Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Survey Methodology
42-1

Release date: June 22, 2016

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology

Catalogue No. 12-001-XPB

A journal published by Statistics Canada

June 2016

•

Volume 42

•

Number 1

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year in electronic format. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (statcan.smj-rte.statcan@canada.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca/SurveyMethodology).

Survey Methodology

A Journal Published by Statistics Canada

Volume 42, Number 1, June 2016

Contents

Regular papers

Short note

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.
∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.
∞

Survey Methodology, June 2016
Vol. 42, No. 1, pp. 1-18
Statistics Canada, Catalogue No. 12-001-X

A generalized Fellegi-Holt paradigm for automatic error localization

Sander Scholtus[1]

Abstract

The aim of automatic editing is to use a computer to detect and amend erroneous values in a data set, without human intervention. Most automatic editing methods that are currently used in official statistics are based on the seminal work of Fellegi and Holt (1976). Applications of this methodology in practice have shown systematic differences between data that are edited manually and automatically, because human editors may perform complex edit operations. In this paper, a generalization of the Fellegi-Holt paradigm is proposed that can incorporate a large class of edit operations in a natural way. In addition, an algorithm is outlined that solves the resulting generalized error localization problem. It is hoped that this generalization may be used to increase the suitability of automatic editing in practice, and hence to improve the efficiency of data editing processes. Some first results on synthetic data are promising in this respect.

Key Words: Automatic editing; Edit operations; Maximum likelihood; Numerical data; Linear edits.

1 Introduction

Data that have been collected for the production of statistics inevitably contain errors. A data editing process is needed to detect and amend these errors, at least in so far as they have an appreciable impact on the quality of statistical output (Granquist and Kovar 1997). Traditionally, data editing has been a manual task, ideally performed by professional editors with extensive subject-matter knowledge. To improve the efficiency, timeliness, and reproducibility of editing, many statistical institutes have attempted to automate parts of this process (Pannekoek, Scholtus and van der Loo 2013). This has resulted in deductive correction methods for *systematic errors* and error localization algorithms for *random errors* (de Waal, Pannekoek and Scholtus 2011, Chapter 1). In this article, I will focus on automatic editing for random errors.

Methods for this task usually proceed by minimally adjusting each record of data, according to some optimization criterion, so that it becomes consistent with a given set of constraints known as *edit rules*, or *edits* for short. Depending on the effectiveness of the optimization criterion and the strength of the edit rules, automatic editing may be used as a partial alternative to traditional manual editing. In practice, automatic editing is applied nearly always in combination with some form of *selective editing*, which means that the most influential errors are treated manually (Hidiroglou and Berthelot 1986; Granquist 1995, 1997; Granquist and Kovar 1997; Lawrence and McKenzie 2000; Hedlin 2003; de Waal et al. 2011).

Most automatic editing methods that are currently used in official statistics are based on the paradigm of Fellegi and Holt (1976): for each record, the smallest subset of variables is identified as erroneous that can be imputed so that the record becomes consistent with the edits. A slight generalization is obtained by assigning so-called *confidence weights* to the variables and minimizing the total weight of the imputed variables. Once this *error localization problem* is solved, suitable new values have to be found in a separate step for the variables that were identified as erroneous. This is the so-called *consistent imputation problem*; see de Waal et al. (2011) and their references. In this article, I will focus on the error localization problem.

1. Sander Scholtus, Statistics Netherlands, Department of Process Development and Methodology, P.O. Box 24500, 2490 HA, The Hague, The Netherlands. E-mail: sshs@cbs.nl.

At *Statistics Netherlands*, error localization based on the Fellegi-Holt paradigm has been a part of the data editing process for Structural Business Statistics (SBS) for over a decade now. In evaluation studies, where the same SBS data were edited both automatically and manually, a number of systematic differences were found between the two editing efforts. Many of these differences could be explained by the fact that human editors performed certain types of adjustments that were suboptimal under the Fellegi-Holt paradigm. For instance, editors sometimes interchanged the values of associated costs and revenues items, or transferred parts of reported amounts between variables.

In practice, the outcome of manual editing is usually taken as the "gold standard" for assessing the quality of automatic editing. A critical evaluation of this assumption is beyond the scope of the present paper; however, see EDIMBUS (2007, pages 34-35). Here I simply note that, by improving the ability of automatic editing methods to mimic the results of manual editing, their usefulness in practice may be increased. In turn, this means that the share of automatic editing may be increased to improve the efficiency of the data editing process (Pannekoek et al. 2013).

To some extent, systematic differences between automatic and manual editing could be prevented by a clever choice of confidence weights. In general, however, the effects of a modification of the confidence weights on the results of automatic editing are difficult to predict. Moreover, if the editors apply a number of different complex adjustments, it might be impossible to model all of them under the Fellegi-Holt paradigm using a single set of confidence weights. Another option is to try to catch errors for which the Fellegi-Holt paradigm is known to provide an unsatisfactory solution at an earlier stage in the data editing process, i.e., during deductive correction of systematic errors through automatic correction rules (de Waal et al. 2011; Scholtus 2011). This approach has practical limitations, however, because it may require a large collection of if-then rules, which would be difficult to design and maintain over time (Chen, Thibaudeau and Winkler 2003). Moreover, it is not self-evident that appropriate correction rules can be found for all errors that do not fit within the Fellegi-Holt paradigm.

In this article, a different approach is suggested. A new definition of the error localization problem is proposed that allows for the possibility that errors affect more than one variable at a time. It is shown that this problem contains error localization under the original Fellegi-Holt paradigm as a special case. Throughout this article, I restrict attention to numerical data and linear edits; a possible extension to categorical and mixed data will be discussed briefly in Section 8.

The remainder of this article is organized as follows. Section 2 briefly reviews relevant previous work done in this area. In Section 3, the concept of an edit operation is introduced and illustrated. The new error localization problem is formulated in terms of these edit operations in Section 4. Section 5 generalizes an existing method for identifying solutions to the Fellegi-Holt-based error localization problem, and this result is used in Section 6 to outline a possible algorithm for solving the new problem. A small simulation study is discussed in Section 7. Finally, some conclusions and questions for further research follow in Section 8.

2 Background and related work

Let $\mathbf{x} = (x_1, \ldots, x_p)' \in \mathbb{R}^p$ be a record of p numerical variables. Suppose that this record has to satisfy k edit rules, in the form of the following system of linear (in)equalities:

$$\mathbf{A}\mathbf{x} + \mathbf{b} \odot \mathbf{0}, \tag{2.1}$$

where $\mathbf{A} = (a_{rj})$ is a $k \times p$ – matrix of coefficients and $\mathbf{b} = (b_1, \ldots, b_k)'$ is a vector of constants. Here and elsewhere, $\mathbf{0}$ represents a vector of zeros of appropriate length; similarly, $\odot$ represents a symbolic vector of operators from the set $\{\geq, \leq, =\}$.

For a given record $\mathbf{x}$ that does not satisfy all edits in (2.1), the Fellegi-Holt-based error localization problem amounts to finding the minimum of

$$\sum_{j=1}^{p} w_j \delta_j, \tag{2.2}$$

with $w_j > 0$ the confidence weight of variable x_j and $\delta_j \in \{0,1\}$, under the condition that the original record can be made consistent with the edits by imputing only those x_j with $\delta_j = 1$ (de Waal et al. 2011, page 66).

Fellegi and Holt (1976) also proposed a method for solving the above error localization problem, based on the generation of a sufficient set of so-called *implied edits* (see below). Unfortunately, the number of implied edits needed by this method is often extremely large in practice. Over the past decades, various dedicated algorithms for the error localization problem have been developed by, among others, Schaffer (1987), Garfinkel, Kunnathur and Liepins (1988), Kovar and Whitridge (1990), Ragsdale and McKeown (1996), de Waal (2003), de Waal and Quere (2003), Riera-Ledesma and Salazar-González (2003, 2007), Bruni (2004), and de Jonge and van der Loo (2014). Early algorithms mostly focused on strengthening the original method of Fellegi and Holt (1976) by reducing the number of required implied edits. More recent algorithms rely on the fact that the error localization problem can be written as a mixed-integer programming problem, which makes it possible to apply standard optimization techniques. See also de Waal and Coutinho (2005) or de Waal et al. (2011) for an overview and comparison of various error localization algorithms.

Implied edits are constraints that follow logically from the original edits (2.1). In the present context (numerical data, linear edits), all relevant implied edits may be generated by a technique called *Fourier-Motzkin elimination* (FM elimination; cf. Williams 1986). FM elimination transforms a system of linear constraints having p variables into a system of implied linear constraints having at most $p-1$ variables; thus, at least one of the original variables is eliminated. For mathematical details, see the appendix.

FM elimination has the following fundamental property: the system of implied constraints is satisfied by the values of the non-eliminated variables if, and only if, there exists a value for the eliminated variable that, together with the other values, satisfies the original system of constraints. In error localization under the Fellegi-Holt paradigm, by repeatedly applying this fundamental property, one may verify whether any particular combination of variables can be imputed to obtain a consistent record, given the original values of the other variables. A clear illustration of this use of FM elimination is provided by the error localization algorithm of de Waal and Quere (2003).

To conclude this section, it is interesting to look briefly at the statistical interpretation of the error localization problem. In fact, in motivating their paradigm for automatic error localization, Fellegi and Holt (1976) did not provide any formal statistical argument. Their reasoning was more intuitive:

> *"The data in each record should be made to satisfy all edits by changing the fewest possible items of data (fields). This we believe to be in agreement with the idea of keeping the maximum amount of original data unchanged, subject to the constraints of the edits, and so manufacturing as little data as possible. At the same time, if errors are comparatively rare, it*

seems more likely that we will identify the truly erroneous fields." (Fellegi and Holt 1976, page 18).

A statistical argument for minimizing the weighted number of imputed variables was provided by Liepins (1980) and Liepins, Garfinkel and Kunnathur (1982), elaborating on earlier results of Naus, Johnson and Montalvo (1972). Suppose that errors occur according to a stochastic process, with each variable x_j being observed in error with a probability p_j that does not depend on its true value and with errors being independent across variables. Suppose furthermore that the confidence weights are defined as follows:

$$w_j = -\log\left(\frac{p_j}{1-p_j}\right). \tag{2.3}$$

Then it can be shown that minimizing expression (2.2) is approximately equivalent to maximizing the likelihood of the unobserved error-free record. Note that these authors tacitly assume that an error always affects one variable at a time.

Alternative error localization procedures that are based more directly on statistical models have been proposed by, e.g., Little and Smith (1987) and Ghosh-Dastidar and Schafer (2006). These procedures use outlier detection techniques and require an explicit model for the true data. Unfortunately, they cannot handle edit rules such as (2.1) in a straightforward manner.

3 Edit operations

Continuing with the notation from Section 2, I define an *edit operation* g to be an affine function of the general form

$$g(\mathbf{x}) = \mathbf{T}\mathbf{x} + \mathbf{S}\boldsymbol{\alpha} + \mathbf{c}, \tag{3.1}$$

where $\mathbf{T}$ and $\mathbf{S}$ are known coefficient matrices of dimensions $p \times p$ and $p \times m$, respectively, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$ is a vector of free parameters that may occur in g, and $\mathbf{c}$ is a $p-$vector of known constants. In the special case that g does not involve any free parameters $(m = 0)$, the second term in (3.1) vanishes. Sometimes, it may be useful to impose one or several linear constraints on the free parameters in g:

$$\mathbf{R}\boldsymbol{\alpha} + \mathbf{d} \odot \mathbf{0}, \tag{3.2}$$

with $\mathbf{R}$ a known matrix, and $\mathbf{d}$ a known vector of constants. (Note: Matrix-vector notation will be used throughout this article because it leads to a concise description of results; however, using matrices to represent edits and edit operations is probably not the most efficient way to implement these results on a computer.)

As a first example, consider the operation that replaces one of the original values in $\mathbf{x}$ by an arbitrary new value (imputation). I will call this an *FH operation*, in view of its central role in automatic editing based on the Fellegi-Holt paradigm. Let $\mathbf{I}$ denote the $p \times p$ identity matrix and $\mathbf{e}_i$ the i^{th} standard basis vector in $\mathbb{R}^p$. The FH operation that imputes the variable x_j is given by (3.1) with $\mathbf{T} = \mathbf{I} - \mathbf{e}_j\mathbf{e}_j'$, $\mathbf{S} = \mathbf{e}_j$, and $\mathbf{c} = \mathbf{0}$. This yields: $g(\mathbf{x}) = \mathbf{x} + \mathbf{e}_j(\alpha - x_j) = (x_1, \ldots, x_{j-1}, \alpha, x_{j+1}, \ldots, x_p)'$, with $\alpha \in \mathbb{R}$ a free parameter that

represents the imputed value. It should be noted that for a record of p variables, p distinct FH operations can be defined.

To further illustrate the concept of an edit operation, some other examples will now be given. For notational convenience, I restrict attention to the case $p = 3$.

- An edit operation that changes the sign of one of the variables:

$$g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

- An edit operation that interchanges the values of two adjacent items:

$$g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_1 \\ x_3 \end{pmatrix}.$$

- An edit operation that transfers an amount between two items, where the amount transferred may equal at most K units in either direction:

$$g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}\alpha + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} x_1 + \alpha \\ x_2 \\ x_3 - \alpha \end{pmatrix}.$$

  with the constraint that $-K \le \alpha \le K$.

- An edit operation that imputes two variables simultaneously using a fixed ratio:

$$g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ x_3 \end{pmatrix}.$$

  with the constraint that $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)'$ satisfies $10\alpha_1 - \alpha_2 = 0$.

Intuitively, an edit operation is supposed to "reverse the effects" of a particular type of error that may have occurred in the observed data. That is to say, if the error associated with edit operation g actually occurred in the observed record $\mathbf{x}$, then $g(\mathbf{x})$ is the record that would have been observed if that error had not occurred. Somewhat more formally, it is assumed here that errors occurring in the data can be modeled by a stochastic "error generating process" $\mathcal{E}$, and that each edit operation acts as a "corrector" for one particular error that can occur under $\mathcal{E}$ (see Remark 4 in the next section).

If the edit operation g contains free parameters, the record $g(\mathbf{x})$ might not be determined uniquely even when the restrictions (2.1) and (3.2) are taken into account. In that case, one has to "impute" values for the free parameters that occur in an edit operation, which in turn means that some of the variables in $\mathbf{x}$

are imputed via the affine transformation given by (3.1). As in traditional Fellegi-Holt-based editing, finding appropriate "imputations" for the free parameters will not be considered part of the error localization problem here. On the other hand, if g does not contain any free parameters, the imputed values in $g(\mathbf{x})$ follow directly from the edit operation itself and the distinction between error localization and imputation is blurred.

In any particular application, only a small subset of potential edit operations of the form (3.1) would have a substantively meaningful interpretation, in the sense that the associated types of errors are known to occur. In what follows, I assume that a finite set of specific edit operations of the form (3.1) has been identified as relevant for a particular application. This will be called the set of *allowed edit operations* for that application. Some suggestions on how to construct this set will be given in Section 8.

4 A generalized error localization problem

Let $\mathcal{G}$ be a finite set of allowed edit operations for a given application of automatic editing. Informally, I propose to generalize the error localization problem of Fellegi and Holt (1976) by replacing "the smallest subset of variables that can be imputed to make the record consistent" with "the shortest sequence of allowed edit operations that can be applied to make the record consistent". To give a formal definition of this generalized error localization problem, some new notation and concepts need to be introduced.

Consider a sequence of points $\mathbf{x} = \mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t = \mathbf{y}$ in $\mathbb{R}^p$. A *path* from $\mathbf{x}$ to $\mathbf{y}$ is defined as a sequence of *distinct* edit operations $g_1, \ldots, g_t \in \mathcal{G}$ such that $\mathbf{x}_n = g_n(\mathbf{x}_{n-1})$ for all $n \in \{1, \ldots, t\}$. (Note: In the case that g_n contains free parameters, one should interpret this equality as "there exist feasible parameter values such that g_n maps $\mathbf{x}_{n-1}$ to $\mathbf{x}_n$".) A path is denoted by $P = [g_1, \ldots, g_t]$. The set of all possible paths from $\mathbf{x}$ to $\mathbf{y}$ is denoted by $\mathcal{P}(\mathbf{x}, \mathbf{y})$. This set may be empty. Later, I will use $\mathcal{P}(\mathbf{x}; G)$ to denote, for a given subset $G \subseteq \mathcal{G}$, the set of all paths starting in $\mathbf{x}$ that consist of the edit operations in G in some order (without specifying the free parameters); if G contains t elements, $\mathcal{P}(\mathbf{x}; G)$ contains $t!$ paths.

To each edit operation $g \in \mathcal{G}$, one can associate a weight $w_g > 0$ that expresses the costs of applying edit operation g. In particular, the weight of an FH operation is to be chosen equal to the confidence weight of the variable that it imputes. Now the *length* of a path $P = [g_1, \ldots, g_t]$ can be defined as the sum of the weights of its constituent edit operations: $\ell(P) = \sum_{n=1}^{t} w_{g_n}$, where, by convention, the empty path has length zero. The *distance* from $\mathbf{x}$ to $\mathbf{y}$ is defined as the length of the shortest path that connects $\mathbf{x}$ to $\mathbf{y}$:

$$d(\mathbf{x}, \mathbf{y}) = \begin{cases} \min\{\ell(P) | P \in \mathcal{P}(\mathbf{x}, \mathbf{y})\} & \text{if } \mathcal{P}(\mathbf{x}, \mathbf{y}) \neq \varnothing, \\ \infty & \text{otherwise.} \end{cases}$$

In general, $d(\mathbf{x}, \mathbf{y})$ satisfies the standard axioms of a metric *except* that it need not be symmetric in $\mathbf{x}$ and $\mathbf{y}$; it is a so-called *quasimetric* (Scholtus 2014). Accordingly, $d(\mathbf{x}, \mathbf{y})$ represents "the distance from $\mathbf{x}$ to $\mathbf{y}$" rather than "the distance between $\mathbf{x}$ and $\mathbf{y}$".

The distance from $\mathbf{x}$ to any closed, non-empty subset $D \subseteq \mathbb{R}^p$ is defined as the distance to the nearest $\mathbf{y} \in D$: $d(\mathbf{x}, D) = \min\{d(\mathbf{x}, \mathbf{y}) | \mathbf{y} \in D\}$. For the purpose of error localization, the closed, non-empty subset of $\mathbb{R}^p$ that is of particular interest is the set D_0 of all points that satisfy (2.1).

I can now formulate the generalized error localization problem.

Problem. Consider a given set of consistent records D_0, a given set of allowed edit operations $\mathcal{G}$, and a given record $\mathbf{x}$. If $d(\mathbf{x}, D_0) = \infty$, then the error localization problem for $\mathbf{x}$ is infeasible. Otherwise, any shortest path leading to a record $\mathbf{y} \in D_0$ such that $d(\mathbf{x}, \mathbf{y}) < \infty$ is called a *feasible solution* to the error localization problem for $\mathbf{x}$. A feasible solution is called *optimal* if it leads to a record $\mathbf{x}^* \in D_0$ such that

$$d(\mathbf{x}, \mathbf{x}^*) = d(\mathbf{x}, D_0). \tag{4.1}$$

Formally, then, the generalized error localization problem consists of finding an optimal path of edit operations.

Remark 1. In general, there may be infinitely many records $\mathbf{x}^*$ in D_0 that satisfy (4.1) and can be reached by the same path of edit operations. To solve the error localization problem, it is sufficient to find an optimal path. Constructing an associated record $\mathbf{x}^* \in D_0$ may then be regarded as a generalization of the consistent imputation problem; cf. the discussion on imputation at the end of Section 3.

Remark 2. The above error localization problem is infeasible for records that cannot be mapped onto D_0 by any combination of distinct edit operations in $\mathcal{G}$. To avoid this situation, $\mathcal{G}$ should be chosen sufficiently large so that $d(\mathbf{x}, D_0) < \infty$ for all $\mathbf{x} \in \mathbb{R}^p$. In what follows, I tacitly assume that $\mathcal{G}$ has this property. An easy way – not necessarily the only way – to achieve this is by letting $\mathcal{G}$ contain at least all FH operations. That this is sufficient follows from the fact that any two points in $\mathbb{R}^p$ are connected by a path that concatenates the FH operations associated with the coordinates on which they differ.

Remark 3. It is not difficult to see that the above error localization problem reduces to the original problem of Fellegi and Holt (1976) in the special case that $\mathcal{G}$ contains only the FH operations.

Remark 4. As with the original Fellegi-Holt-based error localization problem, it can be shown that, under certain assumptions, minimizing $d(\mathbf{x}, \mathbf{y})$ over all $\mathbf{y} \in D_0$ for a given observed record $\mathbf{x}$ is approximately equivalent to maximizing the likelihood of the associated unobserved error-free record. The argument closely follows that of Kruskal (1983, pages 38-39) for the so-called Levenshtein distance in the context of approximate string matching. This requires first of all that the edits (2.1) be hard edits, i.e., failed only by erroneous values. In addition, it must be assumed that the stochastic "error generating process" $\mathcal{E}$ introduced in Section 3 has the following properties:

- There exists a one-to-one correspondence between the set of errors that can occur under $\mathcal{E}$ and the set of allowed edit operations $\mathcal{G}$ that correct them.
- The errors in $\mathcal{E}$ occur independently of each other.
- The error corresponding to operation g occurs with known probability p_g.

Finally, analogous to (2.3), the weights w_g should be chosen according to

$$w_g = -\log\left(\frac{p_g}{1 - p_g}\right). \tag{4.2}$$

Under these assumptions, Scholtus (2014) adapted the argument of Kruskal (1983) to show that the optimal solution to error localization problem (4.1) can be justified as an approximate maximum likelihood

estimator. [Note: The derivation in Scholtus (2014) assumed in addition that all $p_g \ll 1$, in which case $w_g \approx -\log p_g$. This assumption is unnecessary; cf. Liepins (1980).]

5 Implied edits for general edit operations

In this section, a result will be derived that establishes whether a given path of edit operations of the form (3.1) can be used to make a given record consistent with a given system of edit rules (i.e., is a feasible solution to the error localization problem). This result uses the FM elimination technique discussed in Section 2.

Let $\mathbf{x}$ be a given record and let $\mathbf{y}_t$ be any record that can be obtained by applying, in sequence, the edit operations $g_1, \ldots, g_t$ to $\mathbf{x}$:

$$\mathbf{y}_t = g_t \circ g_{t-1} \circ \cdots \circ g_1(\mathbf{x}). \tag{5.1}$$

Write $g_n(\mathbf{x}) = \mathbf{T}_n\mathbf{x} + \mathbf{S}_n\boldsymbol{\alpha}_n + \mathbf{c}_n$, for $n \in \{1, \ldots, t\}$. From (5.1) it follows by induction that

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{T}_1\mathbf{x} + \mathbf{S}_1\boldsymbol{\alpha}_1 + \mathbf{c}_1, \\ \mathbf{y}_2 &= \mathbf{T}_2\mathbf{T}_1\mathbf{x} + \mathbf{S}_2\boldsymbol{\alpha}_2 + \mathbf{c}_2 + \mathbf{T}_2(\mathbf{S}_1\boldsymbol{\alpha}_1 + \mathbf{c}_1), \end{aligned}$$

and, in general,

$$\mathbf{y}_t = \mathbf{T}_t \cdots \mathbf{T}_1\mathbf{x} + \mathbf{S}_t\boldsymbol{\alpha}_t + \mathbf{c}_t + \sum_{n=2}^{t} \mathbf{T}_t \cdots \mathbf{T}_n(\mathbf{S}_{n-1}\boldsymbol{\alpha}_{n-1} + \mathbf{c}_{n-1}), \tag{5.2}$$

where the sum over n is defined to be zero when $t = 1$. Moreover, all terms involving $\mathbf{S}_n\boldsymbol{\alpha}_n$ vanish in these expressions when g_n does not contain any free parameters.

The path of edit operations $P = [g_1, \ldots, g_t]$ can be applied to $\mathbf{x}$ to obtain a record that is consistent with the edits (2.1) if, and only if, there exists a $\mathbf{y}_t$ of the form (5.2) that satisfies $\mathbf{A}\mathbf{y}_t + \mathbf{b} \odot \mathbf{0}$ and all relevant additional restrictions of the form (3.2) on $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t$. Using (5.2), $\mathbf{A}\mathbf{y}_t + \mathbf{b} \odot \mathbf{0}$ can be written as:

$$(\mathbf{A}\mathbf{T}_t \cdots \mathbf{T}_1)\mathbf{x} + (\mathbf{A}\mathbf{S}_t)\boldsymbol{\alpha}_t + \sum_{n=2}^{t} (\mathbf{A}\mathbf{T}_t \cdots \mathbf{T}_n\mathbf{S}_{n-1})\boldsymbol{\alpha}_{n-1} + \mathbf{b}_t \odot \mathbf{0}, \tag{5.3}$$

with $\mathbf{b}_t = \mathbf{b} + \mathbf{A}\mathbf{c}_t + \sum_{n=2}^{t} \mathbf{A}\mathbf{T}_t \cdots \mathbf{T}_n\mathbf{c}_{n-1}$ a vector of constants.

Interestingly, (5.3) and the possible additional restrictions of the form (3.2) constitute a linear system of the form (2.1) on the extended record $(\mathbf{x}', \boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_t')'$. Therefore, FM elimination may be used to remove all free parameters from this system. This yields a system of implied restrictions for $\mathbf{x}$. Moreover, a repeated application of the fundamental property of FM elimination establishes that $\mathbf{x}$ satisfies this system of implied edits if, and only if, there exist parameter values for $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t$ that, together with $\mathbf{x}$, satisfy (5.3) and (3.2). Hence, it follows that the path of edit operations $P = [g_1, \ldots, g_t]$ can lead to a consistent record for $\mathbf{x}$ if, and only if, $\mathbf{x}$ satisfies the system of implied edits obtained by eliminating $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t$ from (5.3) and (if relevant) additional restrictions of the form (3.2).

Example. Consider the following edits in x_1 and x_2:

$$x_1 \geq 0, \tag{5.4}$$

$$x_2 \geq 0, \tag{5.5}$$

$$x_1 + x_2 \leq 5. \tag{5.6}$$

Let g be the edit operation that transfers an amount of at most four units between x_1 and x_2, in either direction: $g((x_1, x_2)') = (x_1 + \alpha, x_2 - \alpha)'$ with $-4 \leq \alpha \leq 4$. For this single edit operation, the system of transformed edits (5.3) is:

$$x_1 + \alpha \geq 0, \tag{5.7}$$

$$x_2 - \alpha \geq 0, \tag{5.8}$$

$$x_1 + x_2 \leq 5. \tag{5.9}$$

I also add the following restrictions of the form (3.2) on α:

$$\alpha \geq -4, \tag{5.10}$$

$$\alpha \leq 4. \tag{5.11}$$

This yields five linear constraints (5.7)-(5.11) on x_1, x_2, and α, from which α may be removed by FM elimination to obtain:

$$x_1 \geq -4, \tag{5.12}$$

$$x_2 \geq -4, \tag{5.13}$$

$$x_1 + x_2 \geq 0, \tag{5.14}$$

$$x_1 + x_2 \leq 5. \tag{5.15}$$

According to the theory, any record $(x_1, x_2)'$ that satisfies (5.12)-(5.15) can be made consistent with the original edits (5.4)-(5.6) by transferring a certain amount $-4 \leq \alpha \leq 4$ between x_1 and x_2. The example record $(x_1, x_2)' = (-2, 3)'$ is inconsistent with the original edit rules (5.4)-(5.6) but satisfies (5.12)-(5.15). This implies that the record can be made consistent with the original edits by applying g. It is easy to see that this is true; any choice $2 \leq \alpha \leq 3$ will do.

It is interesting to note that, for the special case that P consists of the single FH operation that imputes x_j, the transformed system of edits (5.3) is obtained by replacing every occurrence of x_j in the original edits by an unrestricted parameter α. Eliminating α from (5.3) is equivalent in this case to eliminating x_j directly from the original edits. In this sense, the above result generalizes the fundamental property of FM elimination for FH operations to all edit operations of the form (3.1).

In general, the set of records defined by expression (5.2) depends on the way the edit operations are ordered. Thus, two paths consisting of the same set of edit operations in a different order need not yield the same solution to the error localization problem. In this respect, general edit operations differ from FH operations (Scholtus 2014).

6 An error localization algorithm

In this section, I propose a relatively simple algorithm to solve the error localization problem of Section 4, using the theoretical result from the previous section.

Step 0. Let $\mathbf{x}$ be a given record and $\mathcal{G}$ a given set of allowed edit operations. Initialize: $\mathcal{L} := \emptyset$; $\mathcal{B}_0 := \{\emptyset\}$; $W := \infty$; and $t := 1$.

Step 1. Determine all subsets $G \subseteq \mathcal{G}$ of cardinality t that satisfy these conditions:

1. Every subset of $t-1$ elements in G is part of $\mathcal{B}_{t-1}$.
2. It holds that $\sum_{g \in G} w_g \leq W$.

Step 2. For each G found in step 1, construct $\mathcal{P}(\mathbf{x};G)$ and, for each path $P \in \mathcal{P}(\mathbf{x};G)$, evaluate whether it can lead to a consistent record. If so, then:

- if $\ell(P) < W$, define $\mathcal{L} := \{P\}$ and $W := \ell(P)$;
- if $\ell(P) = W$, define $\mathcal{L} := \mathcal{L} \cup \{P\}$.

If *none* of the paths $P \in \mathcal{P}(\mathbf{x};G)$ lead to a consistent record, add G to $\mathcal{B}_t$.

Step 3. If $t < R$ and $\mathcal{B}_t \neq \emptyset$, define $t := t + 1$ and return to step 1.

Figure 6.1 An algorithm that finds all optimal paths of edit operations for problem (4.1).

In practical applications of error localization in official statistics, it is not unusual to have records of over 100 variables. To obtain a problem that is computationally feasible, existing applications of automatic editing based on the Fellegi-Holt paradigm usually specify an upper bound M on the number of variables that may be imputed in a single record (e.g., $M = 12$ or $M = 15$). de Waal and Coutinho (2005) argued that the introduction of such an upper bound is reasonable because a record that requires more than, say, fifteen imputations should be considered unfit for automatic editing anyway. Following this tradition, one can also introduce an upper bound R on the number of distinct edit operations that may be applied to a single record. Even with this additional restriction, the search space of potential solutions to (4.1) will usually be too large in practice to find the optimal solution by an exhaustive search.

Figure 6.1 summarizes the proposed error localization algorithm. Its basic set-up was inspired by the *apriori algorithm* of Agrawal and Srikant (1994) for data mining. Upon completion, the algorithm returns a set $\mathcal{L}$ containing all paths of allowed edit operations that correspond to an optimal solution to (4.1), as well as the optimal path length W. [Note: An error localization problem may have multiple optimal solutions, and it may be beneficial to find all of them (Giles 1988; de Waal et al. 2011, pages 66-67).]

After initialization in step 0, the algorithm cycles through steps 1, 2, and 3 at most R times. In step 1 of the algorithm, the search space is limited by using the following fact: if G has a proper subset $H \subset G$ for which $\mathcal{P}(\mathbf{x};H)$ contains a path that leads to a consistent record, then $\mathcal{P}(\mathbf{x};G)$ can contain only suboptimal solutions. Thus, any set G that has such a subset may be ignored by the algorithm. Similarly, G may also be ignored whenever the total weight of the edit operations in G exceeds the path length of the best feasible solution found so far.

During the t^{th} iteration, the number of subsets G encountered in step 1 of the algorithm equals $\binom{N}{t}$. For each of these subsets, the conditions in step 1 have to be checked. If a subset G passes these checks, in step 2 all $t!$ paths in $\mathcal{P}(\mathbf{x};G)$ are evaluated using the theory of Section 5. The idea behind the apriori

algorithm is that, as t becomes larger, the majority of subsets will not pass the checks in the first step, so that the total amount of computational work remains limited. In the context of data mining, this desirable behavior has indeed been observed in practice. Whether it also occurs in the context of error localization remains to be seen.

One possible improvement to the algorithm can be made by observing that the order in which edit operations are applied does not matter in all cases. Sometimes two paths in $\mathcal{P}(\mathbf{x};G)$ are *equivalent* in the sense that any record that can be reached from $\mathbf{x}$ by the first path can also be reached by the second path, and vice versa. This property defines an equivalence relation on $\mathcal{P}(\mathbf{x};G)$. Let $\tilde{\mathcal{P}}(\mathbf{x};G)$ be a set that contains one representative from each equivalence class of $\mathcal{P}(\mathbf{x};G)$ under this relation. Clearly, the algorithm in Figure 6.1 remains correct if in step 2 the search is limited to $\tilde{\mathcal{P}}(\mathbf{x};G)$ instead of $\mathcal{P}(\mathbf{x};G)$. Scholtus (2014) provides a simple method for constructing $\tilde{\mathcal{P}}(\mathbf{x};G)$ from $\mathcal{P}(\mathbf{x};G)$.

A detailed example illustrating the above algorithm can be found in Scholtus (2014).

7 Simulation study

To test the potential usefulness of the new error localization approach, I conducted a small simulation study, using the R environment for statistical computing (R Development Core Team 2015). A prototype implementation was created in R of the algorithm in Figure 6.1. This prototype made liberal use of the existing functionality for Fellegi-Holt-based automatic editing available in the `editrules` package (van der Loo and de Jonge 2012; de Jonge and van der Loo 2014). The program was not optimized for computational efficiency, but it turned out to work sufficiently fast for the relatively small error localization problems encountered in this simulation study. (Note: The R code used in this study is available from the author upon request.)

The simulation study involved records of five numerical variables that should satisfy the following nine linear edit rules:

$$\begin{aligned} x_1 + x_2 &= x_3, \\ x_3 - x_4 &= x_5, \\ x_j &\geq 0, \qquad j \in \{1,2,3,4\}, \\ x_1 &\geq x_2, \\ x_5 &\geq -0.1x_3, \\ x_5 &\leq 0.5x_3. \end{aligned}$$

Edits of this form might typically be encountered for SBS, as part of a much larger set of edit rules (Scholtus 2014).

I created a random error-free data set of 2,000 records by drawing from a multivariate normal distribution (using the `mvtnorm` package) with the following parameters:

$$\boldsymbol{\mu} = \begin{pmatrix} 500 \\ 250 \\ 750 \\ 600 \\ 150 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 10{,}000 & -1{,}250 & 8{,}750 & 7{,}500 & 1{,}250 \\ -1{,}250 & 5{,}000 & 3{,}750 & 4{,}000 & -250 \\ 8{,}750 & 3{,}750 & 12{,}500 & 11{,}500 & 1{,}000 \\ 7{,}500 & 4{,}000 & 11{,}500 & 11{,}750 & -250 \\ 1{,}250 & -250 & 1{,}000 & -250 & 1{,}250 \end{pmatrix}.$$

Only records that satisfied all of the above edits were added to the data set. Note that $\boldsymbol{\Sigma}$ is a singular covariance matrix that incorporates the two equality edits. Technically, the resulting data follow a so-called truncated multivariate singular normal distribution; see de Waal et al. (2011, pages 318ff) or Tempelman (2007).

Table 7.1 lists the nine allowed edit operations that were considered in this study. Note that the first five lines contain the FH operations for this data set. As indicated in the table, each edit operation has an associated type of error. A synthetic data set to be edited was created by randomly adding errors of these types to the above-mentioned error-free data set. The probability of each type of error is listed in the fourth column of Table 7.1. The associated "ideal" weight according to (4.2) is shown in the last column.

To limit the amount of computational work, I only considered records that required three edit operations or less. Records without errors were also removed. This left 1,025 records to be edited, each containing one, two, or three of the errors listed in Table 7.1.

Table 7.1
Allowed edit operations for the simulation study

name	operation	associated type of error	p_g	w_g
FH1	impute x_1	erroneous value of x_1	0.10	2.20
FH2	impute x_2	erroneous value of x_2	0.08	2.44
FH3	impute x_3	erroneous value of x_3	0.06	2.75
FH4	impute x_4	erroneous value of x_4	0.04	3.18
FH5	impute x_5	erroneous value of x_5	0.02	3.89
IC34	interchange x_3 and x_4	true values of x_3 and x_4 interchanged	0.07	2.59
TF21	transfer an amount from x_2 to x_1	part of the true value of x_1 reported as part of x_2	0.09	2.31
CS4	change the sign of x_4	sign error in x_4	0.11	2.09
CS5	change the sign of x_5	sign error in x_5	0.13	1.90

Several error localization approaches were applied to this data set. First of all, I tested error localization according to the Fellegi-Holt paradigm (i.e., using only the edit operations FH1–FH5) and according to the new paradigm (i.e., using all edit operations in Table 7.1). Both approaches were tested once using the "ideal" weights listed in Table 7.1 and once with all weights equal to 1 ("no weights"). The latter case simulates a situation where the relevant edit operations would be known, but not their respective frequencies. Finally, to test the robustness of the new error localization approach to a lack of information about relevant edit operations, I also applied this approach with one of the non-FH operations in Table 7.1 missing from the set of allowed edit operations.

The quality of error localization was evaluated in two ways. Firstly, I evaluated how well the optimal paths of edit operations found by the algorithm matched the true distribution of errors, using the following contingency table for all $1{,}025 \times 9 = 9{,}225$ combinations of records and edit operations:

Table 7.2
Contingency table of errors and edit operations suggested by the algorithm

	edit operation was suggested	edit operation was not suggested
associated error occurred	*TP*	*FN*
associated error did not occur	*FP*	*TN*

From this table, I computed indicators that measure the proportion of false negatives, false positives, and overall wrong decisions, respectively:

$$\alpha = \frac{FN}{TP + FN}; \quad \beta = \frac{FP}{FP + TN}; \quad \delta = \frac{FN + FP}{TP + FN + FP + TN}.$$

Similar indicators are discussed by de Waal et al. (2011, pages 410-411). I also computed $\bar{\rho} = 1 - \rho$, with ρ the fraction of records in the data set for which the error localization algorithm found exactly the right solution. A good error localization algorithm should have low scores on all four indicators.

It should be noted that the above quality indicators put the original Fellegi-Holt approach at a disadvantage, as this approach does not use all the edit operations listed in Table 7.1. Therefore, I also calculated a second set of quality indicators $\alpha, \beta, \delta,$ and $\bar{\rho}$ that look at erroneous values rather than edit operations. In this case, α measures the proportion of values in the data set that were affected by errors but left unchanged by the optimal solution of the error localization problem, and similarly for the other measures.

Table 7.3 displays the results of the simulation study for both sets of quality indicators. In both cases, a considerable improvement in the quality of the error localization results is seen for the approach that used all edit operations, compared to the approach that used only FH operations. In addition, leaving one relevant edit operation out of the set of allowed edit operations had a negative effect on the quality of error localization. In some cases this effect was quite large – particularly in terms of edit operations used –, but the results of the new error localization approach still remained substantially better than those of the Fellegi-Holt approach. Contrary to expectation, not using different confidence weights actually improved the quality of the error localization results somewhat for this data set under the Fellegi-Holt approach (both sets of indicators) and to some extent also under the new approach (only the second set of indicators). Finally, it is seen that using all edit operations led to an increase in computing time compared to using only FH operations, but this increase was not dramatic.

Table 7.3
Quality of error localization in terms of edit operations used and identified erroneous values; computing time required

	quality indicators (edit operations)				quality indicators (erroneous values)				
approach	α	β	δ	$\bar{\rho}$	α	β	δ	$\bar{\rho}$	**time***
Fellegi-Holt (weights)	74%	12%	23%	80%	19%	10%	13%	32%	46
Fellegi-Holt (no weights)	70%	12%	21%	74%	13%	8%	9%	24%	33
all operations (weights)	14%	3%	5%	24%	10%	5%	7%	17%	98
except IC34	29%	5%	9%	35%	15%	9%	11%	29%	113
except TF21	34%	5%	10%	37%	10%	5%	7%	18%	80
except CS4	28%	6%	9%	39%	10%	5%	7%	17%	80
except CS5	35%	7%	10%	47%	11%	6%	7%	18%	82
all operations (no weights)	27%	5%	8%	36%	6%	4%	5%	13%	99

* Total computing time (in seconds) on a laptop PC with a 2.5 GHz CPU under Windows 7.

8 Conclusion

In this article, a new formulation was proposed of the error localization problem in automatic editing. It was suggested to find the (weighted) minimal number of edit operations needed to make an observed record consistent with the edits. The new error localization problem can be seen as a generalization of the problem proposed in a seminal paper by Fellegi and Holt (1976), because the operation that imputes a new value for one variable at a time is an important special case of an edit operation.

The main focus here has been on developing the mathematical theory behind the new error localization problem. It turns out that FM elimination, a technique that has been used in the past to solve the Fellegi-Holt-based error localization problem, can be applied also in the context of the new problem (Section 5). Nevertheless, the task of solving the new error localization problem is challenging from a computational point of view, at least for the numbers of variables, edits, and edit operations that would be encountered in practical applications at statistical institutes. A possible error localization algorithm was outlined in Section 6. More efficient algorithms probably could and should be developed. Similarly to FM elimination, it may be possible to adapt other ideas that have been used to solve the Fellegi-Holt-based problem to the generalized problem considered here.

The discussion in this article was restricted to numerical data and linear edits. The original Fellegi-Holt paradigm has been applied also to categorical and mixed data. Several authors, including Bruni (2004) and de Jonge and van der Loo (2014), have shown that a large class of edits for mixed data can be re-formulated in terms of numerical data and linear edits, with the additional restriction that some of the variables have to be integer-valued. In principle, this means that the results in this article could be applied also to mixed data. To accommodate the fact that some variables are integer-valued, Pugh's (1992) extension of FM elimination to integers could be used; see also de Waal et al. (2011) for a discussion of this extended elimination technique in the context of Fellegi-Holt-based error localization. It remains to be seen whether this approach is computationally feasible.

Remark 4 in Section 4 hinted at an analogy between error localization in statistical microdata and the field of approximate string matching. In approximate string matching, text strings are compared under the assumption that they may have been partially corrupted (Navarro 2001). Various distance functions have been proposed for this task. The Hamming distance, which counts the number of positions on which two strings differ, may be seen as an analogue of the Fellegi-Holt-based target function (2.2). The generalized error localization problem defined in this paper has its counterpart in the use of the Levenshtein distance or "edit distance" for approximate string matching. It may be interesting to explore this analogy further. In particular, efficient algorithms have been developed for computing edit distances between strings; it might be possible to apply some of the underlying ideas also to the generalized error localization problem.

The new error localization algorithm was applied successfully to a small synthetic data set (Section 7). Overall, the results of this simulation study suggest that the new error localization approach has the potential to achieve a substantial improvement of the quality of automatic editing compared to the approach that is currently used in practice. However, this does require that sufficient information be available to identify all – or at least most – of the relevant edit operations in a particular application. Possible gains in the quality of error localization also have to be weighed in practice against the higher computational demands of the generalized error localization problem.

An obvious candidate for applying the new methodology in practice would be the SBS. However, more research is needed before this method could be applied during regular production. To apply the method in a particular context, it is necessary first to specify the relevant edit operations. Ideally, each edit operation should correspond to a combination of amendments to the data that human editors consider to be a correction for one particular error. In addition, a suitable set of weights w_g has to be determined for these edit operations. This would require information about the relative frequencies of the most common types of amendments made during manual editing. Both aspects could be investigated based on historical data before and after manual editing, editing instructions and other documentation used by the editors, and interviews with editors and/or supervisors of editing.

On a more fundamental level, a question of demarcation arises between deductive correction methods and automatic editing under the new error localization problem. In principle, many known types of error could be resolved either by automatic correction rules or by error localization using edit operations. Each approach has its own advantages and disadvantages (Scholtus 2014). It is likely that some compromise will produce the best results, with some errors handled deductively and others by edit operations. However, it is not obvious how best to make this division in practice.

Ultimately, the aim of the new methodology proposed in this article is to improve the usefulness of automatic editing in practice. So far, the results are promising.

Acknowledgements

The views expressed in this article are those of the author and do not necessarily reflect the policies of *Statistics Netherlands*. The author would like to thank Jeroen Pannekoek, Ton de Waal, and Mark van der Loo for their comments on earlier versions of this article, as well as the Associate Editor and two anonymous referees.

Appendix

Fourier-Motzkin elimination

Consider a system of linear constraints (2.1) and let x_f be the variable to be eliminated. First, suppose that x_f is involved only in inequalities. For ease of exposition, suppose that the edits are normalized so that all inequalities use the $\geq$ operator. The FM elimination method considers all pairs (r,s) of inequalities in which the coefficients of x_f have opposite signs; that is, $a_{rf}a_{sf} < 0$. Suppose without loss of generality that $a_{rf} < 0$ and $a_{sf} > 0$. From the original pair of edits, the following implied constraint is derived:

$$\sum_{j=1}^{p} a_j^* x_j + b^* \geq 0, \tag{A.1}$$

with $a_j^* = a_{sf}a_{rj} - a_{rf}a_{sj}$ and $b^* = a_{sf}b_r - a_{rf}b_s$. Note that $a_f^* = 0$, so x_f is not involved in (A.1). An inequality of the form (A.1) is derived from each of the above-mentioned pairs (r,s). The full implied system of constraints obtained by FM elimination now consists of these derived constraints, together with all original constraints that do not involve x_f.

If there are linear equalities that involve x_f, the above technique could be applied after replacing each linear equality with two equivalent linear inequalities. de Waal and Quere (2003) suggested a more efficient

alternative for this case. Suppose that the r^{th} constraint in (2.1) is an equality that involves x_f. This constraint can be rewritten as

$$x_f = \frac{-1}{a_{rf}}\left(b_r + \sum_{j \neq f} a_{rj} x_j\right). \tag{A.2}$$

By substituting the expression on the right-hand-side of (A.2) for x_f in all other constraints, one again obtains an implied system of constraints that does not involve x_f and that can be rewritten in the form (2.1).

For a proof that FM elimination has the fundamental property mentioned in Section 2, see, e.g., de Waal et al. (2011, pages 69-70).

References

Agrawal, R., and Srikant, R. (1994). *Fast Algorithms for Mining Association Rules*. Technical report, IBM Almaden Research Center, San Jose, California.

Bruni, R. (2004). Discrete models for data imputation. *Discrete Applied Mathematics*, 144, 59-69.

Chen, B., Thibaudeau, Y. and Winkler, W.E. (2003). *A Comparison Study of ACS If-Then-Else, NIM, DISCRETE Edit and Imputation Systems Using ACS Data*. Working Paper No. 7, UN/ECE Work Session on Statistical Data Editing, Madrid.

de Jonge, E., and van der Loo, M. (2014). *Error Localization as a Mixed Integer Problem with the Editrules Package*. Discussion Paper 2014-07, Statistics Netherlands, The Hague. Available at: http://www.cbs.nl.

de Waal, T. (2003). Solving the error localization problem by means of vertex generation. *Survey Methodology*, 29, 1, 71-79.

de Waal, T., and Coutinho, W. (2005). Automatic editing for business surveys: An assessment for selected algorithms. *International Statistical Review*, 73, 73-102.

de Waal, T., and Quere, R. (2003). A fast and simple algorithm for automatic editing of mixed data. *Journal of Official Statistics*, 19, 383-402.

de Waal, T., Pannekoek, J. and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken, New Jersey: John Wiley & Sons, Inc.

EDIMBUS (2007). *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Eurostat manual prepared by ISTAT, Statistics Netherlands, and SFSO.

Fellegi, I.P., and Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

Garfinkel, R.S., Kunnathur, A.S. and Liepins, G.E. (1988). Error localization for erroneous data: Continuous data, linear constraints. *SIAM Journal on Scientific and Statistical Computing*, 9, 922-931.

Ghosh-Dastidar, B., and Schafer, J.L. (2006). Outlier detection and editing procedures for continuous multivariate data. *Journal of Official Statistics*, 22, 487-506.

Giles, P. (1988). A model for generalized edit and imputation of survey data. *The Canadian Journal of Statistics*, 16, 57-73.

Granquist, L. (1995). Improving the traditional editing process. In *Business Survey Methods*, (Eds., B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott), John Wiley & Sons, Inc., 385-401.

Granquist, L. (1997). The new view on editing. *International Statistical Review*, 65, 381-387.

Granquist, L., and Kovar, J. (1997). Editing of survey data: How much is enough? In *Survey Measurement and Process Quality*, (Eds., L.E. Lyberg, P. Biemer, M. Collins, E.D. de Leeuw, C. Dippo, N. Schwartz and D. Trewin), John Wiley & Sons, Inc., 415-435.

Hedlin, D. (2003). Score functions to reduce business survey editing at the U.K. Office for National Statistics. *Journal of Official Statistics*, 19, 177-199.

Hidiroglou, M.A., and Berthelot, J.-M. (1986). Statistical editing and imputation for periodic business surveys. *Survey Methodology*, 12, 1, 73-83.

Kovar, J., and Whitridge, P. (1990). Generalized edit and imputation system; Overview and applications. *Revista Brasileira de Estadistica*, 51, 85-100.

Kruskal, J.B. (1983). An overview of sequence comparison. In *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, (Eds., D. Sankoff and J.B. Kruskal), Addison-Wesley, 1-44.

Lawrence, D., and McKenzie, R. (2000). The general application of significance editing. *Journal of Official Statistics*, 16, 243-253.

Liepins, G.E. (1980). *A Rigorous, Systematic Approach to Automatic Data Editing and its Statistical Basis*. Report ORNL/TM-7126, Oak Ridge National Laboratory.

Liepins, G.E., Garfinkel, R.S. and Kunnathur, A.S. (1982). Error localization for erroneous data: A survey. *TIMS/Studies in the Management Sciences*, 19, 205-219.

Little, R.J.A., and Smith, P.J. (1987). Editing and imputation of quantitative survey data. *Journal of the American Statistical Association*, 82, 58-68.

Naus, J.I., Johnson, T.G. and Montalvo, R. (1972). A probabilistic model for identifying errors in data editing. *Journal of the American Statistical Association*, 67, 943-950.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, 33, 31-88.

Pannekoek, J., Scholtus, S. and van der Loo, M. (2013). Automated and manual data editing: A view on process design and methodology. *Journal of Official Statistics*, 29, 511-537.

Pugh, W. (1992). The omega test: A fast and practical integer programming algorithm for data dependence analysis. *Communications of the ACM*, 35, 102-114.

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: http://www.R-project.org/.

Ragsdale, C.T., and McKeown, P.G. (1996). On solving the continuous data editing problem. *Computers & Operations Research*, 23, 263-273.

Riera-Ledesma, J., and Salazar-González, J.J. (2003). *New Algorithms for the Editing and Imputation Problem*. Working Paper No. 5, UN/ECE Work Session on Statistical Data Editing, Madrid.

Riera-Ledesma, J., and Salazar-González, J.J. (2007). A branch-and-cut algorithm for the continuous error localization problem in data cleaning. *Computers & Operations Research*, 34, 2790-2804.

Schaffer, J. (1987). Procedure for solving the data-editing problem with both continuous and discrete data types. *Naval Research Logistics*, 34, 879-890.

Scholtus, S. (2011). Algorithms for correcting sign errors and rounding errors in business survey data. *Journal of Official Statistics*, 27, 467-490.

Scholtus, S. (2014). *Error Localisation using General Edit Operations*. Discussion Paper 2014-14, Statistics Netherlands, The Hague. Available at: http://www.cbs.nl.

Tempelman, D.C.G. (2007). *Imputation of Restricted Data*. Ph. D. Thesis, University of Groningen. Available at: http://www.cbs.nl.

van der Loo, M., and de Jonge, E. (2012). *Automatic Data Editing with Open Source R*. Working Paper No. 33, UN/ECE Work Session on Statistical Data Editing, Oslo.

Williams, H.P. (1986). Fourier's method of linear programming and its dual. *The American Mathematical Monthly*, 93, 681-695.

Survey Methodology, June 2016
Vol. 42, No. 1, pp. 19-40
Statistics Canada, Catalogue No. 12-001-X

Statistical matching using fractional imputation

Jae Kwang Kim, Emily Berg and Taesung Park[1]

Abstract

Statistical matching is a technique for integrating two or more data sets when information available for matching records for individual participants across data sets is incomplete. Statistical matching can be viewed as a missing data problem where a researcher wants to perform a joint analysis of variables that are never jointly observed. A conditional independence assumption is often used to create imputed data for statistical matching. We consider a general approach to statistical matching using parametric fractional imputation of Kim (2011) to create imputed data under the assumption that the specified model is fully identified. The proposed method does not have a convergent expectation-maximisation (EM) sequence if the model is not identified. We also present variance estimators appropriate for the imputation procedure. We explain how the method applies directly to the analysis of data from split questionnaire designs and measurement error models.

Key Words: Data combination; Data fusion; Hot deck imputation; Split questionnaire design; Measurement error model.

1 Introduction

Survey sampling is a scientific tool for making inference about the target population. However, we often do not collect all the necessary information in a single survey, due to time and cost constraints. In this case, we wish to exploit, as much as possible, information already available from different data sources from the same target population. Statistical matching, sometimes called data fusion (Baker, Harris and O'Brien 1989) or data combination (Ridder and Moffit 2007), aims to integrate two or more data sets when information available for matching records for individual participants across data sets is incomplete. D'Orazio, Zio and Scanu (2006) and Leulescu and Agafitei (2013) provide comprehensive overviews of the statistical matching techniques in survey sampling.

Statistical matching can be viewed as a missing data problem where a researcher wants to perform a joint analysis of variables that are never jointly observed. Moriarity and Scheuren (2001) provide a theoretical framework for statistical matching under a multivariate normality assumption. Rässler (2002) develops multiple imputation techniques for statistical matching with pre-specified parameter values for non-identifiable parameters. Lahiri and Larsen (2005) address regression analysis with linked data. Ridder and Moffit (2007) provide a rigorous treatment of the assumptions and approaches for statistical matching in the context of econometrics.

Statistical matching aims to construct fully augmented data files to perform statistically valid joint analyses. To simplify the setup, suppose that two surveys, Survey A and Survey B, contain partial information about the population. Suppose that we observe x and y_1 from the Survey A sample and observe x and y_2 from the Survey B sample. Table 1.1 illustrates a simple data structure for matching. If the Survey B sample (Sample B) is a subset of the Survey A sample (Sample A), then we can apply record linkage techniques (Herzog, Scheuren and Winkler 2007) to obtain values of y_1 for the survey B sample. However, in many cases, such perfect matching is not possible (for instance, because the samples may contain

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A. E-mail: jkim@iastate.edu; Emily Berg, Department of Statistics, Iowa State University, Ames, Iowa, U.S.A. E-mail: emilyb@iastate.edu; Taesung Park, Department of Statistics, Seoul National University, Seoul, Korea. E-mail: taesungp@gmail.com.

non-overlapping subsets), and we may rely on a probabilistic way of identifying the "statistical twins" from the other sample. That is, we want to create y_1 for each element in sample B by finding the nearest neighbor from Sample A. Nearest neighbor imputation has been discussed by many authors, including Chen and Shao (2001) and Beaumont and Bocci (2009), in the context of missing survey items.

Table 1.1
A simple data structure for matching

	$\boldsymbol{X}$	$\boldsymbol{Y}_1$	$\boldsymbol{Y}_2$
Sample A	o	o	
Sample B	o		o

Finding the nearest neighbor is often based on "how close" they are in terms of x's only. Thus, in many cases, statistical matching is based on the assumption that y_1 and y_2 are independent, conditional on x. That is,

$$y_1 \perp y_2 \,|\, x. \tag{1.1}$$

Assumption (1.1) is often referred to as the conditional independence (CI) assumption and is heavily used in practice.

In this paper, we consider an alternative approach that does not rely on the CI assumption. After we discuss the assumptions in Section 2, we present the proposed methods in Section 3. Furthermore, we consider two extensions, one to split questionnaire designs (in Section 4) and the other to measurement error models (in Section 5). Results from two simulation studies are presented in Section 6. Section 7 concludes the paper.

2 Basic setup

For simplicity of the presentation, we consider the setup of two independent surveys from the same target population consisting of N elements. As discussed in Section 1, suppose that Sample A collects information only on x and y_1 and Sample B collects information only on x and y_2.

To illustrate the idea, suppose for now that (x, y_1, y_2) are generated from a normal distribution such that

$$\begin{pmatrix} x \\ y_1 \\ y_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_x \\ \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{1x} & \sigma_{2x} \\ & \sigma_{11} & \sigma_{12} \\ & & \sigma_{22} \end{pmatrix}\right].$$

Clearly, under the data structure in Table 1.1, the parameter σ_{12} is not estimable from the samples. The conditional independence assumption in (1.1) implies that $\sigma_{12} = \sigma_{1x}\sigma_{2x}/\sigma_{xx}$ and $\rho_{12} = \rho_{1x}\rho_{2x}$. That is, σ_{12} is completely determined from other parameters, rather than estimated directly from the realized samples.

Synthetic data imputation under the conditional independence assumption in this case can be implemented in two steps:

[Step 1] Estimate $f(y_1|x)$ from Sample A, and denote the estimate by $\hat{f}_a(y_1|x)$.

[Step 2] For each element i in Sample B, use the x_i value to generate imputed value(s) of y_1 from $\hat{f}_a(y_1|x_i)$.

Since y_1 values are never observed in Sample B, synthetic values of y_1 are created for all elements in Sample B, leading to synthetic imputation. Haziza (2009) provides a nice review of literature on imputation methodology. Kim and Rao (2012) present a model-assisted approach to synthetic imputation when only x is available in Sample B. Such synthetic imputation completely ignores the observed information in y_2 from Sample B.

Statistical matching based on conditional independence assumes that $\text{Cov}(y_1, y_2|x) = 0$. Thus, the regression of y_2 on x and y_1 using the imputed data from the above synthetic imputation will estimate a zero regression coefficient for y_1. That is, the estimate $\hat{\beta}_2$ for

$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 y_1,$$

will estimate zero. Such analyses can be misleading if CI does not hold. To explain why, we consider an omitted variable regression problem:

$$\begin{aligned} y_1 &= \beta_0^{(1)} + \beta_1^{(1)} x + \beta_2^{(1)} z + e_1 \\ y_2 &= \beta_0^{(2)} + \beta_1^{(2)} x + \beta_2^{(2)} z + e_2 \end{aligned}$$

where z, e_1, e_2 are independent and are not observed. Unless $\beta_2^{(1)} = \beta_2^{(2)} = 0$, the latent variable z is an unobservable confounding factor that explains why $\text{Cov}(y_1, y_2|x) \neq 0$. Thus, the coefficient on y_1 in the population regression of y_2 on x and y_1 is not zero.

Note that the CI assumption is an assumption for model identification. Another identifying assumption is the instrumental variable (IV) assumption, as described in the following remark.

Remark 2.1 *We present a formal description of the IV assumption. First, assume that we can decompose* x *as* $x = (x_1, x_2)$ *such that*

(*i*) $f(y_2|x_1, x_2, y_1) = f(y_2|x_1, y_1)$

(*ii*) $f(y_1|x_2, x_1 = a) \neq f(y_1|x_2, x_1 = b)$

for some $a \neq b$. *Thus,* x_1 *is conditionally independent of* y_2 *given* x_2 *and* y_1 *but* x_1 *is correlated with* y_1 *given* x_2. *Note that* x_2 *may be null or have a degenerate distribution, such as an intercept. The variable* x_1 *satisfying the above two conditions is often called an instrumental variable (IV) for* y_1. *The directed acyclic graph in Figure 2.1 illustrates the dependence structure of a model with an instrumental variable. Ridder and Moffit (2007) used "exclusion restrictions" to describe the instrumental variable assumption. One example where the instrumental variable assumption is reasonable is repeated surveys. In the repeated survey, suppose that* y_t *is the study variable at year* t *and satisfies Markov property*

$$P(y_{t+1} \mid y_1, \ldots, y_t) = P(y_{t+1} \mid y_t),$$

where $P(y_t)$ *denotes a cumulative distribution function. In this case,* y_{t-1} *is an instrumental variable for* y_t*. In fact, any last observation of* $y_s\ (s \le t)$ *is the instrumental variable for* y_t*.*

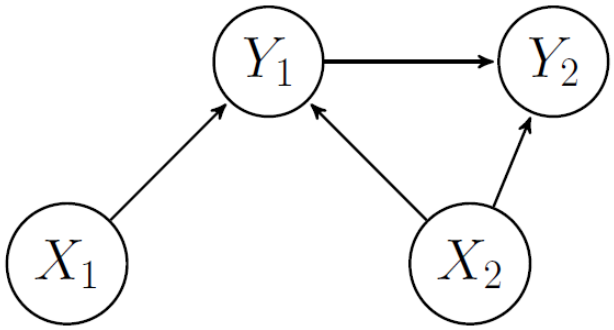

Figure 2.1 Graphical illustration of the dependence structure for a model in which x_1 is an instrumental variable for y_1 and x_2 is an additional covariate in the models for y_2 and y_1.

Under the instrumental variable assumption, one can use two-step regression to estimate the regression parameters of a linear model. The following example presents the basic ideas.

Example 2.1 *Consider the two sample data structure in Table 1.1. We assume the following linear regression model:*

$$y_{2i} = \beta_0 + \beta_1 y_{1i} + \beta_2 x_{2i} + e_i, \tag{2.1}$$

where $e_i \sim (0, \sigma_e^2)$ *and* e_i *is independent of* (x_{1j}, x_{2j}, y_{1j}) *for all* i, j*. In this case, a consistent estimator of* $\beta = (\beta_0, \beta_1, \beta_2)'$ *can be obtained by the two-stage least squares (2SLS) method as follows:*

1. *From Sample A, fit the following "working model" for* y_1

$$y_{1i} = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + u_i, \quad u_i \sim (0, \sigma_u^2) \tag{2.2}$$

to obtain a consistent estimator of $\alpha = (\alpha_0, \alpha_1, \alpha_2)'$ *defined by*

$$\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)' = (X'X)^{-1} X'Y_1$$

where $X = [X_0, X_1, X_2]$ *is a matrix whose* i^{th} *row is* $(1, x_{1i}, x_{2i})$ *and* Y_1 *is a vector with* y_{1i} *being the* i^{th} *component.*

2. *A consistent estimator of* $\beta = (\beta_0, \beta_1, \beta_2)'$ *is obtained by the least squares method for the regression of* y_{2i} *on* $(1, \hat{y}_{1i}, x_{2i})$ *where* $\hat{y}_{1i} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i}$*.*

Asymptotic unbiasedness of the 2SLS estimator under the instrumental variable assumption is discussed in Appendix A. The 2SLS method is not directly applicable if the regression model (2.1) is nonlinear. Also, while the 2SLS method gives estimates of the regression parameters, 2SLS does not provide consistent estimators for more general parameters such as $\theta = \Pr(y_2 < 1 | y_1 < 3)$. Stochastic imputation can provide a solution for estimating a more general class of parameters. We explain how to modify parametric fractional imputation of Kim (2011) to address general purpose estimation in statistical matching problems.

3 Fractional imputation

We now describe the fractional imputation methods for statistical matching without using the CI assumption. The use of fractional imputation for statistical matching was originally presented in Chapter 9 of Kim and Shao (2013) under the IV assumption. In this paper, we present the methodology without requiring the IV assumption. We only assume that the specified model is fully identified. The identifiability of the specified model can be easily checked in the computation of the proposed procedure.

To explain the idea, note that y_1 is missing in Sample B and our goal is to generate y_1 from the conditional distribution of y_1 given the observations. That is, we wish to generate y_1 from

$$f(y_1 | x, y_2) \propto f(y_2 | x, y_1) f(y_1 | x). \tag{3.1}$$

To generate y_1 from (3.1), we can consider the following two-step imputation:

1. Generate y_1^* from $\hat{f}_a(y_1 | x)$.
2. Accept y_1^* if $f(y_2 | x, y_1^*)$ is sufficiently large.

Note that the first step is the usual method under the CI assumption. The second step incorporates the information in y_2. The determination of whether $f(y_2 | x, y_1^*)$ is sufficiently large required for Step 2 is often made by applying a Markov Chain Monte Carlo (MCMC) method such as the Metropolis-Hastings algorithm (Chib and Greenberg 1995). That is, let $y_1^{(t-1)}$ be the current value of y_1 in the Markov Chain. Then, we accept y_1^* with probability

$$R(y_1^*, y_1^{(t-1)}) = \min\left\{1, \frac{f(y_2 | x, y_1^*)}{f(y_2 | x, y_1^{(t-1)})}\right\}.$$

Such algorithms can be computationally cumbersome because of slow convergence of the MCMC algorithm.

Parametric fractional imputation of Kim (2011) enables generating imputed values in (3.1) without requiring MCMC. The following EM algorithm by fractional imputation can be used:

1. For each $i \in B$, generate m imputed values of y_{1i}, denoted by $y_{1i}^{*(1)}, \ldots, y_{1i}^{*(m)}$, from $\hat{f}_a(y_1 | x_i)$, where $\hat{f}_a(y_1 | x)$ denotes the estimated density for the conditional distribution of y_1 given x obtained from Sample A.

2. Let $\hat{\theta}_t$ be the current parameter value of θ in $f(y_2 | x, y_1)$. For the j^{th} imputed value $y_{1i}^{*(j)}$, assign the fractional weight

$$w_{ij(t)}^{*} \propto f\left(y_{2i} | x_i, y_{1i}^{*(j)}; \hat{\theta}_t\right)$$

such that $\sum_{j=1}^{m} w_{ij}^{*} = 1$.

3. Solve the fractionally imputed score equation for θ

$$\sum_{i \in B} w_{ib} \sum_{j=1}^{m} w_{ij(t)}^{*} S\left(\theta; x_i, y_{1i}^{*(j)}, y_{2i}\right) = 0 \tag{3.2}$$

to obtain $\hat{\theta}_{t+1}$, where $S(\theta; x, y_1, y_2) = \partial \log f(y_2 | x, y_1; \theta) / \partial \theta$, and w_{ib} is the sampling weight of unit i in Sample B.

4. Go to Step 2 and continue until convergence.

When the model is identified, the EM sequence obtained from the above PFI method will converge. If the specified model is not identifiable then there is no unique solution to maximizing the observed likelihood and the above EM sequence does not converge. In (3.2), note that, for sufficiently large m,

$$\begin{aligned}\sum_{j=1}^{m} w_{ij(t)}^{*} S\left(\theta; x_i, y_{1i}^{*(j)}, y_{2i}\right) &\cong \frac{\int S(\theta; x_i, y_1, y_{2i}) f\left(y_{2i} | x_i, y_{1i}^{*(j)}; \hat{\theta}_t\right) \hat{f}_a(y_1 | x_i) dy_1}{\int f\left(y_{2i} | x_i, y_{1i}^{*(j)}; \hat{\theta}_t\right) \hat{f}_a(y_1 | x_i) dy_1} \\ &= E\left\{S(\theta; x_i, Y_1, y_{2i}) | x_i, y_{2i}; \hat{\theta}_t\right\}.\end{aligned}$$

If y_{i1} is categorical, then the fractional weight can be constructed by the conditional probability corresponding to the realized imputed value (Ibrahim 1990). Step 2 is used to incorporate observed information of y_{i2} in Sample B. Note that Step 1 is not repeated for each iteration. Only Step 2 and Step 3 are iterated until convergence. Because Step 1 is not iterated, convergence is guaranteed and the observed likelihood increases, as long as the model is identifiable. See Theorem 2 of Kim (2011).

Remark 3.1 *In Section 2, we introduce IV only because this is what it is typically done in the literature to ensure identifiability. The proposed method itself does not rely on this assumption. To illustrate a situation where we can identify the model without introducing the IV assumption, suppose that the model is*

$$\begin{aligned} y_2 &= \beta_0 + \beta_1 x + \beta_2 y_1 + e_2 \\ y_1 &= \alpha_0 + \alpha_1 x + e_1 \end{aligned}$$

with $e_1 \sim N\left(0, x^2 \sigma_1^2\right)$ *and* $e_2 | e_1 \sim N\left(0, \sigma_2^2\right)$. *Then*

$$f(y_2 | x) = \int f(y_2 | x, y_1) f(y_1 | x) dy_1$$

is also a normal distribution with mean $(\beta_0 + \beta_2 \alpha_0) + (\beta_1 + \beta_2 \alpha_1) x$ *and variance* $\sigma_2^2 + \beta_2^2 \sigma_1^2 x^2$. *Under the data structure in Table 1.1, such a model is identified without assuming the IV assumption. The*

assumption of no interaction between y_1 and x in the model for y_2 is key to ensuring the model is identifiable.

Instead of generating $y_{1i}^{*(j)}$ from $\hat{f}_a(y_1|x_i)$, we can consider a hot-deck fractional imputation (HDFI) method, where all the observed values of y_{1i} in Sample A are used as imputed values. In this case, the fractional weights in Step 2 are given by

$$w_{ij}^*(\hat{\theta}_t) \propto w_{ij0}^* f(y_{2i}|x_i, y_{1i}^{*(j)}; \hat{\theta}_t),$$

where

$$w_{ij0}^* = \frac{\hat{f}_a(y_{1j}|x_i)}{\sum_{k\in A} w_{ka}\hat{f}_a(y_{1j}|x_k)}. \tag{3.3}$$

The initial fractional weight w_{ij0}^* in (3.3) is computed by applying importance weighting with

$$\hat{f}_a(y_{1j}) = \int \hat{f}_a(y_{1j}|x)\,\hat{f}_a(x)\,dx \propto \sum_{i\in A} w_{ia}\hat{f}_a(y_{1j}|x_i)$$

as the proposal density for y_{1j}. The M-step is the same as for parametric fractional imputation. See Kim and Yang (2014) for more details on HDFI. In practice, we may use a single imputed value for each unit. In this case, the fractional weights can be used as the selection probability in Probability-Proportional-to-Size (PPS) sampling of size $m = 1$.

For variance estimation, we can either use a linearization method or a resampling method. We first consider variance estimation for the maximum likelihood estimator (MLE) of θ. If we use a parametric model $f(y_1|x) = f(y_1|x;\theta_1)$ and $f(y_2|x,y_1;\theta_2)$, the MLE of $\theta = (\theta_1,\theta_2)$ is obtained by solving

$$[S_1(\theta_1), \bar{S}_2(\theta_1,\theta_2)] = (0,0), \tag{3.4}$$

where $S_1(\theta_1) = \sum_{i\in A} w_{ia}S_{i1}(\theta_1)$, $S_{i1}(\theta_1) = \partial \log f(y_{1i}|x_i;\theta_1)/\partial\theta_1$ is the score function of θ_1,

$$\bar{S}_2(\theta_1,\theta_2) = E\{S_2(\theta_2)|X, Y_2; \theta_1,\theta_2\},$$

$S_2(\theta_2) = \sum_{i\in B} w_{ib}S_{i2}(\theta_2)$, and $S_{i2}(\theta_2) = \partial \log f(y_{2i}|x_i, y_{1i};\theta_2)/\partial\theta_2$ is the score function of θ_2. Note that we can write $\bar{S}_2(\theta_1,\theta_2) = \sum_{i\in B} w_{ib}E\{S_{i2}(\theta_2)|x_i, y_{2i};\theta\}$. Thus,

$$\begin{aligned}
\frac{\partial}{\partial\theta_1'}\bar{S}_2(\theta) &= \sum_{i\in B} w_{ib}\frac{\partial}{\partial\theta_1'}\left[\frac{\int S_{i2}(\theta_2)\,f(y_1|x_i;\theta_1)\,f(y_{2i}|x_i,y_1;\theta_2)\,dy_1}{\int f(y_1|x_i;\theta_1)\,f(y_{2i}|x_i,y_1;\theta_2)\,dy_1}\right] \\
&= \sum_{i\in B} w_{ib}E\{S_{i2}(\theta_2)S_{i1}(\theta_1)'|x_i, y_{2i};\theta\} \\
&\quad -\sum_{i\in B} w_{ib}E\{S_{i2}(\theta_2)|x_i, y_{2i};\theta\}E\{S_{i1}(\theta_1)'|x_i, y_{2i};\theta\}
\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial \theta_2'}\bar{S}_2(\theta) &= \sum_{i\in B} w_{ib}\frac{\partial}{\partial \theta_2'}\left[\frac{\int S_{i2}(\theta_2) f(y_1|x_i;\theta_1) f(y_{2i}|x_i,y_1;\theta_2)dy_1}{\int f(y_1|x_i;\theta_1) f(y_{2i}|x_i,y_1;\theta_2)dy_1}\right] \\ &= \sum_{i\in B} w_{ib} E\left\{\frac{\partial}{\partial \theta_2'} S_{i2}(\theta_2)|x_i,y_{2i};\theta\right\} \\ &+ \sum_{i\in B} w_{ib} E\{S_{i2}(\theta_2) S_{i2}(\theta_2)'|x_i,y_{2i};\theta\} \\ &- \sum_{i\in B} w_{ib} E\{S_{i2}(\theta_2)|x_i,y_{2i};\theta\} E\{S_{2i}(\theta_2)'|x_i,y_{2i};\theta\}.\end{aligned}$$

Now, $\partial\bar{S}_2(\theta)/\partial\theta_1'$ can be consistently estimated by

$$\hat{B}_{21} = \sum_{i\in B} w_{ib}\sum_{j=1}^{m} w_{ij}^* S_{2ij}^*(\hat{\theta}_2)\left\{S_{1ij}^*(\hat{\theta}_1) - \bar{S}_{1i}^*(\hat{\theta}_1)\right\}', \tag{3.5}$$

where $S_{1ij}^*(\hat{\theta}_1) = S_1(\hat{\theta}_1; x_i, y_{1i}^{*(j)})$, $S_{2ij}^*(\hat{\theta}_2) = S_2(\hat{\theta}_2; x_i, y_{1i}^{*(j)}, y_{2i})$, and $\bar{S}_{1i}^*(\hat{\theta}_1) = \sum_{j=1}^m w_{ij}^* S_1(\hat{\theta}_1; x_i, y_{1i}^{*(j)})$. Also, $\partial\bar{S}_2(\theta)/\partial\theta_2'$ can be consistently estimated by

$$-\hat{I}_{22} = \sum_{i\in B} w_{ib}\sum_{j=1}^{m} w_{ij}^* \dot{S}_{2ij}^*(\hat{\theta}_2) - \hat{B}_{22} \tag{3.6}$$

where

$$\hat{B}_{22} = \sum_{i\in B} w_{ib}\sum_{j=1}^{m} w_{ij}^* S_{2ij}^*(\hat{\theta}_2)\left\{S_{2ij}^*(\hat{\theta}_2) - \bar{S}_{2i}^*(\hat{\theta}_2)\right\}',$$

$\dot{S}_{2ij}^*(\theta_2) = \partial S_2(\theta_2; x_i, y_{1i}^{*(j)}, y_{2i})/\partial\theta_2'$ and $\bar{S}_{2i}^*(\theta_2) = \sum_{j=1}^m w_{ij}^* S_{2ij}^*(\theta_2)$.

Using a Taylor expansion with respect to θ_1,

$$\begin{aligned}\bar{S}_2(\hat{\theta}_1,\theta_2) &\cong \bar{S}_2(\theta_1,\theta_2) - E\left\{\frac{\partial}{\partial\theta_1'}\bar{S}_2(\theta)\right\}\left[E\left\{\frac{\partial}{\partial\theta_1'}S_1(\theta_1)\right\}\right]^{-1} S_1(\theta_1) \\ &= \bar{S}_2(\theta) + KS_1(\theta_1),\end{aligned}$$

and we can write

$$V(\hat{\theta}_2) \doteq \left\{E\left(\frac{\partial}{\partial\theta_2'}\bar{S}_2\right)\right\}^{-1} V\{\bar{S}_2(\theta) + KS_1(\theta_1)\}\left\{E\left(\frac{\partial}{\partial\theta_2'}\bar{S}_2\right)\right\}^{-1'}.$$

Writing

$$\bar{S}_2(\theta) = \sum_{i\in B} w_{ib}\bar{s}_{2i}(\theta),$$

with $\bar{s}_{2i}(\theta) = E\{S_{i2}(\theta_2)|x_i, y_{2i}; \theta\}$, a consistent estimator of $V\{\bar{S}_2(\theta)\}$ can be obtained by applying a design-consistent variance estimator to $\sum_{i\in B} w_{ib}\hat{s}_{2i}$ with $\hat{s}_{2i} = \sum_{j=1}^{m} w_{ij}^* S_{2ij}^*(\hat{\theta}_2)$. Under simple random sampling for Sample B, we have

$$\hat{V}\{\bar{S}_2(\theta)\} = n_B^{-2}\sum_{i\in B}\hat{s}_{2i}\hat{s}_{2i}'.$$

Also, $V\{KS_1(\theta_1)\}$ is consistently estimated by

$$\hat{V}_2 = \hat{K}\hat{V}(S_1)\hat{K}',$$

where $\hat{K} = \hat{B}_{21}\hat{I}_{11}^{-1}$, $\hat{B}_{21}$ is defined in (3.5), and $\hat{I}_{11} = -\partial S_1(\theta_1)/\partial\theta_1'$ evaluated at $\theta_1 = \hat{\theta}_1$. Since the two terms $\bar{S}_2(\theta)$ and $S_1(\theta_1)$ are independent, the variance can be estimated by

$$\hat{V}(\hat{\theta}) \doteq \hat{I}_{22}^{-1}\left[\hat{V}\{\bar{S}_2(\theta)\} + \hat{V}_2\right]\hat{I}_{22}^{-1'},$$

where $\hat{I}_{22}$ is defined in (3.6).

More generally, one may consider estimation of a parameter η defined as a root of the census estimating equation $\sum_{i=1}^{N} U(\eta; x_i, y_{1i}, y_{2i}) = 0$. Variance estimation of the FI estimator of η computed from $\sum_{i\in B} w_{ib}\sum_{j=1}^{m} w_{ij}^* U\left(\eta; x_i, y_{1i}^{*(j)}, y_{2i}\right) = 0$ is discussed in Appendix B.

4 Split questionnaire survey design

In Section 3, we consider the situation where Sample A and Sample B are two independent samples from the same target population. We now consider another situation of a split questionnaire design where the original sample S is selected from a target population and then Sample A and Sample B are randomly chosen such that $A \cup B = S$ and $A \cap B = \phi$. We observe (x, y_1) from Sample A and observe (x, y_2) from Sample B. We are interested in creating fully augmented data with observation (x, y_1, y_2) in S.

Such split questionnaire survey designs are gaining popularity because they reduce response burden (Raghunathan and Grizzle 1995; Chipperfield and Steel 2009). Split questionnaire designs have been investigated, for example, for the Consumer Expenditure survey (Gonzalez and Eltinge 2008) and the National Assessment of Educational Progress (NAEP) survey in the US. In applications of split-questionnaire designs, analysts may be interested in multiple parameters such as the mean of y_1 and the mean of y_2, in addition to the coefficient in the regression of y_2 on y_1.

We consider a design where the original Sample S is partitioned into two subsamples: A and B. We assume that x_i is observed for $i \in S$, y_{1i} is collected for $i \in A$ and y_{2i} is collected for $i \in B$. The probability of selection into A or B may depend on x_i but does not depend on y_{1i} or y_{2i}. As a consequence, the design used to select subsample A or B is non-informative for the specified model (Fuller 2009, Chapter 6). We let w_i denote the sampling weight associated with the full sample S. We assume a procedure is available for estimating the variance of an estimator of the form $\hat{Y} = \sum_{i\in S} w_i y_i$, and we denote the variance estimator by $\hat{V}_s\left(\sum_{i\in S} w_i y_i\right)$.

A procedure for obtaining a fully imputed data set is as follows. First, use the procedure of Section 3 to obtain imputed values $\left\{y_{1i}^{*(j)} : i \in B, j = 1,\ldots,m\right\}$ and an estimate, $\hat{\theta}$, of the parameter in the distribution $f(y_2 \mid y_1, x; \theta)$. The estimate $\hat{\theta}$ is obtained by solving

$$\sum_{i \in B} w_i \sum_{j=1}^{m} w_{ij}^{*} S_2\left(\theta; x_i, y_{1i}^{*(j)}, y_{2i}\right) = 0, \tag{4.1}$$

where $S_2(\theta; x, y_1, y_2) = \partial \log f(y_2 \mid y_1, x; \theta)/\partial\theta$. Given $\hat{\theta}$, generate imputed values $y_{2i}^{*(j)} \sim f\left(y_2 \mid y_{1i}, x_i; \hat{\theta}\right)$, for $i \in A$ and $j = 1,\ldots,m$.

Under the assumption that the model is identified, the parameter estimator $\hat{\theta}$ generated by solving (4.1) is fully efficient in the sense that the imputed value of y_{2i} for Sample A leads to no efficiency gain. To see this, note that the score equation using the imputed value of y_{2i} is computed by

$$\sum_{i \in A} w_i m^{-1} \sum_{j=1}^{m} S_2\left(\theta; x_i, y_{1i}, y_{2i}^{*(j)}\right) + \sum_{i \in B} w_i \sum_{j=1}^{m} w_{ij}^{*} S_2\left(\theta; x_i, y_{1i}^{*(j)}, y_{2i}\right) = 0. \tag{4.2}$$

Because $y_{2i}^{*(1)},\ldots,y_{2i}^{*(m)}$ are generated from $f\left(y_2 \mid y_{1i}, x_i; \hat{\theta}\right)$,

$$p\lim_{m \to \infty} \sum_{i \in A} w_i m^{-1} \sum_{j=1}^{m} S_2\left(\theta; x_i, y_{1i}, y_{2i}^{*(j)}\right) = \sum_{i \in A} w_i E\left\{S_2(\theta; x_i, y_{1i}, Y_2) \mid y_{1i}, x_i; \hat{\theta}\right\}.$$

Thus, by the property of score function, the first term of (4.2) evaluated at $\theta = \hat{\theta}$ is close to zero and the solution to (4.2) is essentially the same as the solution to (4.1). That is, there is no efficiency gain in using the imputed value of y_{2i} in computing the MLE for θ in $f(y_2 \mid y_1, x; \theta)$.

However, the imputed values of y_{2i} can improve the efficiency of inferences for parameters in the joint distribution of (y_{1i}, y_{2i}). As a simple example, consider estimation of μ_2, the marginal mean of y_{2i}. Under simple random sampling, the imputed estimator of $\mu = E(Y_2)$ is

$$\hat{\mu}_{I,m} = \frac{1}{n}\left\{\sum_{i \in A}\left(m^{-1}\sum_{j=1}^{m} y_{2i}^{*(j)}\right) + \sum_{i \in B} y_{2i}\right\}, \tag{4.3}$$

where $y_{2i}^{*(1)},\ldots,y_{2i}^{*(m)}$ are generated from $f\left(y_2 \mid y_{1i}, x_i; \hat{\theta}\right)$. For sufficiently large m, we can write

$$\begin{aligned}\hat{\mu}_{I,\infty} &= \frac{1}{n}\left\{\sum_{i \in A} \hat{y}_{2i} + \sum_{i \in B} y_{2i}\right\} \\ &= \frac{1}{n}\left\{\sum_{i \in A} E\left(y_2 \mid y_{1i}, x_i; \hat{\theta}\right) + \sum_{i \in B} y_{2i}\right\}.\end{aligned}$$

Under the setup of Example 2.1, we can express $\hat{y}_{2i} = \hat{\beta}_0 + \hat{\beta}_1 y_{1i} + \hat{\beta}_2 x_{2i}$ where $\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right)$ satisfies

$$\sum_{i \in B}\left(y_{2i} - \hat{\beta}_0 - \hat{\beta}_{1i}\hat{y}_{1i} - \hat{\beta}_2 x_{2i}\right) = 0$$

and $\hat{y}_{1i} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i}$ with $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)$ satisfying $\sum_{i \in A} (y_{1i} - \hat{\alpha}_0 - \hat{\alpha}_1 x_{1i} - \hat{\alpha}_2 x_{2i}) = 0$. Thus, ignoring the smaller order terms, we have

$$V\left(\hat{\mu}_{I,\infty}\right) = \frac{1}{n} V\left(y_2\right) + \left(\frac{1}{n_b} - \frac{1}{n}\right) V\left(y_2 - \hat{y}_2\right)$$

which is smaller than the variance of the direct estimator $\hat{\mu}_b = n_b^{-1} \sum_{i \in B} y_{2i}$.

5 Measurement error models

We now consider an application of statistical matching to the problem of measurement error models. Suppose that we are interested in the parameter θ in the conditional distribution $f(y_2 \mid y_1; \theta)$. In the original sample, instead of observing (y_{1i}, y_{2i}), we observe (x_i, y_{2i}), where x_i is a contaminated version of y_{1i}. Because inference for θ based on (x_i, y_{2i}) may be biased, additional information is needed. One common way to obtain additional information is to collect (x_i, y_{1i}) in an external calibration study. In this case, we observe (x_i, y_{1i}) in Sample A and (x_i, y_{2i}) in Sample B, where Sample A is the calibration sample, and Sample B is the main sample. Guo and Little (2011) discuss an application of external calibration.

The external calibration framework can be expressed as a statistical matching problem. Table 5.1 makes the connection between statistical matching and external calibration explicit. An instrumental variable assumption permits inference for θ based on data with the structure of Table 1.1. In the notation of the measurement error model, the instrumental variable assumption is

$$f(y_{2i} \mid y_{1i}, x_i) = f(y_{2i} \mid y_{1i}) \quad \text{and} \quad f(y_{1i} \mid x_i = a) \neq f(y_{1i} \mid x_i = b), \tag{5.1}$$

for some $a \neq b$. The instrumental variable assumption may be judged reasonable in applications related to error in covariates because the subject-matter model of interest is $f(y_{2i} \mid y_{1i})$, and x_i is a contaminated version of y_{1i} that contains no additional information about y_{2i} given y_{1i}.

Table 5.1
Data structure for measurement error model

	x_i	y_{1i}	y_{2i}
Survey A (calibration study)	o	o	
Survey B (main study)	o		o

For fully parametric $f(y_{2i} \mid y_{1i})$ and $f(y_{1i} \mid x_i)$, one can use parametric fractional imputation to execute the EM algorithm. This method requires evaluating the conditional expectation of the complete-data score function given the observed values. To evaluate the conditional expectation using fractional imputation, we first express the conditional distribution of y_1 given (x, y_2) as,

$$f(y_1 \mid x, y_2) \propto f(y_1 \mid x) f(y_2 \mid y_1). \tag{5.2}$$

We let an estimator $\hat{f}_a(y_{1i}|x_i)$ of $f(y_{1i}|x_i)$ be available from the calibration sample (Sample A). Implementation of the EM algorithm via fractional imputation involves the following steps:

1. For each $i \in B$, generate $y_{1i}^{*(j)}$ from $\hat{f}_a(y_1|x_i)$, for $j = 1,\ldots,m$.
2. Compute the fractional weights

$$w_{ij(t)}^* \propto f\left(y_{2i}|y_{1i}^{*(j)};\hat{\theta}_t\right)$$

with $\sum_{j=1}^m w_{ij(t)}^* = 1$.

3. Update θ by solving

$$\sum_{i\in B} w_{ib} \sum_{j=1}^m w_{ij(t)}^* S\left(\theta; y_{1i}^{*(j)}, y_{2i}\right) = 0,$$

where $S(\theta; y_1, y_2) = \partial \log f(y_2|y_1;\theta)/\partial\theta$.

4. Go to Step 2 until convergence.

The method above requires generating data from $f(y_1|x)$. For some nonlinear models or models with non-constant variances, simulating from the conditional distribution of y_1 given x may require Monte Carlo methods such as accept-reject or Metropolis Hastings. The simulation of Section 6.2 exemplifies a simulation in which the conditional distribution of $y_1|x$ has no closed form expression. In this case, we may consider an alternative approach, which may be computationally simpler. To describe this approach, let $h(y_1|x)$ be the "working" conditional distribution, such as the normal distribution, from which samples are easily generated. We assume that estimates $\hat{f}_a(y_1|x)$ and $\hat{h}_a(y_1|x)$ of $f(y_1|x)$ and $h(y_1|x)$, respectively, are available from Sample A. Implementation of the EM algorithm via fractional imputation then involves the following steps:

1. For each $i \in B$, generate $x_i^{*(j)}$ from $\hat{h}_a(y_1|x_i)$, for $j = 1,\ldots,m$.
2. Compute the fractional weights

$$w_{ij(t)}^* \propto f\left(y_{2i}|y_{1i}^{*(j)};\hat{\theta}_t\right)\hat{f}_a\left(y_{1i}^{*(j)}|x_i\right)\Big/\hat{h}_a\left(y_{1i}^{*(j)}|x_i\right) \tag{5.3}$$

with $\sum_{j=1}^m w_{ij(t)}^* = 1$.

3. Update θ by solving

$$\sum_{i\in B} w_{ib} \sum_{j=1}^m w_{ij(t)}^* S\left(\theta; y_{1i}^{*(j)}, y_{2i}\right) = 0.$$

4. Go to Step 2 until convergence.

Variance estimation is a straightforward application of the linearization method in Section 3. The hot-deck fractional imputation method described in Section 3 with fractional weights defined in (3.3) also extends readily to the measurement error setting.

6 Simulation study

To test our theory, we present two limited simulation studies. The first simulation study considers the setup of combining two independent surveys of partial observation to obtain joint analysis. The second simulation study considers the setup of measurement error models with external calibration.

6.1 Simulation one

To compare the proposed methods with the existing methods, we generate 5,000 Monte Carlo samples of (x_i, y_{1i}, y_{2i}) with size $n = 400$, where

$$\begin{pmatrix} y_{1i} \\ x_i \end{pmatrix} \sim N\left(\begin{bmatrix} 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}\right),$$

$$y_{2i} = \beta_0 + \beta_1 y_{1i} + e_i, \tag{6.1}$$

$e_i \sim N(0, \sigma^2)$, and $\beta = (\beta_0, \beta_1, \sigma^2)' = (1,1,1)'$. Note that, in this setup, we have $f(y_2 \mid x, y_1) = f(y_2 \mid y_1)$ and so the variable x plays the role of the instrumental variable for y_1.

Instead of observing (x_i, y_{1i}, y_{2i}) jointly, we assume that only (y_1, x) are observed in Sample A and only (y_2, x) are observed in Sample B, where Sample A is obtained by taking the first $n_a = 400$ elements and Sample B is obtained by taking the remaining $n_b = 400$ elements from the original sample. We are interested in estimating four parameters: three regression parameters $\beta_0, \beta_1, \sigma^2$ and $\pi = P(y_1 < 2, y_2 < 3)$, the proportion of $y_1 < 2$ and $y_2 < 3$. Four methods are considered in estimating the parameters:

1. Full sample estimation (Full): Uses the complete observation of (y_{1i}, y_{2i}) in Sample B.
2. Stochastic regression imputation (SRI): Use the regression of y_1 on x from Sample A to obtain $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\sigma}_1^2)$, where the regression model is $y_1 = \alpha_0 + \alpha_1 x + e_1$ with $e_1 \sim (0, \sigma_1^2)$. For each $i \in B$, $m = 10$ imputed values are generated by $y_{1i}^{*(j)} = \hat{\alpha}_0 + \hat{\alpha}_1 x_i + e_i^{*(j)}$ where $e_i^{*(j)} \sim N(0, \hat{\sigma}_1^2)$.
3. Parametric fractional imputation (PFI) with $m = 10$ using the instrumental variable assumption.
4. Hot-deck fractional imputation (HDFI) with $m = 10$ using the instrumental variable assumption.

Table 6.1 presents Monte Carlo means and Monte Carlo variances of the point estimators of the four parameters of interest. SRI shows large biases for all parameters considered because it is based on the conditional independence assumption. Both PFI and HDFI provide nearly unbiased estimators for all parameters. Estimators from HDFI are slightly more efficient than those from PFI because the two-step procedure in HDFI uses the full set of respondents in the first step. The theoretical asymptotic variance of $\hat{\beta}_1$ computed from PFI is

$$V(\hat{\beta}_1) \doteq \frac{1}{(0.7)^2}\frac{1}{400}2\left(1 - \frac{0.7^2}{2}\right) + \frac{1}{(0.7)^2}\frac{1}{400}(1 - 0.7^2) \doteq 0.0103$$

which is consistent with the simulation result in Table 6.1. In addition to point estimation, we also compute variance estimators for PFI and HDFI methods. Variance estimators show small relative biases (less than 5% in absolute values) for all parameters. Variance estimation results are not presented here for brevity.

Table 6.1
Monte Carlo means and variances of point estimators from Simulation One. (SRI, stochastic regression imputation; PFI, parametric fractional imputation; HDFI; hot-deck fractional imputation)

Parameter	Method	Mean	Variance
β_0	Full	1.00	0.0123
	SRI	1.90	0.0869
	PFI	1.00	0.0472
	HDFI	1.00	0.0465
β_1	Full	1.00	0.00249
	SRI	0.54	0.01648
	PFI	1.00	0.01031
	HDFI	1.00	0.01026
σ^2	Full	1.00	0.00482
	SRI	1.73	0.01657
	PFI	0.99	0.02411
	HDFI	0.99	0.02270
π	Full	0.374	0.00058
	SRI	0.305	0.00255
	PFI	0.375	0.00059
	HDFI	0.375	0.00057

The proposed method is based on the instrumental variable assumption. To study the sensitivity of the proposed fractional imputation method to violations of the instrumental variable assumption, we performed an additional simulation study. Now, instead of generating y_{2i} from (6.1), we use

$$y_{2i} = 0.5 + y_{1i} + \rho(x_i - 3) + e_i, \tag{6.2}$$

where $e_i \sim N(0,1)$ and ρ can take non-zero values. We use three values of ρ, $\rho \in \{0, 0.1, 0.2\}$, in the sensitivity analysis and apply the same PFI and HDFI procedure that is based on the assumption that x is an instrumental variable for y_1. Such assumption is satisfied for $\rho = 0$, but it is weakly violated for $\rho = 0.1$ or $\rho = 0.2$. Using the fractionally imputed data in sample B, we estimated three parameters, $\theta_1 = E(Y_1)$, θ_2 is the slope for the simple regression of y_2 on y_1, and $\theta_3 = P(y_1 < 2, y_2 < 3)$, the proportion of $y_1 < 2$ and $y_2 < 3$. Table 6.2 presents Monte Carlo means and variances of the point estimators for three parameters under three different models. In Table 6.2, the absolute values of the difference between the fractionally imputed estimator and the full sample estimator increase as the value of ρ increases, which is expected as the instrumental variable assumption is more severely violated for larger values of ρ, but the differences are relatively small for all cases. In particular, the estimator of θ_1 is not affected by the departure from the instrumental variable assumption. This is because the imputation estimator under the incorrect imputation model still provides an unbiased estimator for the population mean as long as the regression imputation model contains an intercept term (Kim and Rao 2012). Thus, this limited sensitivity analysis

suggests that the proposed method seems to provide comparable estimates when the instrumental variable assumption is weakly violated.

Table 6.2
Monte Carlo means and Monte Carlo variances of the two point estimators for sensitivity analysis in Simulation One (Full, full sample estimator; PFI, parametric fractional imputation; HDFI; hot-deck fractional imputation)

Model	Parameter	Method	Mean	Variance
$\rho = 0$	θ_1	Full	2.00	0.00235
		PFI	2.00	0.00352
		HDFI	2.00	0.00249
	θ_2	Full	1.00	0.00249
		PFI	1.00	0.01031
		HDFI	1.00	0.01026
	θ_3	Full	0.43	0.00061
		PFI	0.43	0.00059
		HDFI	0.43	0.00057
$\rho = 0.1$	θ_1	Full	2.00	0.00235
		PFI	2.00	0.00353
		HDFI	2.00	0.00250
	θ_2	Full	1.07	0.00248
		PFI	1.14	0.01091
		HDFI	1.14	0.01081
	θ_3	Full	0.44	0.00061
		PFI	0.45	0.00062
		HDFI	0.45	0.00059
$\rho = 0.2$	θ_1	Full	2.00	0.00235
		PFI	2.00	0.00353
		HDFI	2.00	0.00250
	θ_2	Full	1.14	0.00250
		PFI	1.28	0.01115
		HDFI	1.28	0.01102
	θ_3	Full	0.44	0.00061
		PFI	0.46	0.00066
		HDFI	0.46	0.00062

6.2 Simulation two

In the second simulation study, we consider a binary response variable y_{2i}, where

$$y_{2i} \sim \text{Bernoulli}(p_i), \tag{6.3}$$

$$\text{logit}(p_i) = \gamma_0 + \gamma_1 y_{1i},$$

and $y_{1i} \sim N(\mu_1, \sigma_1^2)$. In the main sample, denoted by B, instead of observing (y_{1i}, y_{2i}), we observe (x_i, y_{2i}), where

$$x_i = \beta_0 + \beta_1 y_{1i} + u_i, \tag{6.4}$$

and $u_i \sim N\left(0, \sigma^2 \mid y_{1i} \mid^{2\alpha}\right)$. We observe (x_i, y_{1i}), $i = 1, \ldots, n_A$ in a calibration sample, denoted by A. For the simulation, $n_A = n_B = 800$, $\gamma_0 = 1$, $\gamma_1 = 1$, $\beta_0 = 0$, $\beta_1 = 1$, $\sigma^2 = 0.25$, $\alpha = 0.4$, $\mu_1 = 0$, and $\sigma_1^2 = 1$. Primary interest is in estimation of γ_1 and testing the null hypothesis that $\gamma_1 = 1$. The Monte Carlo (MC) sample size is 1,000.

We compare the PFI estimators of γ_1 to three other estimators. Because the conditional distribution of y_{1i} given x_i is non-standard, we use the weights of (5.3) to implement PFI, where the proposal distribution $\hat{h}_a(y_{1i} \mid x_i)$ is an estimate of the marginal distribution of y_{1i} based on the data from Sample A. We consider the following four estimators:

1. *PFI*: For PFI, the proposal distribution for generating $y_{1i}^{*(j)}$ is a normal distribution with mean $\hat{\mu}_1$ and variance $\hat{\sigma}_1^2$, where $\hat{\mu}_1$ and $\hat{\sigma}_1^2$ are the maximum likelihood estimates of μ_1 and σ_1^2, respectively, based on Sample A. The fractional weights defined in (5.3) has the form

$$w_{ij}^* \propto \hat{p}_{ij}^{y_{2i}} \left(1 - \hat{p}_{ij}\right)^{1-y_{2i}} \hat{f}_a\left(y_{1i}^{*(j)} \mid x_i\right), \tag{6.5}$$

   where $\hat{p}_{ij} = \left\{1 + \exp\left(-\hat{\gamma}_0 - \hat{\gamma}_1 y_{1i}^{*(j)}\right)\right\}^{-1}$ and $\hat{f}_a(y_{1i} \mid x_i)$ is the estimate of $f(y_{1i} \mid x_i)$ based on maximum likelihood estimation with the Sample A data. The imputation size $m = 800$.

2. *Naive*: A *naive* estimator is the estimator of the slope in the logistic regression of y_{2i} on x_i for $i \in B$.

3. *Bayes*: We use the approach of Guo and Little (2011) to define a Bayes estimator. The model for this simulation differs from the model of Guo and Little (2011) in that the response of interest is binary. We implement GIBBS sampling with JAGS (Plummer 2003), specifying diffuse proper prior distributions for the parameters of the model. Letting

$$\theta_1 = \left(\log\left(\sigma^2\right), \log\left(\sigma_1^2\right), \mu_1, \beta_0, \beta_1, \gamma_0, \gamma_1\right),$$

   we assume a priori that $\theta_1 \sim N\left(0, 10^6 I_7\right)$, where I_7 is a 7×7 identity matrix, and the notation $N(0, V)$ denotes a normal distribution with mean 0 and covariance matrix V. The prior distribution for the power α is uniform on the interval $[-5, 5]$.

   To evaluate convergence, we examine trace plots and potential scale reduction factors defined in Gelman, Carlin, Stern and Rubin (2003) for 10 preliminary simulated data sets. We initiate three MCMC chains, each of length 1,500 from random starting values and discard the first 500 iterations as burn-in. The potential scale reduction factors across the 10 simulated data sets range from 1.0 to 1.1, and the trace plots indicate that the chains mix well. To reduce computing time, we use 1,000 iterations of a single chain for the main simulation, after discarding the first 500 for burn-in.

4. A *Weighted Regression Calibration (WRC)* estimator. The WRC estimator is a modification of the weighted regression calibration estimator defined in Guo and Little (2011) for a binary response variable. The computation for the weighted regression calibration estimator involves the following steps:

   (i) Using ordinary least squares (OLS), regress y_{1i} on x_i for the calibration sample.

(ii) Regress the logarithm of the squared residuals from step (i) on the logarithm of x_i^2 for the calibration sample. Let $\hat{\lambda}$ denote the estimated slope from the regression.

(iii) Using weighted least squares (WLS) with weight $|x_i|^{2\hat{\lambda}}$, regress y_{1i} on x_i for the calibration sample. Let $\hat{\eta}_0$ and $\hat{\eta}_1$ be the estimated intercept and slope, respectively, from the WLS regression.

(iv) For each unit i in the main sample, let $\hat{y}_{1i} = \hat{\eta}_0 + \hat{\eta}_1 x_i$.

(v) The estimate of (γ_0, γ_1) is obtained from the logistic regression of y_{2i} on $\hat{y}_{1i}$ in the main sample.

Table 6.3 contains the MC bias, variance, and MSE of the four estimators of γ_1. The naive estimator has a negative bias because x_i is a contaminated version of y_{1i}. The PFI estimator is superior to the Bayes and WRC estimators.

We compute an estimate of the variance of the PFI estimators of γ_1 using the variance expression based on the linear approximation. We define the MC relative bias as the ratio of the difference between the MC mean of the variance estimator and the MC variance of the estimator to the MC variance of the estimator. The MC relative bias of the variance estimators for PFI is negligible (less than 2% in absolute values).

Table 6.3
Monte Carlo (MC) means, variances, and mean squared errors (MSE) of point estimators of γ_1 from Simulation Two. (PFI, parametric fractional imputation; WRC, weighted regression calibration; MC, Monte Carlo; MSE, mean squared error)

Method	MC Bias	MC Variance	MC MSE
PFI	0.0239	0.0386	0.0392
Naive	-0.2241	0.0239	0.0742
Bayes	0.0406	0.0415	0.0432
WRC	0.112	0.0499	0.0625

7 Concluding remarks

We approach statistical matching as a missing data problem and propose the PFI method to obtain consistent estimators and corresponding variance estimators. Under the assumption that the specified model is fully identified, the proposed method provides the pseudo maximum likelihood estimators of the parameters in the model.

A sufficient condition for model identifiability is the existence of an instrumental variable in the model. The measurement error framework of Section 5 and Section 6.2, where external calibration provides an independent measurement of the true covariate of interest, is a situation in which the study design may be judged to support the instrumental variable assumption. The proposed methodology is applicable without the instrumental variable assumption, as long as the model is identified. If the model is not identifiable, then the EM algorithm for the proposed PFI method does not necessarily converge. In practice, one can treat the specified model as identified if the EM sequence converges. That is, as long as the EM sequence converges,

the resulting analysis is consistent under the specified model. This is one of the main advantages of using the frequentist approach over Bayesian. In the Bayesian approach, it is possible to obtain the posterior values even under non-identified models and the resulting analysis can be misleading.

Testing whether the IV assumption holds in the data at hand is much more difficult, perhaps impossible, under the data structure in Table 1.1. Instead, given the specified model, we can only check whether the model parameters are fully estimable. On the other hand, whether the specified model is a good model for the data at hand is a different story. Model diagnostics and model selection among different identifiable models are certainly important future research topics.

Statistical matching can also be used to evaluate effects of multiple treatments in observational studies. By properly applying statistical matching techniques, we can create an augmented data file of potential outcomes so that causal inference can be investigated with the augmented data file (Morgan and Winship 2007). Such extensions will be presented elsewhere.

Acknowledgements

We thank Professor Yanyuan Ma, an anonymous referee and the Assistant Editor (AE) for very constructive comments. The research of the first author was partially supported by Brain Pool program (131S-1-3-0476) from Korean Federation of Science and Technology Society and by a grant from NSF (MMS-121339). The research of the second author was supported by a Cooperative Agreement between the US Department of Agriculture Natural Resources Conservation Service and Iowa State University. The work of the third author was supported by the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of Science, ICT and Future Planning through the National Research Foundation in Korea.

Appendix

A. Asymptotic unbiasedness of 2SLS estimator

Assume that we observe (y_1, x) in Sample A and observe (y_2, x) in Sample B. To be more rigorous, we can write (y_{1a}, x_a) to denote the observation (y_1, x) in Sample A. Also, we can write (y_{2b}, x_b) to denote the observations in Sample B. In this case, the model can be written as

$$\begin{aligned} y_{1a} &= \phi_0 1_a + \phi_1 x_{1a} + \phi_2 x_{2a} + e_{1a} \\ y_{2b} &= \beta_0 1_b + \beta_1 y_{1b} + \beta_2 x_{2b} + e_{2b} \end{aligned}$$

with $E(e_{1a} \mid x_a) = 0$ and $E(e_{2b} \mid x_b, y_{1b}) = 0$. Note that y_{1b} is not observed from the sample. Instead, we use $\hat{y}_{1b}$ using the OLS estimate obtained from Sample A.

Writing $X_a = [1_a, x_a]$ and $X_b = [1_b, x_b]$, we have $\hat{y}_{1b} = X_b (X_a' X_a)^{-1} X_a' y_{1a} = X_b \hat{\phi}_a$. The 2SLS estimator of $\beta = (\beta_0, \beta_1, \beta_2)'$ is then

$$\hat{\beta}_{2\text{SLS}} = (Z_b' Z_b)^{-1} Z_b' y_{2b}$$

where $Z_b = [1_b, \hat{y}_{1b}, x_{2b}]$. Thus, we have

$$\begin{aligned} \hat{\beta}_{2\text{SLS}} - \beta &= (Z_b' Z_b)^{-1} Z_b' (y_{2b} - Z_b \beta) \\ &= (Z_b' Z_b)^{-1} Z_b' \{\beta_1 (y_{1b} - \hat{y}_{1b}) + e_{2b}\}. \end{aligned} \tag{A.1}$$

We may write

$$y_{1b} = \phi_0 1_b + \phi_1 x_b + e_{1b} = X_b \phi + e_{1b}$$

where $E(e_{1b} | x_b) = 0$. Since

$$\begin{aligned} \hat{y}_{1b} &= X_b (X_a' X_a)^{-1} X_a' y_{1a} \\ &= X_b (X_a' X_a)^{-1} X_a' (X_a \phi + e_{1a}) \\ &= X_b \phi + X_b (X_a' X_a)^{-1} X_a' e_{1a}, \end{aligned}$$

we have

$$y_{1b} - \hat{y}_{1b} = e_{1b} - X_b (X_a' X_a)^{-1} X_a' e_{1a}$$

and (A.1) becomes

$$\hat{\beta}_{2\text{SLS}} - \beta = (Z_b' Z_b)^{-1} Z_b' \left\{\beta_1 e_{1b} - \beta_1 X_b (X_a' X_a)^{-1} X_a' e_{1a} + e_{2b}\right\}. \tag{A.2}$$

Assume that the two samples are independent. Thus, $E(e_{1b} | x_a, x_b, y_{1a}) = 0$. Also, $E\left\{(Z_b' Z_b)^{-1} Z_b' e_{2b} \,\middle|\, x_a, x_b, y_{1a}, y_{1b}\right\} = 0$. Thus,

$$E\left\{\hat{\beta}_{2\text{SLS}} - \beta \,\middle|\, x_a, x_b, y_{1a}\right\} = E\left\{-\beta_1 (Z_b' Z_b)^{-1} Z_b' X_b (X_a' X_a)^{-1} X_a' e_{1a} \,\middle|\, x_a, x_b, y_{1a}\right\}$$

and

$$\begin{aligned} (Z_b' Z_b)^{-1} Z_b' X_b (X_a' X_a)^{-1} X_a' e_{1a} &= (Z_b' Z_b)^{-1} Z_b' \left\{X_b (X_a' X_a)^{-1} X_a' (y_{1a} - X_a \phi)\right\} \\ &= (Z_b' Z_b)^{-1} Z_b' X_b \left(\hat{\phi}_a - \phi\right). \end{aligned}$$

This term has zero expectation asymptotically because $n_b^{-1} Z_b' Z_b$ and $n_b^{-1} Z_b' X_b$ are bounded in probability and $\left(\hat{\phi}_a - \phi\right)$ converges to zero.

B. Variance estimation

Let the parameter of interest be defined by the solution to $U_N(\eta) = \sum_{i=1}^{N} U(\eta; y_{1i}, y_{2i}) = 0$. We assume that $\partial U_N(\eta)/\partial\theta = 0$. Thus, parameter η is priori independent of θ which is the parameter in the data-generating distribution of (x, y_1, y_2).

Under the setup of Section 3, let $\hat{\theta} = \left(\hat{\theta}_1, \hat{\theta}_2\right)$ be the MLE of $\theta = (\theta_1, \theta_2)$ obtained by solving (3.4). Also, let $\hat{\eta}$ be the solution to $\bar{U}\left(\eta | \hat{\theta}\right) = 0$ where

$$\bar{U}(\eta | \theta) = \sum_{i \in B} \sum_{j=1}^{m} w_{ib} w_{ij}^{*} U\left(\eta; y_{1i}^{*(j)}, y_{2i}\right),$$

and

$$w_{ij}^{*} \propto f\left(y_{1i}^{*(j)} \mid x_i; \hat{\theta}_1\right) f\left(y_{2i} \mid y_{1i}^{*(j)}; \hat{\theta}_2\right) \Big/ h\left(y_{1i}^{*(j)} \mid x_i\right)$$

with $\sum_{j=1}^{m} w_{ij}^{*} = 1$. Here, $h(y_1 \mid x)$ is the proposal distribution of generating imputed values of y_1 in the parametric fractional imputation. By introducing the proposal distribution h, we can safely ignore the dependence of imputed values $y_{1i}^{*(j)}$ on the estimated parameter value $\hat{\theta}_1$.

By Taylor linearization,

$$\bar{U}\left(\eta \mid \hat{\theta}\right) \cong \bar{U}\left(\eta \mid \theta\right) + \left(\partial \bar{U} / \partial \theta_1'\right)\left(\hat{\theta}_1 - \theta_1\right) + \left(\partial \bar{U} / \partial \theta_2'\right)\left(\hat{\theta}_2 - \theta_2\right)$$

Note that

$$\hat{\theta}_1 - \theta_1 \cong \left\{I_1\left(\theta_1\right)\right\}^{-1} S_1\left(\theta_1\right)$$

where $I_1\left(\theta_1\right) = -\partial S_1\left(\theta_1\right) / \partial \theta_1'$. Also,

$$\hat{\theta}_2 - \theta_2 \cong \left\{-\frac{\partial}{\partial \theta_2'} \bar{S}_2\left(\theta\right)\right\}^{-1} \bar{S}_2\left(\theta\right)$$

where

$$\bar{S}_2\left(\theta\right) = \sum_{i \in B} \sum_{j=1}^{m} w_i w_{ij}^{*}\left(\theta\right) S_2\left(\theta_2; y_{1i}^{*(j)}, y_{2i}\right).$$

Thus, we can establish

$$\bar{U}\left(\eta \mid \hat{\theta}\right) \cong \bar{U}\left(\eta \mid \theta\right) + K_1 S_1\left(\theta_1\right) + K_2 \bar{S}_2\left(\theta\right),$$

where $K_1 = D_{21} I_{11}^{-1}$ and $K_2 = D_{22} I_{22}^{-1}$ with $I_{11} = -E\left(\partial S_1 / \partial \theta_1'\right)$, $I_{22} = -E\left(\partial \bar{S}_2 / \partial \theta_2'\right)$, $D_{21} = E\left\{U\left(\eta\right) S_1\left(\theta_1\right)'\right\}$ and $D_{22} = E\left\{U\left(\eta\right) S_2\left(\theta_2\right)'\right\}$, we have

$$V\left\{\bar{U}\left(\eta \mid \hat{\theta}\right)\right\} = \tau^{-1}\left\{V_1 + V_2\right\}\tau^{-1'}$$

where $\tau = -E\left\{\partial \bar{U}\left(\eta \mid \theta\right) / \partial \eta'\right\}$,

$$V_1 = V\left\{\sum_{i \in B} w_i\left(\bar{u}_i^{*} + K_2 S_{2i}^{*}\right)\right\},$$

$\bar{u}_i^{*} = E\left[U\left(\hat{\eta}; y_{1i}, y_{2i}\right) \mid y_{2i}; \hat{\theta}\right]$, and $V_2 = V\left\{K_1 \sum_{i \in A} w_i S_{1i}\right\}$. A consistent estimator of each component can be developed similarly to Section 3.

References

Baker, K.H., Harris, P. and O'Brien, J. (1989). Data fusion: An appraisal and experimental evaluation. *Journal of the Market Research Society*, 31, 152-212.

Beaumont, J.-F., and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 37, 3, 400-416.

Chen, J., and Shao, J. (2001). Jackknife variance estimation for nearest neighbor imputation. *Journal of the American Statistical Association*, 96, 453, 260-269.

Chib, S., and Greenberg, E. (1995). Jackknife variance estimation for nearest neighbor imputation. *The American Statistician*, 46, 327-333.

Chipperfield, J.O., and Steel, D.G. (2009). Design and estimation for split questionnaire surveys. *Journal of Official Statistics*, 25, 2, 227-244.

D'Orazio, M., Zio, M.D. and Scanu, M. (2006). *Statistical Matching: Theory and Practice.* Chichester, UK: Wiley.

Fuller, W.A. (2009). *Sampling Statistics*, Hoboken, NJ: John Wiley & Sons, Inc.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003). *Bayesian Data Analysis*, Chapman and Hall Texts in Statistical Science. Chapman and Hall/CRC, second edition.

Gonzalez, J., and Eltinge, J. (2008). Adaptive matrix sampling for the consumer expenditure quarterly interview survey. In *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2081-2088.

Guo, Y., and Little, R.J. (2011). Regression analysis with covariates that have heteroskedastic measurement error. *Statistics Medicine*, 30, 18, 2278-2294.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In *Handbook of Statistics, Volume 29, Sample Surveys: Theory Methods and Inference*, (Eds., C.R. Rao and D. Pfeffermann), 215-246.

Herzog, T.N., Scheuren, F.J. and Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques.* New York: Springer.

Ibrahim, J.G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85, 765-769.

Kim, J.K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98, 119-132.

Kim, J.K., and Rao, J.N.K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika*, 99, 85-100.

Kim, J.K., and Shao, J. (2013). *Statistical Methods in Handling Incomplete Data*, Chapman and Hall/CRC.

Kim, J.K., and Yang, S. (2014). Fractional hot deck imputation for robust inference under item nonresponse in survey sampling. *Survey Methodology*, 40, 2, 211-230.

Lahiri, P., and Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 1265-1275.

Leulescu, A., and Agafitei, M. (2013). Statistical matching: A model based approach for data integration. *Eurostat Methodologies and Working Papers.*

Morgan, S.L., and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* New York, USA: Cambridge University Press.

Moriarity, C., and Scheuren, F. (2001). Statistical matching: A paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics*, 17, 407-422.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Raghunathan, T.E., and Grizzle, J.E. (1995). A split questionnaire design. *Journal of the American Statistical Association*, 90, 54-63.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York: Springer-Verlag.

Ridder, S., and Moffit, R. (2007). The econometrics of data combination. *Handbook of Econometrics*, 5470-5544.

Survey Methodology, June 2016
Vol. 42, No. 1, pp. 41-61
Statistics Canada, Catalogue No. 12-001-X

Comparison of unit level and area level small area estimators

Michael A. Hidiroglou and Yong You[1]

Abstract

In this paper, we compare the EBLUP and pseudo-EBLUP estimators for small area estimation under the nested error regression model and three area level model-based estimators using the Fay-Herriot model. We conduct a design-based simulation study to compare the model-based estimators for unit level and area level models under informative and non-informative sampling. In particular, we are interested in the confidence interval coverage rate of the unit level and area level estimators. We also compare the estimators if the model has been misspecified. Our simulation results show that estimators based on the unit level model perform better than those based on the area level. The pseudo-EBLUP estimator is the best among unit level and area level estimators.

Key Words: Confidence interval; Design consistency; Fay-Herriot model; Informative sampling; Model misspecification; Nested error regression model; Relative root mean squared error (RRMSE); Survey weight.

1 Introduction

Model-based small area estimators have been widely used in practice to provide reliable indirect estimates for small areas in recent years. The model-based estimators are based on explicit models that provide a link to related small areas through supplementary data such as census and administrative records. Small area models can be classified into two broad types: (i) Unit level models that relate the unit values of the study variable to unit-specific auxiliary variables and (ii) Area level models that relate direct estimators of the study variable of the small area to the corresponding area-specific auxiliary variables. In general, area level models are used to improve the direct estimators if unit level data are not available. The sampling set-up is as in Rao (2003). That is, a universe U of size N is split into m non-overlapping small areas U_i of size N_i, where $i = 1,\ldots,m$. Sampling is carried out in each small area using a probabilistic mechanism, resulting in samples s_i of size n_i. The selection probabilities associated with each element $j = 1,\ldots,n_i$ selected in sample s_i is denoted as p_{ij}. The resulting design weights are given by $w_{ij} = n_i^{-1} p_{ij}^{-1}$. In practice, these weights can be adjusted to account for non-response and/or auxiliary information. The resulting weights are known as the survey weights. In this paper, we assume full response to the survey, and no adjustment to the auxiliary data. Direct area level estimates are obtained for each area using the survey weights and unit observations from the area. The survey design can be incorporated into small area models in different ways. In the area level case, direct design-based estimators are modeled directly and the survey variance of the associated direct estimator is introduced into the model via the design-based errors. In the case of the unit level, the observations can be weighted using the survey weight. A number of factors affect the success of using these estimators. Two important factors are whether the assumed model is correct and whether the variable of interest is correlated with the selection probabilities associated with the sampling process, that is, informativeness of the sampling process. In this paper, we compare, via a simulation study, the impact of model misspecification and the informativeness of the sampling design for two basic small area procedures based on unit and area levels in terms of bias, estimated mean squared error and confidence

1. Michael A. Hidiroglou, Business Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6, Canada. E-mail: hidirog@yahoo.ca; Yong You, International Cooperation and Corporate Statistical Methods Division, Statistics Canada, Ottawa, K1A 0T6, Canada. E-mail: yong.you@canada.ca.

interval coverage rates. A sampling design is informative if the selection probabilities p_{ij} are related to the variable of interest y_{ij} even after conditioning on the covariates $\mathbf{x}_{ij}$. In such cases, we have informative sampling in the sense that the population model no longer holds for the sample. Pfeffermann and Sverchkov (2007) accounted for this possibility by adjusting the small area procedures. Verret, Rao and Hidiroglou (2015) simplified the procedure. In this paper, we do not adjust the small area procedures for informativeness, but study their impact.

The paper is structured as follows. The point estimators and associated mean squared error estimators for the unit level and area models are described in Section 2 and in Section 3 respectively. The description of the simulation and results are given in Section 4. This simulation computes the point and associated mean squared errors for a PPSWR (probability proportional to size with replacement) sampling scheme by varying the following two factors: (a) the assumed model is correct or incorrect; and (b) design informativeness varies from being non-significant to being very significant. In Section 5, we give an example using data from Battese, Harter and Fuller (1988) that compares the unit level and area level estimates. Finally, conclusions resulting from this work are presented in Section 6.

2 Unit level model

A basic unit level model for small area estimation is the nested error regression model (Battese et al. 1988) given by $y_{ij} = \mathbf{x}'_{ij}\beta + v_i + e_{ij}, \quad j = 1,\ldots,N_i, i = 1,\ldots,m,$ where y_{ij} is the variable of interest for the j^{th} population unit in the i^{th} small area, $\mathbf{x}_{ij} = \left(x_{ij1},\ldots,x_{ijp}\right)'$ is a $p \times 1$ vector of auxiliary variables, with $x_{ij1} = 1,$ $\boldsymbol{\beta} = \left(\beta_0,\ldots,\beta_{p-1}\right)'$ is a $p \times 1$ vector of regression parameters, and N_i is the number of population units in the i^{th} small area. The random effects v_i are assumed to be independent and identically distributed $(i.i.d.)$ $N\left(0,\sigma_v^2\right)$ and independent of the unit errors $e_{ij},$ which are assumed to be $i.i.d.$ $N\left(0,\sigma_e^2\right).$ Assuming that N_i is large, the parameter of interest is the mean for the i^{th} area, $\bar{Y}_i = N_i^{-1}\sum_{j=1}^{N_i} y_{ij},$ which may be approximated by

$$\theta_i = \bar{\mathbf{X}}'_i\boldsymbol{\beta} + v_i, \tag{2.1}$$

where $\bar{\mathbf{X}}_i = \sum_{j=1}^{N_i}\mathbf{x}_{ij} \Big/ N_i$ is the vector of known population means of the $\mathbf{x}_{ij}$ for the i^{th} area. We assume that samples are drawn independently within each small area according to a specified sampling design. Under non-informative sampling, the sample data $\left(y_{ij}, \mathbf{x}_{ij}\right)$ are assumed to obey the population model, i.e.,

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + v_i + e_{ij}, \quad j = 1,\ldots,n_i, \quad i = 1,\ldots,m, \tag{2.2}$$

where w_{ij} is the basic design weight associated with unit $(i, j),$ and n_i is the sample size in the i^{th} small area.

2.1 EBLUP estimation

The best linear unbiased prediction (BLUP) estimator of small area mean, $\theta_i = \bar{\mathbf{X}}'_i\boldsymbol{\beta} + v_i,$ based on the nested error regression model (2.2) is given by

$$\tilde{\theta}_i = r_i \bar{y}_i + \left(\bar{\mathbf{X}}_i - r_i \bar{\mathbf{x}}_i\right)' \tilde{\boldsymbol{\beta}}, \tag{2.3}$$

where $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} \Big/ n_i$, $\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij} \Big/ n_i$, $r_i = \sigma_v^2 \Big/ \left(\sigma_v^2 + \sigma_e^2 \Big/ n_i\right)$, and

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \bar{\mathbf{x}}_i' \mathbf{V}_i^{-1} \mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^{m} \bar{\mathbf{x}}_i' \mathbf{V}_i^{-1} y_i\right) \equiv \tilde{\boldsymbol{\beta}}\left(\sigma_e^2, \sigma_v^2\right), \tag{2.4}$$

with $\mathbf{x}_i' = \left(x_{i1}, \ldots, x_{in_i}\right)$, $\mathbf{V}_i = \sigma_e^2 \mathbf{I}_{n_i} + \sigma_v^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}'$, $y_i = \left(y_{i1}, \ldots, y_{in_i}\right)'$, $i = 1, \ldots, m$. Both $\tilde{\theta}_i$ and $\tilde{\boldsymbol{\beta}}$ depend on the unknown variance parameters σ_e^2 and σ_v^2. The method of fitting constant can be used to estimate σ_e^2 and σ_v^2, and the resulting estimators are $\hat{\sigma}_e^2 = (n - m - p + 1)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij}^2$, and $\hat{\sigma}_v^2 = \max\left(\tilde{\sigma}_v^2, 0\right)$, where $\tilde{\sigma}_v^2 = n_*^{-1}\left[\sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{u}_{ij}^2 - (n - p)\hat{\sigma}_e^2\right]$, $n_* = n - \operatorname{tr}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1} \sum_{i=1}^{m} n_i^2 \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i'\right]$, $\mathbf{X}' = \left(x_1', \ldots, x_m'\right)$, $n = \sum_{i=1}^{m} n_i$.

The residuals $\left\{\hat{\varepsilon}_{ij}\right\}$ are obtained from the ordinary least squares (OLS) regression of $y_{ij} - \bar{y}_i$ on $\left\{\mathbf{x}_{ij1} - \bar{\mathbf{x}}_{i\cdot 1}, \ldots, \mathbf{x}_{ijp} - \bar{\mathbf{x}}_{i\cdot p}\right\}$ and $\left\{\hat{u}_{ij}\right\}$ are the residuals from the OLS regression of y_{ij} on $\left\{x_{ij1}, \ldots, x_{ijp}\right\}$. See Rao (2003), page 138 for more details.

Replacing σ_e^2 and σ_v^2 by estimators $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$ in equation (2.3), we obtain the EBLUP estimator of small area mean θ_i as

$$\hat{\theta}_i^{\text{EBLUP}} = r_i \bar{y}_i + \left(\bar{\mathbf{X}}_i - \hat{r}_i \bar{\mathbf{x}}_i\right)' \hat{\boldsymbol{\beta}}, \tag{2.5}$$

where $\hat{r}_i = \hat{\sigma}_v^2 \Big/ \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \Big/ n_i\right)$ and $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)$. The mean squared error (MSE) of the EBLUP estimator $\hat{\theta}_i^{\text{EBLUP}}$ is given by

$$\operatorname{MSE}\left(\hat{\theta}_i^{\text{EBLUP}}\right) \approx g_{1i}\left(\sigma_e^2, \sigma_v^2\right) + g_{2i}\left(\sigma_e^2, \sigma_v^2\right) + g_{3i}\left(\sigma_e^2, \sigma_v^2\right),$$

see Prasad and Rao (1990). The g – terms are

$$\begin{aligned} g_{1i}\left(\sigma_e^2, \sigma_v^2\right) &= (1 - r_i)\sigma_v^2, \\ g_{2i}\left(\sigma_e^2, \sigma_v^2\right) &= \left(\bar{\mathbf{X}}_i - r_i \bar{\mathbf{x}}_i\right)' \left(\sum_{i=1}^{m} \mathbf{x}_i' \mathbf{V}_i^{-1} \mathbf{x}_i\right)^{-1} \left(\bar{\mathbf{X}}_i - r_i \bar{\mathbf{x}}_i\right) \end{aligned}$$

and

$$g_{3i}\left(\sigma_e^2, \sigma_v^2\right) = n_i^{-2}\left(\sigma_v^2 + \sigma_e^2 n_i^{-1}\right)^{-3} h\left(\sigma_e^2, \sigma_v^2\right),$$

where $h\left(\sigma_e^2, \sigma_v^2\right) = \sigma_e^4 V\left(\tilde{\sigma}_v^2\right) - 2\sigma_e^2 \sigma_v^2 \operatorname{cov}\left(\hat{\sigma}_e^2, \tilde{\sigma}_v^2\right) + \sigma_v^4 V\left(\hat{\sigma}_e^2\right)$. The variances and covariance of $\hat{\sigma}_e^2$ and $\tilde{\sigma}_v^2$ are given by

$$\begin{aligned} V\left(\hat{\sigma}_e^2\right) &= 2(n - m - p + 1)^{-1} \sigma_e^4 \\ V\left(\tilde{\sigma}_v^2\right) &= 2n_*^{-2}\left[(n - m - p + 1)^{-1}(m - 1)(n - p)\sigma_e^4 + 2n_* \sigma_e^2 \sigma_v^2 + n_{**} \sigma_v^4\right], \end{aligned}$$

and

$$\operatorname{cov}\left(\hat{\sigma}_e^2, \tilde{\sigma}_v^2\right) = -(m - 1) n_*^{-1} V\left(\hat{\sigma}_e^2\right),$$

where $n_{**} = \mathrm{tr}\left(\mathbf{Z}'\mathbf{M}\mathbf{Z}\right)^2$, $\mathbf{M} = \mathbf{I}_n - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$, $\mathbf{Z} = \mathrm{diag}\left(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_m}\right)$.

A second-order unbiased estimator of the MSE (Prasad and Rao 1990) is given by

$$\mathrm{mse}\left(\hat{\theta}_i^{\mathrm{EBLUP}}\right) = g_{1i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + g_{2i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + 2g_{3i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right). \tag{2.6}$$

Note that the EBLUP estimator $\hat{\theta}_i^{\mathrm{EBLUP}}$ given by (2.5) depends on the unit level model (2.2). It is model-unbiased, but it is not design consistent unless the sample design is simple random sampling. If model (2.2) does not hold for the sampled data, then the EBLUP estimator $\hat{\theta}_i^{\mathrm{EBLUP}}$ may be biased, that is, additional bias will be present in the EBLUP estimator due to model misspecification.

2.2 Pseudo-EBLUP estimation

You and Rao (2002) proposed a pseudo-EBLUP estimator of the small area mean θ_i by combining the survey weights and the unit level model (2.2) to achieve design consistency. Let w_{ij} be the weights associated with each unit (i, j). A direct design-based estimator of the small area mean is given by

$$\overline{y}_{iw} = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}} = \sum_{j=1}^{n_i} \tilde{w}_{ij} y_{ij}, \tag{2.7}$$

where $\tilde{w}_{ij} = w_{ij} \Big/ \sum_{j=1}^{n_i} w_{ij} = w_{ij} / w_{i.}$ and $\sum_{j=1}^{n_i} \tilde{w}_{ij} = 1$. The weighted estimator $\overline{y}_{iw}$ is also known as the weighted Hájek estimator. By combining the direct estimator (2.7) and the unit level model (2.2), we can obtain the following aggregated (survey-weighted) area level model

$$\overline{y}_{iw} = \overline{\mathbf{x}}_{iw}'\boldsymbol{\beta} + v_i + \overline{e}_{iw}, \quad i = 1, \ldots, m, \tag{2.8}$$

where $\overline{e}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{ij} e_{ij}$ with $E\left(\overline{e}_{iw}\right) = 0$, $V\left(\overline{e}_{iw}\right) = \sigma_e^2 \sum_{j=1}^{n_i} \tilde{w}_{ij}^2 \equiv \delta_i^2$, and $\overline{\mathbf{x}}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{ij} \mathbf{x}_{ij}$. Note that the regression parameter $\boldsymbol{\beta}$ and the variance components σ_e^2 and σ_v^2 are unknown in model (2.8). Based on model (2.8), assuming that the parameters $\boldsymbol{\beta}$, σ_e^2 and σ_v^2 are known, the BLUP estimator of θ_i is

$$\tilde{\theta}_{iw} = r_{iw}\overline{y}_{iw} + \left(\overline{\mathbf{X}}_i - r_{iw}\overline{\mathbf{x}}_{iw}\right)'\boldsymbol{\beta} = \tilde{\theta}_{iw}\left(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2\right), \tag{2.9}$$

where $r_{iw} = \sigma_v^2 \Big/ \left(\sigma_v^2 + \delta_i^2\right)$. The BLUP estimator $\tilde{\theta}_{iw}$ depends on $\boldsymbol{\beta}$, σ_e^2 and σ_v^2. To estimate the regression parameter, You and Rao (2002) proposed a weighted estimation equation approach, and obtained an estimator of $\boldsymbol{\beta}$ as follows:

$$\tilde{\boldsymbol{\beta}}_w = \left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}\left(\mathbf{x}_{ij} - r_{iw}\overline{\mathbf{x}}_{iw}\right)'\right]^{-1}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\left(\mathbf{x}_{ij} - r_{iw}\overline{\mathbf{x}}_{iw}\right)y_{ij}\right] \equiv \tilde{\boldsymbol{\beta}}_w\left(\sigma_e^2, \sigma_v^2\right).$$

$\tilde{\boldsymbol{\beta}}_w = \tilde{\boldsymbol{\beta}}_w\left(\sigma_e^2, \sigma_v^2\right)$ depends on σ_e^2 and σ_v^2. Replacing σ_e^2 and σ_v^2 in $\tilde{\boldsymbol{\beta}}_w$ by the fitting of constant estimators $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$, $\hat{\boldsymbol{\beta}}_w = \tilde{\boldsymbol{\beta}}_w\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)$ is obtained; See Rao (2003, page 149). Replacing $\boldsymbol{\beta}$, σ_e^2 and σ_v^2 in (2.9) by $\hat{\boldsymbol{\beta}}_w$, $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$, the pseudo-EBLUP estimator for the small area mean θ_i is given by

$$\hat{\theta}_i^{P-\text{EBLUP}} \triangleq \hat{\theta}_{iw} = \hat{r}_{iw}\overline{y}_{iw} + \left(\overline{\mathbf{X}}_i - \hat{r}_{iw}\overline{\mathbf{x}}_{iw}\right)' \hat{\boldsymbol{\beta}}_w. \tag{2.10}$$

As the sample size n_i becomes large, estimator $\hat{\theta}_i^{P-\text{EBLUP}}$ becomes design-consistent. It also has a self-benchmarking property when the weights w_{ij} are calibrated to agree with the known population total. That is, if $\sum_{j=1}^{n_i} w_{ij} = N_i$, $\sum_{i=1}^{m} N_i \hat{\theta}_i^{P-\text{EBLUP}}$ is equal to the direct regression estimator of the overall total,

$$\sum_{i=1}^{m} N_i \hat{\theta}_i^{P-\text{EBLUP}} = \hat{Y}_w + \left(\mathbf{X} - \hat{\mathbf{X}}_w\right)' \hat{\boldsymbol{\beta}}_w,$$

where $\hat{Y}_w = \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij} y_{ij}$, and $\hat{\mathbf{X}}_w = \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}$. For more details, see You and Rao (2002).

The MSE of $\hat{\theta}_i^{P-\text{EBLUP}}$ is given by

$$\text{MSE}\left(\hat{\theta}_i^{P-\text{EBLUP}}\right) \approx g_{1iw}\left(\sigma_e^2, \sigma_v^2\right) + g_{2iw}\left(\sigma_e^2, \sigma_v^2\right) + g_{3iw}\left(\sigma_e^2, \sigma_v^2\right),$$

where $g_{1iw}\left(\sigma_e^2, \sigma_v^2\right) = (1 - r_{iw})\sigma_v^2$, $g_{2iw}\left(\sigma_e^2, \sigma_v^2\right) = \left(\overline{X}_i - r_{iw}\overline{x}_{iw}\right)'\Phi_w\left(\overline{X}_i - r_{iw}\overline{x}_{iw}\right)$. The term Φ_w is

$$\Phi_w \left(\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathbf{x}_{ij} z_{ij}'\right)^{-1} \left(\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathbf{z}_{ij}\mathbf{z}_{ij}'\right)\left[\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathbf{x}_{ij}\mathbf{z}_{ij}'\right)^{-1}\right]' \sigma_e^2$$

$$+ \left(\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathbf{x}_{ij}\mathbf{z}_{ij}'\right)^{-1}\left[\sum_{i=1}^{m}\left(\sum_{j=1}^{n_i}\mathbf{z}_{ij}\right)\left(\sum_{j=1}^{n_i}\mathbf{z}_{ij}\right)'\right]\left[\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}\mathbf{z}_{ij}\right)^{-1}\right]' \sigma_v^2,$$

where $\mathbf{z}_{ij} = w_{ij}\left(\mathbf{x}_{ij} - r_{iw}\overline{\mathbf{x}}_{iw}\right)$, $g_{3iw}\left(\sigma_e^2, \sigma_v^2\right) = r_{iw}(1 - r_{iw})^2 \sigma_e^{-4}\sigma_v^{-2} h\left(\sigma_e^2, \sigma_v^2\right)$. $h\left(\sigma_e^2, \sigma_v^2\right)$ is the same function as in the MSE for the EBLUP estimator given in Section 2.1. A nearly second-order unbiased estimator of the MSE can be obtained as

$$\text{mse}\left(\hat{\theta}_i^{P-\text{EBLUP}}\right) = g_{1iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + g_{2iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + 2g_{3iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right). \tag{2.11}$$

(See Rao 2003, page 150 and You and Rao 2002, page 435). Note that the MSE estimator (2.11) ignores the cross-product terms. Torabi and Rao (2010) obtained the second-order correct MSE estimator including the cross-product terms using linearization and bootstrap methods. There are two cross-product terms. The first one is simple and has a closed form. Although the linearization method performs well, the explicit form for the second cross-product term is very lengthy: furthermore, the formulas based on the linearization procedure are not provided in Torabi and Rao (2010). The bootstrap method always underestimates the true MSE. A double bootstrap method needs to be applied to get an unbiased estimator of the MSE and is computationally extensive. The MSE estimator (2.11) behaves like the linearization estimator of Torabi and Rao (2010) when the variation of the survey weights is small. In the case of self-weighting within areas, one of the cross-product term is zero and the other term is of order $o\left(m^{-1}\right)$. Hence, the MSE estimator (2.11) is nearly unbiased; more discussion is provided in Torabi and Rao (2010). It is for these reasons that these cross-product terms were not included in the MSE estimator given by (2.11) in our study.

Note that under model (2.2) the pseudo-EBLUP estimator $\hat{\theta}_i^{P-\text{EBLUP}}$ is slightly less efficient than the EBLUP estimator $\hat{\theta}_i^{\text{EBLUP}}$. However, the pseudo-EBLUP estimator is design consistent and is therefore

more robust to model misspecification. We will compare the performance of the EBLUP and pseudo-EBLUP estimators through a simulation study.

3 Area level model

The Fay-Herriot model (Fay and Herriot 1979) is a basic area level model widely used in small area estimation to improve the direct survey estimates. The Fay-Herriot model has two components, namely, a sampling model for the direct survey estimates and a linking model for the small area parameters of interest. The sampling model assumes that given the area-specific sample size $n_i > 1$, there exists a direct survey estimator $\hat{\theta}_i^{\mathrm{DIR}}$. The direct survey estimator is design unbiased for the small area parameter θ_i. The sampling model is given by

$$\hat{\theta}_i^{\mathrm{DIR}} = \theta_i + e_i,\ i = 1,\ldots,m, \tag{3.1}$$

where the e_i is the sampling error associated with the direct estimator $\hat{\theta}_i^{\mathrm{DIR}}$ and m is the number of small areas. It is customary in practice to assume that the e_i's are independently normal random variables with mean $E(e_i) = 0$ and sampling variance $\mathrm{var}(e_i) = \sigma_i^2$. The linking model is obtained by assuming that the small area parameter of interest θ_i is related to area level auxiliary variables $\mathbf{z}_i = (\mathrm{z}_{i1},\ldots,\mathrm{z}_{ip})'$ through the following linear regression model

$$\theta_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i,\ i = 1,\ldots,m, \tag{3.2}$$

where $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)'$ is a $p \times 1$ vector of regression coefficients, and the v_i's are area-specific random effects assumed to be $i.i.d.$ with $E(v_i) = 0$ and $\mathrm{var}(v_i) = \sigma_v^2$. The assumption of normality is generally also made, even though it is more difficult to justify the assumption. This assumption is needed to obtain the MSE estimation. The model variance σ_v^2 is unknown and needs to be estimated from the data. The area level random effect v_i capture the unstructured heterogeneity among areas that is not explained by the sampling variances. Combining models (3.1) and (3.2) leads to a linear mixed area level model given by

$$\hat{\theta}_i^{\mathrm{DIR}} = \mathbf{z}_i'\boldsymbol{\beta} + v_i + e_i. \tag{3.3}$$

Model (3.3) involves both design-based random errors e_i and model-based random effects v_i. For the Fay-Herriot model, the sampling variance σ_i^2 is assumed to be known in model (3.3). This is a very strong assumption. Generally smoothed estimators of the sampling variances are used in the Fay-Herriot model and then σ_i^2's are treated as known. However, if direct estimators of sampling variances are used in the Fay-Herriot model, an extra term needs to be added to the MSE estimator to account for the extra variation (Wang and Fuller 2003).

Assuming that the model variance σ_v^2 is known, the best linear unbiased predictor (BLUP) of the small area parameter θ_i can be obtained as

$$\tilde{\theta}_i = \gamma_i\hat{\theta}_i^{\mathrm{DIR}} + (1-\gamma_i)\mathbf{z}_i'\tilde{\boldsymbol{\beta}}_{\mathrm{WLS}}, \tag{3.4}$$

where $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_i^2)$, and $\tilde{\boldsymbol{\beta}}_{\mathrm{WLS}}$ is the weighted least squared (WLS) estimator of $\boldsymbol{\beta}$ given by

$$\tilde{\boldsymbol{\beta}}_{\mathrm{WLS}} = \left[\sum_{i=1}^{m}\left(\sigma_i^2+\sigma_v^2\right)^{-1}\mathbf{z}_i\mathbf{z}_i'\right]^{-1}\left[\sum_{i=1}^{m}\left(\sigma_i^2+\sigma_v^2\right)^{-1}\mathbf{z}_i y_i\right] = \left[\sum_{i=1}^{m}\gamma_i\mathbf{z}_i\mathbf{z}_i'\right]^{-1}\left[\sum_{i=1}^{m}\gamma_i\mathbf{z}_i y_i\right].$$

There are several methods available to estimate the unknown model variance σ_v^2; You (2010) provides a review of these methods. We chose the restricted maximum likelihood (REML) obtained by Cressie (1992) to estimate the model variance under the Fay-Herriot model. Using the scoring algorithm, the REML estimator $\hat{\sigma}_v^2$ is obtained as

$$\sigma_v^{2(k+1)} = \sigma_v^{2(k)} + \left[I_R\left(\sigma_v^{2(k)}\right)\right]^{-1} S_R\left(\sigma_v^{2(k)}\right), \quad \text{for } k = 1,2,\ldots,$$

where $I_R\left(\sigma_v^2\right) = 1/2\,\mathrm{tr}[\mathbf{PP}]$, and $S_R\left(\sigma_v^2\right) = 1/2\,\mathbf{y}'\mathbf{PPy} - 1/2\,\mathrm{tr}[\mathbf{P}]$, and $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{V}^{-1}$. Using a guessing value for $\sigma_v^{2(1)}$ as the starting value, the algorithm converges very fast.

Replacing σ_v^2 in equation (3.4) by the REML estimator $\hat{\sigma}_v^2$, we obtain the EBLUP of the small area parameter θ_i based on the Fay-Herriot model as

$$\hat{\theta}_i^{\mathrm{FH}} = \hat{\gamma}_i\hat{\theta}_i^{\mathrm{DIR}} + (1-\hat{\gamma}_i)\,\mathbf{z}_i'\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}, \tag{3.5}$$

where $\hat{\gamma}_i = \hat{\sigma}_v^2\big/\left(\hat{\sigma}_v^2+\sigma_i^2\right)$. The MSE estimator of $\hat{\theta}_i^{\mathrm{FH}}$ is given by (see Rao 2003)

$$\mathrm{mse}\left(\hat{\theta}_i^{\mathrm{FH}}\right) = g_{1i} + g_{2i} + 2g_{3i}, \tag{3.6}$$

where g_{1i} is the leading term, g_{2i} accounts for the variability due to estimation of the regression parameter β, and g_{3i} is due to the estimation of the model variance. These $g-$terms are defined as follow:

$$g_{1i} = \hat{\gamma}_i\sigma_i^2, g_{2i} = (1-\hat{\gamma}_i)^2\,\mathbf{z}_i'\mathrm{var}\left(\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}\right)\mathbf{z}_i = \hat{\sigma}_v^2\,(1-\hat{\gamma}_i)^2\,\mathbf{z}_i'\left(\sum_{i=1}^{m}\hat{\gamma}_i\mathbf{z}_i\mathbf{z}_i'\right)^{-1}\mathbf{z}_i$$

and $g_{3i} = \left(\sigma_i^2\right)^2\left(\hat{\sigma}_v^2+\sigma_i^2\right)^{-3}\mathrm{var}\left(\hat{\sigma}_v^2\right)$.

The estimated variance of $\hat{\sigma}_v^2$ is given by $\mathrm{var}\left(\hat{\sigma}_v^2\right) = 2\left(\sum_{i=1}^{m}\left(\hat{\sigma}_v^2+\sigma_i^2\right)^{-2}\right)^{-1}$; see Datta and Lahiri (2000).

Up to now we have assumed that the sampling variance σ_i^2 is assumed known in the Fay-Herriot model (3.3). This is a very strong assumption. Usually a direct survey estimator, say s_i^2, of the sampling variance σ_i^2 is available. As these estimated variances can be quite variable, they are smoothed using external models and generalized variance functions: these smoothed variances are denoted as $\tilde{s}_i^2$. The smoothed sampling variance estimates $\tilde{s}_i^2$ are used in the Fay-Herriot model and treated as known. The associated $\mathrm{mse}\left(\hat{\theta}_i^{\mathrm{FH}}\right)$ is obtained by replacing σ_i^2 by $\tilde{s}_i^2$ in equation (3.6). Rivest and Vandal (2003) and Wang and Fuller (2003) considered the small area estimation using the Fay-Herriot model with the direct sampling variance estimates s_i^2 under the assumption that the estimators s_i^2 are independent of the direct survey estimators y_i and $d_i s_i^2 \sim \sigma_i^2\chi_{d_i}^2$, where $d_i = n_i - 1$ and n_i is the sample size for the i^{th} area. When the direct sampling variance estimate s_i^2 is used in the place of the true sampling variance σ_i^2, an extra term accounts for the uncertainty of using s_i^2 is needed in the MSE estimator (3.6), and this term, denoted as g_{4i}, is given by

$$g_{4i} = \frac{4}{n_i - 1} \frac{\hat{\sigma}_v^4 s_i^4}{\left(\hat{\sigma}_v^2 + s_i^2\right)^3};$$

see Rivest and Vandal (2003) and Wang and Fuller (2003) for details.

To apply the Fay-Herriot model, we need to obtain area level direct estimates and the corresponding sampling variance estimates as input values for the Fay-Herriot model. We consider three area level direct estimators; namely, the direct sample mean estimator assuming simple random sampling (SRS), the Horvitz-Thompson estimator (HT), and the weighted Hájek estimator (HA). The weighted Hájek estimator is also used in the pseudo-EBLUP estimator for the unit level model denoted as $\bar{y}_{iw}$ in equation (2.7). Table 3.1 presents these three area level direct estimators and the corresponding sampling variance estimators.

Table 3.1
Area level direct estimators and sampling variances

	Point estimator	Sampling variance estimator
Direct mean (SRS)	$\hat{\theta}_i^{\text{SRS}} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$	$\text{var}\left(\hat{\theta}_i^{\text{SRS}}\right) = \frac{1}{n_i(n_i-1)}\sum_{j=1}^{n_i}\left(y_{ij} - \hat{\theta}_i^{\text{SRS}}\right)^2$
Horvitz-Thompson (HT) estimator	$\hat{\theta}_i^{\text{HT}} = \frac{1}{N_i}\sum_{j=1}^{n_i} w_{ij} y_{ij} = \frac{1}{N_i}\sum_{j=1}^{n_i}\frac{y_{ij}}{n_i p_{ij}}$	$\text{var}\left(\hat{\theta}_i^{\text{HT}}\right) = \frac{1}{N_i^2 n_i(n_i-1)}\sum_{j=1}^{n_i}\left(\frac{y_{ij}}{p_{ij}} - N_i\hat{\theta}_i^{\text{HT}}\right)^2$
Weighted Hájek (HA) estimator	$\hat{\theta}_i^{\text{HA}} = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}} = \frac{1}{\hat{N}_i}\sum_{j=1}^{n_i}\frac{y_{ij}}{n_i p_{ij}}$	$\text{var}\left(\hat{\theta}_i^{\text{HA}}\right) = \frac{1}{\hat{N}_i^2 n_i(n_i-1)}\sum_{j=1}^{n_i}\left(\frac{y_{ij} - \hat{\theta}_i^{\text{HA}}}{p_{ij}}\right)^2$

These area level estimators are used as input values into the Fay-Herriot model. Correspondingly, the three area level model-based estimators are denoted as: FH-SRS, FH-HT, and FH-HA. That is, we replace $\hat{\theta}_i^{\text{DIR}}$ by $\hat{\theta}_i^{\text{SRS}}$, $\hat{\theta}_i^{\text{HT}}$ or $\hat{\theta}_i^{\text{HA}}$ in (3.5) and obtain the corresponding model-based estimator $\hat{\theta}_i^{\text{FH-SRS}}$, $\hat{\theta}_i^{\text{FH-HT}}$ and $\hat{\theta}_i^{\text{FH-HA}}$. The SRS direct estimator $\hat{\theta}_i^{\text{SRS}}$ ignores the sample design and is not design consistent, unless the sample design is based on simple random sampling. Note that $\hat{\theta}_i^{\text{HT}}$ and $\hat{\theta}_i^{\text{HA}}$ are design consistent estimators. It follows that the corresponding model-based estimators $\hat{\theta}_i^{\text{FH-HT}}$ and $\hat{\theta}_i^{\text{FH-HA}}$ are design consistent as the sample size increases. Furthermore, this means that these estimators are robust to model misspecification.

In the next section, we compare the unit level model with the Fay-Herriot model through a simulation study. The statistics used for these comparisons are bias, relative root MSE and confidence intervals of the model-based estimators.

4 Simulation study

4.1 Data generation

To compare the unit level and area level small area estimators, we conducted a design-based simulation study. Following the simulation setup of You, Rao and Kovacevic (2003), we created two finite populations. Each finite population had $m = 30$ areas, and each area consisted of $N_i = 200$ population units. Each finite

population was generated using the unit level model $y_{ij} = \beta_0 + x_{1ij}\beta_1 + v_i + e_{ij}$. The auxiliary variable x_{1ij} was generated from an exponential distribution with mean 4 and variance 8, and the random components were generated from the normal distribution with $v_i \sim N\left(0, \sigma_v^2\right)$, $e_{ij} \sim N\left(0, \sigma_e^2\right)$, where $\sigma_v^2 = 100$ and $\sigma_e^2 = 225$. For the first population, the regression fixed effects were set as $\beta_0 = 50$, $\beta_1 = 10$ for all 30 areas. For the second population, different fixed effects values were used: $\beta_0 = 50$, $\beta_1 = 10$ for areas $m = 1,\ldots,10$; $\beta_0 = 75$, $\beta_1 = 15$ for areas $m = 11,\ldots,20$; $\beta_0 = 100$, $\beta_1 = 20$ for areas $m = 21,\ldots,30$. We had three different means for the fixed effects $\beta_0 + x_{1ij}\beta_1$ in the second population, whereas we only had one in the first population. PPSWR samples within each area were drawn independently from each constructed population. PPSWR sampling was implemented as follows: We first defined a size measure z_{ij} for a given unit (i, j). Using these z_{ij} values, we computed selection probabilities $p_{ij} = z_{ij} \Big/ \sum_j z_{ij}$ for each unit (i, j) and used them to select PPSWR samples of equal size $n_i = n$. Within each generated population, we selected samples of size $n = 10$ and 30. The basic design weight is given by $w_{ij} = n_i^{-1} p_{ij}^{-1}$, so that the standardized weight is $\tilde{w}_{ij} = p_{ij}^{-1} \Big/ \sum_j p_{ij}^{-1}$. We chose the size measure z_{ij} as a linear combination of the auxiliary variable x_{1ij} and data generated from an exponential distribution with mean 4 and variance 16. The correlation coefficient ρ between y_{ij} and the selection probability p_{ij} within each area varied between 0.02 and 0.95. The range of the p_{ij}'s corresponds to non-informative selection $(\rho = 0.02)$ to strongly informative selection $(\rho = 0.95)$ of the PPSWR samples. The sampling is non-informative when the correlation coefficient between y_{ij} and the selection probability p_{ij} is very weak, implying that the sample and the population model coincide. If the selection probability p_{ij} is strongly correlated with the observation y_{ij}, we have informative sampling, and the population model may no longer holds for the sample. For each population, the PPSWR sampling process was repeated $R = 3{,}000$ times. As in Prasad and Rao (1990), the simulation study is design-based as both the populations were generated only once, and repeated samples were generated from the same population.

For unit level modeling, we fitted the nested error regression model to the PPSWR sampling data generated from each population. We obtained the corresponding EBLUP and pseudo-EBLUP estimates and related MSE estimates using the formulas given in Section 2. We then constructed the confidence interval estimates using the squared root of the MSE estimates; details are given in Section 4.2.3. For area level modeling, we first obtained the direct area level estimates $\hat{\theta}_i^{\text{SRS}}$, $\hat{\theta}_i^{\text{HT}}$ and $\hat{\theta}_i^{\text{HA}}$ as well as the corresponding sampling variances. We applied the Fay-Herriot model and obtained the model-based estimators $\hat{\theta}_i^{\text{FH-SRS}}$, $\hat{\theta}_i^{\text{FH-HT}}$ and $\hat{\theta}_i^{\text{FH-HA}}$. The population mean of the auxiliary variable x_{1ij} within each area was used in the Fay-Herriot model as the auxiliary variable. The g_{4i} was added to the MSE estimator to account for the use of unsmoothed sampling variances in the Fay-Herriot model. The corresponding confidence intervals were obtained similarly for the unit level EBLUP and pseudo-EBLUP estimators.

For both unit level and area level model fitting, we used the following two scenarios: Scenario I: correct modeling, where the data were generated from the first population and the fitting models were unit level model (2.2) and area level model (3.3) with common $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Scenario II: incorrect modeling, where the data were generated from the second population with different means for the fixed effects, and the fitting models were the same as in Scenario (I) with common $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Note that under scenario I the

sampling is noninformative when the correct unit level (2.2) is fitted to the sample data to obtain the EBLUP estimator: this is true for any correlation coefficient ρ between y_{ij} and p_{ij}.

4.2 Results

In this section, we compare a number of statistics for the unit level and area level estimates under both scenario I (correct modeling) and scenario II (incorrect modeling).

4.2.1 Comparison within each small area

Figure 4.1 compares the population means with the unit level and area level estimates when $n = 10$ for scenario I. The results are based on a strongly informative sampling design where the correlation coefficient between y_{ij} and the selection probability p_{ij} is $\rho = 0.88$. The model-based estimates are based on the average of $R = 3{,}000$ simulation runs. It is clear from Figure 4.1 that the unit level estimators EBLUP (equation 2.5) and pseudo-EBLUP (equation 2.10) are almost unbiased. The results show that under correct modeling, the sampling is noninformative with respect to unit level model (2.2), and the EBLUP is unbiased. The area level estimator FH-SRS consistently overestimates the population mean, leading to a large bias. The area level estimator FH-HT generally underestimates the population mean and has slightly larger bias than the FH-HA estimator. For $n = 30$, we obtained similar results.

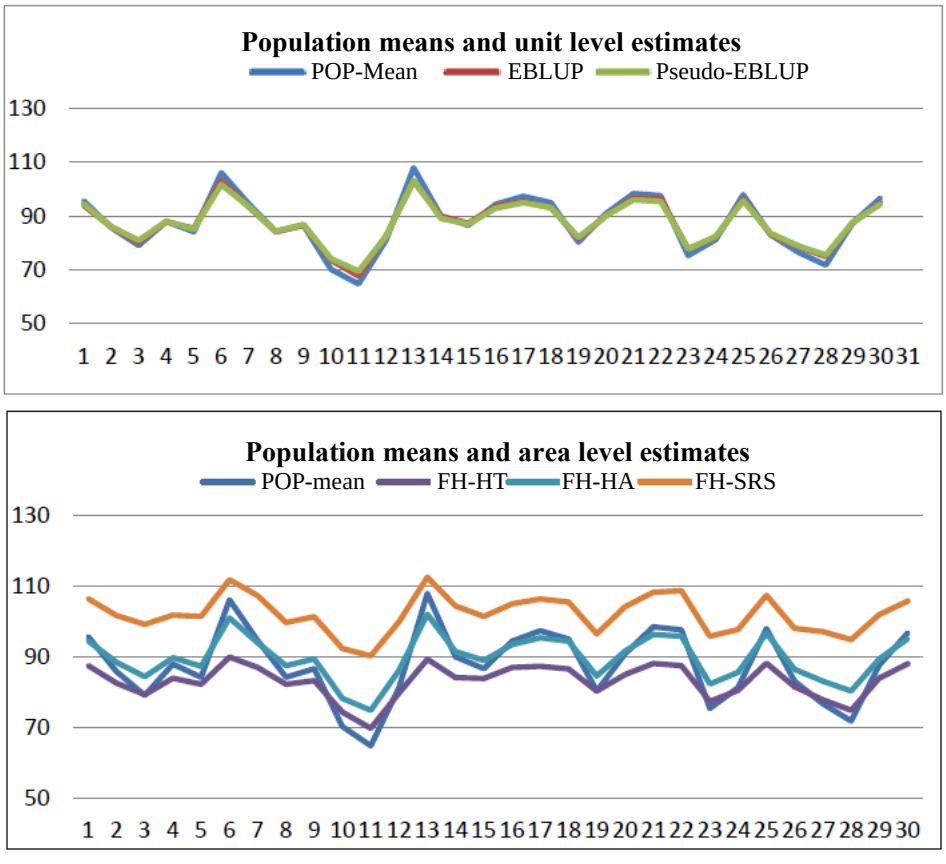

Figure 4.1 Comparison of means under scenario I: $n = 10$.

Figure 4.2 compares the average root mse for both unit level and area level estimators for scenario I when $n = 10$ and $n = 30$. The root mse's are the squared root of the estimated MSE's given in Sections 2 and 3 for the unit level and area level estimators. It is clear that EBLUP and pseudo-EBLUP have much smaller root mse's than the FH area level estimators for both $n = 10$ and $n = 30$. As expected (You and Rao 2002), EBLUP has the smallest root mse and pseudo-EBLUP has slightly larger root mse. For area level estimators, FH-SRS has large root mse and large variations. FH-HT and FH-HA have on average about the same root mse, but FH-HT is more variable than FH-HA as shown in both figures, particularly when sample size $n = 10$. When the sample size $n = 30$, the variability of the root mse's for FH-HT and FH-HA are substantially reduced, but it is clear that FH-HA is more stable than FH-HT.

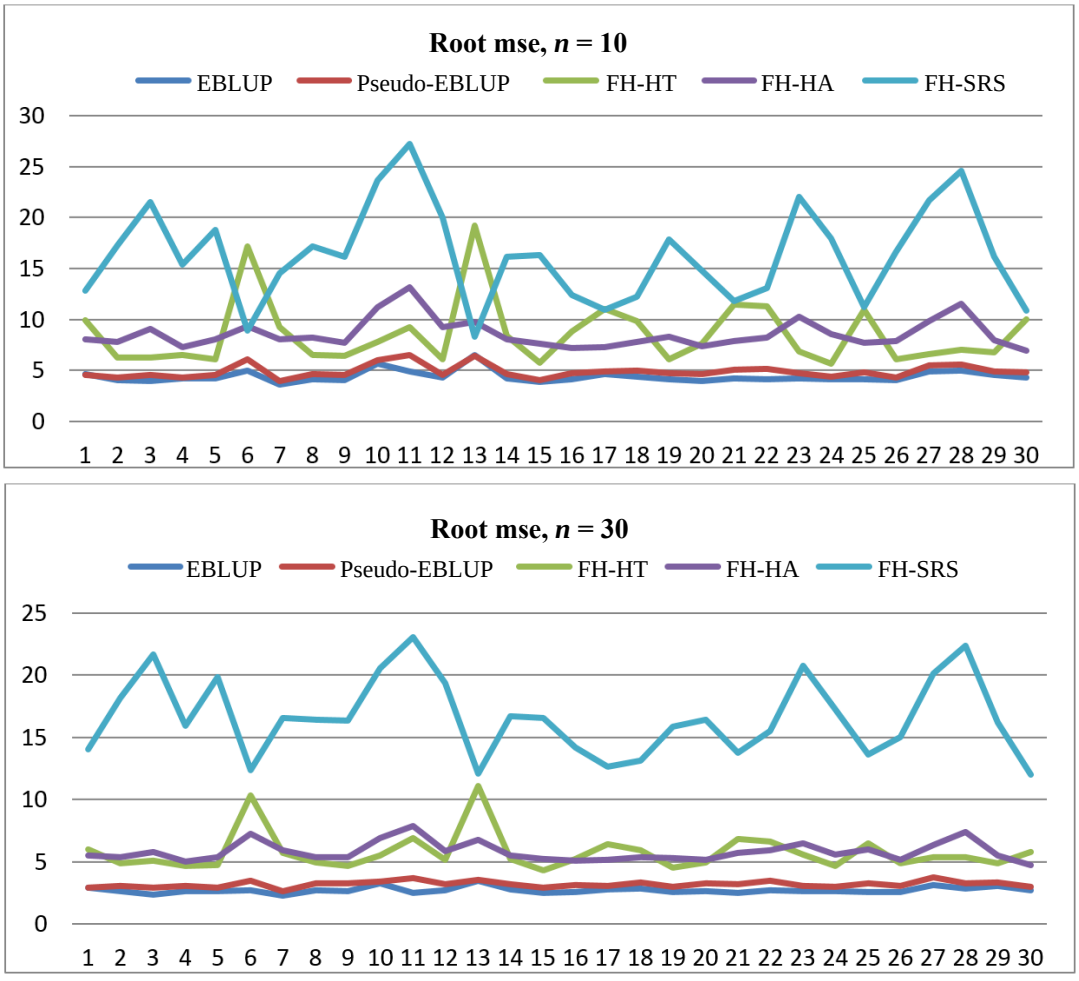

Figure 4.2 Comparison of root mse under scenario I: $n = 10$ and $n = 30$.

Figure 4.3 compares the unit level and area level estimates with the population means when $n = 10$ under scenario II. For unit level models, it is clear that EBLUP both underestimates and overestimates the population mean when the model is misspecified, whereas pseudo-EBLUP is unbiased (the pseudo-EBLUP

estimates and population means overlap in Figure 4.3). For area level estimators, FH-SRS consistently overestimates the true means, while FH-HT has more underestimation than FH-HA as shown when the model is misspecified.

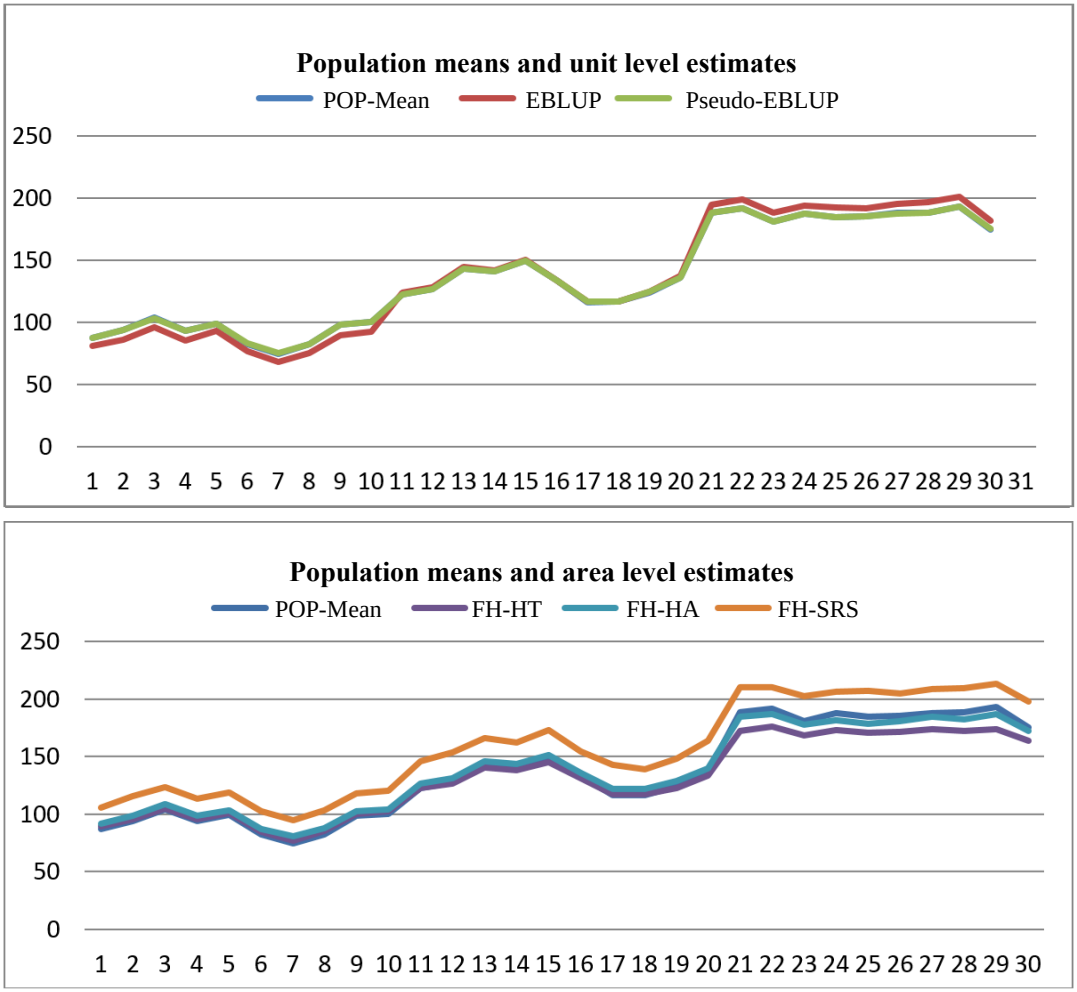

Figure 4.3 Comparison of means under scenario II: incorrect modeling, $n = 10$.

Figure 4.4 compares the root mse's of the unit level and area level estimators for both sample size $n = 10$ and $n = 30$ under incorrect modeling. From Figure 4.4, it can be seen that the pseudo-EBLUP estimator has the smallest root mse under incorrect modeling. EBLUP has very large root mse when the model is misspecified: that is, for areas 1 to 10 and areas 21 to 30, the average root mse is 10.01, whereas for pseudo-EBLUP, the corresponding root mse is 7.38 when the sample size $n = 10$. When the sample size $n = 30$, the average root mse is 8.85 for EBLUP, and only 4.38 for pseudo-EBLUP when the model is misspecified. In summary, the results show that the EBLUP estimator leads to biased estimates with large root mse under incorrect modeling.

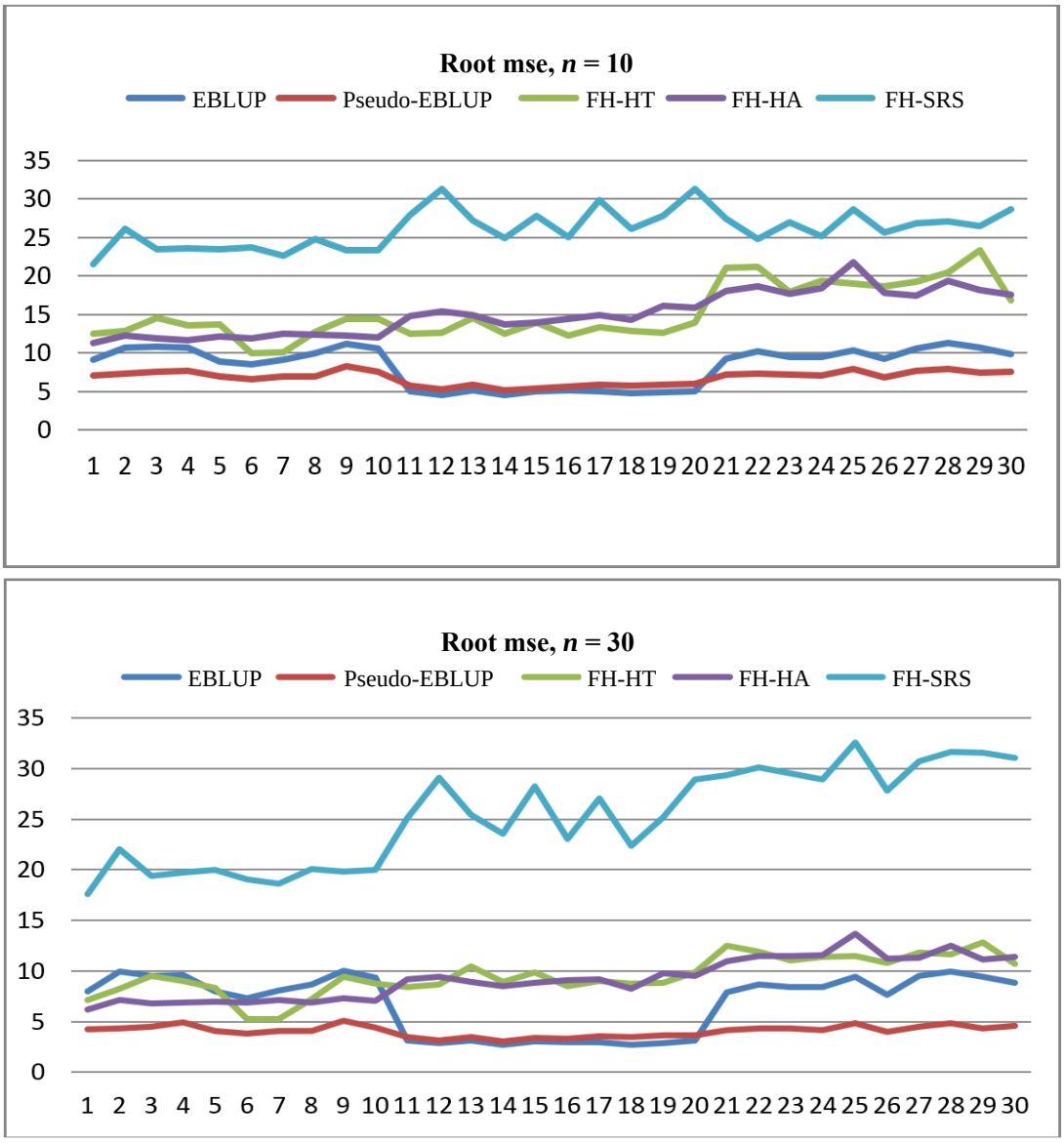

Figure 4.4 Comparison of root mse under scenario II: $n = 10$ and $n = 30$.

4.2.2 Comparison across small areas

To compare the estimators across areas, we considered the average absolute relative bias (ARB) for a specified estimator $\hat{\theta}_i$ of the simulated population mean $\bar{Y}_i$ as $\overline{\text{ARB}} = \left(\sum_{i=1}^{m} \text{ARB}_i\right)\Big/ m$, where

$$\text{ARB}_i = \left| \frac{1}{R} \sum_{r=1}^{R} \frac{\left(\hat{\theta}_i^{(r)} - \bar{Y}_i\right)}{\bar{Y}_i} \right|,$$

and $\hat{\theta}_i^{(r)}$ is the estimate based on the r^{th} simulated sample, $R = 3{,}000, m = 30$. Table 4.1 displays the percentage of the average absolute relative bias $\overline{\text{ARB}}$ of unit level and area level estimators over the 30 area for scenario I. The results are based on samples selected with sample sizes equal to 10 and 30 respectively within each area.

Table 4.1
Average absolute relative bias $\overline{\text{ARB}}\%$ for scenario I

Type	Estimator	$n = 10$	$n = 30$
Unit level	EBLUP	1.71	0.75
	Pseudo-EBLUP	2.14	0.86
Area level	FH-SRS	17.51	18.64
	FH-HT	6.02	3.12
	FH-HA	4.33	2.59

For unit level models, it is clear that if we use the correct model, the sample becomes noninformative with respect to unit level model (2.2), and both EBLUP and pseudo-EBLUP estimators are unbiased. The average absolute relative bias $\overline{\text{ARB}}$ for EBLUP is 1.71% when the sample size $n = 10$ and 0.75% when the sample size $n = 30$. For pseudo-EBLUP, the $\overline{\text{ARB}}$ is 2.14% when $n = 10$ and 0.86% when $n = 30$, respectively. Pseudo-EBLUP has slightly larger bias than EBLUP. For area level models, FH-SRS severely overestimates the means with the average $\overline{\text{ARB}}$ as large as 17.51% when $n = 10$ and 18.6% when $n = 30$. Both area level estimators FH-HT and FH-HA lead to reasonable estimates: (i) The $\overline{\text{ARB}}$ for FH-HT is 6.02% when $n = 10$ and 3.12% when $n = 30$; (ii) The $\overline{\text{ARB}}$ for FH-HA is 4.33% when $n = 10$ and 2.59% when $n = 30$. The FH-HA estimator performs better than the FH-HT estimator. The absolute relative bias for the area level estimators is larger than the one associated with the unit level estimators.

Table 4.2 displays the $\overline{\text{ARB}}$ of the various estimators under scenario II. It is clear that pseudo-EBLUP has a much smaller $\overline{\text{ARB}}$ than EBLUP under incorrect modeling. The $\overline{\text{ARB}}$'s for EBLUP under incorrect modeling are 4.31% $(n = 10)$ and 4.52% $(n = 30)$ respectively. For pseudo-EBLUP, the average $\overline{\text{ARB}}$ is only 0.25% $(n = 10)$ and 0.12% $(n = 30)$. Both FH-HT and FH-HA perform very well. Their average $\overline{\text{ARB}}$'s are 3.91% and 3.48% respectively when $n = 10$. These $\overline{\text{ARB}}$'s decrease to 1.51% and 1.47% when $n = 30$. FH-SRS performs poorly. Both area level estimators FH-HT and FH-HA perform well and these estimators are also design consistent. Again, FH-HA is slightly better than FH-HT in terms of $\overline{\text{ARB}}$. The results show that the use of survey weights in the unit level modeling is very important when the unit level model is incorrectly specified. The pseudo-EBLUP estimator leads to unbiased estimator even when the model is incorrectly specified. It is the best estimator when the model is incorrect.

Table 4.2
Average absolute relative bias $\overline{\text{ARB}}\%$ for scenario II

Type	Estimator	$n = 10$	$n = 30$
Unit level	EBLUP	4.31	4.52
	Pseudo-EBLUP	0.25	0.12
Area level	FH-SRS	17.11	17.87
	FH-HT	3.91	1.51
	FH-HA	3.48	1.47

We now compare the relative root MSE for all the estimators. In particular, we computed both the true simulation relative root MSE (RRMSE) and the estimated relative root MSE based on the MSE estimators. The average true simulation relative root MSE is computed as $\overline{\text{RRMSE}} = \left(\sum_{i=1}^{m} \text{RRMSE}_i\right) \Big/ m$, where

$$\text{RRMSE}_i = \frac{\sqrt{\text{MSE}_i}}{\bar{Y}_i}, \text{ and } \text{MSE}_i = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_i^{(r)} - \bar{Y}_i\right)^2.$$

The average estimated relative root MSE is computed as $\overline{\text{RRmse}} = \left(\sum_{i=1}^{m} \text{RRmse}_i\right)\Big/ m$, where

$$\text{RRmse}_i = \frac{\sqrt{\text{mse}_i}}{\hat{\theta}_i}, \text{ and } \text{mse}_i = \frac{1}{R}\sum_{r=1}^{R}\text{mse}_i^{(r)}, \text{ and } \hat{\theta}_i = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_i^{(r)}.$$

The $\text{mse}_i^{(r)}$ is the estimated MSE of $\hat{\theta}_i^{(r)}$ for the i^{th} area. They are computed using the formulas given in Sections 2 and 3.

Table 4.3 reports the average $\overline{\text{RRMSE}}$ and $\overline{\text{RRmse}}$ over the 30 small areas. When the sample size $n = 10$, $\overline{\text{RRMSE}}$ is 4.98% for EBLUP and 5.49% for the pseudo-EBLUP respectively. As expected (You and Rao 2002), the pseudo-EBLUP has a slightly larger RRMSE than the one associated with EBLUP. Both the unit level EBLUP and pseudo-EBLUP estimators have much smaller RRMSE's than the area level estimators. For area level models, FH-HT and FH-HA perform similarly, with corresponding average true RRMSE equal to 9.72% and 9.68% respectively, when $n = 10$. The FH-SRS performs poorly under informative sampling with the average true RRMSE equal to 18.89% when $n = 10$. Even when $n = 30$, the average RRMSE for FH-SRS is as large as 18.62%. Note that $\overline{\text{RRmse}}$ is very close to its true value.

In summary, the results in Table 4.3 show that the unit level estimators EBLUP and pseudo-EBLUP perform better than the area level estimators FH-HT and FH-HA under correct modeling. Both the area level estimators FH-HT and FH-HA perform reasonably well under informative sampling. As expected, FH-SRS performs poorly.

Table 4.3
Average RRMSE% for scenario I

		$n = 10$		$n = 30$	
Type	**Estimator**	$\overline{\text{RRMSE}}$	$\overline{\text{RRmse}}$	$\overline{\text{RRMSE}}$	$\overline{\text{RRmse}}$
Unit level	EBLUP	4.98	5.09	3.01	3.13
	Pseudo-EBLUP	5.49	5.66	3.58	3.67
Area level	FH-SRS	18.89	17.53	18.62	16.34
	FH-HT	9.72	10.25	6.67	6.69
	FH-HA	9.68	9.71	6.51	6.63

Table 4.4 displays the results of the average RRMSE under scenario II. The pseudo-EBLUP is the most robust estimator and has the smallest $\overline{\text{RRMSE}}$: the $\overline{\text{RRMSE}}$'s are 5.42% and 3.21% for $n = 10$ and $n = 30$ respectively. For the area level estimators, FH-HT and FH-HA perform similarly, whereas FH-SRS performs poorly. When $n = 10$, $\overline{\text{RRMSE}}$ for FH-HT is 11.68% and 11.21% for FH-HA. When $n = 30$, $\overline{\text{RRMSE}}$ is 7.24% for FH-HT and 6.79% for FH-HA. As expected, FH-SRS has large $\overline{\text{RRMSE}}$ under informative sampling. The pseudo-EBLUP performs the best in terms of bias, standard errors and RRMSE under model misspecification. FH-HA is slightly better than FH-HT. The estimated $\overline{\text{RRmse}}$ is very close to the true $\overline{\text{RRMSE}}$ for all estimators.

Table 4.4
Average RRMSE% for scenario II

Type	Estimator	$n = 10$		$n = 30$	
		$\overline{\text{RRMSE}}$	$\overline{\text{RRmse}}$	$\overline{\text{RRMSE}}$	$\overline{\text{RRmse}}$
Unit level	EBLUP	6.78	6.94	5.62	5.81
	Pseudo-EBLUP	5.42	5.45	3.21	3.26
Area level	FH-SRS	19.76	17.43	19.06	16.24
	FH-HT	11.68	11.78	7.24	7.26
	FH-HA	11.21	11.27	6.79	6.91

4.2.3 Comparison of confidence intervals

We now compare the confidence intervals associated with the unit level and area level estimators. The confidence interval is in the form $\text{estimator} \pm z_{\alpha/2}\sqrt{\text{mse}}$, with $z_{\alpha/2}$ denoting the $100(1-\alpha/2)\%$ percentile of the standard normal distribution. For example, the 95% confidence interval of the EBLUP estimator $\hat{\theta}_i^{\text{EBLUP}}$ is obtained as $\hat{\theta}_i^{\text{EBLUP}} \pm 1.96\sqrt{\text{mse}\left(\hat{\theta}_i^{\text{EBLUP}}\right)}$, where $\text{mse}\left(\hat{\theta}_i^{\text{EBLUP}}\right)$ is given by (2.6). The confidence intervals are computed as follows. For a given estimator $\hat{\theta}_i^{(r)}$, $r = 1,\ldots,R$, $i = 1,\ldots,m$, define the indicator variable $I_i^{(r)}$ as:

$$I_i^{(r)} = \begin{cases} 1 \text{ if } \theta_i \subseteq \left(\hat{\theta}_i^{(r)} - 1.96\sqrt{\text{mse}\left(\hat{\theta}_i^{(r)}\right)}, \hat{\theta}_i^{(r)} + 1.96\sqrt{\text{mse}\left(\hat{\theta}_i^{(r)}\right)}\right) \\ 0 \text{ otherwise} \end{cases}.$$

The confidence interval coverage rate is obtained as the average of $I_i^{(r)}$ over the $R = 3{,}000$ simulations. Tables 4.5 and 4.6 present the 95% confidence interval coverage rates for the unit level and area level estimators under scenario I. The correlation coefficient ρ between the selection probabilities p_{ij} and y_{ij} is presented in the first column to reflect the strength of informativeness of the PPS sampling.

Table 4.5
Confidence interval coverage rates under scenario I: $n = 10$

Correlation coefficient (ρ)	EBLUP	Pseudo-EBLUP	FH-SRS	FH-HT	FH-HA
0.95	0.932	0.946	0.618	0.898	0.911
0.88	0.945	0.948	0.649	0.882	0.908
0.75	0.948	0.948	0.705	0.863	0.911
0.51	0.944	0.949	0.825	0.845	0.916
0.28	0.947	0.951	0.901	0.822	0.917
0.12	0.948	0.949	0.924	0.778	0.893
0.02	0.948	0.951	0.925	0.595	0.886
Average rate	*0.945*	*0.949*	*0.792*	*0.812*	*0.906*

We first discuss the coverage properties associated with the unit level estimators EBLUP and pseudo-EBLUP. These tables show that, when the model is correct, the coverage rates for EBLUP and pseudo-EBLUP are quite stable: the pseudo-EBLUP has slightly better coverage rate than EBLUP. When the sample

size $n = 10$, the average coverage rate for EBLUP is 94.5%, and 94.9% for pseudo-EBLUP. When the sample size $n = 30$, it is 93.4% for EBLUP and 94.8% for pseudo-EBLUP. As the sample size increases from $n = 10$ to 30, the coverage rates for EBLUP deteriorate slightly more than those associated with the pseudo-EBLUP. The pseudo-EBLUP estimator is not as much affected by the degree of informativeness caused by the PPS sampling. The relatively stable coverage rates for EBLUP show that the sample is noninformative with respect to the correct unit level model. However, when $n = 30$, EBLUP has slightly lower coverage rate.

Table 4.6
Confidence interval coverage rates under scenario I: $n = 30$

Correlation coefficient (ρ)	EBLUP	Pseudo-EBLUP	FH-SRS	FH-HT	FH-HA
0.95	0.905	0.946	0.265	0.932	0.926
0.88	0.938	0.948	0.286	0.915	0.921
0.75	0.941	0.949	0.377	0.911	0.924
0.51	0.940	0.951	0.625	0.895	0.931
0.28	0.941	0.950	0.806	0.874	0.929
0.12	0.939	0.945	0.923	0.866	0.922
0.02	0.937	0.948	0.937	0.772	0.917
Average rate	*0.934*	*0.948*	*0.603*	*0.881*	*0.924*

We now turn to the coverage rates associated with the area level estimators. As expected, FH-SRS has low coverage rates when the sampling is informative, and the coverage rate increases as the sampling design becomes non-informative. FH-HA has better coverage rate than FH-HT. The coverage rate for FH-HT decreases as the sampling design becomes non-informative. For example, when sample size $n = 10$, the coverage rate for FH-HT is only 59.5% when the sampling is non-informative, compared to 88.6% of the coverage rate for FH-HA. As the sample size increases, the coverage rate for FH-HT and FH-HA improves. The average coverage rate for FH-HA is 90.6% when $n = 10$ and 92.4% when $n = 30$. FH-HT has a lower coverage rate than the one associated with FH-HA. The average coverage rate is only 81.2% for FH-HT when $n = 10$. The coverage rate for FH-SRS is very poor, 61.8%, under informative sampling when $n = 10$ and 26.5% when $n = 30$. As the sample size increases, the coverage rate decreases for FH-SRS under informative sampling. As expected, the coverage rate gradually increases for FH-SRS as the sampling becomes non-informative. Among all the estimators, for both sample size $n = 10$ and $n = 30$, the pseudo-EBLUP has the best coverage rate: FH-HA has the second best coverage rate.

Tables 4.7 and 4.8 present the coverage rates under scenario II. The results show that the EBLUP has low coverage rate under informative sampling, whereas the pseudo-EBLUP has very stable and high coverage rates (all around and over 95%) under both the informative and non-informative sampling. For example, when $n = 10$, EBLUP has 84.6% coverage rate under informative sampling (correlation coefficient is 0.95), and when sample size increases to $n = 30$, EBLUP has an even lower coverage rate of 62.9%. The average coverage rate is 90.4% for $n = 10$ and 79.6% for $n = 30$ for EBLUP under incorrect modeling. The results show that EBLUP is sensitive to the modeling when the sampling is informative. This is because EBLUP is completely model-based and ignores the sample design.

Table 4.7
Confidence interval coverage rates under scenario II: $n = 10$

Correlation coefficient (ρ)	EBLUP	Pseudo-EBLUP	FH-SRS	FH-HT	FH-HA
0.95	0.846	0.965	0.701	0.865	0.896
0.88	0.855	0.964	0.729	0.887	0.893
0.75	0.881	0.962	0.787	0.873	0.898
0.51	0.921	0.961	0.872	0.848	0.898
0.28	0.936	0.961	0.912	0.843	0.887
0.12	0.945	0.955	0.917	0.765	0.867
0.02	0.943	0.951	0.913	0.592	0.838
Average rate	*0.904*	*0.959*	*0.833*	*0.811*	*0.883*

Table 4.8
Confidence interval coverage rates under scenario II: $n = 30$

Correlation coefficient (ρ)	EBLUP	Pseudo-EBLUP	FH-SRS	FH-HT	FH-HA
0.95	0.629	0.969	0.239	0.913	0.923
0.88	0.638	0.965	0.275	0.895	0.919
0.75	0.708	0.964	0.406	0.908	0.923
0.51	0.829	0.963	0.701	0.923	0.926
0.28	0.902	0.964	0.854	0.911	0.921
0.12	0.931	0.958	0.921	0.884	0.912
0.02	0.937	0.953	0.918	0.778	0.894
Average rate	*0.796*	*0.962*	*0.616*	*0.887*	*0.918*

Among the three area level estimators, FH-HA performs the best. The coverage rate for FH-HA is very stable, and the average coverage rate for FH-HA is 88.3% when $n = 10$ and 91.8% when $n = 30$. FH-HT has lower coverage rate when the sampling is very non-informative, particularly when sample size $n = 10$. The average coverage rate for FH-HT is only 81.1% when $n = 10$ and 88.7% when $n = 30$. The results show that FH-HA is superior to FH-HT. FH-SRS performs poorly when the sampling is informative, particularly when the sample size $n = 30$. However, FH-SRS performs relatively well when the sampling becomes non-informative. The average coverage rate for FH-SRS is 83.3% when $n = 10$, but only 61.6% when the sample size $n = 30$.

It is clear that pseudo-EBLUP has very high and stable coverage rate under incorrect modeling. FH-HA also has very stable but slightly lower coverage rate. Both EBLUP and FH-SRS have lower coverage rate as the sample size increases, especially when the sampling is informative.

5 Application to real data

In this section, we compare the unit level and area level estimates through a real data analysis. The data set we studied is the corn and soybean data provided by Battese et al. (1988). They considered the estimation of mean hectares of corn and soybeans per segment for twelve counties in north-central Iowa. Among the

twelve counties, there were three counties with a single sample segment. We combined these three counties into a single one, resulting in 10 counties in our data set with sample size n_i ranging from 2 to 5 in each county. The total number of segments N_i (population size) within each county ranged from 402 to 1,505. Following You and Rao (2002), we assumed simple random sampling within each county, and the basic survey weight was computed as $w_{ij} = N_i / n_i$. For unit level modeling, y_{ij} is the number of hectares of corn (or soybean) in the j^{th} segment of the i^{th} county, the auxiliary variables are the number of pixels classified as corn and soybeans as in Battese et al. (1988). We applied the unit level model to the modified data set and obtained the EBLUP and pseudo-EBLUP estimates. For area level modeling, we first obtained the area level direct sample estimates $\hat{\theta}_i^{\text{SRS}}$ based on the SRS sampling. Next, we applied the Fay-Herriot model to the area level direct estimates and obtained the FH-SRS area level estimates. Figure 5.1 compares the area level direct estimates with the model-based unit level and area level estimates. In terms of point estimation, the EBLUP and pseudo-EBLUP estimates are almost identical as in You and Rao (2002). This is because the unit level model is a correct model for these data (Battese et al. 1988). The model-based area level estimates FH-SRS and the area level direct estimates are quite similar in this example.

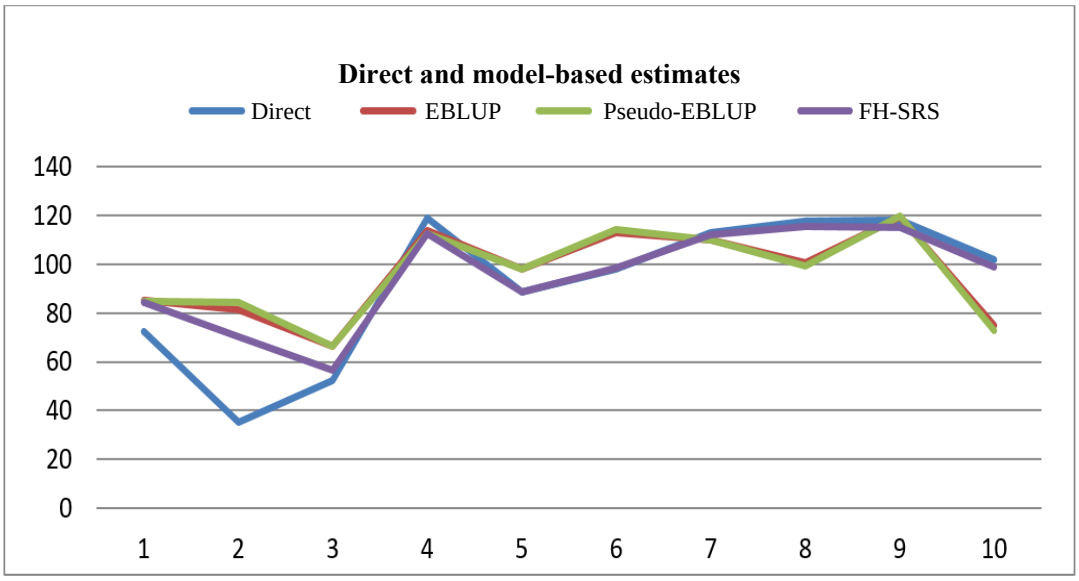

Figure 5.1 Comparison of direct and model-based estimates.

Figure 5.2 compares the standard errors of the direct and model-based estimators. The standard errors of the model-based estimators are the squared root of the estimated MSE. Both the unit level estimators EBLUP and pseudo-EBLUP have small and stable standard errors. As expected, pseudo-EBLUP has slightly larger standard errors than EBLUP. It is clear that the direct and FH-SRS standard errors are very variable and are very unstable. This example shows the effectiveness of the unit level EBLUP and pseudo-EBLUP estimators.

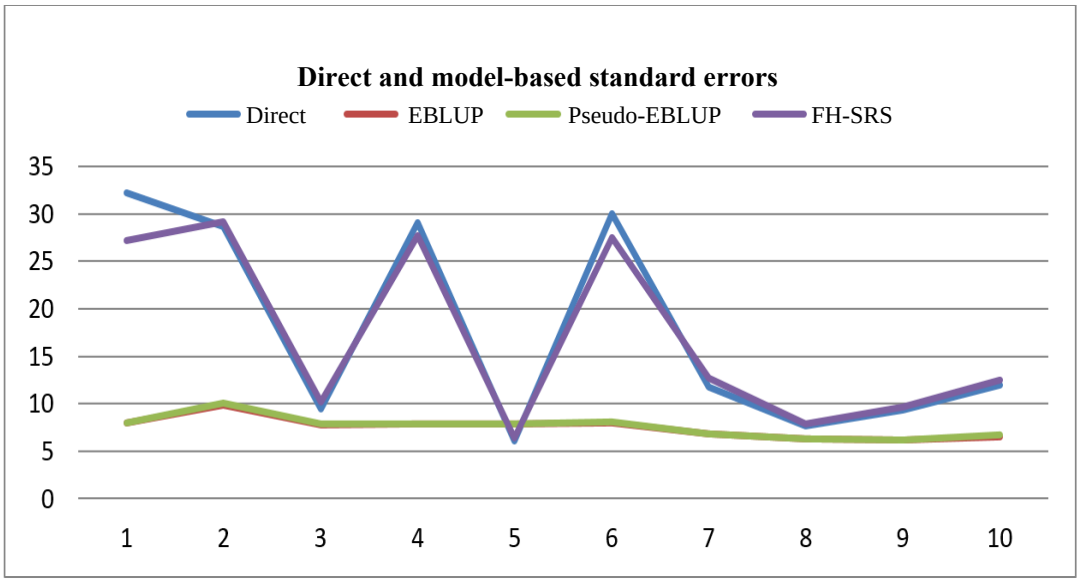

Figure 5.2 Comparison of direct and model-based standard errors.

6 Conclusions

In this paper, we compared performance of the estimators based on the unit level nested error regression model and the area level Fay-Herriot model through a design-based simulation study. We compared the point estimates and coverage rate of confidence intervals of unit level and area level estimators. Overall, the unit level pseudo-EBLUP estimator performs the best in terms of bias and coverage rate under both informative and non-informative sampling. The EBLUP estimator performs well under correct modeling since the sampling is noninformative under correct unit level model (2.2). The pseudo-EBLUP estimator is also quite robust to the model misspecification as well. In practice, we suggest to construct the pseudo-EBLUP estimators using the survey weights and the unit level observations as discussed in Section 2.2. For area level models, FH-HA performs better than FH-HT, and FH-SRS performs poorly. We therefore recommend to construct the weighted HA estimators and then apply the Fay-Herriot model to obtain the corresponding model-based estimators if area level small area estimators are used.

Acknowledgements

The authors would like to thank an Associate Editor and two referees for their suggestions and comments to help to improve the presentation of the results substantially. In particular we would like to thank one referee's very careful and constructive comments.

References

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Cressie, N. (1992). REML estimation in empirical Bayes smoothing of census undercount. *Survey Methodology*, 18, 1, 75-94.

Datta, G.S., and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistics Sinica*, 10, 613-627.

Fay, R.E., and Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 268-277.

Pfeffermann, D., and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102, 480, 1427-1439.

Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rivest, L.-P., and Vandal, N. (2003). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, (Ed., J.N.K. Rao).

Torabi, M., and Rao, J.N.K. (2010). The mean squared error estimators of small area means using survey weights. *The Canadian Journal of Statistics*, 38, 598-608.

Verret, F., Rao, J.N.K. and Hidiroglou, M.A. (2015). Model-based small area estimation under informative sampling. *Survey Methodology*, 41, 2, 333-347.

Wang, J., and Fuller, W.A. (2003). The mean squared error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y. (2010). *Small Area Estimation under the Fay-Herriot Model Using Different Model Variance Estimation Methods and Different Input Sampling Variances*. Methodology branch working paper, SRID-2010-003E, Statistics Canada, Ottawa, Canada.

You, Y., and Rao, J.N.K. (2002). A pseudo empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30, 431-439.

You, Y., Rao, J.N.K. and Kovacevic, M. (2003). Estimating fixed effects and variance components in a random intercept model using survey data. Proceedings: Symposium 2003, *Challenges in Survey Taking for the Next Decade*, Statistics Canada.

Survey Methodology, June 2016
Vol. 42, No. 1, pp. 63-85
Statistics Canada, Catalogue No. 12-001-X

Comparison of some positive variance estimators for the Fay-Herriot small area model

Susana Rubin-Bleuer and Yong You[1]

Abstract

The restricted maximum likelihood (REML) method is generally used to estimate the variance of the random area effect under the Fay-Herriot model (Fay and Herriot 1979) to obtain the empirical best linear unbiased (EBLUP) estimator of a small area mean. When the REML estimate is zero, the weight of the direct sample estimator is zero and the EBLUP becomes a synthetic estimator. This is not often desirable. As a solution to this problem, Li and Lahiri (2011) and Yoshimori and Lahiri (2014) developed adjusted maximum likelihood (ADM) consistent variance estimators which always yield positive variance estimates. Some of the ADM estimators always yield positive estimates but they have a large bias and this affects the estimation of the mean squared error (MSE) of the EBLUP. We propose to use a MIX variance estimator, defined as a combination of the REML and ADM methods. We show that it is unbiased up to the second order and it always yields a positive variance estimate. Furthermore, we propose an MSE estimator under the MIX method and show via a model-based simulation that in many situations, it performs better than other 'Taylor linearization' MSE estimators proposed recently.

Key Words: Variance estimation; Adjusted maximum likelihood; REML; Order of bias; MSE estimation.

1 Introduction

The Fay-Herriot model (Fay and Herriot 1979) is a basic area level model used to estimate small area means, when available direct survey estimates are imprecise due to small sample sizes. In this model, the small area mean is represented by a non-random linear term in the covariates, plus a random area effect. The best linear unbiased prediction (BLUP) estimator of a small area mean, under the Fay-Herriot model, can be obtained by minimizing the mean squared error (MSE) among the class of linear unbiased estimators. The BLUP is a weighted average of the direct survey estimator and the regression-synthetic estimator, with weights depending on the variance of the random area effects, σ_v^2. Usually, this variance has to be estimated from the data under the Fay-Herriot model. The empirical best linear unbiased (EBLUP) estimator of the small area mean is obtained by replacing the variance in the formula of the BLUP with an estimate. There are many well-known methods of variance estimation used in this context but the variance estimator used most often is the restricted maximum likelihood (REML) estimator because it accounts for the loss of degrees of freedom due to estimating the regression coefficient. Furthermore, it is unbiased up to the second order, and it also converges faster in terms of the number of iterations. Despite these important characteristics, occasionally, and particularly when the number of areas, m, is small or moderate, the REML method yields a zero variance estimate. This implies zero weight to the direct survey estimator in the EBLUP formula and hence the EBLUP estimator becomes a regression-synthetic estimator. However, most practitioners are reluctant to use synthetic estimators for small area means, since these ignore the survey based information and are often quite biased. When dealing with real data sets, for which models are never

1. Susana Rubin-Bleuer and Yong You, International Cooperation and Corporate Statistical Methods Division, Statistics Canada. E-mail: susana.rubin-bleuer@canada.ca; yong.you@canada.ca.

perfect, a positive estimate for σ_v^2 reduces the bias of the EBLUP over the synthetic model. Certainly, a positive random effects variance estimate, results in a 'conservative' EBLUP estimator in the sense that it gives a positive weight to the direct survey estimator. Furthermore, it can be viewed as the sum of the regression estimator plus a non-zero term that accounts for part of the 'model bias'. This feature gives rise to a series of variance estimation methods that yield positive estimates.

In this article, we focus on the adjusted likelihood variance estimators developed by Lahiri and Li (2009) and we propose a MIX variance estimator. Our MIX variance estimator is the combination of a REML estimator and any of the adjusted likelihood methods. We also put forward an estimator of the MSE of the EBLUP under the MIX and investigate the theoretical and finite sample properties of both the MIX variance estimator and MSE estimator.

Morris (2006) and Lahiri and Li (2009) proposed adjusted likelihood variance estimators resulting from optimizing the profile and residual likelihood adjusted with a factor $h\left(\sigma_v^2\right), \sigma_v^2 > 0$. Li and Lahiri (2011) proposed two methods of variance estimation (the AM.LL and AR.LL methods, associated with the profile and residual likelihoods respectively) that ensure positive estimates with adjustment factor $h_{\mathrm{LL}}\left(\sigma_v^2\right) = \sigma_v^2$. Yoshimori and Lahiri (2014) proposed two other variance estimators (the AM.YL and AR.YL methods) derived from adjusting the the profile and residual likelihoods with factor

$$h_{\mathrm{YL}}\left(\sigma_v^2\right) = \left\{\arctan\left[\sum_{i=1}^{m} \sigma_v^2 \Big/ \left(\sigma_v^2 + \psi_i\right)\right]\right\}^{1/m}$$

where ψ_i is the sampling variance for the i^{th} area. It is well known that the LL estimators are biased, especially for small or moderate number of areas (see Lahiri and Pramanik 2011). The YL method that adjusts the profile likelihood also leads to a biased estimator of σ_v^2. However the bias of the variance estimator does not affect the MSE of the EBLUP: the second order asymptotic approximation to the MSE shows that the MSE depends on the asymptotic variance and not on the bias of the variance estimator. However, the bias of the variance estimators affects, the Taylor linearization MSE estimators and it can lead to negatively biased MSE estimators. It is desirable then to investigate alternative positive variance estimators.

The method of combining the AM.LL and the REML variance estimators was first mentioned by Yuan (2009) for the Fay-Herriot model. However, Yuan (2009) did not study its properties, empirically or otherwise. Rubin-Bleuer, Yung and Landry (2010, 2011 and 2012) carried out empirical comparisons of a MIX variance estimator under a time series and cross-sectional area level model and Rubin-Bleuer and You (2012) studied the asymptotic and finite sample properties of the MIX variance estimator for the Fay-Herriot model.

Here we formalize the MIX method for the Fay-Herriot model and prove that the MIX variance estimator is unbiased up to the second order. Furthermore, we propose an MSE estimator of the Taylor linearization type. We also examine the empirical performance of the MIX for a small and moderate number of areas. With respect to MSE estimation, Rubin-Bleuer and You (2012) and Molina, Rao and Datta (2015) each proposed a different 'split' MSE estimator under MIX variance estimation. We show that both the Rubin-Bleuer and You (2012) and the Molina et al. (2015) MSE estimators are unbiased up to the second order.

These 'split' MSE estimators were assigned a rule for populations that yielded zero estimates under REML variance estimation, and another rule for populations that yielded positive estimates under REML variance estimation. Both papers mentioned above showed that for a small number of areas, these 'split' estimators behaved well empirically in terms of average relative bias. However this outcome could be misleading, since the MSE estimators are usually negatively biased for populations where the REML variance estimate is zero, and they are positively biased for populations with positive REML estimates: the bias cancels out on average. In view of this issue we propose another MSE estimator, and we compare it to other MSE estimators when conditioned to populations where the REML estimate is zero.

In Section 2, we introduce the Fay-Herriot model, the EBLUP estimator of the small area mean and a second order approximation of the MSE of the EBLUP under the model. In Section 3, we describe the REML estimator and the *.LL and *.YL variance estimators. In Section 4, we present a general MIX variance estimator and we prove that its bias is of the same order as the bias of the REML estimator. We propose an unbiased (up to the second order) estimator of the MSE under the MIX method. In Section 5, we conduct an empirical study to compare the different variance estimators. Note that we defined the MIX variance estimator as a combination between the REML and any of the adjusted likelihood variance estimators, but the MIX variance estimator we chose for this study is the combination of the REML estimator and the AM.LL variance estimator. We selected this combination because Li and Lahiri (2011) reported that the adjusted profile likelihood performed better than adjusted residual likelihood (AR.LL) and because the adjustment factor in the Yoshimori and Lahiri (2014) variance estimators is too close to zero (in log terms), to improve significantly on the REML method. Finally in Section 6, we present the simulation results, analysis and conclusion.

2 EBLUP and MSE of the EBLUP under the Fay-Herriot model

Let $y_i, i = 1,\ldots,m,$ be the direct survey estimators of the small area means $\theta_i, i = 1,\ldots,m.$ The Fay-Herriot model consists of the following sampling and linking models:

$$\textbf{Sampling model: } y_i = \theta_i + e_i, e_i|\theta_i \overset{\text{i.d.}}{\sim} (0,\psi_i), \quad i = 1,\ldots,m, \tag{2.1}$$

$$\textbf{Linking model: } \theta_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i, v_i \overset{\text{i.i.d.}}{\sim} \left(0,\sigma_v^2\right), \sigma_v^2 > 0, \quad i = 1,\ldots,m, \tag{2.2}$$

where e_i are the sampling errors, independently distributed with mean zero and "known" sampling variances ψ_i, $\mathbf{z}_i\,(p\times 1)$ are known vectors of covariate values; $\boldsymbol{\beta}$ is a $p\times 1$ vector of unknown fixed regression coefficients; and v_i are independent and identically distributed random effects with mean zero and model variance σ_v^2. Combining (2.1) and (2.2) we obtain:

$$y_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i + e_i, \quad i = 1,\ldots,m, \tag{2.3}$$

with both model and sampling errors. The $y_i, i = 1,\ldots,m,$ can be viewed as outcomes in the combined design-model space (see Rubin-Bleuer and Schiopu-Kratina 2005).

Under model (2.3), the EBLUP of the small area mean θ_i is given by:

$$\hat{\theta}_i\left(\hat{\sigma}_v^2\right) = \mathbf{z}_i'\hat{\boldsymbol{\beta}}\left(\hat{\sigma}_v^2\right) + \hat{\gamma}_i\left[y_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}\left(\hat{\sigma}_v^2\right)\right] = \hat{\gamma}_i y_i + (1-\hat{\gamma}_i)\,\mathbf{z}_i'\hat{\boldsymbol{\beta}}\left(\hat{\sigma}_v^2\right), \quad i = 1,\ldots,m, \tag{2.4}$$

where $\hat{\sigma}_v^2$ is a consistent estimator of σ_v^2,

$$\hat{\gamma}_i = \hat{\sigma}_v^2 \Big/ \left(\hat{\sigma}_v^2 + \psi_i\right), \quad \text{and} \quad \hat{\boldsymbol{\beta}}\left(\hat{\sigma}_v^2\right) = \left[\sum_{i=1}^m \mathbf{z}_i\mathbf{z}_i' \Big/ \left(\hat{\sigma}_v^2 + \psi_i\right)\right]^{-1}\left[\sum_{i=1}^m \mathbf{z}_i y_i \Big/ \left(\hat{\sigma}_v^2 + \psi_i\right)\right]. \tag{2.5}$$

To calculate the Mean Squared Error (MSE) of the EBLUP, we set the following regularity conditions:

1) The $\boldsymbol{\psi}_i$ are bounded from above and away from zero,
2) The $\mathbf{z}_i, 1 \le i \le m$ are bounded, and
3) $\liminf \boldsymbol{\lambda}_{\min}\left(1/m\sum_i \mathbf{z}_i \cdot \mathbf{z}_i'\right) > 0$ where $\boldsymbol{\lambda}_{\min}(A)$ = minimum eigenvalue of matrix A.

Under normality of the sampling errors e_i associated with model (2.3) and the above regularity conditions, a second order approximation to the MSE is given by:

$$\mathrm{MSE}\left\{\hat{\theta}_i\left(\hat{\sigma}_v^2\right)\right\} = g_{1i}\left(\sigma_v^2\right) + g_{2i}\left(\sigma_v^2\right) + g_{3i}\left(\sigma_v^2\right) + o\left(\frac{1}{m}\right), \tag{2.6}$$

with $g_{1i}\left(\sigma_v^2\right) = \gamma_i\psi_i$, $g_{2i}\left(\sigma_v^2\right) = (1-\gamma_i)^2\,\mathbf{z}_i'\left[\sum_{i=1}^m \mathbf{z}_i\mathbf{z}_i' \Big/ \left(\sigma_v^2 + \psi_i\right)\right]^{-1}\mathbf{z}_i$ and

$$g_{3i}\left(\sigma_v^2\right) = (\psi_i)^2\,\bar{V}\left(\hat{\sigma}_v^2\right) \Big/ \left(\sigma_v^2 + \psi_i\right)^3, \tag{2.7}$$

where $\bar{V}\left(\hat{\sigma}_v^2\right)$ is the asymptotic variance of $\hat{\sigma}_v^2$ (Das, Jiang and Rao 2004).

3 Review of REML and adjusted maximum likelihood methods

3.1 REML method

We consider the combined Fay-Herriot model (2.3) with $\sigma_v^2 > 0$. The REML variance estimator of σ_v^2 is obtained by maximizing the residual likelihood function with respect to σ_v^2 :

$$L_{\mathrm{REML}}\left(\sigma_v^2\right) \propto \left|\left[\sum_{i=1}^m \mathbf{z}_i\mathbf{z}_i' \Big/ \left(\sigma_v^2 + \psi_i\right)\right]\right|^{-1/2} \prod_{i=1}^m \left(\sigma_v^2 + \psi_i\right)^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{y}\right\}$$

where $\mathbf{y} = (y_1,\ldots,y_m)'$, $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{V}^{-1}$, $\mathbf{V} = \mathrm{Var}(\mathbf{y})$, and $\mathbf{Z} = (\mathbf{z}_1,\ldots,\mathbf{z}_m)'$. (Cressie 1992, Datta and Lahiri 2000 and Rao 2003, chapter 6). The REML variance estimator is given by:

$$\hat{\sigma}_{v\mathrm{REML}}^2 = \max\left(\tilde{\sigma}_{v\mathrm{REML}}^2, 0\right), \tag{3.1}$$

where $\tilde{\sigma}^2_{v\text{REML}}$ is the converging value of the REML algorithm. The asymptotic bias and variance of the REML estimator, up to the second order, are respectively given by:

$$\text{Bias}\left(\hat{\sigma}^2_{v\text{REML}}\right) = o\left(\frac{1}{m}\right) \text{ and } V\left(\hat{\sigma}^2_{v\text{REML}}\right) = \frac{2}{\text{tr}\left(\mathbf{V}^{-2}\right)} + o\left(\frac{1}{m}\right). \tag{3.2}$$

A second order unbiased estimator of the MSE of the EBLUP under REML variance estimation is given by (Datta and Lahiri 2000 and Chen and Lahiri 2008, 2011):

$$\text{mse}\left\{\hat{\theta}_i\left(\hat{\sigma}^2_{v\text{REML}}\right)\right\} = \begin{cases} g_{1i}\left(\hat{\sigma}^2_{v\text{REML}}\right) + g_{2i}\left(\hat{\sigma}^2_{v\text{REML}}\right) + 2g_{3i}\left(\hat{\sigma}^2_{v\text{REML}}\right) & \text{if } \hat{\sigma}^2_{v\text{REML}} > 0 \\ g_{2i}(0) & \text{if } \hat{\sigma}^2_{v\text{REML}} = 0. \end{cases} \tag{3.3}$$

Remark 3.1. When $\hat{\sigma}^2_v = 0$, the EBLUP reduces to the synthetic estimator. However, note that when

$$\hat{\sigma}^2_v = 0, g_{1i}\left(\hat{\sigma}^2_v\right) = 0, \quad g_{2i}\left(\hat{\sigma}^2_v\right) = \mathbf{z}_i'\left[\sum_{i=1}^m \mathbf{z}_i\mathbf{z}_i'\Big/\psi_i\right]^{-1}\mathbf{z}_i,$$

and $g_{3i}\left(\hat{\sigma}^2_v\right) = \bar{V}\left(\hat{\sigma}^2_v\right)\Big/\psi_i > 0$, i.e., $\text{mse}\left\{\hat{\theta}_i\left(\hat{\sigma}^2_v\right)\right\}$ is not a continuous function of $\hat{\sigma}^2_v$. We will see in the empirical study that when conditioning on $\left\{\hat{\sigma}^2_v = 0\right\}$, the MSE estimator in (3.3) has significant negative bias, unless the underlying signal to noise ratio $\sigma^2_v\Big/\psi_i$ is negligible.

3.2 Adjusted maximum likelihood methods

The adjusted maximum likelihood variance estimators are derived from optimizing either the profile (AM) or the residual (AR) likelihood adjusted with the factor $h\left(\sigma^2_v\right)$. As noted in the introduction, the AM.LL and AR.LL estimators use the adjustment factor $h_{\text{LL}}\left(\sigma^2_v\right) = \sigma^2_v$, and the AM.YL and AR.YL estimators use the adjustment factor

$$h_{\text{YL}}\left(\sigma^2_v\right) = \left\{\arctan\left[\sum_{i=1}^m \sigma^2_v\Big/\left(\sigma^2_v + \psi_i\right)\right]\right\}^{1/m}.$$

We denote by $\hat{\sigma}^2_{v\text{AM.LL}}$ and $\hat{\sigma}^2_{v\text{AM.YL}}$ the variance estimators obtained by maximizing the adjusted profile likelihood functions, with respect to σ^2_v :

$$L_{\text{AM.}*}\left(\sigma^2_v\right) \propto h\left(\sigma^2_v\right)\cdot\prod\nolimits_{i=1}^m\left(\sigma^2_v + \psi_i\right)^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{y}\right\}, \tag{3.4}$$

where $h\left(\sigma^2_v\right) = h_{\text{LL}}\left(\sigma^2_v\right)$ and $h\left(\sigma^2_v\right) = h_{\text{YL}}\left(\sigma^2_v\right)$ for AM.LL and AM.YL respectively. The matrix $\mathbf{P}$ is as in (3.1). The bias of the AM estimators up to the second order (denoted by $\approx$) is:

$$B\left(\hat{\sigma}^2_{v\text{AM.LL}}\right) \approx \frac{\text{tr}\left\{\mathbf{P} - \mathbf{V}^{-1}\right\} + 2\Big/\sigma^2_v}{\text{tr}\left(\mathbf{V}^{-2}\right)} = O\left(\frac{1}{m}\right) \quad \text{and} \quad B\left(\hat{\sigma}^2_{v\text{AM.YL}}\right) \approx \frac{\text{tr}\left\{\mathbf{P} - \mathbf{V}^{-1}\right\}}{\text{tr}\left(\mathbf{V}^{-2}\right)} = O\left(\frac{1}{m}\right), \tag{3.5}$$

(Li and Lahiri 2011 and Yoshimori and Lahiri 2014). The AR.LL and AR.YL variance estimators, denoted by $\hat{\sigma}^2_{v\text{AR.LL}}$ and $\hat{\sigma}^2_{v\text{AR.YL}}$, are obtained by maximizing the adjusted residual (AR) likelihood functions with respect to σ_v^2 :

$$L_{\text{AR.}*}\left(\sigma_v^2\right) \propto h\left(\sigma_v^2\right) \cdot \left| \sum_{i=1}^{m} \mathbf{z}_i \mathbf{z}_i' \Big/ \left(\sigma_v^2 + \psi_i\right) \right|^{-1/2} \prod_{i=1}^{m} \left(\sigma_v^2 + \psi_i\right)^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{y}\right\} \tag{3.6}$$

where $h\left(\sigma_v^2\right) = h_{\text{LL}}\left(\sigma_v^2\right)$ and $h\left(\sigma_v^2\right) = h_{\text{YL}}\left(\sigma_v^2\right)$ for AR.LL and AR.YL respectively and $\mathbf{P}$ is as in (3.1), The asymptotic bias of the AR estimators are given, respectively by:

$$B\left(\hat{\sigma}^2_{v\text{AR.LL}}\right) \approx \frac{2/\sigma_v^2}{\text{tr}\left(\mathbf{V}^{-2}\right)} = O\left(\frac{1}{m}\right) \text{ and } B\left(\hat{\sigma}^2_{v\text{AR.YL}}\right) = o\left(\frac{1}{m}\right). \tag{3.7}$$

Under the regularity conditions given in Section 2 and $\sigma_v^2 > 0$, the two LL and the two YL variance estimators exist and are $\sqrt{m}$ −consistent (Li and Lahiri 2011 and Yoshimori and Lahiri 2014). Lahiri and co-authors proposed the following MSE estimators:

$$\text{mse}\left\{\hat{\theta}_i\left(\cdot\right)\right\} = g_{1i}\left(\cdot\right) + g_{2i}\left(\cdot\right) + 2g_{3i}\left(\cdot\right) - \psi_i^2 \cdot B\left(\cdot\right)/\left(\cdot + \psi_i\right)^2 \tag{3.8}$$

where the argument in $(\cdot)$ above is either $\hat{\sigma}^2_{v\text{AM.LL}}$, $\hat{\sigma}^2_{v\text{AR.LL}}$ or $\hat{\sigma}^2_{v\text{AM.YL}}$ under AM.LL, AR.LL and AM.YL variance estimation respectively, and under $\hat{\sigma}^2_{v\text{AR.YL}}$:

$$\text{mse}\left\{\hat{\theta}_i\left(\hat{\sigma}^2_{v\text{AR.YL}}\right)\right\} = g_{1i}\left(\hat{\sigma}^2_{v\text{AR.YL}}\right) + g_{2i}\left(\hat{\sigma}^2_{v\text{AR.YL}}\right) + 2g_{3i}\left(\hat{\sigma}^2_{v\text{AR.YL}}\right). \tag{3.9}$$

Estimators (3.8) and (3.9) are unbiased up to the second order.

Remark 3.2. The sampling errors do not need to be normally distributed for the consistency and asymptotic normality of the LL and YL estimators (see, for example, Rubin-Bleuer et al. 2011).

3.3 Optimization algorithms

Given the data, the REML likelihood function may attain its maximum value at $\sigma_v^2 = 0$, even when the true underlying value of $\boldsymbol{\sigma}_v^2$ is positive. On the other hand, the LL and YL likelihoods always attain their maximum value at $\sigma_v^2 > 0$. Yet, the YL residual likelihood is very close to the REML likelihood. Empirical studies show that the scoring algorithm under AR.YL yields $\hat{\sigma}^2_{v\text{AR.YL}}$ in almost as large a percentage as under REML for data sets following a Fay-Herriot model with a small but non-zero true underlying variance. This happens when the scoring algorithm misses the positive maximum value of the AR.YL likelihood and outputs a zero value (see Appendix B for details). To avoid this problem, we use a grid method for optimization (Estevao 2014). In our study, we set the upper boundary of the search interval as $1{,}000 \times \sigma_v^2$, since we know σ_v^2 a priori. For applications with real data we suggest to obtain an initial estimate $\hat{\sigma}^2_{v\text{AM.LL}}$ by the method of scoring and set $1{,}000 \times \hat{\sigma}^2_{v\text{AM.LL}}$ as the upper boundary. Then keep increasing the boundary until the variance estimate lies within the search interval.

4 The MIX variance estimator

4.1 Variance estimation

The MIX variance estimator is a procedure that first calculates the REML variance estimate and only substitutes it by an adjusted likelihood variance estimate if the REML estimate is negative. The MIX variance estimator is always positive and it is unbiased up to a term of order $o(1/m)$. The MIX variance estimator of σ_v^2 is defined by:

$$\hat{\sigma}_{v\text{MIX}}^2 = \begin{cases} \hat{\sigma}_{v\text{REML}}^2 & \text{if } \hat{\sigma}_{v\text{REML}}^2 > 0 \\ \hat{\sigma}_{v\text{adj}}^2 & \text{if } \hat{\sigma}_{v\text{REML}}^2 = 0, \end{cases} \tag{4.1}$$

where $\hat{\sigma}_{v\text{adj}}^2$ is one of the adjusted likelihood estimators defined in Section 3.

Remark 4.1. The MIX variance estimator automatically carries some of the common properties shared by the REML and the adjusted likelihood variance estimator. For example, it is even and translation invariant. Thus, under normality of the sampling errors, the second order approximation (2.6) of the MSE of the EBLUP is also valid: Theorem 4.1 below shows that the MSE of the EBLUP under the MIX variance estimator inherits the same asymptotic properties as the MSE under the REML variance estimator.

Theorem 4.1. Under regularity conditions 1 through 3 given in Section 2, and the assumption that $\sigma_v^2 > 0$, the MSE of the EBLUP under the MIX variance estimator is equal to the MSE under the REML variance estimator up to the second order. The theorem follows from the fact that the asymptotic variance of $\hat{\sigma}_{v\text{MIX}}^2$ coincides with the asymptotic variance of $\hat{\sigma}_{v\text{REML}}^2$ (see Appendix A for details).

Theorem 4.2. Under the conditions of Theorem 4.1, $\text{Bias}\left(\hat{\sigma}_{v\text{MIX}}^2\right) = o(1/m)$. The proof is given in Appendix A.

4.2 MSE estimation

The fact that the MIX estimator, $\hat{\sigma}_{v\text{MIX}}^2$, is unbiased to the second order, is crucial to show that our proposed MSE estimator is also unbiased up to the second order.

Corollary 4.2. The MSE estimator of the EBLUP under $\hat{\sigma}_{v\text{MIX}}^2$ given by:

$$\text{mse}\left[\hat{\theta}_i\left(\hat{\sigma}_{v\text{MIX}}^2\right)\right] = g_{1i}\left(\hat{\sigma}_{v\text{MIX}}^2\right) + g_{2i}\left(\hat{\sigma}_{v\text{MIX}}^2\right) + 2g_{3i}\left(\hat{\sigma}_{v\text{MIX}}^2\right) \tag{4.2}$$

is second order unbiased. Once given that $\hat{\sigma}_{v\text{MIX}}^2$ is second order unbiased, the result follows along the lines of Datta and Lahiri (2000).

4.3 Alternative MSE estimators

In the following the MIX variance estimator is the combination of REML and AM.LL.

Rubin-Bleuer and You (2012) had suggested another MSE estimator, also unbiased up to the second order: a 'split' MSE estimator of the form:

$$\text{mse}^*\left[\hat{\theta}_i\left(\hat{\sigma}^2_{v\text{MIX}}\right)\right] = \begin{cases} g_{1i}\left(\hat{\sigma}^2_{v\text{MIX}}\right) + g_{2i}\left(\hat{\sigma}^2_{v\text{MIX}}\right) + 2g_{3i}\left(\hat{\sigma}^2_{v\text{MIX}}\right) & \text{if } \hat{\sigma}^2_{v\text{MIX}} = \hat{\sigma}^2_{v\text{REML}}, \\ g_{1i}\left(\hat{\sigma}^2_{v\text{MIX}}\right) + g_{2i}\left(\hat{\sigma}^2_{v\text{MIX}}\right) & \\ \quad + 2g_{3i}\left(\hat{\sigma}^2_{v\text{MIX}}\right) - (1-\hat{\gamma}_{i\text{MIX}})^2 \cdot \text{Bias}\left(\hat{\sigma}^2_{v\text{MIX}}\right) & \text{if } \hat{\sigma}^2_{v\text{MIX}} = \hat{\sigma}^2_{v\text{AM.LL}}. \end{cases} \tag{4.3}$$

Estimator mse* has a lower average relative bias (ARB) than the MSE estimator given in (4.2). The lower ARB occurs because the MSE estimates overestimate when REML is positive and underestimate when REML is zero. The mse* estimator is good on average, but for a particular data set the mse* estimator might take on negative values.

Molina et al. (2015) proposed two different MSE estimators for the EBLUP under the MIX: with PT standing for their proposed preliminary test of hypothesis for zero variance these estimators are:

$$\text{mse}_0\left\{\hat{\theta}_i\left(\hat{\sigma}^2_{v\text{MIX}}\right)\right\} = \begin{cases} \text{mse}\left\{\hat{\theta}_i\left(\hat{\sigma}^2_{v\text{REML}}\right)\right\} & \text{if } \hat{\sigma}^2_{v\text{REML}} > 0 \\ g_{2i}(0) & \text{if } \hat{\sigma}^2_{v\text{REML}} = 0 \end{cases} \tag{4.4}$$

and

$$\text{mse}_{\text{PT}}\left\{\hat{\theta}_i\left(\hat{\sigma}^2_{v\text{MIX}}\right)\right\} = \begin{cases} \text{mse}\left\{\hat{\theta}_i\left(\hat{\sigma}^2_{v\text{REML}}\right)\right\} & \text{if } \hat{\sigma}^2_{v\text{REML}} > 0 \text{ and PT rejected} \\ g_{2i}(0) & \text{if } \hat{\sigma}^2_{v\text{REML}} = 0 \text{ or PT not rejected.} \end{cases} \tag{4.5}$$

The rationale for mse_0 and mse_{PT} is based on the MSE of the BLUP with $\sigma^2_v = 0$. Molina et al. (2015) showed in an empirical study that their proposed MSE estimators performed well on average when both σ^2_v and the number of areas m were small.

Remark 4.2. mse_0 and mse_{PT} are also unbiased up to the second order (see Appendix for a brief proof of this property). Our argument against $\text{mse}\left\{\hat{\theta}_i\left(\hat{\sigma}^2_v\right)\right\}$ (in 3.3) is also valid against mse_0 and mse_{PT} : for a moderate number of areas, the % of populations with $\hat{\sigma}^2_{v\text{REML}} = 0$ may be significant even if σ^2_v / ψ_i is not negligible. In this case, the MSE of the EBLUP should account also for the variation due to variance estimation or risk underestimation.

5 Simulation set up and performance measures

5.1 Simulation set-up

We conducted a model-based Monte Carlo simulation, following Rubin-Bleuer and You (2012), to examine the finite sample performance of the various methods. 'Direct' estimates $(y_1, \ldots, y_m)$ with $m = 15, m = 45$ and $m = 100,$ are generated from the Fay-Herriot model in (2.3) with $\boldsymbol{\beta}' = (5, 4, 3, 2, 1)$ and covariates $\mathbf{z}'_i = (1, z_{i2}, \ldots, z_{ip})$, generated once from normal distributions $z_{ik} \sim k + N(1,1)$, $k = 2, \ldots, 5,\ i = 1, \ldots, m,$ and held fixed over the repeated populations. The independent normal random

area effects v_i are generated with variance $\sigma_v^2 = 1$. Independent sampling errors e_i, are generated with sampling variances $\psi_i \triangleq 50/n_i$, where n_i is the sample size for area $i, i = 1,\ldots,m$. There are five sampling variance groups determined by $n_i = 3, 5, 7, 10$ or 15, with signal to noise ratios $\sigma_v^2/\psi_i = 0.06$, 0.1, 0.14, 0.2 and 0.3, respectively. Thus when $m = 100$ there are 20 areas per signal to noise ratio. We first generated 50,000 sets of direct estimators for each case and computed the EBLUP and the true Monte Carlo MSE of the EBLUP using the REML, AM.LL, MIX, AM.YL and AR.YL variance estimators. We did not study AR.LL due to its poor performance reported by Li and Lahiri (2011). Next we generated 10,000 sets of direct estimators independently of the first 50,000. For each generated set, we computed the five variance estimators. For the MIX variance estimator we looked at three of the four linearization type MSE estimators discussed in Section 4. Since the linearization MSE estimators often do not estimate bias accurately, we also considered the parametric bootstrap MSE (PB MSE) estimator adjusted for bias using Pfeffermann and Glickman's (2004) method and the naïve PB MSE estimator with 500 repetitions each (see Appendix B for the construction of the bootstrap). The Monte Carlo performance measures are defined below.

1. The MSE of the EBLUP, $\overline{\text{MSE}}_\ell\left(\hat{\theta}_i\right)$, per sampling variance group:

$$\text{MSE}\left(\hat{\theta}_i\right) = \frac{1}{50{,}000}\sum_{r=1}^{50{,}000}\left(\hat{\theta}_i^{(r)} - \theta_i^{(r)}\right)^2,\ \overline{\text{MSE}}_\ell\left(\hat{\theta}\right) = \frac{5}{m}\sum_{i\in\{j:\psi_j=50/n_\ell\}}\text{MSE}\left(\hat{\theta}_i\right),\ \ \ell = 1,\ldots,5.$$

2. $E\left(\hat{\sigma}_v^2\right) = \sum_{r=1}^{10{,}000}\hat{\sigma}_v^{2(r)}/10{,}000$, $V\left(\hat{\sigma}_v^2\right) = \sum_{r=1}^{10{,}000}\left(\hat{\sigma}_v^{2(r)} - E\left(\hat{\sigma}_v^2\right)\right)^2/10{,}000$, where $\hat{\sigma}_v^{2(r)}$ is the value of $\hat{\sigma}_v^2$ for the r^{th} simulation run $(r = 1,\ldots,10{,}000)$.
3. The Average Relative Bias (ARB) of the MSE per sampling variance group:

$$\text{ARB}_\ell\,(\text{mse}) = \frac{5}{m}\sum_{i\in\{j:\psi_j=50/n_\ell\}}\text{RB}\left(\text{mse}\left(\hat{\theta}_i\right)\right),\ \ \ell = 1,\ldots 5,$$

where $\text{RB}\left(\text{mse}\left(\hat{\theta}_i\right)\right) = \left[\sum_{r=1}^{10{,}000}\text{mse}\left(\hat{\theta}_i^{(r)}\right)/10{,}000 - \text{MSE}\left(\hat{\theta}_i\right)\right]/\text{MSE}\left(\hat{\theta}_i\right)$.

4. The Root Relative MSE of MSE estimators per sampling variance group:

$$\text{RRMSE}_\ell\,(\text{mse}) = \left(\frac{5}{m}\sum_{i\in\{j:\psi_j=50/n_\ell\}}\frac{\sum_{r=1}^{10{,}000}\left(\text{mse}\left(\hat{\theta}_i^{(r)}\right) - \text{MSE}\left(\hat{\theta}_i\right)\right)^2/10{,}000}{\text{MSE}\left(\hat{\theta}_i\right)}\right)^{1/2}.$$

We also examine the bias of the conditional MSE estimators given that $\left\{\hat{\sigma}_{v\text{REML}}^2 = 0\right\}$ because these are the populations for which the positive estimators were developed.

5. The Average Relative Bias of Conditional MSE estimators:

$$\text{ARB}_{\mathbb{C}} = \frac{5}{m}\sum_{i\in\ell}\text{E}\left[\text{mse}\left(\hat{\theta}_i\right)\middle|\,\hat{\sigma}_{v\text{REML}}^2 = 0\right]\Big/\text{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2\middle|\hat{\sigma}_{v\text{REML}}^2 = 0\right] - 1.$$

6 Simulation results and analysis

6.1 Monte Carlo Distribution of the variance estimators

Table 6.1 shows that the REML variance estimator has the lowest bias $\left(\sigma_v^2 = 1\right)$ and the highest variance. The lower efficiency of REML may be due to it not being a smooth function of the data caused by its split definition (3.1). The MIX estimator inherits some of this low efficiency. The other variance estimators have lower variability, higher positive bias but the conditional expectation of AM.YL and AR.YL given $\hat{\sigma}^2_{\text{vREML}} = 0$ is close to zero. The unconditional bias of AM.LL is higher than the unconditional bias of the MIX. By definition of the MIX estimator, the conditional bias of the MIX and AM.LL estimators coincide. The MIX estimator also converges faster than the other estimators. For example, given the probability distribution over the 10,000 variance estimates with $m = 45$, we calculated the probability of estimates lying within an interval containing $\sigma_v^2 = 1$. The probability that the MIX estimates lie between 0.6 and 1.4 is 0.47 whereas the probability that AM.YL estimates lie between 0.6 and 1.4 is 0.16. Furthermore, the probability that MIX estimates are smaller than 0.2 is 0.05 whereas the probability that AM.YL estimates are smaller than 0.2 is 0.53.

Table 6.1
Expectation, variance and conditional expectation and variance of $\hat{\sigma}_v^2$

Method	m	$E\left(\hat{\sigma}_v^2\right)$	$V\left(\hat{\sigma}_v^2\right)$	%REML = 0	$E\left(\hat{\sigma}_v^2 / \text{REML} = 0\right)$	$V\left(\hat{\sigma}_v^2 / \text{REML} = 0\right)$
REML	15	1.48	3.38	43%	N/A	N/A
	45	1.21	1.67	29%	N/A	N/A
	100	1.07	0.81	16%	N/A	N/A
AM.LL	15	2.80	1.37	43%	1.80	0.11
	45	1.88	1.01	29%	0.94	0.03
	100	1.49	0.51	16%	0.63	0.01
MIX	15	2.28	1.87	43%	1.80	0.11
	45	1.48	1.31	29%	0.94	0.03
	100	1.17	0.66	16%	0.63	0.01
AR.YL	15	1.66	2.99	43%	0.27	0.01
	45	1.24	1.72	29%	0.06	0.00
	100	1.08	0.80	16%	0.02	0.00
AM.YL	15	0.52	0.84	43%	0.10	0.00
	45	0.65	0.85	29%	0.03	0.00
	100	0.76	0.59	16%	0.01	0.00

6.2 True MSE of the EBLUP, average relative bias and average root relative MSE of the MSE estimators

All variance estimators are consistent and asymptotically normal with variance converging at the same rate. They differ in their bias: REML, AR.YL and MIX have bias of the order of $o(1/m)$ whereas AM.LL and AM.YL have bias of the order $O(1/m)$. The bias inherent in the last three methods impacts the estimation of the MSE of the EBLUP even for a moderate number of areas.

For $m = 100$, Tables 6.2a and 6.2b show that as σ_v^2 / ψ_i increases, the MSE of the EBLUP decreases and this relationship holds irrespective of the number of areas. We observe that the MSE of $\hat{\theta}_i$ under the

REML and the MIX variance estimators are slightly higher than the rest of the MSEs, due to the higher variability inherent in these variance estimators. Table 6.2a presents results for the Taylor linearization MSE estimator and the two parametric MSE estimators under REML, AM.LL, AR.YL and AM.YL variance estimation. Table 6.2b presents results for the following MSE estimators under the MIX variance estimation: RB_Y1 defined in (4.3), RB_Y2 defined in (4.2), M_et_al, defined in (4.5), PB MSE and naïve PB MSE estimators. Among the Taylor MSE estimators, RB_Y1 and M_et_al under MIX exhibit the lowest bias. Among the bootstrap MSE estimators PB under MIX and Naive PB under AR.YL exhibit the lowest bias. Turning to the RRMSE of the MSE estimators, it decreases as σ_v^2/ψ_i increases. Differences between the RB_Y2 MSE estimator under the MIX and the Taylor MSE estimator under the AM.YL seem small but consistent. While ARB is lower for the RB_Y1, the M_et_al and the Naive MSE estimators under the MIX method than for the RB_Y2 under the MIX, and also lower for the Taylor and the Naive PB under the AR.YL method than for the RB_Y2 under the MIX, the opposite happens in terms of RRMSE. This can be explained in part due to the extreme negative conditional bias exhibited by these MSE estimators (i.e., the RB_Y1 and the M_et_al under the MIX and the Taylor and the Naive PB under the AR.YL method) as shown in Table 6.3. Even for $m = 100$ there is a relatively high proportion (16%) of populations that yield $\hat{\sigma}^2_{v\text{REML}} = 0$ and in these populations, estimates from most variance methods and most MSE estimators are farthest below the true value. That is, for these MSE estimators, the conditional MSE estimators do not fare well. The PB MSE estimator seems to adjust well for bias, but it is more variable than the Naive PB MSE. When we also include the ARB, the RRMSE and the $\text{ARB}_{\mathbb{C}}$ in the evaluation, the RB_Y2 under the MIX method, followed closely by Naïve PB under the MIX seems to perform the best. This may suggest the superiority of RB_Y2 and Naive under MIX for $m = 100$, which is a moderate number of areas for this data.

Table 6.2a
MSE, ARB & RRMSE (percentage) of MSE Estimators, $m = 100$

			Taylor MSE estimator		PB estimator		Naïve PB estimator	
Method	σ_v^2/ψ_i	$\overline{\text{MSE}}$	**ARB**	**RRMSE**	**ARB**	**RRMSE**	**ARB**	**RRMSE**
REML	0.06	135.4	5.1	71.1	-4.4	80.7	1.6	69.9
	0.1	132.1	5.3	64.7	-4.7	74.0	-0.2	63.0
	0.14	119.5	6.0	61.9	-5.5	71.3	-1.8	59.9
	0.2	119.2	6.5	53.6	-5.8	62.4	-3.4	51.7
	0.3	106.6	8.2	46.7	-6.8	55.0	-5.6	44.8
AM.LL	0.06	134.9	6.1	75.4	8.2	66.9	31.3	63.8
	0.1	131.2	6.8	68.1	7.8	59.5	27.5	55.7
	0.14	118.3	8.1	64.6	7.8	55.6	26.5	51.2
	0.2	117.6	8.4	55.4	6.5	46.7	21.6	42.1
	0.3	104.5	10.2	46.7	5.5	38.8	18.2	34.0
AR.YL	0.06	135.4	6.6	69.3	-4.3	80.2	2.1	69.4
	0.1	132.0	7.4	61.9	-4.5	73.4	0.3	62.5
	0.14	119.4	9.0	58.0	-5.3	70.6	-1.2	59.3
	0.2	119.0	10.6	48.2	-5.6	61.8	-2.9	51.1
	0.3	106.4	14.7	38.5	-6.6	54.3	-5.1	44.1
AM.YL	0.06	134.7	10.0	63.2	-12.3	81.0	-19.6	65.9
	0.1	131.3	12.0	56.6	-12.5	75.2	-19.7	61.2
	0.14	118.8	15.0	53.1	-13.7	73.3	-21.4	59.8
	0.2	118.6	18.1	44.8	-13.4	65.2	-20.7	53.5
	0.3	106.4	25.2	38.4	-14.4	58.8	-21.7	48.6

Table 6.2b
MSE, ARB & RRMSE (percentage) of MSE Estimators $m = 100$

			RB_Y1		RB_Y2		M_et_al		PB estimator		Naïve PB estimator	
	σ_v^2/ψ_i	$\overline{\text{MSE}}$	ARB	RRMSE	ARB	RRMSE	ARB	RRMSE	ARB	RRMSE	ARB	RRMSE
MIX	0.06	135.4	2.7	75.7	13.6	63.0	5.2	71.1	-3.0	75.3	8.8	62.4
	0.1	132.1	3.6	68.3	14.9	56.1	5.3	64.7	-3.2	68.3	6.6	55.4
	0.14	119.5	4.9	64.7	16.0	52.4	6.0	61.9	-3.9	65.1	5.3	51.8
	0.2	119.1	6.3	55.2	16.7	43.8	6.5	53.6	-4.4	56.3	2.9	43.7
	0.3	106.5	9.4	46.2	19.9	36.0	8.3	46.7	-5.4	48.6	0.6	36.7

Table 6.3
$\text{MSE}_{\text{C}}\left(\text{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2 \mid \hat{\sigma}^2_{v\text{REML}} = 0\right]\right)$ and ARB_{C} (percentage), $m = 100$

Method	σ_v^2/ψ_i	$\overline{\text{MSE}}_{\text{C}}$	Taylor MSE estimator	PB estimator	Naïve PB estimator
REML	0.06	135.6	-76.5	-98.6	-74.8
	0.1	133.0	-74.5	-94.4	-71.8
	0.14	121.5	-78.6	-98.0	-74.9
	0.2	120.4	-73.1	-89.8	-68.6
	0.3	108.0	-73.6	-88.2	-67.3
AM.LL	0.06	135.0	-92.0	-67.6	-26.1
	0.1	132.2	-85.2	-62.0	-24.6
	0.14	120.2	-85.4	-62.3	-25.4
	0.2	118.8	-74.4	-54.1	-22.1
	0.3	105.9	-65.9	-49.7	-20.6
AR.YL	0.06	135.5	-68.6	-96.9	-73.0
	0.1	132.9	-62.4	-92.6	-70.0
	0.14	121.4	-61.1	-96.1	-73.0
	0.2	120.2	-48.9	-87.9	-66.7
	0.3	107.8	-34.5	-86.1	-65.4
AM.YL	0.06	134.9	-45.9	-88.6	-74.7
	0.1	132.1	-39.4	-85.4	-72.2
	0.14	120.4	-36.0	-89.3	-75.7
	0.2	119.6	-23.6	-82.3	-69.7
	0.3	107.6	-6.5	-81.7	-69.3

			RB_Y1	RB_Y2	M_et_al	PB estimator	Naïve PB estimator
MIX	0.06	135.0	-92.0	-22.0	-76.4	-46.0	-27.0
	0.1	132.2	-85.2	-17.7	-74.3	-42.7	-25.9
	0.14	120.2	-85.4	-15.0	-78.3	-43.3	-27.0
	0.2	118.8	-74.4	-7.6	-72.8	-37.6	-23.9
	0.3	105.9	-65.9	1.5	-73.1	-34.6	-22.6

Tables 6.4a and b below display results for $m = 45$ with 9 areas per σ_v^2/ψ_i. The AM.YL yields MSEs smaller than the MIX, with differences in MSEs of at most 2%. As the number of areas decreases, the bias of the variance estimators increase and the MSE estimators are affected by this. Indeed, the ARB of all MSE estimators have increased. In particular, the ARB of the Taylor MSE estimators under YL and LL variance estimation and the ARB of RB_Y2, have increased by 100% over the ARB with 100 areas. In terms of RRMSE, the Taylor MSE under the AM.YL has slightly lower RRMSE than the RB_Y2 under the MIX method for very small σ_v^2/ψ_i. In general, the variability (in RRMSE) of the RB_Y2 is lower than that of the Taylor under LL and YL estimation and than that of the RB_Y1 and the M_et_al. This may be due in part to the underestimation of the MSEs for the populations with zero REML estimates, which, for $m = 45$,

range around 30% of all populations. Table 6.5 illustrates this better: given $\hat{\sigma}^2_{vREML} = 0$, there is serious underestimation in RB_Y1 and M_et_al.

Table 6.4a
MSE, ARB & RRMSE (percentage) of MSE Estimators, $m = 45$ areas

Method	σ_v^2/ψ_i	$\overline{MSE}$	Taylor MSE estimator		PB estimator		Naïve PB estimator	
			ARB	RRMSE	ARB	RRMSE	ARB	RRMSE
REML	0.06	171.4	11.8	94.7	-4.7	107.0	6.2	89.2
	0.1	174.1	11.9	83.9	-5.3	93.8	3.0	76.2
	0.14	171.3	12.6	74.5	-5.4	81.9	1.1	65.3
	0.2	166.6	13.9	63.4	-5.8	66.7	-1.2	52.0
	0.3	128.9	20.1	63.0	-7.0	61.4	-3.1	46.7
AM.LL	0.06	171.1	15.5	100.0	16.0	84.9	43.5	83.3
	0.1	173.4	16.8	87.0	14.4	71.1	36.7	68.5
	0.14	170.4	17.7	75.7	12.6	59.7	30.7	56.7
	0.2	165.3	18.2	61.7	9.9	46.2	23.5	43.2
	0.3	127.5	25.6	55.0	10.0	39.7	22.6	36.6
AR.YL	0.06	171.1	17.2	89.9	-3.7	105.0	8.0	87.6
	0.1	173.6	19.6	76.9	-4.3	91.8	4.8	74.6
	0.14	170.8	22.6	65.8	-4.4	79.9	2.7	63.7
	0.2	166.0	27.3	53.7	-4.8	64.8	0.3	50.5
	0.3	128.3	43.8	54.8	-5.7	59.3	-1.3	45.0
AM.YL	0.06	167.5	30.2	78.4	-18.0	97.3	-23.8	73.3
	0.1	169.6	36.5	72.2	-18.0	87.7	-23.6	66.7
	0.14	167.0	42.7	69.3	-17.2	78.0	-22.3	59.7
	0.2	162.8	52.1	70.8	-15.8	65.4	-20.3	50.6
	0.3	126.0	81.3	91.1	-18.0	62.3	-22.9	48.4

Table 6.4b
MSE, ARB & RRMSE (percentage) of MSE Estimators, $m = 45$ areas

	σ_v^2/ψ_i	$\overline{MSE}$	RB_Y1		RB_Y2		M_et_al		PB estimator		Naïve PB estimator	
			ARB	RRMSE	ARB	RRMSE	ARB	RRMSE	ARB	RRMSE	ARB	RRMSE
MIX	0.06	171.4	9.8	99.4	31.9	84.0	11.8	94.7	3.5	93.8	21.9	78.5
	0.1	174.0	12.1	86.2	33.2	73.1	11.9	83.9	2.6	80.4	17.5	65.1
	0.14	171.2	14.5	74.9	34.4	64.6	12.6	74.5	2.0	68.7	14.0	54.4
	0.2	166.5	17.7	61.7	36.0	55.8	13.9	63.4	0.7	54.5	9.8	41.8
	0.3	128.9	28.8	57.6	48.8	58.2	20.2	63.1	0.3	48.6	8.7	35.9

Taking into account the ARB, the RRMSE and the $ARB_{\mathbb{C}}$ of the MSE estimators, the Naive PB MSE estimator under the MIX performs the best for larger σ_v^2/ψ_i. Table 6.6 displays performance measures, averaged over the five sampling variance groups, for the three Taylor MSE estimators under the MIX with

data from the same model described in 5.1 but with three different values of σ_v^2. The RB_Y2 performs better when $\sigma_v^2 = 1$, but as σ_v^2 becomes smaller, the M_et_al MSE estimator has an advantage, precisely because it was constructed under the premise that σ_v^2 is approximately zero.

Table 6.5
MSE_C and ARB_C (percentage). $m = 45$ areas

Method	σ_v^2/ψ_i	$\overline{MSE}_C$	Taylor MSE estimator	PB estimator	Naïve PB estimator
REML	0.06	170.2	-64.3	-89.7	-60.7
	0.1	173.0	-62.4	-83.7	-57.1
	0.14	170.2	-58.1	-75.5	-51.8
	0.2	165.8	-51.9	-65.1	-44.8
	0.3	131.1	-59.0	-70.5	-49.2
AM.LL	0.06	170.0	-71.5	-49.0	-3.1
	0.1	172.3	-61.5	-42.1	-2.3
	0.14	169.1	-51.1	-35.7	-2.1
	0.2	164.7	-38.3	-28.3	-1.6
	0.3	129.9	-28.8	-29.1	-3.7
AR.YL	0.06	169.9	-48.3	-86.2	-56.7
	0.1	172.6	-38.0	-80.2	-53.2
	0.14	169.7	-25.9	-72.2	-48.2
	0.2	165.3	-7.4	-61.9	-41.5
	0.3	130.5	19.3	-66.8	-45.5
AM.YL	0.06	166.6	-8.2	-73.5	-60.7
	0.1	168.8	3.8	-70.1	-58.1
	0.14	166.1	16.1	-64.1	-53.3
	0.2	162.2	35.9	-56.1	-46.8
	0.3	128.1	72.8	-62.5	-52.5

			RB_Y1	RB_Y2	M_et_al		
MIX	0.06	170.0	-71.5	6.2	-64.3	-28.1	-4.0
	0.1	172.3	-61.5	13.2	-62.3	-23.8	-3.5
	0.14	169.1	-51.1	18.9	-57.8	-20.0	-3.3
	0.2	164.7	-38.3	26.8	-51.6	-15.7	-2.9
	0.3	129.9	-28.8	40.4	-58.7	-16.7	-5.1

Table 6.6
MSE, ARB, ARB_C and RRMSE (percentage), 45 areas

			RB_Y1			RB_Y2			M_et_al		
%REML = 0	σ_v^2	$\overline{MSE}$	**ARB**	ARB_C	**RRMSE**	**ARB**	ARB_C	**RRMSE**	**ARB**	ARB_C	**RRMSE**
29	1	108	16	-50	75	36	21	66	14	-59	75
48	0.2	99	48	-36	101	113	88	114	47	-38	94
51	0.1	91	58	-33	108	137	107	127	58	-32	100

Tables 6.7a and 6.7b below show the outcomes for $m = 15$ areas with 3 areas per σ_v^2/ψ_i. Differences in MSEs per variance estimation method are at most 5%.

There is no monotone relationship between ARB or RRMSE and σ_v^2/ψ_i, which could be an indication that the second order approximation to estimating the MSE is poor under every method of variance estimation. The ARB of all Taylor MSE estimators under the LL and the YL methods of variance estimation

are unacceptably high and the same is true for the RRMSE. The RB_Y2 under the MIX does not fare well either. The reason for this last outcome is clear: the high % of zero REML estimates (43%) implies the MIX coincides with AM.LL for the zero REML populations. Thus, the MIX has a positive bias for $m = 15$, and the RB_Y2 does not account for this bias. The RB_Y1 accounts for the bias in the MIX, but the bias estimator is not very precise for $m = 15$. The M_et_al MSE estimator almost coincides with the ARB and RRMSE of the Taylor MSE estimator under the REML variance estimation, because by definition they are equal when $\hat{\sigma}^2_{v\text{REML}} = 0$. The $\text{ARB}_{\mathbb{C}}$ of the three Taylor MSE estimators under the MIX is poor. Taking into account all performance measures, the bootstrap MSE estimators perform better than the Taylor MSE estimators. For $m = 15$ areas with 3 areas per σ_v^2/ψ_i, PB under MIX performs the best, followed by the Naive under AR.YL and AM.YL.

Table 6.7a
MSE, ARB & RRMSE (percentage) of MSE estimators, $m = 15$ areas

			Taylor MSE estimator		PB estimator		Naïve PB estimator	
Method	σ_v^2/ψ_i	$\overline{\text{MSE}}$	ARB	RRMSE	ARB	RRMSE	ARB	RRMSE
REML	0.06	584.8	12.6	87.9	1.2	85.9	6.9	64.5
	0.1	376.7	26.5	106.3	2.3	85.6	9.6	62.8
	0.14	352.5	25.2	90.1	0.7	54.1	4.3	39.3
	0.2	209.4	43.0	123.0	0.4	74.0	6.3	51.1
	0.3	198.7	50.6	124.7	-1.0	46.3	2.6	31.5
AM.LL	0.06	589.3	24.1	89.3	13.7	61.2	24.1	65.8
	0.1	380.7	48.3	107.1	19.4	58.6	32.5	62.9
	0.14	355.7	40.2	88.6	10.0	36.2	16.8	38.1
	0.2	212.5	76.3	117.9	17.8	45.1	28.7	47.3
	0.3	200.7	76.5	105.1	10.7	26.9	17.2	27.6
AR.YL	0.06	583.3	23.8	83.3	3.2	79.5	3.2	61.6
	0.1	375.1	53.3	106.7	5.4	78.6	5.4	59.7
	0.14	351.3	53.3	102.7	2.4	49.4	2.4	37.1
	0.2	207.7	107.3	153.1	4.1	66.2	4.1	47.2
	0.3	197.5	142.0	199.4	1.9	41.1	1.9	28.9
AM.YL	0.06	571.4	41.6	103.5	-8.0	61.2	-9.2	43.3
	0.1	363.3	95.0	161.4	-11.3	62.9	-13.2	44.1
	0.14	342.0	97.2	179.7	-6.7	40.4	-7.8	29.3
	0.2	197.0	198.4	274.6	-14.5	58.2	-16.7	41.7
	0.3	191.4	270.2	362.4	-11.5	38.4	-13.1	28.7

Table 6.7b
MSE, ARB & RRMSE (percentage) of MSE estimators, $m = 15$ areas

			RB_Y1		RB_Y2		M_et_al		PB estimator		Naïve PB estimator	
	σ_v^2/ψ_i	$\overline{\text{MSE}}$	ARB	RRMSE	ARB	RRMSE	ARB	RRMSE	%ARB	%RRMSE	%ARB	%RRMSE
MIX	0.06	584.9	21.0	84.7	35.4	93.7	12.6	87.9	10.0	53.8	19.3	62.1
	0.1	377.1	46.0	103.9	68.4	122.6	26.4	106.1	14.8	52.7	26.6	59.9
	0.14	353.0	41.9	91.5	59.4	112.7	25.0	89.9	7.6	33.2	13.7	36.7
	0.2	209.7	83.2	127.8	108.9	155.8	42.8	122.8	14.0	42.5	23.7	46.0
	0.3	198.9	94.8	136.7	117.1	162.2	50.4	124.6	8.7	26.6	14.5	27.7

Summarizing, under the Fay-Herriot model with positive σ_v^2, and among the positive variance estimators under study, the MIX and the AR.YL variance estimators are the only ones with negligible asymptotic bias. The AM.YL and the LL variance estimators have a larger asymptotic bias. On the other hand, our simulation showed that for a moderate number of areas and for populations that yield zero REML estimates, both YL variance estimators were negatively biased, and produced EBLUPs that were close to the synthetic estimator of the mean. In contrast, the MIX, built as the combination of the AM.LL and the REML, was only mildly negatively biased in these populations. Moreover, the unconditional distribution of the MIX approached normality much faster than those of the other variance estimators.

Table 6.8
MSE_C and ARB_C m = 15 areas

Method	σ_v^2/ψ_i	$\overline{MSE_C}$	Taylor MSE estimator			PB estimator	Naïve PB estimator
REML	0.06	594.2		-22.6		-31.7	-16.5
	0.1	381.2		-32.9		-43.2	-22.5
	0.14	345.1		-17.7		-22.7	-10.7
	0.2	212.7		-41.1		-47.3	-25.5
	0.3	197.9		-30.4		-32.7	-17.6
AM.LL	0.06	595.6		-4.1		-5.7	12.1
	0.1	385.7		8.6		-7.0	15.6
	0.14	351.2		18.9		-2.0	10.4
	0.2	216.0		46.4		-5.8	14.4
	0.3	199.5		67.0		-2.9	9.8
AR.YL	0.06	592.2		-0.8		-27.1	-11.0
	0.1	379.7		21.0		-36.5	-14.8
	0.14	344.5		44.0		-18.6	-6.3
	0.2	210.9		98.2		-38.6	-16.4
	0.3	196.6		177.3		-26.1	-11.0
AM.YL	0.06	581.7		30.7		-21.9	-18.0
	0.1	368.6		79.8		-31.5	-25.8
	0.14	333.9		98.3		-15.2	-11.9
	0.2	198.9		198.0		-36.4	-30.0
	0.3	190.0		296.3		-26.2	-21.5
			RB_Y1	**RB_Y2**	**M_et_al**	**PB estimator**	**Naive PB estimator**
MIX	0.06	595.6	-4.1	27.9	-22.9	3.4	17.8
	0.1	385.7	8.6	57.1	-33.7	5.1	22.8
	0.14	351.2	18.9	58.5	-19.1	4.9	14.3
	0.2	216.0	46.4	102.4	-42.0	5.9	20.4
	0.3	199.5	67.0	116.3	-30.9	4.8	13.4

In terms of MSE of the EBLUP, there were considerable gains in precision over the direct estimator, under all methods of variance estimation considered here, even for a small number of areas. The AM.LL and both the AM.YL and the AR.YL variance estimators carried lower variability than the REML and the MIX. It impacted only minimally the MSE of the EBLUP: differences among MSEs for the same signal to noise ratio were small. These differences widened as either the number of areas or the signal to noise ratio decreased. Thus, it may possible that for an extremely low signal to noise ratio, the MSE under MIX would be somewhat larger than under the AM.YL variance estimator.

Under the MIX method of variance estimation, we compared three different Taylor-type MSE estimators and two bootstrap MSE estimators. All three Taylor estimators of the MSE under MIX (RB_Y1, RB_Y2

and M_et_al) are unbiased up to the second order. Also the Taylor-type estimators of the MSE under the LL and the YL are unbiased up to the second order. RB_Y1, AM.LL and AM.YL may yield negative MSE estimates.

The Taylor MSE under the REML method of variance estimation and the M_et_al under the MIX coincide by definition, hence their performance measures have negligible differences (their true MSEs are different, however in our study, for $m = 100,$ the MIX coincided with the REML 84% of the time). For a moderate number of areas, which for this data could be $m = 45$ or $100,$ and for populations that yield zero REML estimates, both the Taylor MSE estimators under the REML and the M_et_al MSE estimators do not account for the variation due to the estimation of σ_v^2 and this is reflected in their very negative $\text{ARB}_{\mathbb{C}}$, which is below -60% for the smaller signal to noise ratios. On the other hand, the RB_Y1 does account for the variation due to the estimation of $\sigma_v^2,$ but its $\text{ARB}_{\mathbb{C}}$ is also very negative: the RB_Y1 is a split MSE estimator that for populations with $\hat{\sigma}^2_{v\text{REML}} = 0,$ it subtracts a factor of the unconditional bias of the AM.LL, which is always positive, whereas a better formula for a split MSE estimator would be to use an estimator of the conditional bias $E\left(\hat{\sigma}_v^2 \big/ \hat{\sigma}^2_{v\text{REML}} = 0\right).$ Indeed, even for a moderate number of areas $(m = 100),$ Table 6.1 shows that the unconditional bias of the MIX is 49% whereas the conditional bias of the MIX is -37%.

The PB MSE estimator under the AR.YL and the MIX methods adjusted well for the bias but paid in terms of variance. Among all the MSE estimators it appears that the Naive Bootstrap MSE estimator performed best, and even better under the MIX variance estimation, when taking into account the three measures ARB, $\text{ARB}_{\mathbb{C}}$ and RRMSE together. We found that for a moderate number of areas, the RB_Y2 had the lowest RRMSE among the Taylor estimators under the MIX method. On the other hand, M_et_al is most reliable when the true underlying variance σ_v^2 is very small: in this case M_et_al is effectively the MSE estimator of the synthetic estimator of the small area mean. We do not recommend relying on the second order approximation to the MSE when m is small: the approximation (2.6) to the MSE does not necessarily hold, the performance measures obtained from our study are very unstable and they may vary from data set to data set.

In conclusion, under the hypothesis of $\sigma_v^2 > 0,$ the relative performances of competing positive variance estimators depend on the size of $\sigma_v^2,$ the signal to noise ratio, the number of areas and the objective function. For a moderate number of areas, the MIX variance estimator appeared to perform better than the LL and the YL estimators in this study; under the MIX method, the Naive PB MSE estimator had the lowest $\text{ARB}_{\mathbb{C}}$ and RRMSE combined; the M_et_al MSE estimator under the MIX variance estimator performed marginally better than the RB_Y1 when the underlying σ_v^2 was very small. However, the percentage of REML zeros yielded under the simulation model shows that an outcome of $\hat{\sigma}^2_{v\text{REML}} = 0$ and/or negative tests of hypothesis do not necessarily mean that σ_v^2 is sufficiently small to rely on M_et_al. In the absence of other information, the Naive PB estimator under the MIX appears to perform better.

Acknowledgements

The authors would like to thank Professor J.N.K. Rao from Carleton University for his useful comments and to Victor Estevao from Statistics Canada for developing the grid maximization especially for this

project. We also would like to thank the reviewers for their careful appraisal of our paper and for their suggestions to improve this paper.

Appendix A

Proof of Theorem 4.1

The asymptotic variance of $\hat{\sigma}^2_{v\text{MIX}}$ is given by: $\overline{V}\left(\hat{\sigma}^2_{v\text{MIX}}\right) = \lim_{m\to\infty} E\left(\hat{\sigma}^2_{v\text{MIX}} - \sigma^2_v\right)^2$

We show that $E\left(\hat{\sigma}^2_{v\text{MIX}} - \sigma^2_v\right)^2 \le E\left(\hat{\sigma}^2_{v\text{REML}} - \sigma^2_v\right)^2 + o(1/m)$ as $m \to \infty$.

$$\begin{aligned} E\left(\hat{\sigma}^2_{v\text{MIX}} - \sigma^2_v\right)^2 &= \int_{\left\{\hat{\sigma}^2_{v\text{REML}} > 0\right\}} \left(\hat{\sigma}^2_{v\text{REML}} - \sigma^2_v\right)^2 dP + \int_{\left\{\hat{\sigma}^2_{v\text{REML}} = 0\right\}} \left(\hat{\sigma}^2_{v\text{AM.LL}} - \sigma^2_v\right)^2 dP \\ &\le \int_{\Omega} \left(\hat{\sigma}^2_{v\text{REML}} - \sigma^2_v\right)^2 dP + \int_{\left\{\hat{\sigma}^2_{v\text{REML}} = 0\right\}} \left(\hat{\sigma}^2_{v\text{AM.LL}} - \sigma^2_v\right)^2 dP = E\left(\hat{\sigma}^2_{v\text{REML}} - \sigma^2_v\right)^2 \\ &\quad + o\left(\frac{1}{m}\right). \end{aligned} \tag{A.1}$$

Indeed, by the Holder and Minkowski inequalities, with any $1 < p < \infty, 1/p + 1/q = 1$, and setting $X \equiv \left(\hat{\sigma}^2_{v\text{AM.LL}} - \sigma^2_v\right)^2 = \text{O}_p(1/m)$ and the indicator $I\left(\hat{\sigma}^2_{v\text{REML}} = 0\right)$ of populations with $\hat{\sigma}^2_{v\text{REML}} = 0$, we have:

$$\begin{aligned} \int_{\left\{\hat{\sigma}^2_{v\text{REML}} < 0\right\}} \left(\hat{\sigma}^2_{v\text{AM.LL}} - \sigma^2_v\right)^2 dP &\le \left(\int_{\Omega} \left(\hat{\sigma}^2_{v\text{AM.LL}} - \sigma^2_v\right)^{2p} dP\right)^{1/p} \cdot \left(P\left\{\hat{\sigma}^2_{v\text{REML}} = 0\right\}\right)^{1/q} \\ &= \left(O\left(\frac{1}{m^p}\right)\right)^{1/p} \cdot (o(1))^{1/q} = o\left(\frac{1}{m}\right), \end{aligned} \tag{A.2}$$

since $\left(\hat{\sigma}^2_{v\text{AM.LL}} - \sigma^2_v\right)^2$ is uniformly bounded and $\hat{\sigma}^2_{v\text{REML}} \xrightarrow{P} \sigma^2_v > 0$. Note that the AM.LL and REML estimators of σ^2_v are uniformly bounded as a consequence of their almost sure convergence to σ^2_v (see, for example, Yuan and Jennrich 1998).

Proof of Theorem 4.2

We denote by $\hat{\boldsymbol{\sigma}}^2_{v\text{ML}}$ the maximum likelihood variance estimator.

We show first that $\hat{\boldsymbol{\sigma}}^2_{v\text{REML}} - \hat{\boldsymbol{\sigma}}^2_{v\text{ML}} = O_p(1/m)$. Let $G_*\left(\boldsymbol{\sigma}^2_v\right) = \partial \log\left(L_*\right)/\partial \boldsymbol{\sigma}^2_v = 0$ be the estimating equation that yields the variance estimator *. Equation (3.4) implies:

$$G_{\text{AM.LL}}\left(\boldsymbol{\sigma}^2_v\right) - G_{\text{ML}}\left(\boldsymbol{\sigma}^2_v\right) = \partial \log \boldsymbol{\sigma}^2_v / \partial\, \boldsymbol{\sigma}^2_v = \frac{1}{m\boldsymbol{\sigma}^2_v} = O\left(\frac{1}{m}\right). \tag{A.3}$$

With $G'_{\text{ML}}(\cdot) \triangleq \left(\partial G_{\text{ML}} / \partial \boldsymbol{\sigma}^2_v\right)(\cdot)$ and $G''_{\text{ML}}(\cdot) \triangleq \left(\partial G'_{\text{ML}} / \partial \boldsymbol{\sigma}^2_v\right)(\cdot)$, equation (A.3) implies:

$$G'_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right) - G'_{\mathrm{AM.LL}}\left(\boldsymbol{\sigma}_v^2\right) = O\left(\frac{1}{m}\right). \tag{A.4}$$

Now, using equation (A.4), the $\sqrt{m}$ −consistency of the ML and AM.LL estimators of $\boldsymbol{\sigma}_v^2$, the two-term Taylor expansion of $G_{\mathrm{ML}}(\cdot)$ and $G_{\mathrm{AM.LL}}(\cdot)$ at $\boldsymbol{\sigma}_v^2$ and $G'_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right) = O(1)$ as $m \to \infty$, the left- hand side in (A.3) is equal to:

$$\begin{aligned}
&= G'_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right)\left(\hat{\boldsymbol{\sigma}}^2_{v\mathrm{ML}} - \boldsymbol{\sigma}_v^2\right) - G'_{\mathrm{AM.LL}}\left(\boldsymbol{\sigma}_v^2\right)\left(\hat{\boldsymbol{\sigma}}^2_{v\mathrm{AM.LL}} - \boldsymbol{\sigma}_v^2\right) + O_p\left(\frac{1}{m}\right) \\
&= G'_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right)\left(\hat{\boldsymbol{\sigma}}^2_{v\mathrm{ML}} - \hat{\boldsymbol{\sigma}}^2_{v\mathrm{AM.LL}}\right) + \left(G'_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right) - G'_{\mathrm{AM.LL}}\left(\boldsymbol{\sigma}_v^2\right)\right)\left(\hat{\boldsymbol{\sigma}}^2_{v\mathrm{AM.LL}} - \boldsymbol{\sigma}_v^2\right) + O_p\left(\frac{1}{m}\right) \\
&= G'_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right)\left(\hat{\boldsymbol{\sigma}}^2_{v\mathrm{ML}} - \hat{\boldsymbol{\sigma}}^2_{v\mathrm{AM.LL}}\right) + O_p\left(\frac{1}{m^{3/2}}\right) + O_p\left(\frac{1}{m}\right).
\end{aligned}$$

The last equality above implies

$$\hat{\boldsymbol{\sigma}}^2_{v\mathrm{AM.LL}} - \hat{\boldsymbol{\sigma}}^2_{v\mathrm{ML}} = O_p\left(\frac{1}{m}\right) \text{ as } m \to \infty. \tag{A.5}$$

Similarly, we establish a relationship between $G_{\mathrm{REML}}\left(\boldsymbol{\sigma}_v^2\right)$ and $G_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right)$: given that $\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{V}^{-1}\right) = O(1)$ follows from conditions 1 through 3 in Section 3 and equation (3.1), we have:

$$G_{\mathrm{REML}}\left(\boldsymbol{\sigma}_v^2\right) - G_{\mathrm{ML}}\left(\boldsymbol{\sigma}_v^2\right) = \frac{1}{m}\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{V}^{-1}\right) = O\left(\frac{1}{m}\right) \text{ as } m \to \infty, \tag{A.6}$$

Equation (A.6) and the same argument as with the AM.LL estimator, imply:

$$\hat{\boldsymbol{\sigma}}^2_{v\mathrm{REML}} - \hat{\boldsymbol{\sigma}}^2_{v\mathrm{ML}} = O_p\left(\frac{1}{m}\right) \text{ as } m \to \infty. \tag{A.7}$$

Equations (A.5) and (A.7) combined, yield:

$$\left(\hat{\sigma}^2_{v\mathrm{REML}} - \hat{\sigma}^2_{v\mathrm{AM.LL}}\right) = O_P\left(\frac{1}{m}\right). \tag{A.8}$$

Now we express the bias of the MIX estimator by:

$$B_{\mathrm{MIX}}\left(\hat{\sigma}^2_{v\mathrm{MIX}}\right) = \int_{\left\{\hat{\sigma}^2_{v\mathrm{REML}} > 0\right\}} \left(\hat{\sigma}^2_{v\mathrm{REML}} - \sigma_v^2\right) dP + \int_{\left\{\hat{\sigma}^2_{v\mathrm{REML}} = 0\right\}} \left(\hat{\sigma}^2_{v\mathrm{AM.LL}} - \sigma_v^2\right) dP.$$

We add and subtract $\int_{\left\{\hat{\sigma}^2_{v\mathrm{REML}} = 0\right\}} \left(\hat{\sigma}^2_{v\mathrm{REML}} - \sigma_v^2\right) dP$ from the right-hand side of the equation above to obtain:

$$\begin{aligned}
B_{\mathrm{MIX}}\left(\hat{\sigma}^2_{v\mathrm{MIX}}\right) &= \int_{\Omega} \left(\hat{\sigma}^2_{v\mathrm{REML}} - \sigma_v^2\right) dP + \int_{\left\{\hat{\sigma}^2_{v\mathrm{REML}} = 0\right\}} \left(\hat{\sigma}^2_{v\mathrm{AM.LL}} - \hat{\sigma}^2_{v\mathrm{REML}}\right) dP \\
&= \mathrm{Bias}\left(\hat{\sigma}^2_{v\mathrm{REML}}\right) + \int_{\left\{\hat{\sigma}^2_{v\mathrm{REML}} = 0\right\}} \left(\hat{\sigma}^2_{v\mathrm{AM.LL}} - \hat{\sigma}^2_{v\mathrm{REML}}\right) dP.
\end{aligned} \tag{A.9}$$

Now, since $\hat{\sigma}^2_{v\text{AM.LL}} - \hat{\sigma}^2_{v\text{REML}}$ is uniformly bounded, we apply the Holder and Minkowski inequality with $p = q = 2$ and equation (A.8) to the last term in (A.9) to obtain:

$$\begin{aligned} B_{\text{MIX}}\left(\hat{\sigma}^2_{v\text{MIX}}\right) &= \text{Bias}\left(\hat{\sigma}^2_{v\text{REML}}\right) + \left(\int_{\Omega}\left(\hat{\sigma}^2_{v\text{AM.LL}} - \hat{\sigma}^2_{v\text{REML}}\right)^2 dP\right)^{1/2} \cdot P\left\{\hat{\sigma}^2_{v\text{REML}} = 0\right\}^{1/2} \\ &= \text{Bias}\left(\hat{\sigma}^2_{v\text{REML}}\right) + O\left(\frac{1}{m}\right) \cdot o(1) = \text{Bias}\left(\hat{\sigma}^2_{v\text{REML}}\right) + o\left(\frac{1}{m}\right). \end{aligned} \tag{A.10}$$

Proof of Remark 4.2: mse_0 is unbiased up to the second order

$$\begin{aligned} E(\text{mse}_0) - \text{MSE}\left(\hat{\boldsymbol{\theta}}_i\right) &= \int_{\left\{\hat{\boldsymbol{\sigma}}^2_{v\text{REML}} > 0\right\}} (g_{1i} + g_{2i} + 2g_{3i})\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) dP + \int_{\left\{\hat{\boldsymbol{\sigma}}^2_{v\text{REML}} = 0\right\}} g_{2i}\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) dP - \text{MSE} \\ &= \left[\int_{\Omega} (g_{1i} + g_{2i} + 2g_{3i})\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) dP - \text{MSE}\right] \\ &\quad + \int_{\left\{\hat{\boldsymbol{\sigma}}^2_{v\text{REML}} = 0\right\}} g_{2i}\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) dP - \int_{\left\{\hat{\boldsymbol{\sigma}}^2_{v\text{REML}} = 0\right\}} (g_{1i} + g_{2i} + 2g_{3i})\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) dP \\ &= \left[\text{o}\left(\frac{1}{m}\right)\right] - \int_{\left\{\hat{\boldsymbol{\sigma}}^2_{v\text{REML}} = 0\right\}} 2g_{3i}\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) dP, \end{aligned} \tag{A.11}$$

since $g_{1i}\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) = g_{1i}(0) = 0$ in $\left\{\hat{\boldsymbol{\sigma}}^2_{v\text{REML}} = 0\right\}$ and $g_{2i}\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right)$ cancels out in (A.11). But

$$g_{3i}\left(\hat{\boldsymbol{\sigma}}^2_{v\text{REML}}\right) = g_{3i}(0) = \frac{\bar{V}(0)}{\boldsymbol{\psi}_i} = O_p\left(\frac{1}{m}\right)$$

and is uniformly bounded under the regularity conditions given in Section 2, hence the last term in (A.11) is also an $\text{o}(1/m)$, which renders mse_0 unbiased up to the second order.

Appendix B

B.1 Comparison between REML and AR.YL using the scoring algorithm

The scoring algorithm could sometimes yield zero estimates for the likelihood of the AR.YL. Indeed, for data sets simulated under the model given in Section 5, with $m = 45$ and $\sigma^2_v = 1$, the REML and AR.YL scoring algorithms yielded 28% and 26% zeros respectively. Figures B.1 to B.3 illustrate the why: the likelihoods correspond to a single population generated under the model with $\boldsymbol{\sigma}^2_v = 1$ for which $\hat{\boldsymbol{\sigma}}^2_{v\text{REML}} = 0$.

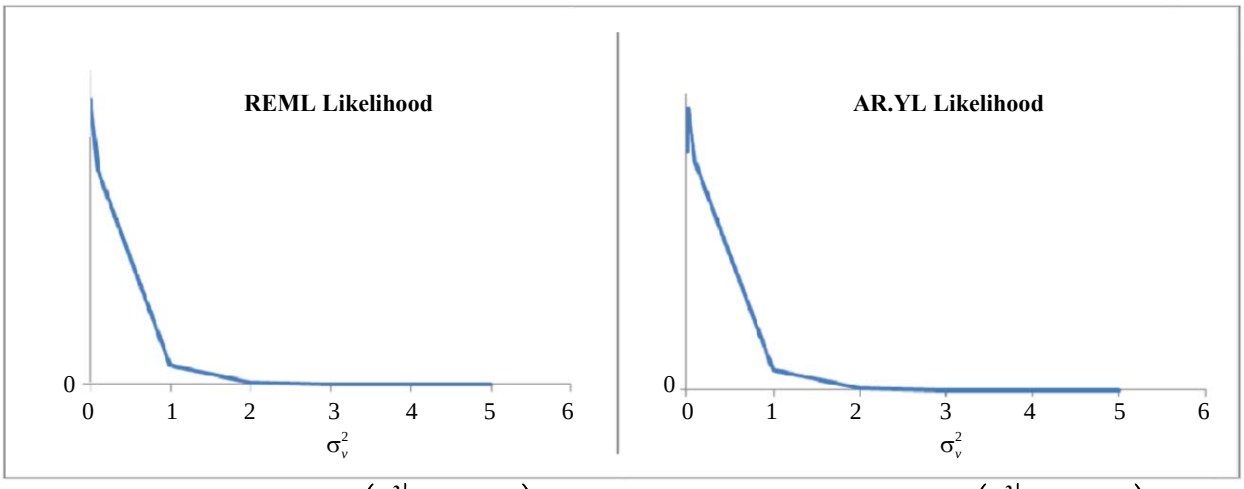

Figure B.1 $L = L_{\text{REML}}\left(\sigma_v^2 \middle| y_1, \ldots, y_{45}\right)$.

Figure B.2 $L = L_{\text{AR.YL}}\left(\sigma_v^2 \middle| y_1, \ldots, y_{45}\right)$.

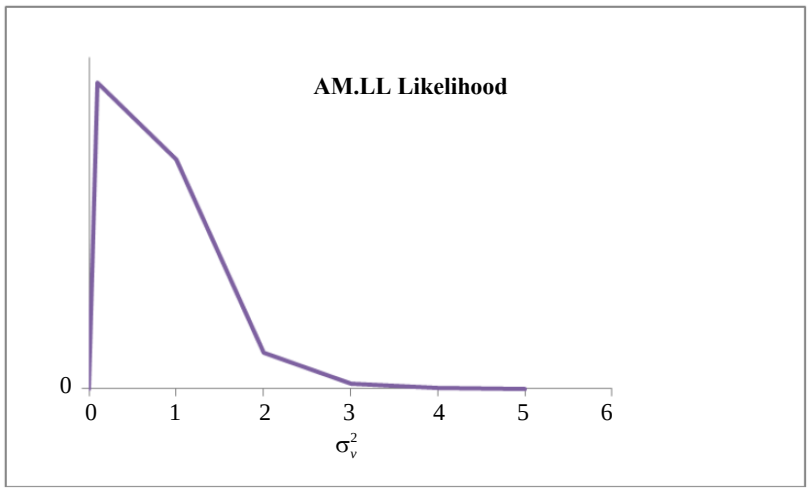

Figure B.3 $L = L_{\text{AM.LL}}\left(\sigma_v^2 \middle| y_1, \ldots, y_{45}\right)$.

Figure B.2 shows that the maximum value of the AR.YL likelihood is very near the border. The scoring algorithm may often miss the maximum and yield a zero value. Figure B.3 shows that the AM.LL likelihood has a maximum value that differentiates better from the border.

B.2 Treatment of zeros in the parametric bootstrap

For each estimate $\hat{\sigma}_v^2 = \hat{\sigma}_v^2\left(\mathbf{y}^{(r)}\right), r = 1, \ldots 10K,$ and each method of variance estimation:

i. Generate a large number B of random area effects $v_i^{(b)} \overset{\text{i.i.d.}}{\sim} N\left(0, \hat{\sigma}_v^2\right), b = 1, \ldots, B,$ and generate, independently of $v_i^{(b)}$, sampling errors $e_i^{(b)} \overset{\text{i.i.d.}}{\sim} N\left(0, \psi_i\right)$, $i = 1, \ldots, m, b = 1, \ldots, B.$ Generate

bootstrap data $y_i^{(b)} = \theta_i^{(b)} + e_i^{(b)}, \theta_i^{(b)} = \mathbf{x}_i'\hat{\beta} + v_i^{(b)}, i = 1,\ldots,m.$ If $\hat{\sigma}^2_{v\text{REML}}\left(y^{(r)}\right) = 0,$ then generate $\left(y_i^{(b)}, \theta_i^{(b)}\right),\ b = 1,\ldots,B,$ from the synthetic model (see also Rao and Molina 2015).

ii. Fit the model to the bootstrap data and obtain $\hat{\sigma}_v^{2(b)}$; for the MIX estimator calculate $\hat{\sigma}^{2(b)}_{v\text{MIX}} = \hat{\sigma}^{2(b)}_{v\text{REML}}$ if $\hat{\sigma}_v^{2(b)}$ is positive and $\hat{\sigma}^{2(b)}_{v\text{MIX}} = \hat{\sigma}^{2(b)}_{v\text{AM}}$ otherwise.

iii. Now obtain $\hat{\beta}^{(b)}$, the corresponding EBLUP $\hat{\theta}_i^{(b)}$, the bootstrap components $g_{1i}^{(b)} = g_{1i}\left(\hat{\sigma}_v^2\left(y^{(b)}\right)\right),\ g_{2i}^{(b)} = g_{2i}\left(\hat{\sigma}_v^2\left(y^{(b)}\right)\right)$ and $\overline{g}_{ji}^{\text{PB}} = B^{-1}\sum_b g_{ji}^{(b)},\ j = 1,2.$

iv. The Naive MSE bootstrap estimator is $\text{mse}_{\text{naive}} = B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_i^{(b)} - \theta_i^{(b)}\right)^2.$

v. The PB MSE estimator (which is adjusted for bias (Pfeffermann and Glickman 2004) is: $\text{mse}_{\text{PB}}\left(\hat{\theta}_i\right) = g_{1i}\left(\hat{\sigma}_v^2\right) + g_{2i}\left(\hat{\sigma}_v^2\right) - \overline{g}_{1i}^{\text{PB}} - \overline{g}_{2i}^{\text{PB}} + \text{mse}_{\text{naive}}.$

vi. To calculate $\text{ARB}_{\mathbb{C}}$, average $\left(\text{mse}_{\text{PB}}^{(r)}\left(\hat{\theta}_i\right) - \text{MSE}\left(\hat{\theta}_i\right)\right)\Big/\text{MSE}\left(\hat{\theta}_i\right)$ over the populations with $(r)\Big/\hat{\sigma}^2_{v\text{REML}}\left(\mathbf{y}^{(r)}\right) = 0$ and do similarly with $\text{ARB}_{\mathbb{C}}$ of $\text{mse}_{\text{naive}}.$

References

Chen, S., and Lahiri, P. (2008). On mean squared prediction error estimation in small area estimation problems. *Communications in Statistics-Theory and Methods,* 37, 1792-1798.

Chen, S., and Lahiri, P. (2011). On the estimation of Mean Squared Prediction Error in small area estimation. *Calcutta Statistical Association Bulletin,* 63, (Special 7th Triennial Proceedings Volume), Nos. 249-252.

Cressie, N. (1992). REML estimation in empirical Bayes smoothing of census undercount. *Survey Methodology*, 18, 1, 75-94.

Das, K., Jiang, J. and Rao, J.N.K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, 32, 2, 818-840.

Datta, G., and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.

Estevao, V. (2014). Grid optimization algorithm for maximum likelihood. Internal report, Statistical Research and Innovation Division (SRID), Statistics Canada.

Fay, R.E., and Herriot, R.A. (1979). Estimation of income from small places: An application of James-Stein Procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Lahiri, P., and Li, H. (2009). Generalized maximum likelihood method in linear mixed models with an application in small area estimation. In *Proceedings of the Federal Committee on Statistical Methodology Research Conference,* available at http//www.fcsm.gov/events/papers2009.html.

Lahiri, P., and Pramanik, S. (2011). Discussion of "Estimating random effects via adjustment for density maximization" by C. Morris and R. Tang. *Statistical Science*, 26, 2, 291-295.

Li, H., and Lahiri, P. (2011). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101, 882-892.

Molina, I., Rao, J.N.K. and Datta, G.S. (2015). Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random effects. *Survey Methodology*, 41, 1, 1-19.

Morris, C.N. (2006). Mixed model prediction and small area estimation (with discussions). *Test*, 15, 72-76.

Pfeffermann, D., and Glickman, H. (2004). Mean squarred error approximation in small area estimation by use of parametric and non-parametric bootstrap. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, Alexandria, VA. 4167-78.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation, second edition*. New York: John Wiley & Sons, Inc.

Rubin-Bleuer, S., and Schiopu-Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *The Annals of Statistics*, 33, 6, 2789-2810.

Rubin-Bleuer, S., and You, Y. (2012). A positive variance estimator for the Fay-Herriot small area model. SRID-2012-009E, Statistical Research and Innovation Division (SRID), Statistics Canada.

Rubin-Bleuer, S., Yung, W. and Landry, S. (2010). Adjusted maximum likelihood method for a small area model accounting for time and area effects. SRID-2010-006E, Statistical Research and Innovation Division (SRID), Statistics Canada.

Rubin-Bleuer, S., Yung, W. and Landry, S. (2011). Adjusted maximum likelihood method for a small area model accounting for time and area effects. Long abstract, *Small Area Estimation*, (SAE 20122) in Trier, Germany, International Statistical Institute Satellite Conference.

Rubin-Bleuer, S., Yung, W. and Landry, S. (2012). Variance Component Estimation through the Adjusted Maximum Likelihood Approach. Presentation at the Conference in Honour of the 75th birthday of J.N.K. Rao Carleton University, May 2012, Ottawa.

Yoshimori, M., and Lahiri, P. (2014). A new adjusted maximum likelihood method for the Fay-Herriot small area model. *Journal of Multivariate Analysis*, 124, 281-294.

Yuan, P. (2009). Comparison of SAE methods of variance estimation. Internal document, Statistical Research and Innovation Division (SRID), Statistics Canada.

Yuan, K.H., and Jennrich, R. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, 65, 2, 245-260.

Survey Methodology, June 2016
Vol. 42, No. 1, pp. 87-120
Statistics Canada, Catalogue No. 12-001-X

A comparison between nonparametric estimators for finite population distribution functions

Leo Pasquazzi and Lucio de Capitani[1]

Abstract

In this work we compare nonparametric estimators for finite population distribution functions based on two types of fitted values: the fitted values from the well-known Kuo estimator and a modified version of them, which incorporates a nonparametric estimate for the mean regression function. For each type of fitted values we consider the corresponding model-based estimator and, after incorporating design weights, the corresponding generalized difference estimator. We show under fairly general conditions that the leading term in the model mean square error is not affected by the modification of the fitted values, even though it slows down the convergence rate for the model bias. Second order terms of the model mean square errors are difficult to obtain and will not be derived in the present paper. It remains thus an open question whether the modified fitted values bring about some benefit from the model-based perspective. We discuss also design-based properties of the estimators and propose a variance estimator for the generalized difference estimator based on the modified fitted values. Finally, we perform a simulation study. The simulation results suggest that the modified fitted values lead to a considerable reduction of the design mean square error if the sample size is small.

Key Words: Finite population sampling; Distribution function estimator; Fitted values; Kuo estimator.

1 Introduction

Since Chambers and Dunstan's seminal paper Chambers and Dunstan (1986), several estimators for finite population distribution functions have been proposed. Most of them are based either on different types of fitted values or on different ways to combine them into an estimator. The estimator proposed by Chambers and Dunstan (1986), for example, is based on fitted values derived from a superpopulation model where the study variable and an auxiliary variable are linked by a linear regression model with independent error components whose variances are assumed to be known. Substituting the fitted values to the unobserved indicator functions in the definition of the population distribution function of the study variable yields the Chambers and Dunstan estimator. Rao, Kovar and Mantel (1990) incorporate design weights into the fitted values of Chambers and Dunstan and use them in a generalized difference estimator. Kuo (1988) uses nonparametric regression to estimate directly the regression relationship between the indicator functions and the auxiliary variable and obtains fitted values that accommodate virtually any superpopolation model. Like Chambers and Dunstan, she substitutes the unobserved indicator functions with their corresponding fitted values and obtains a model-based estimator. Chambers, Dorfman and Wehrly (1993) combine the fitted values of Chambers and Dunstan (1986) and of Kuo (1988) and propose still another model-based estimator that aims to be more efficient than the Kuo estimator if the linear superpopulation model assumed by Chambers and Dunstan is true, and that does not suffer from model misspecification bias otherwise. Following these early works there has been quite a large number of subsequent proposals with the aim to achieve some gain in efficiency with respect to the Horvitz-Thompson estimator, while preserving robustness and sometimes also one or both of the following desirable properties shared by the Horvitz-Thompson estimator: (i) the fact that it is a linear combination of the sample indicator functions with

1. Leo Pasquazzi and Lucio de Capitani, Università degli Studi di Milano-Bicocca, Milan, Italy. E-mail: leo.pasquazzi@unimib.it, lucio.decapitani1@unimib.it.

coefficients that do not depend on the study variable and (ii) the fact that it gives always rise to nondecreasing estimates for the distribution function.

The present work originates from the idea to improve upon the fitted values proposed by Kuo (1988) through incorporation of an estimate for the mean regression function (see Section 2). This idea has been put forward in a recent textbook of Chambers and Clark (2012) and it is based on the assumption of an underlying superpopulation model with smooth regression relationship between the study variable and an auxiliary variable and with smoothly varying error component distributions. According to this idea, the fitted values are the outcome of a two-step procedure: at the first step the mean regression function is estimated through either parametric or nonparametric regression, and at the second step, using the regression residuals from the first step, the distribution functions of the error components are estimated using nonparametric regression in order to accommodate the possibility of smoothly varying error component distributions. Combining both estimates one may compute fitted values for the indicator functions in the definition of the finite population distribution function of the study variable. Chambers and Clark (2012) analyze the model-based estimator that is obtained by substituting the unobserved indicator functions by their corresponding fitted values and they sketch a proof that leads to an expression for the model variance of the resulting estimator. In that proof they assume that the mean regression function is estimated by a consistent estimator and that the contribution from its estimation error to the model variance of the final distribution function estimator can be neglected. In the present work we consider local linear regression for estimating both the model mean regression function and the error component distributions. We provide asymptotic expansions for the model bias and the model variance of the resulting estimator and compare them with those corresponding to the Kuo estimator based on local linear regression. It turns out that the leading terms in the model variances are the same and that, for appropriately chosen bandwidth sequences, the squared model bias of both estimators goes to zero faster than the model variance. To establish which estimator is asymptotically more efficient from the model-based perspective thus requires knowledge of the second order terms of the model variances. The latter however depend on more specific assumptions than those considered in the present work and, at least for the estimator based on the modified fitted values, it seems no easy task to determine the second order terms of the model variances. Which estimator is more efficient from the model-based perspective remains thus an open question.

In addition to the above model-based estimators, we analyze also the generalized difference estimators based on both types of fitted values in their design weighted versions. The results in Section 3 show that the convergence rates of their model biases and their model variances are the same as those of their model-based counterparts. As for design-based properties, they are discussed to some extent in Section 4 along with the issue of variance estimation. It would of course be of interest to derive and compare asymptotic expansions for the design biases and the design variances. Breidt and Opsomer (2000) derive under mild conditions a general expression for the first order term in the design mean square error of local polynomial regression estimators, of which the generalized difference estimator based on the fitted values of Kuo is a special case. The generalized difference estimator based on the modified fitted values does however not fall into this class. In line with Särndal, Swensson and Wretman (1992), we conjecture that under broad conditions the first order term of its design mean square error is the same as the one of the generalized difference estimator based on the fitted values of Kuo. Formal proofs could perhaps be obtained by adapting and extending some of the results in Wang and Opsomer (2011). To test this conjecture and to compare the performance of the generalized difference and the model-based estimators in various settings, we performed a simulation study whose results are presented in Section 5.

2 Definition of the estimators

Let (y_i, x_i) denote the values taken on by a study variable Y and an auxiliary variable X on unit i of a finite population $U := \{1, 2, \ldots, N\}$. Suppose that

$$y_i = m(x_i) + \varepsilon_i, \qquad i \in U, \tag{2.1}$$

where $m(x)$ is a smooth function and where the ε_i's are independent zero mean random variables whose distribution functions $P(\varepsilon_i \le \varepsilon) = G(\varepsilon | x_i)$ depend smoothly on x_i. Let $s \subset U$ be a sample chosen from the population U according to some sample design. As usual in the context of complete auxiliary information we assume that the x_i – values are known for all population units, while the y_i – values are observed only for the population units which belong to the sample s.

To estimate the unknown population distribution function

$$F_N(t) := \frac{1}{N} \sum_{i \in U} I(y_i \le t),$$

Kuo (1988) proposes the estimator given by

$$\hat{F}(t) := \frac{1}{N} \left(\sum_{j \in s} I(y_j \le t) + \sum_{i \notin s} \sum_{j \in s} w_{i,j} I(y_j \le t) \right), \tag{2.2}$$

where in place of $w_{i,j}$ she suggests to use either the local constant regression weights

$$w_{i,j} := \frac{K\left(\frac{x_i - x_j}{\lambda}\right)}{\sum_{k \in s} K\left(\frac{x_i - x_k}{\lambda}\right)}$$

with some (integrable) kernel function in place of $K(u)$ and $\lambda > 0$, or the nearest k neighbor weights

$$w_{i,j} := \begin{cases} 1/k, & \text{if } x_j \text{ is one of the } k \text{ nearest neighbors to } x_i \\ 0, & \text{otherwise.} \end{cases}$$

Note that in the definition $\hat{F}(t)$,

$$\hat{G}_i(t) := \sum_{j \in s} w_{i,j} I(y_j \le t) \tag{2.3}$$

is used as the fitted value in place of the unobserved indicator function $I(y_i \le t)$ for $i \notin s$.

Following an idea put forward in the textbook of Chambers and Clark (2012), we shall analyze an estimator for $F_N(t)$ based on alternative fitted values which incorporate a nonparametric estimate for the mean regression function $m(x)$. The fitted values in question are given by

$$\hat{G}_i^*(t) := \sum_{j \in s} w_{i,j} I(y_j - \hat{m}_j \le t - \hat{m}_i) \tag{2.4}$$

where

$$\hat{m}_i := \sum_{k \in s} w_{i,j} y_j$$

is a nonparametric estimator for $m(x)$ at $x = x_i$, and the resulting estimator for $F_N(t)$ is given by

$$\hat{F}^*(t) := \frac{1}{N}\left(\sum_{j \in s} I\left(y_j \le t\right) + \sum_{i \notin s} \sum_{j \in s} w_{i,j} I\left(y_j - \hat{m}_j \le t - \hat{m}_i\right) \right). \tag{2.5}$$

The fitted values in (2.3) and (2.4), or appropriately modified versions of them which include sample inclusion probabilities in the regression weights $w_{i,j}$, can obviously be computed also for $i \in s$, and they can be employed for example in generalized difference estimators (Särndal et al. 1992, page 221) or in model calibrated estimators (see for example Wu and Sitter 2001; Chen and Wu 2002; Wu 2003; Montanari and Ranalli 2005; Rueda, Martínez, Martínez and Arcos 2007; Rueda, Sànchez-Borrego, Arcos and Martínez 2010). In addition to the model-based estimators in (2.2) and (2.5), we shall thus consider also the generalized difference estimators given by

$$\tilde{F}(t) := \frac{1}{N}\left(\sum_{i \in U} \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j \le t\right) \right) + \sum_{i \in s} \pi_i^{-1} \left(I\left(y_i \le t\right) - \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j \le t\right) \right)$$

and by

$$\tilde{F}^*(t) := \frac{1}{N}\left(\sum_{i \in U} \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j - \tilde{m}_j \le t - \tilde{m}_i\right) \right) + \sum_{i \in s} \pi_i^{-1} \left(I\left(y_i \le t\right) - \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j - \tilde{m}_j \le t - \tilde{m}_i\right) \right)$$

where π_i denotes the first order sample inclusion probabilities, $\tilde{w}_{i,j}$ denotes design weighted regression weights whose definition is given below, and $\tilde{m}_i := \sum_{k \in s} \tilde{w}_{i,k} y_k$. Note that $\tilde{F}(t)$ and $\tilde{F}^*(t)$ are based on design weighted counterparts of the fitted values $\hat{G}_i(t)$ and $\hat{G}_i^*(t)$ which are given by

$$\tilde{G}_i(t) := \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j \le t\right)$$

and

$$\tilde{G}_i^*(t) := \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j - \tilde{m}_j \le t - \tilde{m}_i\right),$$

respectively.

As for the regression weights $w_{i,j}$ and $\tilde{w}_{i,j}$, in the present work we consider local linear regression weights in their place. In what follows $w_{i,j}$ and $\tilde{w}_{i,j}$ are thus defined by

$$w_{i,j} := \frac{1}{n\lambda} K\left(\frac{x_i - x_j}{\lambda}\right) \frac{M_{2,s}(x_i) - \left(\frac{x_i - x_j}{\lambda}\right) M_{1,s}(x_i)}{M_{2,s}(x_i) M_{0,s}(x_i) - M_{1,s}^2(x_i)}$$

and

$$\tilde{w}_{i,j} := \frac{1}{\pi_j n\lambda} K\left(\frac{x_i - x_j}{\lambda}\right) \frac{\tilde{M}_{2,s}(x_i) - \left(\frac{x_i - x_j}{\lambda}\right) \tilde{M}_{1,s}(x_i)}{\tilde{M}_{2,s}(x_i) \tilde{M}_{0,s}(x_i) - \tilde{M}_{1,s}^2(x_i)},$$

where n is the number of units in the sample s,

$$M_{r,s}(x) := \sum_{k \in s} \frac{1}{n\lambda} K\left(\frac{x - x_k}{\lambda}\right)\left(\frac{x - x_k}{\lambda}\right)^r, \qquad r = 0,1,2,$$

and

$$\tilde{M}_{r,s}(x) := \sum_{k \in s} \frac{1}{\pi_k n\lambda} K\left(\frac{x - x_k}{\lambda}\right)\left(\frac{x - x_k}{\lambda}\right)^r, \qquad r = 0,1,2.$$

It is worth noting that the nonparametric estimators of this section are not well-defined if the regression weights $w_{i,j}$ and $\tilde{w}_{i,j}$ included in their definitions are not well-defined. This problem occurs for example when the support of the kernel function $K(u)$ is given by the interval $[-1,1]$ (e.g., uniform kernel, Epanechnikov kernel), and when there are not at least two $j \in s$ such that $|x_i - x_j| < \lambda$. To overcome this problem one can use a kernel function whose support is given by the whole real line (e.g., Gaussian kernel) or choose the bandwidth adaptively. The latter solution may also lead to more efficient estimators (see e.g., Fan and Gijbels 1992). With reference to the estimators $\hat{F}^*(t)$ and $\tilde{F}^*(t)$ based on the modified fitted values, it is moreover worth noting that one could in principle apply different bandwidths and/or regression weights to the y_i – values and to the indicator functions. For the sake of simplicity, in the present work we shall consider neither adaptive bandwidth selection nor the possibility of different regression weights to estimate the mean regression function and the distributions of the error components.

Comparing the definitions of the estimators based on the two types of fitted values, it becomes immediately obvious that $\hat{F}(t)$ and $\tilde{F}(t)$ are easier to compute since they are linear combinations of the observed indicator functions $I(y_j \le t)$. The coefficients of these linear combinations do not depend on the study variable Y and they can therefore be used to estimate averages of other functions than indicator functions, or of functions of several study variables, in particular when there are reasons to believe that the latter are related to the auxiliary variable X. This fact is of particular value to practitioners who want estimates related to several study variables to be consistent with one another. However, there is a strong argument in favor of the estimators $\hat{F}^*(t)$ and $\tilde{F}^*(t)$ based on the modified fitted values too: if $y_i = a + bx_i$ for all $i \in U$, then it follows that $\hat{F}^*(t) = \tilde{F}^*(t) = F_N(t)$ for every sample s such that the estimators are well-defined. One would therefore expect that $\hat{F}^*(t)$ and $\tilde{F}^*(t)$ be more efficient than $\hat{F}(t)$ and $\tilde{F}(t)$ when there is a strong regression relationship between Y and X.

3 Model-based properties

In this section we provide asymptotic expansions for the model bias and the model variance of the estimators introduced in the previous section. The expansions are based on the following assumptions:

(C1) $N \to \infty$ and the sequence of population x_i – values and of sample designs are such that

$$H_{N,s}(x) := \frac{1}{n}\sum_{i \in s} I(x_i \le x)$$

and

$$H_{N,\bar{s}}(x) := \frac{1}{N-n}\sum_{i \notin s} I(x_i \le x)$$

converge to absolutely continuous distribution functions $H_s(x) := \int_a^x h_s(z)dz$ and $H_{\bar{s}}(x) := \int_a^x h_{\bar{s}}(z)dz$, respectively. The support of $H_s(x)$ and $H_{\bar{s}}(x)$ is given by a bounded interval $[a,b]$ and the density functions $h_s(x)$ and $h_{\bar{s}}(x)$ have bounded first derivatives for $x \in (a,b)$. $h_s(x)$ is bounded away from zero.

(C2) The kernel function $K(u)$ is symmetric, has support on $[-1,1]$ and has bounded derivative for $u \in (-1,1)$. The bandwidth sequence λ goes to zero slow enough to make sure that

$$\alpha := \max\left\{\sup_{x\in[a,b]}\left|H_{N,s}(x)-H_s(x)\right|, \sup_{x\in[a,b]}\left|H_{N,\bar{s}}(x)-H_{\bar{s}}(x)\right|\right\}$$

is of order $o(\lambda)$.

(C3) The population y_i – values are generated from model (2.1). The function $m(x)$ is such that

$$\left|m(x)-m(x_0)-m'(x_0)(x-x_0)-\frac{1}{2}m''(x_0)(x-x_0)^2\right| \le C\left|x-x_0\right|^{2+\delta}$$

for some $\delta > 0$, and the family of error component distribution functions $G(\varepsilon|x)$ is such that

$$\begin{aligned}&\Big|G(\varepsilon|x)-G(\varepsilon_0|x_0)-G^{(1,0)}(\varepsilon_0|x_0)(\varepsilon-\varepsilon_0)-G^{(0,1)}(\varepsilon_0|x_0)(x-x_0)\\&-\frac{1}{2}\left(G^{(2,0)}(\varepsilon_0|x_0)(\varepsilon-\varepsilon_0)^2+2G^{(1,1)}(\varepsilon_0|x_0)(\varepsilon-\varepsilon_0)(x-x_0)+G^{(0,2)}(\varepsilon_0|x_0)(x-x_0)^2\right)\Big|\\&\qquad\le C\left(\left|\varepsilon-\varepsilon_0\right|^{2+\delta}+\left|x-x_0\right|^{2+\delta}\right)\end{aligned}$$

for some $C > 0$ and some $\delta > 0$, where

$$G^{(r,s)}(\varepsilon|x) := \partial^{r+s}G(\varepsilon|x)\big/\left(\partial\varepsilon^r\partial x^s\right) \quad \text{for} \quad r,s = 0,1,2.$$

Assumption (C1) poses a restriction on how the sample and nonsample x_i – values are generated. Together with assumption (C2) it makes sure that the estimation errors of the kernel density estimators for $h_s(x)$ and $h_{\bar{s}}(x)$ go to zero uniformly for $x \in [a+\lambda, b-\lambda]$ and that they are uniformly bounded for $x \in [a,b]$. Replacing (C1) by more specific assumptions may allow for relaxing (C2) and for improving the uniform convergence rate for the estimation error of the kernel density estimators (see for example the results in Hansen 2008). Assumption (C3) is finally needed to make sure that the model mean square errors of the two estimators converge to zero. It can be relaxed at the cost of slowing down the convergence rates. In addition to assumptions (C1) to (C3) we shall also need the following assumption (C4) to make sure that the model mean square errors of the generalized difference estimators go to zero:

(C4) The first order sample inclusion probabilities are given by

$$\pi_i := n^* \frac{\pi(x_i)}{\sum_{j\in U}\pi(x_j)}, \qquad i \in U,$$

where n^* is the expected sample size and $\pi(x)$ is a function which is bounded away from zero and has bounded first derivative for $x \in (a,b)$.

Proposition 1. *Under assumptions (C1) to (C3) it follows that:*

$$E\left(\hat{F}(t) - F_N(t)\right) = \lambda^2 \frac{N-n}{N} \frac{\mu_2}{2\mu_0} \int_a^b \Big[G^{(2,0)}(t - m(x)|x)\left(m'(x)\right)^2 - G^{(1,0)}(t - m(x)|x) m''(x) - 2G^{(1,1)}(t - m(x)|x) m'(x) + G^{(0,2)}(t - m(x)|x) \Big] h_{\bar{s}}(x)\, dx + o\left(\lambda^2\right)$$

and

$$\operatorname{var}\left(\hat{F}(t) - F_N(t)\right) = \frac{1}{n}\left(\frac{N-n}{N}\right)^2 \int_a^b \Big[G(t - m(x)|x) - G^2(t - m(x)|x) \Big] \Big[h_{\bar{s}}(x)/h_s(x) \Big] h_{\bar{s}}(x)\, dx + \frac{1}{N-n}\left(\frac{N-n}{N}\right)^2 \int_a^b \Big[G(t - m(x)|x) - G^2(t - m(x)|x) \Big] h_{\bar{s}}(x)\, dx + o\left(n^{-1}\right),$$

where $\mu_r := \int_{-1}^{-1} K(u) u^r du$ *for* $r = 0,1,2.$

Adding assumption (C4) it can be shown that

$$E\left(\tilde{F}(t) - F_N(t)\right) = \lambda^2 \frac{N-n}{N} \frac{\mu_2}{2\mu_0} \int_a^b \Big[G^{(2,0)}(t - m(x)|x)\left(m'(x)\right)^2 - G^{(1,0)}(t - m(x)|x) m''(x) - 2G^{(1,1)}(t - m(x)|x) m'(x) + G^{(0,2)}(t - m(x)|x) \Big] h(x)\, dx + o\left(\lambda^2\right),$$

where

$$h(x) := h_{\bar{s}}(x) + \left(1 - \pi^{-1}(x)\right) h_s(x),$$

and it can be shown that

$$\operatorname{var}\left(\tilde{F}(t) - F_N(t)\right) = \operatorname{var}\left(\hat{F}(t) - F_N(t)\right) + o\left(n^{-1}\right).$$

Proposition 2. *Under assumptions (C1) to (C3) and assuming that*

i) the function

$$\sigma^2(x) := \int_{-\infty}^{\infty} \varepsilon^2 dG(\varepsilon|x)$$

has bounded first derivative for $x \in (a,b)$

ii)

$$\sup_{x \in [a,b]} \int_{-\infty}^{\infty} \varepsilon^4 dG(\varepsilon|x) < \infty,$$

it can be shown that

$$E\left(\hat{F}^{*}(t)-F_N(t)\right) = \lambda^2\frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}(t-m(x)|x)h_{\bar{s}}(x)dx$$
$$+\frac{1}{n\lambda}\frac{N-n}{N}\left[\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}(t-m(x)|x)(t-m(x))h_s^{-1}(x)h_{\bar{s}}(x)dx\right.$$
$$\left.+\frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}(t-m(x)|x)\sigma^2(x)h_s^{-1}(x)h_{\bar{s}}(x)dx\right]+o\left(\lambda^2+(n\lambda)^{-1}\right),$$

where $\kappa := \int_{-1}^{1}K^2(u)du$ *and* $\theta := \int_{-1}^{1}K(v)\int_{-1}^{1}K(u+v)K(u)dudv$, *and it can be shown that*

$$\mathrm{var}\left(\hat{F}^{*}(t)-F_N(t)\right)=\mathrm{var}\left(\hat{F}(t)-F_N(t)\right)+o\left(n^{-1}+\lambda^5\right).$$

Adding assumption (C4) it can also be shown that

$$E\left(\tilde{F}^{*}(t)-F_N(t)\right) = \lambda^2\frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}(t-m(x)|x)h(x)dx$$
$$+\frac{1}{n\lambda}\frac{N-n}{N}\left[\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}(t-m(x)|x)(t-m(x))h_s^{-1}(x)h(x)dx\right.$$
$$\left.+\frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}(t-m(x)|x)\sigma^2(x)h_s^{-1}(x)h(x)dx\right]$$
$$+o\left(\lambda^2+(n\lambda)^{-1}\right)$$

and that

$$\mathrm{var}\left(\tilde{F}^{*}(t)-F_N(t)\right)=\mathrm{var}\left(\hat{F}(t)-F_N(t)\right)+o\left(n^{-1}+\lambda^5\right).$$

The proofs of the Propositions are given in the Appendix. Dorfman and Hall (1993) derived similar expansions for the Kuo estimator with local constant regression weights instead of local linear ones.

Note that in view of the asymptotic expansions it is possible to choose bandwidth sequences λ in such a way as to make sure that the squares of the model biases are of smaller order of magnitude than the corresponding model variances. For the estimators based on the fitted values of Kuo this is achieved whenever $\lambda = o\left(n^{-1/4}\right)$, while for the estimators with the modified fitted values this requires that λ goes to zero faster than $O\left(n^{-1/4}\right)$ and slower than $O\left(n^{-1/2}\right)$. The convergence rates for the model biases of the latter estimators are optimized when $\lambda = O\left(n^{-1/3}\right)$ and in this case the resulting model biases are both of order $O\left(n^{-2/3}\right)$. The model biases for the estimators based on the fitted values of Kuo can be made to converge much faster, depending on the sequences $H_{N,s}(x)$ and $H_{N,\bar{s}}(x)$ and on the bandwidth sequence λ.

Given the above considerations concerning the model biases and given the fact that the leading terms in the model variances are the same for both types of fitted values, it would be of interest to know the second order terms in the model variances in order to establish which estimator is more efficient from the model-based perspective. The proofs in the Appendix suggest however that the second order terms depend on more specific assumptions than (C1) to (C3) and that, in particular for the estimators based on the modified fitted values, they are difficult to determine.

4 Design-based properties

In the previous section we have shown that the model-based estimators $\hat{F}(t)$ and $\hat{F}^*(t)$ are asymptotically model-unbiased and model mean square error consistent. However, they are not design-unbiased in general and therefore they should not be used when the sample inclusion probabilities are not constant. In these cases the generalized difference estimators $\tilde{F}(t)$ and $\tilde{F}^*(t)$ should be used. In fact, it follows from the results in Breidt and Opsomer (2000) that under fairly general conditions $\tilde{F}(t)$ is asymptotically design-unbiased and that its design mean square error is given by

$$E_d\left(\left|\tilde{F}(t)-F_N(t)\right|^2\right)=\frac{1}{N^2}\sum_{i,j\in U}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_i\pi_j}\left[I(y_i\le t)-\bar{G}_i(t)\right]\left[I\left(y_j\le t\right)-\bar{G}_j(t)\right]+o\left(n^{-1}\right),$$

where $E_d(\cdot)$ denotes expectation with respect to the sample design, $\pi_{i,j}$ denotes the joint sample inclusion probability for units i and j (it is understood that $\pi_{i,i}=\pi_i$), and where

$$\bar{G}_i(t):=\sum_{j\in U}\bar{w}_{i,j}I\left(y_j\le t\right).$$

The regression weights $\bar{w}_{i,j}$ in the definition of $\bar{G}_i(t)$ refer to the whole finite population U and are given by

$$\bar{w}_{i,j}:=\frac{1}{N\lambda}K\left(\frac{x_i-x_j}{\lambda}\right)\frac{\bar{M}_{2,s}(x_i)-\left(\frac{x_i-x_j}{\lambda}\right)\bar{M}_{1,s}(x_i)}{\bar{M}_{2,s}(x_i)\bar{M}_{0,s}(x_i)-\bar{M}_{1,s}^2(x_i)},$$

where

$$\bar{M}_{r,s}(x):=\sum_{k\in U}\frac{1}{N\lambda}K\left(\frac{x-x_k}{\lambda}\right)\left(\frac{x-x_k}{\lambda}\right)^r,\qquad r=0,1,2.$$

Moreover, according to Breidt and Opsomer (2000),

$$\tilde{V}\left(\tilde{F}(t)\right):=\frac{1}{N^2}\sum_{i,j\in s}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_{i,j}\pi_i\pi_j}\left[I(y_i\le t)-\tilde{G}_i(t)\right]\left[I\left(y_j\le t\right)-\tilde{G}_j(t)\right]$$

is a consistent estimator for the design mean square error of $\tilde{F}(t)$.

Unfortunately the results in Breidt and Opsomer (2000) cannot be applied to the generalized difference estimator $\tilde{F}^*(t)$ as well, since the latter estimator does not fall into the class of local polynomial regression estimators due to the presence of the regression function estimators $\tilde{m}_i$ and $\tilde{m}_j$ inside the indicator functions in the fitted values $\tilde{G}_i^*(t)$. However, the results for $\tilde{F}(t)$ suggest that in large samples $\tilde{G}_i^*(t)$ and

$$\bar{G}_i^*(t):=\sum_{j\in U}\bar{w}_{i,j}I\left(y_j-\bar{m}_j\le t-\bar{m}_i\right),$$

where $\bar{m}_i := \sum_{j \in U} \bar{w}_{i,j} y_j$, are approximately the same, and that

$$E_d\left(\left|\tilde{F}^*(t) - F_N(t)\right|^2\right) = \frac{1}{N^2} \sum_{i,j \in U} \frac{\pi_{i,j} - \pi_i \pi_j}{\pi_i \pi_j} \left[I(y_i \le t) - \bar{G}_i^*(t)\right]\left[I(y_j \le t) - \bar{G}_j^*(t)\right] + o\left(n^{-1}\right)$$

Based on this conjecture, we tested

$$\tilde{V}\left(\tilde{F}^*(t)\right) := \frac{1}{N^2} \sum_{i,j \in s} \frac{\pi_{i,j} - \pi_i \pi_j}{\pi_{i,j} \pi_i \pi_j} \left[I(y_i \le t) - \tilde{G}_i^*(t)\right]\left[I(y_j \le t) - \tilde{G}_j^*(t)\right].$$

as estimator for the design mean square error of the generalized difference estimator $\tilde{F}^*(t)$ in the simulation study of the following section.

5 Simulation study

In this section we analyze some simulation results. Our goal is to compare efficiency with respect to the sample design of the distribution function estimators introduced in Section 2 and of the variance estimators of Section 4. The simulation results refer to simple random without replacement sampling and to Poisson sampling with unequal inclusion probabilities. As a benchmark, we included also the Horvitz-Thompson distribution function estimator

$$\hat{F}_\pi(t) := \frac{1}{N} \sum_{j \in s} \pi_j^{-1} I\left(y_j \le t\right)$$

and the corresponding variance estimator

$$\tilde{V}\left(\hat{F}_\pi(t)\right) := \frac{1}{N^2} \sum_{i,j \in s} \frac{\pi_{i,j} - \pi_i \pi_j}{\pi_{i,j} \pi_i \pi_j} I(y_i \le t) I\left(y_j \le t\right)$$

in the simulation study.

We considered both artificial and real populations. The former were obtained by generating $N = 1{,}000$ values x_i from i.i.d. uniform random variables with support on the interval $(0,1)$ and by combining them with three types of regression function $m(x)$ and two types of error components ε_i. The regression functions are (i) $m(x) = 0$ (flat), (ii) $m(x) = 10x$ (linear) and (iii) $m(x) = 10x^{1/4}$ (concave), while the error components ε_i are either independent realizations from a unique Student t distribution with $\nu = 5$ d.o.f., or independent realizations from N different shifted noncentral Student t distributions with $\nu = 5$ d.o.f. and with noncentrality parameters given by $\mu = 15x_i$. The shifts applied to the error components in the latter case make sure that the means of the noncentral Student t distributions from which they were generated are zero. The artificial populations are shown in Figure 5.1 to 5.3. As for the real populations, we took the *MU284 Population of Sweden Municipalities* of Särndal et al. (1992) (population size $N = 284$) and considered the natural logarithm of *RMT85 = Revenues from the 1985 municipal taxation (in millions of kronor)* as study variable Y, and the natural logarithm of either *P85 = 1985 population (in thousands)*

or *REV84 = Real estate values according to 1984 assessment (in millions of kronor)* as auxiliary variable X. The real populations are shown in Figure 5.4.

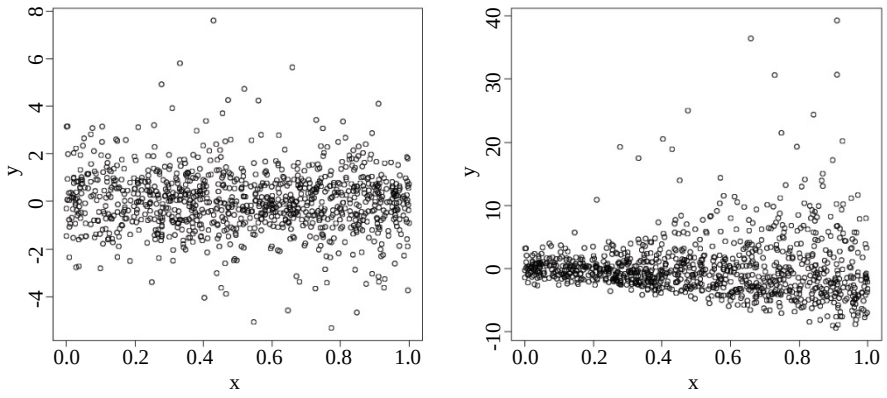

Figure 5.1 Populations generated from $y_i = \varepsilon_i$, where $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$ (left panel) and $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$ (right panel).

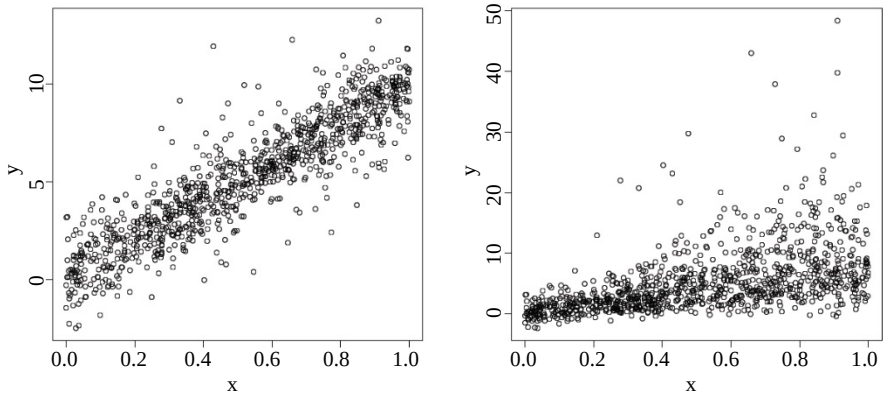

Figure 5.2 Populations generated from $y_i = 10x_i + \varepsilon_i$, where $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$ (left panel) and $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$ (right panel).

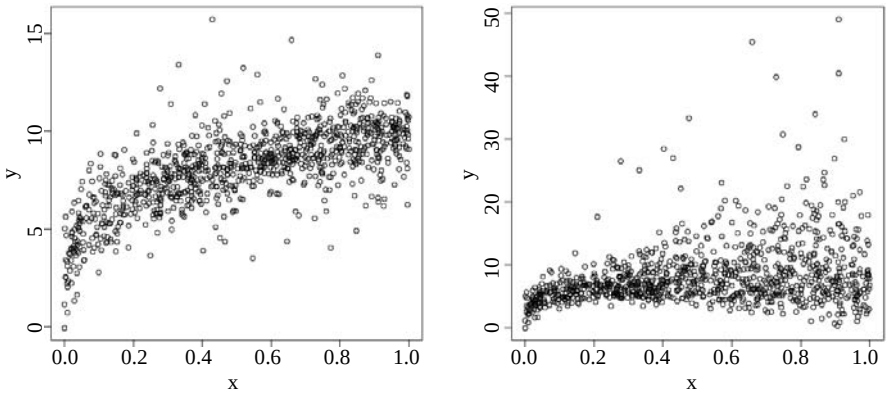

Figure 5.3 Populations generated from $y_i = 10x_i^{1/4} + \varepsilon_i$, where $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$ (left panel) and $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$ (right panel).

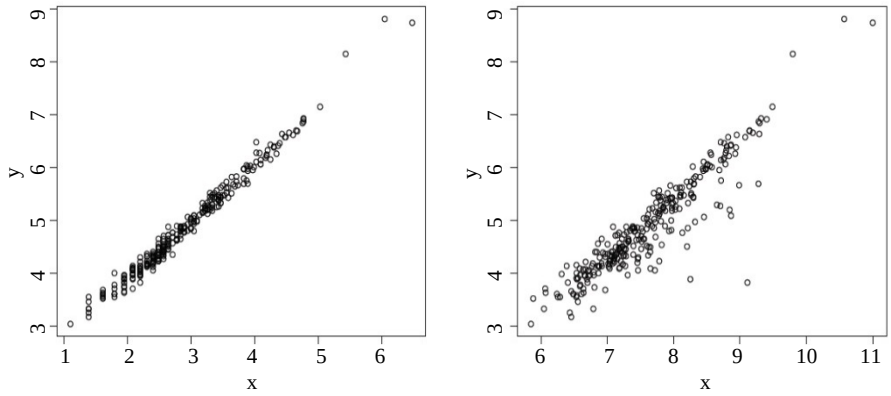

Figure 5.4 ***MU284 Population of Sweden Municipalities*** **of Särndal et al (1992).** $y_i = \ln RMT85_i$ **for the** i^{th} **municipality, and** $x_i = \ln P85_i$ **(left panel) or** $x_i = \ln REV84_i$ **(right panel).**

From each population we selected independently $B = 1{,}000$ samples. When sampling from the artificial populations we set the sample size equal to $n = 100$ in case of simple random without replacement sampling and, in case of Poisson sampling, we set the expected sample size equal to $n^* = 100$ and made the sample inclusion probabilities proportional to the standard deviations of the shifted noncentral Student t distributions of above. When sampling from the real populations, we set the sample size equal to $n = 30$ in case of simple random without replacement sampling. In case of Poisson sampling, we set the expected sample size equal to $n^* = 30$ and made the sample inclusion probabilities proportional to the absolute values of the residuals from the linear least squares regressions of the population y_i values on the population x_i values.

As for the definition of the nonparametric estimators, we used the Epanechnikov kernel function $K(u) := 0.75\left(1-u^2\right)$ with $\lambda = 0.15$ or $\lambda = 0.3$ for the samples taken from the artificial populations, and the Gaussian kernel function $K(u) := 1/\sqrt{2\pi}\, e^{-(1/2)u^2}$ with $\lambda = 1$ or $\lambda = 2$ for the samples taken from the real populations. In the tables with the simulation results the nonparametric estimators corresponding to the small and large bandwidth values are identified with an s (small) or an l (large) in the subscript. We resorted to the Gaussian kernel function for the samples taken from the real populations to avoid singularity problems that occur in case of holes in the sampled set of $x_i -$values. Such holes are much more likely to occur with the real populations than with the artificial ones, because the distributions of the auxiliary variables are asymmetric in the former. In fact, in the artificial populations the nonparametric estimators were well-defined for all the $B = 1{,}000$ samples selected according to the simple random without replacement sampling design. For the Poisson sampling design, on the other hand, 47 among the $B = 1{,}000$ simulated samples were such that the nonparametric estimators with the small bandwidth value could not be computed and just one of these samples was such that the nonparametric estimators with the large bandwidth value were undefined. The simulation results referring to the nonparametric estimators in Tables 5.2 and 5.5 account only for the samples where they were well-defined and thus they are based on a little less than $B = 1{,}000$ realizations.

Tables 5.1 to 5.4 report the simulated bias (BIAS) and the simulated root mean square error (RMSE) for each distribution function estimator at different levels of t at which $F_N(t)$ has been estimated: based, for example, on the values $\tilde{F}_b(t),\ b = 1, 2, \ldots, B,$ taken on by the estimator $\tilde{F}(t)$,

$$\text{BIAS} := \frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_b(t) - F_N(t)\right) \times 10{,}000$$

and

$$\text{RMSE} := \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_b(t) - F_N(t)\right)^2} \times 10{,}000.$$

The RMSE's show that the estimators based on the modified fitted values are usually more efficient. In sampling from the real populations the gain in RMSE is sometimes quite large. As expected, the model-based estimators tend to be more efficient than the generalized difference estimators in case of simple random without replacement sampling when both types of estimator are approximately unbiased. Under the Poisson sampling scheme the BIAS of the model-based estimators increases, but nonetheless they remain competitive. More variability in the sample inclusion probabilities would certainly change this outcome, because it would increase the BIAS of the model-based estimators. The simulation results should therefore not be seen to be in contrast with Johnson, Breidt and Opsomer (2008) who argue in favor of generalized difference estimators (called model-assisted estimators in their paper) as "a good overall choice for distribution function estimators".

Table 5.1
Artificial populations (population size N = 1,000). BIAS and RMSE of distribution function estimators under simple random without replacement sampling. Sample size n = 100

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	BIAS	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. central Student t with $\nu = 5$										
$\hat{F}_s(t)$	6	216	-3	433	31	512	23	434	12	207
$\hat{F}_l(t)$	15	219	10	430	0	502	-10	429	3	213
$\hat{F}_s^*(t)$	6	209	-30	411	22	484	22	414	3	200
$\hat{F}_l^*(t)$	15	214	-9	409	10	477	1	407	-10	207
$\tilde{F}_s(t)$	6	213	8	425	24	504	-4	430	8	207
$\tilde{F}_l(t)$	6	210	10	417	22	494	-8	422	6	206
$\tilde{F}_s^*(t)$	8	213	9	426	25	503	-5	432	5	206
$\tilde{F}_l^*(t)$	7	210	10	417	23	494	-6	424	4	206
$\tilde{F}_\pi(t)$	7	208	11	411	19	489	-5	417	6	200
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\hat{F}_s(t)$	26	225	33	376	8	477	26	419	33	209
$\hat{F}_l(t)$	52	236	23	374	-5	475	38	421	29	213
$\hat{F}_s^*(t)$	20	195	-29	351	-89	471	11	407	30	202
$\hat{F}_l^*(t)$	36	201	-11	357	-94	473	28	410	21	204
$\tilde{F}_s(t)$	8	211	11	370	-7	473	4	415	16	211
$\tilde{F}_l(t)$	5	208	8	367	-5	468	5	411	16	212
$\tilde{F}_s^*(t)$	11	210	11	372	-11	475	4	416	15	210
$\tilde{F}_l^*(t)$	7	208	11	368	-7	468	8	412	15	211
$\tilde{F}_\pi(t)$	1	211	1	391	-6	477	8	399	18	210

Table 5.1 (continued)
Artificial populations (population size N = 1,000). BIAS and RMSE of distribution function estimators under simple random without replacement sampling. Sample size n = 100

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	BIAS	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\hat{F}_s(t)$	32	201	25	275	13	250	-14	264	-36	217
$\hat{F}_l(t)$	114	250	152	304	12	236	-180	312	-86	242
$\hat{F}_s^*(t)$	-50	165	12	226	51	216	26	230	13	172
$\hat{F}_l^*(t)$	-46	155	-14	199	69	195	23	211	17	156
$\tilde{F}_s(t)$	-5	186	4	275	15	248	11	269	-2	201
$\tilde{F}_l(t)$	-5	184	7	274	17	250	5	269	-2	196
$\tilde{F}_s^*(t)$	-10	180	5	275	16	245	14	266	-1	200
$\tilde{F}_l^*(t)$	-9	176	3	272	15	242	13	262	-1	194
$\tilde{F}_\pi(t)$	-7	203	14	413	37	472	17	405	1	206
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\hat{F}_s(t)$	24	204	23	351	27	403	26	382	29	208
$\hat{F}_l(t)$	94	242	135	372	51	392	13	380	15	212
$\hat{F}_s^*(t)$	55	182	-9	301	-18	368	-23	359	37	202
$\hat{F}_l^*(t)$	124	210	-31	278	-63	363	-8	356	48	200
$\tilde{F}_s(t)$	-2	194	-4	349	11	401	18	377	13	208
$\tilde{F}_l(t)$	-2	190	-5	345	12	398	17	374	11	209
$\tilde{F}_s^*(t)$	0	191	-5	352	14	401	20	376	13	207
$\tilde{F}_l^*(t)$	-1	189	-6	344	13	397	18	375	12	209
$\tilde{F}_\pi(t)$	-4	205	-5	401	21	470	24	401	14	207
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\hat{F}_s(t)$	81	207	44	316	17	384	-2	376	23	203
$\hat{F}_l(t)$	138	258	183	356	35	367	-50	374	8	208
$\hat{F}_s^*(t)$	7	146	-14	274	16	352	-8	358	15	197
$\hat{F}_l^*(t)$	9	144	10	246	-2	323	-18	339	24	186
$\tilde{F}_s(t)$	3	175	3	319	10	383	17	374	10	203
$\tilde{F}_l(t)$	0	178	5	316	11	380	17	370	8	202
$\tilde{F}_s^*(t)$	1	167	5	320	12	383	17	374	9	203
$\tilde{F}_l^*(t)$	-1	164	6	316	13	379	20	368	8	201
$\tilde{F}_\pi(t)$	4	209	11	412	25	477	27	422	10	200
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\hat{F}_s(t)$	59	234	95	402	66	455	51	395	26	208
$\hat{F}_l(t)$	94	259	190	441	147	467	98	400	16	212
$\hat{F}_s^*(t)$	30	184	33	343	-123	435	-34	385	40	203
$\hat{F}_l^*(t)$	57	201	58	331	-148	437	2	382	34	203
$\tilde{F}_s(t)$	1	205	7	386	12	449	17	392	13	208
$\tilde{F}_l(t)$	-1	204	0	385	9	445	20	389	11	209
$\tilde{F}_s^*(t)$	3	201	8	389	7	449	13	392	14	207
$\tilde{F}_l^*(t)$	0	198	6	383	9	446	19	390	13	208
$\tilde{F}_\pi(t)$	0	205	-2	399	9	463	25	398	14	208

Table 5.2
Artificial populations (population size $N = 1{,}000$). BIAS and RMSE of distribution function estimators under Poisson sampling with sample inclusion probabilities π_i proportional to the standard deviations of the noncentral Student t distributions with $\nu = 5$ d.o.f. and with noncentrality parameters $\mu = 15x_i$. Expected sample size $n^* = 100$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	BIAS	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. central Student t with $\nu = 5$										
$\hat{F}_s(t)$	-10	252	-11	593	-22	738	-20	743	6	357
$\hat{F}_l(t)$	-1	237	9	543	-15	621	-5	590	11	302
$\hat{F}_s^*(t)$	22	244	-29	485	-3	555	9	515	-17	297
$\hat{F}_l^*(t)$	14	238	-10	492	-5	564	14	524	-1	283
$\tilde{F}_s(t)$	-6	247	0	579	-27	724	-40	736	3	349
$\tilde{F}_l(t)$	-2	231	11	526	-1	598	-10	566	7	285
$\tilde{F}_s^*(t)$	23	248	23	505	-4	562	-27	531	-20	304
$\tilde{F}_l^*(t)$	12	240	20	504	1	573	-13	538	-6	287
$\tilde{F}_\pi(t)$	-6	220	-7	543	-37	741	-44	929	-48	1,058
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\hat{F}_s(t)$	17	164	30	411	4	749	14	590	15	190
$\hat{F}_l(t)$	47	173	19	383	-1	602	57	498	15	187
$\hat{F}_s^*(t)$	21	175	-7	378	-89	554	-11	473	3	192
$\hat{F}_l^*(t)$	29	152	-3	367	-99	555	27	481	3	184
$\tilde{F}_s(t)$	1	159	10	406	-11	737	-5	579	-2	194
$\tilde{F}_l(t)$	1	158	9	388	-5	586	14	482	-1	192
$\tilde{F}_s^*(t)$	14	186	27	409	-3	562	-17	487	-10	200
$\tilde{F}_l^*(t)$	3	160	22	399	-11	566	-5	482	-2	193
$\tilde{F}_\pi(t)$	-3	162	-7	451	-31	738	-29	980	-55	1,067
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\hat{F}_s(t)$	8	461	21	561	-12	259	-18	218	-30	164
$\hat{F}_l(t)$	78	429	183	451	2	248	-161	261	-79	189
$\hat{F}_s^*(t)$	-69	306	12	340	10	267	15	199	6	143
$\hat{F}_l^*(t)$	-59	294	4	302	56	205	15	172	17	124
$\tilde{F}_s(t)$	-25	441	4	560	-10	257	9	219	5	153
$\tilde{F}_l(t)$	-14	372	35	410	-10	262	4	219	5	151
$\tilde{F}_s^*(t)$	-31	333	-2	386	-29	294	4	227	-1	161
$\tilde{F}_l^*(t)$	-20	339	15	372	-10	259	11	215	4	151
$\tilde{F}_\pi(t)$	-15	385	3	746	-37	917	-35	1,004	-48	1,070
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\hat{F}_s(t)$	-4	516	30	671	7	453	11	344	6	182
$\hat{F}_l(t)$	63	409	129	539	61	421	9	341	1	180
$\hat{F}_s^*(t)$	44	300	-29	433	-45	422	-47	345	12	180
$\hat{F}_l^*(t)$	107	314	-41	420	-60	397	-22	323	31	171
$\tilde{F}_s(t)$	-27	502	8	667	-8	450	0	344	-8	185
$\tilde{F}_l(t)$	-10	364	16	510	11	425	-2	345	-7	182
$\tilde{F}_s^*(t)$	-6	325	-9	479	-25	447	-14	356	-10	187
$\tilde{F}_l^*(t)$	-7	332	-9	489	-5	426	-3	344	-6	182
$\tilde{F}_\pi(t)$	-16	349	-2	705	-21	886	-42	1,013	-61	1,069

Table 5.2 (continued)
Artificial populations (population size $N = 1{,}000$). BIAS and RMSE of distribution function estimators under Poisson sampling with sample inclusion probabilities π_i proportional to the standard deviations of the noncentral Student t distributions with $\nu = 5$ d.o.f. and with noncentrality parameters $\mu = 15x_i$. Expected sample size $n^* = 100$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	BIAS	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\hat{F}_s(t)$	36	497	47	629	9	418	-11	320	15	191
$\hat{F}_l(t)$	56	393	186	490	43	383	-48	308	13	184
$\hat{F}_s^*(t)$	-29	276	-19	383	-18	380	-43	335	-1	204
$\hat{F}_l^*(t)$	-29	274	10	355	7	336	-29	290	23	179
$\tilde{F}_s(t)$	-30	475	12	630	4	421	7	317	6	191
$\tilde{F}_l(t)$	-42	336	31	452	11	390	8	312	8	186
$\tilde{F}_s^*(t)$	-31	306	5	429	-18	406	-14	344	-8	210
$\tilde{F}_l^*(t)$	-28	308	14	424	7	387	5	315	7	191
$\tilde{F}_\pi(t)$	-15	380	10	739	-23	891	-37	993	-47	1,064
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\hat{F}_s(t)$	24	308	69	687	53	690	38	406	2	188
$\hat{F}_l(t)$	47	301	131	553	139	561	91	393	-2	186
$\hat{F}_s^*(t)$	15	237	2	435	-135	513	-59	411	12	186
$\hat{F}_l^*(t)$	27	235	18	435	-149	506	-5	374	13	179
$\tilde{F}_s(t)$	-28	274	-8	673	4	688	3	403	-10	191
$\tilde{F}_l(t)$	-29	251	-12	512	17	541	7	395	-9	188
$\tilde{F}_s^*(t)$	-3	255	-12	481	-7	536	-20	422	-12	196
$\tilde{F}_l^*(t)$	-12	251	-16	489	2	538	-4	399	-9	189
$\tilde{F}_\pi(t)$	-10	267	-8	608	-4	860	-38	1,009	-63	1,066

Table 5.3
Real populations (population size $N = 284$). BIAS and RMSE of distribution function estimators under simple random without replacement sampling. Sample size $n = 30$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	BIAS	**RMSE**	**BIAS**	**RMSE**	**RBIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
MU284 population with $Y = \ln RMT85$ and $X = \ln P85$										
$\hat{F}_s(t)$	133	421	339	625	180	529	-265	490	-187	439
$\hat{F}_l(t)$	52	380	67	588	45	555	-63	469	-87	370
$\hat{F}_s^*(t)$	8	81	-154	203	90	130	62	123	6	54
$\hat{F}_l^*(t)$	28	66	-170	212	69	112	57	109	2	50
$\tilde{F}_s(t)$	-28	300	-24	497	8	483	-48	421	-38	319
$\tilde{F}_l(t)$	-28	326	-96	569	-52	544	3	466	1	319
$\tilde{F}_s^*(t)$	26	177	-11	302	0	244	1	308	-18	102
$\tilde{F}_l^*(t)$	29	179	-10	302	-2	243	-1	308	-21	104
$\tilde{F}_\pi(t)$	22	388	-10	771	9	864	5	731	-43	394
MU284 population with $Y = \ln RMT85$ and $X = \ln REV84$										
$\hat{F}_s(t)$	143	449	303	643	138	554	-217	543	-166	446
$\hat{F}_l(t)$	62	395	62	611	36	582	-49	519	-71	376
$\hat{F}_s^*(t)$	-11	204	-32	300	-101	328	42	285	31	155
$\hat{F}_l^*(t)$	36	183	-40	288	-149	345	6	261	34	122
$\tilde{F}_s(t)$	5	340	-22	548	4	557	-30	498	-23	332
$\tilde{F}_l(t)$	-2	349	-78	599	-36	588	10	522	8	331
$\tilde{F}_s^*(t)$	24	303	7	446	-6	494	2	439	-13	209
$\tilde{F}_l^*(t)$	29	304	4	443	-6	495	-1	432	-18	192
$\tilde{F}_\pi(t)$	34	395	1	766	16	880	9	744	-37	398

Table 5.4
Real populations (population size $N = 284$). BIAS and RMSE of distribution function estimators under Poisson sampling with inclusion probabilities proportional to the absolute value of the residuals of the linear regression of the population y_i – values on the population x_i – values. Expected size $n^* = 30$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	BIAS	**RMSE**	**BIAS**	**RMSE**	**RBIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
MU284 population with $Y = \ln RMT85$ and $X = \ln P85$										
$\hat{F}_s(t)$	204	420	485	668	239	519	-412	626	-90	317
$\hat{F}_l(t)$	180	424	417	684	319	614	-239	548	-148	348
$\hat{F}_s^*(t)$	-41	97	-118	199	132	178	40	140	-71	104
$\hat{F}_l^*(t)$	11	70	-147	211	63	128	-25	122	-85	106
$\tilde{F}_s(t)$	24	360	30	649	0	675	-68	614	58	368
$\tilde{F}_l(t)$	9	390	-63	737	-64	774	-7	682	75	414
$\tilde{F}_s^*(t)$	16	184	-14	307	36	283	16	323	-11	103
$\tilde{F}_l^*(t)$	25	187	-15	312	30	286	14	328	-11	112
$\tilde{F}_\pi(t)$	40	445	73	1,983	12	2,498	-43	3,094	-49	3,341
MU284 population with $Y = \ln RMT85$ and $X = \ln REV84$										
$\hat{F}_s(t)$	349	660	1,185	1,373	890	1,059	458	654	-32	270
$\hat{F}_l(t)$	287	601	1,003	1,236	771	989	484	695	42	263
$\hat{F}_s^*(t)$	317	453	739	866	761	879	624	701	159	207
$\hat{F}_l^*(t)$	364	471	720	842	718	824	572	647	96	158
$\tilde{F}_s(t)$	35	488	82	818	-31	772	7	634	-8	326
$\tilde{F}_l(t)$	22	500	3	878	-98	852	40	704	27	354
$\tilde{F}_s^*(t)$	37	317	32	498	-13	513	32	412	7	157
$\tilde{F}_l^*(t)$	51	313	30	498	-30	518	12	411	-10	149
$\tilde{F}_\pi(t)$	32	671	19	1,658	-172	2,354	-173	2,787	-191	2,935

Consider finally the simulation results referring to the variance estimators of Section 4. Tables 5.5 to 5.8 report the relative bias (RBIAS) and the relative root mean square error (RRMSE) for each of them. For example, based on the variance estimates $\tilde{V}_b\left(\tilde{F}(t)\right),\ b = 1,2,\ldots,B,$ obtained from the estimator $\tilde{V}\left(\tilde{F}(t)\right)$,

$$\text{RBIAS} := \frac{1}{B}\sum_{b=1}^{B}\frac{\tilde{V}_b\left(\tilde{F}(t)\right) - V_B\left(\tilde{F}(t)\right)}{V_B\left(\tilde{F}(t)\right)} \times 10{,}000$$

and

$$\text{RRMSE} := \frac{\sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\tilde{V}_b\left(\tilde{F}(t)\right) - V_B\left(\tilde{F}(t)\right)\right)^2}}{V_B\left(\tilde{F}(t)\right)} \times 10{,}000$$

where

$$V_B\left(\tilde{F}(t)\right) := \frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_b(t) - F_N(t)\right)^2.$$

As a benchmark, we report also the RBIAS and RRMSE of the estimator

$$\tilde{V}\left(\tilde{F}_{\pi}(t)\right) := \frac{1}{N^2}\sum_{i,j\in s}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_{i,j}\pi_i\pi_j} I\left(y_i \le t\right) I\left(y_j \le t\right).$$

for the variance of the Horvitz-Thompson estimator.

Table 5.5
Artificial populations (population size $N = 1{,}000$). RBIAS and RRMSE of variance estimators under simple random without replacement sampling. Sample size $n = 100$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	RBIAS	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. central Student t with $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-1,092	32,442	-1,249	3,895	-1,714	3,077	-1,536	3,828	-824	34,601
$\tilde{V}(\tilde{F}_l(t))$	-576	31,726	-603	3,838	-1,122	3,374	-951	3,758	-441	33,055
$\tilde{V}(\tilde{F}_s^*(t))$	-1,091	32,579	-1,292	3,914	-1,708	3,085	-1,640	3,828	-802	34,809
$\tilde{V}(\tilde{F}_l^*(t))$	-556	31,881	-622	3,857	-1,148	3,361	-1,025	3,749	-425	33,184
$\tilde{V}(\tilde{F}_{\pi}(t))$	42	30,952	57	3,928	-592	3,776	-287	3,825	551	33,462
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-1,900	29,622	50	4,707	-917	3,557	-998	3,695	-1,480	29,417
$\tilde{V}(\tilde{F}_l(t))$	-1,359	29,623	535	4,572	-395	3,881	-527	3,736	-1,277	28,267
$\tilde{V}(\tilde{F}_s^*(t))$	-1,832	30,119	-101	4,710	-991	3,530	-1,077	3,704	-1,398	29,927
$\tilde{V}(\tilde{F}_l^*(t))$	-1,362	29,713	465	4,559	-420	3,865	-591	3,718	-1,236	28,489
$\tilde{V}(\tilde{F}_{\pi}(t))$	-351	29,132	1,096	4,215	-78	4,074	574	4,067	-638	29,507
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-2,170	11,624	-1,027	2,480	-816	3,274	-1,424	2,583	-1,946	8,681
$\tilde{V}(\tilde{F}_l(t))$	-1,534	11,605	-529	2,632	-148	2,975	-859	2,590	-1,151	9,015
$\tilde{V}(\tilde{F}_s^*(t))$	-1,765	12,107	-1,108	2,529	-714	3,366	-1,318	2,660	-1,905	8,658
$\tilde{V}(\tilde{F}_l^*(t))$	-1,062	11,948	-671	2,735	-212	3,291	-762	2,785	-1,048	8,590
$\tilde{V}(\tilde{F}_{\pi}(t))$	254	31,545	-52	3,726	136	4,152	267	3,992	35	30,264
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-1,642	25,809	-855	3,541	-1,076	3,038	-1,081	3,030	-1,361	21,157
$\tilde{V}(\tilde{F}_l(t))$	-950	25,692	-323	3,509	-597	3,312	-617	3,164	-1,124	20,231
$\tilde{V}(\tilde{F}_s^*(t))$	-1,385	26,406	-997	3,505	-1,089	3,045	-1,096	3,033	-1,310	21,393
$\tilde{V}(\tilde{F}_l^*(t))$	-832	26,212	-292	3,556	-614	3,317	-716	3,154	-1,135	20,286
$\tilde{V}(\tilde{F}_{\pi}(t))$	105	29,621	507	3,857	209	4,244	425	3,910	-337	29,082
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-2,465	30,612	-1,121	4,594	-1,512	3,183	-1,958	3,076	-863	19,720
$\tilde{V}(\tilde{F}_l(t))$	-1,780	28,103	-663	4,420	-1,092	3,319	-1,491	3,140	-439	18,985
$\tilde{V}(\tilde{F}_s^*(t))$	-2,052	33,980	-1,150	4,619	-1,537	3,217	-1,948	3,127	-954	19,637
$\tilde{V}(\tilde{F}_l^*(t))$	-1,194	33,573	-691	4,472	-1,124	3,368	-1,438	3,228	-357	19,245
$\tilde{V}(\tilde{F}_{\pi}(t))$	-81	30,001	9	3,756	-110	3,996	-598	3,661	440	32,455
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-1,873	29,437	-758	3,759	-621	3,476	-709	3,599	-1,298	27,679
$\tilde{V}(\tilde{F}_l(t))$	-1,267	28,511	-284	3,661	-131	3,758	-321	3,552	-1,075	26,790
$\tilde{V}(\tilde{F}_s^*(t))$	-1,710	30,670	-928	3,741	-628	3,510	-777	3,603	-1,245	27,972
$\tilde{V}(\tilde{F}_l^*(t))$	-939	30,486	-270	3,764	-171	3,803	-375	3,581	-1,014	26,926
$\tilde{V}(\tilde{F}_{\pi}(t))$	178	29,640	599	3,816	533	4,324	590	3,874	-404	28,917

Table 5.6
Artificial populations (population size $N = 1{,}000$). RBIAS and RRMSE of variance estimators under Poisson sampling with sample inclusion probabilities π_i proportional to standard deviation of noncentral Student t distribution with $\nu = 5$ d.f. and with noncentrality parameter $\mu = 15x_i$. Expected sample size $n^* = 100$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	RBIAS	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. central Student t with $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-3,306	65,777	-4,248	8,032	-5,093	4,242	-6,258	4,844	-5,652	32,037
$\tilde{V}(\tilde{F}_l(t))$	-2,048	47,035	-2,656	4,705	-2,434	3,116	-3,310	3,939	-3,092	29,380
$\tilde{V}(\tilde{F}_s^*(t))$	-3,362	36,855	-2,488	4,409	-1,910	3,147	-2,869	3,910	-4,329	23,247
$\tilde{V}(\tilde{F}_l^*(t))$	-2,696	39,509	-2,076	4,450	-1,768	3,163	-2,648	3,811	-3,244	26,343
$\tilde{V}(\tilde{F}_\pi(t))$	113	129,637	259	15,120	618	6,327	193	5,429	273	6,097
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-740	125,975	-2,522	14,864	-5,466	3,658	-4,896	6,691	-1,551	83,262
$\tilde{V}(\tilde{F}_l(t))$	-391	83,047	-1,503	8,946	-2,428	4,099	-2,228	5,526	-1,154	54,680
$\tilde{V}(\tilde{F}_s^*(t))$	-3,260	58,072	-2,649	7,661	-2,260	3,936	-2,795	5,011	-2,116	48,739
$\tilde{V}(\tilde{F}_l^*(t))$	-716	77,935	-2,000	7,979	-1,934	4,235	-2,279	5,243	-1,243	52,531
$\tilde{V}(\tilde{F}_\pi(t))$	666	251,134	-564	26,553	-87	7,344	-2	6,029	407	6,610
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-6,801	7,898	-6,470	4,281	-1,059	22,596	-398	32,401	-1,650	72,632
$\tilde{V}(\tilde{F}_l(t))$	-4,978	5,826	-2,898	4,473	-603	9,530	206	15,226	-1,157	40,466
$\tilde{V}(\tilde{F}_s^*(t))$	-4,520	6,691	-2,710	4,213	-3,245	6,723	-1,156	12,681	-2,458	32,907
$\tilde{V}(\tilde{F}_l^*(t))$	-4,226	6,206	-1,674	5,062	-978	7,874	55	12,781	-1,283	33,737
$\tilde{V}(\tilde{F}_\pi(t))$	-707	47,550	118	7,214	609	4,409	743	4,628	435	4,800
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-7,398	8,847	-6,235	3,667	-2,493	8,171	-1,051	16,299	-1,440	71,943
$\tilde{V}(\tilde{F}_l(t))$	-4,548	9,463	-3,136	3,282	-1,187	4,246	-832	7,638	-982	45,182
$\tilde{V}(\tilde{F}_s^*(t))$	-3,902	11,727	-2,808	3,409	-2,411	3,501	-1,721	6,737	-1,671	41,389
$\tilde{V}(\tilde{F}_l^*(t))$	-3,598	10,771	-2,610	3,462	-1,284	3,988	-852	7,008	-972	43,017
$\tilde{V}(\tilde{F}_\pi(t))$	146	57,044	-42	8,708	520	4,784	214	4,686	390	5,085
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-7,731	8,568	-6,597	3,484	-2,442	7,775	-903	16,067	-1,967	56,480
$\tilde{V}(\tilde{F}_l(t))$	-4,611	9,378	-2,990	3,252	-874	4,119	-347	7,420	-1,310	35,051
$\tilde{V}(\tilde{F}_s^*(t))$	-4,747	11,909	-2,679	3,298	-1,896	3,272	-2,248	5,747	-3,382	27,222
$\tilde{V}(\tilde{F}_l^*(t))$	-4,223	10,380	-2,100	3,494	-788	3,731	-550	5,975	-1,795	29,856
$\tilde{V}(\tilde{F}_\pi(t))$	-428	47,038	-206	7,350	641	4,504	738	4,708	487	4,943
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-4,936	40,696	-6,111	4,579	-5,549	4,035	-1,864	14,381	-1,509	84,892
$\tilde{V}(\tilde{F}_l(t))$	-3,004	29,404	-2,764	3,962	-2,436	3,606	-1,234	7,357	-1,103	53,875
$\tilde{V}(\tilde{F}_s^*(t))$	-4,328	27,704	-2,516	4,235	-2,671	3,332	-2,586	5,955	-1,939	47,601
$\tilde{V}(\tilde{F}_l^*(t))$	-3,454	28,267	-2,263	4,160	-2,329	3,574	-1,433	6,682	-1,171	50,985
$\tilde{V}(\tilde{F}_\pi(t))$	152	98,607	663	12,879	15	5,376	20	5,080	429	5,619

Table 5.7
Real populations (population size $N = 284$). RBIAS and RRMSE of variance estimators under simple random without replacement sampling. Sample size $n = 30$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	RBIAS	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**
MU284 population with $Y = \ln RMT85$ and $X = \ln P85$										
$\tilde{V}(\tilde{F}_s(t))$	-2,853	16,809	-1,700	3,037	-1,554	2,984	-1,100	4,633	-5,503	16,257
$\tilde{V}(\tilde{F}_l(t))$	-1,110	16,374	-1,827	2,760	-1,683	2,847	-927	4,387	-3,016	18,685
$\tilde{V}(\tilde{F}_s^*(t))$	-1,043	19,081	-91	7,728	-448	9,120	-484	7,715	-1,877	65,298
$\tilde{V}(\tilde{F}_l^*(t))$	-424	18,971	104	7,819	-382	9,110	-301	7,799	-1,058	62,968
$\tilde{V}(\tilde{F}_\pi(t))$	-186	29,720	-603	3,901	31	3,971	500	4,383	-74	28,418
MU284 population with $Y = \ln RMT85$ and $X = \ln REV84$										
$\tilde{V}(\tilde{F}_s(t))$	-2,283	16,303	-1,450	3,538	-945	3,526	-1,071	4,300	-4,832	19,401
$\tilde{V}(\tilde{F}_l(t))$	-1,095	16,755	-1,427	3,181	-938	3,390	-780	4,051	-2,753	20,551
$\tilde{V}(\tilde{F}_s^*(t))$	-1,737	14,642	-298	5,648	-546	5,282	-736	5,679	-3,564	38,344
$\tilde{V}(\tilde{F}_l^*(t))$	-1,174	14,111	-27	5,856	-422	5,452	-228	5,974	-1,433	43,923
$\tilde{V}(\tilde{F}_\pi(t))$	-307	28,421	-460	3,963	-344	3,850	112	4,235	-401	27,987

Table 5.8
Real populations (population size $N = 284$). RBIAS and RRMSE of variance estimators under Poisson sampling with inclusion probabilities proportional to the absolute value of the residuals of the linear regression of the population y_i – values on the population x_i – values. Expected size $n^* = 30$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	RBIAS	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**
MU284 population with $Y = \ln RMT85$ and $X = \ln P85$										
$\tilde{V}(\tilde{F}_s(t))$	-3,502	26,342	-1,841	14,037	-2,691	12,087	-3,415	9,674	-5,932	26,823
$\tilde{V}(\tilde{F}_l(t))$	-2,159	27,610	-1,782	14,010	-2,840	12,002	-3,186	10,177	-4,455	26,802
$\tilde{V}(\tilde{F}_s^*(t))$	-434	22,455	515	15,503	-506	31,296	-1,460	23,496	-2,649	78,527
$\tilde{V}(\tilde{F}_l^*(t))$	-80	22,921	677	15,575	-280	33,294	-1,283	26,612	-1,597	72,166
$\tilde{V}(\tilde{F}_\pi(t))$	-294	361,991	522	75,891	43	48,764	-241	36,354	90	32,354
MU284 population with $Y = \ln RMT85$ and $X = \ln REV84$										
$\tilde{V}(\tilde{F}_s(t))$	-5,220	18,699	-3,667	8,749	-3,222	7,537	-3,018	9,279	-4,955	44,597
$\tilde{V}(\tilde{F}_l(t))$	-4,254	20,765	-3,100	9,180	-3,435	7,231	-3,196	8,540	-3,461	43,206
$\tilde{V}(\tilde{F}_s^*(t))$	-2,938	18,922	-1,110	11,828	-1,265	8,726	-1,040	10,963	-3,682	89,262
$\tilde{V}(\tilde{F}_l^*(t))$	-1,938	19,997	-699	12,641	-1,003	9,305	-599	11,545	-1,558	98,798
$\tilde{V}(\tilde{F}_\pi(t))$	-143	128,401	493	33,934	-255	18,473	-91	17,904	327	16,463

As can be seen from the simulation results, the variance estimators suffer from large variability. This problem is shared by the variance estimator for the Horvitz-Thompson estimator, which occasionally exhibits extremely large RRMSE's. It is further interesting to note that while the RBIAS of the variance estimators for the generalized difference estimators is almost always negative and at times rather large in absolute value, the RBIAS of the variance estimator for the Horvitz-Thompson estimator is in most of the considered cases positive.

Acknowledgements

This research was partially supported by the FAR 2014-ATE-0200 grant from University of Milano-Bicocca.

Appendix

Let β denote a sequence of real numbers. Throughout this appendix we shall indicate by $O_{i_1,i_2,\ldots,i_k}(\beta)$ rest terms that may depend on $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ and that are of the same order as the sequence β uniformly for $i_1, i_2, \ldots, i_k \in U$. Formally, $R\left(x_{i_1}, x_{i_2}, \ldots, x_{i_k}\right) = O_{i_1,i_2,\ldots,i_k}(\beta)$ if

$$\sup_{i_1,i_2,\ldots,i_k \in U} \left| R\left(x_{i_1}, x_{i_2}, \ldots, x_{i_k}\right) \right| = O(\beta).$$

Moreover, to simplify the notation, we shall write m_i in place of $m(x_i)$ and σ_i^2 in place of $\sigma^2(x_i)$.

Bias of the model-based Kuo estimator

$$\begin{aligned}
E\left(\hat{F}(t) - F_N(t)\right) &= E\left(\frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left[I\left(\varepsilon_j \le t - m_j\right) - I\left(\varepsilon_i \le t - m_i\right)\right]\right) \\
&= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left[G\left(t - m_j \middle| x_j\right) - G\left(t - m_i \middle| x_i\right)\right] \\
&= \frac{1}{2N}\sum_{i\notin s}\Big[G^{(2,0)}\left(t - m_i \middle| x_i\right)\left(m_i'\right)^2 - G^{(1,0)}\left(t - m_i \middle| x_i\right)m_i'' \\
&\qquad - 2G^{(1,1)}\left(t - m_i \middle| x_i\right)m_i' + G^{(0,2)}\left(t - m_i \middle| x_i\right)\Big]\sum_{j\in s} w_{i,j}\left(x_j - x_i\right)^2 + o\left(\lambda^2\right) \\
&= \lambda^2 \frac{N-n}{N}\frac{\mu_2}{2\mu_0}\int_a^b \Big[G^{(2,0)}\left(t - m(x) \middle| x\right)\left(m'(x)\right)^2 - G^{(1,0)}\left(t - m(x) \middle| x\right)m''(x) \\
&\qquad - 2G^{(1,1)}\left(t - m(x) \middle| x\right)m'(x) + G^{(0,2)}\left(t - m(x) \middle| x\right)\Big]h_{\bar{s}}(x)\,dx + o\left(\lambda^2\right).
\end{aligned}$$

Bias of the generalized difference Kuo estimator

Write

$$\begin{aligned}
\tilde{F}(t) - F_N(t) &= \frac{1}{N}\Bigg\{\sum_{i\notin s}\sum_{j\in s}\tilde{w}_{i,j}\left[I\left(\varepsilon_j \le t - m_j\right) - I\left(\varepsilon_i \le t - m_i\right)\right] \\
&\qquad + \sum_{i\in s}\left(1 - \frac{1}{\pi_i}\right)\sum_{j\in s}\tilde{w}_{i,j}\left[I\left(\varepsilon_j \le t - m_j\right) - I\left(\varepsilon_i \le t - m_i\right)\right]\Bigg\}.
\end{aligned}$$

Similar steps as those seen for $\hat{F}(t)$ show that

$$E\left(\tilde{F}(t)-F_N(t)\right) = \lambda^2 \frac{N-n}{N}\frac{\mu_2}{2\mu_0}\int_a^b \Big[G^{(2,0)}(t-m(x)|x)\left(m'(x)\right)^2 - G^{(1,0)}(t-m(x)|x)m''(x)$$

$$- 2G^{(1,1)}(t-m(x)|x)m'(x) + G^{(0,2)}(t-m(x)|x)\Big]h(x)\,dx + o\left(\lambda^2\right),$$

where

$$h(x) := h_{\bar{s}}(x) + \left(1-\pi^{-1}(x)\right)h_s(x).$$

Variance of the model-based Kuo estimator

$$\begin{aligned}\operatorname{var}\left(\hat{F}(t)-F_N(t)\right) &= \operatorname{var}\left(\frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} I\left(\varepsilon_j \le t-m_j\right) - \frac{1}{N}\sum_{i\notin s} I\left(y_i \le t\right)\right)\\ &= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\left[G\left(t-m_j|x_j\right)-G^2\left(t-m_j|x_j\right)\right]\\ &\quad + \frac{1}{N^2}\sum_{i\notin s}\left[G\left(t-m_i|x_i\right)-G^2\left(t-m_i|x_i\right)\right]\\ &= A_1 + A_2,\end{aligned}$$

where

$$\begin{aligned}A_1 &:= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\left[G\left(t-m_j|x_j\right)-G^2\left(t-m_j|x_j\right)\right]\\ &= \frac{1}{N^2}\sum_{j\in s}\left[G\left(t-m_j|x_j\right)-G^2\left(t-m_j|x_j\right)\right]\left(\sum_{i\notin s} w_{i,j}\right)^2\\ &= \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G(t-m(x)|x)-G^2(t-m(x)|x)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx\\ &\quad + O\left((n\lambda)^{-1}\alpha\right)\end{aligned}$$

and

$$\begin{aligned}A_2 &:= \frac{1}{N^2}\sum_{i\notin s}\left[G\left(t-m_i|x_i\right)-G^2\left(t-m_i|x_i\right)\right]\\ &= \frac{1}{N-n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)|x\right)-G^2\left(t-m(x)|x\right)\right]h_{\bar{s}}(x)\,dx + O\left(n^{-1}\alpha\right).\end{aligned}$$

Thus,

$$\operatorname{var}\left(\hat{F}(t)-F_N(t)\right) = \frac{1}{n}\left(\frac{N-n}{N}\right)^2 \int_a^b \left[G(t-m(x)|x)-G^2(t-m(x)|x)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx$$
$$+\frac{1}{N-n}\left(\frac{N-n}{N}\right)^2 \int_a^b \left[G(t-m(x)|x)-G^2(t-m(x)|x)\right]h_{\bar{s}}(x)\,dx + O\left((n\lambda)^{-1}\alpha\right).$$

Variance of the generalized difference Kuo estimator

Note that

$$\tilde{F}(t)-F_N(t) = \frac{1}{N}\left\{\sum_{j\in s} I\left(y_j \le t\right)\left[\sum_{i\notin s}\tilde{w}_{i,j} - \sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)+\left(\pi_j^{-1}-1\right)\right] - \sum_{i\notin s} I\left(y_i \le t\right)\right\}$$

so that

$$\begin{aligned}\operatorname{var}\left(\tilde{F}(t)-F_N(t)\right) &= \operatorname{var}\left(\frac{1}{N}\sum_{j\in s} I\left(y_j \le t\right)\left[\sum_{i\notin s}\tilde{w}_{i,j}+\left(\pi_j^{-1}-1\right)-\sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)\right]\right)\\ &\quad + \operatorname{var}\left(\frac{1}{N}\sum_{i\notin s} I\left(y_i \le t\right)\right)\\ &= B_1 + A_2,\end{aligned}$$

where A_2 is the same as in the variance of $\hat{F}(t)$, and where

$$\begin{aligned}B_1 &:= \operatorname{var}\left(\frac{1}{N}\sum_{j\in s} I\left(y_j \le t\right)\left[\sum_{i\notin s}\tilde{w}_{i,j}+\left(\pi_j^{-1}-1\right)-\sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)\right]\right)\\ &= \frac{1}{N^2}\sum_{j\in s}\left[G\left(t-m_j|x_j\right)-G^2\left(t-m_j|x_j\right)\right]\left[\sum_{i\notin s}\tilde{w}_{i,j}+\left(\pi_j^{-1}-1\right)-\sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)\right]^2\\ &= \frac{1}{N^2}\sum_{j\in s}\left[G\left(t-m_j|x_j\right)-G^2\left(t-m_j|x_j\right)\right]\left[\sum_{i\notin s}\tilde{w}_{i,j}+\left(\pi_j^{-1}-1\right)\left(1-\sum_{i\in s}\tilde{w}_{i,j}\right)\right]^2 + O\left(\lambda n^{-1}\right)\\ &= \frac{1}{n}\left(\frac{N-n}{N}\right)^2 \int_a^b \left[G(t-m(x)|x)-G^2(t-m(x)|x)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx\\ &\quad + O\left((n\lambda)^{-1}\alpha+\lambda n^{-1}\right)\\ &= A_1 + O\left((n\lambda)^{-1}\alpha+\lambda n^{-1}\right).\end{aligned}$$

Thus,

$$\operatorname{var}\left(\tilde{F}(t)-F_N(t)\right) = \operatorname{var}\left(\hat{F}(t)-F_N(t)\right) + O\left((n\lambda)^{-1}\alpha+\lambda n^{-1}\right).$$

Bias of the model-based estimator with modified fitted values

Let $\hat{\hat{m}}_i := \sum_{k\in s} w_{i,k} m_k$, $c_{i,j} := 1 - w_{j,j} + w_{i,j}$ and

$$d_{i,j} := \frac{1}{c_{i,j}}\left[(1-c_{i,j})(t-m_i)+(\hat{\hat{m}}_j - m_j)-(\hat{\hat{m}}_i - m_i)+\sum_{k\in s,k\neq j}(w_{j,k}-w_{i,k})\varepsilon_k\right].$$

Observe that $w_{i,j} = O_{i,j}\left((n\lambda)^{-1}\right)$ so that

$$y_j - \hat{m}_j \le t - \hat{m}_i$$

is (asymptotically, as soon as $c_{i,j} > 0$) equivalent to

$$\varepsilon_j \le t - m_i + d_{i,j}.$$

Since $d_{i,j}$ does not depend on ε_j, it follows that

$$\begin{aligned} E\left(I\left(y_j - \hat{m}_j \le t - \hat{m}_i\right)\right) &= E\left(I\left(\varepsilon_j \le t - m_i + d_{i,j}\right)\right) \\ &= E\left(E\left(I\left(\varepsilon_j \le t - m_i + d_{i,j}\right)\middle|\varepsilon_k, k\neq j\right)\right) \\ &= E\left(G\left(t - m_i + d_{i,j}\middle| x_j\right)\right). \end{aligned} \tag{A.1}$$

Now, using the fact that

$$d_{i,j} = (1-c_{i,j})(t-m_i)+(\hat{\hat{m}}_j - m_j)-(\hat{\hat{m}}_i - m_i)+\sum_{k\in s,k\neq j}(w_{j,k}-w_{i,k})\varepsilon_k + R(d_{i,j}), \tag{A.2}$$

where

$$E^{1/4}\left(\left|R(d_{i,j})\right|^4\right) = O_{i,j}\left(\lambda n^{-1} + (n\lambda)^{-3/2}\right), \tag{A.3}$$

it is seen from (A.1) that

$$\begin{aligned} E\left(I\left(y_j - \hat{m}_j \le t - \hat{m}_i\right)\right) &= E\left(G\left(t - m_i + d_{i,j}\right)\middle| x_j\right) \\ &= G\left(t - m_i\middle| x_j\right) + G^{(1,0)}\left(t - m_i\middle| x_j\right)E\left(d_{i,j}\right) \\ &\quad + \frac{1}{2}G^{(2,0)}\left(t - m_i\middle| x_j\right)E\left(d_{i,j}^2\right) + o_{i,j}\left(\lambda^4 + (n\lambda)^{-1}\right). \end{aligned} \tag{A.4}$$

Thus,

$$\begin{aligned} E\left(\hat{F}^*(t) - F_N(t)\right) &= E\left(\frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left(I\left(y_j - \hat{m}_j \le t - \hat{m}_i\right) - I(y_i \le t)\right)\right) \\ &= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left[G\left(t - m_i\middle| x_j\right) - G\left(t - m_i\middle| x_i\right)\right] \\ &\quad + \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}G^{(1,0)}\left(t - m_i\middle| x_j\right)E\left(d_{i,j}\right) \\ &\quad + \frac{1}{2N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}G^{(2,0)}\left(t - m_i\middle| x_j\right)E\left(d_{i,j}^2\right) + o\left(\lambda^4 + (n\lambda)^{-1}\right) \\ &:= C_1 + C_2 + C_3 + o\left(\lambda^4 + (n\lambda)^{-1}\right). \end{aligned} \tag{A.5}$$

Consider first C_1 and note that

$$\begin{aligned} C_1 &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left[G\left(t-m_i|\,x_j\right)-G\left(t-m_i|\,x_i\right)\right] \\ &= \frac{1}{2N}\sum_{i\notin s} G^{(0,2)}\left(t-m_i|\,x_i\right)\sum_{j\in s} w_{i,j}\left(x_j-x_i\right)^2+o\left(\lambda^2\right) \\ &= \lambda^2\frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}\left(t-m(x)|\,x\right)h_{\bar{s}}\left(x\right)dx+o\left(\lambda^2\right). \end{aligned}$$

Consider next C_2. (A.2) and (A.3) imply that

$$\begin{aligned} E\left(d_{i,j}\right) &= \left(1-c_{i,j}\right)\left(t-m_i\right)+\left(\hat{\bar{m}}_j-m_j\right)-\left(\hat{\bar{m}}_i-m_i\right)+O_{i,j}\left(\lambda n^{-1}+(n\lambda)^{-3/2}\right) \\ &= \left(w_{j,j}-w_{i,j}\right)\left(t-m_i\right)+m_j''\sum_{k\in s} w_{j,k}\left(x_k-x_j\right)^2-m_i''\sum_{k\in s} w_{i,k}\left(x_k-x_i\right)^2 \\ &\quad +o_{i,j}\left(\lambda^2\right)+O_{i,j}\left(\lambda n^{-1}+(n\lambda)^{-3/2}\right) \\ &= \left(w_{j,j}-w_{i,j}\right)\left(t-m_i\right)+\left(m_j''-m_i''\right)\sum_{k\in s} w_{j,k}\left(x_k-x_j\right)^2 \\ &\quad +m_i''\left(\sum_{k\in s} w_{j,k}\left(x_k-x_j\right)^2-\sum_{k\in s} w_{i,k}\left(x_k-x_i\right)^2\right) \\ &\quad +o_{i,j}\left(\lambda^2\right)+O_{i,j}\left(\lambda n^{-1}+(n\lambda)^{-3/2}\right) \end{aligned}$$

so that

$$C_2 = C_{2,a}+C_{2,b}+C_{2,c}+o\left(\lambda^2\right)+O\left(\lambda n^{-1}+(n\lambda)^{-3/2}\right),$$

where

$$\begin{aligned} C_{2,a} &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}G^{(1,0)}\left(t-m_i|\,x_j\right)\left(w_{j,j}-w_{i,j}\right)\left(t-m_i\right) \\ &= \frac{1}{N}\sum_{i\notin s} G^{(1,0)}\left(t-m_i|\,x_i\right)\left(t-m_i\right)\sum_{j\in s} w_{i,j}\left(w_{j,j}-w_{i,j}\right)+O\left(n^{-1}\right) \\ &= \frac{1}{n\lambda}\frac{N-n}{N}\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}\left(t-m(x)|\,x\right)\left(t-m(x)\right)\left[h_{\bar{s}}\left(x\right)/h_s\left(x\right)\right]dx \\ &\quad +O\left((n\lambda)^{-1}\lambda^{-1}\alpha+n^{-1}\right) \end{aligned}$$

with $\kappa := \int_{-1}^{1} K^2(u)\,du,$

$$\begin{aligned} C_{2,b} &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}G^{(1,0)}\left(t-m_i|\,x_j\right)\left(m_j''-m_i''\right)\sum_{k\in s} w_{j,k}\left(x_k-x_j\right)^2 \\ &= o\left(\lambda^2\right) \end{aligned}$$

and

$$\begin{aligned} C_{2,c} &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(1,0)}\left(t-m_i|\, x_j\right) m_i''\left(\sum_{k\in s} w_{j,k}\left(x_k-x_j\right)^2 - \sum_{k\in s} w_{i,k}\left(x_k-x_i\right)^2\right) \\ &= \frac{1}{N}\sum_{i\notin s} G^{(1,0)}\left(t-m_i|\, x_i\right) m_i''\left(\sum_{j\in s} w_{i,j}\sum_{k\in s} w_{j,k}\left(x_k-x_j\right)^2 - \sum_{k\in s} w_{i,k}\left(x_k-x_i\right)^2\right) + o\left(\lambda^2\right) \\ &= o\left(\lambda^2\right). \end{aligned}$$

Consider finally C_3. Note that from (A.2) and (A.3)

$$E\left(d_{i,j}^2\right) = \sum_{k\in s}\left(w_{j,k}-w_{i,k}\right)^2\sigma_k^2 + O_{i,j}\left(\lambda^4+(n\lambda)^{-2}\right) \tag{A.6}$$

so that

$$\begin{aligned} C_3 &= \frac{1}{2N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(2,0)}\left(t-m_i|\, x_j\right)\sum_{k\in s}\left(w_{j,k}-w_{i,k}\right)^2\sigma_k^2 + O\left(\lambda^4+(n\lambda)^{-2}\right) \\ &= \frac{1}{2N}\sum_{i\notin s} G^{(2,0)}\left(t-m_i|\, x_i\right)\sigma_i^2\sum_{j\in s} w_{i,j}\sum_{k\in s}\left(w_{j,k}-w_{i,k}\right)^2 + o\left((n\lambda)^{-1}\right) + O\left(\lambda^4\right) \\ &= \frac{1}{n\lambda}\frac{N-n}{N}\frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}\left(t-m(x)|\, x\right)\sigma^2(x)\left[h_{\bar{s}}(x)/h_s(x)\right]dx + o\left((n\lambda)^{-1}\right) + O\left(\lambda^4\right) \end{aligned}$$

with $\theta := \int_{-1}^{1} K(v)\int_{-1}^{1} K(u+v)K(u)\,du\,dv$.

Substituting the above expansions for C_1, C_2 and C_3 into (A.5) yields finally

$$\begin{aligned} E\left(\hat{F}^*(t)-F_N(t)\right) &= \lambda^2\frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}\left(t-m(x)|\, x\right)h_{\bar{s}}(x)\,dx \\ &\quad + \frac{1}{n\lambda}\frac{N-n}{N}\left[\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}\left(t-m(x)|\, x\right)\left(t-m(x)\right)h_s^{-1}(x)h_{\bar{s}}(x)\,dx\right. \\ &\qquad \left. + \frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}\left(t-m(x)|\, x\right)\sigma^2(x)h_s^{-1}(x)h_{\bar{s}}(x)\,dx\right] \\ &\quad + o\left(\lambda^2+(n\lambda)^{-1}\right). \end{aligned}$$

Bias of the generalized difference estimator with modified fitted values

Let $\tilde{d}_{i,j}$ be the design-weighted counterpart of $d_{i,j}$ and observe that

$$\begin{aligned} \tilde{F}^*(t)-F_N(t) &= \frac{1}{N}\left[\sum_{i\notin s}\sum_{j\in s}\tilde{w}_{i,j}\left(I\left(\varepsilon_j\le t-m_i+\tilde{d}_{i,j}\right)-I\left(y_i\le t\right)\right)\right. \\ &\qquad \left. + \sum_{i\in s}\left(1-\pi_i^{-1}\right)\sum_{j\in s}\tilde{w}_{i,j}\left(I\left(\varepsilon_j\le t-m_i+\tilde{d}_{i,j}\right)-I\left(y_i\le t\right)\right)\right]. \end{aligned} \tag{A.7}$$

Adapting the proof that leads to (A.4), it is seen that the asymptotic expansion in (A.4) holds also with $\tilde{d}_{i,j}$ in place of $d_{i,j}$. Adapting the remaining part of the proof finally leads to

$$\begin{aligned} E\left(\tilde{F}^*(t)-F_N(t)\right) &= \lambda^2\frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}(t-m(x)|\,x)h(x)\,dx \\ &\quad+\frac{1}{n\lambda}\frac{N-n}{N}\left[\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}(t-m(x)|\,x)(t-m(x))h_s^{-1}(x)h(x)\,dx\right. \\ &\quad\left.+\frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}(t-m(x)|\,x)\sigma^2(x)h_s^{-1}(x)h(x)\,dx\right] \\ &\quad+o\left(\lambda^2+(n\lambda)^{-1}\right), \end{aligned}$$

where

$$h(x):=h_{\bar{s}}(x)+\left(1-\pi^{-1}(x)\right)h_s(x).$$

Variance of the model-based estimator with modified fitted values

Write

$$\hat{F}^*(t)-F_N(t)=\frac{1}{N}\left(\sum_{i\notin s}\sum_{j\in s}w_{i,j}I\left(\varepsilon_j\le t-m_i+d_{i,j}\right)-\sum_{i\notin s}I\left(\varepsilon_i\le t-m_i\right)\right)$$

and observe that

$$\operatorname{var}\left(\hat{F}^*(t)-F_N(t)\right)=D_1+D_2+D_3,$$

where

$$D_1:=\frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s}w_{i_1,j}w_{i_2,j}\operatorname{cov}\left(I\left(\varepsilon_j\le t-m_{i_1}+d_{i_1,j}\right),I\left(\varepsilon_j\le t-m_{i_2}+d_{i_2,j}\right)\right),$$

$$D_2:=\frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j_1\in s}\sum_{j_2\in s,\,j_2\ne j_1}w_{i_1,j_1}w_{i_2,j_2}\times\operatorname{cov}\left(I\left(\varepsilon_{j_1}\le t-m_{i_1}+d_{i_1,j_1}\right),I\left(\varepsilon_{j_2}\le t-m_{i_2}+d_{i_2,j_2}\right)\right)$$

and where $D_3:=A_2$ from the variance of the model-based Kuo estimator.

Consider D_1. Observe that

$$\begin{aligned} \operatorname{cov}\left(I\left(\varepsilon_j\le t-m_{i_1}+d_{i_1,j}\right),I\left(\varepsilon_j\le t-m_{i_2}+d_{i_2,j}\right)\right) &= E\left(G\left(t-m_{i_1}+d_{i_1,j}\wedge t-m_{i_2}+d_{i_2,j}\middle|x_j\right)\right) \\ &\quad-E\left(G\left(t-m_{i_1}+d_{i_1,j}\middle|x_j\right)\right)E\left(G\left(t-m_{i_2}+d_{i_2,j}\middle|x_j\right)\right). \end{aligned} \tag{A.8}$$

Since

$$\left|\left(t-m_{i_1}+d_{i_1,j}\wedge t-m_{i_2}+d_{i_2,j}\right)-\left(t-m_{i_1}\wedge t-m_{i_2}\right)\right|\le\left|d_{i_1,j}\right|+\left|d_{i_2,j}\right|,$$

it follows from (A.6) that

$$E\left(G\left(t-m_{i_1}+d_{i_1,j}\wedge t-m_{i_2}+d_{i_2,j}\middle|\, x_j\right)\right)=G\left(t-m_{i_1}\wedge t-m_{i_2}\middle|\, x_j\right)\quad+O_{i_1,i_2,j}\left(\lambda^2+(n\lambda)^{-1/2}\right). \tag{A.9}$$

Moreover, from (A.1), (A.4) and (A.6) it follows that

$$E\left(G\left(t-m_i+d_{i,j}\middle|\, x_j\right)\right)=G\left(t-m_i\middle|\, x_j\right)+O_{i,j}\left(\lambda^2+(n\lambda)^{-1/2}\right). \tag{A.10}$$

Using (A.9) and (A.10) to get an asymptotic expansion for the covariance in (A.8), and substituting the outcome into the definition of D_1 yields

$$\begin{aligned}
D_1 &:= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\operatorname{cov}\left(I\left(\varepsilon_j\le t-m_{i_1}+d_{i_1,j}\right),I\left(\varepsilon_j\le t-m_{i_2}+d_{i_2,j}\right)\right)\\
&= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\left[E\left(G\left(t-m_{i_1}+d_{i_1,j}\wedge t-m_{i_2}+d_{i_2,j}\middle|\, x_j\right)\right)\right.\\
&\qquad\qquad \left.-E\left(G\left(t-m_{i_1}+d_{i_1,j}\middle|\, x_j\right)\right)E\left(G\left(t-m_{i_2}+d_{i_2,j}\middle|\, x_j\right)\right)\right]\\
&= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\left[G\left(t-m_{i_1}\wedge t-m_{i_2}\middle|\, x_j\right)-G\left(t-m_{i_1}\middle|\, x_j\right)G\left(t-m_{i_2}\middle|\, x_j\right)\right]\\
&\quad+O\left(\lambda^2 n^{-1}+(n\lambda)^{-1/2}n^{-1}\right)\\
&= \frac{1}{N^2}\sum_{j\in s}\left[G\left(t-m_j\middle|\, x_j\right)-G^2\left(t-m_j\middle|\, x_j\right)\right]\left(\sum_{i\notin s} w_{i,j}\right)^2+O\left(\lambda n^{-1}+(n\lambda)^{-1/2}n^{-1}\right)\\
&= \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)\middle|\, x\right)-G^2\left(t-m(x)\middle|\, x\right)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx\\
&\quad+O\left((n\lambda)^{-1}\alpha+n^{-1}\lambda+n^{-1}(n\lambda)^{-1/2}\right).
\end{aligned} \tag{A.11}$$

Consider next

$$D_2 := \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j_1\in s}\sum_{j_2\in s,\, j_2\neq j_1} w_{i_1,j_1}w_{i_2,j_2}\times\operatorname{cov}\left(I\left(\varepsilon_{j_1}\le t-m_{i_1}+d_{i_1,j_1}\right),I\left(\varepsilon_{j_2}\le t-m_{i_2}+d_{i_2,j_2}\right)\right).$$

Since

$$\operatorname{cov}\left(I\left(\varepsilon_{j_1}\le t-m_{i_1}+d_{i_1,j_1}\right),I\left(\varepsilon_{j_2}\le t-m_{i_2}+d_{i_2,j_2}\right)\right)=0$$

if $\left|x_{i_1}-x_{i_2}\right|>2\lambda,$ it follows that rest terms R_{i_1,j_1,i_2,j_2}, whose contribution to the above covariance is of order $O_{i_1,j_1,i_2,j_2}(\beta)$ for some sequence β that goes to zero, contribute to D_2 a term of order $O(\lambda\beta)$. Now, let

$$b_{i,j_1,j_2} := c_{i,j_1}^{-1}\left(w_{j_1,j_2}-w_{i,j_2}\right),$$

$$a_{i,j_1,j_2} := t-m_i+d_{i,j_1}-b_{i,j_1,j_2}\varepsilon_{j_2}$$

and note that

$$t - m_i + d_{i,j_1} = a_{i,j_1,j_2} + b_{i,j_1,j_2}\varepsilon_{j_2}.$$

Since a_{i,j_1,j_2} does not depend on ε_{j_1} and ε_{j_2}, it follows that

$$\begin{aligned}
&E\left(I\left(\varepsilon_{j_1} \le t - m_{i_1} + d_{i_1,j_1}\right) I\left(\varepsilon_{j_2} \le t - m_{i_2} + d_{i_2,j_2}\right)\right) \\
&\quad = E\left(E\left(I\left(\varepsilon_{j_1} \le a_{i_1,j_1,j_2} + b_{i_1,j_1,j_2}\varepsilon_{j_2}\right) I\left(\varepsilon_{j_2} \le a_{i_2,j_2,j_1} + b_{i_2,j_2,j_1}\varepsilon_{j_1}\right) \mid \varepsilon_k, k \ne j_1, j_2\right)\right) \\
&\quad = E\left(\int_{-\infty}^{\varepsilon^*_{i_1,i_2,j_1,j_2}} G\left(a_{i_2,j_2,j_1} + b_{i_2,j_2,j_1}\varepsilon \mid x_{j_2}\right) dG\left(\varepsilon \mid x_{j_1}\right)\right) \\
&\quad + E\left(\int_{-\infty}^{\varepsilon^*_{i_2,i_1,j_2,j_1}} G\left(a_{i_1,j_1,j_2} + b_{i_1,j_1,j_2}\varepsilon \mid x_{j_1}\right) dG\left(\varepsilon \mid x_{j_2}\right)\right) \\
&\quad - E\left(G\left(\varepsilon^*_{i_1,i_2,j_1,j_2} \mid x_{j_1}\right) G\left(\varepsilon^*_{i_2,i_1,j_2,j_1} \mid x_{j_2}\right)\right),
\end{aligned} \tag{A.12}$$

where

$$\varepsilon^*_{i_1,i_2,j_1,j_2} := \frac{a_{i_1,j_1,j_2} + a_{i_2,j_2,j_1} b_{i_1,j_1,j_2}}{1 - b_{i_1,j_1,j_2} b_{i_2,j_2,j_1}}.$$

Note that the two expectations in the third and fourth lines in (A.12) are the same if i_1 and j_1 are interchanged with i_2 and j_2, respectively. Thus it suffices to analyze the first expectation. Using the fact that

$$\varepsilon^*_{i_1,i_2,j_1,j_2} = t - m_{i_1} + d_{i_1,j_1} + b_{i_1,j_1,j_2}\left(t - m_{i_2} - \varepsilon_{j_2}\right) + R\left(\varepsilon^*_{i_1,i_2,j_1,j_2}\right),$$

where

$$E^{1/4}\left(\left|R\left(\varepsilon^*_{i_1,i_2,j_1,j_2}\right)\right|^4\right) = O_{i_1,i_2,j_1,j_2}\left(\lambda n^{-1} + (n\lambda)^{-3/2}\right),$$

it is seen that

$$\begin{aligned}
&E\left(\int_{-\infty}^{\varepsilon^*_{i_1,i_2,j_1,j_2}} G\left(a_{i_2,j_2,j_1} + b_{i_2,j_2,j_1}\varepsilon \mid x_{j_2}\right) dG\left(\varepsilon \mid x_{j_1}\right)\right) \\
&= G\left(t - m_{i_1} \mid x_{j_1}\right) G\left(t - m_{i_2} \mid x_{j_2}\right) \\
&\quad + G^{(1,0)}\left(t - m_{i_1} \mid x_{j_1}\right) G\left(t - m_{i_2} \mid x_{j_2}\right)\left[E\left(d_{i_1,j_1}\right) + b_{i_1,j_1,j_2}\left(t - m_{i_2}\right)\right] \\
&\quad + G^{(1,0)}\left(t - m_{i_2} \mid x_{j_2}\right) G\left(t - m_{i_1} \mid x_{j_1}\right) E\left(d_{i_2,j_2}\right) + G^{(1,0)}\left(t - m_{i_2} \mid x_{j_2}\right) b_{i_2,j_2,j_1} \int_{-\infty}^{t - m_{i_1}} \varepsilon dG\left(\varepsilon \mid x_{j_1}\right) \\
&\quad + \frac{1}{2} G^{(2,0)}\left(t - m_{i_1} \mid x_{j_1}\right) G\left(t - m_{i_2} \mid x_{j_2}\right) E\left(d^2_{i_1,j_1}\right) + \frac{1}{2} G^{(2,0)}\left(t - m_{i_2} \mid x_{j_2}\right) G\left(t - m_{i_1} \mid x_{j_1}\right) E\left(d^2_{i_2,j_2}\right) \\
&\quad + G^{(1,0)}\left(t - m_{i_1} \mid x_{j_1}\right) G^{(1,0)}\left(t - m_{i_2} \mid x_{j_2}\right) E\left(d_{i_1,j_1} d_{i_2,j_2}\right) \\
&\quad + o_{i_1,i_2,j_1,j_2}\left(\lambda^4 + (n\lambda)^{-1}\right),
\end{aligned} \tag{A.13}$$

and that

$$
\begin{aligned}
&E\left(G\left(\varepsilon^{*}_{i_1,i_2,j_1,j_2}\,\middle|\,x_{j_1}\right)G\left(\varepsilon^{*}_{i_2,i_1,j_2,j_1}\,\middle|\,x_{j_2}\right)\right)\\
&\quad= G\left(t-m_{i_1}\,\middle|\,x_{j_1}\right)G\left(t-m_{i_2}\,\middle|\,x_{j_2}\right)\\
&\qquad+G^{(1,0)}\left(t-m_{i_1}\,\middle|\,x_{j_1}\right)G\left(t-m_{i_2}\,\middle|\,x_{j_2}\right)\left[E\left(d_{i_1,j_1}\right)+b_{i_1,j_1,j_2}\left(t-m_{i_2}\right)\right]\\
&\qquad+G^{(1,0)}\left(t-m_{i_2}\,\middle|\,x_{j_2}\right)G\left(t-m_{i_1}\,\middle|\,x_{j_1}\right)\left[E\left(d_{i_2,j_2}\right)+b_{i_2,j_2,j_1}\left(t-m_{i_1}\right)\right]\\
&\qquad+\frac{1}{2}G^{(2,0)}\left(t-m_{i_1}\,\middle|\,x_{j_1}\right)G\left(t-m_{i_2}\,\middle|\,x_{j_2}\right)E\left(d^2_{i_1,j_1}\right)\\
&\qquad+\frac{1}{2}G^{(2,0)}\left(t-m_{i_2}\,\middle|\,x_{j_2}\right)G\left(t-m_{i_1}\,\middle|\,x_{j_1}\right)E\left(d^2_{i_2,j_2}\right)\\
&\qquad+G^{(1,0)}\left(t-m_{i_1}\,\middle|\,x_{j_1}\right)G^{(1,0)}\left(t-m_{i_2}\,\middle|\,x_{j_2}\right)E\left(d_{i_1,j_1}d_{i_2,j_2}\right)\\
&\qquad+o_{i_1,i_2,j_1,j_2}\left(\lambda^4+(n\lambda)^{-1}\right).
\end{aligned}
\tag{A.14}
$$

Using the asymptotic expansions in (A.4), (A.13) and (A.14) yields

$$
\begin{aligned}
&\operatorname{cov}\left(I\left(\varepsilon_{j_1}\le t-m_{i_1}+d_{i_1,j_1}\right),I\left(\varepsilon_{j_2}\le t-m_{i_2}+d_{i_2,j_2}\right)\right)\\
&\quad=G^{(1,0)}\left(t-m_{i_2}\,\middle|\,x_{j_2}\right)b_{i_2,j_2,j_1}\gamma_{i_1,j_1}+G^{(1,0)}\left(t-m_{i_1}\,\middle|\,x_{j_1}\right)b_{i_1,j_1,j_2}\gamma_{i_2,j_2}\\
&\qquad+G^{(1,0)}\left(t-m_{i_1}\,\middle|\,x_{j_1}\right)G^{(1,0)}\left(t-m_{i_2}\,\middle|\,x_{j_2}\right)\operatorname{cov}\left(d_{i_1,j_1},d_{i_2,j_2}\right)\\
&\qquad+o_{i_1,i_2,j_1,j_2}\left(\lambda^4+(n\lambda)^{-1}\right),
\end{aligned}
\tag{A.15}
$$

where

$$\gamma_{i,j}:=\int_{-\infty}^{t-m_i}\varepsilon dG\left(\varepsilon\,|\,x_j\right).$$

Now observe that

$$b_{i,j_1,j_2}=w_{j_1,j_2}-w_{i,j_2}+O_{i,j_1,j_2}\left((n\lambda)^{-2}\right)$$

and that

$$
\begin{aligned}
\operatorname{cov}\left(d_{i_1,j_1},d_{i_2,j_2}\right)&=\frac{1}{c_{i_1,j_1}c_{i_2,j_2}}\sum_{k\in s;k\ne j_1,j_2}\left(w_{j_1,k}-w_{i_1,k}\right)\left(w_{j_2,k}-w_{i_2,k}\right)\sigma_k^2\\
&=\sum_{k\in s}\left(w_{j_1,k}-w_{i_1,k}\right)\left(w_{j_2,k}-w_{i_2,k}\right)\sigma_k^2+O_{i_1,i_2,j_1,j_2}\left((n\lambda)^{-2}\right)
\end{aligned}
$$

so that

$$D_2=2D_{2a}+D_{2b}+o\left(\lambda^5+n^{-1}\right),\tag{A.16}$$

where

$$
\begin{aligned}
D_{2a} &:= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j_1\in s}\sum_{j_2\in s,\, j_2\neq j_1} w_{i_1,j_1}w_{i_2,j_2}G^{(1,0)}\left(t-m_{i_1}\,|\,x_{j_1}\right)\left(w_{j_1,j_2}-w_{i_1,j_2}\right)\gamma_{i_2,j_2} \\
&= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j_1\in s}\sum_{j_2\in s} w_{i_1,j_1}w_{i_2,j_2}G^{(1,0)}\left(t-m_{i_1}\,|\,x_{j_1}\right)\left(w_{j_1,j_2}-w_{i_1,j_2}\right)\gamma_{i_2,j_2} + O\left(n^{-1}(n\lambda)^{-1}\right) \\
&= \frac{1}{N^2}\sum_{j_2\in s} G^{(1,0)}\left(t-m_{j_2}\,|\,x_{j_2}\right)\gamma_{j_2,j_2}\left[\sum_{j_1\in s} w_{j_1,j_2}\sum_{i_1\notin s} w_{i_1,j_1}\sum_{i_2\notin s} w_{i_2,j_2} - \left(\sum_{i\notin s} w_{i,j_2}\right)^2\right] \\
&\quad + O\left(n^{-1}\lambda + n^{-1}(n\lambda)^{-1}\right) \\
&= O\left((n\lambda)^{-1}\alpha + n^{-1}\lambda + n^{-1}(n\lambda)^{-1}\right)
\end{aligned}
\tag{A.17}
$$

and

$$
\begin{aligned}
D_{2b} &:= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j_1\in s}\sum_{j_2\in s,\, j_2\neq j_1} w_{i_1,j_1}w_{i_2,j_2}G^{(1,0)}\left(t-m_{i_1}\,|\,x_{j_1}\right)G^{(1,0)}\left(t-m_{i_2}\,|\,x_{j_2}\right) \\
&\quad \times\sum_{k\in s}\left(w_{j_1,k}-w_{i_1,k}\right)\left(w_{j_2,k}-w_{i_2,k}\right)\sigma_k^2 \\
&= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j_1\in s}\sum_{j_2\in s} w_{i_1,j_1}w_{i_2,j_2}G^{(1,0)}\left(t-m_{i_1}\,|\,x_{j_1}\right)G^{(1,0)}\left(t-m_{i_2}\,|\,x_{j_2}\right) \\
&\quad \times\sum_{k\in s}\left(w_{j_1,k}-w_{i_1,k}\right)\left(w_{j_2,k}-w_{i_2,k}\right)\sigma_k^2 + O\left(n^{-1}(n\lambda)^{-1}\right) \\
&= \frac{1}{N^2}\sum_{k\in s}\sigma_k^2\left[G^{(1,0)}\left(t-m_k\,|\,x_k\right)\right]^2\left(\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left(w_{j,k}-w_{i,k}\right)\right)^2 + O\left(n^{-1}\lambda + n^{-1}(n\lambda)^{-1}\right) \\
&= \frac{1}{N^2}\sum_{k\in s}\sigma_k^2\left[G^{(1,0)}\left(t-m_k\,|\,x_k\right)\right]^2\left(\sum_{j\in s} w_{j,k}\sum_{i\notin s} w_{i,j} - \sum_{i\notin s} w_{i,k}\right)^2 + O\left(n^{-1}\lambda + n^{-1}(n\lambda)^{-1}\right) \\
&= O\left((n\lambda)^{-1}\alpha + n^{-1}\lambda\right).
\end{aligned}
\tag{A.18}
$$

Putting everything together finally yields

$$
\begin{aligned}
\operatorname{var}\left(\hat{F}^*(t) - F_N(t)\right) &= \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)|x\right) - G^2\left(t-m(x)|x\right)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx \\
&\quad + \frac{1}{N-n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)|x\right) - G^2\left(t-m(x)|x\right)\right]h_{\bar{s}}(x)\,dx + o\left(\lambda^5 + n^{-1}\right).
\end{aligned}
$$

Variance of the generalized difference estimator with modified fitted values

In view of (A.7), we shall show that

$$
\operatorname{var}\left(\tilde{F}^*(t) - F_N(t)\right) = \operatorname{var}\left(\hat{F}^*(t) - F_N(t)\right) + o\left(n^{-1}\right) \tag{A.19}
$$

by showing that

$$\operatorname{var}\left(\frac{1}{N}\sum_{i\in s}\left(1-\pi_i^{-1}\right)\sum_{j\in s}\tilde{w}_{i,j}\left(I\left(\varepsilon_j \le t-m_i+\tilde{d}_{i,j}\right)-I\left(y_i \le t\right)\right)\right)=o\left(n^{-1}\right). \tag{A.20}$$

To prove (A.20) observe that the variance on the left hand side may be written as

$$E_1+E_2+E_3-2E_4-2E_5,$$

where

$$E_1 := \frac{1}{N^2}\sum_{i_1\in s}\sum_{i_2\in s}\sum_{j\in s}\tilde{w}_{i_1,j}\tilde{w}_{i_2,j}\left(1-\pi_{i_1}^{-1}\right)\left(1-\pi_{i_2}^{-1}\right)\times\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_{i_1}+\tilde{d}_{i_1,j}\right),I\left(\varepsilon_j \le t-m_{i_2}+\tilde{d}_{i_2,j}\right)\right),$$

$$E_2 := \frac{1}{N^2}\sum_{i_1\in s}\sum_{i_2\in s}\sum_{j_1\in s}\sum_{j_2\in s,\,j_2\ne j_1}\tilde{w}_{i_1,j}\tilde{w}_{i_2,j_2}\left(1-\pi_{i_1}^{-1}\right)\left(1-\pi_{i_2}^{-1}\right)\times\operatorname{cov}\left(I\left(\varepsilon_{j_1} \le t-m_{i_1}+\tilde{d}_{i_1,j_1}\right),I\left(\varepsilon_{j_2} \le t-m_{i_2}+\tilde{d}_{i_2,j_2}\right)\right),$$

$$E_3 := \frac{1}{N^2}\sum_{i\in s}\left(1-\pi_i^{-1}\right)^2\operatorname{var}\left(I\left(\varepsilon_i \le t-m_i\right)\right),$$

$$E_4 := \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\tilde{w}_{i_1,j}\left(1-\pi_i^{-1}\right)\left(1-\pi_j^{-1}\right)\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_i+\tilde{d}_{i,j}\right),I\left(\varepsilon_j \le t-m_j\right)\right),$$

and finally

$$E_5 := \frac{1}{N^2}\sum_{i_1\in s}\sum_{i_2\in s}\sum_{j\in s,\,j\ne i_2}\tilde{w}_{i_1,j}\left(1-\pi_{i_1}^{-1}\right)\left(1-\pi_{i_2}^{-1}\right)\times\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_{i_1}+\tilde{d}_{i_1,j}\right),I\left(\varepsilon_{i_2} \le t-m_{i_2}\right)\right).$$

To begin with, consider E_1 and E_2. Observe that except for (i) the fact that the summation indexes i_1 and i_2 range over s instead of the complement of s in U, (ii) the presence of the factors $\left(1-\pi_i^{-1}\right)$ and (iii) the fact that the $w_{i,j}$'s and the $d_{i,j}$'s are substituted by their design-weighted counterparts $\tilde{w}_{i,j}$ and $\tilde{d}_{i,j}$, E_1 and E_2 are the same as D_1 and D_2 from $\operatorname{var}\left(\hat{F}^*(t)-F_N(t)\right)$, respectively. Adapting the proofs that lead to the asymptotic expansions for D_1 and D_2 shows thus that

$$E_1 = \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)|x\right)-G^2\left(t-m(x)|x\right)\right]\left[1-\pi^{-1}(x)\right]^2 h_s(x)\,dx+o\left(n^{-1}\right)$$

and that

$$E_2 = o\left(\lambda^5+n^{-1}\right).$$

As for E_3 it is immediately seen that

$$E_3 = E_1+o\left(n^{-1}\right),$$

while in order to deal with E_4 and E_5 we shall need asymptotic expansions for

$$\operatorname{cov}\left(I\left(\varepsilon_j \le t - m_{i_1} + \tilde{d}_{i_1,j}\right), I\left(\varepsilon_{i_2} \le t - m_{i_2}\right)\right) \tag{A.21}$$

for the case when $j = i_2$ and the case when $j \neq i_2$. In the former case we may employ arguments similar to those for proving (A.9) and (A.10), which lead to

$$\begin{aligned}&\operatorname{cov}\left(I\left(\varepsilon_j \le t - m_{i_1} + \tilde{d}_{i_1,j}\right), I\left(\varepsilon_j \le t - m_j\right)\right)\\ &\quad = G\left(t - m_{i_1} \wedge t - m_j \,|\, x_j\right) - G\left(t - m_{i_1} \,|\, x_j\right) G\left(t - m_j \,|\, x_j\right) + O\left(\lambda^2 + (n\lambda)^{-1/2}\right).\end{aligned}$$

When $j \neq i_2$, on the other hand, the covariance in (A.21) is different from zero only if $\left|x_j - x_{i_2}\right| \le \lambda$ or $\left|x_{i_1} - x_{i_2}\right| \le \lambda$, and adapting (A.12) it can be shown that

$$\begin{aligned}&E\left(I\left(\varepsilon_j \le t - m_{i_1} + \tilde{d}_{i_1,j}\right) I\left(\varepsilon_{i_2} \le t - m_{i_2}\right)\right)\\ &\quad = E\left(E\left(I\left(\varepsilon_j \le \tilde{a}_{i_1,j,i_2} + \tilde{b}_{i_1,j,i_2}\varepsilon_{i_2}\right) I\left(\varepsilon_{i_2} \le t - m_{i_2}\right) \middle| \varepsilon_k, k \neq i, j\right)\right)\\ &\quad = E\left(\int_{-\infty}^{t-m_{i_2}} G\left(\tilde{a}_{i_1,j,i_2} + \tilde{b}_{i_1,j,i_2}\varepsilon \,\middle|\, x_j\right) dG\left(\varepsilon \,|\, x_{i_2}\right)\right)\\ &\quad = G\left(t - m_{i_1} \,|\, x_j\right) G\left(t - m_{i_2} \,|\, x_{i_2}\right) + G\left(t - m_{i_2} \,|\, x_{i_2}\right) G^{(1,0)}\left(t - m_{i_1} \,|\, x_j\right) E\left(d_{i_1,j}\right)\\ &\qquad + G^{(1,0)}\left(t - m_{i_1} \,|\, x_j\right) \tilde{b}_{i_1,j,i_2}\gamma_{i_2,i_2} + \frac{1}{2} G\left(t - m_{i_2} \,|\, x_{i_2}\right) G^{(2,0)}\left(t - m_{i_1} \,|\, x_j\right) E\left(d_{i_1,j}^2\right)\\ &\qquad + o_{i_1,i_2,j}\left(\lambda^4 + (n\lambda)^{-1}\right),\end{aligned}$$

where $\tilde{a}_{i,j,k}$ and $\tilde{b}_{i,j,k}$ are the design-weighted counterparts of $a_{i,j,k}$ and $b_{i,j,k}$, respectively. Adapting also (A.4) to account for the design-weights, it is seen that

$$\begin{aligned}\operatorname{cov}\left(I\left(\varepsilon_j \le t - m_{i_1} + \tilde{d}_{i_1,j}\right), I\left(\varepsilon_{i_2} \le t - m_{i_2}\right)\right) &= G^{(1,0)}\left(t - m_{i_1} \,|\, x_j\right) \tilde{b}_{i_1,j,i_2}\gamma_{i_2,i_2} + o_{i_1,i_2,j}\left(\lambda^4 + (n\lambda)^{-1}\right)\\ &= G^{(1,0)}\left(t - m_i \,|\, x_j\right)\left(\tilde{w}_{j,i_2} - \tilde{w}_{i_1,i_2}\right)\gamma_{i_2,i_2} + o_{i_1,i_2,j}\left(\lambda^4 + (n\lambda)^{-1}\right)\end{aligned}$$

so that (cfr. the steps that lead to the asymptotic expansions of the terms D_1 and D_2 in the variance of the model-based two-step estimator)

$$E_4 = E_1 + o\left(n^{-1}\right)$$

and

$$E_5 = o\left(\lambda^5 + n^{-1}\right).$$

This completes the proof of (A.20) and thus (A.19) follows.

References

Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals Statistics*, 28(4), 1026-1053.

Chambers, R.L., and Clark, R. (2012). *An Introduction to Model-Based Survey Sampling with Applications*, Oxford Statistical Science Series 37.

Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73(3), 597-604.

Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993). Bias robust estimation in finite populations using non-parametric calibration. *Journal of the American Statistical Association,* 88(421), 268-277.

Chen, J., and Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 12, 1223-1239.

Dorfman, A.H., and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *The Annals of Statistics*, 21(3), 1452-1475.

Fan, J., and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, 20(4), 2008-2036.

Hansen, B.E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24, 726-748.

Johnson, A.A., Breidt, F.J. and Opsomer, J.D. (2008). Estimating distribution functions from survey data using nonparametric regression. *Journal of Statistical Theory and Practice,* 2(3), 419-431.

Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. In *Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, 280-285.

Montanari, G.E., and Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100(472), 1429-1442.

Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77(2), 365-375.

Rueda, M., Martínez, S., Martínez, H. and Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, 137(2), 435-448.

Rueda, M., Sànchez-Borrego, I., Arcos, A. and Martínez, S. (2010). Model-calibration estimation of the distribution function using nonparametric regression. *Metrika*, 71(1), 33-44.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer.

Wang, J.C., and Opsomer, J.D. (2011). On asymptotic normality and variance estimation for nondifferentiable survey estimators. *Biometrika*, 98(1), 91-106.

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90(4), 937-951.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453), 185-193.

Survey Methodology, June 2016
Vol. 42, No. 1, pp. 121-135
Statistics Canada, Catalogue No. 12-001-X

A note on regression estimation with unknown population size

Michael A. Hidiroglou, Jae Kwang Kim and Christian Olivier Nambeu[1]

Abstract

The regression estimator is extensively used in practice because it can improve the reliability of the estimated parameters of interest such as means or totals. It uses control totals of variables known at the population level that are included in the regression set up. In this paper, we investigate the properties of the regression estimator that uses control totals estimated from the sample, as well as those known at the population level. This estimator is compared to the regression estimators that strictly use the known totals both theoretically and via a simulation study.

Key Words: Optimal estimator; Survey sampling; Weighting.

1 Introduction

Regression estimation has been increasingly used in large survey organizations as a means to improve the reliability of the estimators of parameters of interest (such as totals or means) when auxiliary variables are available in the population. A comprehensive overview of the regression estimator in survey sampling can be found in Cassel, Särndal and Wretman (1976) and Fuller (2009) among others. We next illustrate how the regression estimator can be used to estimate the total, $Y = \sum_{i \in U} y_i$ where $U = \{1, \ldots, N\}$ denotes the target population. A sample s of expected size n is selected according to a sampling plan $p(s)$ from U, where π_i is the resulting probability of inclusion of the first order. In the absence of auxiliary variables, we use the Horvitz-Thompson estimator given by $\hat{Y}_\pi = \sum_{i \in s} d_i y_i$ (Horvitz and Thompson 1952) where $d_i = 1/\pi_i$ is referred to as the weight survey associated with unit i. The regression estimator is given by

$$\hat{Y}_{\text{REG}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\mathrm{T}} \hat{\mathbf{B}}, \tag{1.1}$$

where $\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$, $\hat{\mathbf{X}}_\pi = \sum_{i \in s} d_i \mathbf{x}_i$, $\mathbf{x}_i = \left(1, x_{2i}, \ldots, x_{pi}\right)^{\mathrm{T}}$, and $\hat{\mathbf{B}}$ is a $p-$dimensional vector of estimated regression coefficients, which is computed as a function of the observed variables $\left(y_i, \mathbf{x}_i^{\mathrm{T}}\right)^{\mathrm{T}}$ in the sample s.

Note that the components of the vector of population total $\mathbf{X}$ are known for each of the corresponding components variables in the vector $\mathbf{x}_i = \left(1, x_{2i}, \ldots, x_{pi}\right)^{\mathrm{T}}$ used to compute $\hat{\mathbf{B}}$. However, there are instances when we have more observed auxiliary variables in the sample than in the population. Assume that the sample has q observed variables $(q > p)$, and that the p variables in the population are a subset of the q variables observed in the sample. Furthermore, suppose that some of the extra $q - p$ variables in the sample are well correlated with the variable of interest y. Can these extra variables be incorporated in the

1. Michael A. Hidiroglou, Business Survey Methods Division, Statistics Canada, ON, Canada K1A 0T6. E-mail: hidirog@yahoo.ca; Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, IA 50011. E-mail: jkim@iastate.edu; Christian Olivier Nambeu, Business Survey Methods Division, Statistics Canada, ON, Canada K1A 0T6. E-mail: christianolivier.nambeu@canada.ca.

regression estimator so as to make it more efficient? Singh and Raghunath (2011) attempted to respond to that question for the case where $q = p + 1$. Their extra variable in the sample was the intercept. They used it to estimate the unknown population size N by $\hat{N} = \sum_{i \in s} d_i$.

In this article, we compare the estimator proposed by Singh and Raghunath (2011) to other regression estimators when N is known or unknown. In Section 2, we describe standard regression estimators for estimating totals when N is known as well as the regression proposed by Singh and Raghunath (2011) when N is unknown. In Section 3, an alternative estimator is proposed for the case where N is unknown. A simulation study is carried out in Section 4, to illustrate the performance of the various estimators studied in terms of bias and mean square error. Overall conclusions and recommendations are given in Section 5.

2 Regression estimators

Under general regularity conditions (Isaki and Fuller 1982; Montanari 1987), an approximation to the regression estimator (1.1) is

$$\tilde{Y}_{\text{REG}} = \hat{Y}_{\pi} + \left(\mathbf{X} - \hat{\mathbf{X}}_{\pi}\right)^{\text{T}} \mathbf{B}, \tag{2.1}$$

where $\mathbf{B}$ is the limit in probability of $\hat{\mathbf{B}}$ when both the sample and the population sizes tend to infinity. For large samples, the variance of regression estimator (1.1) can be studied via (2.1). Note that $\tilde{Y}_{\text{REG}}$ is unbiased under the sampling plan $p(s)$ and can be re-expressed as:

$$\tilde{Y}_{\text{REG}} = \mathbf{X}^{\text{T}}\mathbf{B} + \sum_{i \in s} d_i E_i, \tag{2.2}$$

where $E_i = y_i - \mathbf{x}_i^{\text{T}}\mathbf{B}$.

The design variance for $\hat{Y}_{\text{REG}}$ can be approximated by

$$\text{AV}_p\left(\hat{Y}_{\text{REG}}\right) = \sum_{i \in U}\sum_{j \in U} \Delta_{ij} \frac{E_i}{\pi_i} \frac{E_j}{\pi_j}, \tag{2.3}$$

where $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$ and π_{ij} is the second order inclusion probability for units i and j. Both the model-assisted (Särndal, Swensson and Wretman 1992) and the optimal-variance (Montanari 1987) approaches can be used to estimate $\mathbf{B}$. They both yield approximately unbiased estimators. In the case of the model-assisted approach, the basic properties (bias and variance terms) are valid even when the model is not correctly specified. Under the optimal-variance approach no assumption is made on the variable of interest.

The model-assisted estimator of Särndal et al. (1992) assumes a working model between the variable of interest (y) and the auxiliary variables $(\mathbf{x})$. The working model is denoted by $m: y_i = \mathbf{x}_i^{\text{T}}\boldsymbol{\beta} + \varepsilon_i$ where $\boldsymbol{\beta}$ is a vector of p unknown parameters, $E_m(\varepsilon_i \mid \mathbf{x}_i) = 0$, $V_m(\varepsilon_i \mid \mathbf{x}_i) = \sigma_i^2$, and $\text{Cov}_m\left(\varepsilon_i, \varepsilon_j \mid \mathbf{x}_i, \mathbf{x}_j\right) = 0, i \neq j$. Under this approach, $\mathbf{B}$ in equation (2.1) is the ordinary least squares estimator of $\boldsymbol{\beta}$ in the population and it is given by

$$\mathbf{B}_{\text{GREG}} = \left(\sum_{i \in U} c_i \mathbf{x}_i \mathbf{x}_i^{\text{T}}\right)^{-1} \left(\sum_{i \in U} c_i \mathbf{x}_i y_i\right), \tag{2.4}$$

where $c_i = \sigma_i^{-2}$. This yields the following estimator for the total Y

$$\hat{Y}_{\text{GREG}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\text{T}} \hat{\mathbf{B}}_{\text{GREG}}, \tag{2.5}$$

where

$$\hat{\mathbf{B}}_{\text{GREG}} = \left(\sum_{i\in s} c_i d_i \mathbf{x}_i \mathbf{x}_i^{\text{T}}\right)^{-1} \left(\sum_{i\in s} c_i d_i \mathbf{x}_i y_i\right). \tag{2.6}$$

The optimal estimator of Montanari (1987), obtained by minimizing the design variance of

$$\tilde{Y}_{\text{REG}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\text{T}} \mathbf{B},$$

is

$$\tilde{Y}_{\text{OPT}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\text{T}} \mathbf{B}_{\text{OPT}}, \tag{2.7}$$

where

$$\begin{aligned} \mathbf{B}_{\text{OPT}} &= \left\{V\left(\hat{\mathbf{X}}_\pi\right)\right\}^{-1} \text{Cov}\left(\hat{\mathbf{X}}_\pi, \hat{Y}_\pi\right) \\ &= \left(\sum_{i\in U}\sum_{j\in U} \Delta_{ij} \frac{\mathbf{x}_i}{\pi_i} \frac{\mathbf{x}_j^{\text{T}}}{\pi_j}\right)^{-1} \left(\sum_{i\in U}\sum_{j\in U} \Delta_{ij} \frac{\mathbf{x}_i}{\pi_i} \frac{y_j}{\pi_j}\right). \end{aligned} \tag{2.8}$$

The optimal estimator for the total Y is estimated by

$$\hat{Y}_{\text{OPT}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\text{T}} \hat{\mathbf{B}}_{\text{OPT}}, \tag{2.9}$$

where

$$\hat{\mathbf{B}}_{\text{OPT}} = \left(\sum_{i\in s}\sum_{j\in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\mathbf{x}_i}{\pi_i} \frac{\mathbf{x}_j^{\text{T}}}{\pi_j}\right)^{-1} \left(\sum_{i\in s}\sum_{j\in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\mathbf{x}_i}{\pi_i} \frac{y_j}{\pi_j}\right). \tag{2.10}$$

Note that the computation of the regression vectors requires that the first component that defines them is invertible. We can ensure this by reducing the number of auxiliary variables that are input into the regression if not much loss in efficiency of the resulting regression estimator is incurred. If, on the other hand, there is a significant loss in efficiency, then we can invert these singular matrices using generalised inverses.

As mentioned in the introduction, not all population totals may be known for each component of the auxiliary vector $\mathbf{x}$. The regression normally uses the auxiliary variables for which a corresponding population total is known. Decomposing $\mathbf{x}_i$ as $\left(1, \mathbf{x}_i^{*\text{T}}\right)^{\text{T}}$ where $\mathbf{x}_i^* = \left(x_{2i}, \ldots, x_{pi}\right)^{\text{T}}$, Singh and Raghunath (2011) proposed a GREG-like estimator that assumes that the regression is based on an intercept and the variable $\mathbf{x}^*$, even though only the population total of the $\mathbf{x}^*$ is known.

For the case that N is not known and that the population total of $\mathbf{x}^*$ is known, their estimator is

$$\hat{Y}_{\text{SREG}} = \hat{Y}_{\pi} + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}_{\pi}^{*}\right)^{\text{T}} \hat{\mathbf{B}}_{2,\text{GREG}}, \tag{2.11}$$

where $\mathbf{X}^{*} = \sum_{i \in U} \mathbf{x}_i^{*}$ and $\hat{\mathbf{X}}_{\pi}^{*} = \sum_{i \in s} d_i \mathbf{x}_i^{*}$. The regression vector of estimated coefficients $\hat{\mathbf{B}}_{2,\text{GREG}}$ is obtained from $\hat{\mathbf{B}}_{\text{GREG}} = \left(\hat{B}_{1,\text{GREG}}, \hat{\mathbf{B}}_{2,\text{GREG}}^{\text{T}}\right)^{\text{T}}$ given by (2.6). The approximate design variance for $\hat{Y}_{\text{SREG}}$ takes the same form as equation (2.3), with $E_i = y_i - \mathbf{x}_i^{*\text{T}} \mathbf{B}_{2,\text{GREG}}$, where

$$\mathbf{B}_{2,\text{GREG}} = \left\{\sum_{i \in U} c_i \left(\mathbf{x}_i^{*} - \bar{\mathbf{X}}_N^{*}\right)\left(\mathbf{x}_i^{*} - \bar{\mathbf{X}}_N^{*}\right)^{\text{T}}\right\}^{-1} \sum_{i \in U} c_i \left(\mathbf{x}_i^{*} - \bar{\mathbf{X}}_N^{*}\right) y_i$$

and $\bar{\mathbf{X}}_N^{*} = \sum_{i \in U} \mathbf{x}_i^{*} / N$.

The properties of (2.11) can be obtained by noting that

$$\begin{aligned}
\hat{Y}_{\text{SREG}} - Y &= \hat{Y}_{\pi} - Y + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}_{\pi}^{*}\right)^{\text{T}} \hat{\mathbf{B}}_{2,\text{GREG}} \\
&= \hat{Y}_{\pi} - Y + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}_{\pi}^{*}\right)^{\text{T}} \mathbf{B}_{2,\text{GREG}} + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}_{\pi}^{*}\right)^{\text{T}} \left(\hat{\mathbf{B}}_{2,\text{GREG}} - \mathbf{B}_{2,\text{GREG}}\right).
\end{aligned}$$

Since $\hat{\mathbf{B}}_{2,\text{GREG}} - \mathbf{B}_{2,\text{GREG}} = O_p\left(n^{-1/2}\right)$ under some regularity conditions discussed in Fuller (2009, Chapter 2), the last term is of smaller order. Thus, ignoring the smaller order terms, we get the following approximation

$$\hat{Y}_{\text{SREG}} - Y \cong \sum_{i \in s} d_i E_i - \sum_{i \in U} E_i, \tag{2.12}$$

where $E_i = y_i - \mathbf{x}_i^{*\text{T}} \mathbf{B}_{2,\text{GREG}}$. Thus, $\hat{Y}_{\text{SREG}}$ is approximately design-unbiased. The asymptotic variance can be computed using

$$V\left\{\sum_{i \in s} d_i E_i - \sum_{i \in U} E_i\right\} = E\left\{\left(\sum_{i \in s} d_i E_i - \sum_{i \in U} E_i\right)^2\right\}.$$

As we can see, the asymptotic variance can be quite large unless $\sum_{i \in U} E_i = 0$.

Remark 2.1 *If* $y_i = a + bx_i$, *we have* $\hat{Y}_{\text{SREG}} - Y = \left(\hat{N}_{\pi} - N\right)a$ *and this implies that* $V\left(\hat{Y}_{\text{SREG}}\right) = a^2 V\left(\hat{N}_{\pi}\right)$. *This means that if* $V\left(\hat{N}_{\pi}\right) > 0$, *we can artificially increases* $a^2 V\left(\hat{N}_{\pi}\right)$, *the variance of* $\hat{Y}_{\text{SREG}}$, *by choosing large values of* a.

Note that the optimal regression estimator using $\mathbf{x}^{*} = \left(x_2, \ldots, x_p\right)^{\text{T}}$ is also approximately design unbiased because

$$\begin{aligned}
\hat{Y}_{\text{OPT}}^{*} - Y &= \hat{Y}_{\pi} - Y + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}_{\pi}^{*}\right)^{\text{T}} \hat{\mathbf{B}}_{\text{OPT}}^{*} \\
&= \hat{Y}_{\pi} - Y + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}_{\pi}^{*}\right)^{\text{T}} \mathbf{B}_{\text{OPT}}^{*} + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}_{\pi}^{*}\right)^{\text{T}} \left(\hat{\mathbf{B}}_{\text{OPT}}^{*} - \mathbf{B}_{\text{OPT}}^{*}\right),
\end{aligned}$$

where $\mathbf{B}_{\text{OPT}}^{*}$ is obtained by replacing $\mathbf{x}_i$ by $\mathbf{x}_i^{*}$ in equation (2.8). Since $\hat{\mathbf{B}}_{\text{OPT}}^{*} - \mathbf{B}_{\text{OPT}}^{*} = O_p\left(n^{-1/2}\right)$ under some regularity conditions discussed in Fuller (2009, Chapter 2), ignoring the smaller order terms we get

$$\hat{Y}^{*}_{\text{OPT}} - Y \cong \hat{Y}_{\pi} - Y + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}^{*}_{\pi}\right)^{\text{T}} \mathbf{B}^{*}_{\text{OPT}}.$$

The asymptotic variance of $\hat{Y}^{*}_{\text{OPT}}$ is smaller than the one associated with $\hat{Y}_{\text{SREG}}$. The reason for this is that the optimal estimator minimizes the asymptotic variance among the class of estimators of the form

$$\hat{Y}_{B} = \hat{Y}_{\pi} + \left(\mathbf{X}^{*} - \hat{\mathbf{X}}^{*}_{\pi}\right)^{\text{T}} \hat{\mathbf{B}} \tag{2.13}$$

indexed by $\hat{\mathbf{B}}$.

3 Alternative regression estimator

We now consider an alternative estimator that does not use the population size (N) information. Rather, it uses the known inclusion probabilities π_i provided that they are known for each unit in the population. Given that $\sum_{i\in U} \pi_i = n$, we can use $\mathbf{z}_i = \left(\pi_i, \mathbf{x}^{*\text{T}}_i\right)^{\text{T}}$ as auxiliary data in the model

$$y_i = \mathbf{z}^{\text{T}}_i \beta + e_i,$$

where $e_i \overset{\text{ind}}{\sim} \left(0, \sigma^2 \pi_i\right)$. This means that the incorporation of the variance structure c_i of the error in the regression vector is given by $c_i = d_i / \sigma^2$. The resulting estimator is given by

$$\hat{Y}_{\text{KREG}} = \hat{Y}_{\pi} + \left(\mathbf{Z} - \hat{\mathbf{Z}}_{\pi}\right)^{\text{T}} \hat{\mathbf{B}}_{\text{KREG}}, \tag{3.1}$$

with $\mathbf{Z} = \sum_{i\in U} \mathbf{z}_i$, $\hat{\mathbf{Z}} = \sum_{i\in s} d_i \mathbf{z}_i$ and

$$\hat{\mathbf{B}}_{\text{KREG}} = \left(\sum_{i\in s} c_i d_i \mathbf{z}_i \mathbf{z}^{\text{T}}_i\right)^{-1} \sum_{i\in s} c_i d_i \mathbf{z}_i y_i. \tag{3.2}$$

This estimator corresponds exactly to the one given by Isaki and Fuller (1982).

Remark 3.1 *By construction,*

$$\sum_{i\in s} d^2_i \left(y_i - \mathbf{z}^{\text{T}}_i \hat{\mathbf{B}}_{\text{KREG}}\right) \mathbf{z}_i = \mathbf{0}.$$

Since π_i *is a component of* $\mathbf{z}_i$, *we have* $\sum_{i\in s} d_i \left(y_i - \mathbf{z}^{\text{T}}_i \hat{\mathbf{B}}_{\text{KREG}}\right) = 0$, *this leads to*

$$\hat{Y}_{\text{KREG}} = \mathbf{Z}^{\text{T}} \hat{\mathbf{B}}_{\text{KREG}}.$$

Thus, $\hat{Y}_{\text{KREG}}$ *is the best linear unbiased predictor of* $Y = \sum_{i=1}^{N} y_i$ *under the model*

$$y_i = \pi_i \beta_1 + \mathbf{x}^{*\text{T}}_i \boldsymbol{\beta}_2 + e_i,$$

where $e_i \sim \left(0, \sigma^2 \pi_i\right)$.

Note that $\hat{\mathbf{B}}_{\text{KREG}}$ can be expressed as $\hat{\mathbf{B}}_{\text{GREG}}$ by setting $c_i = d_i/\sigma^2$ and $\mathbf{x}_i = \mathbf{z}_i$. Thus, the proposed regression estimator can be viewed as a special case of GREG estimator. Using the argument similar to (2.12), we obtain

$$\hat{Y}_{\text{KREG}} - Y \cong \sum_{i\in s} d_i E_i^* - \sum_{i\in U} E_i^*, \tag{3.3}$$

where $E_i^* = y_i - \mathbf{z}_i^{\text{T}}\mathbf{B}_{\text{KREG}}$ and

$$\mathbf{B}_{\text{KREG}} = \left(\sum_{i\in U} c_i \mathbf{z}_i \mathbf{z}_i^{\text{T}}\right)^{-1} \sum_{i\in U} c_i \mathbf{z}_i y_i.$$

The proposed estimator is approximately unbiased and its asymptotic variance

$$V\left\{\sum_{i\in s} d_i \left(y_i - \mathbf{z}_i^{\text{T}}\mathbf{B}_{\text{KREG}}\right)\right\} = \sum_{i\in U}\sum_{j\in U} \Delta_{ij} \frac{E_i^*}{\pi_i}\frac{E_j^*}{\pi_j}$$

is often smaller than the asymptotic variance of Singh and Raghunath (2011)'s estimator.

The optimal version of $\hat{Y}_{\text{KREG}}$ uses $\mathbf{z}_i = \left(\pi_i, \mathbf{x}_i^{*\text{T}}\right)^{\text{T}}$ as auxiliary data. It is given by

$$\hat{Y}_{\text{KOPT}} = \hat{Y}_\pi + \left(\mathbf{Z} - \hat{\mathbf{Z}}_\pi\right)^{\text{T}} \hat{\mathbf{B}}_{\text{KOPT}}, \tag{3.4}$$

where $\hat{\mathbf{B}}_{\text{KOPT}}$ is obtained by substituting $\mathbf{x}_i$ by $\mathbf{z}_i$ in equation (2.10).

Remark 3.2 *For fixed-size sampling designs, we have* $V_p\left(\sum_{i\in s} d_i\pi_i\right) = 0$. *In this case, the optimal regression coefficient vector* $\mathbf{B}_{\text{KOPT}} = V_p\left(\hat{\mathbf{Z}}_\pi\right)^{-1} \text{Cov}_p\left(\hat{\mathbf{Z}}_\pi, \hat{Y}_\pi\right)$ *cannot be computed because the variance-covariance matrix* $V_p\left(\hat{\mathbf{Z}}_\pi\right)$ *is not invertible. Thus, the optimal estimator with* $\mathbf{z}_i = \left(\pi_i, \mathbf{x}_i^{*\text{T}}\right)^{\text{T}}$ *reduces to the optimal estimator* (2.9) *only using* $\mathbf{x}_i^*$.

Remark 3.3 *For random-size sampling designs,* $V_p\left(\sum_{i\in s} d_i\pi_i\right) \geq 0$. *In this case, all of the components of* $\mathbf{z}_i = \left(\pi_i, \mathbf{x}_i^{*\text{T}}\right)^{\text{T}}$ *can be used in the design-optimal regression estimator* (2.9).

A difficulty with using the optimal estimator $\hat{Y}_{\text{KOPT}}$ is that it requires the computation of the joint inclusion probabilities π_{ij}: these may be difficult to compute for certain sampling designs. An estimator that does not require the computation of the joint inclusion probabilities is obtained by assuming that $\pi_{ij} = \pi_i\pi_j$. We refer to this estimator as the pseudo-optimal estimator, $\hat{Y}_{\text{POPT}}$. It is given by

$$\hat{Y}_{\text{POPT}} = \hat{Y}_\pi + \left(\mathbf{Z} - \hat{\mathbf{Z}}_\pi\right)^{\text{T}} \hat{\mathbf{B}}_{\text{POPT}}, \tag{3.5}$$

where

$$\hat{\mathbf{B}}_{\text{POPT}} = \left(\sum_{i\in s} c_i d_i \mathbf{z}_i \mathbf{z}_i^{\text{T}}\right)^{-1} \sum_{i\in s} c_i d_i \mathbf{z}_i y_i$$

and

$$c_i = d_i - 1.$$

In general, the pseudo-optimal estimator $\hat{Y}_{\text{POPT}}$ should yield estimates that are quite close to those produced by $\hat{Y}_{\text{KREG}}$ when the sampling fraction is small. Note that $\hat{Y}_{\text{POPT}}$ is exactly equal to the optimal estimator $\hat{Y}_{\text{KOPT}}$ in the case of Poisson sampling. In this sampling design the inclusion probabilities of units in the sample are independent. The approximate design variance for $\hat{Y}_{\text{KREG}}$, $\hat{Y}_{\text{KOPT}}$ and $\hat{Y}_{\text{POPT}}$ have the same form as the one given in equation (2.3) with the E_i's respectively given by $y_i - \mathbf{z}_i^{\text{T}}\hat{\mathbf{B}}_{\text{KREG}}$, $y_i - \mathbf{z}_i^{\text{T}}\hat{\mathbf{B}}_{\text{KOPT}}$ and $y_i - \mathbf{z}_i^{\text{T}}\hat{\mathbf{B}}_{\text{POPT}}$.

4 Simulations

We carried out two simulation studies. The first one used a dataset provided in the textbook of Rosner (2006) and the second one was based on an artificial population created according to a simple linear regression model. The first simulation assessed the performance of all of the estimators with respect to different sample schemes while the second simulation study focused on the impact of changing the intercept value in the model.

The parameter of interest for these two simulations is the total of the variable of interest y: $Y = \sum_{i \in U} y_i$. All estimators were used $\left(\hat{Y}_{\text{GREG}}, \hat{Y}_{\text{OPT}}, \hat{Y}_{\text{POPT}}, \hat{Y}_{\text{SREG}}, \hat{Y}_{\text{KREG}} \text{ and } \hat{Y}_{\text{KOPT}}\right)$ with the available auxiliary data. Table 4.1 summarizes the auxiliary data and the variance structure of the errors (when applicable) associated with the estimators used in the two studies.

Table 4.1
Estimators used in simulation

***N* known**	***N* unknown**
$\hat{Y}_{\text{GREG2}}$ as defined by (2.5) with $\mathbf{x}_i = (1, x_{2i})^{\text{T}}$ and $c_i = c$	$\hat{Y}_{\text{SREG1}}$ as defined as special case of (2.11) with $\mathbf{x}_i^* = (x_{2i})$
$\hat{Y}_{\text{OPT2}}$ as defined by (2.9) with $\mathbf{x}_i = (1, x_{2i})^{\text{T}}$	$\hat{Y}_{\text{OPT1}}$ as defined by (2.9) with $\mathbf{x}_i = (x_{2i})$
$\hat{Y}_{\text{OPT3}}$ as defined by (2.9) with $\mathbf{x}_i = (1, \pi_i, x_{2i})^{\text{T}}$	$\hat{Y}_{\text{KREG2}}$ as defined by (3.1) with $\mathbf{z}_i = (\pi_i, x_{2i})^{\text{T}}$ and $c_i = d_i / \sigma^2$
$\hat{Y}_{\text{POPT3}}$ as defined by (3.5) with $\mathbf{z}_i = (1, \pi_i, x_{2i})^{\text{T}}$ and $c_i = d_i - 1$	$\hat{Y}_{\text{KOPT2}}$ as defined as (3.4) with $\mathbf{z}_i = (\pi_i, x_{2i})^{\text{T}}$
	$\hat{Y}_{\text{POPT2}}$ as defined as (3.5) with $\mathbf{z}_i = (\pi_i, x_{2i})^{\text{T}}$ and $c_i = d_i - 1$

The performance of all estimators was evaluated based on the relative bias, the Monte Carlo relative efficiency and the approximate relative efficiency. Expressions of these quantities as shown below.

1. *Relative bias:*

$$\mathrm{RB}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{100}{R}\sum_{i=1}^{R}\frac{\left(\hat{Y}_{\mathrm{EST}(r)} - Y\right)}{Y}, \tag{4.1}$$

where $\hat{Y}_{\mathrm{EST}(r)}$ represents one of the estimators presented in Table 4.1 as computed in the r^{th} Monte Carlo sample.

2. *Monte Carlo Relative efficiency*

$$\mathrm{RE}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{\mathrm{MSE}_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{EST}}\right)}{\mathrm{MSE}_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{GREG2}}\right)}, \tag{4.2}$$

where

$$\mathrm{MSE}_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{Y}_{\mathrm{EST}(r)} - Y\right)^2.$$

The RE measures the relative efficiency of the estimator $\hat{Y}_{\mathrm{EST}}$ with respect to $\hat{Y}_{\mathrm{GREG2}}$.

3. *Approximate Relative efficiency*

$$\mathrm{AR}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{\mathrm{AV}_p\left(\hat{Y}_{\mathrm{EST}}\right)}{\mathrm{AV}_p\left(\hat{Y}_{\mathrm{GREG2}}\right)}, \tag{4.3}$$

where

$$\mathrm{AV}_p\left(\hat{Y}_{\mathrm{EST}}\right) = \sum_{i\in U}\sum_{j\in U}\Delta_{ij}\frac{E_i}{\pi_i}\frac{E_j}{\pi_i},$$

is the approximate variance of $\hat{Y}_{\mathrm{EST}}$ with $E_i = y_i - \mathbf{x}_i^{\mathrm{T}}\mathbf{B}_{\mathrm{EST}}$. The approximate relative efficiency (AR) measures the relative gain in efficiency of $\hat{Y}_{\mathrm{EST}}$ with respect to $\hat{Y}_{\mathrm{GREG2}}$ using the population residual obtained by Taylor linearisation. It is expected that RE and AR give comparable results. However, as we will see, this may not be the case.

4.1 Simulation 1

The population was the dataset (FEV.DAT) available on the CD that accompanies the textbook by Rosner (2006). The data file contains 654 records from a study on Childhood Respiratory Disease carried out in Boston. The variables in the file were: age, height, sex (male female), smoking (indicates whether the individual smokes or not) and Forced expiratory volume (FEV). Singh and Raghunath (2011) used the same data set. The parameter of interest is the total height (y) of the population. The variable age (x_1) was used as auxiliary variable in the regression. The variable FEV (x_2) was chosen as the size variable to compute probabilities of selection for the sampling schemes that are considered in this simulation. The two variables sex and smoking were discarded from the simulation. Table 4.2 summarizes the central tendency measures of the three variables in the population. For each variable, the mean and median were similar. This indicates that the three variables have a symmetrical distribution.

Table 4.2
Descriptive statistics of y, x_1 and x_2

	Min	Q1	Median	Mean	Q3	Max
y	46	57	61.5	61.14	65.5	74
x_1	3	8	10	9.931	12	19
x_2	0.79	1.98	2.55	2.64	3.12	5.79

Figure 4.1 displays the relationship between the variable of interest y and the auxiliary variable x_1. The relationship between Height (y) and the age (x_1) appears to be linear but does not go through the origin. The Pearson correlation coefficient between y and x_1 was 0.79.

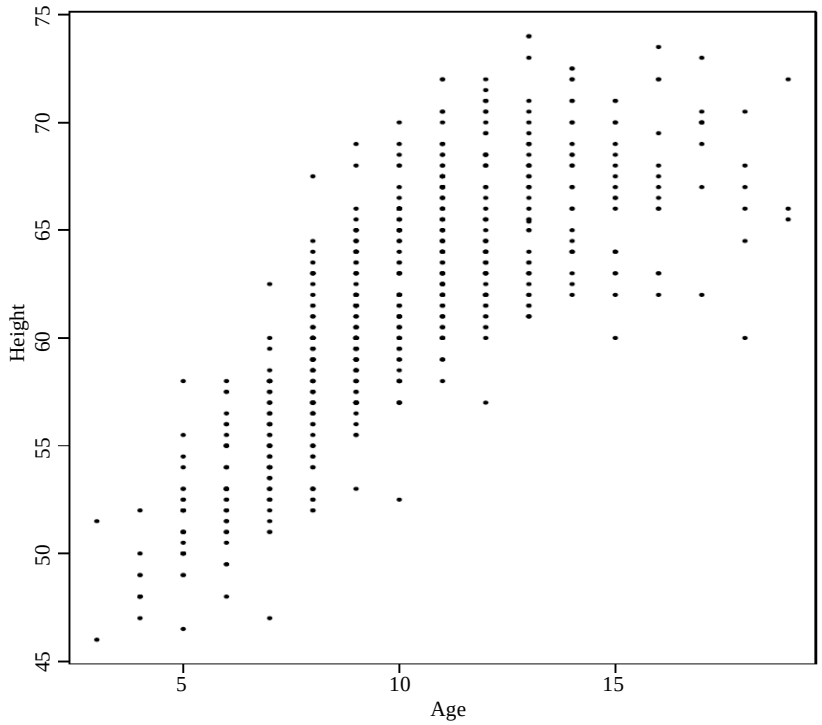

Figure 4.1 Relationship between the variable of interest *Height* and the auxiliary variable *Age*.

The objective of this simulation study was to evaluate the performance of the estimators presented in Table 4.1 using different sampling designs. We considered the Midzuno, the Sampford and the Poisson sampling designs. The variable x_2 were used as a size measure for the three sampling schemes to compute the inclusion probabilities. These sampling designs are as follows:

1. *Midzuno sampling* (see Midzuno 1952): The first unit is sampled with probability p_i and the remaining $n-1$ units are selected as a simple random sampling without replacement from the remaining $N-1$ remaining units in the population. The probabilities of selection p_i for unit i

is given by $p_i = x_{2i} \Big/ \sum_{i \in U} x_{2i}$. The first order inclusion probability for unit i is given by $\pi_i = (N-1)^{-1}\left[(N-n)\, p_i + (n-1)\right]$.

2. *Sampford sampling* (see Sampford 1967): The algorithm for selecting the sample is carried out as follows. The first unit is selected with probability $p_i = x_{2i} \Big/ \sum_{i \in U} x_{2i}$ and the remaining $n-1$ units are selected with replacement with probability $\lambda_i = (1 - np_i)^{-1}\, p_i$. If any of the units are selected more than once, the procedure is repeated until all elements of the sample are different. The probability of inclusion of the first order is given by $\pi_i = np_i$.
3. *Poisson sampling*: Each unit is selected independently, resulting in a random sample size. The probability of selecting unit i is $p_i = x_{2i} \Big/ \sum_{i \in U} x_{2i}$. The inclusion probability associated with unit i is $\pi_i = np_i$. A good description of this procedure can be found in Särndal et al. (1992).

The total of $Y = \sum_{i \in U} y_i$ was the parameter of interest. Based on each of these sampling schemes, we selected $R = 2{,}000$ Monte Carlo samples of size $n = 50$. Estimators in Table 4.1 were then computed for each sample. The performance of the estimators was then assessed using the Relative Bias, the Monte Carlo Relative Efficiency and the Approximate Relative Efficiency as described by the equations (4.1), (4.2) and (4.3) respectively.

4.2 Simulation 1 results

Simulation results are presented in Table 4.3. All estimators studied are approximately unbiased, and their relative bias is smaller than 1%. We discuss separately the approximate relative efficiency (AR) and the relative efficiency (RE) of the estimators when the population size N is known and unknown.

Case 1: *Population size N is known*

We compare the AR and the RE for the following estimators in Table 4.3: $\hat{Y}_{\text{GREG2}}$, $\hat{Y}_{\text{OPT2}}$, $\hat{Y}_{\text{OPT3}}$ and $\hat{Y}_{\text{POPT3}}$ for each of the three sampling designs. We can do so for almost all these estimators except for $\hat{Y}_{\text{OPT3}}$ for the Midzuno and the Sampford sampling schemes. In this case, we cannot compute $\mathbf{B}_{\text{OPT3}}$ for a similar reason as the one described in Remark 3.2.

On the basis of both AR and RE, the pseudo-optimal estimator $\hat{Y}_{\text{OPT3}}$ is the most reliable estimator regardless of the sampling scheme. It is close to the optimal estimator $\hat{Y}_{\text{OPT2}}$ only in terms of AR. Both the RE and the AR of the optimal estimator $\hat{Y}_{\text{OPT2}}$ were not as close as expected under the Midzuno sampling design. The poor behaviour of the RE of the optimal estimator $\hat{Y}_{\text{OPT2}}$ has also been observed by Montanari (1998). Figure 4.2 explains what is happening. We observe that most estimates obtained for the optimal estimator $\hat{Y}_{\text{OPT2}}$ for the 2,000 Monte Carlo samples are close to the mean. However, in some samples, the estimates are quite far from it. This is in contrast to $\hat{Y}_{\text{POPT3}}$ where the values are tightly centered around the mean: note that the associated RE and AR are quite close to one another.

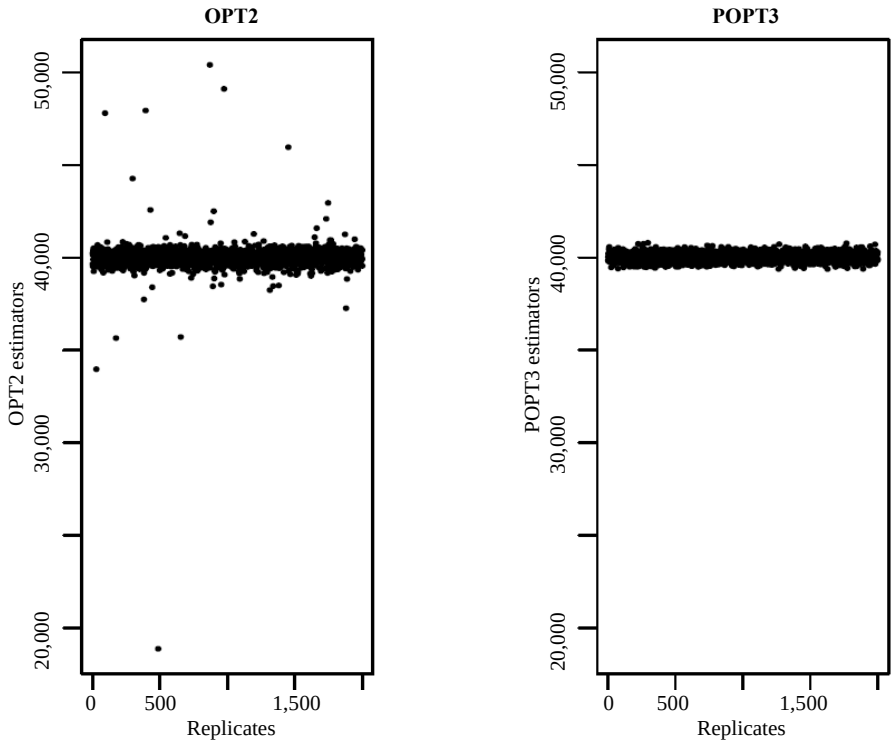

Figure 4.2 Scatter plots of Monte Carlo estimators under the Midzuno Sampling Design.

The optimal estimator $\hat{Y}_{\text{OPT3}}$ is equivalent to the pseudo-optimal estimator $\hat{Y}_{\text{POPT3}}$ in the case of Poisson sampling scheme. Recall that the optimal estimator $\hat{Y}_{\text{OPT2}}$ used $\mathbf{x}_i = (1, x_{2i})^{\text{T}}$ as auxiliary data. The optimal estimator $\hat{Y}_{\text{OPT3}}$ used $\mathbf{x}_i = (1, \pi_i, x_{2i})^{\text{T}}$ as auxiliary data. The addition of the π_i has significantly improved the efficiency of the optimal estimator for the Poisson sampling scheme.

Singh and Raghunath (2011) used $\hat{Y}_{\text{SREG1}}$ when N was known, but did not include it as a control count. Nonetheless, they observed that $\hat{Y}_{\text{SREG1}}$ was quite comparable to $\hat{Y}_{\text{GREG2}}$ in terms of AR and RB for the Midzuno sampling design. The reason for this is that this sampling scheme is quite close to simple random sampling without replacement. However, using these two measures, $\hat{Y}_{\text{SREG1}}$ is by far the worst estimator for the other two sampling schemes.

Case 2: *Population size N is unknown*

Five estimators are reported in Table 4.3 for this case. However, as $\hat{Y}_{\text{KREG2}}$ is quite close to $\hat{Y}_{\text{KOPT2}}$ and $\hat{Y}_{\text{POPT2}}$, we comment on the results obtained for $\hat{Y}_{\text{SREG1}}$, $\hat{Y}_{\text{OPT1}}$ and $\hat{Y}_{\text{KREG2}}$. Estimators $\hat{Y}_{\text{SREG1}}$, $\hat{Y}_{\text{OPT1}}$ and $\hat{Y}_{\text{KREG2}}$ were very similar in terms of relative efficiency and approximate relative efficiency for the Midzuno sampling design. For the Sampford sampling scheme, $\hat{Y}_{\text{OPT1}}$, $\hat{Y}_{\text{KREG2}}$ and $\hat{Y}_{\text{POPT2}}$ were comparable and slightly better than $\hat{Y}_{\text{SREG1}}$. Under the Poisson sampling scheme, $\hat{Y}_{\text{OPT1}}$ and $\hat{Y}_{\text{KREG2}}$ outperformed $\hat{Y}_{\text{SREG1}}$. We can also see that $\hat{Y}_{\text{SREG1}}$ was very inefficient with an RE at least 10 times larger than those associated with $\hat{Y}_{\text{KREG2}}$ or $\hat{Y}_{\text{POPT2}}$. Note that $\hat{Y}_{\text{KREG2}}$ was better than $\hat{Y}_{\text{OPT1}}$: this is reasonable as $\hat{Y}_{\text{KREG2}}$ uses two auxiliary variables whereas $\hat{Y}_{\text{OPT1}}$ uses the single auxiliary variable x_{2i}.

Table 4.3
Comparison of estimators in terms of relative bias and relative efficiencies

		Population size known				Population size unknown				
		$\hat{Y}_{GREG2}$	$\hat{Y}_{OPT2}$	$\hat{Y}_{OPT3}$	$\hat{Y}_{POPT3}$	$\hat{Y}_{SREG1}$	$\hat{Y}_{OPT1}$	$\hat{Y}_{KREG2}$	$\hat{Y}_{KOPT2}$	$\hat{Y}_{POPT2}$
Midzuno	RB (in %)	0.08	0.04		0.07	0.07	0.07	0.07		0.07
	RE	1.00	5.84		0.54	0.94	0.93	0.93		0.93
	AR	1.00	0.55		0.55	0.94	0.93	0.93		0.93
Sampford	RB (in %)	0.11	0.11		0.07	-0.01	0.07	0.02		0.02
	RE	1.00	0.59		0.58	14.72	13.69	13.55		13.56
	AR	1.00	0.55		0.56	15.77	14.39	14.39		14.40
Poisson	RB (in %)	0.11	0.11	0.08	0.08	0.09	0.14	0.16	0.16	0.16
	RE	1.00	0.96	0.57	0.57	160.47	15.49	13.85	13.85	13.85
	AR	1.00	0.96	0.55	0.56	180.36	16.73	14.40	14.39	15.73

Note: We do not provide results for $\hat{Y}_{OPT3}$ and $\hat{Y}_{KOPT2}$ for the Midzuno and Sampford designs because the variance-covariance matrix is not invertible.

4.3 Simulation 2

The performance of the estimators was assessed for different values of the intercept in the model. We restricted ourselves to the Poisson sampling design to illustrate Remark 2.1 in Section 2: that is the efficiency of $\hat{Y}_{SREG}$ deteriorates as the intercept gets bigger. The population was generated according to the following model

$$y_i = a + x_i + e_i. \tag{4.4}$$

The e_i values were generated from a normal distribution with mean 0 and variance $\sigma_i^2 = 1$. The x values were generated according to a chi-square distribution with one degree of freedom. Three populations of size $N = 5{,}000$ were generated using (4.4) with different values of the intercept a. Note that $x-$values were re-generated for each population. The three populations were labelled as A, B and C depending on the intercept used. The intercept values were set to 3, 5 and 10 respectively for populations A, B and C. From each of these populations we drew $R = 2{,}000$ Monte Carlo samples with expected sample size $n = 50$ using the Poisson sampling design. The first inclusion probability was set equal to $\pi_i = nz_i \Big/ \sum_{i \in U} z_i$ for each unit i. The z values were generated according to the following model

$$z_i = 0.5y_i + u_i,$$

where u_i was a random error generated according to an exponential distribution with mean k equals to 0.5 or 1.

4.4 Simulation 2 results

Numerical results are given in Table 4.4 for $k = 1$ and Table 4.5 for $k = 0.5$. All estimators are approximately unbiased with relative biases smaller than 1%.

Case 1: *Population size N is known*

As expected, both optimal estimators $\hat{Y}_{\text{OPT2}}$ and $\hat{Y}_{\text{OPT3}}$ are more efficient than $\hat{Y}_{\text{GREG2}}$. The optimal estimator $\hat{Y}_{\text{OPT2}}$ based on $(1, x_{2i})^{\text{T}}$ is slightly better than $\hat{Y}_{\text{GREG2}}$. The inclusion of the additional variable π_i resulting in $\hat{Y}_{\text{OPT3}}$ yields significant gains in terms of RE and AR: these gains decrease as the intercept gets larger. Once more, $\hat{Y}_{\text{SREG1}}$ is quite inefficient, and as noted in Remark 2.1, this inefficiency increases as the intercept gets larger. The previous observations are valid regardless of k. The efficiency of both optimal estimators $\hat{Y}_{\text{OPT2}}$ and $\hat{Y}_{\text{OPT3}}$ decreases as k gets smaller.

Case 2: *Population size N unknown*

The most efficient estimator is $\hat{Y}_{\text{KREG2}}$. It outperforms $\hat{Y}_{\text{OPT1}}$ as it uses more auxiliary variables. Estimator $\hat{Y}_{\text{SREG1}}$ is by far the most inefficient one. As the intercept in the population model increases, the relative efficiency (both in terms of RE and AR) is fairly stable for $\hat{Y}_{\text{KREG2}}$. On the other hand, the relative efficiencies associated with $\hat{Y}_{\text{SREG1}}$ and $\hat{Y}_{\text{OPT1}}$ deteriorate rapidly, as the intercept in the population model increases. The effect of k on the efficiencies of the estimators is as described when the population size is known.

Table 4.4
Relative bias and relative efficiencies of the estimators for $k = 1$ under Poisson sampling design

Intercept		Population size known				Population size unknown				
		$\hat{Y}_{\text{GREG2}}$	$\hat{Y}_{\text{OPT2}}$	$\hat{Y}_{\text{OPT3}}$	$\hat{Y}_{\text{POPT3}}$	$\hat{Y}_{\text{SREG1}}$	$\hat{Y}_{\text{OPT1}}$	$\hat{Y}_{\text{KREG2}}$	$\hat{Y}_{\text{KOPT2}}$	$\hat{Y}_{\text{POPT2}}$
3	RB (in %)	0.23	0.38	0.56	0.56	0.18	0.77	0.22	0.22	0.22
	RE	1.00	0.95	0.67	0.67	7.72	5.42	0.94	0.94	0.94
	AR	1.00	0.94	0.60	0.98	7.08	5.01	0.85	0.85	0.91
5	RB (in %)	0.04	0.07	0.18	0.18	-0.01	0.67	-0.07	-0.07	-0.07
	RE	1.00	0.99	0.76	0.76	23.91	16.63	1.50	1.50	1.50
	AR	1.00	0.98	0.70	0.73	23.48	16.20	1.45	1.45	1.52
10	RB (in %)	-0.01	-0.02	0.06	0.06	-0.57	0.79	-0.02	-0.02	-0.02
	RE	1.00	1.00	0.80	0.80	88.30	67.47	2.20	2.20	2.20
	AR	1.00	0.99	0.73	0.74	97.92	66.13	2.15	2.15	2.20

Table 4.5
Relative bias and relative efficiencies of the estimators for $k = 0.5$ under Poisson sampling design

Intercept		Population size known				Population size unknown				
		$\hat{Y}_{\text{GREG2}}$	$\hat{Y}_{\text{OPT2}}$	$\hat{Y}_{\text{OPT3}}$	$\hat{Y}_{\text{POPT3}}$	$\hat{Y}_{\text{SREG1}}$	$\hat{Y}_{\text{OPT1}}$	$\hat{Y}_{\text{KREG2}}$	$\hat{Y}_{\text{KOPT2}}$	$\hat{Y}_{\text{POPT2}}$
3	RB (in %)	0.13	0.25	0.42	0.42	-0.18	0.54	-0.02	-0.02	-0.02
	RE	1.00	0.99	0.89	0.89	8.42	5.93	1.78	1.78	1.78
	AR	1.00	0.96	0.83	0.95	8.30	5.83	1.79	1.79	2.10
5	RB (in %)	0.03	0.09	0.22	0.22	0.72	1.49	0.18	0.18	0.18
	RE	1.00	1.00	0.91	0.91	24.35	17.39	3.26	3.26	3.26
	AR	1.00	0.98	0.88	0.94	23.83	16.41	3.15	3.15	3.54
10	RB (in %)	0.06	0.07	0.12	0.12	0.33	1.42	0.13	0.13	0.13
	RE	1.00	1.00	0.96	0.96	98.69	73.93	6.26	6.26	6.26
	AR	1.00	0.99	0.91	0.92	98.65	66.20	5.89	5.89	6.24

5 Conclusions

The regression estimator can be quite efficient if the auxiliary data that it uses are well correlated with the variable of interest. Furthermore, it requires that population totals corresponding to the auxiliary variables are available. In this article, we investigated the behavior of the regression estimator $\left(\hat{Y}_{\text{SREG}}\right)$ proposed by Singh and Raghunath (2011). This estimator uses estimated population count as a control total and the known population totals for the auxiliary variables. We compared it to the Generalized Regression estimator $\left(\hat{Y}_{\text{GREG}}\right)$, its optimal analogue $\left(\hat{Y}_{\text{OPT}}\right)$, and to an alternative estimator $\left(\hat{Y}_{\text{KREG}}\right)$ that uses the first-order inclusion probabilities and auxiliary data for which the population totals are known. As the optimal regression estimator requires the computation of second-order inclusion probabilities, we also included a pseudo-optimal estimator $\left(\hat{Y}_{\text{POPT}}\right)$ that does not require them. We investigated the properties of these estimators in terms of bias and efficiency via a simulation that included various sampling designs, and different values of the intercept in the model for a generated artificial population. We compared the results when the population size was known and unknown.

When the population size is known, the most efficient estimator is the optimal estimator $\hat{Y}_{\text{OPT}}$. However, since this estimator can be unstable, the pseudo-optimal estimator $\hat{Y}_{\text{POPT}}$ is a good alternative to it. This is in line with Rao (1994) who favoured the optimal estimator $\hat{Y}_{\text{POPT}}$ over the Generalized Regression estimator $\hat{Y}_{\text{GREG}}$. The Singh and Raghunath (2011) proposition to use $\hat{Y}_{\text{SREG}}$ is not viable, as it can be quite inefficient. When the population size is not known, the alternative regression estimator $\hat{Y}_{\text{KREG}}$ is the best one to use.

Acknowledgements

The authors kindly acknowledge suggestions for improved readability provided by the Associate Editor and the referees.

References

Cassel, C.M., Särndal, C.-E. and Wretman, J.H. (1976). Some results on generalized difference estimators and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Fuller, W.A. (2009). *Sampling Statistics*. New York: John Wiley & Sons, Inc.

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.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Midzuno, H. (1952). On the sampling system with probability proportional to sum of size. *Annals of the Institute of Statistical Mathematics*, 3, 99-107.

Montanari, G.E. (1987). Post-sampling efficient QR-prediction in large-scale surveys. *International Statistical Review*, 55, 191-202.

Montanari, G.E. (1998). On regression estimation of finite population means. *Survey Methodology*, 24, 1, 69-77.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary data information at the estimation stage. *Journal of Official Statistics*, 10(2), 153-165.

Rosner, B. (2006). *Fundamentals of Biostatistics*. Sixth edition, Duxbury Press.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of section. *Biometrika*, 54, 499-513.

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

Singh, S., and Raghunath, A. (2011). On calibration of design weights. *METRON International Journal of Statistics*, vol. LXIX, 2, 185-205.

Survey Methodology, June 2016
Vol. 42, No. 1, pp. 137-159
Statistics Canada, Catalogue No. 12-001-X

Register-based sampling for household panels

Jan A. van den Brakel[1]

Abstract

In the Netherlands, statistical information about income and wealth is based on two large scale household panels that are completely derived from administrative data. A problem with using households as sampling units in the sample design of panels is the instability of these units over time. Changes in the household composition affect the inclusion probabilities required for design-based and model-assisted inference procedures. Such problems are circumvented in the two aforementioned household panels by sampling persons, who are followed over time. At each period the household members of these sampled persons are included in the sample. This is equivalent to sampling with probabilities proportional to household size where households can be selected more than once but with a maximum equal to the number of household members. In this paper properties of this sample design are described and contrasted with the Generalized Weight Share method for indirect sampling (Lavallée 1995, 2007). Methods are illustrated with an application to the Dutch Regional Income Survey.

Key Words: Probabilities proportional to size; Indirect sampling; Consistent weighting of persons and households; Regional Income Survey; Generalized Weight Share method.

1 Introduction

Statistics Netherlands conducts two important sample surveys to describe the income and wealth situation of the Dutch population. First, the Dutch Regional Income Survey (RIS) provides a description of the income and wealth situation, being accurate at a very detailed regional level. This is accomplished by publishing accurate income distributions for persons and households at a level of neighbourhoods on a yearly basis, using a large sample based on a small set of the main income components derived in a relatively straightforward manner from tax administration. Second, the Income Panel Survey (IPS) publishes yearly income and wealth characteristics of the Dutch population at a more aggregated regional level. This survey is based on a large set of variables using all possible income components of households that can be derived from the available administrative data in the Netherlands. The derivation of the variables for this survey is more time consuming. Therefore the sample size of this survey is considerably smaller than the RIS. Both surveys are designed as a household panel where both person and household based variables about income and wealth are observed.

Households are often considered as the sampling units in panels conducted to collect information at the level of households and persons (Lynn 2009; Smith, Lynn and Elliot 2009). Such panels are used for longitudinal analysis as well as the production of cross-sectional estimates. Using households as the sampling units in a panel design has, however, some major disadvantages due to their instability over time. As time proceeds, households might disintegrate, join or split, new members might enter the households and other members might leave the households for different reasons. Kalton and Brick (1995) explain that these changes can affect the selection probabilities of the households in the sample. Reconstruction of the correct inclusion probabilities of the sampling units is essential to derive correct weights for analysis purposes, in particular if the panel is used for producing cross-sectional estimates.

1. Jan A. van den Brakel, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands and Department of Quantitative Economics, Maastricht University School of Business and Economics, P.O. Box 616, 6200 MD, Maastricht, The Netherlands. E-mail: ja.vandenbrakel@cbs.nl.

Consider a panel where households are selected by means of simple random sampling, say at time $t = 0$. In many panels, people that enter a sampled household at a later stage are also included in the panel. These individuals are called cohabitants by Lavallée (1995). As time proceeds, more and more cohabitants are included in the sample and disturb the equal probability design that is used to select the initial sample (Kalton and Brick 1995). Consider for example household A, which is selected in the sample when the panel started at $t = 0$. If after some period of time this household merges with another household B, which was initially not selected for the panel at time $t = 0$, then the selection probability of this new household is the sum of the selection probabilities of households A and B at time $t = 0$. Not correcting for differences in selection probabilities due to the gradual increasing share of cohabitants in the sample leads to biased inference. Ernst (1989) proposes the Weight Share method to overcome these problems. Lavallée (1995) extends this method to the Generalized Weight Share method as a solution for drawing inference about target populations that are sampled through the use of a frame that refers to a different population.

The RIS and the IPS are both based on a panel and are conducted to collect information about households and persons. To avoid the problems with panels using households as sampling units, an alternative design is applied. Instead of households, so-called core persons are drawn with an equal probability design, who are followed over time. All household members belonging to the household of a core person at each particular period are included in the sample. This results in a sample design where households are drawn proportionally to the household size and households can be selected more than once, but with a maximum that is equal to the household size. This design is an application of indirect sampling (Lavallée 1995, 2007; Deville and Lavallée 2006).

The purpose of this paper is to describe a sample design with an estimation technique that is useful for panels that collect information at person and household level. The methodology employed in this paper is particularly useful for register based sampling, since the core persons are included in the sample indefinitely. The sample design is also useful for Web panels, but might require some form of rotating design to avoid problems with panel attrition. This means that sampling units enter the panel, are observed multiple times and leave the panel according to a pre-specified pattern (Smith et al. 2009). The main contribution of this paper to the existing literature is that explicit expressions for the variance of the target parameters are derived using inclusion expectations instead of inclusion probabilities under the aforementioned sample design. A measure of the minimum accuracy for an estimated income distribution is proposed and explicit expressions for the minimum sample size are derived. The RIS is used throughout the paper to illustrate the described sampling techniques.

The paper is organized as follows. A description of the sample design of the RIS is given in Section 2. In Section 3 the concept of inclusion expectations is introduced as a convenient practical alternative for inclusion probabilities. Subsequently, first and second order inclusion expectations are derived for the proposed sampling design. These inclusion expectations are required to construct the $\pi-$estimator or Horvitz-Thompson (HT) estimator (Narain 1951; Horvitz and Thompson 1952). It is also shown that the same weights can be derived as a special case of the Generalized Weight Share method for indirect sampling (Lavallée 1995, 2007). The key target variables for the RIS are estimated income distributions. In Section 4 formulas for the minimum required sample size are derived based on a precision measure for estimated income distributions. Since households can be selected more than once, an expression for the expected number of unique households is derived in Section 4. The estimation procedure of the RIS is based on linear weighting using the general regression (GREG) estimator (Särndal, Swensson and Wretman 1992) and is

described in Section 5. The integrated weighting method of Lemaître and Dufour (1987), Nieuwenbroek (1993) and Steel and Clark (2007) is applied to obtain equal weights for persons belonging to the same household. In Section 6 variance approximations for the GREG estimator under the proposed sample design are derived. An application to the RIS is provided in Section 7. The paper concludes with a discussion in Section 8.

2 Sampling design

The target population of the RIS is all natural persons residing in the Netherlands. The sample frame is a register containing all natural persons aged 15 years and over residing in the Netherlands as far as they are known to the Tax Office. From this register a stratified simple random sample of so-called core persons is drawn with a sample fraction of 0.16. Neighbourhoods are used as the stratification variable. Although an equal probability design is used, stratified sampling is useful to eliminate the variation between strata and to meet minimum precision requirements for the individual strata. The Netherlands is divided in about 2,830 neighbourhoods with an average size of 5,000 persons aged 15 years and over.

The RIS has been conducted as a panel since 1994. A first requirement for correct cross-sectional inference with this panel is to have correct first and second order inclusion expectations for the sampling units, which are derived in Section 3. A second requirement for correct cross-sectional inference is to keep the panel representative of the target population. To this end, it is determined on a yearly basis which part of the population has entered the target population of the RIS through birth and immigration. From this subpopulation, a stratified simple random sample of core persons with a sample fraction of 0.16 is selected. These core persons are added to the panel of the RIS, with the purpose to maintain a representative sample.

Neighbourhoods are the most detailed level of publication for the RIS and are therefore used as strata. In Section 4 expressions for minimum sample sizes based on precision requirements are derived. Core persons remain in the panel indefinitely. On each survey occasion, all members of the core person's household are also included in the sample. Persons that leave the household of a core person also leave the panel. New persons entering the household of the core person are followed in the panel as long as this person stays in the household of a core person. Information about the household composition of the core persons are obtained from the Municipal Basis Administration (MBA), which is the Dutch government's registry of all residents in the country. Dutch citizens are required by law to report changes in their demographics to their municipalities. The MBA is used in combination with the information from tax administrations to identify household members of the core persons in the sample.

The sample design results in a sample of households where the households are selected with probabilities proportional to the number of persons aged 15 years or older belonging to a household at the current period. Households can be selected more than once, but with a maximum that equals the number of household members aged 15 year or older. In this paper the term core persons is used to refer to the persons that are initially included in the sample and are followed over time in the panel. The term persons is used to refer to the sample obtained if all the household members at a particular period are included in the sample.

The IPS applies a similar sample design with a substantially smaller sampling fraction. The RIS, like the IPS, are register based samples which implies that for each person that is included in the sample, the necessary information for the RIS variables is obtained from the registers of the Tax Office. Core persons

and their household members are therefore not aware that they are included in these samples. This has the advantage that there are no problems with selective non-response and panel attrition. This also makes it possible to include the core persons indefinitely. In the case of a panel where sampling units must complete a questionnaire, some kind of rotating design would be required in order to avoid selectivity bias due to panel attrition. Also, problems with measurement bias associated with data collection where sampling units are asked to complete a questionnaire do not occur. Of course other types of measurement errors are encountered with a survey that is based on registrations (Wallgren and Wallgren 2007). It is assumed that all the required information about income to estimate the target parameters of the RIS and the IPS are available in these registers. Since all the required information is available in a register, a complete enumeration of the population is possible. In the past, however, the IT infrastructure was insufficient to produce timely regional income statistics based on a complete enumeration of the Dutch population. Therefore the RIS was traditionally based on a large sample with a fraction of 0.16 core persons. For the same reason the IPS is traditionally based on a sample of about 80,000 core persons. With the current computational capacity a complete enumeration would still be very demanding but not impossible. The main rationale for conducting this survey as a sample is to maintain the panel for longitudinal analysis that cover time periods from the past where a census was impossible.

3 Inclusion weights

3.1 Weighting with inclusion expectations

For design-based inference, first and second order inclusion probabilities for households and persons are required. Let M denote the number of households in the population, N the number of persons in the population aged 15 years or over and g_k the number of persons aged 15 years or over that belong to the k^{th} household. With the sample design described in Section 2, households k can be included more than once but a maximum of g_k times. This complicates the derivation of inclusion probabilities since the probability of selecting household k is equal to the selection probability of the union of its household members (k, j) aged 15 years and over. This probability is defined as:

$$
\begin{aligned}
P(k \in s) = P\left(\bigcup_{j=1}^{g_k}[(k,j) \in s]\right) &= \sum_{j=1}^{g_k} P((k,j) \in s) \\
&\quad - \sum_{j=1}^{g_k}\sum_{j'=j+1}^{g_k} P([(k,j) \cap (k,j')] \in s) \\
&\quad + \sum_{j=1}^{g_k}\sum_{j'=j+1}^{g_k}\sum_{j''=j+j'+1}^{g_k} P([(k,j) \cap (k,j') \cap (k,j'')] \in s) - \ldots
\end{aligned}
$$

This kind of computation can be avoided by using the concept of inclusion expectations instead of inclusion probabilities. Bethlehem (2009), Chapter 2, generalizes the HT estimator to the concept of inclusion expectation for sampling with replacement. Let a_k denote the number of times that household k is selected in the sample. In the proposed sample design $a_k \in [0,1,\ldots,g_k]$. Let $\mathrm{E}(.)$ denote the expectation with

respect to the sample design. Now $\pi_k = \mathrm{E}(a_k)$ denotes the inclusion expectation of sampling unit k. Since a_k can be larger than one, π_k can also take values larger than one and can therefore no longer be interpreted as an inclusion probability. It can, however, be interpreted as an expectation.

The parameter of interest is the population total, which is defined as

$$t_y = \sum_{k=1}^{M} \sum_{j=1}^{N_k} y_{kj} \equiv \sum_{k=1}^{M} y_k. \tag{3.1}$$

The HT estimator for the population total in (3.1) can be defined as

$$\hat{t}_y = \sum_{k=1}^{M} \frac{a_k y_k}{\pi_k}. \tag{3.2}$$

Since $\mathrm{E}(a_k) = \pi_k$, it follows that this HT estimator is design unbiased. Let $\pi_{kk'}$ denote the inclusion expectation of units k and k', i.e., $\pi_{kk'} = \mathrm{E}(a_k a_{k'})$. The variance of the HT estimator is by definition equal to

$$\begin{aligned}
\mathrm{V}(\hat{t}_y) &= \sum_{k=1}^{M} \sum_{k'=1}^{M} \mathrm{Cov}(a_k a_{k'}) \frac{y_k}{\pi_k} \frac{y_{k'}}{\pi_{k'}} \\
&= \sum_{k=1}^{M} \sum_{k'=1}^{M} [\mathrm{E}(a_k a_{k'}) - \mathrm{E}(a_k)\mathrm{E}(a_{k'})] \frac{y_k}{\pi_k} \frac{y_{k'}}{\pi_{k'}} \\
&= \sum_{k=1}^{M} \sum_{k'=1}^{M} (\pi_{kk'} - \pi_k \pi_{k'}) \frac{y_k}{\pi_k} \frac{y_{k'}}{\pi_{k'}}.
\end{aligned}$$

Note that in the case of sampling without replacement a_k is a dummy taking values zero or one indicating whether unit k is selected in the sample. In this case π_k and $\pi_{kk'}$ are the usual first and second order inclusion probabilities. This illustrates that the standard HT estimator, based on inclusion probabilities, can be extended easily to inclusion expectations. In the case of sample designs where units can be selected more than once, it is more convenient to work with inclusion expectations, since they are derived relatively easily. In the remainder of this subsection, first and second order inclusion expectations for the sample design described in Section 2 are derived.

Core persons are drawn by means of stratified simple random sampling. Since stratification is based on geographical regions, all members of a household k belong to the same stratum h at the moment of drawing core persons. Let N_h denote the number of persons in the population of stratum h aged 15 years or over, n_h the number of core persons selected in the sample from stratum h and g_k the number of persons aged 15 years or over, belonging to household k. Finally, a_{jk} denotes an indicator that is equal to one if person j from household k is selected in the sample and zero otherwise. The first order inclusion expectation of the k^{th} household equals

$$\pi_{kh} = \mathrm{E}(a_k) = E\left(\sum_{j=1}^{g_k} a_{jk}\right) = \sum_{j=1}^{g_k} E(a_{jk}) = g_k \frac{n_h}{N_h}. \tag{3.3}$$

Second order inclusion expectations for households k and k' for $k \neq k'$ belonging to the same stratum h, equal

$$\pi_{kk'} = \mathrm{E}\left(a_k a_{k'}\right) = E\left(\sum_{j=1}^{g_k} a_{jk} \sum_{j'=1}^{g_{k'}} a_{j'k'}\right) = \sum_{j=1}^{g_k}\sum_{j'=1}^{g_{k'}} E\left(a_{jk}a_{j'k'}\right) = g_k g_{k'} \frac{n_h\left(n_h-1\right)}{N_h\left(N_h-1\right)}. \tag{3.4}$$

The second order inclusion expectation for household $k = k'$ from the same stratum h, is given by

$$\begin{aligned} \pi_{kk} &= \mathrm{E}\left(a_k a_k\right) = E\left(\sum_{j=1}^{g_k} a_{jk} \sum_{j'=1}^{g_k} a_{j'k}\right) = E\left(\sum_{j=1}^{g_k} a_{jk} + \sum_{j=1}^{g_k}\sum_{j'\neq j=1}^{g_k} a_{jk}a_{j'k}\right) \\ &= \sum_{j=1}^{g_k} E\left(a_{jk}\right) + \sum_{j=1}^{g_k}\sum_{j'\neq j=1}^{g_k} E\left(a_{jk}a_{j'k}\right) = g_k \frac{n_h}{N_h} + g_k\left(g_k-1\right)\frac{n_h\left(n_h-1\right)}{N_h\left(N_h-1\right)}. \end{aligned} \tag{3.5}$$

Second order inclusion expectations for households k and k' for $k \neq k'$ belonging to two different strata h and h' equal

$$\pi_{kk'} = \mathrm{E}\left(a_k a_{k'}\right) = \mathrm{E}\left(\sum_{j=1}^{g_k} a_{jk} \sum_{j'=1}^{g_{k'}} a_{j'k'}\right) = \sum_{j=1}^{g_k}\sum_{j'=1}^{g_{k'}} E\left(a_{jk}a_{j'k'}\right) = g_{kh} g_{k'h'} \frac{n_h n_{h'}}{N_h N_{h'}}. \tag{3.6}$$

An alternative proof based on the definition of an expected value, which does not use the rule that the expected value of a sum of mutual dependent variables is equal to the sum over the expected values of these variables is given by van den Brakel (2013).

As time proceeds the household composition of the core persons changes, which affects the inclusion expectations of the households in the sample. If sampling fractions differ between strata, the inclusion expectations (3.3) through (3.6) become more complicated and require information of stratum membership for all persons belonging to the household of the core persons. This complication is avoided by choosing a self-weighted sampling design. In this case each household member of a core persons has the same inclusion probability and the only household specific information required to derive household inclusion expectations is the number of persons aged 15 years and over in the household of the core person.

Since all members of a selected household are included in the sample, it follows that the first order inclusion expectations for persons belonging to household k are equal to the first order inclusion expectation of household k defined in (3.3). The second order inclusion expectations for persons from two different households k and k', are equal to (3.4) for two households from the same stratum or (3.6) for two households from two different strata. The second order inclusion expectations for persons from the same household are defined by (3.5).

During the review the question was raised whether the inclusion expectations themselves have a variance that should be taken into account in the variance of HT or GREG estimators when they are based on inclusion expectations instead inclusion probabilities. In the finite population each person and each household has a pre-specified inclusion expectation. For the households observed in the sample these expectations can be calculated exactly without uncertainty since all information required to evaluate the true value of these expectations is available. Substituting inclusion probabilities for expectations, therefore does not result in an additional variance component.

3.2 Generalized Weight Share method

The sample design described in Section 2 can be considered as a special case of indirect sampling (Lavallée 2007). Indirect sampling refers to the situation where the population of interest is sampled through the use of a frame that refers to a different population. Lavallée (1995) develops the Generalized Weight Share method to construct weights for these situations and can be used to derive design weights for households and persons in the sample design described in Section 2.

Following the notation of Lavallée (1995) for the case of indirect sampling, there is a population U^A of size N^A from which a sample s^A of size n is drawn with selection probabilities π_i^A. In addition, there is the target population U^B of size N^B. This population can be divided in M^B clusters. Each cluster k contains N_k^B units, such that $N^B = \sum_{k=1}^{M^B} N_k^B$. The situation for the sample design described in Section 2 is depicted in Figure 3.1. The clusters are households, U^A is the population of persons aged 15 years and over, and U^B is the population of all persons residing in the Netherlands. Persons in U^A and U^B are depicted as circles, households in U^B are depicted as shaded squares, and the circles within a shaded square visualise persons belonging to the same household. Figure 3.1 shows respectively, a single person household, a two person household containing for example a divorced parent with a child younger than 15, a two person household containing two adults without children, and a four person household containing two parents with two children and one of the children is younger than 15 while the other is 15 years or older. The arrows depict the links between the units of U^A and U^B. In the sample design considered in Section 2, each unit in U^A has exactly one unique link with a unit in U^B. Clusters in U^B have at least one link with units in U^A. Links are identified with an indicator variable

$$l_{ij} = \begin{cases} 1 & \text{if there is a link between } i \in U^A \text{ and } \quad j \in U^B \\ 0 & \text{if there is no link between } i \in U^A \text{ and } \quad j \in U^B. \end{cases}$$

If a unit i in U^A is selected in the sample, the entire cluster k to which this unit belongs, is included in the sample. The parameter of interest is the population total in U^B and is similar to (3.1) defined as $t_y = \sum_{k=1}^{M^B} \sum_{j=1}^{N_k^B} y_{kj}$. An estimator for t_y is defined as

$$\hat{t}_y = \sum_{k=1}^{m} \sum_{j=1}^{N_k^B} w_{kj} y_{kj}, \tag{3.7}$$

with m the number of unique clusters (households) included in the sample and w_{kj} the weight attached to each unit j of cluster k. Generally the inverse of the selection probabilities of units (k, j) observed in the sample are used as weights in the HT estimator. In this situation not all units in the sample have a known inclusion probability. Firstly not all units in U^B have a link to U^A. Secondly, as time proceeds household compositions change due to marriages, divorces, departures of children and cohabitation. As a result, as time proceeds, units with a link to U^A enter the clusters in the sample although they are not initially included in the sample drawn from U^A. For these units inclusion probabilities are not necessarily known. They affect, however, the inclusion expectations of the clusters included in the sample. Reconstruction of the inclusion probabilities requires information of selection probabilities of all units in the population at the moment that the sample is drawn. In many practical situations this information is not available.

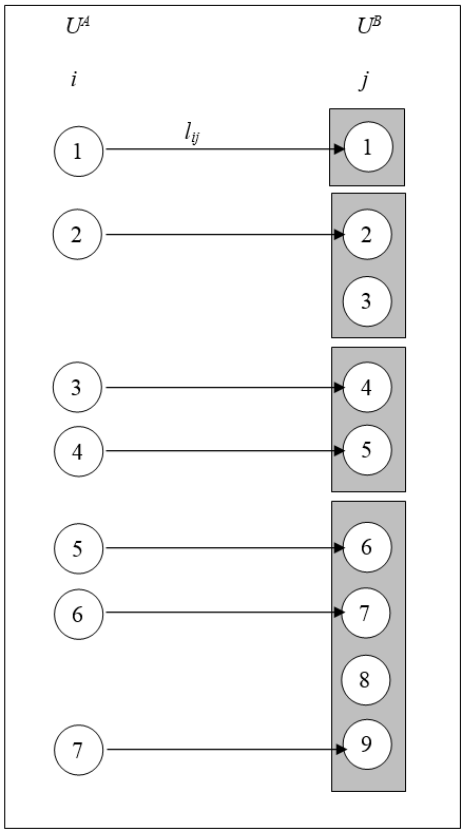

Figure 3.1 Links between units from the sample frame and units from the target population.

The Generalized Weight Share method can be used to derive non-zero weights for all units in the sample. This method starts by deriving initial weights, which are defined as

$$w_{kj}^{*} = \begin{cases} \dfrac{\delta_i^A}{\pi_i^A} & \text{if } (k, j) \text{ has a link with } i \in U^A \\ 0 & \text{otherwise} \end{cases},$$

with δ_i^A an indicator variable that is equal to one if i is included in the sample s^A and zero otherwise. This expression follows directly from Lavallée (1995), equation (2) in combination with the fact that in this application each unit in U^A has exactly one unique link with a unit in U^B, see Figure 3.1. In a second step a so-called basic weight for each cluster k is derived as the mean of all initial weights within each cluster

$$w_k = \frac{\sum_{j=1}^{N_k^B} w_{kj}^{*}}{\sum_{j=1}^{N_k^B} l_{kj}},$$

which follows from Lavallée (1995), equation (7). Finally all persons j that belong to the same household k receive the same weight assigned to their household, i.e., $w_{kj} = w_k$ for all $j \in k$. A proof that the use of the basic weights in (3.7) is an unbiased estimator for the population total is also given by Lavallée (1995).

Let $\sum_{j=1}^{N_k^B} l_{kj} = g_k$ denote the number of persons in household k aged 15 years and older and a_k the number of core persons in household k, i.e., the number of persons in household k that are included in sample s^A. Since s^A is drawn by means of stratified simple random sampling, it follows that $\pi_i^A = n_h^A / N_h^A$ with N_h^A the number of persons aged 15 years and older in the population of stratum h, and n_h^A the number of core persons selected in the sample from stratum h. Then it follows that

$$w_k = \frac{a_k}{g_k} \frac{N_h^A}{n_h^A}. \tag{3.8}$$

Inserting the first order inclusion expectation (3.3) into (3.2) gives the same HT estimator as derived with the Generalized Weight Share method, i.e., inserting (3.8) into (3.7).

The derivation of the inclusion expectations in Subsection 3.1 applies to stratified sampling of households with inclusion expectations proportional to household size and is a special case of the Generalized Weight Share method. An argument to apply a design as outlined in Section 2 is that sampling households proportional to household size is efficient for target variables that are positively correlated with household size.

Lavallée (1995) also provides variance expressions for (3.7) based on the Generalized Weight Share method. This expression is based on the first and second order inclusion probabilities of the sample units drawn from U^A and a transformation of the target variable. As a result the property that clusters are drawn proportional to their size is not made explicit, nor that the fact they are drawn partially with replacement. In Section 6 it is pointed out that the variance expressions in Lavallée (1995) for this application are equal to the variance expressions based on the inclusion expectations derived in (3.3) through (3.6).

4 Sample size determination

The purpose of the RIS is to publish income distributions for households and persons at different geographical levels. Income distributions for households for region or area r are defined as

$$P_{lr} = \frac{M_{lr}}{M_{+r}}, \qquad l = 1, \ldots, L, \tag{4.1}$$

where M_{lr} denotes the number of households from region r, belonging to the l^{th} income category, and $M_{+r} = \sum_l M_{lr}$, the total number of households in area r. This income distribution is estimated as

$$\hat{P}_{lr} = \frac{\hat{M}_{lr}}{M_{+r}}, \qquad l = 1, \ldots, L, \tag{4.2}$$

where $\hat{M}_{lr}$ denotes an appropriate direct estimator for the total number of households from area r, classified to the l^{th} income category. For the moment the HT estimator is assumed as an appropriate estimator for M_{lr}, i.e.,

$$\hat{M}_{lr} = \sum_{h \in r} \sum_{k=1}^{m_h} \frac{y_{khl}}{\pi_k},$$

where $y_{khl} = 1$ if household k from stratum h is classified to the l^{th} income class and $y_{khl} = 0$ otherwise and m_h the total number of households selected in stratum h. In the RIS $L = 10$. Income distributions for persons are defined and estimated similarly to (4.1), (4.2), with M_{lr} the number of persons from area r, belonging to the l^{th} income category. The HT estimator for M_{lr} is now defined as

$$\hat{M}_{lr} = \sum_{h \in r} \sum_{k=1}^{m_h} \frac{1}{\pi_k} \sum_{j=1}^{N_k} y_{kjhl},$$

where $y_{kjhl} = 1$ if person j from household k and stratum h is classified to the l^{th} income class and $y_{kjhl} = 0$ otherwise.

For sample size determination, precision specifications for the estimated income distributions are required. For stratified sampling designs, Neyman allocations are often considered to determine minimum sample sizes and optimal allocations to meet precision requirements at aggregated levels (Cochran 1977). Power allocations are useful to find the right balance between precision requirements for aggregates and strata (Bankier 1988). In this application the minimum sample size is based on precision requirements for the individual strata, i.e., neighbourhoods, which is the most detailed publication level.

If precision requirements are specified for the separate classes of the income distributions, then the income class with the largest population variance determines the minimum required sample size, resulting in unnecessarily large sample sizes. As an alternative the square root of the mean over the variances of the estimated income classes of an income distribution is proposed as a precision measure for the estimated income distributions. With this measure the influence of the most imprecise income class on the minimum sample size will be reduced. The square root of the mean over the variances of the estimated income classes of an income distribution is called the average standard error measure and is defined as

$$s = \sqrt{\frac{1}{L} \sum_{l=1}^{L} V\left(\hat{P}_{lr}\right)}. \tag{4.3}$$

In this section an exact expression for s will be derived as well as an approximation that can be used to estimate the minimum required sample size which does not require information about income distributions or variances.

Since neighbourhoods are the most detailed areas for which income distributions are published, precision requirements for sample size determination are specified at this level. Since neighbourhoods are used as the stratification variable in the sample design, expressions for s can be derived under simple random sampling without replacement of core persons within each neighbourhood. It is proved in the appendix that an expression for the average standard error measure s_h in (4.3) for an income distribution is given by

$$s_h = \sqrt{\frac{1}{L} \frac{N_h - n_h}{n_h} \frac{1}{N_h - 1} \left(\frac{N_h}{M_h^2} \sum_{l=1}^{L} \sum_{k=1}^{M_h} \frac{y_{khl}}{g_{kh}} - \sum_{l=1}^{L} \left(\frac{M_{lh}}{M_h} \right)^2 \right)}, \tag{4.4}$$

with M_h the number of households in stratum h and M_{lh} the number of households in stratum h belonging to the l^{th} income class. Note that if $g_{kh} = 1$ for all households in the population of stratum h, then it follows that $M_h = N_h$ and that formula (4.1) simplifies to

$$V\left(\hat{P}_{lh}\right) = \frac{N_h - n_h}{n_h}\frac{1}{N_h - 1}\left(P_{lh}\left(1 - P_{lh}\right)\right),$$

which can be recognized as the variance of an estimated fraction under simple random sampling without replacement (Cochran 1977, Chapter 3).

Minimum sample size requirements based on (4.4) require information about the income distribution and its variance from preceding periods. Since this information is generally not available at the design phase of a panel, it is useful to have an upper bound for the average standard error measure for the income distribution in (4.4). This is comparable to taking the variance for a parameter defined as a proportion, which reaches a maximum when the proportion is 0.5 for calculating the minimum sample size for a survey. It is shown in the appendix that an upper bound for the average standard error measure s_h for an income distribution, specified in (4.4) is given by

$$s_h \leq \sqrt{\frac{1}{L}\frac{N_h - n_h}{n_h}\frac{1}{N_h - 1}\left(\frac{N_h}{M_h^2}\sum_{t=1}^{T}\frac{M_{th}}{t} - \frac{1}{L}\right)}, \tag{4.5}$$

with M_{th} the number of households of size t in stratum h.

If $g_{kh} = 1$ for all households in the population of stratum h and the number of classes of the income distribution $L = 2,$ then it follows that the approximation for the average standard error measure s_h in (4.5) can be simplified to

$$s_h \leq \sqrt{\frac{N_h - n_h}{n_h}\frac{1}{(N_h - 1)}\frac{1}{4}},$$

which equals the square root of the maximum variance of an estimated fraction at $\hat{P} = 0.5$ under simple random sampling. This illustrates that the approximation for the average standard error measure in (4.5) can be interpreted as a generalization of the approximation of the maximum variance of an estimated fraction at $\hat{P} = 0.5,$ often used in sample size determination. The average standard error measure has its maximum value in the case of an equal distribution of the households over the income categories, i.e., $\hat{P}_{lh} = 1/L$ for $l = 1, \ldots, L.$ In this situation the approximation for s_h is exact, which follows directly from equation (4.3).

Equating the expression for s_h in (4.5) to a pre-specified maximum value, say $\Delta_h,$ results in the following expression for the minimum sample size of core persons

$$n_h \geq \frac{\left(\frac{N_h}{M_h}\right)^2\sum_{t=1}^{T}\frac{M_{th}}{t} - \frac{N_h}{L}}{(N_h - 1)L\Delta_h^2 + \frac{N_h}{M_h^2}\sum_{t=1}^{T}\frac{M_{th}}{t} - \frac{1}{L}}. \tag{4.6}$$

The information required to estimate the minimum sample size is the total number of persons and the total number of equally sized households for neighbourhoods. No information about the expected income distribution or its variance is required. More precise estimates for the minimum sample size can be obtained with the expression in (4.4), but require sample information from, for example, previous periods about the income distributions.

Expression (4.6) gives the minimum sample size for core persons. Subsequently all household members of each core person are included in the sample. As a result, households can be included in the sample more than once and the sample size in terms of unique households and unique persons is random. To plan a survey and control survey costs, it is necessary to know the expected number of unique households and unique persons if a sample of core persons of size n_h is drawn. In the appendix it is proved that the expected number of unique households in a sample of n_h core persons, drawn by means of simple random sampling without replacement from a finite population of size N_h is given by

$$D_h = \sum_{t=1}^{T} M_{th} \left(1 - \frac{\prod_{i=0}^{t-1} (N_h - n_h - i)}{\prod_{i=0}^{t-1} (N_h - i)} \right). \tag{4.7}$$

The expected number of unique persons in a sample of n_h core persons, drawn by means of simple random sampling without replacement from a finite population of size N_h follows directly from (4.7) and is given by

$$D_h^{[p]} = \sum_{t=1}^{T} tM_{th} \left(1 - \frac{\prod_{i=0}^{t-1} (N_h - n_h - i)}{\prod_{i=0}^{t-1} (N_h - i)} \right). \tag{4.8}$$

Since the expected numbers of unique households and persons are random variables, it would be useful to have an uncertainty measure for these expected values. Variance expressions for (4.7) and (4.8) are however not straightforward and therefore left for further research.

Sample size calculations are conducted at the level of neighbourhoods. It was finally decided to select core persons with a sampling fraction of 0.16. With this sample size, the maximum value for the average standard error measure s_h at the level of neighbourhoods amounts to about 0.01 for the estimated household income distributions. With a total population of about 12 million persons, this resulted in a sample size of about 2.1 million core persons and an expected sample size of about 4.6 million unique persons. This sample was drawn in 1994, which was the start of the panel for the Dutch RIS.

5 Linear weighting

For household surveys like the RIS, estimates are required for person characteristics as well as household characteristics. Let t_y denote the total of a target variable y. With linear weighting, an estimator for a person based target variable is defined as

$$\hat{t}_y = \sum_{h=1}^{H} \sum_{k \in 1}^{m_h} \sum_{j \in k} w_{kj} y_{kjh}, \tag{5.1}$$

with y_{kjh} the value of the target variable for person (k, j, h) and w_{kj} a weight for person j belonging to household k. An estimator for a household based target variable is given by

$$\hat{t}_y = \sum_{h=1}^{H} \sum_{k=1}^{m_h} w_k y_{kh}, \tag{5.2}$$

with y_{kh} the value of the target variable for household k from stratum h and w_k a weight for the corresponding household.

Weights are obtained by means of the GREG estimator to use auxiliary variables which are observed in the sample and for which the population totals are known from other sources (Särndal et al. 1992). Consequently, the weights reflect the (unequal) inclusion expectations of the sampling units and an adjustment such that for auxiliary variables the weighted observations sum to the known population totals. Often categorical variables like gender, age, marital status or region are used as auxiliary variables. Due to the fact that the values of auxiliary variables differ from person to person within the same household, different weights can be derived for people from the same household. To ensure that relationships between household variables and person variables are reflected in estimated totals, it is relevant to apply a weighting method which yields one unique household weight for all its household members. If the weights for persons within a household are the same, then household and person based estimates of the same target variables are consistent with each other (for example the total income estimated from households and that from persons). This can be achieved with so-called integrated weighting methods.

Lemaître and Dufour (1987) apply an integrated weighting method at the persons level and replace the original auxiliary variables defined at the person level by the corresponding household mean. In this way, members of the same household have the same inclusion expectation and share the same auxiliary information, and therefore the resulting regression weights are forced to be the same. Nieuwenbroek (1993) proposes a slightly more general approach by applying the linear weighting method at the household level, where the auxiliary information of person based characteristics is aggregated at the household level. Nieuwenbroek (1993) mentions that the linear weighting method at the household level is equal to the linear weighting method of Lemaître and Dufour (1987) at the person level, if the residual variance of the regression model at the household level is chosen proportional to the number of persons within the household. Steel and Clark (2007) and Estevao and Särndal (2006) further generalize the integrated weighting of person and household surveys. Steel and Clark (2007) address the issue of whether the cosmetic benefits of integrated weighting result in an increased design variance of the GREG estimates. They show that large-sample design variances obtained by linear weighting at the household level is less than or equal to the design variance obtained with linear weighting at the person level. For small samples there can be a small increase in the design variance due to integrated weighting. As a result there is little or no loss in efficiency by applying an integrated weighting method.

In this paper the integrated weighting approach at the household level is applied. Let $\mathbf{x}_{kh}$ denote a q-vector containing q auxiliary variables for household k from stratum h. Person based characteristics are aggregated to household totals. The GREG estimator is derived from a linear regression model that specifies the relation between the target variable and the available auxiliary variables for which population totals are known, and is defined as:

$$y_{kh} = \mathbf{x}_{kh}^t \boldsymbol{\beta} + e_{kh}, \quad \text{with} \quad \mathrm{E}_m(e_{kh}) = 0, \quad \mathrm{V}_m(e_{kh}) = \sigma_{kh}^2. \tag{5.3}$$

In (5.3) $\boldsymbol{\beta}$ denotes a vector containing the q regression coefficients of the regression of y_{kh} on $\mathbf{x}_{kh}$ and e_{kh} the residuals and E_m and V_m denote the expectation and variance with respect to the regression model. In this application, the variance structure is taken proportional to the household size, i.e., $\sigma^2_{hk} = g_k\sigma^2$. Nieuwenbroek (1993) shows that in this case the weighting applied at the household level is equal to the method of Lemaître and Dufour (1987).

Regression weights for the households are finally obtained by

$$w_k = \frac{1}{\pi_k}\left(1 + \left(\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}\boldsymbol{\pi}}\right)^t \left(\sum_{k=1}^{m} \frac{\mathbf{x}_{kh}\mathbf{x}_{kh}^t}{\pi_k g_k}\right)^{-1} \frac{\mathbf{x}_{kh}}{g_k}\right),$$

with $\mathbf{t}_{\mathbf{x}}$ a q vector containing the known population totals of the auxiliary variables $\mathbf{x}$, $\hat{\mathbf{t}}_{\mathbf{x}\boldsymbol{\pi}}$ the HT estimator for $\mathbf{t}_{\mathbf{x}}$. The weights calculated at the household level can be used for weighting person based characteristics of the corresponding household members, using formula (5.1) since $w_{kj} = w_k$ for all persons belonging to the same household k.

6 Variance estimation

Parameters of the RIS are estimated as the ratio of two population totals

$$\hat{R} = \frac{\hat{t}_y}{\hat{t}_z}, \tag{6.1}$$

where $\hat{t}_y$ and $\hat{t}_z$ are GREG estimators defined by (5.1) or (5.2) in the case of person-based or household-based target variables, respectively. The variance of (6.1) under a sample design where core persons are drawn by means of stratified simple random sampling, and all household members of these core persons are included in the sample can be approximated by

$$V\left(\hat{R}\right) = \frac{1}{t_z^2}\sum_{h=1}^{H} \frac{N_h^2\left(1 - f_h\right)}{n_h} \frac{1}{N_h - 1}\sum_{k=1}^{N_h}\left(\frac{e_{kh}}{g_k} - \frac{1}{N_h}\sum_{k'=1}^{N_h}\frac{e_{k'h}}{g_{k'}}\right)^2, \tag{6.2}$$

where $f_h = n_h / N_h$, $e_{kh} = \left(y_{kh} - \mathbf{x}_{kh}^t\mathbf{b}_y\right) - R\left(z_{kh} - \mathbf{x}_{kh}^t\mathbf{b}_z\right)$, and $\mathbf{b}_y$ and $\mathbf{b}_z$ are the finite population regression coefficients of the regression of y_{kh} and z_{kh} respectively on $\mathbf{x}_{kh}$. An estimator for the variance specified in (6.2) is given by

$$\hat{V}\left(\hat{R}\right) = \frac{1}{\hat{t}_z^2}\sum_{h=1}^{H}\left(1 - f_h\right)\frac{n_h}{n_h - 1}\sum_{k=1}^{n_h}\left(w_k\hat{e}_k - \frac{1}{n_h}\sum_{k'=1}^{n_h} w_{k'}\hat{e}_{k'h}\right)^2, \tag{6.3}$$

where $\hat{e}_{kh} = \left(y_{kh} - \mathbf{x}_{kh}^t\hat{\mathbf{b}}_y\right) - \hat{R}\left(z_{kh} - \mathbf{x}_{kh}^t\hat{\mathbf{b}}_z\right)$ and $\hat{\mathbf{b}}_y$ and $\hat{\mathbf{b}}_z$ are the HT type estimators for $\mathbf{b}_y$ and $\mathbf{b}_z$. These results follow directly from inserting first and second order inclusion expectations specified in (3.3) through (3.6) in the general approximation for the variance of the ratio of two GREG estimators and its estimator (Särndal et al. 1992, Section 7.13).

The same expressions for the variance can be derived from the variance expressions proposed for the Generalized Weight Share method in the case of indirect sampling. In Lavallée (1995), variance expressions for the HT estimator are based on the sampling design used to select the sample s^A of n units from population U^A with transformed target variables, say z_i. In this application each unit in U^A has exactly one link with a unit in U^B. As a result z_i in Lavallée (1995) is in this case defined as the sum over the target variables of all elements in cluster k, divided by the number of units in cluster k with a link to population U^A, i.e., $z_i = y_k / g_k$ for all $i \in U^A$ that have a link with cluster $k \in U^B$. Inserting the first and second order inclusion probabilities for stratified simple random sampling without replacement and the transformed variables z_i (where the target variable y_k is preplaced by the residual of the regression on the cluster totals e_k) in the variance formula for a ratio gives (6.2). Result (6.3) follows in a similar way.

7 Application

In the RIS, core persons are selected from the population aged 15 years and older through stratified simple random sampling without replacement with a sample fraction of 0.16. In this application results are presented for a large municipality (Rotterdam), a municipality of intermediate size (Enschede) and a small municipality (Sevenum) for three consecutive years 2006, 2007 and 2008. Population and sample sizes for these three municipalities are summarized in Table 7.1.

Table 7.1
Population and sample size RIS for three Dutch municipalities

Municipality	Population		Sample		
	Households	Persons 15 and older	Core persons	Unique households	Unique persons
Rotterdam	293,400	484,000	73,000	67,600	171,400
Enschede	74,200	128,000	19,300	17,600	46,300
Sevenum	2,950	6,100	870	750	2,500

Target variables of interest for the RIS are:

- Income distribution of households in ten classes where the categories are based on ten percentage point quantiles (deciles) of the national distribution using standardized household income (abbreviated as IncDistHh);
- Mean standardized household income (abbreviated as HHinc);
- Mean disposable income of persons that receive income during the 52 weeks of the year (abbreviated as Pinc).

Disposable income of a person is total income of a person minus his or her current taxes. Total income contains earnings, profit, income from capital and savings, and social or other benefits. Standardized household income is defined as the total disposable income of a household corrected for differences in household size and composition. In the literature, this is also known as the equivalised spendable income (OECD 2013).

Estimates for official publications of the RIS are obtained with the GREG estimator using the method of Lemaître and Dufour (1987). Since this survey does not suffer from nonresponse, auxiliary information is used in the estimation for variance reduction and consistency between the marginals of different publication tables. Inclusion expectations are based on the formulas derived in Subsection 3.1. For each municipality the following weighting scheme is applied in the GREG estimator:

$$\text{Age}(7) \times \text{Gender} + \text{Age}(4) \times \text{Gender} \times \text{MaritalStatus}(2) + \text{Address}(2) \times \text{HHsize}(5).$$

All auxiliary variables are categorical. The numbers between brackets denote the number of categories. MaritalStatus distinguishes between people who are married and other forms of marital status. Address distinguishes between addresses where one family is residing and other types of addresses. HHsize stands for household size and distinguishes between households with one, two, three, four, and five or more persons. Estimates for HHinc and Pinc with their standard errors based on the HT estimator, the GREG estimator and the GREG estimator with the method of Lemaître and Dufour (1987) are given in Table 7.2. In Figure 7.1 the income distributions IncDistHh estimated with the HT estimator, GREG estimator and the GREG estimator with the method of Lemaître and Dufour (1987) are plotted with a 95% confidence interval for Rotterdam and Sevenum in 2008. The standard errors for these estimates are compared in a separate histogram. In Figure 7.2 the IncDistHh for Rotterdam and Sevenum estimated with the method of Lemaître and Dufour (1987) are given for 2006, 2007 and 2008. See van den Brakel (2013) for more detailed output of the income distributions.

Table 7.2
Estimation results RIS for Rotterdam (large city), Enschede (intermediate city), and Sevenum (small village), standard errors in brackets

	Variable	Year	HT		GREG		GREG consistent (L&D)	
Rotterdam	HHinc	2006	19,790	(83)	20,134	(80)	20,161	(76)
		2007	22,306	(73)	22,950	(64)	22,866	(64)
		2008	23,750	(78)	24,511	(69)	24,410	(68)
	Pinc	2006	22,074	(94)	22,219	(84)	22,233	(93)
		2007	24,094	(82)	24,362	(75)	24,432	(78)
		2008	25,325	(84)	25,625	(75)	25,705	(78)
Enschede	HHinc	2006	19,810	(128)	20,353	(111)	20,300	(107)
		2007	20,878	(128)	21,716	(107)	21,753	(105)
		2008	22,254	(148)	23,235	(125)	23,237	(123)
	Pinc	2006	20,402	(102)	20,608	(92)	20,590	(92)
		2007	21,387	(115)	21,751	(103)	21,852	(106)
		2008	22,235	(123)	22,659	(110)	22,724	(114)
Sevenum	HHinc	2006	25,696	(799)	25,698	(734)	25,968	(711)
		2007	28,207	(618)	28,901	(520)	29,026	(490)
		2008	31,466	(795)	32,372	(715)	32,536	(694)
	Pinc	2006	21,328	(466)	21,680	(428)	21,712	(428)
		2007	24,056	(456)	24,219	(396)	24,459	(393)
		2008	24,980	(468)	25,482	(426)	25,644	(455)

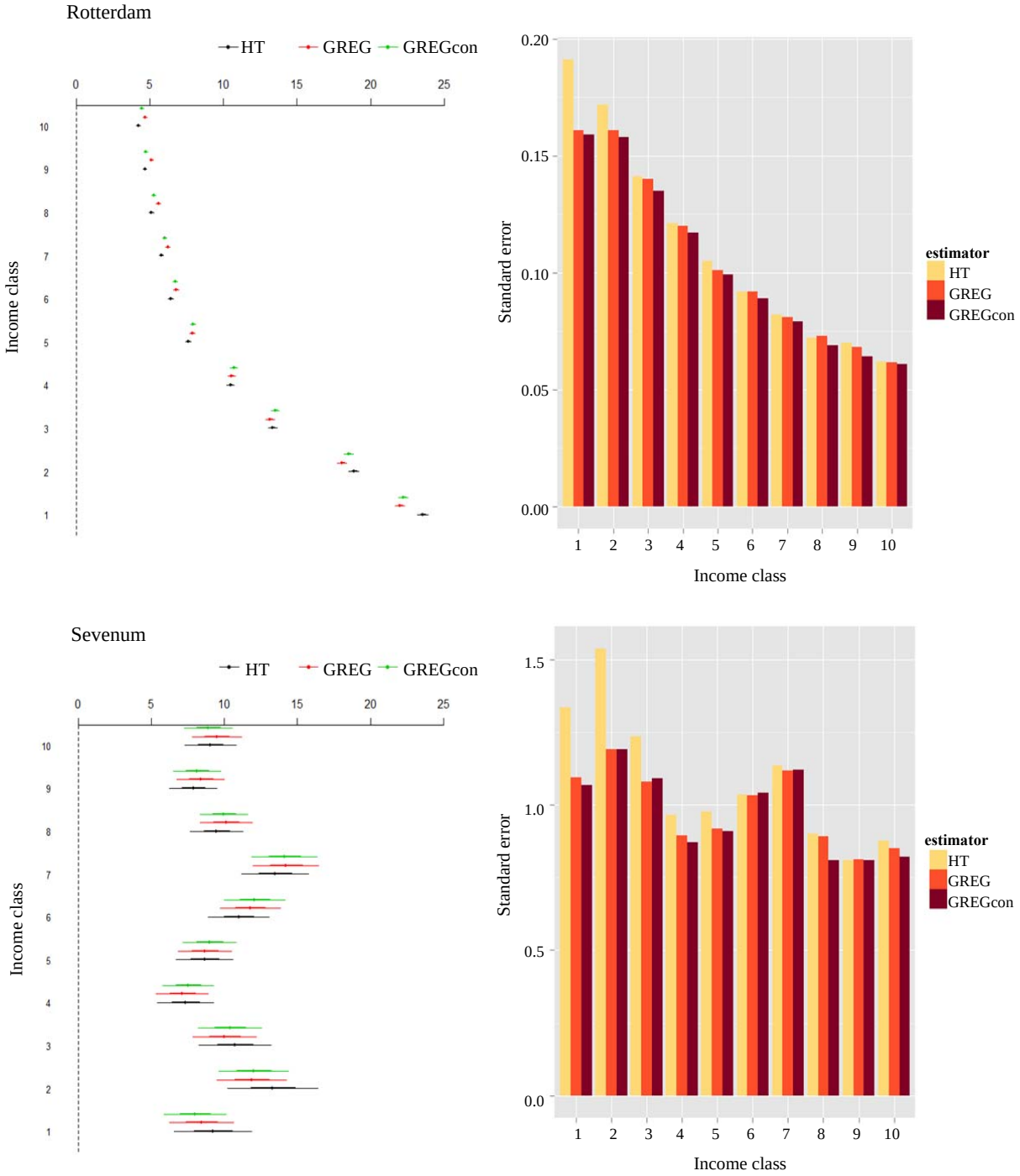

Figure 7.1 IncDistHh in percentages for Rotterdam and Sevenum (left panels) with Horvitz-Thompson estimator, GREG estimator and integrated GREG estimator (GREGcon), with 95% confidence intervals. Standard errors of the corresponding estimators are plotted in the right panels.

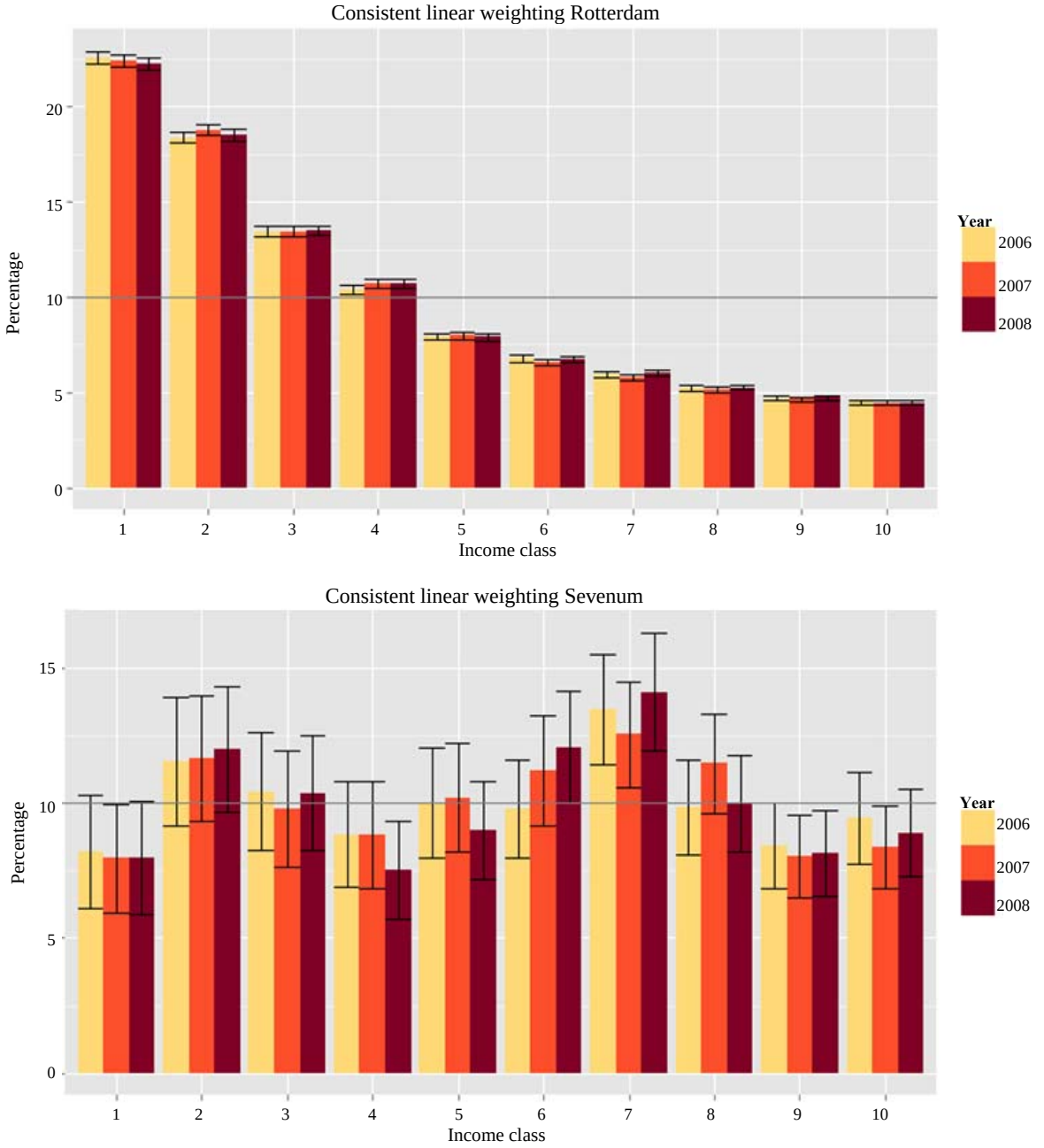

Figure 7.2 IncDistHh in percentages for Rotterdam (upper panel) and Sevenum (lower panel) estimated with integrated weighting for 2006, 2007 and 2008 with 95% confidence intervals. Grey line refers to the national income distribution.

The observed income distributions in Figures 7.1 and 7.2 are a result of the demographic compositions in both municipalities. Rotterdam is a city where the fraction of households in low income categories are above the national average, since the fractions in the first three categories are above 10%. The fraction of households in higher income categories, on the other hand, are below the national average, since these fractions are below 10%. This is a typical distribution for a large university city with a high fraction of

non-western immigrants. Sevenum on the other hand is a small village close to a large industrial city. Such villages typically have small fractions of immigrants, no students and large fractions of households with one or two people that receive income during 52 weeks of the year. This explains why the fraction of households in the lowest income category is below the national average and the fraction of households in the higher income categories (6, 7 and 8) is above the national average. Sevenum is a village that does not attract extreme rich households.

Since HHinc and Pinc are based on different income definitions and since Pinc is the average over the domains of people that receive income during 52 weeks of the year, the differences between the two means vary between municipalities. For a large university city like Rotterdam, the mean standardized household income is typically smaller compared to the mean of disposable personal income averaged over people that receive income during 52 weeks of the year. Other cities with large universities show a similar picture. In a small but rich village like Sevenum, the situation is the other way around.

Another remarkable result is that in Rotterdam and Enschede the difference between the HT estimator and the GREG estimator is relatively large compared to the standard errors. Given the large sample size and the fact that there is no nonresponse, these differences are expected to be smaller. A possible explanation is that Rotterdam and Enschede are large university cities. Students are often identified in the tax register (used as the sample frame) in a different way than they appear in the population register (used to derive population distributions of the auxiliary variables), in particular with respect to their household situation.

For each municipality there is a steady increase over time in the mean of the income for households and persons. Also the income distributions for each municipality show a stable pattern over the years. This can be expected if a panel is applied in combination with large sample sizes to estimate phenomena that are not very volatile in time.

Comparing GREG estimates with and without using the method of Lemaître and Dufour (1987) shows that standard errors of estimated household parameters are smaller if the method of Lemaître and Dufour (1987) is applied. This is particularly visible for the mean household income in the small sample of Sevenum. For estimated person based parameters, on the other hand, the method of Lemaître and Dufour (1987) slightly increases the standard error compared to the regular GREG estimator. This suggests that the assumed variance structure for the residuals in the underlying regression model in the case of integrated weighting better fits the household-based variables than the person-based variables.

8 Discussion

Households, due to their instability over time, are inappropriate as sampling units in panels conducted to collect information at the level of households or persons. In this paper, a sample design is proposed where persons are drawn through a self-weighted sample design. At each point in time, the household members of these so-called core persons are included in the sample. This results in a sample where households can be drawn more than once but with a maximum that is equal to the household size. Households are included with expectations proportional to the household size. First and second order inclusion expectations for households are derived under an equal probability sample design for selecting core persons. These inclusion expectations can be used in a similar way to the more common inclusion probabilities in design-based and model-assisted inference.

The sample design in this paper is a special case of indirect sampling (Lavallée 1995, 2007). In the case of a self-weighted sample design it is shown that first and second order inclusion expectations for this sample design can be derived in a relatively straightforward manner from the household composition of the core persons at each point in time. In the case of more complex sample designs, the Generalized Weight Share method (Lavallée 1995, 2007), is required to construct inclusion weights at each point in time.

The advantage of the proposed sample design is that the estimation procedure is simpler than the Generalized Weight Share method. The design is particularly useful if core persons are selected with a self-weighted sampling design. If, due to, e.g., minimum precision and maximum cost requirements, an unequal probability design for the selection of core persons is required, then the Generalized Weight Share method is required. Since core persons remain in the panel indefinitely, this sample design is particularly appropriate for register-based household panels where all the required information is derived from administrative data. For interview-based household panels some kind of rotating design is required to cope with problems like panel attrition.

In the paper the so called average standard error measure, defined as the square root of the mean over the variances of the estimated income classes of an income distribution, is proposed as a precision measure for minimum sample size determination. It is shown that the maximum value of this precision measure corresponds with a distribution where the proportions in the categories are equal. It is also shown that this result can be seen as generalization of the variance of a fraction taking its maximum value at 0.5. An expression for the minimum required sample size to meet a pre-specified precision for estimated distributions is derived. Since households can be included more than once in the sample, an expression for the expected number of unique households in a sample is also derived.

A topic for further research is to combine this mean standard error measure with a Neyman allocation or power allocations to have expressions for the minimum sample size based on precision requirements for estimated distributions at aggregates of strata. This results in an unequal inclusion probability design for the core persons and requires the Generalized Weight Share method for deriving appropriate weights.

In the context of household surveys and panels, weighting procedures that enforce equal regression weights for persons within the same household are relevant in order to enforce consistency between person based and household based estimates. In this paper an integrated weighting approach based on Lemaître and Dufour (1987) is applied to the RIS. In this application standard errors obtained with Lemaître and Dufour (1987) are smaller than a non-integrated weighting procedure for household based estimates. For person based estimates, standard errors can be slightly larger. These results are in line with Steel and Clark (2007), who show that the large-sample design variance of integrated weighting at the household level is smaller than or equal to the design variance obtained with non-integrated weighting at the person level. In their simulation they also report small increases of the design-variances due to integrated weighting in the case of small sample sizes.

Integrated weighting of Lemaître and Dufour (1987) at the household level is obtained by assuming a variance structure for the residuals that is proportional to the household size (Nieuwenbroek 1993). If household characteristics are proportional to household size, then it can be anticipated that such a variance structure better explains the variation of the household variables in the population compared to a variance structure that assumes equal residual variance for the households. For person based variables such a variance

structure might be less efficient but the additional advantage of integrated weighting is that totals for household and person based income, which can be derived directly from their means, are consistent.

Acknowledgements

The views expressed in this paper are those of the author and do not reflect the policies of Statistics Netherlands. The author is grateful to the Associate Editor and the unknown referees for giving constructive comments on two former drafts of the paper. The author also thanks Drs. M. van den Brakel-Hofmans for making the RIS data available.

Proof of equation (4.4)

An expression for the variance of the estimated fraction of households in income class l can be derived from the general expression for the variance of the HT estimator (Särndal et al. 1992, Section 2.8):

$$V\left(\hat{P}_{lh}\right) = \frac{1}{M_h^2}\sum_{k=1}^{M_h}\sum_{k'=1}^{M_h}\left(\pi_{kk'h} - \pi_{kh}\pi_{k'h}\right)\frac{y_{khl}}{\pi_{kh}}\frac{y_{k'hl}}{\pi_{k'h}}. \tag{A.1}$$

Inserting first and second order inclusion expectations specified in (3.3) through (3.6), and taking advantage of the property that $y_{khl} = y_{khl}^2$ since the values of the target variable are restricted to zero or one, it follows after some algebra that (A.1) can be simplified to

$$V\left(\hat{P}_{lh}\right) = \frac{N_h - n_h}{n_h}\frac{1}{N_h - 1}\left(\frac{N_h}{M_h^2}\sum_{k=1}^{M_h}\frac{y_{khl}}{g_{kh}} - \left(\frac{M_{lh}}{M_h}\right)^2\right). \tag{A.2}$$

Result (4.4) is obtained by inserting (A.2) into (4.3).

The *population* of households in stratum h can be divided into T subpopulations of equally sized households. Let M_{th} denote the number of households of size t in stratum h. Now it follows for the double summation between brackets for the expression of s in (4.4) that

$$\sum_{l=1}^{L}\sum_{k=1}^{M_h}\frac{y_{khl}}{g_{kh}} = \sum_{l=1}^{L}\sum_{t=1}^{T}\sum_{k=1}^{M_{th}}\frac{y_{khl}}{t} = \sum_{t=1}^{T}\frac{M_{th}}{t}. \tag{A.3}$$

According to the Cauchy-Schwartz inequality (Cochran 1977, Section 5.5) it follows for the single summation between brackets for the expression of s_h in (4.4) that

$$\sum_{l=1}^{L}\left(\frac{M_{lh}}{M_h}\right)^2 = \sum_{l=1}^{L}P_{lh}^2 \geq \frac{1}{L}. \tag{A.4}$$

Result (4.5) is obtained by inserting (A.3) and (A.4) in the expression for s in (4.4).

Let $\tilde{\pi}_{tkh}$ denote the inclusion probability for household k from stratum h of size t. Since equally sized households share the same first order probabilities, it follows that $\tilde{\pi}_{tkh} = \tilde{\pi}_{tk'h} \equiv \tilde{\pi}_{th}$. Let I_{tkh} denote an

indicator variable, taking value 1 if household k from stratum h of size t is included in the sample and zero otherwise. The expected number of unique households can be derived as

$$D_h = \mathrm{E}\left(\sum_{t=1}^{T}\sum_{k=1}^{M_{th}} I_{tkh}\right) = \sum_{t=1}^{T} M_{th}\tilde{\pi}_{th}$$

$$= \sum_{t=1}^{T} M_{th}\left(1 - \frac{\binom{N_h - t}{n_h}}{\binom{N_h}{n_h}}\right) = \sum_{t=1}^{T} M_{th}\left(1 - \frac{(N_h - n_h)(N_h - n_h - 1)\ldots(N_h - n_h - t + 1)}{N_h(N_h - 1)\ldots(N_h - t + 1)}\right).$$

References

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.

Bethlehem, J.G. (2009). *Applied Survey Methods*, New Jersey: John Wiley & Sons, Inc.

Cochran, W.G. (1977). *Sampling Techniques*, New York: John Wiley & Sons, Inc.

Deville, J.-C., and Lavallée, P. (2006). Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology*, 32, 2, 165-176.

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley & Sons, Inc., 135-159.

Estevao, V.M., and Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74, 127-147.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Ass*ociation, 47, 663-685.

Kalton, G., and Brick, J.M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21, 1, 33-44.

Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 1, 25-32.

Lavallée, P. (2007). *Indirect Sampling*, New York: Springer Verlag.

Lemaître, G., and Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 2, 199-207.

Lynn, P. (2009). Methods for longitudinal surveys. In *Methodology of Longitudinal Surveys*, (Ed., P. Lynn), Wiley, Chichester, 1-19.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.

Nieuwenbroek, N.J. (1993). An integrated method for weighting characteristics of persons and households using the linear regression estimator. Research paper, BPA nr: 8555-93-M1-1, Statistics Netherlands, Heerlen.

OECD (2013). *OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth*. OECD publishing, http://dx.doi.org/10.1787/9789264194830-en.

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

Smith, P., Lynn, P. and Elliot, D. (2009). Sample design for longitudinal surveys. In *Methodology of Longitudinal Surveys*, (Ed., P. Lynn), Wiley, Chichester, 21-33.

Steel, D.G., and Clark, R.G. (2007). Person-level and household-level regression estimation in household surveys. *Survey Methodology*, 33, 1, 51-60.

van den Brakel, J.A. (2013). Sampling and estimation techniques for household panels. Discussion paper 2013-15, Statistics Netherlands, Heerlen. http://www.cbs.nl/NR/rdonlyres/B4F85FB9-52F2-4B8A-94C4-56DA43F2250D/0/201315x10pub.pdf.

Wallgren, A., and Wallgren, B. (2007). *Register-Based Statistics: Administrative Data for Statistical Purposes*. New York: John Wiley & Sons, Inc.

Survey Methodology, June 2016
Vol. 42, No. 1, pp. 161-177
Statistics Canada, Catalogue No. 12-001-X

Nonresponse adjustments with misspecified models in stratified designs

Ismael Flores Cervantes and J. Michael Brick[1]

Abstract

Adjusting the base weights using weighting classes is a standard approach for dealing with unit nonresponse. A common approach is to create nonresponse adjustments that are weighted by the inverse of the assumed response propensity of respondents within weighting classes under a quasi-randomization approach. Little and Vartivarian (2003) questioned the value of weighting the adjustment factor. In practice the models assumed are misspecified, so it is critical to understand the impact of weighting might have in this case. This paper describes the effects on nonresponse adjusted estimates of means and totals for population and domains computed using the weighted and unweighted inverse of the response propensities in stratified simple random sample designs. The performance of these estimators under different conditions such as different sample allocation, response mechanism, and population structure is evaluated. The findings show that for the scenarios considered the weighted adjustment has substantial advantages for estimating totals and using an unweighted adjustment may lead to serious biases except in very limited cases. Furthermore, unlike the unweighted estimates, the weighted estimates are not sensitive to how the sample is allocated.

Key Words: Nonresponse; Stratification; Sampling weights; Weighting classes reweighting.

1 Introduction

Adjusting the base weights for unit nonresponse using weighting classes is a standard approach to survey weighting, but the adjustments are not done in the same way by all researchers or survey organizations. Little and Vartivarian (2003), hereafter referred to as L&V, observed that using a nonresponse adjustment factor that is weighted by the inverse of the probability of selection appears to be the most common approach. They also pointed out that using design weights to compute a weighted nonresponse adjustment does not eliminate nonresponse bias in estimates of the mean of the population when the response mechanism is not specified correctly by the weighting adjustment model. L&V then conducted a simulation study using a simple stratified sample design to examine the effect of weighting the nonresponse adjustment factors. They concluded that weighting the nonresponse adjustment has little or no value.

Theoretical justifications for nonresponse adjustment require that either the response mechanism or the target variable must be modeled correctly to eliminate nonresponse bias; we are not aware of any theory that suggests that weighting by the inverse of the probability of selection completely eliminates bias when the model is misspecified (e.g., Kalton 1983; Little 1986; Little and Rubin 2002; Särndal and Lundström 2005). In this regard, the importance of modeling for nonresponse adjustment urged by L&V is essential for good statistical practice. However, correctly specifying a highly predictive model is an ideal that cannot be achieved in most surveys because of the complexity of the phenomenon and because powerful auxiliary variables rarely exist. The search for better auxiliary data for this modeling has fueled research into paradata, but the models using these data still have relatively poor correlations with response propensities (Kreuter,

1. Ismael Flores Cervantes and J. Michael Brick, Westat, 1600 Research Blvd, Rockville, Maryland 20850. E-mail: ismaelflorescervantes@westat.com.

Olson, Wagner, Yan, Ezzati-Rice, Casas-Cordero, Lemay, Peytchev, Groves and Raghunathan 2010). In practice, imperfect models are used and nonresponse bias is never completely eliminated.

Consequently, understanding the effects of nonresponse adjustment methods and whether there is any value to weighting the nonresponse adjustment with an incorrectly specified response model is important. Even though a message of L&V was the need to include design variables in the nonresponse modeling, some researchers appear to have concluded that weighting the adjustment has no role (e.g., Chadborn, Baster, Delpech, Sabin, Sinka, Rice and Evans 2005; Haukoos and Newgard 2007). However, L&V's conclusion that weighting the nonresponse adjustment factor is either incorrect or inefficient was based on comparisons to correctly specified models that always produce unbiased estimates. Their suggestion to condition on the design variables (in their setting the design variable was the stratum) resulted in identical weighted and unweighted estimators. Their simulations are also centered on a specific stratified sample design and they only consider estimating means. As discussed below, these are substantial limitations and the conclusions that some have drawn that weighting the adjustment is inappropriate need to be reconsidered.

Following L&V, researchers have examined the effects of weighting in other cases. Sukasih, Jang, Vartivarian, Cohen and Zhang (2009) compared nonresponse adjustments with and without weights by simulation within the context of a specific survey. West (2009) used simulation to study estimates of population means under more complex sample designs that featured clustering and differential sampling rates. Both of these studies concluded that weighting the nonresponse adjustments by the design weights was beneficial compared to using an unweighted approach, even though the differences due to weighting were not large. Kott (2012) assessed the robustness of the adjustments theoretically and described the conditions under which the various estimators for population means had greater protection against nonresponse bias; he recommended a weighted approach. Related research has been conducted on the need for weighting for estimating response propensity model coefficients (Wun, Ezzati-Rice, Diaz-Tena and Greenblatt 2007; Grau, Potter, Williams and Diaz-Tena 2006), but this line of research is sufficiently different that we do not discuss it here.

In this article, we explore the effect of weighting nonresponse adjustments when the nonresponse model is imperfect. In Section 2, we expand on the L&V results by looking at estimators for totals and domain means and totals; L&V only considered overall means. Using the same population and basic simulation setting of L&V, we also explore the effect of different sample allocation to the strata while L&V used one sample allocation. The results of the simulations presented in Section 3 show that there are important differences in the properties of the weighted and unweighted estimators and these vary by how the sample is allocated. We explain the behaviors of the estimators using simple approximations to show why they differ. Although weighting the adjustment factor does not always give estimates with lower bias and root mean square error when compared to estimates from the unweighted alternative, it has substantial benefits for estimates of totals and provides protection against large errors that may arise with an unweighted approach. As a result, we recommend a weighted approach when the true response mechanism is not fully known. Conclusions are presented in Section 4.

2 Setting

Survey weights compensate for different types of missing data – sampling or base weights adjust for those that are not sampled, noncoverage adjustment weights account for those that are not in the sampling frame, and nonresponse adjustment weights compensate for those that are sampled but do not respond. We focus on nonresponse adjustment weights and the effect of using the base weights in creating the nonresponse adjustments.

We begin with the unadjusted Horvitz-Thompson estimator of the total

$$\hat{y}_{un} = \sum_s R_i d_i y_i, \tag{2.1}$$

where d_i is the inverse of the probability of selection of unit $i, R_i = 1$ if unit i responds and $= 0$ otherwise, and the sum is over the units in sample s. The ratio mean is $\hat{\bar{y}}_{un} = \hat{y}_{un} \big/ \sum_s R_i d_i$. If all the sample data are observed and the frame is complete, then $E(\hat{y}_{un}) = Y$, and the ratio mean is consistent for $\bar{Y}$.

When there is unit nonresponse, we assume that response is a random variable and the probability or propensity of response $(\phi_i = \Pr(R_i = 1))$ is like the probability from an additional phase of sampling (Särndal, Swensson and Wretman 1992). If we assume $\phi_i > 0$ for all i, then the nonresponse bias of an estimated ratio mean under the stochastic model is

$$\text{bias}\left(\hat{\bar{y}}_{un}\right) \approx \bar{\phi}^{-1}\sigma_\phi \sigma_y \rho_{\phi,y}, \tag{2.2}$$

where $\bar{\phi}$ is the population mean of the response propensities, σ_ϕ is the standard deviation of ϕ, σ_y is the standard deviation of y, $\rho_{\phi,y}$ is the correlation between ϕ and y (Bethlehem 1988). The estimated respondent mean is unbiased if ϕ and y are uncorrelated. Brick and Jones (2008) extend these results to other types of statistics and estimators.

To reduce nonresponse bias, auxiliary variables associated with the sample can be used to support nonresponse adjustments to the base weights. The adjustments can be implemented by modeling either the distribution of ϕ or y, or both using the auxiliaries. We are specifically interested in modeling the response mechanism.

The estimated response propensities are applied as if they were the actual probabilities of responding. In other words, the nonresponse adjustment factor is the inverse of the estimated propensity of responding for sampled unit $i\left(\hat{\phi}_i\right)$. The response propensity can be estimated by a variety of methods such as logistic regression, but most surveys form mutually exclusive groups called weighting classes or response homogeneity groups which contain units with similar estimated propensities and adjust the weights in each group or class by a common factor, say $\hat{f}_c = \hat{\phi}_c^{-1}$ for all $i \in c$ (Särndal et al. 1992, and Little 1986). When this approach is used, the adjusted estimator is called a weighting class estimator and is

$$\hat{y}_{wc} = \sum_c \sum_{i \in s_c} R_{ci} d_{ci} \hat{f}_c y_{ci}, \tag{2.3}$$

where $c = 1, 2, \ldots, C$ are the nonresponse adjustment classes and $i \in s_c$ is a sampled unit in class c.

The specific issue we address is the effect of weighting the adjustment factor. The unweighted factor is

$$\hat{f}_c^u = \frac{\sum_{i \in s_c} \delta_{ci}}{\sum_{i \in s_c} R_{ci}\delta_{ci}} = \frac{n_{c+}}{r_{c+}}$$

where $\delta_{ci} = 1$ if $i \in c$ and $\delta_{ci} = 0$ if $i \notin c$, and n_{c+} and r_{c+} are the number of sampled and responding units in class c. The weighted adjustment factor is

$$\hat{f}_c^w = \frac{\sum_{i \in s_c} d_{ci}}{\sum_{i \in s_c} R_{ci}d_{ci}} = \frac{\hat{N}_c}{\hat{N}'_c},$$

where $\hat{N}_c = \sum_{i \in s_c} d_{ci}$ and $\hat{N}'_c = \sum_{i \in s_c} R_{ci}d_{ci}$. The factors correspond to the unweighted and weighted response rates, respectively. Substituting the factors into the estimator (2.3) yields two alternative estimators (2.4) and (2.5) of the total population. These are both weighting class estimators but we have changed notation to emphasize whether the weighted or unweighted response rate is used.

$$\hat{y}_{urr} = \sum_c \hat{f}_c^u \sum_{i \in r_c} d_{ci}y_{ci} = \sum_c \frac{n_{c+}}{r_{c+}} \sum_{i \in r_c} d_{ci}y_{ci}, \tag{2.4}$$

$$\hat{y}_{wrr} = \sum_c \hat{f}_c^w \sum_{i \in r_c} d_{ci}y_{ci} = \sum_c \frac{\hat{N}_c}{\hat{N}'_c} \sum_{i \in r_c} d_{ci}y_{ci}. \tag{2.5}$$

These two estimators are the building blocks for all the types of statistics that we consider in the simulation study. For example, estimators of means, domain means, and ratios are simple functions of estimators (2.4) and (2.5).

To be consistent with the structure, notation, and simulations in L&V, we restrict our study to the same population with a stratified simple random sample where two strata are defined by the binary design variable, Z, and two nonresponse adjustment classes are defined by a binary auxiliary variable, C, that cross the strata as shown in Table 2.1. We replaced the X used in L&V with C for weighting cell as introduced above to easily identify the nonresponse adjustment cell. Consistent with L&V, the population size is set at $N = 10{,}000$.

Table 2.1
Population counts by strata *Z* and nonresponse adjustment cell *C*

	Nonresponse adjustment cell	
Sampling strata	***C* = 0**	***C* = 1**
Z = 0	3,064	3,931
Z = 1	2,079	926

Source: Little and Vartivarian (2003) who used *X* instead of *C*.

The variable of interest, Y, is a binary variable with the probability that $Y = 1$ defined by a logistic model with $\text{logit}(Y = 1|C, Z) = 0.5 + \gamma_C (C - \bar{C}) + \gamma_Z (Z - \bar{Z}) + \gamma_{CZ} (C - \bar{C})(Z - \bar{Z})$. The response variable R is also binary with the probability of $R = 1$ generated from a logistic model with $\text{logit}(R|C, Z) = 0.5 + \beta_C (C - \bar{C}) + \beta_Z (Z - \bar{Z}) + \beta_{CZ} (C - \bar{C})(Z - \bar{Z})$. Different populations and

response propensities are generated depending on the values of $\gamma_C, \gamma_Z, \gamma_{CZ}, \beta_C, \beta_Z$ and β_{CZ} as shown in Table 2.2. We have adopted the generalized linear model notation L&V used to make comparison to their work easier. The tabled values are the same populations and response variables that L&V generated by assigning values to $(\gamma_C, \gamma_Z, \gamma_{CZ}, \beta_C, \beta_Z, \beta_{CZ})$. In the notation $[A]^B$ in Table 2.2, the population (Y) or the response propensity (R) are indicated by the superscript B while the parameters and interactions of the model for the distribution of the population or response are indicated by A inside the brackets. For example, the additive logistic model that generates the distribution of Y within the sampling stratum Z and nonresponse cell C is indicated by $[C + Z]^Y$. Similarly, models where R depends on C only, Z only or neither C nor Z are denoted by $[C]^R, [Z]^R$, and $[C + Z]^R$ respectively. L&V give more details on their rationale for choosing these populations and response models.

Table 2.2
Models for outcome variable, *Y*, and probability of response, *R*

Model for *Y* (Variable of interest)	Model for *R* (Response propensity)	Parameters		
		γ_C, β_C	γ_Z, β_Z	γ_{CZ}, β_{CZ}
$[CZ]^Y$	$[CZ]^R$	2	2	2
$[C + Z]^Y$	$[C + Z]^R$	2	2	0
$[C]^Y$	$[C]^R$	2	0	0
$[Z]^Y$	$[Z]^R$	0	2	0
$[\phi]^Y$	$[\phi]^R$	0	0	0

Source: Little and Vartivarian (2003).

L&V computed estimates of means that are, in our notation,

$$\hat{\bar{y}}_{urr} = \frac{\hat{y}_{urr}}{\sum_c \hat{f}_c^u \sum_{i \in s_c} R_{ci} d_{ci}} = \frac{\hat{y}_{urr}}{\sum_c \hat{f}_c^u \hat{N}'_c}, \tag{2.6}$$

and

$$\hat{\bar{y}}_{wrr} = \frac{\hat{y}_{wrr}}{\sum_c \hat{f}_c^w \sum_{i \in s_c} R_{ci} d_{ci}} = \frac{\hat{y}_{wrr}}{\sum_c \hat{N}_c}. \tag{2.7}$$

The denominators of the means are estimates of the population size N. In estimator (2.7), the denominator is a constant and equal N, but in estimator (2.6) the denominator is a random variable. In the simulation setting with the stratified simple random sample design described below, or in any design where $\sum_{i \in s} d_i = N$ for every s, the estimator (2.7) reduces to the linear estimator $\hat{\bar{y}}_{wrr} = N^{-1} \hat{y}_{wrr}$; whereas (2.6) is a ratio estimator. This is an important point we return to later.

Domain means may have properties that differ from overall means because the denominators of the weighted and unweighted domain means are both random variables. One exception is when the domains match the sampling strata and therefore both the domain sizes and stratum sizes are known. L&V did not discuss domains, so these estimates are not studied in their simulation. We create domains by randomly

generating a random variable ν_i from a uniform (0, 1) distribution, and defining the membership function $\tau(a) = 1$ if $a < 0$ and $\tau(a) = 0$ if $a \geq 0$. Domain means of 50% were created by substituting $d_{ci}^{*} = \tau(\nu_i - 0.5)\, d_{ci}$ into expressions (2.6) and (2.7) to produce the estimators $\hat{\bar{y}}_{urr,0.5}$ and $\hat{\bar{y}}_{wrr,0.5}$, respectively. Weighted and unweighted estimators of domain totals $\hat{y}_{urr,0.5}$ and $\hat{y}_{wrr,0.5}$ were formed similarly. We used the same device to create 25 percent domain means and 25 percent domain totals. Since we are interested in the effect of the nonresponse adjustments in means computed as ratio estimators, other domains such as those defined close to 100 percent of the population were excluded from the analysis because the denominator of the domain means approaches the constant population total N and the mean becomes a linear estimator. Domains closer to 0 percent were excluded because of small sample sizes.

3 Findings

The simulation was done in R (R Development Core Team 2011) with 10,000 draws (L&V used 1,000 draws). We evaluated the estimators by computing the root mean squared error (rmse) and the bias of the estimates, where the bias and rmse are measured in deviations from the population quantities as done in L&V. We used the same total sample size of 312 that they used in their simulation, but with different sample allocation or relative sampling rates between strata. We replicated all 25 configurations in L&V and these results are found in Table S-1 in the supplemental materials. Table S-2 in supplemental materials also includes the 25 configurations but presents the relative bias of unweighted and weighted means and totals, as well as ratios of variances and rmse of unweighted to weighted estimates. The relative bias and ratios of variances and rmse facilitate the comparisons between the estimates. These materials include their estimated simulation errors, which are all relatively small. For those estimators and sampling rates given in L&V, our results are consistent with their published values within simulation error. We begin by examining the bias of the estimators.

3.1 Bias

There are two situations where theoretical results exist and are well-known (Little and Rubin 2002). One is when the propensity to respond is the same in all cells – missing completely at random (MCAR); MCAR corresponds to the model $[\phi]^R = (\beta_c = 0, \beta_z = 0, \beta_{cz} = 0)$ in the last row in Table 2.2. With MCAR, the unweighted and weighted adjustment factors are equal in expectation, and both produce unbiased estimates. The simulation results in Table V of L&V paper (rows 5, 10, 15, 20, and 25) confirm this observation. The second situation is when the response propensity is independent of the strata, corresponding to missing at random (MAR) with the response model $[\phi]^C = (\beta_c = 2, \beta_z = 0, \beta_{cz} = 0)$ in the third row of Table 2.2. We refer to these situations as MAR because the bias of the estimator does not depend on whether the information about Z is used in the model. Here again, the weighted and unweighted estimates are both unbiased and the adjustments are equal in expectation. The simulation results in Table V of L&V (rows 3, 8, 13, 18, and 23) confirm this observation empirically.

To focus on the situation in which the model is incorrectly specified, we do not present the simulations results for the MCAR and MAR situations in this document, but these results can be found in the supplemental materials. An important point is that even though the weighted and unweighted adjustments for the MCAR and MAR models are equal in expectation, they are not identical. Sukasih et al. (2009) simulated the two approaches under MAR models and stated a preference for the weighted approach largely due to the lower variability in the estimates of total across simulations even though both were unbiased.

As mentioned before, our simulations vary the sampling rates while keeping the overall sample size fixed at 312; L&V used a single sampling rate. When the sampling rates are the same across strata (i.e., the sample is proportionally allocated to the strata), then the sampling weights are the same for the two strata and consequently the weighted and unweighted estimators are identical. The proportional allocation sampling rate plays a visible role in our presentation because the two estimates must converge at this point.

Figure 3.1 (left panel) is a graph of the simulation results for the bias of the weighted and unweighted estimator of the total for $[CZ]^Y$ and $[C + Z]^R$. We chose this configuration (row 2 in L&V's tables) because the simulations in L&V showed the unweighted mean had lower bias and rmse than the weighted mean for this case. The horizontal axis shows the relative sampling rate computed as the ratio of the sampling rate of $Z = 0$ to $Z = 1$ or $N_0 n_0^{-1} / \left(N_1 n_1^{-1}\right)$. The relative sampling rate used by L&V was about 2.25. It is immediately apparent that the bias of the weighted estimator is essentially constant across different sampling rates while the bias of the unweighted estimator varies substantially with the relative sampling rate. The bias of the unweighted estimators of the total can be more than two times the bias of the weighted estimator for some sampling rates. Both estimators are biased for almost all relative sampling rates, and the estimator that has the lower bias depends on this rate. When the relative sampling rates are equal (proportional allocation) the unweighted and weighted estimators have the same bias, as expected. However, in practice, it is not generally possible to recognize the effect the sampling rate has on the bias and choose in advance the adjustment method to reduce the bias for a specific sample.

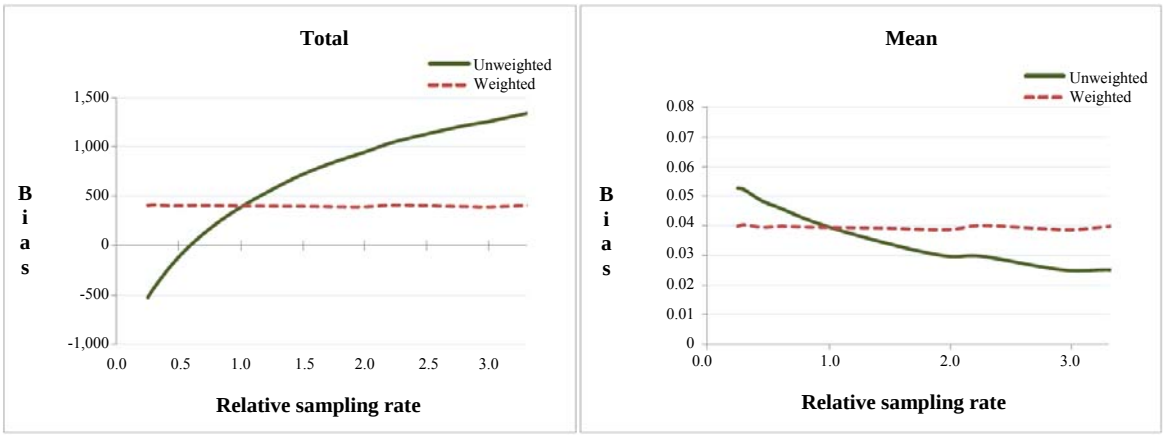

Figure 3.1 Bias for weighted and unweighted estimator for the population model $[CZ]^Y$ and response propensity model $[C+Z]^R$, where the left panel is for the total and right panel is for the mean.

To understand these findings, we applied some standard approximations that hold reasonably well in this situation (i.e., $E\left(\eta^{-1}\right) \approx E^{-1}\left(\eta\right)\right)$. The approximate expected value for the weighted estimator is

$$E\hat{y}_{wrr} \approx \sum_z \sum_c \frac{N_c}{\left(\sum_z \phi_{cz} N_{cz}\right)} \phi_{cz} Y_{cz}, \tag{3.1}$$

where Y_{cz} is the population total in cell cz. Similarly, the approximate expected value for the unweighted estimator is

$$E\hat{y}_{urr} \approx \sum_z \sum_c \frac{\left(\sum_z N_z n_z^{-1} N_{cz}\right)}{\left(\sum_z \phi_{cz} N_z n_z^{-1} N_{cz}\right)} \phi_{cz} Y_{cz}. \tag{3.2}$$

If ϕ_{cz} is a constant (MCAR) or ϕ_{cz} is constant within weighting cells (MAR), then both estimators are unbiased to this order of approximation and consistent with known theory. When the sampling rates are the same across strata, the two estimators have the same expected value (as noted above they are identical in this case). More importantly, these approximations show the expectation of the weighted estimator is not dependent on the sampling rate, but the expectation of the unweighted estimator is. This explains the patterns shown in the Figure 3.1.

Some details of the simulation estimates for this configuration are shown in Table 3.1 for selected sampling rates. As noted above, the full simulation results for all configurations and sampling rates used to create the figures can be found in the supplemental materials. These materials include the relative biases, ratios of variances and ratios of rmse which are better indicators for assessing the impact of the adjustments on the estimates. We observed that all configurations with biased estimates of totals have biases that are lower for the weighted estimator on one side of the relative sampling rate of 1 and are higher on the other side. All configurations exhibit an approximately constant bias for the weighted estimator of the total across the relative sampling rates, but the bias of the unweighted estimator varies by relative sampling rate.

Next, we turn to estimated means – the only estimators considered by L&V. The right panel of Figure 3.1 shows the bias for the weighted estimator is again independent of the relative sampling rate while the bias of the unweighted estimator varies with the sampling rate. L&V used a sampling rate of 2.25 so this explains why they found the unweighted estimator had a lower bias for the mean in their simulation. Two points are worth noting. First, the biases for the means for both adjustment methods are all relatively small, especially when compared to the potential relative biases of the totals with the unweighted estimator in the panel on the left. Second, there is no way to identify if a particular estimate would fall on the left or right of the relative sampling rate of 1. Table 3.1 shows the estimated biases for this configuration.

The graphs also show a relationship that is somewhat surprising; the relative sampling rates where the unweighted estimator of the total has a lower bias are those where the unweighted estimator of the mean has a higher bias. In other words, the means behave differently from the totals because the unweighted mean is a ratio while the weighted mean is not. As a result, the relative bias (rb = bias/estimate) of the unweighted estimator of the mean is not equal to the relative bias of the unweighted estimator of the total (the relationship holds for the weighted estimator). We approximate the relative bias by

$$rb\left(\hat{\bar{y}}_{urr}\right) \approx \frac{1 + rb\left(\hat{y}_{urr}\right)}{1 + rb\left(\hat{N}_{urr}\right)},$$

where $\hat{N}_{urr}$ is the unweighted estimator of the total (where $y_i = 1$ for all i). This approximation holds well in this situation since $\text{cov}\left(\hat{\bar{y}}_{urr}, \hat{N}_{urr}\right) / E\left(\hat{N}_{urr}\right) \approx 0$. Thus, the relative bias of the unweighted mean is reduced whenever the biases of the numerator and denominator are positively correlated.

Now, consider domain estimates – estimators not studied in L&V. The biases for the weighted and unweighted domain total estimators and the relationships with the biases of the unweighted estimators varying by the relative sampling rate are the same as observed for the overall totals (see Table 3.1). This follows because domain totals are still totals and approximations (3.1) and (3.2) still apply. The domain means are also given in the table and they too exhibit the same pattern of biases as shown in Figure 3.1 for the full sample mean. It is worth noting that the relative biases for the mean estimates (overall and for the domains) do not vary much, with most relative biases in the range of 5 to 7 percent.

Table 3.1
Bias (times 10,000), root mean square error (times 10,000) and variance of weighted and unweighted estimators of means and total of the full sample and domains, configuration $[CZ]^Y$, $[C+Z]^R$ with various sampling rates

				Relative sampling rate				
	Characteristic	**Domain**	**Adjustment**	**0.30**	**0.44**	**1.00**	**2.25**	**3.30**
Bias	Mean	Full	urr	515	491	404	301	248
			wrr	398	403	404	404	394
		50%	urr	513	501	411	307	257
			wrr	397	414	410	410	401
		25%	urr	523	498	407	298	252
			wrr	408	411	407	400	395
	Total	Full	urr	-419	-184	401	1,058	1,335
			wrr	398	403	404	404	394
		50%	urr	-214	-89	205	535	673
			wrr	194	205	206	207	200
		25%	urr	-107	-48	101	264	335
			wrr	97	98	102	101	100
Rmse	Mean	Full	urr	643	614	546	536	566
			wrr	553	547	545	587	616
		50%	urr	758	726	669	699	778
			wrr	687	671	669	728	794
		25%	urr	949	898	863	952	1,062
			wrr	895	859	863	955	1,041
	Total	Full	urr	537	376	543	1,183	1,485
			wrr	553	547	545	587	616
		50%	urr	371	311	393	714	888
			wrr	399	392	394	449	494
		25%	urr	255	233	282	451	553
			wrr	285	273	283	328	365
Variance	Mean	Full	urr	15	14	14	20	26
			wrr	15	14	14	18	22
		50%	urr	32	28	28	40	54
			wrr	32	28	28	37	47
		25%	urr	64	57	59	83	107
			wrr	64	58	59	76	93
	Total	Full	urr	11	11	14	28	43
			wrr	15	14	14	18	22
		50%	urr	9	9	11	23	34
			wrr	12	11	11	16	21
		25%	urr	5	5	7	14	20
			wrr	7	7	7	10	12

3.2 Root mean square error

Despite the small sample size used in the simulations (312 before nonresponse) and the relatively modest relative bias of the estimates for means, the bias is still a large component of the rmse. For example, the bias

accounts for 56 (unweighted) to 69 (weighted) percent of the rmse for the estimate of the mean in the $[CZ]^Y$ and $[C + Z]^R$ configuration using the L&V sampling rate. With larger sample sizes that are common in large sample surveys, the bias is often the dominant component of the rmse (Brick 2013).

Figure 3.2 shows the rmse for the estimated total (left panel) and for the mean (right panel) using the same configuration used in the previous figure. The rmse for the total for the weighted estimator is approximately constant and smaller than the rmse for the unweighted estimator, except when the relative sampling rate is about 0.5 which corresponds to the region with very low bias for the unweighted estimator as shown in Figure 3.1. However, when the relative sampling rate is greater than one, the rmse of the unweighted estimator of the total is much larger than the rmse of the weighted estimator (it can be as much as twice the rmse of the weighted estimator for some sampling rates). In contrast, for the estimates of the mean shown in Figure 3.2 (right panel), the rmse of both the weighted and unweighted estimators are similar in magnitude, and the symmetry around the proportional allocation rate remains. Even though L&V point out the unweighted estimator has a lower rmse (at the relative sampling rate of 2.25), we view the rmse of both estimators to be approximately equal across the range of relative sampling rates.

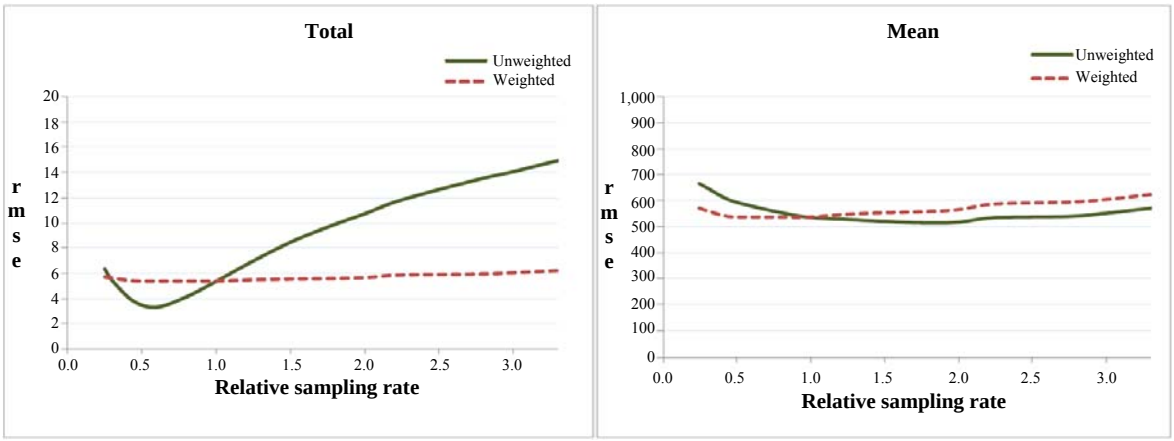

Figure 3.2 Root mean square error for weighted and unweighted estimator when $[CZ]^Y$ and $[C+Z]^R$, where the left panel is for the total (rmse is in millions) and the right panel is for the mean.

Figure 3.3 shows the rmse for the estimated 50% domain mean (left panel) and for the 25% domain mean (right panel) again using $[CZ]^Y$ and $[C + Z]^R$. Looking at the three graphs of the rmse of the means (for the overall mean, the 50% domain mean, and the 25% domain mean) reveals the effect of the ratio estimator. As the percentage in the domain decreases from 100% to 25%, the weighted estimator becomes more like an unconditional ratio estimator and the correlation between the numerator and denominator reduces the rmse of the estimate. As a result, the rmse of the weighted and unweighted estimators are very similar for the domain estimators. Even though the weighted estimator has a lower rmse at each of the relative sampling rates compared to the unweighted one for the 25% domain mean, the two estimators are essentially equivalent in terms of rmse. The slight advantage of the unweighted estimator pointed out by L&V for the full population mean for this configuration vanishes for domain means where the weighted estimator is also a ratio estimator.

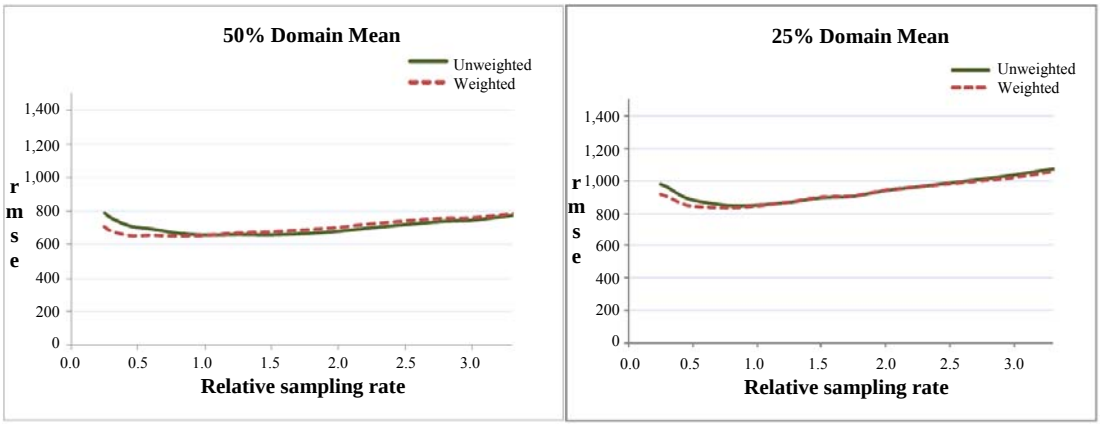

Figure 3.3 Root mean square error for weighted and unweighted estimator when $[CZ]^Y$ and $[C+Z]^R$, where the left panel is for the 50% domain mean and the right panel is for the 25% domain mean.

3.3 Variance

A general concern about nonresponse adjustment factors is that when the factors are based on a small number of respondents they may increase the variance of the estimates (Kalton 1983; Tremblay 1986). L&V suggest weighting the nonresponse adjustment factors may be responsible for greater variance inflation than using the unweighted factors. The figures above show that this did not occur in this simulation. Figure 3.4 shows the ratio of the unweighted estimator's variance to that of the weighted estimates for the full population mean and total and the 50% domain total for the $[CZ]^Y$ and $[C + Z]^R$ configuration. For the mean, the variance ratio is nearly equal to one over all the relative sampling rates showing no inflation of variance for the weighted estimator compared to the unweighted estimator. For totals, the ratio is less than unity for relative sampling rates less than 1 and greater than 1 for relative sampling rates greater than unity. The same relationship holds true for the 50% domain total. These results suggest that weighting the adjustment is not the source of large factors that can inflate the variance of the estimates. A prudent approach is to examine the size of nonresponse factors, irrespective of whether they are weighted or unweighted.

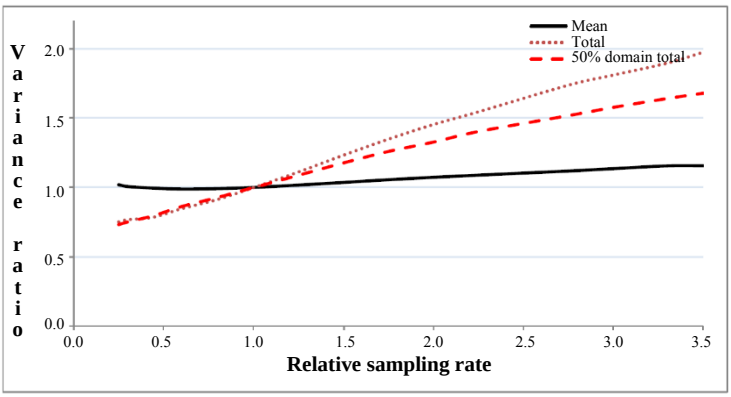

Figure 3.4 Ratio of variance of unweighted to weighted estimates of the mean, total and 50% domain total for $[CZ]^Y$ and $[C+Z]^R$.

Table 3.2 gives the simulation results for another configuration, $[CZ]^Y$ and $[CZ]^R$, that was favorable to the unweighted adjustment in L&V (the first row in their tables). In contrast, Table 3.3 gives simulation results for $[C + Z]^Y$ and $[C + Z]^R$ which is a configuration that was favorable to the weighted adjustment. The results for both of these configurations show the same general patterns as discussed above for $[CZ]^Y$ and $[C + Z]^R$.

Table 3.2
Bias (times 10,000), root mean square error (times 10,000) and variance of weighted and unweighted estimators of means and total of the full sample and domains, configuration $[CZ]^Y$, $[CZ]^R$ with various sampling rates

				Relative sampling rate				
	Characteristic	**Domain**	**Adjustment**	**0.30**	**0.44**	**1.00**	**2.25**	**3.30**
Bias	Mean	Full	urr	329	329	289	255	237
			wrr	294	299	289	298	298
		50%	urr	334	341	293	251	238
			wrr	299	311	293	294	298
		25%	urr	336	344	306	257	247
			wrr	302	314	306	299	307
	Total	Full	urr	-412	-187	287	732	901
			wrr	294	299	289	298	298
		50%	urr	-209	-91	145	367	455
			wrr	143	152	146	149	154
		25%	urr	-103	-46	72	184	230
			wrr	74	76	73	75	79
Rmse	Mean	Full	urr	530	507	476	501	533
			wrr	505	487	476	520	554
		50%	urr	684	653	616	664	732
			wrr	666	638	616	674	740
		25%	urr	911	859	832	920	1,016
			wrr	900	849	832	920	1,011
	Total	Full	urr	550	395	474	886	1,078
			wrr	505	487	476	520	554
		50%	urr	385	326	373	575	696
			wrr	394	375	373	425	475
		25%	urr	263	244	278	390	464
			wrr	285	274	278	321	361
Variance	Mean	Full	urr	17	15	14	19	23
			wrr	17	15	14	18	22
		50%	urr	36	31	30	38	48
			wrr	36	31	30	37	46
		25%	urr	73	63	61	79	98
			wrr	73	63	61	76	94
	Total	Full	urr	14	12	14	25	35
			wrr	17	15	14	18	22
		50%	urr	11	10	12	20	28
			wrr	14	12	12	16	20
		25%	urr	6	6	7	12	16
			wrr	8	7	7	10	13

Table 3.3
Bias (times 10,000), root mean square error (times 10,000) and variance of weighted and unweighted estimators of means and total of the full sample and domains, configuration $[C+Z]^Y$, $[C+Z]^R$ with various sampling rates

				Relative sampling rate				
	Characteristic	**Domain**	**Adjustment**	**0.30**	**0.44**	**1.00**	**2.25**	**3.30**
Bias	Mean	Full	urr	763	735	654	566	529
			wrr	665	661	654	654	652
		50%	urr	773	737	653	564	532
			wrr	677	664	653	651	656
		25%	urr	773	739	659	574	513
			wrr	679	668	659	660	636
	Total	Full	urr	-272	-8	651	1,411	1,744
			wrr	665	661	654	654	652
		50%	urr	-133	-6	326	711	875
			wrr	336	328	328	332	328
		25%	urr	-69	-2	157	359	438
			wrr	165	166	158	168	165
Rmse	Mean	Full	urr	854	818	745	699	711
			wrr	767	753	745	764	790
		50%	urr	951	901	827	816	863
			wrr	877	845	826	863	912
		25%	urr	1,101	1,046	981	1,023	1,098
			wrr	1,044	1,004	981	1,045	1,107
	Total	Full	urr	426	313	741	1,503	1,868
			wrr	767	753	745	764	790
		50%	urr	334	300	475	867	1,071
			wrr	489	470	476	529	575
		25%	urr	246	240	314	530	649
			wrr	320	316	314	372	409
Variance	Mean	Full	urr	15	13	13	17	23
			wrr	15	13	13	16	20
		50%	urr	31	27	26	35	46
			wrr	31	28	26	32	40
		25%	urr	62	56	54	73	95
			wrr	63	57	54	67	83
	Total	Full	urr	11	10	13	27	45
			wrr	15	13	13	16	20
		50%	urr	10	9	12	25	39
			wrr	13	12	12	17	22
		25%	urr	6	6	7	15	23
			wrr	8	7	8	11	14

3.4 Estimating population size

A particular type of estimate studied by Sukasih et al. (2009) is the estimate of the number of units in a population. We refer to this as an estimate of population size where the population size is just an estimate of a total where $y_i = 1$ for all i. It can be estimated for a domain by assigning all units outside the domain $y_i = 0$. In the simple stratified sample design studied here, the weighted estimator always reproduces the overall total population size $N = 10{,}000$, but the unweighted estimator does not. Since this situation clearly favors the weighted estimator, we instead examine the domain population size estimates.

Suppose we are estimating the number of units in a domain or subgroup that have a value below a percentile defined by a characteristic for the total population (e.g., national median income). This type of

statistic is extremely important in surveys because estimates of the population size for domains are often key outcome statistics. For example, an estimate of this type is the total number of persons with an income below the poverty line or the low income line (Kovačević and Yung 1997).

The L&V analysis did not consider estimates for domains sizes or means, so there is not an explicit variable that can be used to define a subpopulation. To avoid complicating this analysis, we illustrate the performance of the two estimators using an artificial domain created by randomly selecting half of the population (i.e., 50% domain). Similar to the analysis in previous sections we computed weighted and unweighted totals and means for the 50% domain. Even though we know the size for this domain beforehand for this example (i.e., 50 percent of the total population), the analysis is still valid. In practice, the domain size would not be known.

When estimating a statistic such as the population size in a domain, both the weighted and unweighted estimators of domain population size are unbiased when the data are MCAR or MAR, as noted by Sukasih et al. (2009). Furthermore, the rmse errors of the weighted and unweighted estimators are approximately equal in this case as confirmed in the simulations.

When the data are not MAR, the situation may be very different. The weighted estimator of a domain population size is approximately unbiased for all relative sampling rates and all configurations, but the unweighted estimator is always biased except when it is identical to the weighted estimator (at a relative sampling rate of 1). As a consequence the rmse of the unweighted estimator for the domain size is often considerably greater than that of the weighted estimator. Figure 3.5 shows that the rmse of the unweighted estimator of the 50% domain size for $[CZ]^Y$ and $[C + Z]^R$ is substantially greater than that of the weighted estimator for most relative sampling rates (as much as twice the rmse of the weighted estimator). The only exception is when the two estimators are approximately equal (near proportional allocation).

The weighted estimator of domain sizes thus has a substantial advantage over the unweighted estimator for all of the missing data mechanisms in L&V that are not MCAR or MAR.

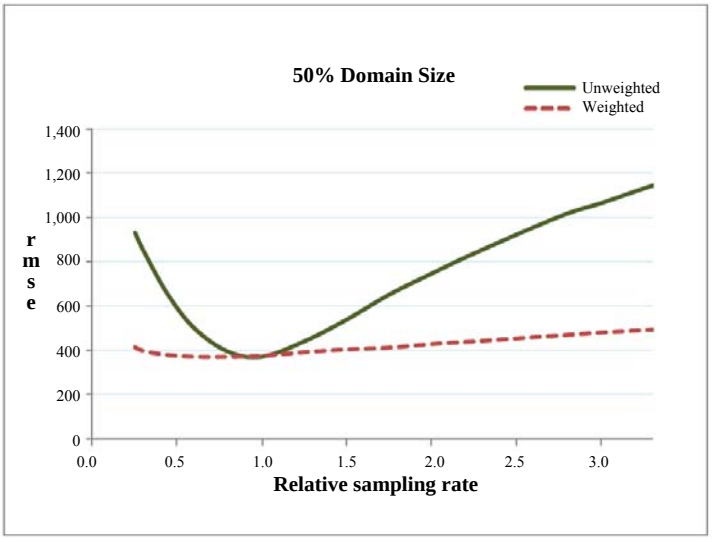

Figure 3.5 Root mean square error (rmse) for 50% domain size weighted and unweighted estimators when $[CZ]^Y$ and $[C+Z]^R$.

4 Conclusions

Nearly every survey suffers from nonresponse so the method for adjusting the base weights for unit nonresponse is an important topic. L&V appropriately noted that using design weights to compute a weighted nonresponse adjustment does not eliminate nonresponse bias when the response mechanism is not specified correctly in the weighting adjustment model. However, their simulation study suggested to at least some researchers that an unweighted adjustment might be more appropriate than a weighted adjustment more generally. The results from our evaluation, using the same setting as in L&V, contradict this perception. We explored the differences between the unweighted and weighted estimators when the adjustment model is misspecified in more detail using the L&V setting by including different sampling rates and estimates of totals and domains in addition to the means discussed in L&V.

These expanded simulations show that the unweighted and weighted adjustments do have different properties. The bias of the weighted estimator of totals means in stratified simple random sample designs is approximately constant irrespective of the sampling rate but the bias of the unweighted estimator depends on the sampling rate. In contrast, the bias of the unweighted estimator of the total is substantially larger than that of the weighted estimator for some sampling rates. For means, the bias and the rmse of the two estimators are not very different including those configurations that L&V described as favoring the unweighted estimator. The same general conclusions hold for estimates of domain means and totals as the weighted mean becomes more of a ratio estimate for domains and this influences its behavior somewhat.

We also looked at estimating domain sizes. With this type of statistic, the rmse of the weighted estimator is almost uniformly lower than the rmse of the unweighted estimator when the data are not MAR in the simulation settings. The differences are due to the bias in the unweighted estimator of the domain size, and this bias causes the unweighted estimator to have a substantially greater rmse compared to the weighted estimator for some sampling rates.

Imperfect models are used in most surveys so the nonresponse adjustment method is important. The expanded simulation findings we present show the weighted adjustment has substantial advantages for some estimates and for some sampling rates when compared to the unweighted adjustment. In particular, any survey with this design that produces estimates of totals and statistics other than just means appears to benefit by weighting the adjustment. Of course, weighting the adjustment does not remove bias; weighting does diminish the magnitude of the bias in many situations and for many of the estimators we examined. The bias of the weighted estimator also is not sensitive to the relative sampling rate, but the bias of the unweighted estimator is sensitive. The potential disadvantage of an increase in the variance of the estimate using the weighted adjustment did not arise in these simulations, and can be avoided by inspecting the adjustment factors, as should also be done with an unweighted adjustment. Finally, the results of this study highlight the potential problem of generalizing from simulations. Although simulations are valuable to demonstrate a specific point, generalizing simulation findings more broadly can be misleading especially when the findings are highly dependent on the conditions of the model being simulated.

References

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics,* 29(2), 329-353.

Brick, J., and Jones, M. (2008). Propensity to respond and nonresponse bias. *Metron-International Journal of Statistics*, LXVI, 51-73.

Chadborn, T.R., Baster, K., Delpech, V., Sabin, C.A., Sinka, K., Rice, B.D. and Evans, B. (2005). No time to wait: How many HIV-infected homosexual men are diagnosed late and consequently die? (England and Wales, 1993-2002). *Aids*, 19(5), 513-520.

Grau, E., Potter, F., Williams, S. and Diaz-Tena, N. (2006). Nonresponse adjustment using logistic regression: To weight or not to weight? *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3073-3080.

Haukoos, J.S., and Newgard, C.D. (2007). Advanced statistics: Missing data in clinical research - part 1: An introduction and conceptual framework. *Academic Emergency Medicine*, 14(7), 662-668.

Kalton, G. (1983). *Introduction to Survey Sampling, SAGE University Paper 35.* Thousand Oaks, CA: SAGE Publications.

Kott, P. (2012). Why one should incorporate the design weights when adjusting for unit nonresponse using response homogeneity groups. *Survey Methodology*, 38, 1, 95-99.

Kovačević, M., and Yung, W. (1997). Variance estimation for measures of income inequality and polarization - An empirical study. *Survey Methodology*, 23, 1, 41-52.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M. and Raghunathan, T.E. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response: Examples from multiple surveys. *Journal of the Royal Statistical Society Series A*, Royal Statistical Society, 173(2), 389-407.

Little, R.J. (1986). Survey nonresponse adjustments. *International Statistical Review*, 54, 139-157.

Little, R.J., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data (2nd Ed.)*. New York: John Wiley & Sons, Inc.

Little, R., and Vartivarian, S. (2003). On weighting the rates in nonresponse weights. *Statistics in Medicine*, 22, 1589-1599.

R Development Core Team (2011). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. doi: http://www.R-project.org.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* Chichester, England: John Wiley & Sons, Inc.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer.

Sukasih, A., Jang, D., Vartivarian, S., Cohen, S. and Zhang, F. (2009). A simulation study to compare weighting methods for nonresponses in the National Survey of Recent College Graduates. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Retrieved October 21, 2013, from www.amstat.org/sections/srms/proceedings/y2009/Files/304345.pdf.

Tremblay, V. (1986). Practical criteria for definition of weighting classes. *Survey Methodology,* 12, 1, 85-97.

West, B.T. (2009). A simulation study of alternative weighting class adjustments for nonresponse when estimating a population mean from complex sample survey data. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Retrieved October 21, 2013, from www.amstat.org/sections/srms/proceedings/y2009/Files/305394.pdf.

Wun, L.-M., Ezzati-Rice, T.M., Diaz-Tena, N. and Greenblatt, J. (2007). On modelling response propensity for dwelling unit (DU) level non-response adjustment in the Medical Expenditure Panel Survey (MEPS). *Statistics in Medicine*, 26(8), 1875-1884.

Survey Methodology, June 2016
Vol. 42, No. 1, pp. 179-187
Statistics Canada, Catalogue No. 12-001-X

A short note on quantile and expectile estimation in unequal probability samples

Linda Schulze Waltrup and Göran Kauermann[1]

Abstract

The estimation of quantiles is an important topic not only in the regression framework, but also in sampling theory. A natural alternative or addition to quantiles are expectiles. Expectiles as a generalization of the mean have become popular during the last years as they not only give a more detailed picture of the data than the ordinary mean, but also can serve as a basis to calculate quantiles by using their close relationship. We show, how to estimate expectiles under sampling with unequal probabilities and how expectiles can be used to estimate the distribution function. The resulting fitted distribution function estimator can be inverted leading to quantile estimates. We run a simulation study to investigate and compare the efficiency of the expectile based estimator.

Key Words: Quantiles; Expectiles; Probability proportional to size; Design-based; Auxiliary variable; Distribution function.

1 Introduction

Quantile estimation and quantile regression have seen a number of new developments in recent years with Koenker (2005) as a central reference. The principle idea is thereby to estimate an inverted cumulative distribution function, generally called the quantile function $Q(\alpha) = F^{-1}(\alpha)$ for $\alpha \in (0,1)$, where the 0.5 quantile $Q(0.5)$, the median, plays a central role. For survey data tracing from an unequal probability sample with known probabilities of inclusion Kuk (1988) shows how to estimate quantiles taking the inclusion probabilities into account. The central idea is to estimate a distribution function of the variable of interest and invert this in a second step to obtain the quantile function. Chambers and Dunstan (1986) propose a model-based estimator of the distribution function. Rao, Kovar and Mantel (1990) propose a design-based estimator of the cumulative distribution function using auxiliary information. Bayesian approaches in this direction have recently been proposed in Chen, Elliott, and Little (2010) and Chen, Elliott, and Little (2012).

Quantile estimation results from minimizing an L_1 loss function as demonstrated in Koenker (2005). If the L_1 loss is replaced by the L_2 loss function one obtains so called expectiles as introduced in Aigner, Amemiya and Poirier (1976) or Newey and Powell (1987). For $\alpha \in (0,1)$, this leads to the expectile function $M(\alpha)$ which, like the quantile function $Q(\alpha)$, uniquely defines the cumulative distribution function $F(y)$. Expectiles are relatively easy to estimate and they have recently gained some interest, see e.g., Schnabel and Eilers (2009), Pratesi, Ranalli, and Salvati (2009), Sobotka and Kneib (2012) and Guo and Härdle (2013). However since expectiles lack a simple interpretation their acceptance and usage in statistics is less developed than quantiles, see Kneib (2013). Quantiles and expectiles are connected in that a unique and invertible transformation function $h_y : [0,1] \to [0,1]$ exists so that $M(h(\alpha)) = Q(\alpha)$, see Yao and Tong (1996) and De Rossi and Harvey (2009). This connection can be used to estimate quantiles

1. Linda Schulze Waltrup, Business Administration and Social Sciences, Ludwig Maximilian University of Munich, Ludwigstraße 33, 80539 Munich, Germany. E-mail: lschulze_waltrup@stat.uni-muenchen.de; Göran Kauermann, Business Administration and Social Sciences, Ludwig Maximilian University of Munich, Ludwigstraße 33, 80539 Munich, Germany. E-mail: goeran.kauermann@stat.uni-muenchen.de.

from a set of fitted expectiles. The idea has been used in Schulze Waltrup, Sobotka, Kneib and Kauermann (2014) and the authors show empirically that the resulting quantiles can be more efficient than empirical quantiles, even if a smoothing step is applied to the latter (see Jones 1992). An intuitive explanation for this is that expectiles account for all the data while quantiles based on the empirical distribution function only take the left (or the right) hand side of the data into account. That is, the median is defined by the 50% left (or 50% right) part of the data while the mean (as 50% expectile) is a function of all data points. In this note we extend these findings and demonstrate how expectiles can be estimated for unequal probability samples and how to obtain a fitted distribution function from fitted expectiles.

The paper is organized as follows. In Section 2 we give the necessary notation and discuss quantile regression in unequal probability sampling. This is extended in Section 3 towards expectile estimation. Section 4 utilizes the connection between expectiles and quantiles and demonstrates how to derive quantiles from fitted expectiles. Section 5 demonstrates in simulations the efficiency gain in quantiles derived from expectiles and a discussion concludes the paper in Section 6.

2 Quantile estimation

We consider a finite population with N elements and a continuous survey variable Y. We are interested in quantiles of the cumulative distribution function $F(y) = \sum_{i=1}^{N} 1\{Y_i \leq y\}/N$ and define as

$$Q(\alpha) = \inf\left\{\arg\min_{q} \sum_{i=1}^{N} w_\alpha (Y_i - q)\,|\,Y_i - q\,|\right\} \tag{2.1}$$

the Quantile function of Y (see Koenker 2005), where

$$w_\alpha(\varepsilon) = \begin{cases} \alpha & \text{for } \varepsilon > 0 \\ 1-\alpha & \text{for } \varepsilon \leq 0. \end{cases}$$

The "inf" argument in (2.1) is required in finite populations since the "arg min" is not unique. We draw a sample from the population with known inclusion probabilities $\pi_i,\ i = 1,\ldots,N$. Denoting by $y_1,\ldots,y_n$ the resulting sample, we estimate the quantile function by replacing (2.1) through its weighted sample version

$$\hat{Q}_N(\alpha) = \inf\left\{\arg\min_{q} \sum_{j=1}^{n} \frac{1}{\pi_j} w_{\alpha,j}\,|\,y_j - q\,|\right\} \tag{2.2}$$

with $w_{\alpha,j} = w_\alpha(y_j - q)$ as defined above. It is easy to see that the sum in (2.2) is a design-unbiased estimate for the sum in $Q(\alpha)$ given in (2.1). Nonetheless, because we take the "arg min" it follows that $\hat{Q}_N(\alpha)$ is not unbiased for $Q(\alpha)$. We therefore look at consistency statements for $\hat{Q}_N(\alpha)$ as follows. Let $R_i(q) = w_\alpha(y_i - q)\,|\,y_i - q\,|$ and

$$\bar{R}_N(q) := \frac{1}{N}\sum_i R_i(q).$$

We draw a sample from $R_i(q), i = 1, \ldots, N$ and assume we apply a consistent sampling scheme in that

$$\bar{r}_n(q) := \frac{1}{N}\sum_{j=1}^{n}\frac{1}{\pi_j} r_j(q)$$

is design-consistent for $\bar{R}_N(q)$, where $r_j(q)$ denotes the sample of $R_i(q)$. Note that $r_j(q)$ and hence $\bar{r}_n(q)$, $R_i(q)$ and $\bar{R}_N(q)$ also depend on α which is suppressed in the notation for readability. Let q_0 be the minimum of $\bar{R}_N(q)$ which is not necessarily unique due to the finite structure of the population. We can take the "inf" argument, i.e., $q_0 = \inf\{\arg\min \bar{R}_N(q)\}$, but for simplicity we assume a superpopulation model (see Isaki and Fuller 1982) by considering the finite population to be a sample from an infinite superpopulation. In the latter we assume that survey variable Y has a continuous cumulative distribution function so q_0 results in a unique α quantile. We get for $\delta > 0$

$$P(\bar{r}_n(q_0) < \bar{r}_n(q_0 - \delta)) \Leftrightarrow P\left(\frac{1}{N}\sum_{j=1}^{n}\frac{1}{\pi_j}\{r_j(q_0) - r_j(q_0 - \delta)\} < 0\right).$$

Note that the argument in the probability statement is a design-consistent estimate for $\bar{R}_N(q_0) - \bar{R}_N(q_0 - \delta)$, which is less than zero since q_0 is the minimum of $\bar{R}_N(\cdot)$. Hence, the probability tends to one in the sense of design consistency defined in Isaki and Fuller (1982). The same holds of course for $\delta < 0$. With this statement we may conclude that the estimated minimum $\hat{q}_0 = \arg\min \sum_{j=1}^{n} 1/\pi_j \; r_j(q)$ is a design-consistent estimate for q_0 so that $\hat{Q}_N(\alpha)$ in (2.2) is in turn design-consistent for $Q_N(\alpha)$. It is easily shown that $\hat{Q}_N(\alpha)$ is the inverse of the normed weighted cumulative distribution function

$$\hat{F}_N(y) := \frac{\sum_{j=1}^{n} 1\{y_j \le y\}/\pi_j}{\sum_{j=1}^{n} 1/\pi_j}$$

using the same notation as in Kuk (1988). Note that $\hat{F}_N(y)$ is the Hajek (1971) estimate of the cumulative distribution function (see also Rao and Wu 2009) and as such not a Horvitz-Thompson estimate. As a consequence $\hat{Q}_N(\alpha)$ is not design-unbiased. Nonetheless, $\hat{F}_N(y)$ is a valid distribution function, and hence it can be considered as normalized version of the Lahiri or Horvitz-Thompson estimator of the distribution function (see Lahiri 1951) which is denoted by

$$\hat{F}_L(y) := \frac{1}{N}\sum_{j=1}^{n} 1/\pi_j 1\{y_j \le y\}.$$

Kuk (1988) proposes to replace $\hat{F}_L(\cdot)$ with alternative estimates of the distribution function: Instead of estimating the distribution function itself he suggests to estimate the complementary proportion $\hat{S}_R(y)$ which then leads to the estimate $\hat{F}_R(y)$ defined through

$$\hat{F}_R(y) = 1 - \hat{S}_R(y) = 1 - \frac{1}{N}\sum_{j=1}^{n} 1/\pi_j 1\{y_j > y\}.$$

Resulting directly from these definitions we can express $\hat{F}_R(\cdot)$ in terms of $\hat{F}_N(\cdot)$ through

$$\hat{F}_R = 1 - \frac{1}{N}\sum_{j=1}^{n} 1/\pi_j + \hat{F}_L \quad \text{and} \quad \hat{F}_L = \frac{\sum_{j=1}^{n} 1/\pi_j}{N}\hat{F}_N. \tag{2.3}$$

Kuk (1988) shows that, under sampling with unequal probabilities, estimation of the median derived from $\hat{F}_R$ outperforms median estimates derived from $\hat{F}_N$ and $\hat{F}_L$ in terms of mean squared estimation error. Note that the estimators $\hat{F}_N$, $\hat{F}_L$ and $\hat{F}_R$ coincide in the case of simple random sampling without replacement where $\pi_j = \pi = n/N$.

3 Expectile estimation

An alternative to quantiles are expectiles. The expectile function $M(\alpha)$ is thereby defined by replacing the L_1 loss in (2.1) by the L_2 loss leading to

$$M(\alpha) = \arg\min_m \left\{ \sum_{i=1}^{N} w_\alpha (Y_i - m)(Y_i - m)^2 \right\}. \tag{3.1}$$

Note that $M(\alpha)$ is continuous in α even for finite populations. Moreover $M(0.5)$ equals the mean value $\bar{Y} = \sum_{i=1}^{N} Y_i / N$. Using the sample $y_1, \ldots, y_n$ with inclusion probabilities $\pi_1, \ldots, \pi_n$ we can estimate $M(\alpha)$ by replacing the sum in (2.2) by its sample version, i.e.,

$$\hat{M}(\alpha) = \arg\min_m \left\{ \sum_{j=1}^{n} \frac{1}{\pi_j} w_{\alpha,j} (y_j - m)^2 \right\}$$

with $w_{\alpha,j}$ as defined above. It is easy to see that the sum in $\hat{M}(\alpha)$ is a design-unbiased estimate for the sum in $M(\alpha)$. The estimate itself is however not design-unbiased like for the quantile function above. However the same arguments as for $Q_N(\alpha)$ in (2.2) may be used to establish design-consistency.

4 From expectiles to the distribution function

Both, the quantile function $Q(\alpha)$ and the expectile function $M(\alpha)$ uniquely define a distribution function $F(.)$. While $Q(\alpha)$ is just the inversion of $F(.)$ the relation between $M(\alpha)$ and $F(.)$ is more complicated. Following Schnabel and Eilers (2009) and Yao and Tong (1996), we have the relation

$$M(\alpha) = \frac{(1-\alpha) G(M(\alpha)) + \alpha \{M(0.5) - G(M(\alpha))\}}{(1-\alpha) F(M(\alpha)) + \alpha \{1 - F(M(\alpha))\}}, \tag{4.1}$$

where $G(m)$ is the moment function defined through $G(m) = \sum_{i=1}^{N} Y_i 1\{Y_i \le m\}/N$. Expression (4.1) gives the unique relation of function $M(\alpha)$ to the distribution function $F(.)$. The idea is now to solve (4.1) for $F(.)$, that is to express the distribution $F(.)$ in terms of the expectile function $M(.)$. Apparently,

this is not possible in analytic form but it may be calculated numerically. To do so, we evaluate the fitted function $\hat{M}(\alpha)$ at a dense set of values $0 < \alpha_1 < \alpha_2 \ldots < \alpha_L < 1$ and denote the fitted values as $\hat{m}_l = \hat{M}(\alpha_l)$. We also define left and right bounds through $\hat{m}_o = \hat{m}_1 - c_0$ and $\hat{m}_{L+1} = \hat{m}_L + c_{L+1}$, where c_0 and c_L are some constants to be defined by the user. For instance, one may set $c_0 = \hat{m}_2 - \hat{m}_1$ and $c_{L+1} = \hat{m}_L - \hat{m}_{L-1}$. By doing so we derive fitted values for the cumulative distribution function $F(.)$ at $\hat{m}_l$ which we write as $\hat{F}_l := \hat{F}(\hat{m}_l) = \sum_{j=1}^{l} \hat{\delta}_j$ for non-negative steps $\hat{\delta}_j \geq 0, j = 1, \ldots, L$ with $\sum_{j=1}^{L} \hat{\delta}_j \leq 1$. We define $\hat{\delta}_{L+1} = 1 - \sum_{l=1}^{L} \hat{\delta}_l$ to make $\hat{F}(.)$ a distribution function. Assuming a uniform distribution between the dense supporting points $\hat{m}_l$ we may express the moment function $G(.)$ by simple stepwise integration as

$$\hat{G}_l := \hat{G}(\hat{m}_l) = \int_{-\infty}^{m_l} x \, d\hat{F}(x) = \sum_{j=1}^{l} \hat{d}_j \hat{\delta}_l,$$

where $\hat{d}_j = (\hat{m}_j - \hat{m}_{j-1})/2$ with the constraint that $\hat{G}_{L+1} = \hat{M}(0.5)$ and $\hat{M}(0.5) = \sum_{j=1}^{n}(y_j/\pi_j) \Big/ \sum_{j=1}^{n}(1/\pi_j)$. With the steps $\hat{\delta}_l, l = 1, \ldots, L$ we can now re-express (4.1) as

$$\hat{m}_l = \frac{(1-\alpha)\sum_{j=1}^{l} \hat{d}_j \hat{\delta}_j + \alpha\left(\hat{M}(0.5) - \sum_{j=1}^{l} \hat{d}_j \hat{\delta}_j\right)}{(1-\alpha)\sum_{j=1}^{l} \hat{\delta}_j + \alpha\left(1 - \sum_{j=1}^{l} \hat{\delta}_j\right)}, \quad l = 1, \ldots, L,$$

which is then be solved for $\hat{\delta}_1, \ldots, \hat{\delta}_L$. This is a numerical exercise which is conceptually relatively straightforward. Details can be found in Schulze Waltrup et al. (2014). Once we have calculated $\hat{\delta}_1, \ldots, \hat{\delta}_L$ we have an estimate for the cumulative distribution function which is denoted as $\hat{F}_N^M(y) = \sum_{l:\hat{m}_l < y} \hat{\delta}_l$. We may also invert $\hat{F}_N^M(.)$ which leads to a fitted quantile function which we denote with $\hat{Q}_N^M(\alpha)$.

As Kuk (1988) shows, both theoretically and empirically, $\hat{F}_R(.)$ is more efficient than $\hat{F}_N(.)$. We make use of this relationship and apply it to $\hat{F}_N^M(.)$, which yields the estimator

$$\hat{F}_R^M := 1 - \frac{1}{N}\sum_{j=1}^{n} 1/\pi_j + \frac{\sum_{j=1}^{n} 1/\pi_j}{N} \hat{F}_N^M.$$

In the next section we compare the quantiles calculated from the expectile based estimator $\hat{F}_R^M$ with quantiles calculated from $\hat{F}_R$. Note that neither $\hat{F}_R^M$ nor $\hat{F}_R$ are proper distribution functions since they are not normed to take values between 0 and 1.

5 Simulations

We run a small simulation study to show the performance of the expectile based estimates. In the following, we make use of the Mizuno sampling method (see Midzuno 1952) and define the inclusion

probabilities π_j proportional to a measure of size x, see R package "sampling" by Tillé and Matei (2015). We examine two data sets also used in Kuk (1988). The first data set (Dwellings) contains two variables, the number of dwelling units (X), and the number of rented units (Y), which are highly correlated (with a correlation of 0.97); see also Kish (1965). The second data set (Villages) includes information on the population (X) and on the number of workers in household industry (Y) for 128 villages in India; see Murthy (1967). In the second data set the correlation between Y and X is 0.54. In order to compare our simulation results with the results of Kuk (1988) we choose the same sample size of $n = 30$ (from a total population of $N = 270$ for the Dwellings data and $N = 128$ for the Villages data).

We compare quantiles defined by inversion of $\hat{F}_R$ with quantiles defined by inversion of $\hat{F}_R^M$. In Table 5.1 we give the root mean squared error (RMSE) and the relative efficiency for specified quantiles. We note that the median for the village data and for the Dwelling data also upper quantiles derived from expectiles yield increased efficiency. Also the efficiency gain does not hold uniformly as we observe a loss of efficiency for lower quantiles.

Table 5.1
Comparison of mean squared error on a basis of 500 replications

	α	quantiles $\sqrt{\text{MSE}\left(\hat{Q}_R(\alpha)\right)}$	quantiles from expectiles $\sqrt{\text{MSE}\left(\hat{Q}_R^M(\alpha)\right)}$	relative efficiency $\frac{\sqrt{\text{MSE}\left(\hat{Q}_R^M(\alpha)\right)}}{\sqrt{\text{MSE}\left(\hat{Q}_R(\alpha)\right)}}$
Dwellings	0.1	2.57	2.76	1.07
	0.25	1.77	1.97	1.11
	0.5	2.45	2.35	0.96
	0.75	3.15	2.91	0.92
	0.9	4.20	3.43	0.82
Villages	0.1	5.52	6.65	1.21
	0.25	11.41	10.31	0.90
	0.5	12.29	11.69	0.95
	0.75	16.24	15.41	0.95
	0.9	13.31	18.34	1.38

To obtain more insight we run a simulation scenario which involves a larger sample size of $n = 100$ selected from populations of sizes $N = 1{,}000$ and $N = 10{,}000$. We draw Y and X from a bivariate log standard normal distribution with $\mu = 0$ and $\sigma = 1$. The variables Y and X are drawn such that the correlation between the variables is equal to 0.9. We again calculate the root mean squared error for a range of α values and show the relative efficiency of the expectile based approach in Figure 5.1. For better visual presentation we show a smoothed version of the relative efficiency. We notice a reduction in the root mean squared error for both cases $N = 1{,}000$ and $N = 10{,}000$. We may conclude that the expectiles can easily be fitted in unequal probability sampling and the relation between expectiles and the distribution function can be used numerically to calculate quantiles with increased efficiency. This efficiency gain holds for upper quantiles only, that is for α bounded away from zero. Note however that the sampling scheme is such that large values of Y are sampled with higher probability, reflecting that the sampling scheme aims to get more reliable estimates for the right hand side of the distribution function, i.e., for large quantiles. If we are

interested in small quantiles we should use a different samling scheme by giving individuals with small values of Y an increased inclusion probability. In this case the behavior shown in Figure 5.1 would be mirrored with respect to α.

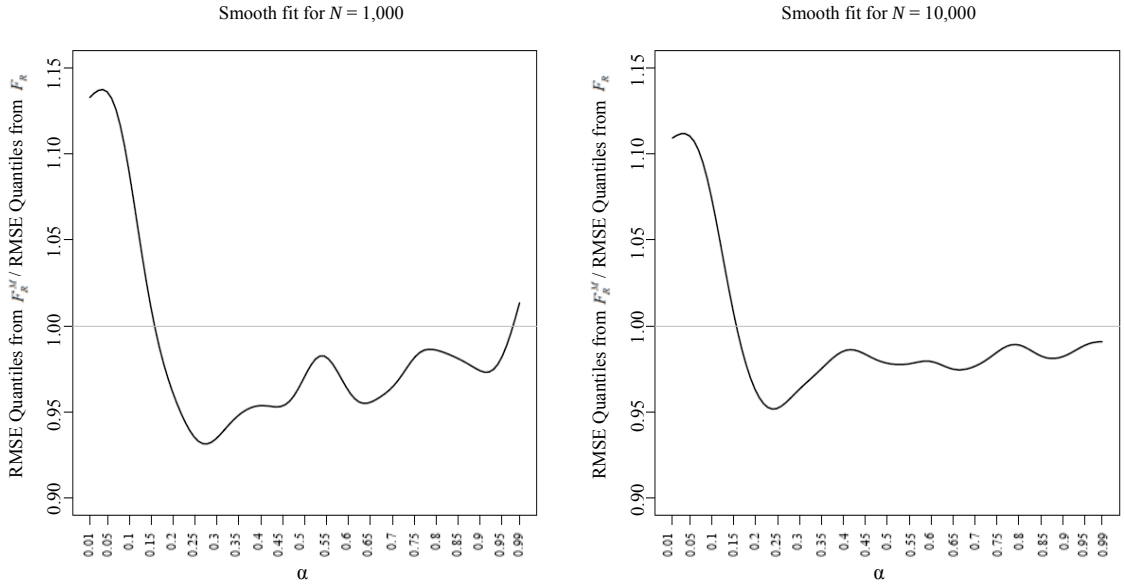

Figure 5.1 Relative Root Mean Squared Error (RMSE) of quantiles and quantiles from expectiles for the Probability Proportional to Size (PPS) design calculated from 500 repetitions (left: N = 1,000, right: N = 10,000).

6 Discussion

In Section 4 we extended the toolbox of expectiles to the estimation of distribution functions in the framework of unequal probability sampling. We defined expectiles for unequal probability samples. When comparing quantiles based on $\hat{F}_R$ with quantiles based on the expectile based estimator $\hat{F}_R^M$, we observed that the proposed estimator performs well in comparison to existing methods. The calculation of empirical expectiles is implemented in the open source software R (see R Core Team 2014) and can be found in the R-package expectreg by Sobotka, Schnabel, and Schulze Waltrup (2013). The calculation of the expectile based distribution function estimator $\hat{F}_N^M$ is also part of the R-package expectreg. The calculation of $\hat{F}_R^M$ is, however, more demanding as the calculation of $\hat{F}_R$ because it involves three steps: First, we calculate the weighted expectiles as described in Section 3; second, we estimate $\hat{F}_R^N$, and in a third step, we derive $\hat{F}_R^M$ from $\hat{F}_R^N$ (see Section 4). In the Log-Normal-Simulation it takes about 2-3 seconds for N = 1,000 to calculate $\hat{F}_R^M$ whereas the computational effort of $\hat{F}_R$ is barely noticeable.

Acknowledgements

Both authors acknowledge financial support provided by the Deutsche Forschungsgemeinschaft DFG (KA 1188/7-1).

References

Aigner, D.J., Amemiya, T. and Poirier, D.J. (1976). On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function. *International Economic Review*, 17(2), 377-396.

Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73(3), 597-604.

Chen, Q., Elliott, M.R. and Little, R.J.A. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Survey Methodology*, 36, 1, 23-34.

Chen, Q., Elliott, M.R. and Little, R.J.A. (2012). Bayesian inference for finite population quantiles from unequal probability samples. *Survey Methodology*, 38, 2, 203-214.

De Rossi, G., and Harvey, A. (2009). Quantiles, expectiles and splines. Nonparametric and robust methods in econometrics. *Journal of Econometrics*, 152(2), 179-185.

Guo, M., and Härdle, W. (2013). Simultaneous confidence bands for expectile functions. *AStA - Advances in Statistical Analysis*, 96(4), 517-541.

Hajek, J. (1971). Comment on "An essay on the logical foundations of survey sampling, part one". *The Foundations of Survey Sampling*, 236.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Jones, M. (1992). Estimating densities, quantiles, quantile densities and density quantiles. *Annals of the Institute of Statistical Mathematics*, 44(4), 721-727.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kneib, T. (2013). Beyond mean regression (with discussion and rejoinder). *Statistical Modelling*, 13(4), 275-385.

Koenker, R. (2005). *Quantile Regression, Econometric Society Monographs*. Cambridge: Cambridge University Press.

Kuk, A.Y.C. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika*, 75(1), 97-103.

Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimates. *Bulletin of the International Statistical Institute*, (33), 133-140.

Midzuno, H. (1952). On the sampling system with probability proportional to sum of size. *Annals of the Institute of Statistical Mathematics*, 3, 99-107.

Murthy, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

Newey, W.K., and Powell, J.L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55(4), 819-847.

Pratesi, M., Ranalli, M. and Salvati, N. (2009). Nonparametric M-quantile regression using penalised splines. *Journal of Nonparametric Statistics*, 21(3), 287-304.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao, J., and Wu, C. (2009). Empirical likelihood methods. *Handbook of Statistics*, 29B, 189-207.

Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77(2), 365-375.

Schnabel, S.K., and Eilers, P.H. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis*, 53(12), 4168-4177.

Schulze Waltrup, L., Sobotka, F., Kneib, T. and Kauermann, G. (2014). Expectile and quantile regression - David and Goliath? *Statistical Modelling*, 15, 433-456.

Sobotka, F., and Kneib, T. (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis*, 56(4), 755-767.

Sobotka, F., Schnabel, S. and Schulze Waltrup, L. (2013). *Expectreg: Expectile and Quantile Regression*. With contributions from P. Eilers, T. Kneib and G. Kauermann, R package version 0.38.

Tillé, Y., and Matei, A. (2015). *Sampling: Survey Sampling. R package, version 2.7*. https://cran.r-project.org/web/packages/sampling/index.html.

Yao, Q., and Tong, H. (1996). Asymmetric least squares regression estimation: A nonparametric approach. *Journal of Nonparametric Statistics*, 6(2-3), 273-292.

ADDENDUM

Model-assisted optimal allocation for planned domains using composite estimation

Wilford B. Molefe and Robert Graham Clark
Volume 41, number 2, (December 2015), 377-387

The second paragraph of page 378 of our paper reviews the 2012 paper of Choudhry, Rao and Hidiroglou. Our paragraph as worded implies a criticism of this paper which we did not intend, and we take this opportunity to correct and clarify our review. The CVs we referred to were in Table 5 of Choudhry et al. (2012), and the heading of this table clearly indicated that the CVs were of composite estimators, rather than being unspecified as we incorrectly stated. We also suggested that some CVs in this table were surprisingly high. This would be the case if the CVs (actually relative root mean squared errors, following a common convention) were calculated using the approximation of Longford (2006) or our closely related anticipated mean squared errors. However, Choudhry et al. (2012) used a different (and more standard) estimator of mean squared error, and the high values are not surprising in this light.

We also stated that Choudhry et al. (2012) did not investigate whether other designs such as power allocation can give lower values of Longford's criteria. This was correct, and motivated the research on this question in our paper. However we should have made clear that Choudhry et al. (2012) did consider square root allocation, a special case of power allocation, in terms of other criteria, such as setting small area CV tolerances.

References

Choudhry, G.H., Rao, J.N.K. and Hidiroglou, M.A. (2012). On sample allocation for efficient domain estimation. *Survey Methodology*, 38, 1, 23-29.

Longford, N.T. (2006). Sample size calculation for small-area estimation. *Survey Methodology*, 32, 1, 87-96.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 31, No. 4, 2015

All inquires about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 32, No. 1, 2016

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 44, No. 1, March/mars 2016

CONTENTS TABLE DES MATIÈRES

Volume 44, No. 2, June/juin 2016

GUIDELINES FOR MANUSCRIPTS

Authors are invited to submit their articles in English or French in electronic form to the Editor, (statcan.smj-rte.statcan@canada.ca). Before submitting the article, please examine a recent issue of *Survey Methodology* (Vol. 39, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word with MathType for the mathematical expressions. A pdf or paper copy may be required for formulas and figures.

1. Layout

1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The documents should be divided into numbered sections with suitable verbal titles.

1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.

3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.

3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered with arabic numerals on the right if they are to be referred to later. Use a two-level numbering system based on the section of the paper. For example, equation (4.2) is the second important equation in section 4.

3.4 Write fractions in the text using a solidus.

3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).

3.6 If possible, avoid using bold characters in formulae.

4. Figures and Tables

4.1 All figures and tables should be numbered with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables. Use a two-level numbering system based on the section of the paper. For example, table 3.1 is the first table in section 3.

4.2 A detailed textual description of figures may be required for accessibility purposes if the message conveyed by the image is not sufficiently explained in the text.

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, page 164).

5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.