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

Survey Methodology

Model-assisted sample design is minimax for model-based prediction

by Robert Graham Clark

Release date: June 30, 2020

Statistics Canada Statistique Canada

Canada

How to obtain more information

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

You can also contact us by

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

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

Depository Services Program

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

Standards of service to the public

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

Note of appreciation

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

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, June 2020
Vol. 46, No. 1, pp. 77-91
Statistics Canada, Catalogue No. 12-001-X

Model-assisted sample design is minimax for model-based prediction

Robert Graham Clark[1]

Abstract

Probability sampling designs are sometimes used in conjunction with model-based predictors of finite population quantities. These designs should minimize the anticipated variance (AV), which is the variance over both the superpopulation and sampling processes, of the predictor of interest. The AV-optimal design is well known for model-assisted estimators which attain the Godambe-Joshi lower bound for the AV of design-unbiased estimators. However, no optimal probability designs have been found for model-based prediction, except under conditions such that the model-based and model-assisted estimators coincide; these cases can be limiting. This paper shows that the Godambe-Joshi lower bound is an *upper* bound for the AV of the best linear unbiased estimator of a population total, where the upper bound is over the space of all covariate sets. Therefore model-assisted optimal designs are a sensible choice for model-based prediction when there is uncertainty about the form of the final model, as there often would be prior to conducting the survey. Simulations confirm the result over a range of scenarios, including when the relationship between the target and auxiliary variables is nonlinear and modeled using splines. The AV is lowest relative to the bound when an important design variable is not associated with the target variable.

Key Words: Anticipated variance; Model-based inference; Probability sampling; Sample surveys.

1 Introduction

Model-based inference about finite population totals relies on an assumed model and usually does not make reference to the sampling plan. Probability sampling, where every unit i has a known probability of selection $\pi_i > 0$, is not strictly necessary, but is often used anyway, because it "eliminates conscious and unconscious bias" (Valliant, Dever and Kreuter, 2013, page 310) and ensures the non-informativeness of sampling which is required for most model-based procedures (Chambers and Clark, 2012, page 12). Särndal, Swensson and Wretman (1992, page 534) note that "proponents of model-based inference advocate randomized selection of the sample as a safeguard against selection bias, but the randomization probabilities play no role in the inference". See also Lohr (2010, page 263), Chambers and Clark (2012, page 92) and Scott, Brewer and Ho (1978) who suggest probability sample designs for model-based predictions. For a review of the model-based approach, see also Valliant, Dorman and Royall (2000).

Model-assisted inference (e.g., Särndal et al., 1992) is an alternative approach, where estimators are design-unbiased (at least asymptotically), that is, unbiased over repeated probability sampling from any fixed population. Subject to this constraint, they minimize the anticipated variance (AV), which is the variance over both repeated realisations of the population from a model and repeated probability sampling. In models with independent errors, the lowest possible AV amongst such estimators (for any given probability sample design) is the Godambe-Joshi lower bound (GJLB) (Godambe and Joshi, 1965). The lower bound is asymptotically achieved for linear models by the well known generalized regression estimator.

1. Robert Graham Clark, Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, Australia. E-mail: robert.clark@anu.edu.au.

Model-assisted designs are probability sample designs which are intended to minimize the AV of the generalized regression estimator (or, equivalently, to minimize the GJLB). These AV-optimal sample designs have been derived for model-assisted inference. In particular, the sample design which minimizes the GJLB for fixed expected sample size for models with independence has probability proportional to the square root of the model error variance for each unit (σ_i) (e.g., Särndal et al., 1992); this will be called a PPσ design. The PPCσ design is a generalization allowing for unequal unit costs (Steel and Clark, 2014). There are no analogous results on optimal probability sampling for model-based prediction, except under strong conditions, a gap that this paper partially fills. Isaki and Fuller (1982) suggested the design-estimation strategy of using the PPσ design for model-based prediction, showing that this design is optimal when selection probabilities and their squares are in the column space of the matrix of covariates. This condition comes at a price, as will be seen in the simulation in this paper.

Optimal non-probability samples have been derived for model-based best linear unbiased predictors (BLUPs) under linear models. These tend to be somewhat extreme designs, where the units with the largest, or the largest and smallest, values of auxiliary variables are chosen (e.g., Royall, 1970). Robust model-based balanced designs have been developed, where one or more sample moments of auxiliary variables are equal to the corresponding population moments (Royall and Herson, 1973), while "over-balanced designs" meet a different constraint on the sample moments (Scott et al., 1978). Another balanced design was proposed by Kott (1986). These designs are robust to families of polynomial alternatives to a working linear model. They are not probability designs, although probability designs have been proposed in order to approximately meet balancing constraints (Valliant et al., 2000, Section 3.4). Exactly balanced probability designs have also been proposed (Tillé, 2006). The choice of balancing or over-balancing strategy depends on which set of polynomial alternatives is postulated. In another non-probability approach, Welsh and Wiens (2013) find the sample which minimizes the maximum model-based variance in a neighbourhood of a working model.

This article derives an asymptotic upper bound for the AV of the BLUP under probability sampling. The AV is the most relevant quantity for probability sample design even in the model-based framework, because averaging over all possible samples is appropriate in advance of sample selection. The bound is applicable to any probability sample design, and is over the space of possible covariate sets. This is useful for sample design in practice, because the precise model to be used is not decided until after data have been collected. For example, some design variables might not be included in the model if the sample data suggests that they have little relevance for the variable whose total is being estimated, but this would not be clear prior to surveying. Or splines might be used, with the number and placement of knots guided by the sample data. It turns out that the upper bound is the GJLB. This implies that model-assisted designs, such as PPσ and PPCσ, are minimax strategies for model-based estimation. The upper bound is an equality when the model has a particular property, which is satisfied when the model is sufficiently rich and includes all design variables.

Other researchers have considered the relationship between the BLUP and the model-assisted generalized regression estimator, including conditions under which these two estimators are identical (e.g., Isaki and Fuller, 1982; Tam, 1988) and modifications to the BLUP so that it is equivalent to a generalized regression estimator at the expense of its optimality under the model (e.g., Brewer, Hanif and Tam, 1988; Brewer, 1999; Nedyalkova and Tillé, 2008). The results here are new because:

- Existing results do not cater for situations where both: the surveyor wants to use the BLUP because it is model-optimal, and the BLUP and the generalized regression estimators are not equal. The case where the two estimators are equal is shown here to be, in a particular sense, the worst case for the BLUP.
- An expression for the AV of the BLUP is derived and shown explicitly to be less than or equal to the upper bound. The result seems intuitively reasonable, given that the GJLB is attained by the design-consistent generalized regression estimator, whereas the BLUP is not subject to the constraint of design-based consistency. However, it is not at all obvious from the expression for the BLUP's AV that the upper bound applies, so it is useful to have an explicit result.
- The interpretation is made that the upper bound is over the space of possible choices for the model covariates $\mathbf{x}$. Thus, the upper bound is relevant when the sample designer is unsure what model will ultimately be adopted once data have been collected.

Section 2 contains the key theoretical results. Section 3 confirms and illustrates the main result in a simulation study with a variable of interest Y and two auxiliary variables: x_1 (continuous) and x_2 (binary). The expected value of Y conditional on these variables is defined by a linear and a sinusoidal term in x_1. It does not depend on x_2. The probabilities of selection are a function of both x_1 and x_2. BLUPs are calculated based on the model with lowest Bayesian Information Criterion (BIC) from a set including the simple linear model in x_1 and splines in x_1 of various degrees, both with and without x_2. The ratio of the simulation prediction mean squared error (MSE) of the BLUP to the GJLB is either less than or equal to 1 or just above 1 across a range of scenarios. Section 4 is a discussion.

Much of the literature comparing model-assisted and model-based estimators and inference has focussed on bias due to mis-specified models when either (a) the mean function is incorrect, or (b) some design variables are inappropriately excluded. See for example Hansen, Madow and Tepping (1983) and the reworking of their simulation study in Valliant et al. (2000, Section 3.4). The simulation in Section 3 considers (a) and (b) to some extent, but this isn't the main focus of the paper. The aim here is to see whether a committed model-based statistician can use the GJLB as an upper bound for the AV for sample design purposes, rather than to adjudicate between model-based and design-based inference. It is assumed that a sufficiently good model can be identified using the sample data; this process would be aided by a design which minimizes the maximal AV over the space of all linear models. A PPσ or PPCσ design is recommended when there is considerable uncertainty over the form of the final model.

2 Upper bound for the AV of the BLUP

Let $U = \{1, \ldots, N\}$ denote the finite population. For unit $i \in U$, the variable of interest is y_i and the p-vector of auxiliary variables is $\mathbf{x}_i$. The sample (of size n) is s and the non-sample set is $r = U - s$. Auxiliary variables are observed for all $i \in U$ while y_i is observed for $i \in s$. The aim is to predict $t_y = \sum_{i \in U} y_i$. The probabilities of selection are $\pi_i = P[i \in s \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_N, y_1, \ldots, y_N] = P[i \in s \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_N] > 0$; they are assumed to be a function of the population values of $\mathbf{x}_i$. Let $\mathbf{t}_x = \sum_{i \in U} \mathbf{x}_i$ and $\mathbf{t}_{xr} = \sum_{i \in r} \mathbf{x}_i$.

The n by p matrix of sample values of $\mathbf{x}$, which has rows $\mathbf{x}_i^T$, is denoted X_s. The $N - n$ by p matrix of non-sample values of $\mathbf{x}$ is X_r. The vector of sample values of y is $\mathbf{y}_s$.

The following linear model M is assumed:

$$E_M [y_i] = \boldsymbol{\beta}^T \mathbf{x}_i \tag{2.1}$$

$$\mathrm{var}_M [y_i] = \sigma_i^2 = \sigma^2 v_i \tag{2.2}$$

$$\mathrm{cov}_M [y_i, y_j] = 0 \tag{2.3}$$

for $i, j \in U$ with $i \neq j$. The subscripts M in E_M, var_M and cov_M indicate distributions over repeated realisations of the population values from the model. It is generally assumed that v_i are known, i.e., the error variances are known up to a constant of proportionality. For example, in business surveys, v_i might be a measure of business size, or the square root thereof. The unknown parameters are $\boldsymbol{\beta}$ and σ^2. The values of $\mathbf{x}_i$ are considered to be fixed.

The best linear unbiased predictor (BLUP) (denoted $\hat{t}_y$) for a generalization of model M is stated in Chapter 2 of Valliant et al. (2000). Its model-based prediction variance is

$$\mathrm{var}_M \left(\hat{t}_y - t_y\right) = \mathbf{t}_{xr}^T \left(\sum_{i \in s} \sigma_i^{-2} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \mathbf{t}_{xr} + \sum_{i \in r} \sigma_i^2. \tag{2.4}$$

(This can be obtained as a special case of Result 2.2.2 on page 29 of Valliant et al., 2000.)

The anticipated variance is defined as $\mathrm{AV}\left(\hat{t}_y - t_y\right) = E_M E_p \left(\hat{t}_y - t_y\right)^2$ (Isaki and Fuller, 1982). As $\hat{t}_y$ is model-unbiased, its AV is equal to

$$\mathrm{AV} = E_p \mathrm{var}_M \left(\hat{t}_y - t_y\right).$$

Theorem 1 will derive an approximation for this AV. The asymptotic framework is based on the design-based asymptotics of Isaki and Fuller (1982). It is assumed that there is a countably infinite population $i = 1, 2, \ldots$. A sequence of finite populations U_t is defined by $U_t = \{1, \ldots, N_t\}$ where $N_1 < N_2 < \ldots$. For each t, a sample s_t of size n_t is selected from U_t by arbitrary probability design with probabilities of selection $\pi_{i(t)} = P[i \in s_t]$. Following (2.7) of Isaki and Fuller (1982), it is assumed that

$$0 < \lambda_1 < \pi_{i(t)} < \lambda_2 \tag{2.5}$$

for some constants λ_1 and λ_2. Isaki and Fuller (1982) note that AVs of estimators of total are typically $O(n_t)$ (which is equivalent to $O(N_t)$) and the totals themselves are also $O(n_t)$. Population means will be denoted as $\bar{\mathbf{X}}_t = N_t^{-1}\sum_{U_t}\mathbf{x}_i$ and the inverse probability weighted estimator of $\bar{\mathbf{X}}$ is $\hat{\bar{\mathbf{X}}}_\pi = N_t^{-1}\sum_{s_t}\pi_{i(t)}^{-1}\mathbf{x}_i$ (and similarly for Y and other variables).

Two new variables are defined for each unit i by $\mathbf{u}_{i(t)} = \pi_{i(t)}\mathbf{x}_i$ (a p-vector) and $\mathbf{v}_{i(t)} = \pi_{i(t)}\sigma_i^{-2}\mathbf{x}_i\mathbf{x}_i^T$ (a p by p matrix). Their population means are $\bar{\mathbf{U}}_t$ and $\bar{\mathbf{V}}_t$ with inverse probability estimators $\hat{\bar{\mathbf{U}}}_{t\pi}$ and $\hat{\bar{\mathbf{V}}}_{t\pi}$.

Theorem 1. *It is assumed that*

$$\lim_{t\to\infty} E_p\left\{\hat{\bar{\mathbf{U}}}_{t\pi}^T\hat{\bar{\mathbf{V}}}_{t\pi}^{-1}\hat{\bar{\mathbf{U}}}_{t\pi} - \bar{\mathbf{U}}_t^T\bar{\mathbf{V}}_t^{-1}\bar{\mathbf{U}}_t\right\} = 0. \tag{2.6}$$

Then

$$E_p var_M\left(\hat{t}_y - t_y\right) = AV + o(n_t). \tag{2.7}$$

where

$$AV = \sum_U(1-\pi_i)\,\mathbf{x}_i^T\left(\sum_U\pi_i\sigma_i^{-2}\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\sum_U(1-\pi_i)\,\mathbf{x}_i + \sum_U(1-\pi_i)\,\sigma_i^2. \tag{2.8}$$

Notes on Theorem 1

- Assumption (2.6) is reminiscent of Result (3.24) of Isaki and Fuller (1982), but there is an important difference. In Isaki and Fuller (1982), unit variables depend only on i, but here $\mathbf{u}_{i(t)}$ and $\mathbf{v}_{i(t)}$ depend on both i and t as they both have a factor $\pi_{i(t)}$. However, $\pi_{i(t)}$ are bounded by (2.5), so the condition is plausible; it would not be if $\pi_{i(t)}$ could be arbitrarily close to zero.
- It is clear that assumption (2.6) is satisfied if $\hat{\bar{\mathbf{U}}}_{t\pi}$ and $\hat{\bar{\mathbf{V}}}_{t\pi}$ are consistent in design probability for $\bar{\mathbf{U}}_t$ and $\bar{\mathbf{V}}_t$, and $\hat{\bar{\mathbf{V}}}_{t\pi}$ is invertible in a neighbourhood of $\bar{\mathbf{V}}_t$. As noted by Isaki and Fuller (1982) in a comment on their condition (3.12), a invertibility requirement of this sort seems reasonable "for any discussion of regression estimation".

An upper bound for the asymptotic AV over all possible choices of the auxiliary vector $\mathbf{x}_i$ will now be derived. This allows for uncertainty about which auxiliary variables will ultimately be included in the model, since this decision is typically only made after data is collected. For example, the full set of variables used in the design might or might not end up being in the model, or spline functions of covariates might be included with knots based on the sample data. Theorem 2 states the upper bound.

Theorem 2. *Let AV be the asymptotic AV defined by (2.8). If $\sum_U\pi_i\sigma_i^{-2}\mathbf{x}_i\mathbf{x}_i^T$ is invertible and $\pi_i > 0$ for all $i \in U$, then*

$$AV \le \sum_U\left(\pi_i^{-1} - 1\right)\sigma_i^2 \tag{2.9}$$

with strict equality if and only if there exists a p-vector $\boldsymbol{\lambda}$ such that

$$\left(\pi_i^{-1} - 1\right) \sigma_i^2 = \boldsymbol{\lambda}^T \mathbf{x}_i \tag{2.10}$$

for all $i \in U$.

The right hand side of (2.9) is the well known Godambe-Joshi lower bound (Godambe and Joshi, 1965) for the AV of design-unbiased estimators. Here it is an upper bound over model space for model-based BLUPs.

Suppose the total cost of running the survey is $\sum_s C_i$ plus fixed costs, where C_i is the cost associated with surveying unit i. Then the expected cost is $C_E = \sum_U C_i \pi_i$. The sample design which minimizes the upper bound in (2.9) subject to fixed cost is the PPCσ design which has

$$\pi_i \propto \sigma_i \big/ \sqrt{C_i} \tag{2.11}$$

(Steel and Clark, 2014 who generalize Särndal et al., 1992, page 452) to allow for unequal costs. Theorem 2 means that (2.11) is a minimax design when there is uncertainty about the form of the model. Note that only the first order inclusion probabilities affect the AV and the bound, but these do not fully specify the design. Samples can be selected using these inclusion probabilities in a variety of ways (Tillé, 2006), including balanced probability sampling (Nedyalkova and Tillé, 2012) which improves the robustness to model mis-specification.

The condition for equality, (2.10), is equivalent to a well known condition for the BLUP to be equal to the generalized regression estimator (formula 3 of Tam, 1988). Tam (1988) argued for the use of sample designs such that (2.10) is satisfied, such as PPσ (provided that the model includes an intercept). Nedyalkova and Tillé (2008), building on a result from Royall (1992), showed that PPσ is model-based-optimal under equal costs when both v_i and $\sqrt{v_i}$ are linear functions of $\mathbf{x}_i$, a condition called *explainable variances*. Brewer et al. (1988) noted that (2.10) can also be satisfied if the estimation model includes an instrumental variable, which is a suitable function of the selection probabilities. However, there are many circumstances under which (2.10) is not satisfied, because some auxiliary variables are omitted from the final model, because multiple variables of interest have different variance structures (ruling out PPσ), because there are unequal costs, or because instrumental variables are eschewed due to the loss of efficiency they entail. Theorem 2 shows that the GJLB is an upper bound under these circumstances which are not covered by the results of these authors. The PPCσ design in (2.11) is a minimax design in this more general setting.

3 Simulation study

A simulation study was conducted to compare the AV of the BLUP and its upper bound in situations where Y has a nonlinear relationship with a continuous auxiliary variable x_1. A second auxiliary variable, x_2, is binary and independent of x_1 and Y. Probabilities of selection depend on x_1 and x_2 in

various ways. For each scenario, 5,000 populations and samples were generated, with a population size of 6,000 and sample sizes of 500 and 1,500, and with a population size of 100,000 and a sample size of 25,000. All code is available at www.github.com/rgcstats/AVLB.

3.1 Simulation of populations

The population values of x_1 were the $\{j/(101): \ j = 1, \ldots, 100\}$ quantiles of a lognormal distribution with mean $-1/32$ and standard deviation 0.25, with equal frequencies of each of these 100 values. This means that x_1 are positive and right-skewed with a mean of 1 and a range of 0.54 to 1.74. The values of x_1 were non-stochastic and discretized in order to speed up computation, to simplify the generation of smooth models (see below), and to facilitate comparison of AVs to the GJLB by making the GJLB constant across simulations. A second binary variable x_2 took on the values 0 and 1 with equal frequency within each value of x_1, so that the two covariates were orthogonal.

Conditional on x_1 and x_2, the population values of Y were generated independently as

$$E_M Y_i = \mu(x_{1i}) + \varepsilon_i \tag{3.1}$$

where

$$\varepsilon_i \sim N(0, 0.25x_{1i}). \tag{3.2}$$

The mean function $\mu(.)$ was a smooth but nonlinear function,

$$\mu(x_1) = 4x_1 + \sin(x_1 2\pi/h), \tag{3.3}$$

consisting of a linear term and a sinusoidal term with period h. When h is large, $\mu()$ is close to linear over the range of x_1 in the population, while for h small there are frequent cycles in the function. Figure 3.1 shows the mean function for the periods used in the simulation (0.5, 1, 2, 5).

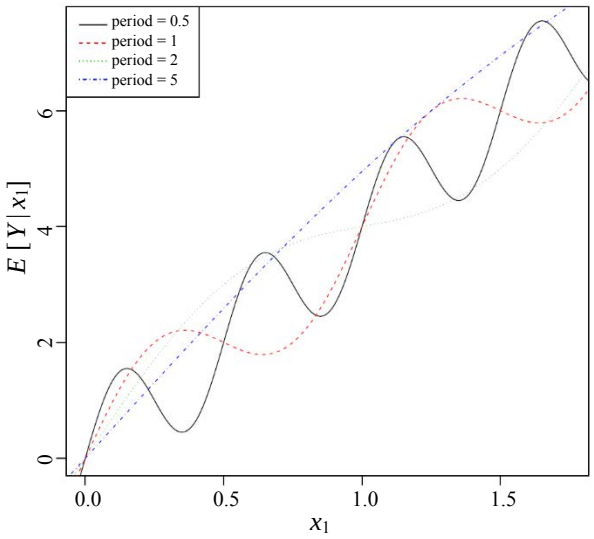

Figure 3.1 $\mu(x_1) = E[Y \mid x_1]$ for the periods used in the simulation study.

3.2 Simulated sampling

The probabilities of selection, π_i, were set to

$$\pi_i \propto x_{1i}^b (1 + cx_{2i}) \tag{3.4}$$

where b was 0.5, 1 or 2, reflecting light, medium or high dependence on x_{1i}. The values of c were 0, 0.5 or 1.5, reflecting no, medium or high dependence on x_{2i}. (Other values of c were also used for the purposes of Figure 3.2 only.)

The second auxiliary variable, x_2, is unrelated to Y but may affect the selection probabilities. One might expect the BLUP to do better compared to the GJLB when probabilities of selection depend on x_2, since the BLUP may be based on a model omitting x_2, potentially leading to lower variance. Of course, there is also the possibility of the working model omitting x_1, leading to the BLUP being biased, however this never occurred in any of the simulations. To explore robustness to incorrect omission of x_1, it would be of interest to consider relationships weaker than those shown in Figure 3.1, but this was beyond the scope of the paper.

Inclusion probabilities are forced to obey the proportionality in (3.4) but are truncated above at 1 and below at 1/40 and scaled such that they add to the required sample size after truncation. Samples are selected by unequal probability systematic sampling with random ordering using the `sampling` package in R (Tillé and Matei, 2016).

3.3 Estimation of the population total of Y

A linear model in x_1 and spline models in x_1 with between 1 and 10 interior knots are fitted to each sample. (See for example Breidt, Claeskens and Opsomer, 2005 for the use of splines in model-assisted survey estimation.) Another 11 models are defined by also including x_2 as an additive covariate. The model with the lowest BIC is then used to calculate a BLUP of t_y. This model selection step would be expected to increase the variability of this predictor. The BLUP based on the simple linear model in x_1 is also calculated. The process is repeated with working models including the correct variance specification $\text{var}_M(y_i) \propto x_{1i}$ and a mis-specification $\text{var}_M(y_i) \propto x_{1i}^2$.

3.4 Simulation results

Tables 3.1-3.4 show the ratios of the prediction MSEs of various BLUP estimators to the GJLB (the right hand side of equation 2.9) for various sample designs and choices of $\mu(x_1)$. The prediction MSEs are the means over all simulations of $(\hat{t}_y - t_y)^2$ so they are with respect to both model and design, as are the results in Theorems 1 and 2.

Table 3.1a evaluates the BLUP corresponding to the lowest-BIC model for a sample size of 500 with correctly specified variances. Nine sample designs are shown corresponding to three choices for b (low/medium/high dependency of the selection probabilities on x_1) and c (no/some/high dependency of selection probabilities on x_2). The period h of the sinusoidal component of $\mu(x_1)$ is also shown (see equation 3.3). The table shows that:

- The ratio is always either less than or equal to 1 or slightly above 1, consistent with Theorem 2. Its range is 0.815 to 1.086.
- The ratio decreases slightly as h increases. So, when the true $E_M(y \mid x_1)$ is close to linear, the model-based BLUP has lower MSE relative to the GJLB, while for more nonlinear models the ratio is closer to 1.
- The ratio depends on b to a degree, although the pattern depends on the other parameters.
- The ratio decreases dramatically as c increases, with reductions of up to 20% from $c = 0$. This shows that the BLUP does much better relative to the GJLB when there is a covariate x_2 which is relevant in the design but not relevant to Y so that it can be omitted from the estimation model.

Table 3.1b shows results for a much larger sample size of 25,000 from a population of 100,000. These results were included to see whether the ratios are less than or equal to 1 for large n as predicted by the theory in Section 2. A larger number of simulations were also used for this panel (15,000 rather than 5,000). Results are only shown for $c = 0$ since these were the designs with the highest ratios in Table 3.1a. The ratios in 3.1b range from 0.984 to 1.011. The values slightly above 1 may reflect that the working spline model does not perfectly capture the sinusoidal functions used to generate the data.

Table 3.2 shows the same scenarios as Table 3.1a except that the BLUP is based on a mis-specified variance model with $\sigma_i^2 \propto x_i^2$ (the generating model has $\sigma_i^2 \propto x_i$). The ratios of the MSE of the lowest-BIC-model BLUP to the GJLB are on the whole slightly higher than Table 3.1 (generally by less than one percentage point). The ratio is still almost always less than or equal to 1, with a maximum value of 1.089.

Table 3.3 is similar to Table 3.1a except that the sample size is 1,500 rather than 500. The ratios are almost always lower than in Table 3.1a. The maximum ratio is 1.030.

Table 3.4 shows results for the BLUP based on the simple linear model containing only x_1 with mis-specified variance. The sample size is 500. When the period is 5, so that the true model is virtually linear in x_1, this BLUP does very well. The ratios are then always below 1.1 and can be as low as 0.790. When the period is 2, there is visible curvature in $E(y \mid x_1)$ (as shown in Figure 3.1), but the simple BLUP still does well, with all ratios less than 1.2. However, for periods 0.5 and 1, the ratios are well above 1, with a maximum value of 3.4. This shows the substantial bias of the BLUP when the model is badly mis-specified.

The extent to which the selection probabilities depend on x_2 is the major factor determining the ratio of the MSE to the GJLB, as shown by Tables 3.1-3.3. Figure 3.2 shows this phenomenon in more detail for correctly specified variance. The sample size is 1,500 with PPσ sampling so that $\pi_i \propto x_{1i}^{0.5}$. Values of c (0, 0.25, ..., 3) are on the x-axis and results are shown for different periods h. The figure shows that the ratio is slightly above 1 for $c = 0$ and decreases smoothly with c, to about 0.6 when $c = 3$. Higher periods h (reflecting a smoother relationship between Y and x_1) are also associated with lower ratios, but the differences are so small as to be almost indiscernible in Figure 3.2.

Table 3.1
Ratios of MSE of BLUP based on lowest BIC spline model to Godambe-Joshi Lower Bound for sample sizes of 500 and 25,000 with variance correctly specified. Probabilities of selection are proportional to $x_1^b(1 + cx_2)$. The period h controls the smoothness of $E[Y \mid x_1]$

sample design		period (h)			
b	c	0.5	1	2	5
(a) sample size of 500					
0.5	0	1.086	1.064	1.052	1.057
0.5	0.5	0.990	0.963	0.951	0.956
0.5	1.5	0.840	0.827	0.808	0.815
1	0	1.033	1.015	0.997	1.006
1	0.5	1.006	0.992	0.973	0.985
1	1.5	0.877	0.859	0.854	0.858
2	0	1.080	1.063	1.035	1.081
2	0.5	1.046	1.021	0.996	1.039
2	1.5	0.870	0.853	0.839	0.856
(b) sample size of 25,000					
0.5	0	1.011	1.011	1.011	1.010
1	0	1.007	1.006	1.006	1.006
2	0	0.985	0.985	0.984	0.984

Table 3.2
Ratios of MSE of BLUP based on lowest BIC model to Godambe-Joshi Lower Bound for sample size of 500 with variance mis-specified. Probabilities of selection are proportional to $x_1^b(1 + cx_2)$. The period h controls the smoothness of $E[Y \mid x_1]$

sample design		period (h)			
b	c	0.5	1	2	5
0.5	0	1.089	1.068	1.055	1.061
0.5	0.5	0.996	0.967	0.959	0.962
0.5	1.5	0.852	0.830	0.819	0.825
1	0	1.033	1.012	0.998	1.001
1	0.5	1.006	0.992	0.978	0.988
1	1.5	0.886	0.863	0.854	0.862
2	0	1.078	1.061	1.035	1.047
2	0.5	1.048	1.017	0.997	1.010
2	1.5	0.878	0.857	0.842	0.856

Table 3.3
Ratios of MSE of BLUP based on lowest BIC model to Godambe-Joshi Lower Bound for sample size of 1,500 with variance correctly specified. Probabilities of selection are proportional to $x_1^b(1 + cx_2)$. The period h controls the smoothness of $E[Y \mid x_1]$

sample design		period (h)			
b	c	0.5	1	2	5
0.5	0	1.030	1.023	1.019	1.022
0.5	0.5	0.928	0.920	0.918	0.922
0.5	1.5	0.762	0.754	0.750	0.749
1	0	0.940	0.936	0.930	0.932
1	0.5	0.979	0.972	0.966	0.968
1	1.5	0.821	0.817	0.810	0.809
2	0	1.028	1.008	0.995	1.020
2	0.5	0.962	0.948	0.941	0.969
2	1.5	0.798	0.798	0.786	0.804

Table 3.4
Ratios of MSE of BLUP based on simple linear model to Godambe-Joshi Lower Bound for sample size of 500 with variance mis-specified. Probabilities of selection are proportional to $x_1^b(1 + cx_2)$. The period h controls the smoothness of $E[Y \mid x_1]$

sample design		period (h)			
b	c	0.5	1	2	5
0.5	0	3.083	2.037	1.109	1.055
0.5	0.5	2.918	1.860	1.002	0.958
0.5	1.5	2.452	1.533	0.840	0.790
1	0	3.086	1.812	1.052	1.007
1	0.5	3.006	1.792	1.016	0.979
1	1.5	2.537	1.500	0.880	0.838
2	0	3.423	2.900	1.174	1.080
2	0.5	3.243	2.693	1.111	1.025
2	1.5	2.689	2.291	0.926	0.829

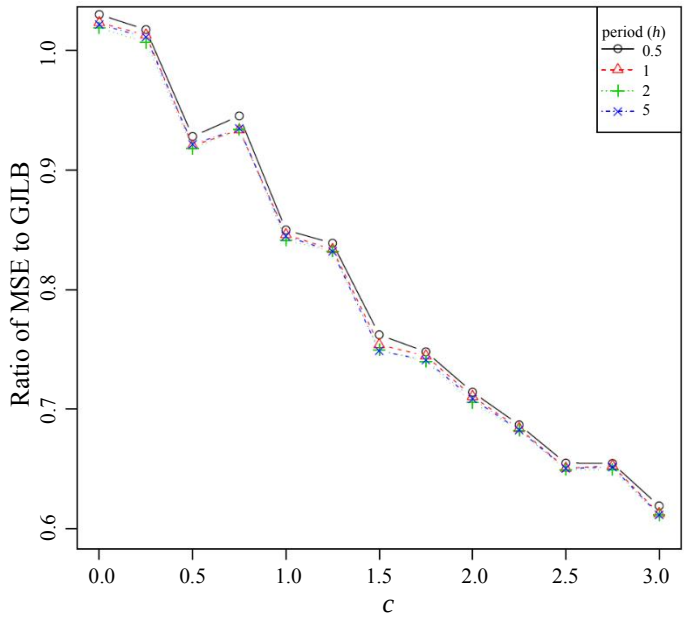

Figure 3.2 Ratios of MSE of BLUP based on lowest BIC model to Godambe-Joshi Lower Bound for sample size of 1,500 with variance correctly specified, vs c, where probabilities of selection are proportional to $x_1^b(1 + cx_2)$ with $b = 1$. The period (h) controls the smoothness of $E[Y \mid x_1]$.

4 Discussion

The Godambe-Joshi lower bound is shown here to be an *upper* bound for the AVs of BLUPs based on a correct model. Simulation MSEs of BLUPs based on an adaptively chosen spline or linear model are consistently less than the GJLB or just above it, even when variances are mis-specified. The MSEs are well below the bound if an important design variable does not figure in the model.

The upper bound result relies on the BLUP being model-unbiased. BLUPs based on a badly mis-specified model had MSEs well above the bound in the simulation study. Choosing a working model with minimum BIC out of a class including spline models avoided this problem.

Once data are available, model-based inference conditions on the sample selected. Probability sampling nevertheless has many advantages even though it is not the basis of inference (e.g., Särndal et al., 1992; Valliant et al., 2013). At the design stage, the AV is then the most relevant objective, because it averages over all the possible samples which may then be selected. The upper bound derived here is relevant at the design stage because there would usually be considerable uncertainty about the form of the model which will ultimately be adopted (there may be exceptions when there are historical or related data to support specification of a model or where one is willing to trust that the true model lies in a class of polynomial or other specific alternative models).

A sensible strategy in practice would be:

i. Set π_i so as to give low values for the upper bound $\sum_{i\in U}\left(\pi_i^{-1}-1\right)v_i$ where the model variances are proportional to v_i (or a weighted combination of the upper bounds for multiple variables of interest) while also respecting cost and practical considerations. If there is a single variable of interest, and unit costs are proportional to C_i, then $\pi_i \propto \sqrt{v_i/C_i}$ is recommended as it is a minimax strategy.

ii. Once the sample is selected and data are available, choose a regression model based on this data.

iii. Estimate population totals using the BLUPs under the selected model.

iv. This may or may not result in condition (2.10) being satisfied, depending on the costs C_i and the auxiliary variables selected in the final model.

All optimal sample design results, whether model-based, design-based or model-assisted, rely on knowledge of the relative unit or stratum residual variances. This appears to be unavoidable. This paper helps when the form of the mean model is not known in advance by giving an upper bound over the space of models for the mean. There does not appear to be a correspondingly useful bound over possible variance models, so the form of the variance model must be guessed or assumed.

Acknowledgements

I would like to thank Stephen Haslett for his encouragement of methodological research amongst competing priorities. Thorough reviews by two referees and an associate editor substantially improved the paper. This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI Australia), an NCRIS enabled capability supported by the Australian Government.

Appendix

Proof of Theorem 1

From (2.4),

$$E_p \mathrm{var}_M\left(\hat{t}_y - t_y\right) = E_p\left\{\mathbf{t}_{xr}^T\left(\sum_{i\in s}\sigma_i^{-2}\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\mathbf{t}_{xr}\right\} + \sum_{i\in U}(1-\pi_i)\,\sigma_i^2. \qquad \text{(A.1)}$$

Making use of the definitions of $\hat{\bar{\mathbf{U}}}_\pi$ and $\hat{\bar{\mathbf{V}}}_\pi$, and assumption (2.6), the first term of (A.1) becomes

$$\begin{aligned} E_p\left\{\mathbf{t}_{xr}^T\left(\sum_{i\in s}\sigma_i^{-2}\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\mathbf{t}_{xr}\right\} &= E_p\left\{\left(N_t\bar{\mathbf{X}}-N_t\hat{\bar{\mathbf{U}}}_\pi\right)^T\left(N_t\hat{\bar{\mathbf{V}}}_\pi\right)^{-1}\left(N_t\bar{\mathbf{X}}-N_t\hat{\bar{\mathbf{U}}}_\pi\right)\right\} \\ &= N_tE_p\left\{\left(\bar{\mathbf{X}}-\hat{\bar{\mathbf{U}}}_\pi\right)^T\hat{\bar{\mathbf{V}}}_\pi^{-1}\left(\bar{\mathbf{X}}-\hat{\bar{\mathbf{U}}}_\pi\right)\right\} \\ &= N_t\left\{\left(\bar{\mathbf{X}}-\bar{\mathbf{U}}\right)^T\bar{\mathbf{V}}^{-1}\left(\bar{\mathbf{X}}-\bar{\mathbf{U}}\right)+o(1)\right\}. \end{aligned} \tag{A.2}$$

The result follows immediately from (A.1) and (A.2).

Lemma 1: Let $a_1,\ldots,a_n$ and $b_1,\ldots,b_n$ be scalars where $b_i>0$ for all i. Let $\mathbf{x}_1,\ldots,\mathbf{x}_n$ be p-vectors. Then

$$\left(\sum_{i=1}^N a_i\mathbf{x}_i\right)^T\left(\sum_{i=1}^N b_i\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\left(\sum_{i=1}^N a_i\mathbf{x}_i\right)\le\sum_{i=1}^N a_i^2/b_i \tag{A.3}$$

provided the matrix inverse exists. Equality in (A.3) obtains if and only if

$$a_ib_i^{-1}=\boldsymbol{\lambda}^T\mathbf{x}_i \tag{A.4}$$

for all $i=1,\ldots,n$ for some p-vector $\boldsymbol{\lambda}$.

Proof of Lemma 1

Let $b=\sum_{i=1}^n b_i$. Let X be a discrete random variable taking on the values a_i/b_i. Let $\mathbf{Y}$ be a discrete random variable taking on the values $\mathbf{x}_i$, for $i=1,\ldots,n$. Let $P[Y=\mathbf{x}_i, X=a_i/b_i]=b_i/b$ for $i=1,\ldots,n$. Write $M_1\le M_2$ if M_1-M_2 is negative semi-definite for any matrices M_1 and M_2. Theorem 1 of Tripathi (1999) states that for any random vectors $\mathbf{X}$ and $\mathbf{Y}$,

$$E[\mathbf{X}\mathbf{Y}^T]\{E[\mathbf{Y}\mathbf{Y}^T]\}^{-1}E[\mathbf{Y}\mathbf{X}^T]\le E[\mathbf{X}\mathbf{X}^T] \tag{A.5}$$

provided the matrix inverse exists. With my definition of X and $\mathbf{Y}$, (A.5) becomes

$$\sum_{i=1}^N b_ib^{-1}a_ib_i^{-1}\mathbf{x}_i^T\left\{\sum_{i=1}^N b_ib^{-1}\mathbf{x}_i\mathbf{x}_i^T\right\}^{-1}\sum_{i=1}^N b_ib^{-1}a_ib_i^{-1}\mathbf{x}_i\le\sum_{i=1}^N b_ib^{-1}a_i^2b_i^{-2} \tag{A.6}$$

which leads directly to (A.3).Tripathi (1999) states that the equality is sharp if

$$\mathbf{X}^T\boldsymbol{\lambda}_1+\mathbf{Y}^T\boldsymbol{\lambda}_2=0 \tag{A.7}$$

with probability 1 for some λ_1 of the same dimension as $\mathbf{X}$ and λ_2 of the same dimension as $\mathbf{Y}$. Here, (A.7) becomes

$$a_ib_i^{-1}\lambda_1=\mathbf{x}_i^T\boldsymbol{\lambda}_2$$

for all i, which is equivalent to (A.4).

Proof of Theorem 2

Let $a_i = 1 - \pi_i$ and $b_i = \pi_i \sigma_i^{-2}$. From Lemma 1,

$$\begin{aligned}
\text{AV} &= \sum_U (1-\pi_i)\,\mathbf{x}_i^T \left(\sum_U \pi_i \sigma_i^{-2} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_U (1-\pi_i)\,\mathbf{x}_i^T + \sum_U (1-\pi_i)\,\sigma_i^2 \\
&\le \sum_U (1-\pi_i)^2\,\pi_i^{-1}\sigma_i^2 + \sum_U (1-\pi_i)\,\sigma_i^2 \\
&= \sum_U (1-\pi_i)\,\pi_i^{-1}\sigma_i^2\,(1-\pi_i+\pi_i) \\
&= \sum_U (\pi_i^{-1}-1)\,\sigma_i^2
\end{aligned}$$

with strict equality if and only if

$$\boldsymbol{\lambda}^T \mathbf{x}_i = a_i b_i^{-1} = (\pi_i^{-1} - 1)\,\sigma_i^2$$

for some vector $\boldsymbol{\lambda}$.

References

Breidt, F.J., Claeskens, G. and Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, 92(4), 831-846.

Brewer, K.R.W. (1999). Cosmetic calibration with unequal probability sampling. *Survey Methodology*, 25, 2, 205-212. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1999002/article/4883-eng.pdf.

Brewer, K.R.W., Hanif, M. and Tam, S.M. (1988). How nearly can model-based prediction and design-based estimation be reconciled? *Journal of the American Statistical Association*, 83, 128-132.

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

Godambe, V.P., and Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations 1. *Annals of Mathematical Statistics*, 36, 1707-1722.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Kott, P.S. (1986). When a mean-of-ratios is the best linear unbiased estimator under a model. *The American Statistician*, 40(3), 202-204.

Lohr, S.L. (2010). *Sampling: Design and Analysis, 2nd edition*. Boston: Brooks-Cole.

Nedyalkova, D., and Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika*, 95(3), 521-537.

Nedyalkova, D., and Tillé, Y. (2012). Bias-robustness and efficiency of model-based inference in survey sampling. *Statistica Sinica*, 22(2), 777-794.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2), 377-387.

Royall, R.M. (1992). Robustness and optimal design under prediction models for finite populations. *Survey Methodology*, 18, 2, 179-185. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1992002/article/14488-eng.pdf.

Royall, R.M., and Herson, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68(344), 880-889.

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

Scott, A.J., Brewer, K.R.W. and Ho, E.W.H. (1978). Finite population sampling and robust estimation. *Journal of the American Statistical Association*, 73(362), 359-361.

Steel, D.G., and Clark, R.G. (2014). Potential gains from using unit level cost information in a model-assisted framework. *Survey Methodology*, 40, 2, 231-242. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14110-eng.pdf.

Tam, S.M. (1988). Asymptotically design-unbiased predictors in survey sampling. *Biometrika*, 75(1), 175-177.

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer.

Tillé, Y., and Matei, A. (2016). *Sampling: Survey Sampling*. R package version 2.8. https://CRAN.R-project.org/package=sampling.

Tripathi, G. (1999). A matrix extension of the Cauchy-Schwarz inequality. *Economics Letters*, 63(1), 1-3.

Valliant, R., Dever, J.A. and Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.

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

Welsh, A.H., and Wiens, D.P. (2013). Robust model-based sampling designs. *Statistics and Computing*, 23(6), 689-701, November. https://doi.org/10.1007/s11222-012-9339-3.