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

Survey Methodology

The missing information principle – A paradigm for analysis of messy sample survey data

by Raymond L. Chambers

Release date: January 3, 2024

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 infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

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.

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, December 2023
Vol. 49, No. 2, pp. 219-256
Statistics Canada, Catalogue No. 12-001-X

The missing information principle – A paradigm for analysis of messy sample survey data

Raymond L. Chambers[1]

Abstract

Sample surveys, as a tool for policy development and evaluation and for scientific, social and economic research, have been employed for over a century. In that time, they have primarily served as tools for collecting data for enumerative purposes. Estimation of these characteristics has been typically based on weighting and repeated sampling, or design-based, inference. However, sample data have also been used for modelling the unobservable processes that gave rise to the finite population data. This type of use has been termed analytic, and often involves integrating the sample data with data from secondary sources.

Alternative approaches to inference in these situations, drawing inspiration from mainstream statistical modelling, have been strongly promoted. The principal focus of these alternatives has been on allowing for informative sampling. Modern survey sampling, though, is more focussed on situations where the sample data are in fact part of a more complex set of data sources all carrying relevant information about the process of interest. When an efficient modelling method such as maximum likelihood is preferred, the issue becomes one of how it should be modified to account for both complex sampling designs and multiple data sources. Here application of the Missing Information Principle provides a clear way forward.

In this paper I review how this principle has been applied to resolve so-called "messy" data analysis issues in sampling. I also discuss a scenario that is a consequence of the rapid growth in auxiliary data sources for survey data analysis. This is where sampled records from one accessible source or register are linked to records from another less accessible source, with values of the response variable of interest drawn from this second source, and where a key output is small area estimates for the response variable for domains defined on the first source.

Key Words: Maximum likelihood; Combined data; Informative sampling; Nondeterministic linkage; Small area estimation.

1. Introduction

1.1 Descriptive and analytical inference with multiple data sources

Over the last century, sample surveys have become the primary method by which data are collected for analysis of social and economic processes, and empirical analysis of survey data is often the way theories are developed and investigated. While there has been a large increase in recent years in the data available from registers and administrative and business systems, these data are often limited. The number of variables for which information is gathered is often small, the definition of the variables may not be what is required, the data may be out-of-date, they may be available only in aggregate form and coverage of the population may be limited. A survey can be used to collect information on many variables at the individual person or business level, using relevant definitions in a consistent manner. This allows great flexibility in the estimates produced and the analyses possible.

How the sample is defined in a sample survey can vary considerably. Probability-based samples based on complex designs that reflect the heterogeneity and complex structures of the population of interest

1. Raymond L. Chambers, University of Wollongong, NSW, 2522, Australia. E-mail: possumgong@gmail.com.

represent one extreme. These designs may use auxiliary data available concerning the population at the time of selection. At the other extreme are convenience samples (Galloway, 2005) whose relationship to the population they are supposed to represent is problematic. In many cases external information, both on the process of sample selection as well as the comparability of the selected sample with the target population in terms of some known characteristics, is available to the analyst. Traditionally this auxiliary information has been used to improve sample design and estimation. With the increasing amount of data available in administrative and business databases, the sample survey now has a significant new role in supplementing such data sources so that they can be fully exploited. In particular, we now regularly face situations where data from multiple data sources need to be integrated for inference.

In order to address this issue of integrated inference, I will distinguish between statistical analysis aimed at estimating the value of an observable population quantity (e.g., the population average value of a variable) and analysis aimed at summarising the relationship between population variables in terms of a statistical model (e.g., a regression model). In the former case it is clear that the value of the population quantity becomes more and more "known" as the sample size increases, with the value known precisely (at least in theory) when the population is completely enumerated. This type of analysis is referred to variously as enumerative, predictive, descriptive or finite population inference. In the latter case the model is an abstract concept, corresponding to an idealisation of how the values of the different variables in the model relate to one another across the entire population. The "true" model is never known precisely, irrespective of how large a sample is used in the survey. This type of analysis is referred to as analytic inference. It is usually carried out by fitting the assumed statistical model to the survey data, with the nature and strength of the population relationship then summarised from the estimated values of the model parameters (e.g., the estimated regression coefficients).

Unfortunately, two quite distinct modes of inference exist for these two cases. If the target is a finite population quantity (e.g., a population average) the inferential framework is based on repeated sampling of the population, i.e., the population values for the variable of interest are held fixed. This is often referred to as *design-based* inference. On the other hand, if the target is a parameter of a statistical model for the population of interest (e.g., a regression parameter), then inference is typically *model-based*, i.e., it is with respect to potential population values that could have arisen under the true model. Little (2012) has described this state of affairs as "inferential schizophrenia" since the distinction between a target of inference that corresponds to a finite population quantity (e.g., the small area mean of a variable Y) and one that is the parameter of a model for the finite population values (e.g., the model expectation of Y) is very blurred. Model-based prediction theory (both frequentist and Bayesian) overcome this by explicitly allowing for the impact of the sample design in model-based inference.

In this paper I will adopt the frequentist interpretation of this framework, focussing on the use of the Missing Information Principle (the MIP) for model-based analytic inference in "messy" data situations where data from multiple sources are available. Other Waksberg Award papers have also discussed model-based analytic inference from sample survey data (Scott, 2006; Rao, 2005; Pfeffermann, 2011), but none have zeroed in on the MIP and its use in the "messy" data that often arise in an integrated data context.

1.2 Weighting and complex sample designs

Most surveys use complex sample designs, reflecting and exploiting the heterogeneity of their target populations. The complexities introduced by the interaction of the survey design with these sources of heterogeneity are often difficult to handle using standard statistical methods. One particular issue that often arises is the role of sample selection probabilities in the analysis of the survey data. Complex sample designs typically result in unequal selection probabilities, one consequence of which is that the distributions observed in the sample can be very different from those in the population from which the sample was selected. There is confusion amongst survey users, as well as among non-survey statisticians (Gelman, 2007), about whether one should use selection probabilities in analysis, and if so, how this should be done in order to effectively "capture" the information about the sample design, and its effects, that they contain. A standard strategy is to weight for unequal selection probabilities when analysing survey data. The purpose of such weighting is to compensate for differences between the sample and population introduced by the sampling scheme (Pfeffermann, 1993). Using weighted summation of the sample data is attractive because it lends itself naturally to the estimation of linear parameters such as averages and totals, which are the primary objectives of many sample surveys, and also because linear estimators are very straightforward to build into survey estimation systems.

There are two main approaches to constructing weights. Often this is via the reciprocals of the selection probabilities. Such weights have a long history in descriptive surveys but may also be incorporated in model fitting, for example by pseudo-likelihood methods (Skinner, Holt and Smith, 1989). In these methods estimating equations that assume simple random sampling are modified to incorporate the survey weights. Second, weights may incorporate auxiliary information concerning the population, for example in post-stratification and regression estimation (Bethlehem and Keller, 1987; Chambers, 1996). In this case a multiple regression model for predicting survey variables by auxiliary variables is used to define the weights, with a widely used example being Generalized Regression (GREG) estimation (Särndal, Swensson and Wretman, 1992). A variant of this approach is calibration estimation (Deville and Särndal, 1992; Särndal, 2007), where weights are constructed so that they are close to the inverses of the selection probabilities while at the same time allowing the weighted estimates to agree with selected population moments of key auxiliary variables.

The other principal concern with analysis involving sample survey data is how to make efficient use of external or auxiliary information when carrying out this analysis. Often, auxiliary information available about the population from a variety of sources (census, administrative registers, other surveys) is used to produce benchmarks that are used to constrain the survey estimates. Benchmarks are values of population characteristics or external estimates of these characteristics that are more reliable than unconstrained estimates derived from the sample data. Although many population benchmarks are often available, well-known results on model over-fitting indicate that the number of benchmarks used in constraining survey weights should be limited to prevent instability of the resulting estimates. This leads to the conclusion that one should limit the number of calibration constraints imposed on the weights. The issue is particularly

important in multipurpose surveys, where the amount of related population information, and hence number of constraints, can be large (Bardsley and Chambers, 1984; Chambers, 1996). There are important practical advantages if the same weights are used for different estimates, and this can be achieved in model-based weighting if the same benchmarks are used for different survey variables. But this comes at a price. In particular, the resulting estimates can be inefficient, because the weights become very variable due to model over-parameterisation. Optimal methods for selecting the appropriate amount of external information to use in weighting have been discussed in the literature, but are limited to linear situations (Silva and Skinner, 1997; Clark and Chambers, 2008). Little is known about the extension to nonlinear situations or to the role of calibration information in analytic inference.

1.3 Analysis of complex survey data

Much attention has been devoted to the analysis of complex data over the last three decades. For example, in the 1990s the UK ESRC Research Programme on Analysis of Large and Complex Datasets focussed on the development of methods for the statistical modelling of the complex data collected in social science investigations. Since then, there has been a rise in interest in the theoretical foundations of inference based on sample survey data (Krieger and Pfeffermann, 1992; Breckling, Chambers, Dorfman, Tam and Welsh, 1994; Dorfman, Chambers and Wang, 2002; Little, 2003; Chambers and Skinner, 2003; Pfeffermann, 2011; Chambers, Steel, Wang and Welsh, 2012; Little, 2022). In particular, it is now accepted that statistical methods that assume that the distribution of the sample data and the distribution of the population data are identical generally lead to biased inference, since they take no account of either the complex sample design or the availability of auxiliary data.

There are three frameworks for frequentist inference that are generally used to deal with this problem.

- *Pseudo-likelihood*: This is a hybrid approach, with the unknown sufficient statistics in the population level likelihood estimating equations replaced by sample-weighted estimators (Kish and Frankel, 1974; Binder, 1983; Godambe and Thompson, 1986). The role of sample weights is therefore to adjust for differences between the sample distribution and the finite population distribution (Pfeffermann, 1993). Such weights usually have no connection with the variance structure of the data and so can lead to considerable inefficiency under the model. Furthermore, the weights used in practice are themselves adjusted, sometimes substantially, in order to integrate external population information (e.g., via calibration). However, weighted methods are very simple to implement and so are widely used.
- *Sample likelihood*: An explicit model for the distribution of the sample data, based on the use of Bayes theorem to integrate the population model and the sampling procedure, is used to develop a likelihood (Krieger and Pfeffermann, 1992; Pfeffermann, Krieger and Rinott, 1998; Pfeffermann and Sverchkov, 1999, 2003). This sample likelihood approach is typically more efficient than pseudo-likelihood. However, since it focuses on the distribution of the sample data as the basis

for inference, rather than on the population, it ignores non-sample information from auxiliary integrated data, making it less than fully efficient.

- *Maximum likelihood*: This is a fully efficient approach, where the auxiliary information and the sampling design are directly accounted for in the likelihood, itself defined by a joint model for the survey variables, the sampling and non-response processes and the auxiliary information. The basic methodology is set out in Breckling et al. (1994), while Chambers et al. (2012) provides a comprehensive development of maximum likelihood ideas in sample surveys. Adopting a maximum likelihood approach is conceptually appealing, but applying it to complex data requires care. Complex models are needed for the type of economic and social populations surveyed in practice, and the likelihood has to incorporate information about the sampling scheme and any auxiliary information about the population. A further difficulty arises in secondary analysis. Here the analyst does not always have access to information on how the data were obtained, and so suitable approximations need to be derived. Chambers, Dorfman and Wang (1998) consider likelihood-based analysis where sample design information is not provided, and in Section 5 of this paper I discuss the case where the analysis data set contains linked records but the analyst does not have access to the original data sets used in the linkage process.

1.4 Multiple surveys and auxiliary data

Extension of these approaches to multiple surveys and multiple auxiliary data sources is a relatively unexplored area of research, although the problem is well known, going back to the pioneering work of Patterson (1950) on composite estimation. Merkouris (2004) develops an integrated set of calibrated weights for use with the combined data from two independent surveys that measure the same variable of interest but use different types of auxiliary information. Elliott and Davis (2005) also consider the problem of weighting combined data on the same variable collected in two independent surveys, but allow one of these surveys to have a "higher quality" measurement process than the other, leading to a propensity-based adjustment to the original sample weights for the records from the second, "lower quality", survey. In contrast to both these approaches, which are aimed at re-weighting the combined data set, and hence concerned with marginal analysis of the same variable using this combined data set, Strauss, Carroll, Bortnick, Menkedick, and Schultz (2001) consider how one would go about modelling a joint distribution using the combined data from two independent surveys. In particular, these authors focus on the situation where each survey contributes a different variable to this joint distribution, but there exists a third, typically much smaller, survey with information on both that can be used to create an estimate of this joint distribution by combining the joint information in the small survey with the marginal information in the two larger surveys.

All of the references in the preceding paragraph address real issues that arise with integration of external auxiliary information with data from sample surveys. However, the approaches taken are problem specific and do not fit within a common inferential framework. When combined with efficient prediction, the MIP provides such a framework and can be used to develop solutions to a wide range of combining surveys

issues. In this paper I aim to show why this is the case in three important, but relatively straightforward, areas of application, and in doing so provide the reader with insight about a useful tool for tackling inference using what may be referred to as "messy" data.

1.5 Summary of the paper

This paper is meant to provide an overview of the MIP and its application, rather than a detailed methodological development. Consequently, I provide an informal definition of the MIP in Section 2 and in Section 3 I illustrate its use in combining data from two sources for the purpose of estimating a population regression relationship. In Section 4 I discuss the concept of informative sampling and use two simple examples to show how the MIP provides an appropriate framework for modelling sample data in this situation, while in Section 5 I show how the MIP can be used to suggest an efficient way of modelling linked data from two sources when the linkages can contain errors. At the close of this section I go on to show how these methods can then be used for small area estimation when records in different small area can be erroneously linked. Finally, in Section 6 I provide an overview and discussion of other applications where using the MIP has proved useful as well as some potential generalisations.

2. The missing information principle and its use

2.1 Messy data structures

What are "the data"? From a classical statistical perspective the answer to this question might be typically characterized as a transparent "window" on the population of interest. But the real world is messier. There are multiple sources of data with varying levels of aggregation, suggesting that a more accurate characterization is a distorted window on the target population, plus (perhaps) clearer windows on related populations. To illustrate this, consider some examples:

Example 1: Values of Y from register A and values of X from register B plus values of both variables from a sample of records taken from a linked version of the two registers. The aim is to use the sample data plus the data from the two registers to model the $Y - X$ relationship at register level (Imbens and Lancaster, 1994; Handcock, Rendall and Cheadle, 2005).

Example 2: Values of Y plus auxiliary variables X and C from survey A plus values of the same variable Y plus auxiliary variables Z and C from survey B. Estimates of the population total of Y based on a combined sample are required (Merkouris, 2004).

Example 3: Values of "accurately measured" variables Y and X from a small survey A and values of a "rough approximation" to X from a much larger survey B are available. This information is to be used to calculate small area estimates of Y (Elliott and Davis, 2005).

Example 4: The analyst wishes to fit a model relating variables Y, X and Z at a national level. She has access to values of variables Y and Z from a large national survey plus values of correlated variable X and the same variable Z from another, distinct, large national survey plus values of Y, X and Z from a small, non-representative, third survey (Strauss et al., 2001).

2.2 Using the Missing Information Principle to combine data sources

The examples in the previous subsection illustrate pooling of multiple data sources, and all present problems for the analyst. However, they can be tackled by application of the Missing Information Principle or MIP. In particular, suppose that we can identify a model for the distribution of Y in the target population, and this model is characterized by a parameter θ. Suppose further that the data that we have for estimating this parameter are a mix of individual Y -values, values of other, related, variables, summary statistics, metadata (e.g., data definitions), paradata (e.g., information about how the data were obtained, sample weights, auxiliary data for the target population), related data from other surveys and other populations and so on. Opposed to this reality, the data we'd like to have for likelihood inference are data that define an ideal "rectangular" dataset containing representative data for the target population and related populations.

A naïve approach in this situation is to assume that the population model for Y also applies to the sample values of this variable, and so the maximum likelihood estimate for θ can be calculated by maximising the sample contribution to the population likelihood. The corresponding "face value" maximum likelihood estimate for θ is generally incorrect since the sampling method underpinning the population model (typically simple random sampling) will not be the one underpinning the sample data. Furthermore, the available data includes data from other sources that also contain information about θ. The appropriate likelihood should therefore also take account of this information in order to arrive at the "full information" maximum likelihood estimate for θ.

The MIP provides a route for going from a simple likelihood analysis based on the ideal dataset to the correct likelihood analysis for the actual data that are available. In particular, it states that likelihood-based inference using a "messy" observed dataset $\mathbf{D}_s$ can be achieved by carrying out likelihood-based inference using a larger "ideal" dataset $\mathbf{D}_U$ with the likelihood estimating equations defined by $\mathbf{D}_U$ replaced by their expected values given $\mathbf{D}_s$. Note that it doesn't matter what $\mathbf{D}_U$ is here. The only requirements are that $\mathbf{D}_s$ (the data we have) is a subset of $\mathbf{D}_U$ (the data we would like to have), and that likelihood inference using $\mathbf{D}_U$ is straightforward. The MIP was first articulated by Orchard and Woodbury (1972) in the context of inference with missing data, and is closely related to the widely used EM algorithm (Dempster, Laird and Rubin, 1977). Its application to analysis of survey data was first described in Breckling et al. (1994). In our subsequent book, Chambers et al. (2012), we provide a comprehensive examination of how working within a MIP-based inferential framework leads to the maximum likelihood estimator (MLE) in a wide variety of messy data situations. In particular, the discussion in Sections 3 and 4 below summarises key aspects of this development by showing how the MIP can be used to fit simple population models to combined survey data.

In Section 5 I expand on this by showing how a difficult to compute MIP-based solution for one particular problem can be approximated by a much easier to compute MIP-based solution to a closely related problem.

In order to apply the MIP, we work with the population distribution of *all* the data available to the survey analyst. This can be illustrated by considering the simple scenario where there is a single survey sample with non-response, linked to an auxiliary variables dataset, with a single analysis variable Y. We use upper case to denote population quantities and lower case to denote sample quantities. Let $\mathbf{y}_{\text{resp}}$ denote the vector of survey respondents' values of Y and let $\mathbf{r}_s$ denote the vector of response indicators for the sampled population units. We use $\mathbf{S}_U$ to denote the vector of sample inclusion indicators for the surveyed population. The matrix of population values of the auxiliary variables, which can include cluster or PSU indicators, is denoted $\mathbf{Z}_U$. Let κ denote a vector of known population summary statistics. The available data are $\mathbf{D}_s = \{\mathbf{y}_{\text{resp}}, \mathbf{r}_s, \mathbf{S}_U, \mathbf{Z}_U, \kappa\}$. In addition to $\mathbf{y}_{\text{resp}}$, the quantities $\mathbf{r}_s, \mathbf{S}_U, \mathbf{Z}_U, \kappa$ potentially also contain information about θ. In contrast, the ideal "rectangular" data are $\mathbf{D}_U = \{\mathbf{Y}_U, \mathbf{R}_U, \mathbf{S}_U, \mathbf{Z}_U, \kappa\}$ with a density $f(\mathbf{D}_U; \Theta)$ that is straightforward to write down, and θ is then either a component of Θ or defined by a $1-1$ transformation of the components of Θ. In either case if we can compute the MLE for Θ, we can write down the MLE for θ. Note that the likelihood generated by $\mathbf{D}_U$ is much easier to write down if $\mathbf{R}_U$ or $\mathbf{S}_U$ (or both) are ancillary for θ given $\mathbf{Z}_U$ and κ. That is, the distribution of $\mathbf{Y}_U$ and that of $\mathbf{R}_U$ and $\mathbf{S}_U$ are mutually independent given $\mathbf{Z}_U$ and κ.

There are two basic quantities used in likelihood inference. They are the score function, i.e., the derivative with respect to θ of the logarithm of the likelihood function, and the information function, i.e., the negative of the derivative of the score function with respect to θ. The MLE is typically defined as a zero of the score function, while an estimate of the variance of MLE is the inverse of the information function evaluated at the MLE.

Let $\partial_x f$ denote a vector of first order partial derivatives with respect to x and let $\partial_{xx} f$ denote the matrix of second order partial derivatives with respect to x. Then the MIP corresponds to two identities, proofs of which are set out in Lemma 2.1 of Chambers et al. (2012).

The score identity: Provided the ideal data $\mathbf{D}_U$ include the available data $\mathbf{D}_s$, the available data score sc_s for the parameter Θ of the distribution of $\mathbf{D}_U$ is the conditional expectation, given these data, of the ideal data score sc_U for Θ, i.e.,

$$\text{sc}_s = E\{\partial_\Theta \log f(\mathbf{D}_U; \Theta) \,|\, \mathbf{D}_s\} = E_s(\text{sc}_U).$$

The information identity: The available data information info_s for Θ is the negative of the matrix of partial derivatives for the components of the available data score sc_s. This matrix can be written as the conditional expectation, given the available data, of the ideal data information info_U for Θ minus the corresponding conditional variance of the ideal data score sc_U, i.e.,

$$\text{info}_s = E\{-\partial_{\Theta\Theta} \log f(\mathbf{D}_U; \Theta) \,|\, \mathbf{D}_s\} - \text{Var}\{\partial_\Theta \log f(\mathbf{D}_U; \Theta) \,|\, \mathbf{D}_s\} = E_s(\text{info}_U) - \text{Var}_s(\text{sc}_U).$$

Note that the conditional expectations and variance in the score and information identities above are with respect to the distribution of the ideal data $\mathbf{D}_U$. Also, in many applications sc_s (and hence info_s) turns out to be a function of $\mathbf{D}_s^{\text{obs}} = \{\mathbf{y}_{\text{resp}}, \mathbf{r}_s, \mathbf{S}_U, \kappa\}$ rather than of $\mathbf{D}_s = \{\mathbf{y}_{\text{resp}}, \mathbf{r}_s, \mathbf{S}_U, \mathbf{Z}_U, \kappa\}$. In such cases it is not difficult to see that the score and information identities still hold, but with $\mathbf{D}_s$ replaced by $\mathbf{D}_s^{\text{obs}}$. See Result 2 in Chambers et al. (1998).

The MIP is sometimes taken as referring to the information identity only, since the conditional variance term in this identity corresponds to the loss of information about Θ due to observing the available data $\mathbf{D}_s$ and not the ideal data $\mathbf{D}_U$. In this paper I take the MIP as being defined by both identities. However, it is the score identity that I find the most useful when faced with a messy data situation since it leads to parameter estimates for a population level model. The information identity can be used to obtain uncertainty estimates by inverting the observed information, but these estimates can be obtained in a variety of other ways including direct differentiation of sc_s as well as via bootstrap simulation of the fitted population level model.

In effect, it is the score identity that specifies the MLE based on the available data, while it is the information identity that shows us how much information about the parameter of interest we actually have given the available data. This is not dissimilar to the way sc_U is used to define a pseudo-likelihood estimator while the "observed information" about this estimator in the available data is given by the inverse of its estimated design variance.

The use of the score identity in the MIP to obtain the MLEs given the available data is an example of the application of what may be termed the "Prediction Principle", which is based on the fact that the minimum mean squared error predictor of the value of an unobserved random variable (say Y) given the value of another random variable (say X) is $E(Y|X)$. This principle underpins the model-based approach to sampling inference, starting with the seminal contributions of Royall (1970) and Royall (1976). In the score identity Y corresponds to sc_U and X corresponds to $\mathbf{D}_s$. So the best predictor of the solution to $\text{sc}_U = 0$ is the solution to $\text{sc}_s = E\{\text{sc}_U | \mathbf{D}_s\} = 0$. Furthermore, since the ideal data score sc_U is a function of the sufficient statistics for Θ defined by $\mathbf{D}_U$, the score identity also tells us that when sc_U is a linear function of these sufficient statistics the best approximation to sc_U given the available data $\mathbf{D}_s$ is obtained by replacing these population sufficient statistics in sc_U by their expected values given $\mathbf{D}_s$.

3. Combining survey data and marginal population information – Comparing the MIP with calibrated weighting

3.1 Calibration weighting in surveys

Likelihood analysis based on the MIP is a general and powerful way of incorporating external information into inference. However, its usefulness depends on our ability to construct models that capture the dependence between the survey variables and this external information at some "ideal" level and that

also allow straightforward conditioning on the available data. Calibrated weighting is a widely used method of survey estimation that also allows external information to be used, typically in the form of calibration constraints that ensure the weighted survey data are capable of exactly reproducing known finite population quantities. In most cases, the population quantities of interest are totals associated with auxiliary variables and so we consider constraints of the form $\mathbf{w}_s'\mathbf{Z}_s = \mathbf{1}_U'\mathbf{Z}_U$ where $\mathbf{Z}_U$ is the matrix of the N population values of a set of survey variables with known population totals, $\mathbf{Z}_s$ is the corresponding matrix of the n sample values, $\mathbf{w}_s$ is a vector of sample weights and $\mathbf{1}_U$ is a unitary N-vector.

Deville and Särndal (1992) introduced the idea of using calibrated sample weights w_i that are closest to the expansion weights π_i^{-1} where π_i denotes the sample inclusion probability of population unit *i*. There are a variety of metrics for measuring closeness that can be used for this purpose, but the most popular is the chi square metric $Q = (\mathbf{w}_s - \boldsymbol{\pi}_s^{-1})'\,\Omega(\mathbf{w}_s - \boldsymbol{\pi}_s^{-1})$ where $\boldsymbol{\pi}_s^{-1}$ is the vector of expansion weights and Ω is a positive definite matrix chosen by the analyst to reflect heteroskedasticity in the population values of the survey variables. Minimising Q subject to calibration leads to weights

$$\mathbf{w}_s^{\text{cal}} = \boldsymbol{\pi}_s^{-1} + \Omega^{-1}\mathbf{Z}_s(\mathbf{Z}_s'\Omega^{-1}\mathbf{Z}_s)^{-1}(\mathbf{Z}_U'\mathbf{1}_U - \mathbf{Z}_s'\boldsymbol{\pi}_s^{-1}).$$

An alternative take on calibration is to view it as ensuring model-unbiased linear prediction of population totals (Valliant, Dorfman and Royall, 2000; Chambers and Clark, 2012). This is because weighting implies the use of a linear predictor $\mathbf{w}_s'\mathbf{y}_s$ for the population total $\mathbf{1}_U'\mathbf{Y}_U$, and if $\mathbf{Y}_U = \mathbf{Z}_U\beta + \mathbf{e}_U$ with $E(\mathbf{e}_U \,|\, \mathbf{Z}_U) = \mathbf{0}_U$, then under non-informative sampling given $\mathbf{Z}_U$ any set of weights that are calibrated on $\mathbf{Z}_U$ will also define an unbiased predictor of $\mathbf{1}_U'\mathbf{Y}_U$ under this linear model. So calibration is a good thing – provided the linear model assumption is valid.

3.2 Application to data from two sources

Consider the case where the population U is such that the values y_i and x_i of two scalar variables, Y and X are stored on separate registers, each of size N. A simple random sample s of n units from one register is linked to the other via a unique common identifier, thus defining n matched (y_i, x_i) pairs. Our aim is to use these linked sample data, plus auxiliary information corresponding to the population averages of Y and X from each register, to estimate the parameters α, β and σ^2 that characterise the population linear regression model $y_i = \alpha + \beta x_i + \sigma e_i$ where the errors e_i are distributed as independent and identically distributed (iid) Gaussian random variables with zero mean and unit variance.

The classical approach to fitting a population regression model like the one above given sample data is to use a pseudo-likelihood approach. See Kish and Frankel (1974), Binder (1983), Godambe and Thompson (1986) and Pfeffermann (1993). This is usually motivated as follows. Let $\mathbf{Y}_U$ and $\mathbf{X}_U$ denote the vectors of population values of Y and X, with $f(\mathbf{Y}_U \,|\, \mathbf{X}_U; \theta)$ denoting the conditional probability density of these population values. Then, if the pair $(\mathbf{Y}_U, \mathbf{X}_U)$ were to be observed, θ would be estimated as a solution to

$\text{sc}_U = \partial_\theta \log f(\mathbf{Y}_U \,|\, \mathbf{X}_U; \theta) = 0.$ But for any specified value of θ, sc_U defines a finite population parameter (the "census score") that we can estimate using the sample data and the sample weights, say by sc_w. The maximum pseudo-likelihood estimator (MPLE) of θ is then the solution to the estimating equation $\text{sc}_w = 0.$

Note that this approach does not specify how the sample weights should be constructed, only that $\text{sc}_w = 0$ defines a "design-consistent" estimator of sc_U for any permissible value of θ. In particular, calibration weights can be used. For the case described above it is clear that there are three calibration constraints, defined by the population size N, the population mean of X and the population mean of Y. Substituting $\mathbf{Z}_U = [\mathbf{1}_U \; \mathbf{Y}_U \; \mathbf{X}_U]$ and setting Ω equal to the identity matrix of order N, calibration weights that satisfy these three constraints are given by

$$w_s^{\text{cal}} = \frac{N}{n}\mathbf{1}_s + N[\mathbf{1}_s \; \mathbf{y}_s \; \mathbf{x}_s] \begin{bmatrix} \mathbf{1}_s'\mathbf{1}_s & \mathbf{y}_s'\mathbf{1}_s & \mathbf{x}_s'\mathbf{1}_s \\ \mathbf{1}_s'\mathbf{y}_s & \mathbf{y}_s'\mathbf{y}_s & \mathbf{x}_s'\mathbf{y}_s \\ \mathbf{1}_s'\mathbf{x}_s & \mathbf{y}_s'\mathbf{x}_s & \mathbf{x}_s'\mathbf{x}_s \end{bmatrix}^{-1} \begin{pmatrix} 0 \\ \bar{y}_U - \bar{y}_s \\ \bar{x}_U - \bar{x}_s \end{pmatrix}.$$

Put $\bar{x}_{ws}^{\text{cal}} = N^{-1}\sum_s w_i^{\text{cal}} x_i$ and $\bar{y}_{ws}^{\text{cal}} = N^{-1}\sum_s w_i^{\text{cal}} y_i.$ The corresponding calibrated MPLEs are then

$$\begin{aligned}
\hat{\beta}_{\text{CALmple}} &= \left(\sum_s w_i^{\text{cal}} x_i (x_i - \bar{x}_{ws}^{\text{cal}})\right)^{-1} \sum_s w_i^{\text{cal}} x_i (y_i - \bar{y}_{ws}^{\text{cal}}) \\
\hat{\alpha}_{\text{CALmple}} &= \bar{y}_{ws}^{\text{cal}} - \hat{\beta}_{\text{CALmple}} \bar{x}_{ws}^{\text{cal}} \\
\hat{\sigma}^2_{\text{CALmple}} &= N^{-1}\sum_s w_i^{\text{cal}} (y_i - \hat{\alpha}_{\text{CALmple}} - \hat{\beta}_{\text{CALmple}} x_i)^2.
\end{aligned}$$

The alternative to this hybrid calibration-MPLE approach is to use the MIP. To start, note that the face value MLEs (i.e., MLEs based on an assumption of simple random sampling and no auxiliary information) for α, β and σ^2 are

$$\begin{aligned}
\hat{\beta}_{\text{FVmle}} &= \frac{\sum_s (x_i - \bar{x}_s)(y_i - \bar{y}_s)}{\sum_s (x_i - \bar{x}_s)^2} \\
\hat{\alpha}_{\text{FVmle}} &= \bar{y}_s - \hat{\beta}_{\text{FVmle}} \bar{x}_s \\
\hat{\sigma}^2_{\text{FVmle}} &= n^{-1}\sum_s (y_i - \hat{\alpha}_{\text{FVmle}} - \hat{\beta}_{\text{FVmle}} x_i)^2.
\end{aligned}$$

However, there is extra information. In particular, we know the population means $\bar{y}_U$ and $\bar{x}_U$ of Y and X. So the face value MLEs are no longer full information MLEs. The latter can be computed using the MIP. Given the Gaussian assumption, the components of the ideal (i.e., population) data score function are

$$\begin{aligned}
\text{sc}_{1U} &= \frac{1}{\sigma^2}\sum_U (y_i - \alpha - \beta x_i) \\
\text{sc}_{2U} &= \frac{1}{\sigma^2}\sum_U x_i (y_i - \alpha - \beta x_i) \\
\text{sc}_{3U} &= \frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_U (y_i - \alpha - \beta x_i)^2.
\end{aligned}$$

Let a subscript of $U-s$ denote the non-sampled part of the population. The corresponding components of the available data score function are then obtained by replacing the components of the ideal data score function by their conditional expectations given the sample values of Y and X and the values $\bar{y}_{U-s}$ and $\bar{x}_{U-s}$ of the non-sample means of Y and X. In order to do this, note that for non-sampled unit i our Gaussian assumption for the error term in the population regression model, plus the fact of random sampling, allows us to write

$$\begin{pmatrix} y_i \\ \bar{y}_{U-s} \end{pmatrix} \Big| \mathbf{X}_{U-s} \sim N\left\{ \begin{pmatrix} \alpha+\beta x_i \\ \alpha+\beta\bar{x}_{U-s} \end{pmatrix}, \begin{bmatrix} \sigma^2 & (N-n)^{-1}\sigma^2 \\ (N-n)^{-1}\sigma^2 & (N-n)^{-1}\sigma^2 \end{bmatrix} \right\}$$

where $\mathbf{X}_{U-s}$ denotes the non-sampled values of X. It is straightforward to see that the conditional distribution of Y given X, $\bar{y}_{U-s}$ and $\bar{x}_{U-s}$ is then

$$y_i \,|\, x_i, \bar{x}_{U-s}, \bar{y}_{U-s} \sim N\left\{ \bar{y}_{U-s} + \beta(x_i - \bar{x}_{U-s}), \sigma^2\left(1 - \frac{1}{N-n}\right) \right\}$$

and so the MIP-based available data score function has components

$$\begin{aligned}
\mathrm{sc}_{1s} &= \frac{1}{\sigma^2}\left\{ \sum\nolimits_s (y_i - \alpha - \beta x_i) + (N-n)\left(\bar{y}_{U-s} - \alpha - \beta\bar{x}_{U-s}\right) \right\} \\
\mathrm{sc}_{2s} &= \frac{1}{\sigma^2}\left\{ \sum\nolimits_s x_i (y_i - \alpha - \beta x_i) + (N-n)\,\bar{x}_{U-s}\left(\bar{y}_{U-s} - \alpha - \beta\bar{x}_{U-s}\right) \right\} \\
\mathrm{sc}_{3s} &= -\frac{(n+1)}{2\sigma^2} + \frac{1}{2\sigma^4}\left\{ \sum\nolimits_s (y_i - \alpha - \beta x_i)^2 + (N-n)\left(\bar{y}_{U-s} - \alpha - \beta\bar{x}_{U-s}\right)^2 \right\}.
\end{aligned}$$

The MIP-based MLEs are obtained by setting these score components to zero and solving for α, β and σ^2. The solutions are

$$\hat{\beta}_{\mathrm{MIPmle}} = \frac{\sum_s (x_i - \bar{x}_s)(y_i - \bar{y}_s) + n\bar{x}_s(\bar{y}_s - \bar{y}_U) + (N-n)\,\bar{x}_{U-s}(\bar{y}_{U-s} - \bar{y}_U)}{\sum_s (x_i - \bar{x}_s)^2 + n\bar{x}_s(\bar{x}_s - \bar{x}_U) + (N-n)\,\bar{x}_{U-s}(\bar{x}_{U-s} - \bar{x}_U)}$$

$$\hat{\alpha}_{\mathrm{MIPmle}} = \bar{y}_U - \hat{\beta}_{\mathrm{MIPmle}}\bar{x}_U$$

and

$$\hat{\sigma}^2_{\mathrm{MIPmle}} = \frac{1}{n+1}\sum\nolimits_s \left(y_i - \hat{\alpha}_{\mathrm{MIPmle}} - \hat{\beta}_{\mathrm{MIPmle}} x_i \right)^2 + (N-n)\left(\bar{y}_{U-s} - \hat{\alpha}_{\mathrm{MIPmle}} - \hat{\beta}_{\mathrm{MIPmle}}\bar{x}_{U-s} \right)^2.$$

These are just the weighted least squares (WLS) estimators of these parameters based on an extended sample consisting of the data values in s (each with weight equal to one) plus an additional data value (with weight equal to $N-n$) defined by the known non-sample means $\bar{y}_{U-s}$ and $\bar{x}_{U-s}$. Standard WLS variance estimation methods can therefore be applied. Furthermore, these MIP-based MLEs depend only on the sample values of Y and X and on the population means of Y and X and not also on the individual values in the population vector $\mathbf{X}_U$ so they are also the available data MLEs when all one has is auxiliary summary information corresponding to the population means of Y and X.

Related results are reported in Handcock, Rendall and Cheadle (2005) who tackle the problem by maximising the face value likelihood generated by the sample values of Y and X subject to the constraint $\bar{y}_U = \hat{\alpha} + \hat{\beta}\bar{x}_U$. This leads to the estimators

$$\hat{\beta}_{\text{con}} = \frac{\sum_s (x_i - \bar{x}_s)(y_i - \bar{y}_s) + n(\bar{x}_s - \bar{x}_U)(\bar{y}_s - \bar{y}_U)}{\sum_s (x_i - \bar{x}_r)^2 + n(\bar{x}_s - \bar{x}_U)^2}$$

$$\hat{\alpha}_{\text{con}} = \bar{y}_U - \hat{\beta}_{\text{con}}\bar{x}_U$$

and

$$\hat{\sigma}^2_{\text{con}} = n^{-1}\sum_s \left(y_i - \hat{\alpha}_{\text{con}} - \hat{\beta}_{\text{con}}x_i\right)^2.$$

In general, the differences between these constraint-based estimators and the MIP-based MLEs defined earlier will be small.

3.3 Imprecise benchmarks

So far, the population benchmarks $\bar{y}_U$ and $\bar{x}_U$ have been assumed to be precise. However, this is not always true. For example, they could be estimated from survey data themselves, albeit from surveys with much larger samples, and so may in fact have errors. This can arise, for example, if census coverage is incomplete, and so census outputs are adjusted for coverage error. It can also be the case that we have access to estimates derived from another larger survey rather than census values for these benchmarks. As long as the error or imprecision of such estimation is small, our MIP-based MLEs above are still valid. Asymptotically, if the benchmark estimates $(\tilde{y}_U, \tilde{x}_U)$ for $(\bar{y}_U, \bar{x}_U)$ satisfy $\tilde{y}_U = \bar{y}_U + o_p(n^{-1/2})$ and $\tilde{x}_U = \bar{x}_U + o_p(n^{-1/2})$, and if they are used in place of $(\bar{y}_U, \bar{x}_U)$, then it is easily seen that the resulting estimators $(\tilde{\alpha}_{\text{MIPmle}}, \tilde{\beta}_{\text{MIPmle}})$ are asymptotically equivalent to $(\hat{\alpha}_{\text{MIPmle}}, \hat{\beta}_{\text{MIPmle}})$ apart from a negligible error of $o_p(n^{-1/2})$. However, this conclusion is not valid for $\tilde{\sigma}^2_{\text{MIPmle}}$ unless a generally higher order accuracy of $o_p(n^{1/2}/N)$ for the benchmark estimates is assumed.

Intuitively, one expects the extra information from knowing $(\bar{y}_U, \bar{x}_U)$ to contribute mainly to estimation of the intercept term α in the population regression model. To see that this is the case we write down the variances of $\hat{\beta}_{\text{MIPmle}}$ and $\hat{\alpha}_{\text{MIPmle}}$. This can be done by differentiating the available data score function components, changing signs and evaluating at these MLEs to get the observed information matrix for the regression model parameters. This matrix can then be inverted to get the (asymptotic) variances and covariances of these MLEs. Alternatively, exploiting their equivalence to a WLS fit, we can obtain the variances of $\hat{\alpha}_{\text{MIPmle}}$ and $\hat{\beta}_{\text{MIPmle}}$ directly. These are

$$\text{Var}(\hat{\alpha}_{\text{MIPmle}}) = n^{-1}\sigma^2\left(\frac{\bar{x}_s^{(2)} - (1 - nN^{-1})(\bar{x}_s^{(2)} - \bar{x}_{U-s}^2)}{\bar{x}_s^{(2)} - \bar{x}_{U-s}^2 + nN^{-1}(\bar{x}_{U-s}^2 - \bar{x}_U^2)}\right)$$

$$\text{Var}(\hat{\beta}_{\text{MIPmle}}) = \frac{n^{-1}\sigma^2}{\bar{x}_s^{(2)} - \bar{x}_{U-s}^2 + Nn^{-1}(\bar{x}_{U-s}^2 - \bar{x}_U^2)} = \frac{n^{-1}\sigma^2}{\bar{x}_s^{(2)} - \bar{x}_s^2 + N^{-1}(N-n)(\bar{x}_s - \bar{x}_{U-s})^2}.$$

Here $\bar{x}_s^{(2)}$ is the mean of the squares of the sample X-values. It can be shown that $\text{Var}(\hat{\beta}_{\text{MIPmle}}) \leq \text{Var}(\hat{\beta}_{\text{FVmle}})$, with equality only if $\bar{x}_s = \bar{x}_{U-s}$. Similarly $\text{Var}(\hat{\alpha}_{\text{MIPmle}}) \leq \text{Var}(\hat{\alpha}_{\text{FVmle}})$, with equality only if $\bar{x}_s^{(2)} = \bar{x}_s \bar{x}_{U-s}$, which is extremely unlikely in practice. This confirms our intuition above.

3.4 Comparing the efficiency of MIP with that of calibration for data integration

How much more efficient is using a MIP-based approach to data integration compared with a calibration-based approach? Some insight can be obtained from the results of a small model-based simulation study set out in Tables 3.1 and 3.2. Here population values were generated as $y_i = 5 + x_i + e_i$ with $x_i = e^{z_i}$ and z_i and e_i generated independently of one another as standard Gaussian. A total of 1,000 simulations were carried out for each scenario, corresponding to choice of N, n and the degree of imprecision in the benchmarks. Sampling in Table 3.1 was carried out using simple random sampling without replacement (SRSWOR) and three levels of imprecision in the population benchmarks were examined – no error in the benchmarks, benchmarks subject to census-level error (benchmark equal to true value plus a random error with standard deviation equal to the actual marginal standard deviation multiplied by $N^{-1/2}$) and benchmarks subject to larger survey error (benchmark equal to true value plus a random error with standard deviation equal to the actual marginal standard deviation multiplied by $(N/5)^{-1/2}$).

The values shown in Table 3.1 are relative efficiencies, defined as the ratio of the 5% trimmed RMSE of a reference estimator to the corresponding 5% trimmed RMSE of the estimator of interest, expressed as a percentage. Values over 100 therefore indicate superior relative efficiency for the alternative estimator. A trimmed RMSE was used to measure efficiency in order to avoid distortions caused by a small number of outlying error values generated in the simulations. The reference estimator in Table 3.1 is the face value MLE under SRSWOR. It is clear that the MIP-based MLEs perform well. In contrast, the MPLEs based on calibrated weights are consistently less efficient for all three parameters of interest, even when the benchmarks contain errors. It is only when the benchmark errors are relatively large that the efficiency of the MIP-based MLEs falls below that of the face value MLEs.

Table 3.2 shows the relative performances of the same estimators as in Table 3.1, but this time where an informative sampling method is used. In particular, the sample data here are selected with inclusion probabilities that are approximately proportional to their Y-values, and the reference estimation method is MPLE, with weights defined by inverse sample inclusion probabilities under probability proportional to Y (PPY) sampling. In contrast the calibrated MPLE is based on calibrated versions of these sample weights while the MIP-based MLE is the same as in Table 3.1, i.e., it based on an assumption of SRSWOR. This allows one to investigate the degree to which incorporation of auxiliary population information helps protect against bias induced by misspecification of the sampling method. The gains from using the MIP-based MLE, even under a misspecified sampling method, are very clear. In contrast, the MPLE based on calibration weights is much less efficient, even though it is based on essentially unbiased sampling weights.

Table 3.1
Linear population model under SRSWOR and population benchmarks of varying quality.

		$N = 500$ $n = 20$	$N = 1{,}000$ $n = 50$	$N = 5{,}000$ $n = 200$
		Benchmarks Known Precisely		
α	CALmple	103	127	143
	MIPmle	134	145	150
β	CALmple	81	90	96
	MIPmle	106	102	101
σ^2	CALmple	84	94	99
	MIPmle	102	100	100
		Benchmarks Subject to Census-level Error		
α	CALmple	84	101	112
	MIPmle	116	111	116
β	CALmple	73	89	96
	MIPmle	104	100	100
σ^2	CALmple	78	88	97
	MIPmle	103	101	100
		Benchmarks Subject to Larger Survey Error		
α	CALmple	64	71	75
	MIPmle	86	80	78
β	CALmple	71	84	93
	MIPmle	100	95	100
σ^2	CALmple	63	77	94
	MIPmle	99	94	99

Notes: Values shown are per cent relative efficiencies with respect to 5% trimmed RMSE of the face value MLE under SRSWOR. CALmple denotes the MPLE based on calibrated weights, while MIPmle denotes the MIP-based MLE.
MIP = Missing information principle; MLE = Maximum likelihood estimator; MPLE = Maximum pseudo-likelihood estimator; RMSE = Root mean square error; SRSWOR = Simple random sampling without replacement.

Table 3.2
Linear population model under PPY sampling and population benchmarks of varying quality.

		$N = 500$ $n = 20$	$N = 1{,}000$ $n = 50$	$N = 5{,}000$ $n = 200$
		Benchmarks Known Precisely		
α	CALmple	118	143	159
	MIPmle	201	210	222
β	CALmple	63	73	81
	MIPmle	109	110	117
σ^2	CALmple	78	89	91
	MIPmle	106	106	111
		Benchmarks Subject to Census-level Error		
α	CALmple	98	120	135
	MIPmle	136	139	152
β	CALmple	65	70	77
	MIPmle	107	112	121
σ^2	CALmple	77	82	90
	MIPmle	108	107	109
		Benchmarks Subject to Larger Survey Error		
α	CALmple	69	74	89
	MIPmle	84	76	82
β	CALmple	54	57	66
	MIPmle	103	107	117
σ^2	CALmple	62	71	87
	MIPmle	99	101	102

Notes: Values shown are per cent relative efficiencies with respect to 5% trimmed RMSE of expansion-weighted MPLE under PPY sampling. CALmple denotes the MPLE based on calibrated weights, while MIPmle denotes the MIP-based MLE.
MIP = Missing information principle; MLE = Maximum likelihood estimator; MPLE = Maximum pseudo-likelihood estimator; PPY = probability proportional to Y; RMSE = Root mean square error.

The results set out in Tables 3.1 and 3.2 provide some evidence for claiming that parameter estimation based on application of the MIP is more efficient, and sometimes considerably more efficient, than parameter estimation based on maximum pseudo-likelihood, particularly when this approach is based on calibrated weights. As one might expect, the MIP-based estimate of α benefits most from the auxiliary information. However there are non-negligible gains for MIP-based estimates of β and σ^2 as well.

Why is the use of calibration weights so inefficient here? One answer follows from taking a model-based perspective on calibration. Recollect that calibrated weighting implicitly assumes a linear model linking Y and the variables defining the calibration constraints. But one of those constraints involves Y, implying an over-parameterised model. It is known (Bardsley and Chambers, 1984) that such models lead to highly variable weights and inefficient inference.

3.5 Another example of the use of the MIP for analysis based on integrated data sources

Integration of information from external sources when analysing survey data can arise in many different ways, and using the MIP as a general-purpose tool for these situations can be beneficial. For example, Merkouris (2004) describes a situation where independent generalized regression (GREG) estimators of the population total of a variable Y based on data from multiple surveys need to be efficiently combined. The solution that is put forward in this paper is to form an efficiently weighted average of these different GREG estimators, where the efficient weights are based on a common auxiliary variable, say C, measured in the different surveys. But a MIP-based approach is also possible. To illustrate, suppose that there are just two surveys, say A and B, with survey A using calibrated weights based on constraints defined by auxiliary variables X and C and survey B using calibrated weights based on constraints defined by auxiliary variables Z and C. From a model-based perspective, the ideal data set would be where all three auxiliaries are measured in both surveys, in which case the data from both surveys could be stacked and values of Y fitted to a linear model with three covariates (X, C and Z). Parameters of this model can therefore be estimated using the MIP, with unknown values of Z in survey A replaced by their conditional expectations and unknown values of X in survey B replaced by their conditional expectations. The model-based regression estimator of the population total of Y using the combined data from both surveys is then just N times the fitted value of the three parameter model at the population means of X, C and Z. This is very similar to data fusion (Raessler, 2004).

4. Using the MIP under informative sampling

4.1 What do we mean by saying that a method of sampling is informative?

In Section 3 above I assumed that the method of sampling was simple random, so that $\mathbf{S}_U$ and $\mathbf{Y}_U$ are conditionally independent given $\mathbf{Z}_U$. This allowed the sample label set s to be treated as fixed since $\mathbf{S}_U$ is

then ancillary for the parameters of the ideal data model. If $\mathbf{S}_U$ is not ancillary, application of the MIP requires one to model the joint distribution of the ideal data and the outcome of the sampling process. This is specific to the method used to select the sample, and so it is impossible to provide general results. Instead, in this section I provide some insight into the use of the MIP under informative sampling by showing how two special cases of informative sampling impact on inference in the case of a very simple single parameter population distribution. These simple examples illustrate how using the MIP to integrate the information in $\mathbf{S}_U$ in these situations can substantially improve inference. Before doing this, however, it is useful to be a little clearer about what we mean when we say a method of sampling is informative.

Broadly speaking, sampling is informative if distributions of population and sample values of Y are different (Pfeffermann, 1993). However, after conditioning on a population auxiliary, the two distributions can be the same. Sampling is non-informative (informative) for inference about the distribution of Y *given some information* if the associated conditional probability of observing a particular value of Y given a random population draw is equal (not equal) to the same conditional probability given the value of a random sample draw. That is, informative/non-informative status depends on what is being conditioned on. In particular, suppose that we have complete response, so $\mathbf{r}_s$ contains no information and our sample values of Y are $\mathbf{y}_s$. This allows us to concentrate on the impact of conditioning on $\mathbf{S}_U$ and $\mathbf{Z}_U$. In the same way that the concepts of Missing Completely At Random, Missing At Random and Non-Ignorable Missingness are defined in the missing data literature (Rubin, 1976; Little and Rubin, 1987; Little, 2003), we can define

Completely Non-Informative Sampling: The distribution of $\mathbf{Y}_U$ is independent of $\mathbf{S}_U$ and $\mathbf{Z}_U$, so the marginal distribution of $\mathbf{y}_s$ contains all relevant information for θ.

Non-Informative Sampling Given $\mathbf{Z}_U$: The distribution of $\mathbf{Y}_U \left| \mathbf{Z}_U \right.$ is independent of that of $\mathbf{S}_U \left| \mathbf{Z}_U \right.$ (i.e., $\mathbf{S}_U$ is ancillary for θ given $\mathbf{Z}_U$), so we have the same parameters for distributions of $\mathbf{y}_s \left| \mathbf{Z}_U \right.$ and $\mathbf{Y}_U \left| \mathbf{Z}_U \right.$ and the parameter of interest θ depends on the parameters of joint distribution of $\mathbf{y}_s$ and $\mathbf{Z}_U$ (i.e., the parameters of distribution of $\mathbf{y}_s \left| \mathbf{Z}_U \right.$ and the parameters of the marginal distribution of $\mathbf{Z}_U$). Here we can ignore the sampling process in likelihood-based inference but *cannot throw away* $\mathbf{Z}_U$ *information*.

Informative Sampling: Here $\mathbf{Y}_U$, $\mathbf{S}_U$ and $\mathbf{Z}_U$ are jointly dependent and the parameter of interest θ can depend on all the parameters of the joint distribution of these quantities. An immediate consequence is that the conditional distributions of $\mathbf{Y}_U$ and $\mathbf{y}_s$ given the auxiliary information $\mathbf{Z}_U$ can be very different, and so inference about the parameters of $\mathbf{Y}_U \left| \mathbf{Z}_U \right.$ cannot just focus on the likelihood generated by the conditional distribution of $\mathbf{y}_s \left| \mathbf{Z}_U \right.$.

It should be clear from the above that informativeness of the sampling method depends on the auxiliary information available to the survey data analyst, and how much this information "explains" the outcome of the sampling process. For example, cluster and multi-stage sampling can be modelled when the auxiliary information includes indicators for the population groupings corresponding to sampling units at the different stages of sampling. A sampling method that is informative in one situation may not be informative in another. For example, even if the sampling mechanism is entirely determined by the auxiliary information,

this mechanism can be informative if we do not (or cannot) include it in our survey data, as often happens in secondary data analysis. Furthermore, even if $\mathbf{Z}_U$ is included in the available data, the sampling method can still be informative if it also depends on variables not included in $\mathbf{Z}_U$ that are correlated with those in $\mathbf{D}_U$. In this case, including the outcome $\mathbf{S}_U$ of the sampling mechanism as part of $\mathbf{D}_U$ requires us to specify the joint density of $\mathbf{Y}_U$ and $\mathbf{S}_U$ given $\mathbf{Z}_U$. From this we can determine the distribution of $\mathbf{y}_s$ and hence write down the likelihood for θ. The traditional approach to likelihood inference under informative sampling achieves this by directly specifying the distribution of $\mathbf{Y}_S$ given $\mathbf{Z}_U$, where S is a random subset of U with distribution determined by the outcome $\mathbf{S}_U$. An alternative is to use the MIP to specify the score and information functions directly.

4.2 Applying the MIP to size-biased and cut-off sampling

In order to illustrate the use of the MIP in this context, first note that a commonly used model for informative sampling is where sample inclusion depends directly on $\mathbf{Y}_U$. Two ways this can happen are when inclusion probabilities are functions of Y and where there is cut-off sampling on Y. The key ideas for dealing with these two situations are straightforwardly developed by assuming that the N population values of Y are independent and identically distributed draws from a single parameter exponential distribution with marginal density $f(y;\theta)=\theta\exp(-\theta y)$, allowing one to obtain explicit results for both cases. In what follows I therefore make this assumption, with the target of inference then being $\mu=E(Y)=\theta^{-1}$.

First, suppose that the sample of n units is selected using size-biased sampling with *known* inclusion probabilities,

$$\pi_i = \frac{n(y_i+\alpha z_i)}{N(\bar{y}_U+\alpha\bar{z}_U)}$$

but where α is unknown. Here z_i is a auxiliary "size" value associated with population unit i and there is complete response. It is easy to see that $(\pi_i N)\,\bar{y}_U+(\pi_i N\bar{z}_U-nz_i)\,\alpha=ny_i$, and so provided $n\geq 2$, values of $\bar{y}_U$ and α are deducible from the sample values of Y and their known inclusion probabilities. Consequently the available data are the sample Y values $\{y_i; i\in s\}$ and $\bar{y}_U$. Applying the MIP, it immediately follows that the available data score for θ is

$$\mathrm{sc}_s = E\left\{\sum\nolimits_U(\theta^{-1}-y_i)\;\middle|\;\{y_i;i\in s\},\bar{y}_U\right\}=N\left\{\theta^{-1}-E_s(\bar{y}_U)\right\}=N(\theta^{-1}-\bar{y}_U).$$

The MIP-based MLE for μ is then $\hat{\mu}_{\mathrm{MIPmle}}=\bar{y}_U$, i.e., the ideal data MLE. Similarly, the available data information for θ is the population information for this parameter, $N\theta^{-2}$, and, since $\mu=\theta^{-1}$, the estimated variance of $\hat{\mu}_{\mathrm{MIPmle}}$ is $N^{-1}\bar{y}_U^2$.

Next consider what happens under cut-off sampling. Again assume complete response and population values distributed as one parameter exponential, with mean μ the target of inference. But now suppose that

the vector of sample inclusion indicators is random, corresponding to all population units with $y_i > K$, for known K. Then

$$E(\bar{y}_U \,|\, \mathbf{y}_s, K) \;=\; N^{-1}\left\{n\bar{y}_s + (N-n)\,E(Y \,|\, Y \le K)\right\} \;=\; N^{-1}\left\{n\bar{y}_s + (N-n)\left(\theta^{-1} - Ke^{-\theta K}(1-e^{-\theta K})^{-1}\right)\right\}$$

and so the available data score for θ is

$$\text{sc}_s = n(\theta^{-1} - \bar{y}_s) + (N-n)\,(Ke^{-\theta K})\,(1-e^{-\theta K})^{-1}.$$

There is no analytic solution to setting this score function to zero, but it is easy to obtain numerically. If we let $\hat{\theta}_{\text{MIPmle}}$ denote this solution then the MIP-based MLE of μ is

$$\hat{\mu}_{\text{MIPmle}} \;=\; \bar{y}_s - (N-n)\,n^{-1}(Ke^{-K\hat{\theta}_{\text{MIPmle}}})\,(1-e^{-K\hat{\theta}_{\text{MIPmle}}})^{-1}.$$

Here it is easiest to obtain the available data information for θ by direct differentiation of the available data score for this parameter, i.e., as $\text{info}_s = n\theta^{-2} + (N-n)\,K^2e^{-\theta K}(1-e^{-\theta K})^{-2}$, and so a large sample estimate of the variance of $\hat{\mu}_{\text{MIPmle}}$ is $\hat{V}(\hat{\mu}_{\text{MIPmle}}) = \left(\theta^4 \text{info}_s \,\big|_{\theta=\hat{\theta}_{\text{MIPmle}}}\right)^{-1}$.

4.3 Maximum pseudo-likelihood under size-biased and cut-off sampling

Alternatively, we could adopt a maximum pseudo-likelihood approach for both sampling methods above. For the case of size-biased sampling the maximum pseudo-likelihood estimator (MPLE) of θ, obtained as the zero of the sample weighted estimate of the ideal data score, is the inverse of the Hájek estimator for the population mean of Y. It immediately follows that the MPLE $\hat{\mu}_{\text{MPLE}}$ of μ is this Hájek estimator. Clearly $\hat{\mu}_{\text{MPLE}}$ is suboptimal – we know the inclusion probabilities π_i, so we know $\bar{y}_U$, which is the ideal data MLE. However, $\hat{\mu}_{\text{MPLE}}$ is approximately unbiased in large populations since

$$E(\hat{\mu}_{\text{MPLE}}) \;=\; E\left\{E\left(\frac{\sum_U I_i y_i (y_i + \alpha z_i)^{-1}}{\sum_U I_i (y_i + \alpha z_i)^{-1}} \,\Big|\, \mathbf{Y}_U\right)\right\} \;\approx\; E\left(\frac{n(\bar{y}_U + \alpha\bar{z}_U)^{-1}\bar{y}_U}{n(\bar{y}_U + \alpha\bar{z}_U)^{-1}}\right) = E(\bar{y}_U).$$

What about the case of cut-off sampling? Since pseudo-likelihood depends essentially on design consistency for its validity, and since this is turn requires that all population units have a non-zero chance of sample inclusion, it is clear that there is no MPLE for μ under cut-off sampling.

4.4 Maximum sample likelihood under size-biased and cut-off sampling

The other well-known approach to inference under informative sampling is to maximize the sample likelihood. This is a model-based methodology (Krieger and Pfeffermann, 1992; Pfeffermann, Krieger and Rinott, 1998; Pfeffermann and Sverchkov, 2003) motivated by inferential methods for size-biased sampling that approximate the probability density f_s of the sample values making up $\mathbf{y}_s$ as a function of the probability density f_U of the population values making up $\mathbf{Y}_U$ and the sampling weights. In particular, Bayes Theorem is used to obtain the probability density of a randomly chosen *sample* value y_i as

$$f_s(y_i;\omega,\theta) = f_U(y_i \,|\, i \in s) \;=\; \frac{\Pr(i \in s \,|\, y_i;\omega)\, f_U(y_i;\theta)}{\Pr(i \in s;\omega,\theta)}$$

where ω is a nuisance parameter that characterizes the sample selection method. The estimator for the parameter θ of interest is then defined by maximising the "sample likelihood" for θ,

$$\mathrm{SL}(\theta, \omega; \mathbf{y}_s) = \prod_{i \in s} \frac{\Pr(i \in s \mid y_i; \omega)\, f_U(y_i; \theta)}{\Pr(i \in s; \omega, \theta)}$$

as a function of θ. Note that under this approach one needs to independently estimate the nuisance parameter ω.

Applying the sample likelihood approach to estimation of the mean of an exponential distribution under the size-biased sampling scheme above, we first note the large sample approximation

$$\log\{\mathrm{SL}(\theta; \mathbf{y}_s)\} \propto \log\left\{\prod_{i \in s} \frac{(y_i + \alpha z_i)}{(\theta^{-1} + \alpha \bar{z}_U)} \theta \exp(-\theta y_i)\right\} = n\log(\theta) - n\log(\theta^{-1} + \alpha \bar{z}_U) - \theta n \bar{y}_s + C$$

where C does not depend on θ. Differentiating with respect to θ leads to the sample likelihood score

$$\mathrm{sc}_{\mathrm{SL}} = n\theta^{-1}\left(1 + (1 + \alpha \bar{z}_U \theta)^{-1}\right) - n\bar{y}_s.$$

When $\alpha = 0$ we have an explicit solution to $\mathrm{sc}_{\mathrm{SL}} = 0$ given by $\theta = 2/\bar{y}_s$, implying a maximum sample likelihood estimator (MSLE) for μ of the form $\hat{\mu}_{\mathrm{MSLE}} = \bar{y}_s/2$. No explicit solution exists for $\mathrm{sc}_{\mathrm{SL}} = 0$ when $\alpha > 0$, so numerical methods need to be used. Also, when $\alpha = 0$, it is straightforward to show that in large samples $E(\bar{y}_s) \approx 2\mu$, so $\hat{\mu}_{\mathrm{MSLE}}$ is approximately unbiased.

A simple MSLE for μ also exists under cut-off sampling. Here $\Pr(i \in s) = \Pr(y_i > K) = \exp(-\theta K)$, so, up to an additive constant, the log of the sample likelihood for θ under cut-off sampling is $n\log(\theta) - \theta n(\bar{y}_s - K)$. It is easy to see that then the MSLE of μ is $\hat{\mu}_{\mathrm{MSLE}} = \bar{y}_s - K$. This estimator is unbiased.

4.5 Comparison of MIPmle, MPLE and MSLE

Once again, small scale simulation results help put some perspective on how much efficiency is lost by using pseudo-likelihood or sample likelihood instead of full information likelihood based on application of the MIP with data from a single parameter exponential population. Tables 4.1 and 4.2 show bias and RMSE for size-biased sampling with $\alpha > 0$ while Tables 4.3 and 4.4 show similar results for cut-off sampling. In all cases these results are based on 1,000 independent simulations. They show that a MIP-based MLE (MIPmle) is consistently preferable to estimation using maximum sample likelihood (MSLE) or maximum pseudo-likelihood (MPLE). Tables 4.1 and 4.2 also show that, as expected, maximum sample likelihood outperforms maximum pseudo-likelihood. These results are in line with what has been observed in other studies where using maximum pseudo-likelihood (by far the most prevalent method of parametric estimation with survey data) is inefficient (Dorfman, Chambers and Wang, 2002). Its only advantage would appear to be its simplicity and the widespread availability of software.

Table 4.1
Size-biased sampling from a single parameter exponential population of size $N = 5{,}000$ with $\theta = 1$ and with $n = 100$.

ρ	Bias			RMSE		
	MIPmle	MPLE	MSLE	MIPmle	MPLE	MSLE
0.05	-0.0006	0.0040	-0.0018	0.0145	0.1424	0.0950
0.25	-0.0002	0.0070	0.0009	0.0138	0.1147	0.0827
0.50	0.0004	0.0094	0.0017	0.0139	0.1186	0.0813
0.75	0.0004	0.0028	-0.0084	0.0140	0.1091	0.0763
0.95	-0.0002	0.0074	-0.0047	0.0145	0.1134	0.0713

Notes: Values of the auxiliary variable Z were generated as a single parameter exponential with $\theta = 1$ and with $\rho = \text{Cor}(Y, Y + \alpha Z)$ where $\alpha = \sqrt{\rho^{-2} - 1}$, so increasing (decreasing) correlation implied decreasing (increasing) α.
MIPmle = MIP-based MLE; MLE = Maximum likelihood estimator; MPLE = Maximum pseudo-likelihood estimator; MSLE = Maximum sample likelihood estimator; RMSE = Root mean square error.

Table 4.2
Same scenario as in Table 4.1 except that $\rho = 0.5$ and the impact of increasing n is shown.

n	Bias			RMSE		
	MIPmle	MPLE	MSLE	MIPmle	MPLE	MSLE
10	-0.0007	0.0611	-0.0139	0.0138	0.3457	0.2539
30	0.0002	0.0255	-0.0061	0.0147	0.1986	0.1448
100	0.0004	0.0094	0.0017	0.0139	0.1186	0.0813
300	-0.0005	-0.0090	-0.0090	0.0142	0.0650	0.0460
900	-0.0001	-0.0344	-0.0267	0.0144	0.0511	0.0371

Notes: MIPmle = MIP-based MLE; MLE = Maximum likelihood estimator; MPLE = Maximum pseudo-likelihood estimator; MSLE = Maximum sample likelihood estimator; RMSE = Root mean square error.

Table 4.3
Cut-off sampling from a single parameter exponential distribution of size $N = 5{,}000$ with $\theta = 1$ and with cut-off $K = 2$.

μ	$E(n)$	Bias		RMSE	
		MIPmle	MSLE	MIPmle	MSLE
0.4343	50	0.0108	-0.0016	0.0400	0.0646
0.5112	100	0.0016	-0.0031	0.0197	0.0515
0.7109	300	-0.0003	0.0020	0.0133	0.0393
1.0172	700	-0.0008	0.0002	0.0162	0.0387
1.6612	1,500	-0.0002	0.0010	0.0240	0.0424

Notes: The impact of increasing expected sample size is shown.
MIPmle = MIP-based MLE; MLE = Maximum likelihood estimator; MSLE = Maximum sample likelihood estimator; RMSE = Root mean square error.

Table 4.4
Same scenario as in Table 4.3 except that the cut-off K changes, with θ modified to ensure expected sample sizes are as shown.

K	$E(n)$	Bias		RMSE	
		MIPmle	MSLE	MIPmle	MSLE
5	50	0.0254	-0.0020	0.0925	0.1522
4	100	0.0038	-0.0025	0.0429	0.1045
3	300	-0.0003	-0.0041	0.0202	0.0607
2	700	0.0002	-0.0005	0.0160	0.0381
1	1,500	0.0002	-0.0004	0.0122	0.0210

Notes: The average value of θ is 1.0048.
MIPmle = MIP-based MLE; MLE = Maximum likelihood estimator; MSLE = Maximum sample likelihood estimator; RMSE = Root mean square error.

4.6 Other examples of the use of the MIP under informative sampling

There are other situations where sampling is informative because the sample design itself is informative. For example, size stratification (i.e., stratification based on a size variable Z correlated with Y) is informative when the size stratum boundaries are known, but the analyst does not have access to non-sampled population values of Z. This would often be the case in secondary analysis. For example, Dorfman, Chambers and Wang (2002) describe how the MIP can be used to approximate maximum likelihood estimates when Z and Y coincide. In a small-scale simulation study they show that using a MIP-based approach in this case leads to significant gains in efficiency compared with a maximum pseudo-likelihood approach using stratification weights. And even if Z and Y differ, it will usually be the case that they are highly correlated, in which case knowing that the non-sampled units within a stratum have values of Z that lie between known bounds provides the analyst with information that can be used to modify inference about the stratum mean of Y and hence its overall population mean.

It has already been noted that using a maximum pseudo-likelihood approach can be inefficient. However, a powerful argument for its use in the past has been that it is design consistent, and so robust to misspecification of the population distribution of Y. But this is usually relative to the use of a face value maximum likelihood approach, which ignores the information in the sample design and implicitly assumes simple random sampling. To illustrate, consider the following scenario, based on Examples 2 and 3 in Binder and Roberts (2003). Suppose that our assumed or working model for the population values of Y is that they are independently and identically distributed as Gaussian with mean μ and with variance σ^2. The sample design is stratified sampling based on an auxiliary size variable Z. In particular, there are two strata, with stratum 1 (low values of Z) sampled disproportionately less than stratum 2 (high values of Z). In this case the face value MLE of μ is the unweighted sample mean $\bar{y}_s$, ignoring the disproportionate stratification. However, the MPLE is the stratified sample mean $\bar{y}_{st} = N^{-1}(N_1\bar{y}_{s1} + N_2\bar{y}_{s2})$, where $\bar{y}_{sj}$ is the sample mean in stratum j.

Now suppose that the working Gaussian model of a common mean and variance is misspecified, and in reality it is the conditional distribution of Y given Z that is Gaussian, with $E(Y|Z) = \beta Z$ and $\mathrm{Var}(Y|Z) = \gamma^2 Z$. We refer to this as the "true" model below. The target parameter μ (the marginal mean of Y across the population) under this true model is then $\mu = E\left(E(Y|Z)\right) = \beta E(Z)$. When N is large it is reasonable to approximate it by $\beta E(\bar{z}_U) = \beta N^{-1}\left(N_1\eta_1 + N_2\eta_2\right)$ where η_1, η_2 are the means of Z in strata 1 and 2 respectively, with corresponding variances ω_1^2, ω_2^2. Consequently, under the true model, $E(\bar{y}_s) = \beta n^{-1}(n_1\eta_1 + n_2\eta_2)$ and $E(\bar{y}_{st}) = \beta N^{-1}(N_1\eta_1 + N_2\eta_2)$, while

$$\mathrm{Var}(\bar{y}_s) = n^{-2}\left(n_1\left(\gamma^2 + \omega_1^2\beta^2\right) + n_2\left(\gamma^2 + \omega_2^2\beta^2\right)\right)$$

and

$$\mathrm{Var}(\bar{y}_{st}) = N^{-2}\left(N_1^2 n_1^{-1}\left(\gamma^2 + \omega_1^2\beta^2\right) + N_2^2 n_2^{-1}\left(\gamma^2 + \omega_2^2\beta^2\right)\right).$$

Typically $\omega_1^2 < \omega_2^2$ so since $n_2/n < N_2/N$ we see that that $\bar{y}_{st}$ is unbiased and has smaller variance than $\bar{y}_s$ under the true model.

However, the face value MLE $\bar{y}_s$ makes no use of the available data on the size variable Z. At a minimum this corresponds to the sample values of Z, which then allows the true regression model to be identified, and β estimated by $\hat{\beta} = \bar{y}_s \bar{z}_s^{-1}$. Applying the MIP to this situation, and still using the working model to define the ideal data score, the MIP-based MLE is the solution of the available data score equation (here r denotes the set of non-sampled population units)

$$\sum_s (y_i - \mu) + \sum_r \left\{ E\left(y_i \,|\, z_i\right) - \mu \right\} \;=\; 0.$$

In order to proceed further, one needs to be more specific about what one knows about the non-sampled Z values. If this is just the stratum population sizes then a reasonable assumption is that the expected value and variance of Y vary between the strata, and $\bar{y}_{st}$ is the MIP-based MLE. If in addition one knows the size stratum boundaries then the approach discussed following Table 4.4 can be adopted. However, suppose that one also knows the population average $\bar{z}_U$ of Z. It is easy to see that the MIP-based MLE is then $\hat{\mu}_{\text{MIPmle}} = \hat{\beta}\bar{z}_U$. Furthermore, this MIP-based MLE is unbiased under the true model since $E(\hat{\mu}_{\text{MIPmle}}) = \beta N^{-1}(N_1\eta_1 + N_2\eta_2)$, with

$$\text{Var}\,(\hat{\mu}_{\text{MIPmle}}) \;=\; n^{-1}\gamma^2 E(\bar{z}_U^2/\bar{z}_s^2) + N^{-2}\beta^2\left(N_1\omega_1^2 + N_2\omega_2^2\right).$$

Also, since $E(\bar{z}_U^2/\bar{z}_s^2) < 1$ under the specified stratified sample design,

$$\text{Var}\,(\bar{y}_{st}) - \text{Var}\,(\hat{\mu}_{\text{MIPmle}}) \;>\; \gamma^2 \sum_{j=1}^{2} n_j^{-1}\left(\frac{N_j^2}{N^2} - \frac{n_j^2}{n^2}\right) \;+\; \beta^2 \sum_{j=1}^{2} n_j^{-1}\omega_j^2\left(\frac{N_j^2}{N^2}\right)\left(1 - \frac{n_j}{N_j}\right).$$

The expression on right hand side above will typically be positive. This is supported by the small-scale simulation results shown in Table 4.5. That is, although the face value MLE is biased and inefficient under the true model, the MIP-based MLE that takes the information on Z into account is unbiased under this model and usually more efficient than the MPLE.

Table 4.5
Simulation results for 1,000 independent repetitions of $N = 1{,}000$, $n = 100$, Z distributed as single parameter exponential with mean 4, $\beta = 1$, $\gamma = 0.1$ and a sample design with two strata defined by values below/above the population mean of Z.

n_1	n_2	Bias			RMSE		
		$\bar{y}_s$	$\bar{y}_{st}$	$\hat{\mu}_{\text{MIPmle}}$	$\bar{y}_s$	$\bar{y}_{st}$	$\hat{\mu}_{\text{MIPmle}}$
10	90	3.3461	-0.0036	-0.0033	3.3745	0.2843	0.1321
25	75	2.3982	-0.0069	-0.0059	2.4319	0.2494	0.1280
50	50	0.8554	0.0138	0.0062	0.9167	0.2546	0.1232
propn allocation		0.0017	0.0021	0.0011	0.2698	0.2700	0.1333
75	25	-0.7444	0.0044	-0.0047	0.7825	0.3220	0.1313
90	10	-1.6860	0.0287	-0.0044	1.6957	0.4878	0.1294

Notes: MIP = Missing information principle; MIPmle = MIP-based MLE; MLE = Maximum likelihood estimator; RMSE = Root mean square error.

5. Modelling using non-deterministically linked data

5.1 Using the MIP when there is data linkage error

Data linkage is the joining of two or more administrative or survey datasets using statistical matching (ADRN, 2012). A key feature of many linkage applications is a clear separation between the linkage process and subsequent analysis of the linked data. Typically, this separation is for reasons of confidentiality, in the sense that the linking agencies often use confidential data in their record matching. See Harron, Goldstein and Dibben (2016). The linked data set that eventuates is then made available to analysts but the information used in the linking is not. This is a *secondary analysis* situation.

My focus here is on the bias in this analysis due to incorrect links, i.e., the bias that may arise if the linkage is not accurate enough (Lahiri and Larsen, 2005; Chambers, 2009; Kim and Chambers, 2012). Many of the linked data sets that are created are based on non-deterministic linking, where there is uncertainty about whether the data values in the linked record are actually for the same population unit. I also focus on the simplest scenario, where two population registers, denoted $\mathscr{y}$ and $\mathscr{X}$, are linked, and the analyst, who has full access to the $\mathscr{X}$-register, is provided with a sample of the records from the linked register. In particular, the $\mathscr{y}$-register contains values of a scalar random variable Y and the $\mathscr{X}$-register contains values of vector random variable X.

Suppose that we are interested in modelling the conditional distribution of Y given X. This is straightforward given a random sample of correctly linked (Y, X) values. But, we do not have such a sample. Instead we have a sample of linked values (Y^*, X) where $Y^* = Y$ if the linkage is correct, but possibly not if it is incorrect. I say "possibly" here because depending on the scale of measurement of Y, we can have $Y^* = Y$ even if the linkage is incorrect.

To proceed further I introduce a set of sandbox assumptions that simplify further analysis. They are unlikely to hold in practice, but serve to define a useful "working model" for inference. These are

- Both registers contain N records, with no duplications, and linkage is $1-1$ and complete. That is, all records in both registers are linkable, and no record in one register can be (eventually) linked to more than one record in the other register;
- There is a categorical "blocking" variable B recorded on both registers, measured without error on both, and taking Q distinct values $q = 1, 2, \ldots, Q$ such that all matching takes place within a block, i.e., records in both registers with the same value of B;
- The records on the linked data set are indexed in exactly the same way as they are indexed on the $\mathscr{X}$-register.

Suppose there are M_q records in each register with $B = q$ (so $N = \sum_q M_q$). Our second assumption above then constrains linkage errors to only occur within "blocks". Let $\mathbf{y}_q$ and $\mathbf{y}_q^*$ denote the original and linked values of Y in block q. Under $1-1$ and complete linkage it immediately follows that $\mathbf{y}_q^* = \mathbf{A}_q \mathbf{y}_q$ where $\mathbf{A}_q$ is an unknown random permutation matrix of order M_q, i.e., entries of $\mathbf{A}_q$ are either zero or one,

with a value of one occurring just once in each row and column. Let $\mathbf{X}$ denote the matrix of X values in the $\mathscr{X}$ register. Put $E(\mathbf{A}_q \,|\, \mathbf{X}) = \mathbf{T}_q$ and suppose that we have non-informative linkage given $\mathbf{X}$. That is, $\mathbf{A}_q$ is independent of $\mathbf{y}_q$ given $\mathbf{X}$ and so

$$E(\mathbf{y}_q^* \,|\, \mathbf{X}) \;=\; \mathbf{T}_q E(\mathbf{y}_q \,|\, \mathbf{X}; \theta)$$

where θ denotes the vector of parameters of the conditional distribution of Y given $\mathbf{X}$. These parameters are our primary target of inference.

An efficient linkage process should ensure that correct linkages within a block are more likely than incorrect linkages. We therefore impose the restrictive but practically useful assumption that linkage errors are exchangeable within a block, i.e., the probability of a record being correctly linked in block q is λ_q while the probability that it is incorrectly linked is η_q. We refer to this as an Exchangeable Linkage Errors or ELE model. Under $1-1$ and complete linkage the ELE model implies $\lambda_q + (M_q - 1)\,\eta_q = 1$. It immediately follows that $\mathbf{T}_q = (\lambda_q - \eta_q)\,\mathbf{I}_q + \eta_q \mathbf{1}_q \mathbf{1}_q'$, where $\mathbf{I}_q$ is the identity matrix of order M_q and $\mathbf{1}_q$ is the unitary vector of length M_q.

However, we do not have access to the full linked register. Instead we have a random sample s of n records from this linked register. We extend the idea of non-informative linking by assuming that the random processes underpinning sample selection and linking are mutually independent and that these processes are both non-informative for the parameters of the distribution of Y given X. This ensures that linkage of a sample to a register is stochastically equivalent to sampling from a completely linked register. This register can be partitioned into Q blocks with block q itself partitioned into m_q sampled values followed by $M_q - m_q$ non-sampled (and hence unobserved) linked values. Following standard practice, we use subscripts of s and r to denote a partition into sampled and non-sampled values. Consequently $\mathbf{A}_{sq}$ is the matrix defined by those rows of $\mathbf{A}_q$ that correspond to sampled units, with $\mathbf{A}_{ssq}$ denoting those columns of $\mathbf{A}_{sq}$ that correspond to sampled units, and so on. We can then write the vector of sampled linked values in block q as $\mathbf{y}_{sq}^* = \mathbf{A}_{sq}\mathbf{y}_q$, with $\mathbf{T}_{sq} = E(\mathbf{A}_{sq} \,|\, \mathbf{X}_q) = [\mathbf{T}_{ssq}\ \mathbf{T}_{srq}]$ where $\mathbf{T}_{ssq} = (\lambda_q - \eta_q)\,\mathbf{I}_{sq} + \eta_q \mathbf{1}_{sq}\mathbf{1}_{sq}'$ and $\mathbf{T}_{srq} = \eta_q \mathbf{1}_{sq}\mathbf{1}_{rq}'$.

In order to carry out a maximum likelihood analysis in this situation we need to specify a model for the conditional distribution of Y given X. We assume a simple linear model with Gaussian errors for this purpose, i.e., we put $\mathbf{y} \,|\, \mathbf{X} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$ where $\mathbf{I}$ is the identity matrix of order N. However, our data consist of a sample of values of Y from linked records. In order to apply the MIP in this case, we need to specify the conditional distribution of the correctly linked population values of Y given these linked values. Now $E(\mathbf{y}_q \,|\, \mathbf{X}) = \mathbf{X}_q\beta$ and $\mathrm{Var}(\mathbf{y}_q \,|\, \mathbf{X}) = \sigma^2\mathbf{I}_q$, while under ELE $E(\mathbf{y}_{sq}^* \,|\, \mathbf{X}) = \mathbf{T}_{sq}\mathbf{X}_q\beta$ and $\mathrm{Var}(\mathbf{y}_{sq}^* \,|\, \mathbf{X}) = \Sigma_{sq}^* = \sigma^2\mathbf{I}_{sq} + \mathbf{V}_{sq}$. Here $\mathbf{V}_{sq}$ is the sample component of $\mathbf{V}_q = \mathrm{Var}(\mathbf{A}_q\mathbf{X}_q\beta \,|\, \mathbf{X})$ and represents the increased heterogeneity in $\mathbf{y}_{sq}^*$ caused by incorrect linkage. Chambers (2009) shows that $\mathbf{V}_q$ is well approximated by the diagonal matrix with i^{th} diagonal term $(1-\lambda_q)\left(\lambda_q(f_i - \bar{f}_q)^2 + \bar{f}_q^{(2)} - (\bar{f}_q)^2\right)$ where $f_i = \mathbf{x}_i'\beta$ and $\bar{f}_q^{(2)}$ is the average of the f_i^2 in block q. Finally, we note that

$$\mathrm{Cov}(\mathbf{y}_q, \mathbf{y}_{sq}^* \,|\, \mathbf{X}) \;=\; \mathrm{Cov}(\mathbf{y}_q, \mathbf{A}_{sq}\mathbf{y}_q \,|\, \mathbf{X}) \;=\; \mathrm{Cov}(\mathbf{y}_q, \mathbf{y}_q \,|\, \mathbf{X})\,\mathbf{T}_{sq}' \;=\; \sigma^2\mathbf{T}_{sq}'.$$

It is tempting to conclude from this (as I have done in the past) that the joint distribution of $\mathbf{y}_q$ and $\mathbf{y}^*_{sq}$ given $\mathbf{X}$ is then multivariate Gaussian with these moments. However, as pointed out by Zhang and Tuoto (2021), since the support of $\mathbf{y}^*_{sq}$ is just $\mathbf{y}_q$ this clearly cannot be true. However, if we are prepared to *approximate* this joint distribution by a Gaussian copula with the same first and second moments, then the MIP can be used to construct a corresponding approximation to the MLEs for the parameters β and σ^2 of the conditional distribution of $\mathbf{y}_q$ given $\mathbf{y}^*_{sq}$. This argument turns out to be surprisingly fruitful.

Put $\mathbf{D}_{sq} = \mathbf{T}'_{sq}(\sigma^2\mathbf{I}_{sq} + \mathbf{V}_{sq})^{-1}$. Then

$$E(\mathbf{y}_q \,|\, \mathbf{y}^*_{sq}, \mathbf{X}) \;=\; \mathbf{X}_q\beta + \sigma^2\mathbf{D}_{sq}(\mathbf{y}^*_{sq} - \mathbf{T}_{sq}\mathbf{X}_q\beta) \;=\; \mathbf{a}_{sq}$$

and

$$\mathrm{Var}(\mathbf{y}_q \,|\, \mathbf{y}^*_{sq}, \mathbf{X}) \;=\; \sigma^2\mathbf{I}_q - \sigma^4\mathbf{D}_{sq}\mathbf{T}_{sq} \;=\; \mathbf{B}_{sq}.$$

That is, we can write $\mathbf{y}_q \,|\, \mathbf{y}^*_{sq}, \mathbf{X} \sim \mathbf{a}_{sq} + \mathbf{B}^{1/2}_{sq}\mathbf{g}_q$ where $\mathbf{g}_q \sim N(\mathbf{0}_q, \mathbf{I}_q)$. Next since $\mathbf{y}^*_{sq} = \mathbf{A}_{sq}\mathbf{y}_q$, the ideal data in block q is the set $\{\mathbf{y}_q, \mathbf{X}_q\}$ while the available data is the set $\{\mathbf{y}^*_{sq}, \mathbf{X}_q\}$, when we treat $\mathbf{A}_q$, and hence $\mathbf{A}_{sq}$, as ancillary. In order to use the MIP we therefore first note that since $\mathbf{y}_q \sim N(\mathbf{X}_q\beta, \sigma^2\mathbf{I}_q)$, the score functions for β and σ^2 based on the ideal data are

$$\mathrm{sc}_U(\beta) \;=\; \sigma^{-2}\sum_{q=1}^{Q}\mathbf{X}'_q(\mathbf{y}_q - \mathbf{X}_q\beta)$$

$$\mathrm{sc}_U(\sigma^2_e) \;=\; \sigma^{-4}\sum_{q=1}^{Q}(\mathbf{y}_q - \mathbf{X}_q\beta)'(\mathbf{y}_q - \mathbf{X}_q\beta) - N\sigma^{-2} \;=\; \sigma^{-4}\sum_{q=1}^{Q}\left\{\mathbf{y}'_q\mathbf{y}_q - 2\beta'\mathbf{X}'_q\mathbf{y}_q + \beta'\mathbf{X}'_q\mathbf{X}_q\beta\right\} - N\sigma^2$$

so the MIP-based score function for β using the available data is

$$\mathrm{sc}_s(\beta) \;=\; \sigma^{-2}\sum_{q=1}^{Q}\mathbf{X}'_q\left(E(\mathbf{y}_q \,|\, \mathbf{y}^*_{sq}, \mathbf{X}) - \mathbf{X}_q\beta\right) \;=\; \sum_{q=1}^{Q}\mathbf{X}'_q\mathbf{D}_{sq}\left(\mathbf{y}^*_{sq} - \mathbf{T}_{sq}\mathbf{X}_q\beta\right).$$

In order to define the corresponding MIP-based score for σ^2 we first note that

$$E(\mathbf{y}'_q\mathbf{y}_q \,|\, \mathbf{y}^*_{sq}, \mathbf{X}) \;=\; E\left(\left(\mathbf{a}_{sq} + \mathbf{B}^{1/2}_{sq}\mathbf{g}_q\right)'\left(\mathbf{a}_{sq} + \mathbf{B}^{1/2}_{sq}\mathbf{g}_q\right)\right) \;=\; \mathbf{a}'_{sq}\mathbf{a}_{sq} + \mathrm{tr}(\mathbf{B}_{sq})$$

and

$$E(\beta'\mathbf{X}'_q\mathbf{y}_q \,|\, \mathbf{y}^*_{sq}, \mathbf{X}) \;=\; \beta'\mathbf{X}'_q\mathbf{a}_{sq}.$$

This leads to a MIP-based available data score for σ^2 of the form

$$\begin{aligned}\mathrm{sc}_s\left(\sigma^2\right) \;&= \sigma^{-4}\sum_{q=1}^{Q}\left\{\left(\mathbf{a}_{sq} - \mathbf{X}_q\beta\right)'\left(\mathbf{a}_{sq} - \mathbf{X}_q\beta\right) + \mathrm{tr}(\mathbf{B}_{sq})\right\} - N\sigma^2 \\ &= \sum_{q=1}^{Q}\left\{\left(\mathbf{y}^*_{sq} - \mathbf{T}_{sq}\mathbf{X}_q\beta\right)'\mathbf{D}'_{sq}\mathbf{D}_{sq}\left(\mathbf{y}^*_{sq} - \mathbf{T}_{sq}\mathbf{X}_q\beta\right) + \mathrm{tr}\left(\mathbf{B}_{sq}\right)\right\} - N\sigma^2.\end{aligned}$$

Formal representations for the resulting estimators of β and σ^2 are obtained by setting these available data scores to zero and solving for these parameters. This leads to

$$\hat{\beta}_{\text{MIPmle}} = \left[\sum_{q=1}^{Q}\mathbf{X}_q'\mathbf{D}_{sq}\mathbf{T}_{sq}\mathbf{X}_q\right]^{-1}\left[\sum_{q=1}^{Q}\mathbf{X}_q'\mathbf{D}_{sq}\mathbf{y}_{sq}^*\right]$$

and

$$\hat{\sigma}^2_{\text{MIPmle}} = N^{-1}\sum_{q=1}^{Q}\left\{\left(\mathbf{y}_{sq}^*-\mathbf{T}_{sq}\mathbf{X}_q\beta\right)'\mathbf{D}_{sq}'\mathbf{D}_{sq}\left(\mathbf{y}_{sq}^*-\mathbf{T}_{sq}\mathbf{X}_q\beta\right)+\text{tr}\left(\mathbf{B}_{sq}\right)\right\}.$$

Since $\mathbf{D}_{sq}$ (and hence $\mathbf{B}_{sq}$) is a function of β and σ^2, the above estimators are computed iteratively. They also require one to know (or at least have a good estimate of) the correct linkage probabilities in each block. This issue is discussed in more detail in Chambers and Diniz da Silva (2019) and highlights the importance of the simultaneous release of paradata about the linking process when linked data are released for secondary analysis. An important practical point that also needs to be made here is that the block size M_q will usually be very large, making computation of block-dimensioned quantities like $\mathbf{D}_{sq}$ and $\mathbf{B}_{sq}$ time consuming. So in the development below I introduce a further approximation, replacing $\mathbf{y}_q$ by $\mathbf{y}_{sq}$ in the ideal data.

5.2 Application to small area estimation using non-deterministically linked data

Probably the most common application of model-based ideas in survey sampling is small area estimation or SAE. That is, where the sampled population is partitioned into D non-overlapping domains such that each domain is represented in the sample, but where the domain sample sizes are small, and sometimes even zero. It is standard to refer to these domains then as "small areas", where "small" is actually a reference to the domain sample size. See Rao and Molina (2015) for a comprehensive discussion of methods that have been proposed for estimation of domain-specific quantities in this situation, with the most common target being the domain average of a variable Y measured on the sampled population units.

Here I focus on the special case where Y is not measured directly on the sample but is obtained by linking the sample frame to another population register and then integrating the data from this register with the data directly obtained from the sampled units. This type of sample data acquisition is now increasingly common. Analysis variables in this integrated data set can exhibit increased heteroskedasticity (compared with an ideal data set where linkage is perfect) when records are incorrectly linked. This has the potential to lead to biased small area inference. See Briscolini, Di Consiglio, Liseo, Tancredi and Tuoto (2018).

In order to show how MIP-based ideas coupled with an ELE linkage errors specification can be used in SAE I assume that the population distribution of Y given a set of covariates X is adequately modelled at the unit level by a linear mixed model with Gaussian random effects of the form

$$y_j = \mathbf{x}_j'\beta+\mathbf{z}_j'\mathbf{u}+e_j$$

where j indexes individual population units, e_j denotes individual model error, $\mathbf{u}$ denotes a set of random area effects and $\mathbf{z}_j$ is a covariate characterising the impact of these area random effects on an individual population unit. The most common specification for this model in SAE is a random intercepts specification, where we associate a random effect u_i with each area i and $\mathbf{u}$ denotes the vector of these effects. In this case $\mathbf{z}_j$ is the vector that "picks out" the area in which unit j is located. It is standard to assume non-informative sampling within each area, in which case the sample data on Y can be written in matrix form as

$$\mathbf{y}_s = \mathbf{X}_s\beta + \mathbf{Z}_s\mathbf{u} + \mathbf{e}_s$$

where $\mathbf{e}_s$ is a n-vector of uncorrelated zero mean Gaussian random variables with common variance σ_e^2, $\mathbf{u}$ is a D-vector of uncorrelated zero mean Gaussian random variables with variance σ_u^2, and $\mathbf{e}_s$ and $\mathbf{u}$ are distributed independently.

With linked data spread across Q blocks, however, we see

$$\mathbf{y}_s^* = \begin{pmatrix} \mathbf{y}_{s1}^* = \mathbf{A}_{s1}\mathbf{y}_1 \\ \mathbf{y}_{s2}^* = \mathbf{A}_{s2}\mathbf{y}_2 \\ \vdots \\ \mathbf{y}_{sQ}^* = \mathbf{A}_{sQ}\mathbf{y}_Q \end{pmatrix} = \mathbf{A}_s\mathbf{y}$$

where $\mathbf{A}_s = \text{diag}(\mathbf{A}_{sq})$ and $\mathbf{y}$ denotes the vector of actual (but unknown) Y -values in the population. Put $\mathbf{T}_s = \text{diag}(\mathbf{T}_{sq})$ and $\mathbf{J}_s = \text{diag}(\mathbf{J}_{sq})$, where $\mathbf{T}_{sq} = (\lambda_q - \gamma_q)\,\mathbf{J}_{sq} + \gamma_q \mathbf{1}_{m_q}\mathbf{1}'_{M_q}$ and $\mathbf{J}_{sq} = \left[\mathbf{I}_{m_q} \,\middle|\, \mathbf{0}_{m_q\times(M_q-m_q)}\right]$. Then

$$E\left(\mathbf{y}_s^* \,\middle|\, \mathbf{X}, \mathbf{Z}\right) = \begin{pmatrix} \mathbf{T}_{s1}\mathbf{X}_1 \\ \mathbf{T}_{s2}\mathbf{X}_2 \\ \vdots \\ \mathbf{T}_{sQ}\mathbf{X}_Q \end{pmatrix}\beta = \mathbf{T}_s\mathbf{X}\beta$$

$$\text{Var}(\mathbf{y}_s^* \,|\, \mathbf{X}, \mathbf{Z}) = \left[\left(\sigma_e^2\mathbf{I}_{sq} + \mathbf{V}_{sq} + \sigma_u^2\mathbf{K}_{sq}\right)1(p=q) + (\sigma_u^2\mathbf{T}_{sp}\mathbf{Z}_p\mathbf{Z}'_q\mathbf{T}'_{sq})\,1(p\neq q)\right] = \Sigma_s^*$$

and

$$\text{Cov}(\mathbf{y}_s, \mathbf{y}_s^* \,|\, \mathbf{X}, \mathbf{Z}) = \text{Cov}(\mathbf{y}_s, \mathbf{A}_s\mathbf{y} \,|\, \mathbf{X}, \mathbf{Z}) = (\sigma_e^2\mathbf{J}_s + \sigma_u^2\mathbf{Z}_s\mathbf{Z}')\,\mathbf{T}'_s = \mathbf{C}_s\mathbf{T}'_s$$

where $1(w)$ equals one if statement w is true and is zero otherwise, and $\mathbf{K}_{sq}$ (see Samart and Chambers, 2014) represents the extra heterogeneity in $\mathbf{y}_s^*$ due to incorrect linkage of units in the same block but in different areas. Put $\text{Var}\left(\mathbf{y}_s \,|\, \mathbf{X}, \mathbf{Z}\right) = \Sigma_s = \sigma_e^2\mathbf{I}_s + \sigma_u^2\mathbf{Z}_s\mathbf{Z}'_s$. Making the same Gaussian copula assumption as before, we can then write $\mathbf{y}_s \,|\, \mathbf{y}_s^*, \mathbf{X}, \mathbf{Z} \sim \mathbf{a}_s + \mathbf{B}_s^{1/2}\mathbf{g}_s$, with

$$\mathbf{a}_s = E(\mathbf{y}_s \,|\, \mathbf{y}_s^*, \mathbf{X}, \mathbf{Z}) = \mathbf{X}_s\beta + \mathbf{C}_s\mathbf{T}'_s(\Sigma_s^*)^{-1}(\mathbf{y}_s^* - \mathbf{T}_s\mathbf{X}\beta)$$

and

$$\mathbf{B}_s = \text{Var}(\mathbf{y}_s \,|\, \mathbf{y}_s^*, \mathbf{X}, \mathbf{Z}) = \Sigma_s - \mathbf{C}_s\mathbf{T}_s'(\Sigma_s^*)^{-1}\mathbf{T}_s\mathbf{C}_s'.$$

Since $\mathbf{y}_s^* = \mathbf{A}_s\mathbf{y}$ the ideal data set underpinning the use of the MIP in this situation would normally include the population vector $\mathbf{y}$ and, as in the previous development, application of the MIP would then proceed using the properties of the conditional distribution $\mathbf{y} \,|\, \mathbf{y}_s^*, \mathbf{X}, \mathbf{Z}$. However, this involves manipulating N-dimensioned quantities, which is usually impractical. I therefore introduce a further approximation that replaces $\mathbf{y}$ by $\mathbf{y}_s$ in the ideal data set. This has the immediate effect of replacing N-dimensioned quantities by n-dimensioned quantities in the score identity, which now depends on the first and second moments of the conditional distribution $\mathbf{y}_s \,|\, \mathbf{y}_s^*, \mathbf{X}, \mathbf{Z}$ derived above. These can now be used to approximate the score functions for β, σ_u^2 and σ_e^2 based on the linked sample data, replacing the score functions for these parameters based on the ideal data set by their conditional expectations given the actual (i.e., linked) sample data. This process is the same as that already outlined for the simple regression case earlier so no details are provided here. Instead, I note that the popular maximum likelihood version of the Empirical Best Linear Predictor (EBLUP) of the mean $\bar{y}_i$ of Y in area i is of the form $\bar{y}_i^{\text{EBLUP}} = \bar{\mathbf{x}}_i'\hat{\beta} + \bar{\mathbf{z}}_i'\hat{\mathbf{u}}$ where $\hat{\beta}$ is the MLE for β and $\hat{\mathbf{u}}$ is the minimum MSE linear predictor for the vector of area effects $\mathbf{u}$ when β, σ_e^2 and σ_u^2 are replaced by their MLEs. However, given linked sample data, the minimum MSE predictor of $\mathbf{u}$ is its conditional expectation given these data. Under the Gaussian copula assumption this is

$$E(\mathbf{u} \,|\, \mathbf{y}_s^*, \mathbf{X}, \mathbf{Z}) = \sigma_u^2\mathbf{Z}'\mathbf{T}_s'(\sigma_s^*)^{-1}(\mathbf{y}_s^* - \mathbf{T}_s\mathbf{X}\beta).$$

When MIP-based MLE approximations for β, σ_e^2 and σ_u^2 are substituted in this expression we obtain a linkage error corrected predictor of the random effects vector $\mathbf{u}$, which we denote by $\hat{\mathbf{u}}_{\text{MIP}}$. Combining this with the MIP-based MLE approximation $\hat{\beta}_{\text{MIP}}$ for β one can then compute a MIP-based predictor for the mean of Y in area i as $\bar{y}_i^{\text{MIP}} = \bar{\mathbf{x}}_i'\hat{\beta}_{\text{MIP}} + \bar{\mathbf{z}}_i'\hat{\mathbf{u}}_{\text{MIP}}$.

I can illustrate the gains from using this MIP-based approach to SAE based on linked data via a small simulation. This assumes linkage errors follow an ELE model with known parameters (see Chambers (2009) for how one deals with an ELE model with estimated parameters). The target population consisted of $D =$ 40 areas, with an average area population size of 500, so $N = 20{,}000$. A random intercepts model was used to generate the ideal data values of Y for unit j in area i according to $y_j = 100 + 5x_j + u_i + e_j$, where the values of x_j were generated as independent and identically distributed lognormal with a log scale mean of log(4.5)-0.5 and a log scale variance of 0.5. The area random effects u_i were independently generated as Gaussian with mean zero and variance $\sigma_u^2 = 2$ while the individual random effects e_j were independently generated as Gaussian with mean zero and variance $\sigma_e^2 = 7$. The actual linked values of Y were then generated by independent repetitions of an ELE model within $Q = 40$ blocks covering the population of interest. The blocks were defined independently of the small areas of interest, with $\lambda_q = 1$ in blocks 1-10, $\lambda_q = 0.95$ in blocks 11-20, $\lambda_q = 0.9$ in blocks 21-30 and $\lambda_q = 0.85$ in blocks 31-40. Blocks contained units from multiple small areas, with a block including units from an average of 5 small areas. As a consequence there was across area linkage error. Independent simple random samples were taken from each area, with area sample sizes ranging from 5 to 40 with an average of 25, so $n = 1{,}000$ and the linked sample values of Y as well as the population values of X were then used to fit the random intercepts model.

The above scenario was independently simulated 100 times. In each simulation estimates of the model parameters were calculated for the ideal case (no linkage error for sample values) and for the naive case (linkage error ignored), in both cases via REML using the function *lmer* in R (R Core Team, 2019). Estimates were also calculated using the MIP-based approach described above, using the naive estimates as starting values. Table 5.1 shows the average values and RMSEs over the 100 simulations, while the boxplots in Figure 5.1 show the distributions of these parameter estimates over the same simulations. Observe that the measurement error due to linkage error causes naive estimates of the fixed effects to be biased, reflecting the fact that linkage error shrinks slope parameters towards zero, with naive estimates of between area variation reduced and corresponding estimates of within area variation greatly increased. This is exactly what one expects. The MIP-based estimates do not suffer from these problems.

In addition, EBLUP-type estimates of the population average of Y in each small area were calculated in each simulation, using the same parameter estimation methods (Ideal, Naive and MIP). For each small area and each simulation, the squared error and the absolute error of these EBLUP-type estimates were also computed. Figure 5.2 shows the boxplots of their corresponding mean squared error and mean absolute error values over the simulations for each area and for each parameter estimation method. These are denoted Area-MSE and Area-MAE respectively.

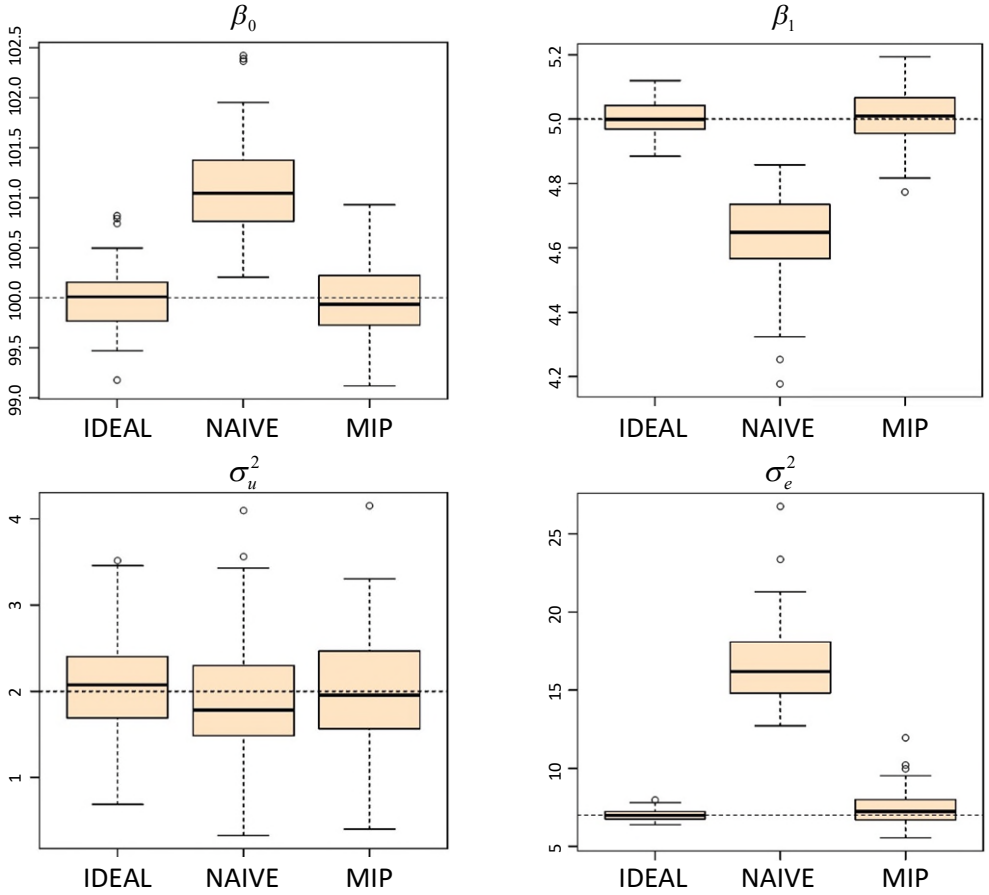
These results show that a method for fitting a mixed model that allows for linkage error can lead to significant improvement over a naive approach that ignores linkage error. This is consistent with results presented in Samart and Chambers (2014), Briscolini et al. (2018) and Salvati, Fabrizi, Ranalli and Chambers (2021). Of these, it is only the first paper where linkage errors are allowed between distinct small areas. Note that the approach described in that paper is not based on use of the MIP but on direct development of the likelihood function generated by the linked data followed by approximation of the relevant score functions. It also assumes balanced data (all block by area cells have sample) in order to obtain a formula for $\mathbf{K}_{sq}$. The same formula was used in the simulation reported here. In related research not presented here, Nicola Salvati and Enrico Fabrizi have empirically demonstrated that the small area estimates generated by the Samart-Chambers approach are less efficient than those generated by the MIP approach.

Table 5.1
Simulation results for 100 independent repetitions of the ELE linkage error scenario with $N = 20{,}000$, $n = 1{,}000$ and $Q = 40$ blocks.

Parameter	Average			RMSE		
(True value)	Ideal	Naive	MIP	Ideal	Naive	MIP
β_0 (100)	99.993	101.115	99.981	0.296	1.206	0.358
β_1 (5)	5.004	4.637	5.006	0.051	0.384	0.078
σ_u^2 (2)	2.078	1.889	2.050	0.560	0.656	0.649
σ_u^2 (7)	7.016	16.651	7.438	0.326	9.975	1.166

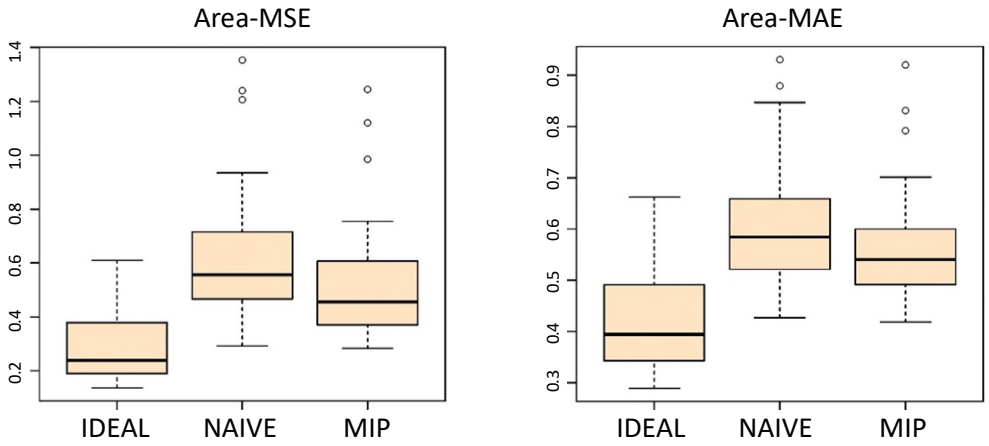
Notes: Independent SRSWOR samples were taken in each of $D = 40$ areas with sample sizes ranging between 5 and 40.
ELE = Exchangeable linkage errors; MIP = Missing information principle; RMSE = Root mean square error; SRSWOR = Simple random sampling without replacement.

Figure 5.1 Boxplots showing the distributions of parameter estimate values in the simulations of the ELE linkage error scenario.



Notes: Dotted horizontal line shows the true value of each parameter. ELE = Exchangeable linkage errors; MIP = Missing information principle.

Figure 5.2 Boxplots showing the distributions of area specific mean squared error (Area-MSE) and mean absolute error (Area-MAE) for the 40 areas in the ELE linkage error simulations.



Notes: ELE = Exchangeable linkage errors; MIP = Missing information principle.

Finally, it should be reiterated that application of the MIP approach with linked data is numerically intensive. This is because use of the ELE model to characterise linkage errors means that computations are effectively performed over all records in each of the blocks making up the $\mathscr{X}$-register. A practical implementation of the MIP algorithm that can handle large-scale linked population registers (which can contain millions of records) is an ongoing research project.

6. Discussion

I never knew Joseph Waksberg, but I certainly knew of him. Lohr (2021) describes a research career at the US Census Bureau and at Westat that made important contributions to many areas of survey methodology, including two that I subsequently became actively involved with – census coverage adjustment and calibration of survey weights. However, it was Waksberg's work on design and estimation using multiple frames that aligns most closely with the aims of this paper, since at their core these are about making maximum use of the information in combined data structures. In particular, his work shows us how to design sampling strategies that take advantage of this complexity to produce efficient estimates that relate to the population underpinning the combined data.

My aim in this paper has not been design but estimation, and in particular the use of the Missing Information Principle as a guide for defining parametric estimators when modelling messy data. In the context of a multiple frames scenario with random sampling of each frame, the ideal data are the values associated with the union of the distinct population frames, and the estimating equations for model parameters given multiple frame sample data are defined by replacing the sufficient statistics in the ideal data score function by their conditional expectations given these available data. When these sufficient statistics are linear in functions of the ideal data values this usually corresponds to replacing function values for individual units by their conditional expectations given the information derived from their (potentially multiple) frame memberships. This can be a complex specification process, requiring different models for different amounts of frame overlap.

There are many other messy data situations where application of the MIP leads to useful insights. Thus, Steel, Beh and Chambers (2004) report on how it can be used in likelihood-based inference with ecological data, i.e., where parameters of a joint distribution are of interest, but where the available data only provide information about marginal distributions. Here also having access to a very small sample taken from the joint distribution can have a very large impact in terms of improving the quality of inference. Another important application area where sampling is clearly informative is case-control sampling, see Prentice and Pyke (1979). Following the approach of Scott and Wild (1997) in this situation, Chambers and Wang (2008) use the MIP to develop MLEs for the parameters of a logistic regression model given case-control data. The simulation results they report show substantial improvement in efficiency over the standard approach for this model, which assumes simple random sampling in the fitting process and then discards the intercept

estimate. The MIP has been employed for efficient design as well, with Chipperfield, Barr and Steel (2018) using it in the context of efficient split questionnaire design.

The book Chambers et al. (2012) contains many more examples of application of the MIP as well as much more detailed developments of the results sketched out in this paper. In particular these show the information functions based on the available data. However, as I stated at the start of this paper, I believe that it is the score identity component of the MIP that is most useful since it shows how estimation should proceed. Uncertainty estimation given these estimates can be derived from the information function, but they can also be derived via more direct Taylor Series approximations or via numerically intensive methods such as bootstrapping.

When one views the score identity in the MIP from a more abstract perspective, it is clear that it is a special case of estimation based on the conditional expectation of a convenient estimating function. Consequently, if one generalises from the standard frequentist likelihood focus of this paper, it is interesting to note that an equivalent formulation of the score identity has been developed for estimating functions based on quasi-likelihood (Lin, Steel and Chambers, 2004). That is, there is scope for extension of the use of a MIP-based approach to estimation in messy data situations, for example those based on nonparametric likelihood approaches like empirical likelihood. Whether this leads to further insights remains to be seen, however. In any case such nonparametric extensions will also require methods for calculating the nonparametric equivalent of the conditional expectation operator reflecting the available information, perhaps via constrained parametric simulation. It will be interesting to see whether the development of these generalisations of the MIP will then allow it to accommodate the types of "large" machine learning models that are becoming more common.

I am very grateful to the Waksberg Award Committee for giving me this opportunity to prepare this paper for Survey Methodology. Hopefully it will encourage other statisticians working with messy data to investigate whether the MIP (and its potential generalisations) can be a useful tool for making inferences based on these data.

Finally, I would like to dedicate the preparation of this paper to my three friends and former colleagues from the University of Southampton, Fred Smith, Chris Skinner and Tim Holt, who have all now sadly passed away. Without the impetus of their groundbreaking book, Skinner, Holt and Smith (1989), and their insight and support during my years at Southampton, many of my personal contributions reported in this paper would not have been possible.

References

ADRN (2012). The UK Administrative Data Research Network. Improving Access for Research and Policy. Report from the Administrative Data Taskforce.

Bardsley, P., and Chambers, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

Bethlehem, J.G., and Keller, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Binder, D.A., and Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters. *Analysis of Survey Data*, (Eds., R.L. Chambers and C.J. Skinner). Chichester: Wiley.

Breckling, J.U., Chambers, R.L., Dorfman, A.H., Tam, S.M. and Welsh, A.H. (1994). Maximum likelihood inference from survey data. *International Statistical Review*, 62, 349-363.

Briscolini, D., Di Consiglio, L., Liseo, B., Tancredi, A. and Tuoto, T. (2018). New methods for small area estimation with linkage uncertainty. *International Journal of Approximate Reasoning*, 94, 30-42.

Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

Chambers, R.L. (2009). Regression analysis of probability-linked data. *Statisphere*, 4. Available from https://ndhadeliver.natlib.govt.nz/delivery/DeliveryManagerServlet?dps_pid=FL1356824.

Chambers, R.L., and Clark, R.G. (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford: Oxford University Press.

Chambers, R., and Diniz da Silva, A. (2019). Improved secondary analysis for linked data: A framework and an illustration. *Journal of the Royal Statistical Society, Series A*, 183, 37-59.

Chambers, R.L., Dorfman, A.H. and Wang, S. (1998). Limited information likelihood analysis of survey data. *Journal of the Royal Statistical Society, Series B*, 60, 397-412.

Chambers, R.L., and Skinner, C.J. Eds. (2003). *Analysis of Survey Data*. Chichester: John Wiley & Sons, Inc.

Chambers, R.L., Steel, D.G., Wang, S. and Welsh, A.H. (2012). *Maximum Likelihood Estimation for Sample Surveys*. CRC Press: Boca Raton, Florida.

Chambers, R., and Wang, S. (2008). Maximum likelihood logistic regression with auxiliary information. Working Paper 12-08, Centre for Statistical and Survey Methodology, University of Wollongong.

Chipperfield, J.O., Barr, M.L. and Steel, D.G. (2018). Split questionnaire designs: Collecting only the data you need through MCAR and MAR designs. *Journal of Applied Statistics*, 45, 1465-1475. DOI 10.1080/02664763.2017.1375085.

Clark, R.G., and Chambers, R.L. (2008). Adaptive calibration for prediction of finite population totals. *Survey Methodology*, 34, 2, 163-172. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008002/article/10757-eng.pdf.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39, 1-37.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Dorfman, A., Chambers, R. and Wang, S. (2002). Are survey weights necessary? The maximum likelihood approach to sample survey inference. *Proceedings of the 162nd Annual Meeting of the American Statistical Association*, New York, August 11-15.

Elliott, M.R., and Davis, W.W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: Combining data from two surveys. *Applied Statistics*, 54, 595-609.

Galloway, A. (2005). Non-probability sampling. *Encyclopedia of Social Measurement*, 2, (Ed., K. Kempf-Leonard), Elsevier, 859-864.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22, 153-164.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of super populations and survey population: Their relationship and estimation. *International Statistical Review*, 54, 37-59.

Handcock, M., Rendall, M. and Cheadle, J. (2005). Improved regression estimation of a multivariate relationship with population data on the bivariate relationship. *Sociological Methodology*, 35, 291-334.

Harron, K., Goldstein, H. and Dibben, C. (Editors) (2016). *Methodological Developments in Data Linkage*. Chichester: Wiley.

Imbens, G.W., and Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *Review of Economic Studies*, 61, 655-680.

Kim, G., and Chambers, R. (2012). Regression analysis under incomplete linkage. *Computational Statistics and Data Analysis*, 56, 2756-2770.

Kish, L., and Frankel, M.R. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society, Series B*, 36, 1-37.

Krieger, A.M., and Pfeffermann, D. (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology*, 18, 2, 225-239. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1992002/article/14484-eng.pdf.

Lahiri, P., and Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 222-230.

Lin, Y.-X., Steel, D. and Chambers, R.L. (2004). Restricted quasi-score estimating functions for sample survey data. *Stochastic Methods and Their Applications* (*Journal of Applied Probability*, 41A, (Eds., J. Gani and E. Seneta). Sheffield: Applied Probability Trust.

Little, R.J.A. (2003). The Bayesian approach to sample survey inference. Chapter 4, *Analysis of Survey Data*, (Eds., R.L. Chambers and C.J. Skinner). Chichester: Wiley.

Little, R.J.A. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics (with discussion). *Journal of Official Statistics*, 28, 309-372.

Little, R.J.A. (2022). Bayes, buttressed by design-based ideas, is the best overarching paradigm for sample survey inference. *Survey Methodology*, 48, 2, 257-281. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00001-eng.pdf.

Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics, 1st Edition. New York: John Wiley & Sons, Inc.

Lohr, S.L. (2021). Multiple-frame surveys for a multiple-data-source world. *Survey Methodology*, 47, 2, 229-263. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00008-eng.pdf.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.

Orchard, T., and Woodbury, M.A. (1972). A missing information principle: Theory and application. *Proc. 6th Berkeley Symp. Math. Statist*., 1, 697-715.

Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society Series, Series B*, 12, 241-255.

Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. *International Statistical Review*, 61, 317-337.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology*, 37, 2, 115-136. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011002/article/11602-eng.pdf.

Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.

Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā, Series B*, 61, 166-186.

Pfeffermann, D., and Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. *Analysis of Survey Data*, (Eds., C. Skinner and R. Chambers), New York: John Wiley & Sons, Inc., 175-195.

Prentice, R.L., and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403-411.

R Core Team (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL: https://www.R-project.org/.

Raessler, S. (2004). Data fusion: Identification problems, validity, and multiple imputation. *Austrian Journal of Statistics*, 33, 153-171.

Rao, J.N.K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, 31, 2, 117-138. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9040-eng.pdf.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. Hoboken, NJ: Wiley.

Royall, R.M. (1970). On finite population sampling under certain linear regression models. *Biometrika*, 57, 377-387.

Royall, R.M. (1976). The least squares linear approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 53, 581-592.

Salvati, N., Fabrizi, E., Ranalli, M.G. and Chambers, R. (2021). Small area estimation with linked data. *Journal of the Royal Statistical Society, Series B*, 83, 78-107.

Samart, K., and Chambers, R. (2014). Linear regression with nested errors using probability-linked data. *Australian and New Zealand Journal of Statistics*, 56, 27-46.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33, 2, 99-119. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10488-eng.pdf.

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

Scott, A. (2006). Population-based case control studies. *Survey Methodology*, 32, 2, 123-132. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9546-eng.pdf.

Scott, A.J., and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84, 57-71.

Silva, P.L.D.N., and Skinner, C.J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23, 1, 23-32. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997001/article/3102-eng.pdf.

Skinner, C.J., Holt, D. and Smith, T.M.F. (Editors) (1989). *Analysis of Complex Surveys*. Chichester: Wiley.

Steel, D.G., Beh, E.J. and Chambers, R.L. (2004). The information in aggregate data. *Ecological Inference: New Methodological Strategies*, (Eds., G. King, O. Rosen and M. Tanner). Cambridge: Cambridge University Press.

Strauss, W.J., Carroll, R.J., Bortnick, S.M., Menkedick, J.R. and Schultz, B.D. (2001). Combining datasets to predict the effects of regulation of environmental lead exposure in housing stock. *Biometrics*, 57, 203-210.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*, New York: John Wiley & Sons, Inc.

Zhang, L.-C., and Tuoto, T. (2021). Linkage-data linear regression. *Journal of the Royal Statistical Society: Series A*, 184, 522-547.