of variance estimators.

Demnati and Rao (2002) provide a simplified method for obtaining variance estimators, and illustrate its use for ratio imputation. This is done by taking the derivative of the expression from Step 1 with respect to the sampling weights, and results in formulas for Step 2 that are equivalent to the Taylor linearization technique but, in many cases, simplifies the derivations.

2.3 Scope of the Presented Work

We now turn to the research involved in this thesis.

2.3.1 Motivation

The motivation for this research is to develop an approximately unbiased variance estimator for use in monthly production for the Monthly Wholesale and Retail Trade Survey (MWRTS) of Statistics Canada. This survey is described in more detail in Section 4.2.1, but of note at this point is the fact that the survey uses a stratified simple random sampling design, and a combination of imputation methods, including ratio and regression, to impute data for non-responding units.

Many of the methods described above could be applied to the present case. Multiple imputation is not a viable option in our case due to the desire to maintain a simple

database structure (one complete dataset), and it is not directly applicable to deterministic imputation, as used in the MWRTS. The use of the composite imputation strategy makes the model-assisted approach less appealing since it was developed for individual imputation methods, although the theory could potentially be extended to the present case. The adjusted jackknife and response path reversal method appear to be most promising. However, since they yield asymptotically equivalent results, the relative ease of computations, the validity with non-negligible sampling fractions, and the straightforward application to combined imputation methods under the response path reversal method make it the method of choice for the current application.

2.3.2 Development Required

If the strata sample sizes are sufficiently large, the strata themselves may be used as imputation cells. Note that we then have self-weighting (i.e. equal sampling weights) within cells under stratified simple random sampling. However, in some business surveys, strata sample sizes are often as small as 8 or 10 so the number of within stratum donors for imputation tends to be small. Hence, strata are collapsed together to form imputation groups. The sampling weights within imputation cells now are not constant, (i.e. the sample is not self-weighting within cells). We derive basic results for individual imputation strategies in the class of regression imputation which incorporate sampling weights, when they vary within imputation groups. Formulas need to be developed for this framework by extending the results of Shao and Steel (1999).

We conduct simulation studies to verify the results obtained for non-self-weighting regression imputation. Through these simulation studies, we will also explore properties of the point and variance estimates under model mis-specification. Since in practice, all of the data required to study the non-response phenomenon is not available, indirect methods to estimate response probabilities are often used, and we are unable to confirm that the response mechanism is ignorable. If the auxiliary data are highly correlated with the variable of interest, the results of INR and NINR will be similar, but we study the performance of the estimators and variance estimators under various non-response models to determine whether they are robust to response mechanism mis-specification. As well, we will explore the impact of deviations in the population data from the model implied by the imputation. This will be done in terms of the first order of the imputation model (the expectation under the model) and the second order of the imputation model (the variance under the model) to explore the impact on the estimation of totals and their variances.

The reverse approach is also useful for studying the bias of the estimators. Although the bias is not discussed in Shao and Steel (1999), we will consider the bias under various conditions. Since the bias is not easily estimated from the sample, we will derive expressions for the bias based on quantities and relationships in the population to shed light on the population characteristics that lead to biased estimates.

As noted before, the derivation of the model-assisted variance estimator requires the

assumption of a model, with respect to which the estimator v_{2M} is derived. In the formulation of the design-based estimator, a different assumption is required, that of a MCAR mechanism. Due to these assumptions, different results are anticipated under different types of model mis-specification for the design-based and model-assisted versions of the variance estimates. The design-based and model-based methods will be evaluated and compared under these types of model-misspecification to determine their robustness.

We seek to develop a variance estimator that can be implemented in conjunction with the generalized software at Statistics Canada used to produce estimates of totals, that is, computed through an adjustment to micro data or a post-estimation process. The estimator will be evaluated in terms of its expectation, variance, and coverage properties through simulation studies. The individual terms of the variance estimator will be analyzed to determine their importance.

We also apply procedures to derive variance estimators for the Monthly Wholesale and Retail Trade Survey conducted by Statistics Canada and apply our results to obtain variance estimators for recent survey data.

Chapter 3

REGRESSION IMPUTATION

3.1 Introduction

In this chapter, regression imputation is considered. Regression imputation is presented in terms of the model used for motivation (referred to as the imputation model). Expressions are derived for the bias of the regression-imputed estimator, and associated variance estimators are developed. The regression-imputed estimator and associated variance estimators are evaluated through a simulation study described in section 3.1.1, under varying response rates, response mechanisms, and using data generated by different models (referred to as the data models). In some cases the data model corresponds to the imputation model and in other cases does not. We first describe the design of the simulation study.

3.1.1 Design of Simulation Study

A simulation study was conducted to evaluate the point and variance estimators corresponding to different imputation methods. These simulations were carried out

using variables of interest from different data models, under different response mechanisms and different response rates. Artificial data were generated to reflect characteristics typical of business surveys (skewed distribution, and relatively small sample sizes). A population of 1,025 units was created, with values of an auxiliary variable, x_1 , generated by a gamma model with mean 550,000 and standard deviation 250,000. A second auxiliary variable, x_2 , was then generated with expectation x_{1hi} and standard deviation $0.2x_{1hi}$. The population was then divided into three strata; stratum 1 contains the 25 units with the largest values of x_2 , stratum 2 contains the units with the next 250 largest values of x_2 , and stratum 3 contains the remaining 750 units (those with the smallest values of x_2). Several variables of interest were generated from different data models for study under each imputation method.

Probabilities of response were assigned to each unit using three classes of models;

- (i) Three missing completely at random (MCAR) models were used to assign equal response probabilities to each unit in the population, one for each overall probability of 0.5, 0.7 and 0.9.
- (ii) Three ignorable nonresponse (INR) models were used to assign unequal probabilities of response to each unit in the population, depending on their values of the auxiliary variable x_1 . Under the INR model, the unit response probabilities are assigned according to a monotonic increasing function of x_1 . The units in the population are ranked by x_1 , and equally spaced response probabilities from 0.2 to 0.8 are assigned to each unit for an overall probability of 0.5, from 0.5 to 0.9 for an

overall probability of response of 0.7, and from 0.8 to 1.0 for an overall probability of response of 0.9.

(iii) Three non ignorable nonresponse (NINR) models were used to assign unequal probabilities of response to each unit in the population, depending on their values of each variable of interest, y_{hi} . Under the NINR model, the unit response probabilities are assigned according to a monotonic increasing function of y_{hi} . The units in the population are ranked by y_{hi} and equally spaced response probabilities from 0.2 to 0.8 are assigned to each unit for an overall probability of 0.5, from 0.5 to 0.9 for an overall probability of response of 0.7, and from 0.8 to 1.0 for an overall probability of response of 0.9.

From the population, $K=10,000$ stratified simple random samples were drawn without replacement with a sample size of 25 from each stratum. Under this disproportionate sample allocation, we obtain sampling weights of 1, 10 and 30 for units in stratum 1, 2 and 3 respectively.

For each sample, k , a set of respondents was generated through independent Bernoulli trials with unit probability of response, p_{hi} , corresponding to each response mechanism (MCAR, INR or NINR) and average probability of response (0.5, 0.7, and 0.9). For the generated non-responding units, values of the variable of interest were imputed according to the specified imputation method. From each simulation, k , the estimate, $\hat{Y}_k$, of the population total for the variable of interest, $Y = \sum_U y_{hi}$, was calculated using the imputed Horvitz-Thompson estimator given in (2.3).

The bias ratio (BR) of the estimator, for a given response mechanism and response rate is calculated as

$$BR = B/SE, \quad (3.1)$$

where $B = \sum_k(\hat{Y}_k - Y)/K$ is the bias of the estimator and $SE = \sqrt{\sum_k(\hat{Y}_k - \bar{\hat{Y}})^2/K}$ is the standard error of the estimator with $\bar{\hat{Y}} = \sum_k \hat{Y}_k/K$. The relative root mean squared error ($RRMSE$) of the estimator, for a given response mechanism and response rate, is calculated as

$$RRMSE = \sqrt{MSE}/Y, \quad (3.2)$$

where $MSE = \sum_k(\hat{Y}_k - Y)^2/K$ is the mean squared error.

Several associated variance estimators, v_m , were used to estimate the variance of the imputed point estimator. The relative bias, $RB(v_m)$, of each variance estimator is given by $Bias(v_m)/MSE(\hat{Y})$. It is calculated as

$$RB(v_m) = \frac{\sum_k(v_{mk} - MSE)}{K \cdot MSE}, \quad (3.3)$$

where v_{mk} denotes the value of the variance estimator v_m for simulation k . The coverage rate, $CR(v_m)$, of the 95% confidence intervals based on each variance estimator, v_m , for a given response mechanism and response rate is

$$CR(v_m) = \sum_k I(\hat{Y}_k - 2\sqrt{v_{mk}} < Y < \hat{Y}_k + 2\sqrt{v_{mk}})/K, \quad (3.4)$$

where $I(z)$ is an indicator function equal to 1 if z is true, and equal to 0 if z is false.

3.2 The Regression-Imputed Estimator

The model believed to underly the data that could be used to motivate regression imputation is given by

$$y_{hi} = \mathbf{x}_{hi}'\mathbf{B} + \epsilon_{hi},$$

where $\mathbf{x}_{hi} = (1, x_{1hi}, x_{2hi}, \dots, x_{qhi})'$ is a column vector of auxiliary variables, $\mathbf{B} = (B_0, B_1, \dots, B_q)'$ is a column vector of parameters and the ϵ_{hi} follow a distribution with

$$E(\epsilon_{hi}) = 0, E(\epsilon_{hi}^2) = \sigma_{hi}^2, \text{ and } E(\epsilon_{hi}, \epsilon_{jk}) = 0 \text{ for } hi \neq jk$$

Under this model, different assumptions on the form of the variance function σ_{hi}^2 are possible but we restrict our discussion to models with equal variance, $V(\epsilon_{hi}^2) = \sigma^2$. Regression imputation is only possible when auxiliary data, $\mathbf{x}_{hi}$, are available for each unit in the sample, and it is most effective when a strong linear relationship exists between the auxiliary variables, $x_1, x_2, \dots, x_q$ and the variable of interest, y . Note that

(i) the mean model is a special case of the regression model with

$$q = 0, B_0 = \mu, \text{ and } \sigma_{hi}^2 = \sigma^2.$$

(ii) the ratio model is a special case of the regression model with

$$q = 1, B_0 = 0, B_1 = R, \text{ and } \sigma_{hi}^2 = x_{1hi}\sigma^2.$$

Because of this relationship, the results derived for regression imputation can be applied to mean or ratio imputation. By drawing an analogy between imputation using auxiliary data and estimation using auxiliary data (described in Sarndal (1992)) we

believe that regression imputation will be most efficient when the variable of interest is well described by the model and the auxiliary data. This analogy leads us to believe that regression imputation will be more efficient than mean imputation when the linear relationship between the auxiliary variables and the variable of interest is moderately strong. Also, regression imputation will be more efficient than ratio imputation if the regression model yields predicted values more highly correlated with the variable of interest than the ratio model.

In order to carry out regression imputation, we estimate the parameter B by the weighted least-squares estimate from the stratified random sample:

$$\hat{B}_r = \hat{T}_r^{-1} \hat{t}_r, \quad (3.5)$$

where $\hat{T}_r = \sum_s w_{hi} a_{hi} x_{hi} x_{hi}'$, and $\hat{t}_r = \sum_s w_{hi} a_{hi} x_{hi} y_{hi}$. We then impute each non-respondent value with $y_{hi}^* = x_{hi}' \hat{B}_r$. Thus, we obtain a completed dataset with values

$$y_{hi}^I = a_{hi} y_{hi} + (1 - a_{hi}) x_{hi}' \hat{B}_r,$$

where $a_{hi} = 1$ if unit (hi) is a respondent and $a_{hi} = 0$ otherwise. The regression-imputed estimator of the total, Y , can then be expressed as

$$\hat{Y}^I = \left(\sum_s w_{hi} x_{hi}' \right) \hat{B}_r, \quad (3.6)$$

with B_r as above if it is true that $\sum_s w_{hi} a_{hi} (y_{hi} - x_{hi}' B_r) = 0$.

3.2.1 Bias of the estimator

The approximate bias of the regression-imputed estimator (3.6) is given by

$$B(\hat{Y}^I) = E(\mathbf{X}'\hat{\mathbf{B}}_r) - Y \approx \mathbf{X}'\mathbf{B}_p - Y, \quad (3.7)$$

where $\mathbf{B}_p = [\sum_U p_{hi} \mathbf{x}_{hi} \mathbf{x}_{hi}']^{-1} \sum_U \mathbf{x}_{hi} y_{hi} p_{hi}$, $\mathbf{X}' = \sum_U \mathbf{x}_{hi}'$, $Y = \sum_U y_{hi}$, and the expectation $E(\cdot)$ is with respect to both sampling and response. The approximation (3.7) holds for reasonably large sample sizes. It follows from (3.7) that the bias of the regression-imputed estimator is approximately zero whenever the finite population covariance between the imputation errors, $y_{hi} - \mathbf{x}_{hi}'\mathbf{B}$, and the response probabilities, p_{hi} , is zero, or the response mechanism is uniform, i.e. $p_{hi} = p$ for all (hi) . Without one of these conditions satisfied, the relative bias of the imputed estimator $\hat{Y}^I$ can be non-negligible with small samples sizes or a non-ignorable non-response mechanism.

We now present some empirical results on the regression-imputed estimator (3.6).

(i) *Imputation Model Corresponds to Data Model*

Results presented in this section were obtained from regression imputation for non-response (using a model with an intercept and x_1 and x_2 as explanatory variables) when the variable of interest was generated using a regression model (with constant error variance). Results are presented for three models, created to study the effects of various degrees of model fit.

Model 1 was used to generate values of a variable of interest, y_1 , for each unit in the population: y_{1hi} was generated with expectation $100,000 + 0.7x_{1hi} + 0.3x_{2hi}$ and standard deviation 175,000. This results in a Pearson correlation coefficient of 0.9

between y_1 and the predicted value under the model.

Model 2 was used to generate values of a variable of interest, y_2 , for each unit in the population: y_{2hi} was generated with expectation $100,000 + 0.7x_{1hi} + 0.3x_{2hi}$ and standard deviation 500,000. This results in a Pearson correlation coefficient of 0.6 between y_2 and the predicted value under the model.

Model 3 was used to generate values of a variable of interest, y_3 , for each unit in the population: y_{3hi} was generated with expectation $100,000 + 0.7x_{1hi} + 0.3x_{2hi}$ and standard deviation 1,250,000. This results in a Pearson correlation coefficient of 0.3 between y_3 and the predicted value under the model.

Table 3.1 gives the bias ratio (BR) and relative root mean squared error ($RRMSE$) of the regression imputed estimators of the population totals of y_1, y_2 , and y_3 under various response rates and a MCAR mechanism (i.e., $p_{hi} = p$).

Table 3.1: Regression Imputation*: Properties of Point Estimator under MCAR

Variable of Interest	Response Rate	BR (%)	RRMSE (%)
y_1	0.5	0.27	5.87
	0.7	-0.19	5.43
	0.9	-0.24	5.13
y_2	0.5	0.24	13.78
	0.7	1.05	11.56
	0.9	0.81	10.33
y_3	0.5	-0.15	27.96
	0.7	-0.59	23.55
	0.9	-0.20	20.39

* Imputation model corresponds to data model

Table 3.1 shows that the regression-imputed point estimator has a small bias ratio ($\leq 2\%$) regardless of response rate, for each of the variables of interest. The relative root mean squared error ($RRMSE$) of the point estimator decreases as the response rate increases. Similarly, for a given response rate, $RRMSE$ decreases as the correlation between y and the predicted value under the data model increases.

Table 3.2 gives the bias ratio (BR) and relative root mean squared error ($RRMSE$) of the regression imputed estimators of the population totals of y_1 , y_2 , and y_3 under various response rates and the INR mechanism specified in section 3.1.1.

Table 3.2: Regression Imputation*: Properties of Point Estimator under INR

Variable of Interest	Response Rate	BR (%)	RRMSE (%)
y_1	0.5	1.51	6.06
	0.7	0.17	5.49
	0.9	-0.19	5.16
y_2	0.5	5.57	14.40
	0.7	2.48	11.84
	0.9	0.69	10.37
y_3	0.5	-6.25	30.35
	0.7	-2.49	24.06
	0.9	-0.79	20.59

* Imputation model corresponds to data model

Table 3.2 shows that the bias ratio of the point estimator is small (less than 5% in absolute value) for all three variables when the response rate is 0.7 or above. The bias ratios of the estimate of the totals of y_2 and y_3 are somewhat larger. The rel-

ative root mean squared error of the point estimator decreases as the response rate increases. Similarly, for a given response rate, RRMSE decreases as the correlation between y and the predicted value under the data model increases.

Table 3.3 gives the bias ratio (BR) and relative root mean squared error($RRMSE$) of the regression imputed estimators of the population totals of y_1, y_2 , and y_3 under various response rates and the NINR mechanism specified in section 3.1.1.

Table 3.3: Regression Imputation*: Properties of Point Estimator under NINR

Variable of Interest	Response Rate	BR (%)	RRMSE (%)
y_1	0.5	57.97	6.84
	0.7	26.88	5.65
	0.9	10.67	5.21
y_2	0.5	127.96	21.56
	0.7	66.84	14.09
	0.9	28.36	10.75
y_3	0.5	151.81	48.78
	0.7	82.65	30.22
	0.9	36.18	21.77

* Imputation model corresponds to data model

Table 3.3 shows that the point estimator is heavily biased under NINR. The bias ratio increases as the response rate decreases, and is higher when the variable of interest is not highly correlated with the predicted value under the data model: with a response rate of 0.5 the bias ratio is 152% for imputation of y_3 , compared to 58% for imputation of y_1 . Note that even with a response rate of 0.9, the bias ratio is above 25% for the variables of y_2 and y_3 . The relative root mean squared error of

the point estimator decreases as the response rate increases. Similarly, for a given response rate, RRMSE decreases as the response rate increases.

(ii) *Imputation Model does not Correspond to Data Model*

Results presented in this section were obtained from regression imputation for non-response (using x_1 and x_2 as auxiliary variables) when the variable of interest was generated using a model other than the linear regression model.

Model 4 was used to generate values of a variable of interest, y_4 , for each unit in the population: y_{4hi} was generated with expectation 487,000 and standard deviation 250,000. This corresponds to the mean model.

Model 5 was used to generate values of a variable of interest, y_5 , for each unit in the population: y_{5hi} was generated with expectation $200,000 + 0.000003x_{1hi}^2$ and standard deviation 1,250,000. This results in a Pearson correlation coefficient of 0.6 between y_5 and the predicted value under the data model. This model does not fall under the class of a linear regression model because of the quadratic term in x_1 .

Model 6 was used to generate values of a variable of interest, y_6 , for each unit in the population: y_{6hi} was generated with expectation $0.9x_{1hi} + 0.45x_{2hi}$ and standard deviation $2,000(0.9x_{1hi} + 0.45x_{2hi})$. This results in a Pearson correlation coefficient of 0.6 between y_6 and the predicted value under the data model. This model falls under the class of a linear regression model, but differs from the imputation model in terms of the error variance; in the imputation model, the error variance is assumed to be constant, whereas in Model 10 the error variance for each unit is dependent on the auxiliary values.

Table 3.4 gives the bias ratio (BR) and relative root mean squared error($RRMSE$) of the regression imputed estimators of the population totals of y_4 , y_5 , and y_6 under various response rates and a MCAR mechanism.

Table 3.4: Regression Imputation**: Properties of Point Estimator under MCAR

Variable of Interest	Response Rate	BR (%)	RRMSE (%)
y_4	0.5	-1.16	8.96
	0.7	-0.73	7.34
	0.9	-1.89	6.38
y_5	0.5	-1.48	26.05
	0.7	-0.73	22.27
	0.9	0.23	20.21
y_6	0.5	1.37	14.72
	0.7	0.71	12.76
	0.9	1.23	11.64

** Imputation model does not correspond to data model

Table 3.4 shows that the bias ratio of the regression-imputed estimator is small ($\leq 2\%$ in absolute value) for all variables of interest and all response rates under MCAR. The relative root mean squared error for the regression-imputed estimator of the total of each variable increases as the response rate decreases.

Table 3.5 gives the bias ratio (BR) and relative root mean squared error($RRMSE$) of the regression-imputed estimators of the population totals of y_4 , y_5 , and y_6 under various response rates and the INR specified in section 3.1.1.

Table 3.5 shows that the bias ratio of the regression-imputed estimator is above 10%

Table 3.5: Regression Imputation**: Properties of Point Estimator under INR

Variable of Interest	Response Rate	BR (%)	RRMSE (%)
y_4	0.5	0.77	9.29
	0.7	-0.59	7.38
	0.9	-1.42	6.35
y_5	0.5	-16.02	27.28
	0.7	-6.94	22.45
	0.9	-1.46	20.22
y_6	0.5	0.91	14.82
	0.7	0.56	12.71
	0.9	0.94	11.56

** Imputation model does not correspond to data model

in absolute value for y_5 when the response rate is low: 16% with a response rate of 0.5. This bias is associated with a covariance between the imputation residuals and the probabilities of response (generally, units with positive imputation residuals are more likely to respond). Since the probability of response is higher for units with higher values of x_{1hi} , we have the following relationship: $E_r E_s(x_{hi}' \hat{B}_r) < y_{hi}$. Thus, when we use the estimated regression co-efficients to impute the non-respondents, this yields a negative bias in the regression-imputed estimators. The bias ratios of the estimator of the total for y_4 and y_6 are small. Since data models used to generate y_4 and y_6 can be expressed in the form of a linear regression, regression imputation leads to approximately unbiased results; that is, the covariance between imputation residuals and probabilities of response is approximately zero.

Table 3.6 gives the bias ratio (BR) and relative root mean squared error($RRMSE$)

of the regression imputed estimators of the population totals of y_4, y_5 , and y_6 under various response rates and the NINR mechanism specified in section 3.1.1. Table 3.6

Table 3.6: Regression Imputation**: Properties of Point Estimator under NINR

Variable of Interest	Response Rate	BR (%)	RRMSE (%)
y_4	0.5	162.14	16.11
	0.7	88.31	9.69
	0.9	38.09	6.81
y_5	0.5	121.24	38.21
	0.7	63.74	26.09
	0.9	27.18	20.97
y_6	0.5	118.17	21.67
	0.7	61.13	14.72
	0.9	26.99	11.99

** Imputation model does not correspond to data model

shows that the bias ratio of the regression-imputed estimator is large for all the characteristics, y_4, y_5 and y_6 , especially when the response rate is low. The bias ratios observed for y_4, y_5 and y_6 are associated with the covariance between the imputation residuals, $y_{hi} - x_{hi}'\hat{B}_r$, and the probabilities of response (generally, units with negative imputation residuals are more likely to respond). Since the probability of response is higher for units with higher values of y_{hi} , we have the following relationship: $E_r E_s(x_{hi}'\hat{B}_r) > y_{hi}$. Thus, when we use the estimated regression parameter $\hat{B}_r$ to impute the non-respondents, this yields a positive bias in the regression-imputed estimators.

3.2.2 Variance Estimation

The fact that regression imputation tends to preserve the distribution of item values, leads us to believe that the absolute relative bias of v_0 given in 2.4, under regression imputation should be similar to the absolute relative bias of v_0 under ratio imputation; that is, smaller than the absolute relative bias under mean imputation. However, we do expect that the variance estimator v_0 will lead to an underestimation of the variance. Formulas for variance estimators which properly take into account the variability of imputed values are developed in Appendix A.

To estimate $E_r V_s[\hat{Y}^I]$ we use a Taylor linearization technique which gives the estimator

$$v_1 = \sum_s \left(1 - \frac{n_h}{N_h}\right) \frac{N_h^2}{n_h} \frac{(\hat{\xi}_{hi} - \hat{\xi}_{h\cdot})^2}{(n_h - 1)}, \quad (3.8)$$

where $\hat{\xi}_{hi} = y_{hi}^I + a_{hi} \hat{c}_{hi} \hat{e}_{hi}$, $\hat{c}_{hi} = a_{hi} [\sum_s w_{hi} (1 - a_{hi}) \mathbf{x}_{hi}'] \hat{\mathbf{T}}_r^{-1} \mathbf{x}_{hi}$, $\hat{\xi}_{h\cdot} = \sum_{s_h} \hat{\xi}_{hi} / n_h$, and $\hat{e}_{hi} = y_{hi} - \mathbf{x}_{hi}' \hat{\mathbf{B}}_r$. We can then use v_1 to estimate the total variance if the sampling fractions, $f_h = n_h / N_h$, are small.

The design-based estimator of $V_r E_s[\hat{Y}^I]$ is valid under a MCAR mechanism regardless of the imputation model, and it is calculated as

$$v_{2D} = N \hat{p} (1 - \hat{p}) \bar{s}_d^2, \quad (3.9)$$

with $\bar{s}_d^2 = \sum_s w_{hi} a_{hi} (\hat{c}_{1hi} + 1)^2 \hat{e}_{hi}^2 / \sum_s w_{hi} a_{hi}$, $\hat{p} = \sum_s w_{hi} a_{hi} / N$, and $\hat{e}_{hi}$ and $\hat{c}_{1hi}$ as above. We can use $v_D = v_1 + v_{2D}$ to estimate the total variance.

The model-assisted estimator of $V_r E_s[\hat{Y}^I]$ is valid under an INR mechanism when the imputation model is true, and it is calculated as

$$v_{2M} = s_d^2 \sum_s w_{hi} (a_{hi} \hat{c}_{1hi}^2 + (1 - a_{hi})), \quad (3.10)$$

with $s_d^2 = \sum_s w_{hi} a_{hi} \hat{\epsilon}_{hi}^2 / \sum_s w_{hi} a_{hi}$ and $\hat{c}_{1hi}$ as above. We can use $v_M = v_1 + v_{2M}$ to estimate the total variance.

(i) *Imputation model corresponds to Data Model*

Results presented in this section were obtained from regression imputation for non-response (using x_1 and x_2 as the auxiliary variables) when the variable of interest was generated using a regression model. Results are presented for three variables of interest, y_1 , y_2 , and y_3 described in section 3.2.1.

Table 3.7 gives the relative bias (RB) of the variance estimators of the regression-imputed estimators of the population totals of y_1 , y_2 , and y_3 under various response rates, and a MCAR mechanism. Recall that the correlations of y_1 , y_2 , and y_3 with the predicted value under the regression model are 0.9, 0.6 and 0.3 respectively.

As expected, Table 3.7 shows a negative relative bias (RB) for the naive variance estimator v_0 which increases in absolute value as the response rate decreases. Also, the absolute relative bias, $|RB|$, of v_0 becomes larger as the correlation between the predicted value under the imputation model and the variable of interest decreases (absolute relative bias of 40%, 70% and 75% for variables of interest with correlation 0.9, 0.6 and 0.3 respectively and a response rate of 0.5). Table 3.7 also shows negative relative biases for v_1 , v_D , and v_M , but $|RB|$ is small ($\leq 13\%$) even with a

Table 3.7: Regression Imputation*: Relative Bias (%) of Variance Estimators under MCAR

Variable of Interest	Response Rate	$RB(v_0)$	$RB(v_1)$	$RB(v_M)$	$RB(v_D)$
y_1	0.5	-39.48	-5.43	-4.42	-4.42
	0.7	-23.60	-3.63	-3.13	-3.13
	0.9	-8.01	-1.44	-1.29	-1.29
y_2	0.5	-69.72	-12.10	-10.38	-10.38
	0.7	-46.22	-5.32	-4.27	-4.27
	0.9	-17.76	-2.84	-2.50	-2.50
y_3	0.5	-74.90	-9.63	-7.68	-7.67
	0.7	-52.59	-7.04	-5.87	-5.87
	0.9	-19.65	-1.89	-1.49	-1.49

* Imputation model corresponds to data model

response rate of 0.5; the absolute relative bias of v_1 is slightly larger than that of v_D and v_M . The variance estimators v_M , and v_D have very similar relative biases for each variable of interest and each response rate. This is due to the fact that the difference between them has an expected value approximately equal to zero under MCAR.

Table 3.8 gives the relative bias (RB) of the variance estimators of the regression-imputed estimators of the population totals of y_1 , y_2 , and y_3 under various response rates, and the INR mechanism specified in section 3.1.1.

Table 3.8 shows a negative relative bias for the "naive" variance estimator, v_0 , which increases in absolute value as the response rate decreases. The relative bias becomes larger in absolute value as the correlation between the auxiliary variable and the variable of interest decreases. The table also shows a negative bias for the variance

Table 3.8: Regression Imputation*: Relative Bias (%) of Variance Estimators under INR

Variable of Interest	Response Rate	$RB(v_0)$	$RB(v_1)$	$RB(v_M)$	$RB(v_D)$
y_1	0.5	-44.58	-7.29	-6.18	-6.17
	0.7	-25.98	-3.92	-3.40	-3.39
	0.9	-9.30	-1.87	-1.72	-1.72
y_2	0.5	-74.46	-11.95	-10.06	-10.09
	0.7	-50.08	-6.06	-4.99	-4.99
	0.9	-19.54	-2.52	-2.17	-2.17
y_3	0.5	-80.28	-13.37	-11.42	-11.36
	0.7	-56.14	-6.23	-5.03	-5.02
	0.9	-22.36	-2.41	-2.00	-2.00

* Imputation model corresponds to data model

estimator v_1 , but its absolute relative bias ($|RB|$) is small compared to that of v_0 : $|RB|$ of v_1 is less than 15% regardless of the response rate or variable of interest. The relative bias of v_D and v_M are of a similar magnitude.

Table 3.9 gives the relative bias (RB) of the variance estimators of the regression-imputed estimators of the population totals of y_1 , y_2 , and y_3 under various response rates, and the NINR mechanism specified in 3.1.1.

Table 3.9 shows that the variance estimator, v_0 , is negatively biased under NINR, and the bias increases in absolute value with a decrease in response rate, and with a decrease in the correlation between the variable of interest and the predicted value under the imputation model. The table also shows a negative bias of a similar size for the other variance estimators, v_1 , v_D , and v_M . The absolute relative bias of these variance estimators is large with lower response rates, and with a lower cor-

Table 3.9: Regression Imputation*: Relative Bias (%) of Variance Estimators under NINR

Variable of Interest	Response Rate	$RB(v_0)$	$RB(v_1)$	$RB(v_M)$	$RB(v_D)$
y_1	0.5	-59.22	-30.80	-29.99	-29.97
	0.7	-31.23	-10.15	-9.67	-9.67
	0.9	-11.45	-3.96	-3.81	-3.81
y_2	0.5	-89.74	-66.12	-65.44	-65.45
	0.7	-65.53	-36.30	-35.59	-35.60
	0.9	-25.13	-9.90	-9.59	-9.59
y_3	0.5	-92.91	-72.37	-71.78	-71.78
	0.7	-72.10	-43.57	-42.87	-42.87
	0.9	-29.90	-13.35	-13.00	-13.00

* Imputation model corresponds to data model

relation between the variable of interest and the auxiliary variable (relative bias of approximately -66% and -72% when the correlation is 0.6 and 0.3 respectively with a response rate of 0.5). This bias is largely due to the contribution of the bias of the regression-imputed point estimator to its mean squared error under NINR which is not accounted in variance estimation.

Table 3.10 gives the Coverage Rates (CR) of 95% confidence intervals based on the variance estimators of the regression-imputed estimators of the population totals of y_1 , y_2 , and y_3 under various response rates, and a MCAR mechanism.

Table 3.10 shows that the confidence intervals constructed using v_0 have coverage rates relatively far below the nominal level for lower response rates (≤ 0.7) when the correlation between the variable of interest and the predicted value under the imputation model is not large (≤ 0.6): For example, the coverage rate drops to

Table 3.10: Regression Imputation*: Coverage Rates of 95% Confidence Intervals of Variance Estimators under MCAR

Variable of Interest	Response Rate	$CR(v_0)$	$CR(v_1)$	$CR(v_M)$	$CR(v_D)$
y_1	0.5	86.28	93.46	93.59	93.59
	0.7	90.72	93.96	94.04	94.04
	0.9	93.21	94.26	94.28	94.28
y_2	0.5	71.60	91.52	91.75	91.75
	0.7	83.72	93.55	93.70	93.70
	0.9	91.60	93.90	93.98	93.98
y_3	0.5	67.13	92.06	92.43	92.42
	0.7	81.53	93.34	93.50	93.50
	0.9	91.60	94.31	94.36	94.36

* Imputation model corresponds to data model

67% for the v_0 -based 95% confidence interval for the population total of y_3 when the response rate is 0.5. In general, the coverage rates of the confidence intervals constructed with v_1 , v_D and v_M are closer to the nominal level of 95%.

Table 3.11 gives the Coverage Rates (CR) of 95% confidence intervals based on the variance estimators of the population totals of y_1 , y_2 , and y_3 under various response rates, and the INR mechanism specified in 3.1.1.

Table 3.11 shows that the confidence intervals constructed using v_0 have coverage rates below the nominal level for lower response rates (≤ 0.7) and when the correlation between the variable of interest and the predicted value under the imputation model is not large (≤ 0.6): For example, the coverage rate drops to 63% for the v_0 -based 95% confidence interval for the population total of y_3 when the response rate is 0.5. The coverage rates of confidence intervals based on the other three variance

Table 3.11: Regression Imputation*: Coverage Rates of 95% Confidence Intervals of Variance Estimators under INR

Variable of Interest	Response Rate	$CR(v_0)$	$CR(v_1)$	$CR(v_M)$	$CR(v_D)$
y_1	0.5	84.31	93.15	93.37	93.36
	0.7	89.97	93.84	93.90	93.90
	0.9	93.26	94.41	94.42	94.42
y_2	0.5	66.91	91.21	91.52	91.54
	0.7	82.89	93.20	93.32	93.31
	0.9	91.40	93.89	93.97	93.97
y_3	0.5	62.85	91.13	91.42	91.39
	0.7	79.83	93.15	93.32	93.29
	0.9	90.85	93.88	93.90	93.90

* Imputation model corresponds to data model

estimators are closer to the nominal level of 95% for all response rates and variables of interest.

Table 3.12 gives the Coverage Rates (CR) of 95% confidence intervals based on the variance estimators of the population totals of y_1 , y_2 , and y_3 under various response rates, and the NINR mechanism specified in 3.1.1.

Table 3.12 shows that the coverage rates of the confidence intervals constructed using v_0 have coverage rates well below the nominal level for response rates below 0.9 and when the correlation between the variable of interest and the auxiliary variable is not large. The coverage rate drops to 28% for the v_0 -based 95% confidence interval for the population total of y_3 when the response rate is 0.5. The coverage rates of confidence intervals based on the other three variance estimators are close to the nominal level ($\geq 92\%$) with a response rate of 0.9, but the coverage rates

Table 3.12: Regression Imputation*: Coverage Rates of 95% Confidence Intervals of Variance Estimators under NINR

Variable of Interest	Response Rate	$CR(v_0)$	$CR(v_1)$	$CR(v_M)$	$CR(v_D)$
y_1	0.5	77.74	88.81	88.99	88.98
	0.7	88.89	93.30	93.36	93.36
	0.9	92.82	94.07	94.08	94.08
y_2	0.5	38.01	67.77	68.32	68.30
	0.7	73.13	86.45	86.63	86.63
	0.9	90.13	92.98	93.07	93.07
y_3	0.5	28.15	61.14	61.95	61.93
	0.7	67.89	84.58	84.89	84.89
	0.9	89.54	92.68	92.74	92.74

* Imputation model corresponds to data model

decrease sharply as the response rate decreases when the correlation between the variable of interest and the auxiliary variable used in imputation is not large (≤ 0.6).

(ii) *Imputation Model does not Correspond to Data Model*

Results presented in this section were obtained from regression imputation for non-response (using x_1 and x_2 as the auxiliary variables) when the variable of interest was generated using a model other than the regression model. Results are presented for the three variables of interest, y_4 , y_5 , and y_6 described in section 3.2.1.

Table 3.13 gives the relative bias (RB) of the variance estimators of the regression-imputed estimators of the population totals of y_4 , y_5 , and y_6 under various response rates, and a MCAR mechanism.

Table 3.13 shows that the naive variance estimator, v_0 , is negatively biased, and

Table 3.13: Regression Imputation**: Relative Bias (%) of Variance Estimators under MCAR

Variable of Interest	Response Rate	$RB(v_0)$	$RB(v_1)$	$RB(v_M)$	$RB(v_D)$
y_4	0.5	-75.16	-11.42	-9.33	-9.31
	0.7	-51.97	-5.02	-3.72	-3.72
	0.9	-19.91	-1.58	-1.13	-1.13
y_5	0.5	-59.55	-9.66	-8.09	-8.09
	0.7	-35.92	-1.68	-0.78	-0.78
	0.9	-9.79	2.56	2.85	2.85
y_6	0.5	-58.06	-7.98	-6.39	-6.36
	0.7	-36.50	-2.94	-2.04	-2.03
	0.9	-12.66	-0.91	-0.63	-0.63

** Imputation model does not correspond to data model

$|RB|$ increases as response rate decreases: $|RB|$ larger than 50% for each variable of interest when the response rate is 0.5. The absolute relative bias of v_1 is much smaller than that of v_0 . The other variance estimators v_M and v_D have smaller absolute relative biases than v_1 , especially when the response rate is low (≤ 0.7). The variance estimators v_M and v_D have small absolute relative biases ($\leq 10\%$) and their relative biases are similar (relative bias of v_S is generally larger in absolute value) for each variable of interest and each response rate. This is due to the fact that their expected values are approximately equal under MCAR.

Table 3.14 gives the relative bias (RB) of the variance estimators of the regression-imputed estimators of the population totals of y_4 , y_5 , and y_6 under various response rates, and the INR mechanism specified in section 3.1.1.

Table 3.14 shows that the naive variance estimator, v_0 , is negatively biased, with

Table 3.14: Regression Imputation**: Relative Bias (%) of Variance Estimators under INR

Variable of Interest	Response Rate	$RB(v_0)$	$RB(v_1)$	$RB(v_M)$	$RB(v_D)$
y_4	0.5	-77.42	-9.64	-7.36	-7.35
	0.7	-52.41	-4.07	-2.71	-2.72
	0.9	-18.88	-0.32	0.14	0.13
y_5	0.5	-59.74	-10.56	-8.88	-8.88
	0.7	-34.19	-0.91	0.05	0.04
	0.9	-8.46	2.77	3.06	3.05
y_6	0.5	-56.71	-4.41	-2.35	-2.56
	0.7	-33.77	-2.33	-1.31	-1.38
	0.9	-10.01	-0.03	0.27	0.26

** Imputation model does not correspond to data model

large $|RB|$ under lower response rates: $|RB|$ larger than 55% for each variable of interest when the response rate is 0.5. The absolute relative bias of v_1 is much smaller than that of v_0 . The variance estimators v_M and v_D have smaller absolute relative biases than v_0 and v_1 in general, but the absolute relative biases of v_M and v_D increase as the response rate decreases. Small differences are observed in the absolute relative biases of v_M and v_D when the response rate is low. In general, $|RB(v_D)| \approx |RB(v_M)|$ for each response rate and variable of interest. Although the assumption required for valid variance estimators is violated for these estimators under INR, the dominance of the first term in the variance, v_1 leads to less pronounced errors.

Table 3.15 gives the relative bias (RB) of the variance estimators of the regression-imputed estimators of the population totals of y_4 , y_5 , and y_6 under various response

rates, and the NINR mechanism specified in section 3.1.1.

Table 3.15: Regression Imputation**: Relative Bias (%) of Variance Estimators under NINR

Variable of Interest	Response Rate	$RB(v_0)$	$RB(v_1)$	$RB(v_M)$	$RB(v_D)$
y_4	0.5	-93.27	-76.08	-75.52	-75.52
	0.7	-73.22	-47.02	-46.30	-46.30
	0.9	-30.04	-14.17	-13.78	-13.79
y_5	0.5	-82.78	-62.05	-61.40	-61.39
	0.7	-53.92	-30.29	-29.65	-29.65
	0.9	-16.12	-5.46	-5.21	-5.21
y_6	0.5	-81.69	-61.71	-61.01	-61.06
	0.7	-52.05	-29.36	-28.69	-28.72
	0.9	-17.21	-7.34	-7.08	-7.09

** Imputation model does not correspond to data model

Table 3.15 shows that the naive variance estimator, v_0 , is negatively biased, with large $|RB|$ under lower response rates: $|RB|$ larger than 80% for each variable of interest when the response rate is 0.5. The absolute relative bias of v_1 is smaller than that of v_0 but is still high (60% or higher) for low response rates(0.5). The three variance estimators v_1 , v_M and v_D have absolute relative biases smaller than v_0 , and are similar in size (the absolute relative bias of v_1 is slightly larger than that of the others). The relative bias of the variance estimators is largely due to the contribution of the bias of the regression-imputed estimator under NINR to its mean squared error, which is not accounted in variance estimation.

Table 3.16 gives the Coverage Rates (CR) of 95% confidence intervals based on the variance estimators of the regression-imputed estimators of the population to-

totals of y_4 , y_5 , and y_6 under various response rates, and a MCAR mechanism.

Table 3.16: Regression Imputation**: Coverage Rates of 95% Confidence Intervals of Variance Estimators under MCAR

Variable of Interest	Response Rate	$CR(v_0)$	$CR(v_1)$	$CR(v_M)$	$CR(v_D)$
y_4	0.5	66.18	91.90	92.23	92.23
	0.7	81.76	93.59	93.81	93.80
	0.9	91.17	94.20	94.24	94.24
y_5	0.5	76.78	92.72	92.95	92.94
	0.7	86.84	94.09	94.20	94.20
	0.9	92.69	94.59	94.62	94.62
y_6	0.5	77.62	92.49	92.77	92.76
	0.7	86.77	93.90	94.02	94.02
	0.9	92.25	94.10	94.13	94.12

** Imputation model does not correspond to data model

Table 3.16 shows that the coverage rates of confidence intervals constructed using v_0 are below the nominal level of 95%, and are farther from the nominal level (approximately 65-80%) when the response rate is low (0.5). The coverage rates of the other variance estimators are close to the nominal level of 95% for all variables of interest and all response rates, and are highest when the response rates are high.

Table 3.17 gives the Coverage Rates (CR) of 95% confidence intervals based on the variance estimators of the population totals of y_4 , y_5 , and y_6 under various response rates, and the INR mechanism specified in section 3.1.1.

Table 3.17 shows that the coverage rates of confidence intervals constructed using v_0 are below the nominal level, and are far below the nominal level (approximately 65-80%) when the response rate is 0.5. The coverage rates of the other three variance

Table 3.17: Regression Imputation**: Coverage Rates of 95% Confidence Intervals of Variance Estimators under INR

Variable of Interest	Response Rate	$CR(v_0)$	$CR(v_1)$	$CR(v_M)$	$CR(v_D)$
y_4	0.5	64.65	91.95	92.28	92.28
	0.7	81.91	93.64	93.80	93.80
	0.9	91.60	94.46	94.53	94.53
y_5	0.5	76.36	92.12	92.38	92.38
	0.7	87.10	93.93	94.08	94.08
	0.9	92.94	94.53	94.55	94.55
y_6	0.5	78.89	93.19	93.43	93.38
	0.7	87.24	93.67	93.80	93.79
	0.9	92.51	94.30	94.33	94.32

** Imputation model does not correspond to data model

estimators are closer to the nominal level when the response rate is high(0.9). However, the coverage rates for the population total of y_5 decrease for all four variance estimators as the response rate decreases due to the bias of the regression-imputed estimator which is not accounted in variance estimation, and the negative bias of the variance estimates.

Table 3.18 gives the Coverage Rates (CR) of 95% confidence intervals based on the variance estimators of the population totals of y_4, y_5 , and y_6 under various response rates, and the NINR mechanism specified in section 3.1.1.

Table 3.18 shows that the coverage rates of confidence intervals constructed using v_0 are always below the nominal level, and are far below the nominal level (approximately 25-55%) when the response rate is 0.5. The coverage rates of the other variance estimators also decrease with a decrease in response rate(55-75% with a

Table 3.18: Regression Imputation**: Coverage Rates of 95% Confidence Intervals of Variance Estimators under NINR

Variable of Interest	Response Rate	$CR(v_0)$	$CR(v_1)$	$CR(v_M)$	$CR(v_D)$
y_4	0.5	25.83	56.54	57.24	57.23
	0.7	66.36	82.61	82.89	82.87
	0.9	89.05	92.39	92.45	92.45
y_5	0.5	51.01	73.27	73.80	73.80
	0.7	81.09	89.50	89.64	89.64
	0.9	92.68	94.35	94.37	94.37
y_6	0.5	53.12	74.87	75.43	75.37
	0.7	82.81	90.52	90.73	90.74
	0.9	92.06	93.83	93.85	93.85

** Imputation model does not correspond to data model

response rate of 0.5), although not as sharply as for v_0 . The decrease in coverage rates of the confidence interval is due to the large relative bias in the regression-imputed estimator under NINR and the underestimation of variance associated with the variance estimators (especially v_0).

3.3 Summary of Findings

In regression imputation, both the response mechanism and the mechanism used to generate the data determine the validity of the point and variance estimators. Note that the variance of the point estimator is lower when the data are generated by a regression model.

Bias of the Estimator:

If the response mechanism is MCAR, the point estimator resulting from regression imputation is approximately unbiased. If the response mechanism is INR, and the data model is a form of the regression model, the point estimator resulting from regression imputation is approximately unbiased. However, if the response mechanism is INR and the data model is not the regression model, or the response mechanism is NINR the estimator can be heavily biased. To reduce the risk of bias, if the response mechanism cannot be verified to be MCAR, caution is advised in using regression imputation to treat non-response. If the response mechanism is INR, extra attention should be given to the development of the imputation model, as erroneously specifying the imputation model could lead to a bias in the estimates. The bias is especially large if the response rate is low, or the correlation between the variable of interest and the predicted values under the imputation model is low. If the response mechanism is NINR, a bias will result in the imputed point estimator regardless of the specification of the imputation model. Given that auxiliary data are available (the variable used in imputation), it may be possible to correct for the bias introduced under NINR (for example using an extension of the method proposed in Rancourt, Sarndal and Lee (1995)). It is important to note that the correction will again be based on another model. This model may be implicit or explicit, but is required to model the imputation errors conditional on the auxiliary variable. If this model fails to hold, the bias-corrected estimator may perform as poorly or worse than the regression-imputed point estimator.

Variance Estimation:

If regression imputation is used, variance estimation should be carried out to reflect the imputed values so the "naive" variance estimator is not recommended. In most situations, the performances of v_1 , v_D and v_M are much better than v_0 in terms of bias, stability (assessed through the *RSE*) and confidence interval coverage rates. The variance estimators v_D and v_M perform slightly better in terms of bias and coverage rates than v_1 . If the imputation model holds the effect of the response mechanism on variance estimation is summarized as follows: under MCAR and INR mechanisms, the variance estimators v_1 , v_D and v_M performed well in terms of bias and coverage rates, however none performed well under the NINR mechanism (with low response rates). If the imputation model does not hold with respect to its mean, the variance estimators still perform well in terms of relative bias and coverage rates under MCAR, but their performance deteriorates under INR and NINR mechanisms due to the bias of the point estimator. If the imputation model holds with respect to its mean, but does not hold with respect to the error variance, the performance of the variance estimators is similar to when the imputation model holds; they perform well unless the response mechanism is NINR with relatively low response rates.

Design-based vs. Model-assisted Variance Estimation:

In the simulations, the difference between v_M and v_D was very small. This is because the two estimators only differ in their estimate of $V_r E_s(\hat{Y}^I)$ which contributes a very small portion of the total variance due to the low sampling fractions used in the simulations. An analysis carried out conditional on $\bar{x}_r$ could be used to compare the performance of v_M and v_D , and the differences would be more pronounced. This

difference is not apparent from the unconditional results. As well, in a situation with larger sampling fractions such as a census, the difference would be more pronounced. We recommend the use of v_M since it is developed in the context of an INR mechanism, and is also valid under MCAR. Since this estimator is only valid when the imputation model holds, extensive model validation is recommended to ensure good properties. If the response mechanism is NINR the performance of all of the variance estimators studied is poor.

Chapter 4

APPLICATION OF RESULTS

In this chapter, we consider the estimator resulting from composite imputation; that is, when several imputation methods have been used to impute values for non-respondents. We consider the composite-imputed point estimator in terms of its bias and variance, and develop associated variance estimators. In Section 4.1, we derive an expression for the bias of the estimator under general composite imputation strategies, and outline the procedure to develop associated variance estimators. In Section 4.2 we study the composite-imputed estimator used in a large-scale survey at Statistics Canada; we evaluate the potential for bias from the point estimator, and develop variance estimators. We also present results from applying these variance estimators to survey data.

4.1 Composite Imputation

In practical situations, a composite imputation strategy is often used where a number of different imputation methods are used to impute different non-responding units, or imputed values themselves are used as auxiliary variables in imputation. This is often necessary due to unavailability of auxiliary data because of non-response, linkage problems, or some other cause. The method used to impute each unit is determined according to a pre-determined order of preference; for example, if the auxiliary variables required to impute a non-responding unit via method 1 are not available, we impute via method 2, and so on. The order of preference for the imputation methods is often developed through analytical studies carried out to compare properties of imputed values resulting from each method (such as bias and variance) and the properties are used in ranking the imputation methods to develop the order of preference.

4.1.1 Bias of the Estimator

As with imputation by an individual method, a bias can be introduced by composite imputation. The bias introduced is a function of the bias arising from each individual imputation method included in the strategy. In this framework, the bias is given by

$$B(\hat{Y}_I) = \sum_U \sum_m p_{mhi} \epsilon_{mhi}, \quad (4.1)$$

where p_{mhi} is the probability that the i th unit in the h th stratum is imputed using imputation method m and ϵ_{mhi} is the imputation residual for the i th unit in the h th stratum under imputation method m .

It follows from (4.1) that the bias of the composite imputed estimator is zero if (i) $\sum_U \epsilon_{mhi} = 0$ for each method, m , and (ii) the covariance between the probability that a unit is imputed by the method, p_{mhi} , and the residual from the imputation model, ϵ_{mhi} , is zero. As we have seen in previous sections, this would be the case for mean, ratio or regression imputation under MCAR regardless of the imputation model or under INR when the imputation model corresponds to the data model. The probability of a unit being imputed by method m depends on three factors; the probability of response, the imputation strategy (the specified order of preference), and the availability of auxiliary data. In order to study the bias, not only do we need to assume a model to describe the probability of response for each unit (hi), but also a model to describe the probability of availability of each auxiliary variables, $\mathbf{x}_{hi}$, is required since non-response to these variables determines the method used to impute non-responding units.

We can also consider the bias of the estimator by assuming that a given data model is true. Since we assume that all of the data are generated by the same data model (regardless of whether the auxiliary variables are available or not), a bias can be introduced when an imputation method based on a model other than the data model is used to impute a given unit. A logical assumption is that the population model corresponds to the imputation model underlying the imputation method that would be used if all the auxiliary data were available: $y_{hi} = \mathbf{x}_{hi}'\mathbf{B}^* + \epsilon_{hi}^*$ where $\mathbf{x}_{hi} = (x_{1hi}, \dots, x_{qhi})$ is a column vector of auxiliary variables and $\mathbf{B}^* = (\beta_0^*, \dots, \beta_q^*)$

is a row vector of parameters, $E(\epsilon_{hi}^*) = 0$, $V(\epsilon_{hi}^*) = \sigma_{hi}^{*2}$ and $COV(\epsilon_{hi}^*, \epsilon_{jk}^*) = 0$ for $(hi) \neq (jk)$. If each imputation method, m , can be expressed in the form of a regression model with co-efficients $B^m = (B_0^m, \dots, B^m)'$ the bias is

$$B_{model}(\hat{Y}^I) = \sum_{(m>1)} \sum_U p_{mhi} \mathbf{x}_{hi}' (B^m - B^*), \quad (4.2)$$

where p_{mhi} denotes the probability that method m is used to impute unit (hi) .

From (4.2) it follows that a bias may be introduced whenever imputation by a method other than that corresponding to the data model is used. As with individual imputation methods if p_{mhi} is constant for all m , the bias will be approximately zero. This special case corresponds to MCAR for the variable of interest, and each auxiliary variable.

4.1.2 Variance Estimation

In order to estimate the variance of a composite-imputed estimator, the same general procedure can be used as for a single imputation strategy, although the derivation is more tedious when several imputation methods are used. We follow the three steps outlined in Shao and Steel (1999):

In the next section we apply this procedure to the composite-imputed estimator used in a large-scale survey conducted by Statistics Canada, the Monthly Wholesale and Retail Trade Survey, and present the results. We also present variance estimates from the survey data and compare the results with other variance estimates.

4.2 Application to Survey Data

4.2.1 The Monthly Wholesale Retail Trade Survey

The Monthly Wholesale Retail Trade Survey (MWRTS), conducted by Statistics Canada, measures the total sales and inventories for a sample of wholesale and retail trade businesses in Canada on a monthly basis. This information is used to estimate the total monthly sales and total inventories for separate industrial groups and geographic regions in Canada. The estimates from the MWRTS are used in many ways. The monthly totals are used in the calculation of the national Gross Domestic Product (GDP), and in calculating provincial equalization payments. As well, they are used by industry analysts as an important economic indicator.

Target Population and Survey Frame

The target population for the MWRTS consists of all businesses involved in Wholesale or Retail trade in Canada. The survey frame is derived from the Central Frame Data Base of Statistics Canada which stores information on businesses in Canada. In addition to providing a list of active businesses in Canada through unique statistical identifiers, this database provides information to classify each business in terms of industry, geography, and annual revenues. In particular, for the MWRTS, the province of domicile, the North American Industrial Classification System (NAICS) code, and the estimate of annual revenues are used.

Sample Design

The MWRTS uses a stratified design, and draws a simple random sample from each

stratum independently. Strata are formed according to Geography, Industry and Size using variables obtained from the survey frame; Province of domicile for Geographic stratification, North American Industrial Classification System (NAICS) codes for industrial stratification, and the estimated annual revenues for Size stratification. The sub-population belonging to each Geography and Industry are referred to as a cell. In each cell, the estimated annual revenues are used to divide the population into three size strata using the method elaborated in Lavalley and Hidioglou (1988). In each cell, the strata of largest companies are selected in the sample with certainty. The companies in other strata are subjected to sampling at fractions determined using an allocation with stratum sample size proportional to the stratum total of estimated annual revenues. The same businesses remain in the sample month after month; that is no sample rotation is applied.

Data Collection

Data collection is carried out using several methods. For some companies, a questionnaire is mailed, to be completed and returned via mail. For other companies, a Computer Assisted Telephone Interview is conducted to obtain responses. Typically, respondent data is obtained for approximately 80 to 90% of the sampled units.

Treatment of Non-Response

Non-response to the survey is treated using an imputation system developed specifically for the MWRTS. The system uses a composite imputation strategy involving several imputation methods, depending on the availability of auxiliary data. The

imputation strategy is described in more detail in section 4.2.2.

Estimation

In order to produce estimates of total monthly sales and inventories for the population, the Horvitz-Thompson estimator is used. In this chapter, we limit our discussion to the estimation of total sales, although the results could be extended to the estimation of total inventories. Variance estimation uses v_0 given in (2.4); that is, the values obtained through imputation are treated as if they are respondent values. It is well known that this leads to an underestimation of the variance. We derive alternative variance estimators in section 4.2.4.

4.2.2 Imputation Strategy

For the MWRTS, a composite imputation strategy is used to take advantage of highly correlated auxiliary data from previous occasions of the survey and from the survey frame. The imputation method is chosen depending on the availability of auxiliary information.

We will use the following notation throughout this chapter. Let y_{hi} denote the sales for unit (hi) collected by the survey for the current month, x_{1hi} denote the sales for unit (hi) collected by the survey for the previous month, x_{2hi} denote the sales for unit (hi) collected by the survey for the same calendar month of the previous reference year, and x_{3hi} denote the estimate of annual revenue for unit (hi) for the current year (derived from administrative data). The response indicator vari-

able, a_{hi} , used in individual imputation methods is replaced by a vector of indicator variables, $\mathbf{a}_{hi} = (a_{0hi}, a_{1hi}, a_{2hi})$, where $a_{0hi} = 1$ if y_{hi} is reported, and 0 otherwise, $a_{1hi} = 1$ if x_{1hi} is reported and 0 otherwise, and $a_{2hi} = 1$ if x_{2hi} is reported and 0 otherwise. Note that x_{3hi} is available for every unit (hi) in the population.

The imputation strategy is as follows:

- If y_{hi} is reported ($a_{0hi} = 1$), no imputation is required ,

If y_{hi} is not reported, we consider the following cases:

- If both x_{1hi} and x_{2hi} are available ($(1 - a_{0hi})a_{1hi}a_{2hi} = 1$), then the value is imputed via Imputation Model 1: regression imputation with both x_{1hi} and x_{2hi} as auxiliary variables,
- If only x_{1hi} is available ($(1 - a_{0hi})a_{1hi}(1 - a_{2hi}) = 1$), then the value is imputed via Imputation Model 2: ratio imputation with x_{1hi} as the auxiliary variable
- If only x_{2hi} is available ($(1 - a_{0hi})(1 - a_{1hi})a_{2hi} = 1$), then the value is imputed via Imputation Model 3: ratio imputation with x_{2hi} as the auxiliary variable
- If neither x_{1hi} nor x_{2hi} is available ($(1 - a_{0hi})(1 - a_{1hi})(1 - a_{2hi}) = 1$), then the value is imputed via Imputation Model 4: ratio imputation with x_{3hi} as the auxiliary variable.

In estimating the model parameters for each imputation model, we can use the indicator variables $\mathbf{a}_{hi}$ to identify which units are used in the calculations: $a_{0hi}a_{1hi}a_{2hi} = 1$ indicates that the record is used in the estimation of the parameters of imputation model 1, $a_{0hi}a_{1hi} = 1$ indicates that the record is used in the estimation of the

parameter of imputation model 2, $a_{0hi}a_{2hi} = 1$ indicates that the record is used in the estimation of the parameter of imputation model 3, and $a_{0hi} = 1$ indicates that the record is used in the estimation of the parameter of imputation model 4. Note that according to the definition of a_{hi} , if a value of x_{1hi} is imputed, $a_{1hi} = 0$, and if a value of x_{2hi} is imputed, $a_{2hi} = 0$; that is, units with imputed values from previous occasions of the survey are not used to estimate imputation model parameters.

Assumptions

For the MWRTS, the imputation strategy is complicated by a collapsing strategy used in forming the imputation groups. The sample design yields strata with small sample sizes, and in some cases, the strata are not homogeneous in terms of detailed industry classifications. To form imputation groups that are homogeneous enough to yield accurate imputations, some imputation groups are created which cut across strata. If a pre-specified number of recipients are not available within these groups, they are grouped together to form larger imputation groups until the minimum number of responding units is available within each imputation group. In this situation, the imputation groups themselves are dependent on sampling and response. Formulas have not been developed to reflect the process where imputation groups are aggregated until a minimum number of respondents is available (i.e. the imputation groups themselves are dependent on the sample). This is addressed as an area for future work in section 5.2. Typically, imputation can be carried out at the cell (Industry group by Geographic region) level, so for the development of variance estimators, we assume that imputation groups are mutually exclusive, and are formed at the cell level.

4.2.3 Bias of the Estimator

The composite-imputed estimator for each cell is then expressed as

$$\hat{Y}^I = \sum_s w_{hi} [a_{0hi} y_{hi} + (1 - a_{0hi})(a_{1hi} a_{2hi} \mathbf{x}_{hi}' \hat{\mathbf{B}}_r + a_{1hi}(1 - a_{2hi}) \hat{R}_{r1} x_{1hi} + (1 - a_{1hi}) a_{2hi} \hat{R}_{r2} x_{2hi} + (1 - a_{1hi})(1 - a_{2hi}) \hat{R}_{r3} x_{3hi})],$$

where $\hat{\mathbf{B}}_r = \hat{\mathbf{T}}_r^{-1} \hat{\mathbf{t}}_r$, $\hat{\mathbf{T}}_r = \sum_s a_{0hi} a_{1hi} a_{2hi} w_{hi} \mathbf{x}_{hi} \mathbf{x}_{hi}'$, $\hat{\mathbf{t}}_r = \sum_s a_{0hi} a_{1hi} a_{2hi} w_{hi} \mathbf{x}_{hi} y_{hi}$, $\hat{R}_{r1} = \frac{\sum_s a_{0hi} a_{1hi} w_{hi} y_{hi}}{\sum_s a_{0hi} a_{1hi} w_{hi} x_{1hi}}$, $\hat{R}_{r2} = \frac{\sum_s a_{0hi} a_{2hi} w_{hi} y_{hi}}{\sum_s a_{0hi} a_{2hi} w_{hi} x_{2hi}}$, and $\hat{R}_{r3} = \frac{\sum_s a_{0hi} w_{hi} y_{hi}}{\sum_s a_{0hi} w_{hi} x_{3hi}}$.

The bias of the imputed estimator is given by

$$B(\hat{Y}_I) = \sum_U (p_{1hi} \epsilon_{1hi} + p_{2hi} \epsilon_{2hi} + p_{3hi} \epsilon_{3hi} + p_{4hi} \epsilon_{4hi}), \quad (4.3)$$

where $p_{1hi} = E((1 - a_{0hi}) a_{1hi} a_{2hi})$, $p_{2hi} = E((1 - a_{0hi}) a_{1hi} (1 - a_{2hi}))$, $p_{3hi} = E((1 - a_{0hi})(1 - a_{1hi}) a_{2hi})$, and $p_{4hi} = E((1 - a_{0hi})(1 - a_{1hi})(1 - a_{2hi}))$. As well, $\epsilon_{1hi} = y_{hi} - (\sum_U p_{(1)hi} \mathbf{x}_{hi}') [\sum_U p_{(1)hi} \mathbf{x}_{hi} \mathbf{x}_{hi}'] (\sum_U p_{(1)hi} \mathbf{x}_{hi} y_{hi})$, $\epsilon_{2hi} = y_{hi} - \frac{\sum_U p_{(2)hi} y_{hi}}{\sum_U p_{(2)hi} x_{1hi}} x_{1hi}$, $\epsilon_{3hi} = y_{hi} - \frac{\sum_U p_{(3)hi} y_{hi}}{\sum_U p_{(3)hi} x_{2hi}} x_{2hi}$, $\epsilon_{4hi} = y_{hi} - \frac{\sum_U p_{(4)hi} y_{hi}}{\sum_U p_{(4)hi} x_{3hi}} x_{3hi}$ with $p_{(1)hi} = E(a_{0hi} a_{1hi} a_{2hi})$, $p_{(2)hi} = E(a_{0hi} a_{1hi})$, $p_{(3)hi} = E(a_{0hi} a_{2hi})$, $p_{(4)hi} = E(a_{0hi})$. That is, the probability that unit (hi) imputed by method m is p_{mhi} and the probability that unit (hi) is available for use in estimating the parameters of imputation model m is $p_{(m)hi}$.

It follows from (4.3) that if $\sum_U \epsilon_{mhi} = 0$ for all m and the finite population covariance between p_{mhi} and ϵ_{mhi} is zero for each imputation method m , the resulting estimator will be unbiased. Since the p_{mhi} are functions of the response indicators to previous occasions of the survey, if the response probabilities are independent for

each occasion of the survey, and the mechanism is MCAR or IGNR for each occasion of the survey (even if the probabilities associated with different occasions of the survey differ) our estimator will be approximately unbiased. This is true because under these assumptions, p_{mhi} can be expressed as the product of probabilities of response for each occasion of the survey, and under MCAR or IGNR we have seen that $\sum_U p_{mhi} \epsilon_{mhi} \approx 0$ for regression imputation.

Model Bias

If we assume that the population data are generated by Model 1, and the response mechanism is ignorable, we can express the bias of the estimator as

$$B_{model}(\hat{Y}^I) = \sum_U [p_{2hi} \mathbf{x}_{hi}' (\mathbf{B}_1 - \mathbf{B}_2) + p_{3hi} \mathbf{x}_{hi}' (\mathbf{B}_1 - \mathbf{B}_3) + p_{4hi} \mathbf{x}_{hi}' (\mathbf{B}_1 - \mathbf{B}_4)], \quad (4.4)$$

where $\mathbf{B}_1 = (\sum_U \mathbf{x}_{hi} \mathbf{x}_{hi}') (\sum_U \mathbf{x}_{hi} y_{hi})$, $\mathbf{B}_2 = (0, \frac{\sum_U y_{hi}}{\sum_U x_{1hi}}, 0, 0)'$, $\mathbf{B}_3 = (0, 0, \frac{\sum_U y_{hi}}{\sum_U x_{2hi}}, 0)'$, and $\mathbf{B}_4 = (0, 0, 0, \frac{\sum_U y_{hi}}{\sum_U x_{3hi}})'$. Estimates of (4.4) that use only the data from respondents are only valid under strong model assumptions.

4.2.4 Variance Estimation

We follow the approach elaborated in Shao and Steel (1999).

Step 1: Obtain an expression for the imputed estimator as a function of estimated totals

Note that $\hat{Y}_I \approx \sum_s w_{hi} \hat{\xi}_{hi}$, where $\hat{\xi}_{hi} = y_{hi}^I + a_{0hi} a_{1hi} a_{2hi} \hat{c}_{1hi} \hat{\epsilon}_{1hi} + a_{0hi} a_{1hi} \hat{c}_2 \hat{\epsilon}_{2hi} + a_{0hi} a_{2hi} \hat{c}_3 \hat{\epsilon}_{3hi} + a_{0hi} \hat{c}_4 \hat{\epsilon}_{4hi}$, $\hat{c}_{1hi} = [\sum_s w_{hi} (1 - a_{0hi} a_{1hi} a_{2hi}) \mathbf{x}_{hi}'] [\sum_s w_{hi} a_{0hi} a_{1hi} a_{2hi} \mathbf{x}_{hi} \mathbf{x}_{hi}']^{-1} \mathbf{x}_{hi}$,

$\hat{c}_2 = \sum_s w_{hi}(1 - a_{0hi})a_{1hi}(1 - a_{2hi})x_{1hi} / \sum_s w_{hi}a_{0hi}a_{1hi}x_{1hi}$, $\hat{c}_3 = \sum_s w_{hi}(1 - a_{0hi})(1 - a_{1hi})a_{2hi}x_{1hi} / \sum_s w_{hi}a_{0hi}a_{2hi}x_{2hi}$, $\hat{c}_4 = \sum_s w_{hi}(1 - a_{0hi})(1 - a_{1hi})(1 - a_{2hi})x_{3hi} / \sum_s w_{hi}a_{0hi}x_{3hi}$,
 $\hat{e}_{1hi} = y_{hi} - \mathbf{x}_{hi}'\hat{\mathbf{B}}_r$, $\hat{e}_{2hi} = y_{hi} - \hat{R}_{r1}x_{1hi}$, $\hat{e}_{3hi} = y_{hi} - \hat{R}_{r2}x_{2hi}$ and $\hat{e}_{4hi} = y_{hi} - \hat{R}_{r3}x_{3hi}$.
 Note that $\sum_s w_{hi}a_{0hi}a_{1hi}a_{2hi}\hat{c}_{1hi}\hat{e}_{1hi} = 0$, $\sum_s w_{hi}a_{0hi}a_{1hi}\hat{c}_2\hat{e}_{2hi} = 0$, $\sum_s w_{hi}a_{0hi}a_{2hi}\hat{c}_3\hat{e}_{3hi} = 0$ and $\sum_s w_{hi}a_{0hi}\hat{c}_4\hat{e}_{4hi} = 0$.

Step 2: Obtain an expression for the estimator of $E_r V_s(\hat{Y}^I)$

Using a Taylor linearization technique, we estimate $E_r V_s(\hat{Y}^I)$ by

$$v_1 = \sum_s \left(1 - \frac{n_h}{N_h}\right) \frac{N_h^2}{n_h} \frac{(\hat{\xi}_{hi} - \hat{\xi}_h)^2}{(n_h - 1)}, \quad (4.5)$$

with $\hat{\xi}_{hi}$ as above, and $\hat{\xi}_h = \sum_{s_h} \hat{\xi}_{hi} / n_h$.

Step 3: Obtain an expression for the estimator of $V_r E_s(\hat{Y}^I)$ within the design-based or model-assisted framework.

As developed in Appendix B, the design-based estimator of $V_r E_s(\hat{Y}^I)$ requires the assumption of MCAR, and is calculated as

$$\begin{aligned}
 v_{2D} = & \frac{1}{\hat{p}_{(1)}} \sum_s w_{hi}a_{0hi}a_{1hi}a_{2hi}\hat{e}_{1hi}^2(\hat{c}_{1hi}^2\hat{p}_{(1)}(1 - \hat{p}_{(1)}) + \hat{p}_{(1)}(1 - \hat{p}_{(1)}) + 2\hat{c}_{1hi}\hat{p}_{(1)}\hat{p}_{(1)}) \\
 & + \frac{\hat{p}_2}{\hat{p}_{(2)}} \sum_s w_{hi}a_{0hi}a_{1hi}\hat{e}_{2hi}^2(\hat{c}_2 + 1) + \frac{\hat{p}_3}{\hat{p}_{(3)}} \sum_s w_{hi}a_{0hi}a_{2hi}\hat{e}_{3hi}^2(\hat{c}_3 + 1) \\
 & + \frac{\hat{p}_4}{\hat{p}_{(4)}} \sum_s w_{hi}a_{0hi}\hat{e}_{4hi}^2(\hat{c}_4 + 1) + 2 \sum_s w_{hi}\hat{c}_{1hi}\hat{c}_2\hat{e}_{1hi}\hat{e}_{2hi} \\
 & + 2 \sum_s w_{hi}\hat{c}_{1hi}\hat{c}_3\hat{e}_{1hi}\hat{e}_{3hi} + 2 \sum_s w_{hi}\hat{c}_{1hi}\hat{c}_4\hat{e}_{1hi}\hat{e}_{4hi} \\
 & + 2 \sum_s w_{hi}\hat{c}_2\hat{c}_3\hat{e}_{2hi}\hat{e}_{3hi} + 2 \sum_s w_{hi}\hat{c}_2\hat{c}_4\hat{e}_{2hi}\hat{e}_{4hi} + 2 \sum_s w_{hi}\hat{c}_3\hat{c}_4\hat{e}_{3hi}\hat{e}_{4hi}, \quad (4.6)
 \end{aligned}$$

where $\hat{p}_1 = \sum_s w_{hi}(1 - a_{0hi})a_{1hi}a_{2hi}/N$, $\hat{p}_2 = \sum_s w_{hi}(1 - a_{0hi})a_{1hi}(1 - a_{2hi})/N$,
 $\hat{p}_3 = \sum_s w_{hi}(1 - a_{0hi})(1 - a_{1hi})a_{2hi}/N$, $\hat{p}_4 = \sum_s w_{hi}(1 - a_{0hi})(1 - a_{1hi})(1 - a_{2hi})/N$,

$\hat{p}_{(1)} = \sum_s w_{hi} a_{0hi} a_{1hi} a_{2hi} / N$, $\hat{p}_{(2)} = \sum_s w_{hi} a_{0hi} a_{1hi} / N$, $\hat{p}_{(3)} = \sum_s w_{hi} a_{0hi} a_{2hi} / N$, and $\hat{p}_{(4)} = \sum_s w_{hi} a_{0hi} / N$. We can then use $v_D = v_1 + v_{2D}$ to estimate the total variance.

In order to develop the model-assisted variance estimator, we first give the explicit form of the imputation model. Let $\mathbf{u}_{hi} = [u_{1hi}, u_{2hi}, u_{3hi}, u_{4hi}] = [x_{hi}, x_{1hi}, x_{2hi}, x_{3hi}]$.

$$E(y_{hi} | u_{1hi}) = \mathbf{u}_{1hi}' \mathbf{B},$$

$$E(y_{hi} | u_{2hi}) = \gamma_2 u_{2hi},$$

$$E(y_{hi} | u_{3hi}) = \gamma_3 u_{3hi},$$

$$E(y_{hi} | u_{4hi}) = \gamma_4 u_{4hi},$$

$$V(y_{hi} | \mathbf{u}_i) = \sigma^2 V(\mathbf{u}_i), \text{ and}$$

$$COV(y_{hi}, y_{jk}) = 0 \text{ for } (hi) \neq (jk).$$

The model-assisted estimator of $V_r E_s(\hat{Y}^I)$ requires the assumption of an INR or MCAR mechanism, and the further assumption that the imputation model holds. Its derivation is detailed in Appendix B, and it is given by

$$v_{2M} = \hat{\sigma}^2 \sum_s w_{hi} [(1 - a_{0hi}) + \hat{c}_{1hi}^2 a_{0hi} a_{1hi} a_{2hi} + \hat{c}_2^2 a_{0hi} a_{1hi} + \hat{c}_3^2 a_{0hi} a_{2hi} + \hat{c}_4^2 a_{0hi}] \quad (4.7)$$

with $\hat{c}_{1hi}$, $\hat{c}_2$, $\hat{c}_3$, and $\hat{c}_4$ as above. We can then use $v_M = v_1 + v_{2M}$ to estimate the total variance.

Although the model-assisted estimator takes a much simpler form, the derivation

of $\hat{\sigma}^2$ is not obvious. When we consider an individual imputation method, $\hat{\sigma}^2$ is estimated with respect to the imputation model. In the present case, given that a number of individual imputation models are used, each one could be used to estimate σ^2 . Specifically, under Model 1: $\hat{\sigma}_1^2 = s_{d1}^2 = \frac{\sum_s w_{hi} a_{0hi} a_{1hi} a_{2hi} (y_{hi} - \mathbf{x}_{hi}' \hat{\mathbf{B}}_r)^2}{\sum_s w_{hi} a_{0hi} a_{1hi} a_{2hi}}$, under Model 2: $\hat{\sigma}_2^2 = s_{d2}^2 = \frac{\sum_s w_{hi} a_{0hi} a_{1hi} (y_{hi} - \hat{R}_{r1} x_{1hi})^2}{\sum_s w_{hi} a_{0hi} a_{1hi}}$, under Model 3: $\hat{\sigma}_3^2 = s_{d3}^2 = \frac{\sum_s w_{hi} a_{0hi} a_{2hi} (y_{hi} - \hat{R}_{r2} x_{2hi})^2}{\sum_s w_{hi} a_{0hi} a_{2hi}}$, and under Model 4: $\hat{\sigma}_4^2 = s_{d4}^2 = \frac{\sum_s w_{hi} a_{0hi} (y_{hi} - \hat{R}_{r3} x_{3hi})^2}{\sum_s w_{hi} a_{0hi}}$. The specification of the constant model variance in the imputation model leads to the use of the pooled model error estimator. It is calculated as

$$\begin{aligned} \hat{S}_p^2 &= \frac{\sum_s w_{hi} a_{0hi} (a_{1hi} a_{2hi} \hat{\epsilon}_{1hi}^2 + a_{1hi} \hat{\epsilon}_{2hi}^2 + a_{2hi} \hat{\epsilon}_{3hi}^2 + \hat{\epsilon}_{4hi}^2)}{\sum_s w_{hi} a_{0hi} (a_{1hi} a_{2hi} + a_{1hi} + a_{2hi} + 1) - 5} \\ &\approx \frac{s_{d1}^2 \hat{p}_{[1]} + s_{d2}^2 \hat{p}_{[2]} + s_{d3}^2 \hat{p}_{[3]} + s_{d4}^2 \hat{p}_{[4]}}{\hat{p}_{[1]} + \hat{p}_{[2]} + \hat{p}_{[3]} + \hat{p}_{[4]}}. \end{aligned}$$

Comparison of the Model-Assisted and Design-Based Variance Estimators

The assumptions used in developing the model-assisted and design-based variance estimators are different. For the design-based variance estimator, we only need to assume that the response mechanism is MCAR (the imputation models are not required to be true). For the model-assisted estimator, we assume that the response mechanism is MCAR or INR, and that the imputation model is true. We have seen that under the assumption of MCAR, the model-assisted and design-based estimators are similar for regression imputation. We expect that they will also be similar when a combination of linear regression based imputation methods is used.

If the response mechanism is not MCAR, the design-based estimator is not valid, so we recommend the use of the design-based variance estimator only when the data

supports a MCAR response mechanism. If the imputation model does not hold, the model-assisted estimator is not valid, so we recommend the use of the model-assisted variance estimator only when the data support the validity of the imputation model; both in terms of expectation and model variance. In this imputation model, the variance conditional on the auxiliary information is equal under each imputation model. This assumption is somewhat dubious in the situation where different imputation methods yield imputed values of differing precisions. For example, by design the regression model using x_1 and x_2 will yield s_{d1}^2 less than or equal to that of the ratio model using only x_1 , s_{d2}^2 , so imputation of non-response by the regression method should yield less variable results.

Upon inspection of the model-assisted estimator of $V_r E_s(\hat{Y}^I)$ it is clear that the importance of s_{dm}^2 for each method m in the pooled variance estimator depends on $p_{(m)}$; the fraction of units in the population for which data are available to estimate the parameters of imputation model m . In the design-based estimator of $V_r E_s(\hat{Y}^I)$ the importance of s_{dm}^2 for each method m in the variance estimator depends on both $p_{(m)}$ and p_m ; the fraction of units in the population for which data are available to estimate the parameters of imputation model m , and the fraction of units imputed by imputation method m .

We now apply our results to data from the Monthly Wholesale and Retail Trade Survey (MWRTS) of Statistics Canada.

4.2.5 Population Characteristics

In this section, we present some characteristics of the survey population for the MWRTS. We present estimates of the response rates, preliminary investigations into the response mechanism, and evaluation of the model fit of each imputation model. These characteristics are intended to illustrate the presence or absence of conditions which would lead to bias in the composite-imputed point estimator, or the associated variance estimators.

Estimated Response Rate

Estimates of the response rate were calculated for each cell $g = 1, \dots, G$ and each month $t = 1, \dots, T$ as $RR_{gt} = \sum_{s_{gt}} w_{hi} a_{0hi} / \sum_{s_{gt}} w_{hi}$ where s_{gt} is the set of units included in the sample in cell g for month t , and w_{hi} is the sampling weight for unit (hi). The overall average of the response rates, $RR_{..} = \sum_g \sum_t RR_{gt} / (MT)$, across all cells and all months is approximately 0.85. This is a high response rate, but we note that it varies between cells: the average over the 12 months for individual cells, $RR_g = \sum_t RR_{gt} / T$ ranges from 0.74 to 0.94. Table 4.1 gives the frequency distribution of the estimated response rates (RR_{gt}). Included in the table are 1500 estimated response probabilities; one observation for each of the 12 months in each of the 125 cells.

Table 4.1 confirms that the response rate is typically quite high for this survey in most cells; we note that the response rate is below 0.75 in only 4% of the observations, and is never below 0.65. This limits (but does not eliminate) the potential for non-response bias in the point estimates.

Table 4.1: MWRTS data - Distribution of response rates by cell

Response Rate	Frequency (n)	Frequency (%)
0.950 - 1.000	18	1.2
0.900 - 0.949	232	15.5
0.850 - 0.899	588	39.2
0.800 - 0.849	426	28.4
0.750 - 0.799	172	11.5
0.700 - 0.749	59	3.9
0.650 - 0.699	5	0.3
< 0.650	0	0.0
TOTAL	1500	100.0

Evaluation of Model Fit

We consider each model individually to assess its fit, both in terms of the expectation and variance of the imputation model. Visual inspection of the data supports the conclusions that the imputation models describe the data well in most cases. Table 4.2 gives the frequencies of the estimated Pearson correlation between the reported values, y , and the values predicted by each imputation model m , $\hat{y}_m^I$. Given

Table 4.2: MWRTS data - $\hat{\rho}(y, \hat{y}_m^I)$ under each imputation model m

Correlation	Model 1	Model 2	Model 3	Model 4
$\rho < 0$	0.5	1.3	2.3	0.4
$0 < \rho \leq 0.4$	0.6	2.0	3.0	1.1
$0.4 < \rho \leq 0.6$	2.7	2.4	4.1	7.1
$0.6 < \rho \leq 0.8$	2.5	3.4	7.2	21.3
$0.8 < \rho \leq 1.0$	93.8	90.9	83.4	70.1
Average	0.950	0.923	0.885	0.841

the high correlations for most cells in Table 4.2, the imputation models generally yield predicted values highly correlated with the observed values. In situations where the data do not appear to be well described by the models, the data do not suggest that quadratic terms or intercepts should be added to the model to improve the fit. As well, the imputation residuals from the regression model in many cases could reasonably have been generated by a constant variance model, however there are suggestions of increasing variance for the residuals from the ratio and regression models. We note from Table 4.2 that the different correlations between the variable of interest and the predicted values under each model suggest that the assumption of constant model variance may not be valid.

Verification of Response Mechanism

Since both the industry and the geographic region are built into the definition of a cell, the units in each cell are homogeneous with respect to industry and geography, but are not necessarily homogeneous with respect to annual revenues. Thus, it is feasible that the probability of response for a units within a cell may be dependent on the units size; that is INR or NINR. Given the high correlations between the variable of interest and the auxiliary variables, the effective difference between INR and NINR should be minimal. To investigate if the annual revenues of a unit is correlated with its probability to respond, a logistic regression analysis was carried out to assess the model fit of the following model within each cell:

$$\log\left(\frac{p_{hi}}{1 - p_{hi}}\right) = \beta_0 + \beta_1 x_{3hi}.$$

If the fit of this model is good, this would suggest that the response mechanism is not MCAR. A Wald Chi-square test was used to determine if the term β_1 is significantly different from zero in the model for each month of data, and for each cell. This gives a total of 1,500 hypothesis tests. Of these, in only 260 cases the co-efficient β_1 was significantly different from zero at 5% level of significance. We have ignored the difficulty with multiple tests, but the above result indicates that, in general, the evidence is not strong that the response mechanism is INR. However, what we are interested in is the response mechanism within each cell (since the variance is estimated independently in each).

Table 4.3 shows the distribution of the number of months for which the term β_1 was significantly different from zero for the 125 different cells.

Further investigation of the data revealed that 10 of the 12 cells for which all twelve months showed a significant β_1 were from the same industry. In this industry, a large number of units with $x_{3hi} = 0$ are classified as respondents. These are units that are in the survey population, and are selected in the sample, but are not actually contacted because they are out of business (therefore they are treated as respondents since their data contains no variability from non-response). Nonetheless, there is evidence that the response mechanism is not MCAR in some cells; that is, the mechanism may be INR or NINR in cells with more months where β_1 is significantly different from zero.

In general, the fit of the regression model is very strong for most cells and most

Table 4.3: MWRTS data - Number of months with significant β_1 in logistic regression

Number of Months with significant β_1	Number of Cells	Percent of Cells (%)
0	78	62.4
1	14	11.2
2	5	4.0
3	1	0.8
4	2	1.6
5	2	1.6
6	2	1.6
7	1	0.8
8	0	0.0
9	1	0.8
10	1	0.8
11	3	2.4
12	12	9.6
Total	125	100.0

months. The ratio model using x_1 and the ratio model using x_2 also yield correlations above 0.8 in over 80% of the cells. Given these correlations, it is clear that for respondents, the vast majority of the variability in the variable of interest is explained by the auxiliary data. As we observed in our simulations, as the correlation between the auxiliary data and the variable of interest increases, the bias introduced by NINR mechanisms decreases. As well, the correlations themselves support the sequence of preference for the imputation models (i.e. the model yielding imputed values of the highest quality is chosen first, and so on).

Summary

In general, the MWRTS data supports the following findings:

- The response rates are typically close to 85% , however in some cells and some months, can be as low as 70% .
- The fit of the imputation models is generally very good; The correlation between the variable of interest and the predicted value under the imputation models is usually above 0.8. However, each model yields lower correlations with the variable of interest in individual cells.
- The response mechanism in most cells appears to be MCAR. In some cases, the data suggest that the response mechanism is INR, but due to the unavailability of data we are unable to confirm that it is not, in fact, NINR. If the response mechanism is NINR, we rely on the high correlation between observed values and predicted values under each imputation model to reduce the impact of non-response on the point estimates.

Based on the theoretical and empirical results of Chapter 3, with a response rate above 0.7, a correlation between observed and predicted values above 0.8 for each imputation model, and a response mechanism of MCAR, our point estimates will be approximately unbiased under ratio or regression imputation methods, so they will also be approximately unbiased using a combination of these methods. For those cells with a suggestion of response mechanisms other than MCAR, the good model fit and relatively high response rate should lead to a limited bias, even with an INR response mechanism.

Variance Estimation

In this section, overall results from the MWRTS are considered, and detailed results are presented for several individual cells selected to illustrate characteristics that impact variance estimation. The variance estimates are presented, and explanations of the results are offered. Recall that the data, in general, suggest that the response mechanism is approximately MCAR, and that the variances of imputed values under each model are not equivalent. We will therefore use the design-based estimator, v_D as the baseline for comparisons since we are more confident in the assumptions associated with the design-based variance estimator than we are in the assumptions associated with the model-assisted variance estimator. We will compare the results from the naive variance estimator, v_0 , and the model-assisted variance estimator, v_m , to the design-based estimator, v_D . We denote the value of the variance estimators v_0 , v_1 , and v_D for a given month t and cell g as v_0^{tg} , v_1^{tg} and v_D^{tg} respectively. For each cell, we calculate the average ratio of v_0 to v_D as $\frac{1}{12} \sum_t \frac{v_0^{tg}}{v_D^{tg}}$, and the average ratio of v_1 to v_D as $\frac{1}{12} \sum_t \frac{v_1^{tg}}{v_D^{tg}}$.

Table 4.4 provides a summary of the underestimation resulting from the use of v_0 or v_1 to estimate the total variance of the estimator for each industry (note that each industry is formed by aggregating together independent cells). The entries in this table are the average values of the ratio of v_0 to v_D and v_1 to v_D for each cell.

From Table 4.4, it is clear that the variance underestimation resulting from v_0 or

Table 4.4: MWRTS Estimates - Average Ratios of v_0 to v_D and v_1 to v_D

Industry	Average of v_0/v_D	Average of v_1/v_D
<i>A</i>	0.656	0.967
<i>B</i>	1.003	0.999
<i>C</i>	0.821	0.993
<i>D</i>	0.924	0.975
<i>E</i>	0.816	0.973
<i>F</i>	0.854	0.988
<i>G</i>	0.919	0.976
<i>H</i>	0.965	0.999
<i>I</i>	0.727	0.940
<i>J</i>	0.397	0.851
<i>K</i>	0.926	0.992
<i>L</i>	0.851	0.986
<i>M</i>	0.775	0.983
<i>N</i>	0.836	0.880
<i>O</i>	0.686	0.998
Total	0.810	0.967

v_1 is not constant across industries. Overall, the variance underestimation resulting from the use of v_0 ranges from 0 to 60% and the variance underestimation resulting from the use of v_1 ranges from 0 to 15%. At the national level, the underestimation of variance associated with v_0 is 19%, and the underestimation of variance associated with v_1 is 3%.

As noted in Chapter 3, the underestimation of variance depends on a number of factors including the response rate, the sampling fraction and the error variance under the imputation model. We now present some results from particular cells selected to illustrate these factors, and their impact on estimates of variance. In these examples,

the sampling fraction for each cell g for month t is defined as $f^{gt} = \sum_{h_g} n_h / \sum_{h_g} N_h$ and the response rate is calculated as $p_0^{gt} = \sum_{s_{gt}} w_{hi} a_{0hi} / \sum_{s_{gt}} w_{hi}$ where h_g denotes the set of strata included in cell g and s_{gt} denotes the set of sampled units in the cell g for month t .

Example 1: Table 4.5 gives the estimates of variance v_0 , v_1 , and v_D for the estimate of total sales for Industry M in Geographic Region 6 for each month of reference year 2000 and the ratio of v_0 to v_D and v_1 to v_D for each month.

Table 4.5: MWRTS Variance Estimates: Example 1

Month	v_0	$\frac{v_0}{v_D}$	v_1	$\frac{v_1}{v_D}$	v_D
JAN	2.050	.770	2.658	.998	2.663
FEB	2.160	.850	2.534	.997	2.541
MAR	1.698	.807	2.093	.994	2.105
APR	2.229	.855	2.570	.985	2.608
MAY	4.147	.853	4.827	.993	4.861
JUN	2.142	.856	2.479	.990	2.503
JUL	2.394	.897	2.575	.964	2.670
AUG	1.350	.785	1.591	.925	1.719
SEP	3.910	.784	4.851	.973	4.986
OCT	3.358	.775	4.213	.972	4.333
NOV	2.873	.772	3.528	.948	3.720
DEC	8.431	.790	10.107	.947	10.678
Average	-	0.816	-	0.974	-

* Variance estimates are in 10E17

Table 4.5 shows that on average, the underestimation of variance through the use of v_0 and v_1 are approximately 18% and 3% respectively. This cell was chosen for

the example since the variance underestimation, and population characteristics are typical of the averages over all industries and all months. The underestimation for this particular cell varies from month to month; the underestimation using v_0 varies from 10% to 23% and the underestimation using v_1 varies from 0.2% to 7.5%. Further examination of the data reveals that the overall sampling fraction f^{gt} for this cell is approximately 7% for each month. The response rate, p_0^{gt} , varies between months from 0.70 to 0.81, and the Pearson correlations between y and each auxiliary variable x_1 , x_2 and x_3 is above 0.9 for each month.

Example 2: Table 4.6 gives the variance estimates v_0 , v_1 , and v_D for the estimate of total sales for Industry J in Geographic Region 8 for each month of reference year 2000 and the ratio of v_0 to v_D and v_1 to v_D for each month.

Table 4.6 shows that on average, the variance underestimation through the use of v_0 is approximately 60% and through v_1 is approximately 0.6%. The underestimation for v_0 is much larger than the average across all industries for v_0 . The underestimation for v_0 is of a relatively constant magnitude; it varies from 48% to 69%. The underestimation for v_1 is never larger than 1.2%. Most population characteristics for this cell are similar to the averages in the population, however the correlation between y and x_3 is lower than average (approximately 0.7 for some months). The relatively large imputation residuals from Method 3 lead to more within stratum variability among the $\hat{\xi}_{hi}$'s than among the y_{hi} 's, resulting in a value of v_1 much larger than that of v_0 .

Table 4.6: MWRTS Variance Estimates: Example 2

Month	v_0	$\frac{v_0}{v_D}$	v_1	$\frac{v_1}{v_D}$	v_D
JAN	4.676	.493	9.483	.999	9.491
FEB	4.406	.324	13.528	.995	13.600
MAR	5.685	.312	18.100	.995	18.192
APR	4.823	.475	10.039	.988	10.160
MAY	7.056	.364	19.273	.994	19.392
JUN	5.621	.334	16.756	.994	16.851
JUL	6.726	.344	19.398	.992	19.555
AUG	6.513	.377	17.204	.995	17.299
SEP	4.423	.512	8.583	.994	8.638
OCT	6.836	.379	17.917	.992	18.056
NOV	6.604	.524	12.524	.993	12.609
DEC	4.795	.368	12.963	.996	13.020
Average	-	0.401	-	0.994	-

* Variance estimates are in 10E15

Example 3: Table 4.7 gives variance estimates v_0 , v_1 , and v_D for the estimate of total sales for Industry K in Geographic Region 6 for each month of reference year 2000 and the ratio of v_0 to v_D and v_1 to v_D for each month.

Table 4.7 shows that on average, the underestimation of variance through the use of v_0 is approximately 4.3% and through the use of v_1 is approximately 0.2% for this cell. This is much lower than the average across all industries for v_0 and v_1 . The underestimation for v_0 is of a relatively constant magnitude; the underestimation varies from 2% to 10.8% and the underestimation of v_1 is never larger than 1.7%. For this cell, population characteristics are close to the average over all cells, but the correlation between y and x_3 is somewhat lower (above 0.75 for all months except January with correlation of 0.12). Due to the relatively infrequent use of imputation

Table 4.7: MWRTS Variance Estimates: Example 3

Month	v_0	$\frac{v_0}{v_D}$	v_1	$\frac{v_1}{v_D}$	v_D
JAN	5.557	.892	6.122	.983	6.230
FEB	5.043	.980	5.143	.999	5.1487
MAR	11.984	.981	12.212	1.00	12.218
APR	19.657	.951	20.666	1.00	20.675
MAY	23.908	.955	25.030	.999	25.043
JUN	19.696	.966	20.377	1.00	20.387
JUL	15.811	.946	16.703	.999	16.712
AUG	14.334	.962	14.889	.999	14.899
SEP	12.530	.940	13.320	1.00	13.323
OCT	11.623	.959	12.115	.999	12.121
NOV	12.321	.954	12.903	.999	12.911
DEC	5.6955	.995	5.7189	.999	5.725
Average	-	0.957	-	0.998	-

* Variance estimates are in 10E18

method 4 in all months except for January, and the low variability associated with values imputed under Methods 1, 2 and 3, the underestimation of variance from v_0 is minor for this cell. This example is intended to demonstrate the possibility for the variance underestimation to vary over time due to the changes in population characteristics.

Example 4: Table 4.8 gives the variance estimates v_0 , v_1 , and v_D for the estimate of total sales for Industry N in Geographic Region 12 for each month of reference year 2000 and the ratio of v_0 to v_D and v_1 to v_D for each month. It shows that on average, the underestimation of variance through the use of v_0 is approximately 16.5% and through v_1 is approximately 12% for this cell. This value is similar to

Table 4.8: MWRTS Variance Estimates: Example 4

Month	v_0	$\frac{v_0}{v_D}$	v_1	$\frac{v_1}{v_D}$	v_D
JAN	1.823	.868	1.987	.947	2.100
FEB	1.234	.905	1.240	.909	1.363
MAR	1.581	.882	1.628	.908	1.792
APR	1.844	.804	1.962	.855	2.293
MAY	2.106	.781	2.231	.828	2.696
JUN	2.715	.788	2.926	.849	3.445
JUL	2.758	.828	2.939	.882	3.330
AUG	2.965	.810	3.160	.863	3.662
SEP	1.714	.813	1.791	.849	2.109
OCT	2.248	.797	2.398	.850	2.820
NOV	2.882	.836	3.141	.911	3.449
DEC	3.525	.919	3.501	.913	3.834
Average	-	0.835	-	0.880	-

* Variance estimates are in 10E15

the average across all industries for v_0 , and much higher than the average across all industries for v_1 (implying that the second component of the variance estimate is important). The underestimation varies from month to month, but is of a relatively constant magnitude; the underestimation of v_0 varies from 8% to 21% and the underestimation of v_1 varies from 5.4% to 17.3%. Most population characteristics for this cell are close to the average across all cells, but the overall sampling fraction, f^{tg} , for this cell is approximately 55%. The difference between v_0 and v_1 is relatively small, but the difference between v_1 and v_D is large; that is, the contribution of v_{2D} is large, which is to be expected when the sampling fractions are non-negligible.

Example 5: Table 4.9 gives the variance estimates v_0 , v_1 , and v_{2D} for the estimate of

Table 4.9: MWRTS Variance Estimation: Example 5

Month	v_0	$\frac{v_0}{v_D}$	v_1	$\frac{v_1}{v_D}$	v_D
JAN	2.262	.879	2.572	.999	2.574
FEB	1.994	1.16	1.713	.998	1.716
MAR	2.121	1.13	1.883	.999	1.884
APR	2.800	1.11	2.517	.999	2.519
MAY	3.433	1.12	3.072	.999	3.077
JUN	3.318	1.09	3.034	.999	3.036
JUL	3.443	1.01	3.407	1.00	3.409
AUG	4.066	1.15	3.541	.999	3.545
SEP	3.042	1.13	2.699	1.00	2.700
OCT	3.417	1.17	2.913	.999	2.915
NOV	3.202	1.23	2.598	.999	2.600
DEC	4.552	1.03	4.409	.996	4.426
Average	-	1.101	-	0.999	-

* Variance estimates are in 10E15

total sales for Industry B in Geographic Region 6 for each month of reference year 2000 and the ratio of v_0 to v_{2D} and v_1 to v_{2D} for each month.

Table 4.9 shows that on average, the overestimation of variance through the use of v_0 is approximately 10% and the underestimation through the use of v_1 is approximately 0.1% for this cell. These values are quite different from the averages across all industries for v_0 and v_1 . The overestimation of v_0 and the underestimation of v_1 are relatively constant from month to month; v_0 yields underestimates of 12% to an overestimation of 123% while the underestimation of v_1 is virtually zero for each month. Most population characteristics for this cell are close to the average over all cells, but the correlation between y and the auxiliary variable x_3 is slightly lower; as low as 0.8 for some months. The apparent overestimation for v_0 results mainly

from one unit, for which the value of $\hat{\xi}_{hi}$ is much closer to its stratum mean than is y_{hi} , which yields a value of v_1 smaller than that of v_0 .

Comparison of Design-based and Model-assisted variance estimators

We have noted that under regression imputation, the design-based and model-assisted variance estimators are very similar. We also noted that these assumptions are easily violated (especially the assumption of constant variance). In general, with small sampling fractions, the second term in the population variance, $V_r E_s(\hat{Y}^I)$, is small compared to the first term, $E_r V_s(\hat{Y}^I)$, so we calculated ratios to isolate the difference in this term. Table 4.10 gives the average value (over the twelve months) of the ratio of the model-assisted and design-based estimators of $V_r E_s(\hat{Y}^I)$.

Table 4.10 shows that the ratios of v_{2M} to v_{2D} are quite variable, ranging from 0.477 to 1.710. These differences can be attributed to violations in the assumptions inherent in one or more of the variance estimators (that is, a response mechanism other than MCAR for the design-based estimator, or model mis-specification for the model-assisted variance estimator).

4.3 Summary of Findings

Bias of the Estimator

Our analysis of population characteristics suggests that the response rate is high and that the imputation models typically yield imputed values that are highly cor-

Table 4.10: MWRTS Variance Estimates: Average Ratios of v_{2M} to v_{2D}

Industry	v_{2M}/v_{2D}
<i>A</i>	0.927
<i>B</i>	0.640
<i>C</i>	1.710
<i>D</i>	1.280
<i>E</i>	1.228
<i>F</i>	1.242
<i>G</i>	0.643
<i>H</i>	0.815
<i>I</i>	0.822
<i>J</i>	0.477
<i>K</i>	0.828
<i>L</i>	0.591
<i>M</i>	0.776
<i>N</i>	0.573
<i>O</i>	1.169
Total	0.614

related with the variable of interest. The analysis of the response mechanism was not conclusive; it suggested that in some cells, the response mechanism was ignorable. In any case, we expect the bias to be negligible given the high response rate and quality of imputed values, with a non-response mechanism within cells of Missing Completely at Random or Ignorable Non-Response.

Variance Estimation

We found that overall, the use of the naive variance estimator that treats imputed values as if true values, v_0 , led to an underestimation of variance in the order of 20%. The underestimation varied between cells, but in general was high when (i)

the sampling fraction is high, (ii) the response rate is low, and (iii) the correlation between auxiliary variables and the variable of interest is low. This was confirmed through the examples studied, and agrees with the theoretical results and results of our simulation studies for individual methods. In general, the performance of the simplified variance estimator, v_1 , obtained from the reverse approach is better than that of v_0 . In most cells, the variance estimator v_1 is close to the correct variance estimator $v_D = v_1 + v_{2D}$. Where a difference is observed, the sampling fraction is high. We therefore recommend the use of the design-based variance estimator, v_D , since it is designed to handle non-negligible sampling fractions.

Design-based vs. Model-assisted Variance Estimation:

The comparisons of the variance estimators obtained by the model-assisted and design-based approaches were not consistent across industries. In some cases, the model-assisted variance estimate is higher, and in others it is lower. Due to our doubts in the validity of the imputation model (particularly the specified constant variance), we recommend the use of the variance estimator, $v_D = v_1 + v_{2D}$ derived under the design-based approach.

Chapter 5

CONCLUSIONS

In this chapter, we summarize the findings from earlier chapters and identify areas for future research.

5.1 Summary of Findings

Regression Imputation: Point Estimation

We should be careful in generalizing the results of the simulations to the general classes of response mechanisms and data models. The magnitudes of the bias of the point estimator and the variance estimator depend heavily on the models used to generate the data and response probabilities. It would be possible to demonstrate a wide range of biases simply by changing values of intercepts or slope parameters in generating the data, or by manipulating the form of the function used to generate the response probabilities under ignorable non-response or non-ignorable non-response. As well, the relative contribution of the component v_2 to the total variance estimate, v , is heavily dependent on the sample design. If the sampling fractions are negligible, the component v_2 is expected to represent a very small portion of the total

variance estimate. In the simulations we have used a sample design with population and sample sizes typical of a business survey where the sample typically include a non-negligible portion of the population, but the results could be quite different for other sample designs (such as a census or sample survey with very low sampling fractions). With these restrictions, the following conclusions are drawn:

- If the response mechanism is missing completely at random, the point estimates will be approximately unbiased, whether the data follow the model implied by imputation or not. The variance of the estimator is reduced with an increased response rate, or with the use of more highly correlated auxiliary data in the imputation.
- If the response mechanism is ignorable, we should make extensive use of these variables in imputation to produce an approximately unbiased estimator. If the imputation model does not correspond to the population data (i.e. the form of the imputation model does not correspond to the data model) a bias can be introduced.
- If the response mechanism is non-ignorable, a bias is introduced by regression imputation methods. The magnitude of the absolute bias increases with lower response rates, and lower correlations between the variable of interest and the predicted values under the imputation model. In this situation, bias correction procedures would be recommended, although these often rely on further modeling. If in addition, the imputation model does not correspond to the population data, the estimators can be expected to perform very poorly.

Regression Imputation: Variance Estimation

- We note that the magnitude of the underestimation of variance from the use of

v_0 depends on the following factors: (i) the response rate - with a lower response rate, the variance underestimation associated with v_0 increases, (ii) the fit of the imputation model - imputation models yielding higher correlations with the variable of interest generally yield a less important variance underestimation from the naive variance estimator, v_0 , and (iii) the sampling fractions - with negligible sampling fractions, the variability arising from sampling dominates the overall variance, so the underestimation resulting from v_0 is less important with smaller sampling fractions.

- The underestimation associated with the use of v_1 depends mainly on the sampling fraction. The variance estimator v_1 results in improvements over v_0 (the bias of the variance estimator v_1 is generally smaller in absolute value than that of v_0), and when the sampling fractions are low, v_1 serves as a good estimator of the total variance.
- Including the estimator of $V_r E_s[\hat{Y}^I]$ (as in v_{2D} or v_{2M}) leads to performance gains in variance estimation. In situations where the response rate is high, or the sampling fraction is low, the contribution of this term is relatively minor, but the estimators v_D and v_M outperform v_0 and v_1 under Missing Completely at Random and Ignorable Non-Response mechanisms.
- None of the variance estimators perform well in terms of relative bias or confidence interval coverage rates when the response mechanism is non-ignorable.
- The variance estimators derived under the model-assisted approach, v_M , and the design-based approach, v_D , are similar because v_1 dominates the estimators. In situations where the assumptions are not met (i.e. MCAR for design-based and

the variance with respect to the model for model-assisted), the results are slightly different for the two approaches, but with regards to the relative bias and coverage rates of confidence intervals, the performance of the two approaches are comparable.

- In order to produce unbiased estimators and variance estimators the model underlying the imputation must be valid. It is thus very important to validate the model extensively. If it becomes clear after imputation that the model is violated, post-imputation methods may be used to compensate for the bias introduced. However, it would be preferable to modify the imputation procedure and re-impute values under the new imputation model.
- The impact of imputation on variance estimation is reduced when the quality of imputation is high, and the imputation rate is low. In order to achieve this, it may be necessary to dedicate significant resources to development of the imputation procedure, obtaining quality auxiliary variables, and some means of reducing non-response (such as follow-up of non-respondents, etc.).

Composite Imputation: Point Estimation

The study of population characteristics of the Monthly Wholesale and Retail Trade Survey data indicate that:

- the response probability is high,
- the response mechanism is quite possibly missing completely at random, and
- the imputation models yielded predicted values that are highly correlated with the variable of interest.

These conditions should yield approximately unbiased estimators of population to-

tals.

Composite Imputation: Variance Estimation

The application to MWRTS data revealed the following:

- Overall, the variance underestimation through the use of the naive variance estimator, v_0 , is approximately 20%. This underestimation varies considerably between domains (industries).
- By examining individual imputation cells, we confirmed that the variance underestimation associated with the naive variance estimator, v_0 , under composite imputation is most significant when (i) the response rate is low (ii) the fit of the imputation model(s) are poor, and (iii) the sampling fractions are high.
- In general, the performance of v_1 was quite close to that of v_D , due to the relatively small sampling fractions however in some domains v_1 was much smaller than v_D .
- The performance of v_M was quite different from v_D , and we suggest that this is due to violations of assumptions used in the development of v_M (particularly the imputation model).
- We recommend the use of v_D to estimate the variance as the assumptions required in its development are more plausible (the data do not strongly suggest that the response mechanism is not MCAR).

5.2 Future Work

A number of aspects that are typical of imputation in large-scale business surveys are not considered in this thesis. These would complicate the estimation of variance, but could likely be incorporated into the procedure of Shao and Steel (1999).

5.2.1 Model-assisted variance estimation under composite imputation

We find some of the assumptions in the model-assisted approach to be strong, particularly the assumption of constant variance. In practice, we often find that the magnitude of the imputation residuals is correlated with the auxiliary variable. In fact, this is the situation that motivates ratio imputation (model variance proportional to the auxiliary variable). An extension of the method is required in order to incorporate more complex model variance structures.

5.2.2 Collapsing of Imputation Cells

In some instances, when imputation groups are relatively small, or non-response rates are relatively high, it is not possible to obtain enough respondents in order to reliably estimate the parameters of the imputation model. When this occurs in practice, often imputation groups are collapsed together. This strategy generally introduces a bias if the imputation groups being collapsed differ with respect to the model.

5.2.3 Stratification, Imputation Groups and Domain Estimation

In this paper we have considered only the case where the stratification, imputation groups and domains of estimation correspond perfectly, or at least the imputation groups and domains of estimation correspond, and can be formed as groupings of strata. In practice, surveys are usually designed to produce efficient estimates at the domain level, so as much as possible, strata are designed to include units from only one domain. This is often not achieved for a number of reasons such as frame mis-classifications, and small sample sizes. It is fairly common to use stratification information to form imputation groups (which are not necessarily groups of strata), but to use collected classifications to form domains of estimation. With non-response, it is not only the value of the variable of interest that is missing, but the information that determines the domain membership of the unit (i.e. a confirmed industry or geography) is missing. In practice, assumptions are made to deal with this lack of information (such as assuming that the industry or geography used in stratification are correct), but the impacts of these assumptions are not well known. This introduces complexity to the point and variance estimation methods, for which formulas have not yet been developed, and the extent of the impact depends on the frequency of mis-classification, and the correspondence between strata, imputation groups and domains of interest.

Bibliography

- [1] Binder, D. and Sun, W. (1996) "Frequency valid multiple imputation for surveys with a complex design", in *Proceedings of the Section on Survey Research Methods*, U.S. Bureau of the Census, 281-286.
- [2] Cochran, W. (1977), *Sampling Techniques.*, 3rd ed. New York: Wiley
- [3] Demnati, A. and Rao, J.N.K. (2002) "Linearization variance estimators for survey data with missing responses", in *Proceedings of the Annual Meeting of the Survey Methods Section*, Statistical Society of Canada, 2002.
- [4] Fay, R. E. (1991), "A design-based perspective on missing data variance", in *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 429-440.
- [5] Gagnon, F., Lee, H., Rancourt, E. and Sarndal, C.E. (1996), "Estimating the variance of the generalized regression estimator in the presence of imputation for the Generalized Estimation System", in *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 151-156.

- [6] Greenless, J.S., Reece, W.S., and Zieschang, K.D. (1982), "Imputation of missing values when the probability of response depends on the variable being imputed", *Journal of the American Statistical Association*, 77, 251-261.
- [7] Haziza, D. and Rao, J. N. K. (2003), "Inference for population means under unweighted imputation for missing survey data", *Survey Methodology*, (to appear).
- [8] Hidiroglou, M., and Lavallee, P. (1988), "On the stratification of skewed populations", *Survey Methodology*, 14, 22-43.
- [9] Horvitz, D. G., and Thompson, D. J. (1952) "A generalization of sampling without replacement from a finite universe", *Journal of the American Statistical Association*, 47, 663-685.
- [10] Kalton, G., and Kasprzyk, D. (1986), "The treatment of missing survey data", *Survey Methodology*, 12, 1-16.
- [11] Kovar, J., and Whitridge, P. (1995), "Imputation of business survey data", *Business Survey Methods*, Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., and Kott, P.S. (editors), 403-23, New York: John Wiley and Sons.
- [12] Lee, H., and Rancourt, E. (1994), "Experiments with variance estimation from survey data with imputed values", *Journal of Official Statistics*, 10, 231-243.
- [13] Lee, H., Rancourt, E., and Sarndal, C.E. (1994), "Variance estimation from survey data under single value imputation", *Survey Methodology*, 20, 137-47.

- [14] Little, R.J.A., and Rubin, D.B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley
- [15] Rancourt, E., Lee, H., and Sarndal, C.E. (1993), "Variance estimation under more than one imputation method", in *Proceedings of the International Conference on Establishment Statistics*, American Statistical Association, 374-379.
- [16] Rancourt, E., Lee, H., and Sarndal, C.E. (1994), "Bias corrections for survey estimates from data with ratio imputed values for confounded nonresponse", *Survey Methodology*, 20, 137-147.
- [17] Rao, J.N.K., and Shao, J. (1992), "Jackknife variance estimation with survey data under hot-deck imputation", *Biometrika*, 79, 811-22.
- [18] Rao, J.N.K., and Sitter, R. (1995), "Variance estimation under two-phase sampling with application to imputation for missing data", *Biometrika*, 82, 453-60.
- [19] Rao, J.N.K. (1996), "On variance estimation with imputed survey data (with discussion)", *Journal of the American Statistical Association*, 91, 499-506.
- [20] Rao, J.N.K. (2000), "Variance estimation in the presence of imputation for missing data", in *Proceedings of the Second International Conference on Establishment Surveys*, 599-608.
- [21] Rubin, Donald B. (1986), "Basic ideas of multiple imputation for nonresponse", *Survey Methodology*, 12, 37-47.
- [22] Sarndal, C.E. (1992), "Methods for estimating the precision of survey estimates when imputation has been used", *Survey Methodology*, 18, 241-252.

- [23] Sarndal, C.E., Swensson, B., and Wretman, J. (1992), *Model-Assisted Survey Sampling*, New York: Springer-Verlag

- [24] Shao, J., and Steel, P. (1999), "Variance estimation for survey data with composite imputation and non-negligible sampling fractions", *Journal of the American Statistical Association*, 94, 254-265.

- [25] Statistics Canada (2000), "Standards and Guidelines for Reporting of Nonresponse Rates", Ottawa, Canada.

Appendix A

REGRESSION IMPUTATION

Here we develop the variance estimators for the regression-imputed estimator following the reverse approach procedure elaborated in Shao and Steel (1999). We use the indicator variable a_{hi} to identify whether units report a value of the variable of interest: $a_{hi} = 1$ if y_{hi} is reported, 0 otherwise. The imputation strategy is outlined as follows:

- If y_{hi} is reported ($a_{hi} = 1$), then $y_{hi}^I = y_{hi}$
- Otherwise ($a_{hi} = 0$), $y_{hi}^I = \mathbf{x}_{hi}' \hat{\mathbf{B}}_r$, where $\mathbf{x}_{hi} = (1, x_{1hi}, x_{2hi}, x_{qhi})'$ and $\hat{\mathbf{B}}_r = \hat{\mathbf{T}}_r^{-1}(\sum_s w_{hi} a_{hi} \mathbf{x}_{hi} y_{hi})$ with $\hat{\mathbf{T}}_r = \sum_s w_{hi} a_{hi} \mathbf{x}_{hi} \mathbf{x}_{hi}'$.

Step 1: Re-express the estimator

The regression-imputed Horvitz-Thompson estimator can be expressed as

$$\hat{Y}^I = \sum_s w_{hi} (a_{hi} y_{hi} + (1 - a_{hi}) \mathbf{x}_{hi}' \hat{\mathbf{B}}_r) \quad (\text{A.1})$$

with $\hat{\mathbf{B}}_r$ as above.

Step 2: Derive an Estimator for $E_r V_s(\hat{Y}^I)$

The estimator for $E_r V_s(\hat{Y}^I)$ is obtained via a Taylor linearization technique. It is derived by re-expressing the regression-imputed estimator:

$$\begin{aligned}\hat{Y}^I &= \sum_s w_{hi}(a_{hi}y_{hi} + (1 - a_{hi})\mathbf{x}_{hi}'\hat{\mathbf{B}}_r) \\ &= \sum_s w_{hi}(a_{hi}y_{hi} + (1 - a_{hi})\mathbf{x}_{hi}'\tilde{\mathbf{B}}) \\ &\quad + \sum_s w_{hi}(1 - a_{hi})\mathbf{x}_{hi}'(\hat{\mathbf{B}}_r - \tilde{\mathbf{B}})\end{aligned}$$

where $\tilde{\mathbf{B}} = \tilde{\mathbf{T}}^{-1}(\sum_U a_{hi}\mathbf{x}_{hi}y_{hi})$, and $\tilde{\mathbf{T}} = \sum_U a_{hi}\mathbf{x}_{hi}\mathbf{x}_{hi}'$. Now,

$$\begin{aligned}\hat{\mathbf{B}}_r - \tilde{\mathbf{B}} &= \hat{\mathbf{T}}_r^{-1} \sum_s w_{hi}a_{hi}\mathbf{x}_{hi}(y_{hi} - \mathbf{x}_{hi}'\tilde{\mathbf{B}}) \\ &= \hat{\mathbf{T}}_r^{-1} \sum_s w_{hi}a_{hi}\mathbf{x}_{hi}\tilde{\epsilon}_{hi}\end{aligned}$$

with $\tilde{\epsilon}_{hi} = y_{hi} - \mathbf{x}_{hi}'\tilde{\mathbf{B}}$. Note $E_s(\sum_s w_{hi}a_{hi}\mathbf{x}_{hi}\tilde{\epsilon}_{hi}) = \sum_U a_{hi}\mathbf{x}_{hi}\tilde{\epsilon}_{hi} = 0$. Hence, by Slutsky's theorem,

$$\sum_s w_{hi}(1 - a_{hi})\mathbf{x}_{hi}'(\hat{\mathbf{B}}_r - \tilde{\mathbf{B}}) \approx (\sum_U (1 - a_{hi})\mathbf{x}_{hi}')\tilde{\mathbf{T}}^{-1}(\sum_s w_{hi}a_{hi}\mathbf{x}_{hi}\tilde{\epsilon}_{hi})$$

Therefore,

$$\hat{Y}^I \approx \sum_s w_{hi}\tilde{\xi}_{hi}$$

where $\tilde{\xi}_{hi} = a_{hi}y_{hi} + (1 - a_{hi})\mathbf{x}_{hi}'\tilde{\mathbf{B}} + (\sum_U (1 - a_{hi})\mathbf{x}_{hi}')\tilde{\mathbf{T}}^{-1}a_{hi}\mathbf{x}_{hi}\tilde{\epsilon}_{hi}$. A standard estimator of $V_s(\sum_s w_{hi}\tilde{\xi}_{hi})$ is

$$\sum_s (1 - \frac{n_h}{N_h}) \frac{N_h^2 (\tilde{\xi}_{hi} - \tilde{\xi}_h)^2}{n_h (n_h - 1)},$$

where $\tilde{\xi}_h = \sum_{s_h} \tilde{\xi}_{hi}/n_h$. This estimator depends on the unknown $\tilde{\xi}_{hi}$, so we estimate $\tilde{\xi}_{hi}$ with $\hat{\xi}_{hi} = a_{hi}y_{hi} + (1 - a_{hi})\mathbf{x}_{hi}'\hat{\mathbf{B}}_r + (\sum_s w_{hi}(1 - a_{hi})\mathbf{x}_{hi}')\hat{\mathbf{T}}_r^{-1}a_{hi}\mathbf{x}_{hi}\hat{\epsilon}_{hi}$ where $\hat{\epsilon}_{hi} = y_{hi} - \mathbf{x}_{hi}'\hat{\mathbf{B}}_r$. Our estimate of $E_r V_s(\hat{Y}^I)$ is now given by

$$v_1 = \sum_s (1 - \frac{n_h}{N_h}) \frac{N_h^2 (\hat{\xi}_{hi} - \hat{\xi}_h)^2}{n_h (n_h - 1)} \quad (\text{A.2})$$

with $\hat{\xi}_h = \sum_{s_h} \hat{\xi}_{hi}/n_h$.

Step 3: Derive an Estimator for $V_r E_s(\hat{Y}^I)$

The regression-imputed estimator in (A.1) can be expressed as

$$\hat{Y}^I = \sum_s w_{hi} y_{hi} (a_{hi} + \hat{c}_{hi} a_{hi}),$$

where $\hat{c}_{hi} = (\sum_s w_{hi} (1 - a_{hi}) \mathbf{x}_{hi}') \hat{\mathbf{T}}_r^{-1} \mathbf{x}_{hi}$ with $\hat{\mathbf{T}}_r = (\sum_s w_{hi} a_{hi} \mathbf{x}_{hi} \mathbf{x}_{hi}')$. Taking expectation with respect to sampling, and noting that $E_s(\hat{c}_{hi}) \approx \tilde{c}_{hi} = (\sum_U (1 - a_{hi}) \mathbf{x}_{hi}') \tilde{\mathbf{T}}^{-1} \mathbf{x}_{hi}$ we have

$$\begin{aligned} E_s(\hat{Y}^I - Y) &\approx \sum_U (\tilde{c}_{hi} a_{hi} - (1 - a_{hi})) y_{hi} \\ &= \sum_U (\tilde{c}_{hi} a_{hi} - (1 - a_{hi})) \tilde{\epsilon}_{hi} + \sum_U (\tilde{c}_{hi} a_{hi} - (1 - a_{hi})) \mathbf{x}_{hi}' \tilde{\mathbf{B}} \end{aligned}$$

Since $\sum_U (\tilde{c}_{hi} a_{hi} - (1 - a_{hi})) \mathbf{x}_{hi}' \tilde{\mathbf{B}} = 0$, we have

$$E_s(\hat{Y}^I - Y) \approx \sum_U ((\tilde{c}_{hi} + 1) a_{hi} - 1) \tilde{\epsilon}_{hi} \quad (\text{A.3})$$

Design-based Variance Estimator

From (A.3), we derive $V_r E_s(\hat{Y}^I - Y)$ through the design-based approach, assuming an MCAR mechanism on the response indicators, a_{hi} . Under MCAR, the a_{hi} 's are independent for each unit (hi), have common expectation, p , and common variance, $p(1 - p)$. Noting that $E_r(\tilde{c}_{hi}) = c_{hi}$ where $c_{hi} = \frac{(1-p)}{p} (\sum_U \mathbf{x}_{hi}') \mathbf{T}^{-1} \mathbf{x}_{hi}$ with $\mathbf{T} = (\sum_U \mathbf{x}_{hi} \mathbf{x}_{hi}')$.

$$\begin{aligned} V_r E_s(\hat{Y}^I - Y) &= V_r (\sum_U ((\tilde{c}_{hi} + 1) a_{hi} - 1) \tilde{\epsilon}_{hi}) \\ &\approx V_r ((\sum_U (c_{hi} + 1) a_{hi} - 1) \tilde{\epsilon}_{hi}) \\ &= \sum_U \epsilon_{hi}^2 (c_{hi} + 1)^2 V_r(a_{hi}) \\ &= \sum_U \epsilon_{hi}^2 (c_{hi} + 1)^2 p(1 - p) \end{aligned}$$

with $\epsilon_{hi} = y_{hi} - \mathbf{x}_{hi}'\mathbf{T}^{-1}(\sum_U y_{hi}\mathbf{x}_{hi})$. We estimate this quantity by

$$\begin{aligned} v_{2D} &= \sum_s w_{hi} a_{hi} \hat{\epsilon}_{hi}^2 (\hat{c}_{hi} + 1)^2 \hat{p}(1 - \hat{p}) \frac{N}{\sum_s w_{hi} a_{hi}} \\ &= N \hat{p}(1 - \hat{p}) \tilde{s}_d^2 \end{aligned}$$

where $\hat{p} = \sum_s w_{hi} a_{hi} / N$ and $\tilde{s}_d^2 = \sum_s w_{hi} a_{hi} \hat{\epsilon}_{hi}^2 (\hat{c}_{hi} + 1)^2 / \sum_s w_{hi} a_{hi}$.

Model-assisted Variance Estimator

For the development of the model-assisted variance estimator, we assume that the response mechanism is INR and that the imputation model is true. The imputation model is specified as follows:

$$\mathbf{u}_{hi} = (\mathbf{x}_{hi}),$$

$$E(y_{hi} | \mathbf{u}_{hi}) = \mathbf{u}_{hi}' \boldsymbol{\beta},$$

$$V(y_{hi} | \mathbf{u}_{hi}) = \sigma^2 V(\mathbf{u}_{hi}), \text{ and}$$

$$COV(y_{hi}, y_{jk}) = 0 \text{ for } (hi) \neq (jk).$$

From (A.3) it follows that

$$\begin{aligned} E_r V_m(E_s(\hat{Y}^I - Y)) &= E_r(\sum_U V_m(\epsilon_{hi} | \mathbf{u}_i) ((\tilde{c}_{hi} + 1)a_{hi} - 1)^2) \\ &\approx E_r(\sum_U V_m(\epsilon_{hi} | \mathbf{u}_i) (\tilde{c}_{hi}^2 a_{hi} - (1 - a_{hi}))) \\ &= \sigma^2 E_r(\sum_U V_m(\mathbf{u}_{hi}) (\tilde{c}_{hi}^2 a_{hi} - (1 - a_{hi}))) \end{aligned}$$

We estimate this quantity by

$$v_{2M} = \hat{\sigma}^2 \sum_s w_{hi} a_{vhi} v(\mathbf{u}_i) (\hat{c}_{hi}^2 a_{hi} - (1 - a_{hi})) * (N / \sum_s w_{hi} a_{vhi})$$

where $a_{vhi} = 1$ for all (hi) where we can estimate $v(u_i)$. Setting $v(u_i) \equiv 1$ implies $a_{vhi} \equiv 1$ and gives the following estimator:

$$v_{2M} = \hat{\sigma}^2 \sum_s w_{hi} (\hat{c}_{hi}^2 a_{hi} - (1 - a_{hi})).$$

Substituting $s_d^2 = \sum_s w_{hi} a_{hi} \epsilon_{hi}^2 / \sum_s w_{hi} a_{hi}$ for $\hat{\sigma}^2$ gives

$$v_{2M} = s_d^2 \sum_s w_{hi} (\hat{c}_{hi}^2 a_{hi} - (1 - a_{hi})).$$

Appendix B

COMPOSITE IMPUTATION FOR THE MWRTS

In this appendix, we develop variance estimators for the composite-imputed estimator used in the Monthly Wholesale and Retail Trade Survey (MWRTS) conducted by Statistics Canada. We follow the reverse approach elaborated in Shao and Steel (1999). We use the following indicator variables: $a_{0hi} = 1$ if y_{hi} is reported and 0 otherwise, $a_{1hi} = 1$ if x_{1hi} is reported and 0 otherwise and $a_{2hi} = 1$ if x_{2hi} is reported and 0 otherwise. Note that x_{3hi} is available for every unit in the population. The imputation strategy is outlined as follows:

- If y_{hi} is reported ($a_{0hi} = 1$), then $y_{hi}^I = y_{hi}$
- Otherwise, if x_{1hi} and x_{2hi} are available ($(1 - a_{0hi})a_{1hi}a_{2hi} = 1$), then $y_{hi}^I = \mathbf{x}_{hi}' \hat{\mathbf{B}}_r$
- Otherwise, if x_{1hi} is available ($(1 - a_{0hi})a_{1hi}(1 - a_{2hi}) = 1$) then $y_{hi}^I = \hat{R}_{r1}x_{1hi}$
- Otherwise, if x_{2hi} is available ($(1 - a_{0hi})(1 - a_{1hi})a_{2hi} = 1$) then $y_{hi}^I = \hat{R}_{r2}x_{2hi}$
- Otherwise ($(1 - a_{0hi})(1 - a_{1hi})(1 - a_{2hi}) = 1$), $y_{hi}^I = \hat{R}_{r3}x_{3hi}$

where $\hat{\mathbf{B}}_r = (\sum_s w_{hi} a_{0hi} a_{1hi} a_{2hi} \mathbf{x}_{hi} \mathbf{x}_{hi}')^{-1} (\sum_s w_{hi} a_{0hi} a_{1hi} a_{2hi} \mathbf{x}_{hi} y_{hi})$,

$$\hat{R}_{r1} = \frac{\sum_s w_{hi} a_{0hi} a_{1hi} y_{hi}}{\sum_s w_{hi} a_{0hi} a_{1hi} x_{1hi}}, \hat{R}_{r2} = \frac{\sum_s w_{hi} a_{0hi} a_{2hi} y_{hi}}{\sum_s w_{hi} a_{0hi} a_{2hi} x_{2hi}}, \text{ and } \hat{R}_{r3} = \frac{\sum_s w_{hi} a_{0hi} y_{hi}}{\sum_s w_{hi} a_{0hi} x_{3hi}}.$$

Step 1: Re-express the estimator

The composite-imputed Horvitz-Thompson estimator is expressed as

$$\begin{aligned}\hat{Y}^I &= \sum_s w_{hi} (a_{0hi} y_{hi} + (1 - a_{0hi}) a_{1hi} a_{2hi} \mathbf{x}_{hi}' \hat{\mathbf{B}}_r + (1 - a_{0hi}) a_{1hi} (1 - a_{2hi}) \hat{R}_{r1} x_{1hi} \\ &\quad + (1 - a_{0hi}) (1 - a_{1hi}) a_{2hi} \hat{R}_{r2} x_{2hi} + (1 - a_{0hi}) (1 - a_{1hi}) (1 - a_{2hi}) \hat{R}_{r3} x_{3hi}), \\ &= \sum_s w_{hi} y_{hi}^I\end{aligned}$$

Step 2: Derive an Estimator for $E_r V_s(\hat{Y}^I)$

We first note through derivations similar to that used in Step 2 of Appendix A that

$$\bullet \sum_s w_{hi} (1 - a_{0hi}) a_{1hi} a_{2hi} \mathbf{x}_{hi}' \hat{\mathbf{B}}_r \approx \sum_s w_{hi} (1 - a_{0hi}) a_{1hi} a_{2hi} \mathbf{x}_{hi}' \tilde{\mathbf{B}} + \sum_s w_{hi} a_{0hi} a_{1hi} a_{2hi} \tilde{c}_{1hi} \tilde{\epsilon}_{1hi}$$

with $\tilde{c}_{1hi} = (\sum_U (1 - a_{0hi}) a_{1hi} a_{2hi} \mathbf{x}_{hi}) (\sum_U a_{0hi} a_{1hi} a_{2hi} \mathbf{x}_{hi}' \mathbf{x}_{hi})^{-1} \mathbf{x}_{hi}$ and $\tilde{\epsilon}_{1hi} = y_{hi} - \mathbf{x}_{hi}' \tilde{\mathbf{B}}$ with $\tilde{\mathbf{B}} = (\sum_U a_{0hi} a_{1hi} a_{2hi} \mathbf{x}_{hi} \mathbf{x}_{hi}')^{-1} (\sum_U a_{0hi} a_{1hi} a_{2hi} \mathbf{x}_{hi} y_{hi})$.

$$\begin{aligned}\bullet \sum_s w_{hi} (1 - a_{0hi}) a_{1hi} (1 - a_{2hi}) \hat{R}_{r1} x_{1hi} &\approx \sum_s w_{hi} (1 - a_{0hi}) a_{1hi} (1 - a_{2hi}) \tilde{R}_1 x_{1hi} + \\ \sum_s w_{hi} a_{0hi} a_{1hi} \tilde{c}_2 \tilde{\epsilon}_{2hi} &\text{ with } \tilde{c}_2 = \frac{\sum_U (1 - a_{0hi}) a_{1hi} (1 - a_{2hi}) x_{1hi}}{\sum_U a_{0hi} a_{1hi} x_{1hi}} \text{ and } \tilde{\epsilon}_{2hi} = y_{hi} - \tilde{R}_1 x_{1hi} \text{ with} \\ \tilde{R}_1 &= \frac{\sum_U a_{0hi} a_{1hi} y_{hi}}{\sum_U a_{0hi} a_{1hi} x_{1hi}}.\end{aligned}$$

$$\begin{aligned}\bullet \sum_s w_{hi} (1 - a_{0hi}) (1 - a_{1hi}) a_{2hi} \hat{R}_{r2} x_{2hi} &\approx \sum_s w_{hi} (1 - a_{0hi}) (1 - a_{1hi}) a_{2hi} \tilde{R}_2 x_{2hi} + \\ \sum_s w_{hi} a_{0hi} a_{2hi} \tilde{c}_3 \tilde{\epsilon}_{3hi} &\text{ with } \tilde{c}_3 = \frac{\sum_U (1 - a_{0hi}) (1 - a_{1hi}) a_{2hi} x_{2hi}}{\sum_U a_{0hi} a_{2hi} x_{2hi}} \text{ and } \tilde{\epsilon}_{3hi} = y_{hi} - \tilde{R}_2 x_{2hi} \text{ with} \\ \tilde{R}_2 &= \frac{\sum_U a_{0hi} a_{2hi} y_{hi}}{\sum_U a_{0hi} a_{2hi} x_{2hi}}.\end{aligned}$$

$$\begin{aligned}\bullet \sum_s w_{hi} (1 - a_{0hi}) (1 - a_{1hi}) (1 - a_{2hi}) \hat{R}_{r3} x_{3hi} &\approx \sum_s w_{hi} (1 - a_{0hi}) (1 - a_{1hi}) (1 - a_{2hi}) \tilde{R}_3 x_{3hi} + \\ \sum_s w_{hi} a_{0hi} \tilde{c}_4 \tilde{\epsilon}_{4hi} &\text{ with } \tilde{c}_4 = \frac{\sum_U (1 - a_{0hi}) (1 - a_{1hi}) (1 - a_{2hi}) x_{3hi}}{\sum_U a_{0hi} x_{3hi}} \text{ and } \tilde{\epsilon}_{4hi} = y_{hi} - \tilde{R}_3 x_{3hi} \text{ with}\end{aligned}$$

$$\tilde{R}_3 = \frac{\sum_U a_{0hi} y_{hi}}{\sum_U a_{0hi} x_{3hi}}.$$

Using these results, the estimate of $E_r V_s(\hat{Y}^I)$ is obtained by a Taylor linearization technique. It is derived by re-expressing the composite-imputed estimator as

$$\begin{aligned}\hat{Y}^I &= \sum_s w_{hi} a_{0hi} y_{hi}^I \\ &\approx \sum_s w_{hi} \tilde{\xi}_{hi}\end{aligned}$$

$$\begin{aligned}\text{where } \tilde{\xi}_{hi} &= a_{0hi} y_{hi} + (1 - a_{0hi}) a_{1hi} a_{2hi} \mathbf{x}_{hi}' \tilde{\mathbf{B}} + (1 - a_{0hi}) a_{1hi} (1 - a_{2hi}) \tilde{R}_{r1} x_{1hi} \\ &\quad + (1 - a_{0hi}) (1 - a_{1hi}) a_{2hi} \tilde{R}_{r2} x_{2hi} + (1 - a_{0hi}) (1 - a_{1hi}) (1 - a_{2hi}) \tilde{R}_{r3} x_{3hi} \\ &\quad + a_{0hi} a_{1hi} a_{2hi} \tilde{c}_{1hi} \tilde{\epsilon}_{1hi} + a_{0hi} a_{1hi} (1 - a_{2hi}) \tilde{c}_2 \tilde{\epsilon}_{2hi} + a_{0hi} (1 - a_{1hi}) a_{2hi} \tilde{c}_3 \tilde{\epsilon}_{3hi} \\ &\quad + a_{0hi} (1 - a_{1hi}) (1 - a_{2hi}) \tilde{c}_4 \tilde{\epsilon}_{4hi}.\end{aligned}$$

A standard estimator of $E_r V_s(\hat{Y}^I)$ is

$$\sum_s \left(1 - \frac{n_h}{N_h}\right) \frac{N_h^2 (\tilde{\xi}_{hi} - \tilde{\xi}_h)^2}{n_h (n_h - 1)}$$

where $\tilde{\xi}_h = \sum_{s_h} \tilde{\xi}_{hi} / n_h$. Since this estimator depends on the unknown $\tilde{\xi}_{hi}$'s we estimate $\tilde{\xi}_{hi}$ with $\hat{\xi}_{hi} = y_{hi}^I + a_{0hi} a_{1hi} a_{2hi} \hat{c}_{1hi} \hat{\epsilon}_{1hi} + a_{0hi} a_{1hi} (1 - a_{2hi}) \hat{c}_2 \hat{\epsilon}_{2hi} + a_{0hi} (1 - a_{1hi}) a_{2hi} \hat{c}_3 \hat{\epsilon}_{3hi} + a_{0hi} (1 - a_{1hi}) (1 - a_{2hi}) \hat{c}_4 \hat{\epsilon}_{4hi}$ where $\hat{c}_{1hi} = (\sum_s w_{hi} (1 - a_{0hi}) a_{1hi} a_{2hi} \mathbf{x}_{hi}') / (\sum_s w_{hi} a_{0hi} a_{1hi} a_{2hi} x_{1hi})$, $\hat{c}_2 = \sum_s w_{hi} (1 - a_{0hi}) a_{1hi} (1 - a_{2hi}) x_{1hi} / \sum_s w_{hi} a_{0hi} a_{1hi} x_{1hi}$, $\hat{c}_3 = \sum_s w_{hi} (1 - a_{0hi}) (1 - a_{1hi}) a_{2hi} x_{1hi} / \sum_s w_{hi} a_{0hi} a_{2hi} x_{2hi}$, $\hat{c}_4 = \sum_s w_{hi} (1 - a_{0hi}) (1 - a_{1hi}) (1 - a_{2hi}) x_{3hi} / \sum_s w_{hi} a_{0hi} x_{3hi}$, $\hat{\epsilon}_{1hi} = y_{hi} - \mathbf{x}_{hi}' \hat{\mathbf{B}}_r$, $\hat{\epsilon}_{2hi} = y_{hi} - \hat{R}_{r1} x_{1hi}$, $\hat{\epsilon}_{3hi} = y_{hi} - \hat{R}_{r2} x_{2hi}$ and $\hat{\epsilon}_{4hi} = y_{hi} - \hat{R}_{r3} x_{3hi}$.

Substituting $\hat{\xi}_{hi}$ for $\tilde{\xi}_{hi}$ our estimate of $E_r V_s(\hat{Y}^I)$ is

$$v_1 = \sum_s \left(1 - \frac{n_h}{N_h}\right) \frac{N_h^2 (\hat{\xi}_{hi} - \hat{\xi}_h)^2}{n_h (n_h - 1)}, \quad (\text{B.1})$$

with $\hat{\xi}_{hi}$ as above and $\hat{\xi}_h = \sum_{s_h} \hat{\xi}_{hi} / n_h$.

Step 3: Derive an Estimator for $V_r E_s(\hat{Y}^I)$

Taking the expectation with respect to sampling, we have

$$\begin{aligned}
E_s(\hat{Y}^I - Y) &\approx \sum_U (\tilde{c}_{1hi} a_{0hi} a_{1hi} a_{2hi} - (1 - a_{0hi}) a_{1hi} a_{2hi}) \tilde{\epsilon}_{1hi} \\
&\quad + \sum_U (\tilde{c}_{2hi} a_{0hi} a_{1hi} - (1 - a_{0hi}) a_{1hi} (1 - a_{2hi})) \tilde{\epsilon}_{2hi} \\
&\quad + \sum_U (\tilde{c}_{3hi} a_{0hi} a_{2hi} - (1 - a_{0hi}) (1 - a_{1hi}) a_{2hi}) \tilde{\epsilon}_{3hi} \\
&\quad + \sum_U (\tilde{c}_{4hi} a_{0hi} - (1 - a_{0hi}) (1 - a_{1hi}) (1 - a_{2hi})) \tilde{\epsilon}_{4hi}, \\
&= T1 + T2 + T3 + T4
\end{aligned}$$

where $T1, T2, T3$, and $T4$ to denote the first, second, third and fourth term of the expression respectively.

Design-based Variance Estimator

From the above expression, we derive the variance estimator through the design-based approach. Under this approach, we assume an MCAR mechanism. Under MCAR the response indicators, a_{0hi} , a_{1hi} , and a_{2hi} are independent over (hi) . For our purposes, we denote the properties of interest of these variables using the following notation: $E_r(a_{0hi} a_{1hi} a_{2hi}) = p_{(1)}$ and $V_r(a_{0hi} a_{1hi} a_{2hi}) = p_{(1)}(1 - p_{(1)})$; $E_r((1 - a_{0hi}) a_{1hi} a_{2hi}) = p_1$ and $V_r((1 - a_{0hi}) a_{1hi} a_{2hi}) = p_1(1 - p_1)$; $E_r((1 - a_{0hi}) a_{1hi} (1 - a_{2hi})) = p_2$ and $V_r((1 - a_{0hi}) a_{1hi} (1 - a_{2hi})) = p_2(1 - p_2)$; $E_r(a_{0hi} a_{1hi}) = p_{(2)}$ and $V_r(a_{0hi} a_{1hi}) = p_{(2)}(1 - p_{(2)})$; $E_r((1 - a_{0hi}) (1 - a_{1hi}) a_{2hi}) = p_3$ and $V_r((1 - a_{0hi}) (1 - a_{1hi}) a_{2hi}) = p_3(1 - p_3)$; $E_r(a_{0hi} a_{2hi}) = p_{(3)}$ and $V_r(a_{0hi} a_{2hi}) = p_{(3)}(1 - p_{(3)})$; $E_r((1 - a_{0hi}) (1 - a_{1hi}) (1 - a_{2hi})) = p_4$ and $V_r((1 - a_{0hi}) (1 - a_{1hi}) (1 - a_{2hi})) = p_4(1 - p_4)$; $E_r(a_{0hi}) = p_{(4)}$ and $V_r(a_{0hi}) = p_{(4)}(1 - p_{(4)})$.

$$\begin{aligned}
V_r E_s(\hat{Y}_I - Y) &= V_r(T1) + V_r(T2) + V_r(T3) + V_r(T4) \\
&\quad + 2COV_r(T1, T2) + 2COV_r(T1, T3) + 2COV_r(T1, T4) \\
&\quad + 2COV_r(T2, T3) + 2COV_r(T2, T4) + 2COV_r(T3, T4)
\end{aligned}$$

We define the following quantities: $c_{1hi} = (\sum_U p_1 \mathbf{x}_{hi}')(\sum_U p_{(1)} \mathbf{x}_{hi} \mathbf{x}_{hi}')^{-1} \mathbf{x}_{hi}$, $c_2 = p_2/p_{(2)}$, $c_3 = p_3/p_{(3)}$, $c_4 = p_4/p_{(4)}$, and note that $\tilde{c}_{1hi} \approx c_{1hi}$, $\tilde{c}_2 \approx c_2$, $\tilde{c}_3 \approx c_3$ and $\tilde{c}_4 \approx c_4$

Examining $V_r(T1)$, we find that

$$\begin{aligned} V_r(T1) &= V_r(\sum_U (\tilde{c}_{1hi} a_{0hi} a_{1hi} a_{2hi} - (1 - a_{0hi}) a_{1hi} a_{2hi}) \tilde{\epsilon}_{1hi}) \\ &= \sum_U \epsilon_{1hi}^2 V_r(\tilde{c}_{1hi} a_{0hi} a_{1hi} a_{2hi} + (1 - a_{0hi}) a_{1hi} a_{2hi}) \\ &= \sum_U \epsilon_{1hi}^2 (E_r((\tilde{c}_{1hi} a_{0hi} a_{1hi} a_{2hi} - (1 - a_{0hi}) a_{1hi} a_{2hi})^2) \\ &\quad - E_r(\tilde{c}_{1hi} a_{0hi} a_{1hi} a_{2hi} - (1 - a_{0hi}) a_{1hi} a_{2hi})^2) \\ &= \sum_U \epsilon_{1hi}^2 E_r(\tilde{c}_{1hi}^2 (p_{(1)}(1 - p_{(1)}) + p_1(1 - p_1) + 2\tilde{c}_{1hi} p_{(1)} p_1)) \\ &\approx \sum_U \epsilon_{1hi}^2 (c_{1hi}^2 (p_{(1)}(1 - p_{(1)}) + p_1(1 - p_1) - 2c_{1hi} p_{(1)} p_1) \end{aligned}$$

where $\epsilon_{1hi} = y_{hi} - \mathbf{x}_{hi}'(\sum_U \mathbf{x}_{hi} \mathbf{x}_{hi}')(\sum_U y_{hi} \mathbf{x}_{hi})$. Similarly, $V_r(T2) \approx \sum_U \epsilon_{2hi}^2 (c_2 + 1)p_2$ with $\epsilon_{2hi} = y_{hi} - \sum_U y_{hi} / \sum_U x_{1hi}$, $V_r(T3) \approx \sum_U \epsilon_{3hi}^2 (c_3 + 1)p_3$ with $\epsilon_{3hi} = y_{hi} - \sum_U y_{hi} / \sum_U x_{2hi}$, and $V_r(T4) \approx \sum_U \epsilon_{4hi}^2 (c_4 + 1)p_4$ with $\epsilon_{4hi} = y_{hi} - \sum_U y_{hi} / \sum_U x_{3hi}$.

The covariance terms simplify as follows:

$$\begin{aligned} COV_r(T1, T2) &= E_r(\sum_U \tilde{\epsilon}_{1hi} (\tilde{c}_{1hi} a_{0hi} a_{1hi} a_{2hi} + (1 - a_{0hi}) a_{1hi} a_{2hi}) \cdot \\ &\quad \sum_U \tilde{\epsilon}_{2hi} (\tilde{c}_2 a_{0hi} a_{1hi} - (1 - a_{0hi}) a_{1hi} (1 - a_{2hi}))) \\ &\quad - E_r(\sum_U \tilde{\epsilon}_{1hi} (\tilde{c}_{1hi} a_{0hi} a_{1hi} a_{2hi} - (1 - a_{0hi}) a_{1hi} a_{2hi})) \cdot \\ &\quad E_r(\sum_U \tilde{\epsilon}_{2hi} (\tilde{c}_2 a_{0hi} a_{1hi} - (1 - a_{0hi}) a_{1hi} (1 - a_{2hi}))) \\ &= E_r(\sum_U \tilde{\epsilon}_{1hi} \tilde{\epsilon}_{2hi} (\tilde{c}_{1hi} a_{0hi} a_{1hi} a_{2hi} - (1 - a_{0hi}) a_{1hi} a_{2hi}) \cdot \\ &\quad (\tilde{c}_2 a_{0hi} a_{1hi} - (1 - a_{0hi}) a_{1hi} (1 - a_{2hi}))) \end{aligned}$$

since $E_r(\sum_U \tilde{\epsilon}_{2hi} (\tilde{c}_2 a_{0hi} a_{1hi} + (1 - a_{0hi}) a_{1hi} (1 - a_{2hi}))) = 0$. Thus, since $a_{0hi}(1 - a_{0hi}) \equiv 0$ and $a_{2hi}(1 - a_{2hi}) \equiv 0$,

$$COV_r(T1, T2) = E_r(\sum_U \tilde{\epsilon}_{1hi} \tilde{\epsilon}_{2hi} \tilde{c}_{1hi} \tilde{c}_2 a_{0hi} a_{1hi} a_{2hi})$$

Under MCAR, $E_r(a_{0hi} a_{1hi} a_{2hi}) = p_{(1)}$, so

$$COV_r(T1, T2) = p_{(1)} \sum_U \epsilon_{1hi} \epsilon_{2hi} c_{1hi} c_2.$$

Similarly, $COV(T1, T3) = p_{(1)} \sum_U \epsilon_{1hi} \epsilon_{3hi} c_{1hi} c_3$, $COV(T1, T4) = p_{(1)} \sum_U \epsilon_{1hi} \epsilon_{4hi} c_{1hi} c_4$,
 $COV(T2, T3) = p_{(1)} \sum_U \epsilon_{2hi} \epsilon_{3hi} c_2 c_3$, $COV(T2, T4) = p_{(2)} \sum_U \epsilon_{2hi} \epsilon_{4hi} c_2 c_4$ and $COV(T3, T4) = p_{(3)} \sum_U \epsilon_{3hi} \epsilon_{4hi} c_3 c_4$.

Therefore,

$$\begin{aligned} V_r E_s(\hat{Y}_I - Y) &= \sum_U \epsilon_{1hi}^2 (c_{1hi}^2 p_{(1)} (1 - p_{(1)}) + p_1 (1 - p_1) - 2c_{1hi} p_{(1)} p_1) \\ &\quad + \sum_U \epsilon_{2hi}^2 (c_2 + 1) p_2 + \sum_U \epsilon_{3hi}^2 (c_3 + 1) p_3 \\ &\quad + \sum_U \epsilon_{4hi}^2 (c_4 + 1) p_4 + 2p_{(1)} \sum_U \epsilon_{1hi} \epsilon_{2hi} c_{1hi} c_2 + 2p_{(1)} \sum_U \epsilon_{1hi} \epsilon_{3hi} c_{1hi} c_3 \\ &\quad + 2p_{(1)} \sum_U \epsilon_{1hi} \epsilon_{4hi} c_{1hi} c_4 + 2p_{(1)} \sum_U \epsilon_{2hi} \epsilon_{3hi} c_2 c_3 \\ &\quad + 2p_{(2)} \sum_U \epsilon_{2hi} \epsilon_{4hi} c_2 c_4 + 2p_{(3)} \sum_U \epsilon_{3hi} \epsilon_{4hi} c_3 c_4 \end{aligned}$$

We estimate this quantity by

$$\begin{aligned} v_{2D} &= \frac{1}{\hat{p}_{(1)}} \sum_s w_{hi} a_{0hi} a_{1hi} a_{2hi} \hat{\epsilon}_{1hi}^2 (\hat{c}_{1hi}^2 \hat{p}_{(1)} (1 - \hat{p}_{(1)}) + \hat{p}_1 (1 - \hat{p}_1) - 2\hat{c}_{1hi} \hat{p}_{(1)} \hat{p}_1) \\ &\quad + \frac{\hat{p}_2}{\hat{p}_{(2)}} \sum_s w_{hi} a_{0hi} a_{1hi} \hat{\epsilon}_{2hi}^2 (\hat{c}_2 + 1) + \frac{\hat{p}_3}{\hat{p}_{(3)}} \sum_s w_{hi} a_{0hi} a_{2hi} \hat{\epsilon}_{3hi}^2 (\hat{c}_3 + 1) \\ &\quad + \frac{\hat{p}_4}{\hat{p}_{(4)}} \sum_s w_{hi} a_{0hi} \hat{\epsilon}_{4hi}^2 (\hat{c}_4 + 1) + 2 \sum_s w_{hi} \hat{c}_{1hi} \hat{c}_2 \hat{\epsilon}_{1hi} \hat{\epsilon}_{2hi} \\ &\quad + 2 \sum_s w_{hi} \hat{c}_{1hi} \hat{c}_3 \hat{\epsilon}_{1hi} \hat{\epsilon}_{3hi} + 2 \sum_s w_{hi} \hat{c}_{1hi} \hat{c}_4 \hat{\epsilon}_{1hi} \hat{\epsilon}_{4hi} \\ &\quad + 2 \sum_s w_{hi} \hat{c}_2 \hat{c}_3 \hat{\epsilon}_{2hi} \hat{\epsilon}_{3hi} + 2 \sum_s w_{hi} \hat{c}_2 \hat{c}_4 \hat{\epsilon}_{2hi} \hat{\epsilon}_{4hi} + 2 \sum_s w_{hi} \hat{c}_3 \hat{c}_4 \hat{\epsilon}_{3hi} \hat{\epsilon}_{4hi} \end{aligned}$$

where $\hat{p}_1 = \sum_s w_{hi} (1 - a_{0hi}) a_{1hi} a_{2hi} / N$, $\hat{p}_{(1)} = \sum_s w_{hi} a_{0hi} a_{1hi} a_{2hi} / N$, $\hat{p}_2 = \sum_s w_{hi} (1 - a_{0hi}) a_{1hi} (1 - a_{2hi}) / N$, $\hat{p}_{(2)} = \sum_s w_{hi} a_{0hi} a_{1hi} / N$, $\hat{p}_3 = \sum_s w_{hi} (1 - a_{0hi}) (1 - a_{1hi}) a_{2hi} / N$, $\hat{p}_{(3)} = \sum_s w_{hi} a_{0hi} a_{2hi} / N$, $\hat{p}_4 = \sum_s w_{hi} (1 - a_{0hi}) (1 - a_{1hi}) (1 - a_{2hi}) / N$ and $\hat{p}_{(4)} = \sum_s w_{hi} a_{0hi} / N$.

Model-assisted Variance Estimator

For the development of the model-assisted variance estimator, we assume that the response mechanism is INR and that the imputation model is true. The imputation

model is specified as follows: Let $\mathbf{u}_{hi} = (u_{1hi}, u_{2hi}, u_{3hi}, u_{4hi}) = (x_{hi}, x_{1hi}, x_{2hi}, x_{3hi})$.

$$E(y_{hi}|u_{1i}) = \mathbf{u}_{1hi}'\beta$$

$$E(y_{hi}|u_{2i}) = \gamma_2 u_{2hi},$$

$$E(y_{hi}|u_{3i}) = \gamma_3 u_{3hi},$$

$$E(y_{hi}|u_{4i}) = \gamma_4 u_{4hi},$$

$$V(y_{hi}|\mathbf{u}_i) = \sigma^2 v(\mathbf{u}_{hi}), \text{ and}$$

$$COV(y_{hi}, y_{jk}) = 0 \text{ for } (hi) \neq (jk).$$

Therefore,

$$\begin{aligned} E_r V_m(E_s(\hat{Y}^I - Y)) &= E_r(\sum_U (1 - a_{0hi}) a_{1hi} a_{2hi} V_m(\epsilon_{1hi}|\mathbf{u}_{hi}) \\ &\quad + \sum_U (1 - a_{0hi}) a_{1hi} (1 - a_{2hi}) V_m(\epsilon_{2hi}|\mathbf{u}_{hi}) \\ &\quad + \sum_U (1 - a_{0hi}) (1 - a_{1hi}) a_{2hi} V_m(\epsilon_{3hi}|\mathbf{u}_{hi}) \\ &\quad + \sum_U (1 - a_{0hi}) (1 - a_{1hi}) (1 - a_{2hi}) V_m(\epsilon_{4hi}|\mathbf{u}_{hi}) \\ &\quad + \sum_U V_m(a_{0hi} a_{1hi} a_{2hi} \hat{c}_{1hi} \epsilon_{1hi} + a_{0hi} a_{1hi} \hat{c}_2 \epsilon_{2hi} \\ &\quad + a_{0hi} a_{2hi} \hat{c}_3 \epsilon_{3hi} + a_{0hi} \hat{c}_4 \epsilon_{4hi}|\mathbf{u}_{hi})) \\ &= E_r(\sum_U (1 - a_{0hi}) a_{1hi} a_{2hi} + \sum_U (1 - a_{0hi}) a_{1hi} (1 - a_{2hi}) \\ &\quad + \sum_U (1 - a_{0hi}) (1 - a_{1hi}) a_{2hi} + \sum_U (1 - a_{0hi}) (1 - a_{1hi}) (1 - a_{2hi}) \\ &\quad + \sum_U a_{0hi} a_{1hi} a_{2hi} \hat{c}_{1hi}^2 + a_{0hi} a_{1hi} \hat{c}_2^2 + a_{0hi} a_{2hi} \hat{c}_3^2 + a_{0hi} \hat{c}_4^2) \sigma^2 V_m(y_{hi}|\mathbf{u}_{hi})) \\ &= E_r(\sum_U ((1 - a_{0hi}) + \hat{c}_{1hi}^2 a_{0hi} a_{1hi} a_{2hi} + \hat{c}_2^2 a_{0hi} a_{1hi} \\ &\quad + \hat{c}_3^2 a_{0hi} a_{2hi} + \hat{c}_4^2 a_{0hi}) \sigma^2 V_m(\mathbf{u}_{hi}))) \end{aligned}$$

We estimate this expression by

$$\begin{aligned} v_{2M} &= \hat{\sigma}^2 \sum_s w_{hi} a_{vhi} v(\mathbf{u}_{hi}) ((1 - a_{0hi}) + \hat{c}_{1hi}^2 a_{0hi} a_{1hi} a_{2hi} \\ &\quad + \hat{c}_2^2 a_{0hi} a_{1hi} + \hat{c}_3^2 a_{0hi} a_{2hi} + \hat{c}_4^2 a_{0hi}) * (N / \sum_s w_{hi} a_{vhi}) \end{aligned}$$

where a_{vhi} is the indicator for all units where we can estimate $v(u_{hi})$. Setting $v(u_i) \equiv 1$ implies $a_{vhi} \equiv 1$ and gives the following estimator:

$$v_{2M} = \hat{\sigma}^2 \sum_s w_{hi} ((1 - a_{0hi}) + \hat{c}_{1hi}^2 a_{0hi} a_{1hi} a_{2hi} + \hat{c}_2^2 a_{0hi} a_{1hi} + \hat{c}_3^2 a_{0hi} a_{2hi} + \hat{c}_4^2 a_{0hi})$$

We use the pooled variance estimator, s_p^2 , to estimate $\hat{\sigma}^2$.

$$s_p^2 = \frac{\sum_s w_{hi} a_{0hi} (a_{1hi} a_{2hi} \hat{\epsilon}_{1hi}^2 + a_{1hi} \hat{\epsilon}_{2hi}^2 + a_{2hi} \hat{\epsilon}_{3hi}^2 + \hat{\epsilon}_{4hi}^2)}{\sum_s w_{hi} a_{0hi} (a_{1hi} a_{2hi} + a_{1hi} + a_{2hi} + 1) - 5}$$

Thus,

$$v_{2M} = s_p^2 \sum_s w_{hi} ((1 - a_{0hi}) + \hat{c}_{1hi}^2 a_{0hi} a_{1hi} a_{2hi} + \hat{c}_2^2 a_{0hi} a_{1hi} + \hat{c}_3^2 a_{0hi} a_{2hi} + \hat{c}_4^2 a_{0hi})$$