Design-unbiased statistical learning in survey sampling

Luis Sanguiao Sande[1] and Li-Chun Zhang[2,3,4]

[1] *Instituto Nacional de Estadística (luis.sanguiao.sande@ine.es)*
[2] *University of Southampton (email: L.Zhang@soton.ac.uk)*
[3] *Statistisk sentralbyraa*
[4] *Universitetet i Oslo*

Abstract: Design-consistent model-assisted estimation has become the standard practice in survey sampling. However, design consistency remains to be established for many machine-learning techniques that can potentially be very powerful assisting models. We propose a subsampling Rao-Blackwell method, and develop a statistical learning theory for *exactly* design-unbiased estimation with the help of linear or non-linear prediction models. Our approach makes use of classic ideas from Statistical Science as well as the rapidly growing field of Machine Learning. Provided rich auxiliary information, it can yield considerable efficiency gains over standard linear model-assisted methods, while ensuring valid estimation for the given target population, which is robust against potential mis-specifications of the assisting model, even if the design consistency of following the standard recipe for plug-in model-assisted estimator cannot be established.

Keywords: Rao-Blackwellisation, bagging, pq-unbiasedness, stability conditions

1 Introduction

To make use of the available auxiliary information, approximately design-unbiased model-assisted estimation has become the standard practice in survey sampling, following the influential works such as Särndal et al. (1992), Deville and Särndal (1992). Breidt and Opsomer (2017) review the "general recipe", and give examples of many Machine Learning (ML) techniques that can be or have been embedded in the model-assisted framework, such as Kernel methods, splines or neural networks. McConville and Toth (2019) observe that many of these sophisticated estimators "are rarely ever used by statistical agencies to produce official statistics", because the underlying "models are often ill suited to the available auxiliary data."

McConville and Toth (2019) promote intuitively the regression tree estimator as a post-stratification estimator, where the post-strata are selected by the recursive partitioning algorithm that generates the regression tree based on the available data. The technical development of the regression tree estimator has taken many years. Gordon and

Olshen (1978, 1980) establish the consistency of recursive partitioning algorithms, given independent and identically distributed (IID) training data. Toth and Eltinge (2011) extend their result, allowing for finite-population sampling *in addition* to the IID super-population model. McConville and Toth (2019) adapt the method of proof of Toth and Eltinge (2019) to establish the design-consistency of the post-stratification estimator corresponding to the sample-trained regression tree.

A closely related model is random forest (RF), consisting of many trees built on randomly selected subsets of all the observations. Breidt and Opsomer (2017) note that they "are not aware of direct uses of random forests in a model-assisted survey estimator". While the recent result of McConville and Toth (2019) concerning the regression tree estimator is encouraging, it cannot be applied directly to RF, for which design consistency needs to be worked out separately. Meanwhile, without one's being able to do so, it remains 'unjustified' to apply the standard recipe of model-assisted estimation to RF, even in a situation where one is quite certain that the RF model improves the estimation efficiency compared to a linear model.

In this paper, we propose a *subsampling Rao-Blackwell (SRB)* method for *exactly* design-unbiased survey estimation given any assisting prediction models. The main attraction of the unbiased approach is that it allows one to safely incorporate any suitable ML models, such as RF, even if one is unable to prove its design consistency technically, because variance reduction (e.g. compared to using a linear model) can be verified in many ways empirically. Note that the algorithmic nature and the flexibility of many modern prediction techniques (including RF) are particularly valuable in the presence of rich feature (or auxiliary) data, while saving one the trouble of building complex models manually. Yet, there is always a danger that ML can result in misspecified models, as with any model building endeavour (including building a linear regression model). This is why it is helpful that the SRB method can provide the much-needed assurance, in situations where the consistency of the standard recipe for model-assisted estimation is either in doubt or simply cannot be proved technically.

The SRB method makes use of three classic ideas from Statistical Science and ML. On the one hand, the training-test split of the sample of observations in ML generates *errors* in the test set rather than *residuals*, conditional on the training dataset, which as we shall explain is the key to achieving exact design-unbiasedness. For model-assisted survey estimation we use this idea to remove the finite-sample bias. On the other hand, Rao-Blackwellisation (Rao, 1945; Blackwell, 1947) and model-assisted estimation (Cassel et al., 1976) are powerful ideas in Statistics and survey sampling, which we apply to ML techniques to obtain design-unbiased survey estimators at the aggregated levels. Monte Carlo (MC) versions of the SRB method can be used in cases where exact Rao-Blackwellisation is computationally too costly. The MC-SRB method is still exactly design-unbiased, but somewhat less efficient due to the additional MC error. In practice, one can easily balance between the numerical efficiency of the MC-SRB method against the statistical efficiency

of the corresponding exact SRB method.

We refer to the amalgamation as *statistical learning* for survey sampling, since the term model-assisted estimation is entrenched with the existing recipe and its reliance on asymptotic unbiasedness (Breidt and Opsomer, 2017), whereas the focus of aggregate-level estimation (i.e. prediction for groups of units) is nonstandard in the ML literature.

Some general remarks about the proposed approach may be worth mentioning immediately. First, under finite-population sampling, the minimal sufficient statistic consisting of the unordered set of distinct *and* enumerated sample observations is not complete. Nevertheless, the RB method can be able to improve unbiased estimators e.g. under adaptive cluster sampling (Thompson, 1990). In the current context of making use of auxiliary information, it will be shown that applying the RB method generally improves the subsample unbiased estimator that is constructed as a simple difference estimator given any assisting model.

Next, there is a natural affinity between SRB and bootstrap aggregating (i.e. bagging), where the latter is an important existing idea (Breiman, 1996a; 1996b) in the ML literature and practice. As we shall explain, SRB leads generally to design-unbiased aggregate-level estimation, which does not result from bagging alone; and it reduces the variance of the subsample unbiased estimator, even when it is based on a stable learner, which bagging does only for unstable learners. The idea of using SRB for achieving design-unbiased prediction (for groups of units) can hopefully raise the awareness of the role of probability sampling and potentially be useful to the practitioners of ML beyond survey sampling.

Since the standard linear-model assisted estimator is design consistent under fairly mild conditions (e.g. Särndal et al., 1992), and one may consider lasso linear regression (McConville et al., 2017) in cases of rich auxiliary data, there is no need to adopt any alternative estimator unless it clearly has a smaller variance compared to using the linear model. It follows that consistency is not a worrisome issue for unbiased statistical learning in practice. Still, an unbiased estimator is consistent if its variance tends to 0 under a suitable asymptotic setting. Breidt and Opsomer (2017) summarise the key condition for consistent model-assisted estimation in terms of the predictor trained on the whole population data, whereas it is common in the ML literature to formulate various *stability* conditions for non-parametric models. Bousquet and Elisseeff (2002) establish uniform stability condition for Regularisation algorithms. Mukherjee et al. (2006) pay special attention to empirical risk minimisation algorithms. These works are directed at *individual-level* predictors from independent and identically distributed (IID) data. Different definitions of stability are needed under the design-based approach to aggregate-level estimation. We shall formulate general stability conditions for design-consistent statistical learning under both simple random sampling and arbitrary unequal probability sampling designs, which can provide additional venues to the proof of consistency.

Finally, let it be clear that we do not propose to replace the standard recipe with the SRB method in every situation. When an estimator is consistent, following the recipe

of plugging in the full-sample trained ML predictor, it may be more efficient than the unbiased SRB method in terms of the mean squared error (MSE), in which case one is free to adopt the consistent estimator instead of the unbiased one. Future research will show how to make the most out of the SRB method through the choice of the subsampling scheme. But even without being optimally tuned, the SRB method allows one to safely gain from model-assisted survey estimation, in situations where the consistency of the standard recipe cannot be guaranteed.

The rest of the paper is organised as follows. In Section 2, by way of introduction we describe the SRB method that uses an assisting linear model. The underlying ideas of design-unbiased statistical learning are explained, as well as the differences to the standard model-assisted generalised regression estimation. Some basic methods of variance estimation are outlined, where a novel jackknife variance estimator is developed for the SRB method. We move on to arbitrary ML techniques in Section 3. The similarity and difference between SRB and bagging are explored. Moreover, we investigate and prove the stability conditions for design-consistency of SRB method that uses non-parametric algorithmic prediction models. Two simulation studies are presented in Section 4, which illustrate the potential gains of the proposed unbiased statistical learning approach, compared to standard linear model-assisted or model-based approaches. A brief summary and topics for future research will be given in Section 5.

2 Unbiased linear estimation

In this section we consider unbiased linear estimation in survey sampling, which builds on generalised regression (GREG) estimation (Särndal et al. 1992). The GREG estimator is the most common estimation method in practical survey sampling. It is consistent under mild regularity conditions, and is often more efficient than exactly unbiased Horvitz-Thompson (HT) estimation (Horvitz and Thompson, 1952). The proposed subsampling Rao-Blackwellisation (SRB) method removes the finite-sample bias of GREG generally, whose relative efficiency is comparable to the standard GREG estimator.

2.1 Bias correction by subsampling

Let s be a sample (of size n) selected from the population U of size N, with probability $p(s)$, where $\sum_s p(s) = 1$ over all possible samples under a given sampling design. Let $\pi_i = \Pr(i \in s) > 0$ be the sample inclusion probability, for each $i \in U$. Let y_i be a survey variable, for $i \in U$, with unknown population total $Y = \sum_{i \in U} y_i$.

Let the assisting linear model expectation of y_i be given by $\mu(x_i) = x_i^\top \beta$, where x_i is the vector of covariates for each $i \in U$. Let $\mu(x_i, s) = x_i^\top b$ be the estimator of $\mu(x_i)$, where $b = (\sum_{i \in s} x_i x_i^\top / \pi_i)^{-1} \sum_{i \in s} x_i y_i / \pi_i$ is a weighted least squares (WLS) estimator of β. It is possible to attach additional heteroscedasticity weights in the WLS; but the development

below is invariant to such variations, so that it is more convenient to simply ignore it in the notation. Let $X = \sum_{i \in U} x_i$. The GREG estimator of Y is given as

$$\widehat{Y}_{GR} = X^\top b + \sum_{i \in s} (y_i - x_i^\top b)/\pi_i$$

While $\widehat{Y}_{GR}$ is design-consistent under mild regularity conditions (e.g. Särndal et al. 1992), as $n, N \to \infty$, it is usually biased given finite sample size n, except in special cases such as when $x_i \equiv 1$ and $\pi_i \equiv n/N$, where $\mu(x, s) = \sum_{i \in s} y_i/n = \bar{y}_s$ and $\widehat{Y}_{GR} = N\bar{y}_s$.

To remove the potential finite-sample bias of $\widehat{Y}_{GR}$, consider subsampling of $s_1 \subset s$, with known probability $q(s_1|s)$, such as SRS with fixed $n_1 = |s_1|$, where $\sum_{s_1} q(s_1|s) = 1$. The induced probability of selecting s_1 from U is given by

$$p_1(s_1) = \sum_{s: s_1 \subset s} q(s_1|s) p(s)$$

where $\pi_{1i} = \Pr(i \in s_1)$ is the corresponding inclusion probability for $i \in U$. Let $s_2 = s \setminus s_1$ be the complement of s_1 in s. Let the conditional sampling probability of s given s_1 be

$$p_2(s_2|s_1) = p(s) q(s_1|s)/p_1(s_1)$$

and let $\pi_{2i} = \Pr(i \in s_2|s_1)$ be the corresponding conditional inclusion probability in s_2 for $i \in U \setminus s_1$. Let $\mu(x_i, s_1) = x_i^\top b_1$ be the estimate of $\mu(x_i)$ based on the sub-sample s_1, where $b_1 = (\sum_{i \in s_1} x_i x_i^\top/\pi_{1i})^{-1} \sum_{i \in s_1} x_i y_i/\pi_{1i}$. Let

$$\widehat{Y}_1 = \sum_{i \in s_1} y_i + \sum_{i \in U \setminus s_1} x_i^\top b_1 + \sum_{i \in s_2} (y_i - x_i^\top b_1)/\pi_{2i} \tag{1}$$

In other words, it is the sum of y_i in s_1 and a difference estimator of the remaining population total based on s_2, via $x_i^\top b_1$ that does not depend on the observations in s_2.

Proposition 1 The estimator $\widehat{Y}_1$ is conditionally unbiased for Y over $p_2(s_2|s_1)$ given s_1, denoted by $E_2(\widehat{Y}_1|s_1) = Y$, as well as unconditionally over $p(s)$, denoted by $E_p(\widehat{Y}_1) = Y$.
Proof: As $\mu(x_i, s_1)$ is fixed for any $i \in U \setminus s_1$ given s_1, the last two terms on the right-hand side of (1) is unbiased for $Y - \sum_{i \in s_1} y_i$ given s_1. It follows that $\widehat{Y}_1$ is conditionally unbiased for Y given s_1; hence, design-unbiased over $p(s)$ unconditionally as well. □

Example: Simple random sampling (SRS) Suppose SRS without replacement of s from U, and s_1 from s with fixed size $n_1 = |s_1|$, such that $\pi_{1i} = n_1/N$ and $\pi_{2i} = (n - n_1)/(N - n_1)$. In the special case of $x_i \equiv 1$, b_1 is the sample mean in s_1, and

$$\widehat{Y}_1 = n_1 b_1 + (N - n_1) \sum_{i \in s_2} y_i/(n - n_1)$$

which amounts to using the sample mean in s_2 to estimate the population mean outside of the given s_1, i.e., instead of using the sample mean in s for the whole population mean. Thus, $\widehat{Y}_1$ achieves unbiasedness generally, but at a cost of increased variance.

2.2 Rao-Blackwellisation

One can reduce the variance of $\widehat{Y}_1$ by the Rao-Blackwell method (Rao, 1945; Blackwell, 1947). The minimal sufficient statistic in the finite population sampling setting is simply $D_s = \{(i, y_i) : i \in s\}$. Applying the RB method to $\widehat{Y}_1$ by (1) yields $\widehat{Y}^*_{GR}$, which is given by the conditional expectation of $\widehat{Y}_1$ given D_s, i.e.

$$\widehat{Y}^*_{GR} = E_q(\widehat{Y}_1|D_s) = E_q(\widehat{Y}_1|s) \tag{2}$$

where the expectation is evaluated with respect to $q(s_1|s)$, and the second expression is leaner as long as one keeps in mind that $\{y_i : i \in U\}$ are treated as fixed constants associated with the distinctive units.

Proposition 2 The estimator $\widehat{Y}^*_{GR}$ is design-unbiased for Y, denoted by $E(\widehat{Y}^*_{GR}) = Y$.
Proof: By construction, the combined randomisation distribution induced by p and q is the same as that induced by p_1 and p_2, for any $s_1 \cup s_2 = s$ and $s_1 \cap s_2 = \emptyset$. Thus,

$$E(\widehat{Y}^*_{GR}) = E_p(\widehat{Y}^*_{GR}) = E_p\big(E_q(\widehat{Y}_1|s)\big) = E_1\big(E_2(\widehat{Y}_1|s_1)\big) = E_1\big(Y\big) = Y \qquad \square$$

Next, for the variance of $\widehat{Y}^*_{GR}$ over $p(s)$, i.e. $V(\widehat{Y}^*_{GR}) = V_p(\widehat{Y}^*_{GR})$, we notice

$$\begin{aligned} V(\widehat{Y}_1) &= E_p\big(V_q(\widehat{Y}_1|s)\big) + V_p\big(E_q(\widehat{Y}_1|s)\big) = E_p\big(V_q(\widehat{Y}_1|s)\big) + V_p(\widehat{Y}^*_{GR}) \\ V(\widehat{Y}_1) &= E_1\big(V_2(\widehat{Y}_1|s_1)\big) + V_1\big(E_2(\widehat{Y}_1|s_1)\big) = E_1\big(V_2(\widehat{Y}_1|s_1)\big) \end{aligned}$$

since $E_2(\widehat{Y}_1|s_1) = Y$. Juxtaposing the two expressions of $V(\widehat{Y}_1)$ above, we obtain

$$V(\widehat{Y}^*_{GR}) = V_p(\widehat{Y}^*_{GR}) = E_1\big(V_2(\widehat{Y}_1|s_1)\big) - E_p\big(V_q(\widehat{Y}_1|s)\big) \tag{3}$$

where $E_p\big(V_q(\widehat{Y}_1|s)\big)$ is the variance reduction compared to $\widehat{Y}_1$.

Proposition 3 Provided unbiased variance estimator $\widehat{V}(\widehat{Y}_1)$ with respect to $p_2(s_2|s_1)$, i.e. $E_2\big(\widehat{V}(\widehat{Y}_1)\big) = V_2(\widehat{Y}_1|s_1)$, a design-unbiased variance estimator for $\widehat{Y}^*_{GR}$ is given by

$$\widehat{V}(\widehat{Y}^*_{GR}) = \widehat{V}(\widehat{Y}_1) - V_q(\widehat{Y}_1|s)$$

Proof: By stipulation, we have $E\big(\widehat{V}(\widehat{Y}_1)\big) = E_1\Big(E_2\big(\widehat{V}(\widehat{Y}_1)\big)\Big) = E_1\Big(V_2(\widehat{Y}_1|s_1)\Big)$, which is the first term on the right-hand side of (3). The result follows immediately. $\square$

Example: SRS, cont'd In the special case of $x_i \equiv 1$ and $n_1 = n - 1$, we have

$$\widehat{Y}_{1(i)} = n_1 \bar{y}_{(i)} + (N - n_1) y_i$$

if $s_2 = \{i\}$ and $\bar{y}_{(i)}$ denotes the mean in $s_1 = s \setminus \{i\}$. The RB estimator follows as

$$\widehat{Y}^*_{GR} = \frac{1}{n} \sum_{i \in s} \widehat{Y}_{1(i)} = \frac{n_1}{n} \sum_{i \in s} \bar{y}_{(i)} + \frac{N}{n} \sum_{i \in s} y_i - \frac{n_1}{n} \sum_{i \in s} y_i = \frac{N}{n} \sum_{i \in s} y_i$$

which is the usual unbiased full-sample expansion estimator in this case. The RB method thus recovers the lost efficiency of any $\widehat{Y}_{1(i)}$ on its own.

Let $X_1 = \sum_{i \in s_1} x_i$, and $X_1^c = \sum_{i \notin s_1} x_i = X - X_1$. To express $\widehat{Y}^*_{GR}$ as a linear combination of $\{y_i : i \in s\}$, we rewrite $\widehat{Y}_1$ as

$$\begin{aligned}
\widehat{Y}_1 &= \sum_{i \in s_1} y_i + \sum_{i \in s_2} \frac{y_i}{\pi_{2i}} + (X - X_1)^\top b_1 - \sum_{i \in s_2} \frac{x_i^\top b_1}{\pi_{2i}} \\
&= \sum_{i \in s_1} y_i + \sum_{i \in s_2} \frac{y_i}{\pi_{2i}} + (X_1^c - \widehat{X}_1^c)^\top b_1 = \sum_{i \in s} w_i y_i
\end{aligned}$$

where

$$w_i = \begin{cases} 1 + (X_1^c - \widehat{X}_1^c)^\top \big(\sum_{i \in s_1} x_i x_i^\top / \pi_{1i} \big)^{-1} x_i / \pi_{1i} & \text{if } i \in s_1 \\ 1/\pi_{2i} & \text{if } i \in s_2 \end{cases}$$

It follows that the SRB estimator (2) can be given as a linear estimator

$$\widehat{Y}^*_{GR} = \sum_{i \in s} w_i^* y_i \qquad \text{where} \quad w_i^* = E_q(w_i | s) \tag{4}$$

This has an important practical advantage that $\{w_i^* : i \in s\}$ can be applied to produce numerically consistent cross-tabulation of multiple survey variables of interest.

In the case of SRS of s_1 with $n_1 = n - 1$, the SRB weight w_i^* in (4) is the average of w_i's over n possible subsamples s_1, for a given unit $i \in s$, where $w_i = 1/\pi_{2i}$ when s_1 does not include the unit i, otherwise w_i is the corresponding GREG weight for $Y_1^c = \sum_{k \notin s_1} y_k$, which is different for each of the rest $n - 1$ subsamples that includes the unit i.

2.3 Relative efficiency to GREG

Let $B = E_p(b)$ and $e_i = y_i - x_i^\top B$ for $i \in U$. Expanding the GREG estimator $\widehat{Y}_{GR}$ around (Y, X, B) yields

$$\widehat{Y}_{GR} \approx \sum_{i \in s} \frac{e_i}{\pi_i} + B^\top X$$

For $\widehat{Y}_1$, the first two terms on the right-hand side of (1) becomes $X^\top b_1$ if there exists a vector λ such that $x_i^\top \lambda \equiv 1$, in which case $\widehat{Y}_1$ is a function of $(\widehat{Y}_1^c, \widehat{X}_1^c, b_1)$, i.e.

$$\widehat{Y}_1 = X^\top b_1 + \widehat{Y}_1^c - b_1^\top \widehat{X}_1^c = \widehat{Y}_1^c + b_1^\top (X - \widehat{X}_1^c)$$

where $\widehat{Y}_1^c = \sum_{i \in s_2} y_i/\pi_{2i}$ is conditionally unbiased for $Y_1^c = \sum_{i \notin s_1} y_i$ given s_1, and similarly $\widehat{X}_1^c = \sum_{i \in s_2} x_i/\pi_{2i}$ for $X_1^c = \sum_{i \notin s_1} x_i$. Let $Y_+^c = E_1(Y_1^c)$ and $X_+^c = E_1(X_1^c)$. We have $E_1(b_1) \approx B$, since b_1 and b aim at the same population parameter, especially if n_1 is close to n. In any case, expanding $\widehat{Y}_1$ around (Y_+^c, X_+^c, B) yields

$$\begin{aligned}\widehat{Y}_1 &\approx \Big(Y_+^c + B^\top (X - X_+^c)\Big) + \Big((\widehat{Y}_1^c - Y_+^c) + (b_1 - B)^\top (X - X_+^c) - B^\top (\widehat{X}_1^c - X_+^c)\Big) \\ &= \big(\widehat{Y}_1^c - B^\top \widehat{X}_1^c\big) + b_1^\top (X - X_+^c) + B^\top X_+^c\end{aligned}$$

and

$$\widehat{Y}_1^c - B^\top \widehat{X}_1^c = \sum_{i \in s_2} e_i/\pi_{2i} = \sum_{i \in s} \delta_{2i}^* e_i/\pi_{2i}$$

where $\delta_{2i}^* = 1$ if $i \in s_2$ and 0 of $i \in s_1$. Thus, we obtain

$$\widehat{Y}_{GR}^* = E_q(\widehat{Y}_1|s) \approx \sum_{i \in s} E_q\big(\frac{\delta_{2i}^*}{\pi_{2i}}|s\big)\pi_i \frac{e_i}{\pi_i} + E_q(b_1|s)^\top (X - X_+^c) + B^\top X_+^c \tag{5}$$

Notice that $B^\top X_+^c$ is a constant. Thus, compared to $\widehat{Y}_{GR}$, the variance of $\widehat{Y}_{GR}^*$ involves that of $E_q(b_1|s)$ in addition. As $n, N \to \infty$, the first term on the right-hand side of (5) is $O_p(N/\sqrt{n})$ provided $\pi_i E_q(\delta_{2i}^*/\pi_{2i}) = O_p(1)$, whereas the second term is $O_p(\sqrt{n})$ if $n_1/n = O(1)$ provided the usual regularity conditions for GREG. As long as the sampling fraction n/N is small, the first term will dominate, in which case the variance of the SRB estimator $\widehat{Y}_{GR}^*$ is of the same order as that of the GREG estimator $\widehat{Y}_{GR}$.

Example: SRS, cont'd Let $n_1 = n - k$, where $k = |s_2|$. We have

$$\pi_i \pi_{2i}^{-1} E_q(\delta_{2i}^*|s) = \frac{n}{N} \cdot \frac{N - n_1}{k} \cdot \frac{(n-1)!}{(k-1)!(n-k)!} \cdot \frac{k!(n-k)!}{n!} = 1 - \frac{n_1}{N}$$

Let S_e^2 be the population variance of $\{e_i : i \in U\}$. The variance of the first-term in (5) is

$$V_p\Big(\sum_{i \in s} E_q\big(\frac{\delta_{2i}^*}{\pi_{2i}}|s\big)\pi_i \frac{e_i}{\pi_i}\Big) = N^2\big(1 - \frac{n}{N}\big)\frac{S_e^2}{n}\big(1 - \frac{n_1}{N}\big)^2$$

which is actually smaller than the approximate variance of the GREG estimator under SRS, although the difference will not be noteworthy in practical terms, if the sampling fraction n/N is small, since $1 - n/N < 1 - n_1/N < 1$. Meanwhile, due to the additional variance of $E_q(b_1|s)$, the estimator $\widehat{Y}_{GR}^*$ by unbiased SRB method can possibly have a

larger variance than the biased GREG (with general x_i). It seems that one should use large n_1 if possible, to keep the additional variance due to $E_q(b_1|s)$ small.

2.4 Delete-one RB method

The largest possible size of s_1 is $n_1 = n - 1$. We refer to Rao-Blackwellisation based on SRS of s_1 with $n_1 = n - 1$ as the *delete-one* (or leave-one-out, LOO) RB method. The conditional sampling design $p_2(s_2|s_1)$ is not measurable in this case, in that one cannot have an unbiased variance estimator $\widehat{V}(\widehat{Y}_1)$ based on a single observation y_j in $s_2 = \{j\}$. For an approximate variance estimator, we reconsider the basic case where $\{y_1, ..., y_n\}$ form a sample of independent and identically distributed (IID) observations, in order to develop an analogy to the classic jackknife variance estimation (Tukey, 1958).

Denote by θ the population mean that is also the expectation of each y_i, for $i = 1, ..., n$. As before, let $\bar{y}_{(j)}$ denote the mean in the subsample $s_1 = s \setminus \{j\}$. Following (1), let

$$\hat{\theta}_{(j)} = \frac{n-1}{N}\bar{y}_{(j)} + \big(1 - \frac{n-1}{N}\big)y_j$$

be the delete-j estimator of θ, where y_j acts as an unbiased estimator of the population mean outside s_1. The RB method yields the whole sample mean, denoted by

$$\hat{\theta}^* = \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_{(j)} = \frac{1}{n}\sum_{j=1}^{n} y_j = \bar{y}$$

Observe that we have $\hat{\theta}^* = \sum_{j=1}^{n} z_{(j)}/n$, where

$$z_{(j)} = \frac{1}{N-n}\big(N\hat{\theta}_{(j)} - n\hat{\theta}^*\big) = y_j \tag{6}$$

Thus, the RB estimator $\hat{\theta}^*$ is the mean of an IID sample of observations $z_{(j)}$, for $j = 1, ..., n$, as in the development of classic jackknife variance estimation, so that we obtain

$$\widehat{V}(\hat{\theta}^*) = \frac{1}{n(n-1)}\sum_{j=1}^{n}\big(z_{(j)} - \hat{\theta}^*\big)^2$$

Notice that, in this case, the IID observations used for the classic development of jackknife method are given by $z_{(j)} = n\hat{\theta} - (n-1)\hat{\theta}_{(j)} = y_j$ instead of (6), where $\hat{\theta}_{(j)} = \bar{y}_{(j)}$.

For the delete-one RB method based on (1) and (2) given auxiliary $\{x_i : i \in U\}$, we have $\pi_{1i} = \pi_i(n-1)/n$, such that the estimator b_1 can be denoted by $b_{(j)}$, based on $s_1 = s \setminus \{j\}$, where it is simply the delete-j jackknife regression coefficients estimator.

Rewrite the corresponding population total estimator $\widehat{Y}_1$ by (1) as

$$\widehat{Y}_{(j)} = X^\top b_{(j)} + \frac{y_j}{\pi_{2j}} - \frac{x_j^\top b_{(j)}}{\pi_{2j}}$$

such that the RB method yields $\widehat{Y}^*_{GR}$ by (2), as the mean of $\widehat{Y}_{(j)}$ over $j = 1, ..., n$. We propose a jackknife variance estimator for $\widehat{Y}^*_{GR}$, given by

$$\widehat{V}(\widehat{Y}^*_{GR}) = \frac{N^2}{n(n-1)} \sum_{j=1}^{n} \big(z_{(j)} - \frac{1}{n}\sum_{i=1}^{n} z_{(i)}\big)^2 \tag{7}$$

where

$$z_{(j)} = \frac{1}{N-n}\big(\widehat{Y}_{(j)} - \frac{n}{N}\widehat{Y}^*_{GR}\big)$$

Notice that it may be the case under general unequal probability sampling that the conditional inclusion probability π_{2j} given $s_1 = s \setminus \{j\}$ is not exactly known. However, in many situations where the sampling fraction is low, it is reasonable that

$$\pi_{2j} \approx \frac{\pi_j}{\sum_{i \notin s_1} \pi_i} = \frac{\pi_j}{n - \sum_{i \in s_1} \pi_i} \approx \frac{\pi_j}{n(1 - n_1/N)}$$

An approximate delete-one RB estimator following (2) can then be given as

$$\widetilde{Y}^*_{GR} = X^\top \sum_{j=1}^{n} \frac{b_{(j)}}{n} + \big(1 - \frac{n_1}{N}\big) \sum_{j=1}^{n} \frac{y_j}{\pi_j} - \big(1 - \frac{n_1}{N}\big) \sum_{j=1}^{n} \frac{x_j^\top}{\pi_j} b_{(j)} \tag{8}$$

with $\widetilde{Y}_{(j)}$ for jackknife variance estimation on replacing $1/\pi_{2j}$ by $n(1 - n_1/N)/\pi_j$. Meanwhile, the delete-one jackknife replicates of GREG $\widehat{Y}_{GR}$ can be written as

$$\widehat{Y}^{(j)}_{GR} = X^\top b_{(j)} + \frac{n}{n-1}\big(\sum_{i \neq j} \frac{y_i}{\pi_i} - \sum_{i \neq j} \frac{x_i^\top b_{(j)}}{\pi_i}\big)$$

$$\widehat{Y}^{(\cdot)}_{GR} = \frac{1}{n}\sum_{j=1}^{n} \widehat{Y}^{(j)}_{GR} = X^\top \sum_{j=1}^{n} \frac{b_{(j)}}{n} + \sum_{i=1}^{n} \frac{y_i}{\pi_i} - \sum_{i=1}^{n} \frac{x_i^\top}{\pi_i}\big(\sum_{j \neq i} \frac{b_{(j)}}{n-1}\big)$$

The estimator $\widehat{Y}^{(\cdot)}_{GR}$ is quite close to $\widetilde{Y}^*_{GR}$ by (8); indeed, identical apart from $1 - n_1/N$ in the special case of $x_i^\top/\pi_i = N/n$. This is not surprising, since the jackknife-based $\widehat{Y}^{(\cdot)}_{GR}$ is an alternative for reducing the bias of the GREG estimator. The difference is that, provided π_{2j} is known, the proposed SRB method will be exactly design-unbiased, but not the jackknife-based $\widehat{Y}^{(\cdot)}_{GR}$. Finally, the resemblance between $\widetilde{Y}^*_{GR}$ and $\widehat{Y}^{(\cdot)}_{GR}$ is another indication that the relative efficiency of the delete-one RB method is usually not a concern compared to the standard GREG estimator $\widehat{Y}_{GR}$.

2.5 Monte Carlo RB

Exact Rao-Blackwellisation can be computationally expensive, when the cardinality of the subsample space (of s_1) is large. Instead of calculating the SRB estimator exactly, consider the Monte Carlo (MC) SRB estimator given as follows:

$$\widehat{Y}_{GR}^{K} = K^{-1} \sum_{k=1}^{K} \widehat{Y}_{1k} \tag{9}$$

where $\widehat{Y}_{1k}$ is the estimator $\widehat{Y}_1$ based on the kth subsample, for $k = 1, ..., K$, which are realisations of s_1 from $q(s_1|s)$, such that $\widehat{Y}_{GR}^{K}$ is a Monte Carlo approximation of $\widehat{Y}_{GR}^{*}$.

Proposition 4 The estimator $\widehat{Y}_{GR}^{K}$ is design-unbiased for Y, denoted by $E(\widehat{Y}_{GR}^{K}) = Y$.
Proof: The result follows from $E(\widehat{Y}_{1k}) = Y$. □

Adopting a computationally manageable K entails an increase of variance, i.e. $V_q(\widehat{Y}_{GR}^{K}|s)$, compared to $\widehat{Y}_{GR}^{*}$, so that the variance of $\widehat{Y}_{GR}^{K}$ is given by

$$V(\widehat{Y}_{GR}^{K}) = E_1\big(V_2(\widehat{Y}_1|s_1)\big) - E_p\big(V_q(\widehat{Y}_1|s)\big) + E_p\big(V_q(\widehat{Y}_{GR}^{K}|s)\big) \tag{10}$$

Due to the IID construction of $\widehat{Y}_{1k}$, an unbiased estimator of $V_q(\widehat{Y}_{GR}^{K}|s)$ is given by

$$\widehat{V}_q(\widehat{Y}_{GR}^{K}|s) = \frac{1}{K(K-1)} \sum_{k=1}^{K} (\widehat{Y}_{1k} - \widehat{Y}_{GR}^{K})^2$$

This allows one to control the statistical efficiency of the MC-SRB method, i.e. the choice of K is acceptable when $\widehat{V}_q(\widehat{Y}_{GR}^{K}|s)$ is deemed small enough in practical terms.

Proposition 5 Provided unbiased variance estimator $\widehat{V}(\widehat{Y}_{1k})$ with respect to $p_2(s_2|s_1)$, i.e. $E_2\big(\widehat{V}(\widehat{Y}_{1k})\big) = V_2(\widehat{Y}_1|s_1)$, a design-unbiased variance estimator for $\widehat{Y}_{GR}^{K}$ is given by

$$\widehat{V}(\widehat{Y}_{GR}^{K}) = \frac{1}{K} \sum_{k=1}^{K} \widehat{V}(\widehat{Y}_{1k}) - \frac{1}{K} \sum_{k=1}^{K} (\widehat{Y}_{1k} - \widehat{Y}_{GR}^{K})^2$$

Proof: Due to the IID construction of $\widehat{Y}_{1k}$, $K^{-1} \sum_{k=1}^{K} \widehat{V}(\widehat{Y}_{1k})$ is an unbiased estimator of the first term on the right-hand side of (10), while $(K-1)^{-1} \sum_{k=1}^{K} (\widehat{Y}_{1k} - \widehat{Y}_{GR}^{K})^2$ is an unbiased estimator of the second term. The result follows. □

Finally, for the delete-one RB method, where unbiased variance estimator $\widehat{V}(\widehat{Y}_1)$ is not available now that $|s_2| = 1$, a practical option is to first apply the jackknife variance estimator (7) to the K samples, as if $\widehat{Y}_{GR}^{K}$ where the exact SRB estimator $\widehat{Y}_{GR}^{*}$, and then add to it the extra term $\widehat{V}_q(\widehat{Y}_{GR}^{K}|s)$ for the additional Monte Carlo error. This would allow one to use the Monte Carlo delete-one RB method in general.

3 Unbiased learning generally

In this section we consider design-unbiased estimation in survey sampling, which builds on arbitrary ML technique that can be non-linear as well as non-parametric.

3.1 Design-unbiased ML for survey sampling

Denote by M the model or algorithm that aims to predict y_i given x_i. Let s_1 be the training set, and $s_2 = s \setminus s_1$ the test set. Let $\widehat{M}$ be the trained model based on $\{(x_i, y_i) : i \in s_1\}$, yielding $\mu(x_i, s_1)$ as the corresponding predictor of y_i given x_i. Apply the trained model to $i \in s_2$ yields the *prediction errors* of $\widehat{M}$ *conditional on* s_1, denoted by $e_i = y_i - \mu(x_i, s_1)$. In contrast, the *same* discrepancy is referred to as the *residuals* of $\widehat{M}$, when it is calculated for $i \in s_1$, denoted by $\hat{e}_i = y_i - \mu(x_i, s_1)$, including when the training set s_1 is equal to s. In standard ML, the errors in the test set are used to select different trained algorithms, or to assess how well a trained algorithm can be expected to perform when applied to the units with unknown y_i's.

From an inference point of view, a basic problem with the standard ML approach above arises because one needs to be able to 'extrapolate' the information in $\{e_i : i \in s_2\}$ to the units outside s, in order for supervised learning to have any value at all. This is simply because $\{y_i : i \in s\}$ are all observed and prediction in any form is *unnecessary* for $i \in s$. No matter how the training-test split is carried out, one cannot ensure valid $\mu(x_k, s_1)$ for $k \notin s$, *unless* s is selected from the entire reference set of units, i.e. the population U, in some non-informative (or representative) manner. This is the well-known problem of observational studies in statistical science, which is sometimes recast as the problem of concept drift in the ML literature (e.g. Tsymbal, 2004).

A *design pq-unbiased* approach to *ML-assisted estimation* of population total Y can be achieved with respective to

(i) a probability sample s from U, with probability $p(s)$, and

(ii) a probabilistic scheme $q(s_1|s)$ for the training-test split (s_1, s_2) given s.

Explicitly, let $\widehat{Y}_{1M}$ be the estimator of Y obtained from the realised sample s and sub-sample s_1 given the model M. It is said to be design pq-unbiased for Y, provided

$$E_{pq}(\widehat{Y}_{1M}) = \sum_{s} p(s) \sum_{s_1 \subset s} q(s_1|s) \widehat{Y}_{1M} = Y$$

where E_{pq} is the expectation of $\widehat{Y}_{1M}$ over all possible (s, s_1). Replacing the linear predictor $x_i^\top b_1$ in (1) by any predictor $\mu(x_i, s_1)$ trained on s_1, we obtain

$$\widehat{Y}_{1M} = \sum_{i \in s_1} y_i + \sum_{i \in U \setminus s_1} \mu(x_i, s_1) + \sum_{i \in s_2} e_i / \pi_{2i} \tag{11}$$

Proposition 6 $\widehat{Y}_{1M}$ by (11) is design pq-unbiased for Y using an arbitrary model M.

The proof is parallel to that for $\widehat{Y}_1$ by (1), only that $\mu(x, s_1)$ is now based on any chosen model M. It is important to point out that the purpose here is to estimate Y at the population level, instead of individual prediction *per se*. Indeed, $\widehat{Y}_{1M}$ is design-unbiased, regardless M is a strong or weak learner. The underlying probabilistic mechanism consists of two necessary elements: $p(s)$ ensures valid extrapolation of learning to the units outside s, since otherwise completely model-based prediction $\sum_{i\in s} y_i + \sum_{i\in U\setminus s} \mu(x_i, s)$ has no guaranteed relevance to Y no matter how the training set s_1 is chosen or how M is selected, whereas subsampling $q(s_1|s)$ is required to be able to project the errors in s_2 to the aggregated level, since projecting the residuals in s in the manner of GREG estimator $\widehat{Y}_{GR}$ (i.e. without the training-test split) would not achieve unbiasedness exactly.

3.2 Subsampling RB and bootstrap aggregating

There is a natural affinity between the subsampling RB method and bootstrap aggregating (i.e. *bagging*). Bagging is originally devised to improve unstable leaners (Breiman, 1996a; 1996b) for individual prediction, where the aggregation averages the learner over bootstrap replicates of the training set. The argument can be adapted to design-based population-level estimation. Let $\widehat{Y}_M = \psi(s; M)$ be an ML-assisted estimator of Y, which varies over different samples s. Insofar as $\{y_i : i \in U\}$ are treated as unknown constants and $\widehat{Y}_M$ is uniquely determined given $\{(y_i, x_i) : i \in s\}$, the only variation of $\widehat{Y}_M$ derives from that of the sample s. For some model M, such as regression tree with random feature selection, there exists an extra variation of $\widehat{Y}_M$ given s. In any case, let the expectation of $\widehat{Y}_M$ be

$$\psi_M = E(\widehat{Y}_M) = E\big(\psi(s; M)\big)$$

over all possible s and additional randomness given s. We have

$$E\big((\widehat{Y}_M - Y)^2\big) = (\psi_M - Y)^2 + E\big((\widehat{Y}_M - \psi_M)^2\big)$$

since $E(\widehat{Y}_M - \psi_M) = 0$ by definition. Thus, ψ_M has always a smaller mean squared error than $\widehat{Y}_M$. Notice that in reality bagging is "caught in two currents" (Breiman, 1994): the improvement can be appreciable if $\widehat{Y}_M$ is unstable, whereas the additional estimation of ψ_M by bagging may not be worthwhile if $\widehat{Y}_M$ is a stable learner to start with.

It is clear from the above that, while it can reduce the variance of unbagged predictor, bagging does not affect the potential bias, now that it aims at replacing $\widehat{Y}_M = \psi(s; M)$ by its expectation ψ_M. The SRB method is more effectual than bagging in the following sense: on the one hand, it leads generally to design-unbiased estimation of Y, which does not result from bagging alone; on the other hand, Rao-Blackwellising $\widehat{Y}_{1M}$ reduces its variance, even when it is based on a stable learner, such as $\mu(x_i, s_1) = x_i^\top b_1$, which

bagging does only for unstable $\widehat{Y}_M$. Replacing $\widehat{Y}_1$ in (2) with $\widehat{Y}_{1M}$ given by (11) generally, the SRB ML-assisted estimator of Y is given by

$$\widehat{Y}_M^* = E_q(\widehat{Y}_{1M}|s) \tag{12}$$

Proposition 7 The SRB ML-assisted estimator $\widehat{Y}_M^*$ by (12) is design p-unbiased, or simply design-unbiased, for Y using an arbitrary model M.

The proof is exactly parallel to that for $\widehat{Y}_{GR}^*$ by (2). Notice that Rao-Blackwellisation of $\widehat{Y}_{1M}$ with respect to $q(s_1|s)$ can accommodate straightforwardly any additional variation given s_1 due to the chosen model M. For example, given a subsample s_1, one can grow a regression tree with random feature selection. Despite the resulting $\mu(x, s_1)$ is not fixed for the given s_1, the corresponding $\widehat{Y}_{1M}$ is still design pq-unbiased, because it is conditionally unbiased for Y given s_1 *and* the outcome of random feature selection, and Y is a constant with respect to subsampling of s_1 and random feature selection given s_1.

Finally, MC-SRB is operationally similar to bagging, involving about the same amount of computation effort. In bagging, one draws a bootstrap replicate sample from s; whereas in SRB, one resamples s_1 from s according to $q(s_1|s)$. In either case, one trains the model based on the resample. Repeating the two steps K times yields the bagged predictor by bagging, and the MC-SRB estimator by subsampling RB. The choice of K balances between numerical and statistical efficiency.

3.3 Design consistency

Provided $n_2 = |s_2| \geq 2$, let $\widehat{V}(\widehat{Y}_{1M,k})$ be an unbiased variance estimator with respect to $p_2(s_2|s_1)$, i.e. $E_2\big(\widehat{V}(\widehat{Y}_{1M,k})\big) = V_2(\widehat{Y}_{1M,k})$, for K subsamples $k = 1, ..., K$. A design-unbiased variance estimator for MC-SRB estimator $\widehat{Y}_M^K$ is given by

$$\widehat{V}(\widehat{Y}_M^K) = \frac{1}{K}\sum_{k=1}^{K}\widehat{V}(\widehat{Y}_{1M,k}) - \frac{1}{K}\sum_{k=1}^{K}(\widehat{Y}_{1M,k} - \widehat{Y}_M^K)^2 \tag{13}$$

similarly as for $\widehat{Y}_{GR}^K$. It is an open question at this stage how to determine the efficient subsampling scheme $q(s_1|s)$, including the choice n_1. But given the simplicity and practical advantage of the delete-one GREG-assisted $\widehat{Y}_{GR}^K$, any other ML-assisted estimator would not be worth considering, *unless* it has clearly a smaller estimated variance.

Design consistency is thus not a worrisome issue for design-unbiased statistical learning in practice. Still, an unbiased estimator is consistent, provided its sampling variance tends to 0 under a suitable asymptotic setting. As mentioned in Section 1, below we develop stability conditions first for the delete-one RB estimator (12) under the special case of SRS and, then, under general unequal probability sampling design.

3.3.1 Stability condition: SRS

Let $s_j = s \setminus \{j\}$ be the delete-j sample. Let $s_{ij} = s \setminus \{i, j\}$ be the delete-ij sample. Let $\mu(x, s_j)$ be the predictor given x, which is trained on s_j, and $\mu(x, s_{ij})$ that on s_{ij}. We define $\mu(x, s)$ to be *twice q-stable*, if

$$\mu(x_k, s_j) - \mu(x_k, s) \xrightarrow{P} 0 \quad \text{and} \quad \mu(x_k, s_{ij}) - \mu(x_k, s_j) \xrightarrow{P} 0 \tag{14}$$

i.e. convergence in probability, as $n, N \to \infty$ asymptotically, for any $i, j \in s$ and $k \in U$, where s_j results from delete-one q-sampling from s, and s_{ij} from recursive q-sampling where one randomly deletes $i \in s_j$. Notice that the first part of (14) is analogous to the 'point-wise hypothesis leave-one-out stability' of Mukherjee et al. (2006).

Theorem 1: The delete-one RB estimator (12) is consistent for population mean $\bar{Y}$ under SRS, as $n, N \to \infty$, given twice q-stability and $y_i - \mu(x_i, s) = O(1)$ for any s.
For a proof see Appendix A.

3.3.2 Stability condition: Unequal probability sampling

For general unequal probability sampling, we define the following the stability conditions. First, we define $\mu(x, s)$ to be simply *q-stable* if, for any $j \in s$ and $k \in U$, we have

$$\mu(x_k, s_j) - \mu(x_k, s) \xrightarrow{P} 0 \tag{15}$$

asymptotically as $n, N \to \infty$, where s_j results from delete-one subsampling $q(s_j|s)$. Next, we define $\mu(x, s)$ to be *p-stable* for the delete-one RB method, if

$$\frac{1}{n} \sum_{j \in s} \frac{\widehat{N}_j}{N} \mu(x_j, s) - \frac{1}{N} \sum_{k \in U} \mu(x_k, s) \xrightarrow{P} 0 \tag{16}$$

where $\widehat{N}_j = \pi_{2j}^{-1} + (n-1)$ is an estimator of N based on $s_2 = \{j\}$. Notice that, given q-stability (15), it is possible to replace p-stability (16) by a pq-stability condition

$$\frac{1}{n} \sum_{j \in s} \frac{\widehat{N}_j}{N} \mu(x_j, s_j) - \frac{1}{N} \sum_{k \in U} \mu(x_k, s_j) \xrightarrow{P} 0$$

which reduces to $\sum_{j \in s} \mu(x_j, s_j)/n - \sum_{k \in U} \mu(x_k, s_j)/N \xrightarrow{P} 0$ under SRS, and resembles the IID 'expected-leave-one-out stability' of Mukherjee et al. (2006): the first term above is the empirical average in the observed set in both definitions, whereas for the second term here we replace averaging over $\mathcal{Z}$ in the IID setting by that over the population distribution function, which places point mass $1/N$ on each $k \in U$.

Some regularity condition on the sampling design $p(s)$ is needed to for the general

situation. Let $\widehat{\bar{Y}}_j = y_j \widehat{N}_j / N$ be the *leave-one-out (LOO) HT* estimator of population mean $\bar{Y}$ based on $s_2 = \{j\}$. We define the sampling design to be *LOO-consistent*, if

$$\frac{1}{n} \sum_{j \in s} \widehat{\bar{Y}}_j \ \stackrel{P}{\to} \ \bar{Y} \tag{17}$$

asymptotically as $n, N \to \infty$. The condition is specified for the LOO-RB-HT estimator, where $\sum_{j \in s} \widehat{\bar{Y}}_j / n = E_q(\widehat{\bar{Y}}_j | s)$. Under SRS, $\sum_{j \in s} \widehat{\bar{Y}}_j / n = \sum_{j \in s} y_j / n$ is the sample mean, which converges to $\bar{Y}$ in probability, provided $y_i = O(1)$ for all $i \in U$. We emphasise that the condition (17) concerns only the sampling design $p(s)$, since it is formulated in terms of the y-values alone, i.e. based on an 'empty' predictor, so to speak.

Theorem 2: The delete-one RB estimator (12) is consistent for population mean $\bar{Y}$, as $n, N \to \infty$, provided q- and p-stabilities, and LOO-consistent sampling design $p(s)$.
Proof: Given the delete-j sample s_j under any general sampling design, we can write

$$\begin{aligned}
\widehat{Y}_{1M}(s_j) &= \sum_{i \in s} y_i + \sum_{k \notin s} \mu(x_k, s_j) + (\pi_{2j}^{-1} - 1)\big(y_j - \mu(x_j, s_j)\big) \\
&= \sum_{i \in s} \big(y_i - \mu(x_i, s_j)\big) + \sum_{k \in U} \mu(x_k, s_j) + (\pi_{2j}^{-1} - 1)\big(y_j - \mu(x_j, s_j)\big)
\end{aligned}$$

where π_{2j} is the conditional probability of selecting j from s_j^c given s_j. Given q-stability (15), i.e. $\mu(x, s_j) = \mu(x, s) + o_p(1)$, the RB-estimator of the population mean is

$$\begin{aligned}
\widehat{\bar{Y}}_M^* &= \frac{1}{N} \sum_{k \in U} \mu(x_k, s) + \frac{1}{n} \sum_{j \in s} \frac{\widehat{N}_j}{N} \big(y_j - \mu(x_j, s)\big) + o_p(1) \\
&= \{\frac{1}{N} \sum_{k \in U} \mu(x_k, s) - \frac{1}{n} \sum_{j \in s} \frac{\widehat{N}_j}{N} \mu(x_j, s)\} + \big[\frac{1}{n} \sum_{j \in s} \widehat{\bar{Y}}_j\big] + o_p(1) \ .
\end{aligned}$$

The result follows from applying the p-stability condition (16) to the expression in the brackets $\{\}$, and the LOO-consistency condition (17) to that in $[\ ]$. $\square$

4 Simulations

Below we present and discuss some simulation results of the delete-one RB (or LOO-RB) method, and the associated jackknife variance estimator described in Section 2.4. The HT and some GREG estimators are computed for comparisons. The target is always the population mean (denoted by θ) in a given set-up. The simulations proceed as follows.

- B samples (usually $B = 100$) are drawn independently from the given fixed population according to a specified sampling design.

- We obtain an estimate $\hat{\theta}_{(b)}$ based on each sample, for $b = 1, ..., B$. In particular, for the LOO-RB method, we calculate its associated jackknife variance estimate $v_{(b)}$.

- An estimate of $E(\hat{\theta})$ over repeated sampling is $\hat{\bar{\theta}} = \sum_{b=1}^{B} \hat{\theta}_{(b)}/B$, with associated Monte Carlo error $\sqrt{v/B}$, where $v = \sum_{b=1}^{B}(\hat{\theta}_{(b)} - \hat{\bar{\theta}})^2/(B-1)$. An estimate of its bias is $\hat{\bar{\theta}} - \theta$; an estimate of its root mean squared error (RMSE) is $\{\sum_{b=1}^{B}(\hat{\theta}_{(b)} - \theta)^2/B\}^{1/2}$.

- Similarly for the bias and RMSE of the variance estimator $v_{(b)}$, except that the true variance of the LOO-RB method is unknown and is replaced by its estimate v.

Now that the HT estimator and the LOO-RB methods are unbiased, an inspection of their respective simulation-based bias estimates and the associated Monte Carlo errors can usually provide adequate information, in order to judge whether a certain conclusion of the results is warranted given the actual number of simulations.

4.1 Simulations with synthetic data

The GREG estimator has become the standard-bearer in practical survey sampling in the past three decades. Using simple simulations below, we would like to gain some basic appreciation of the pros and cons of the corresponding LOO-RB-GREG estimator, given by (2), under the proposed unbiased learning approach. Small synthetic populations are generated based on only two regressors. The first regressor x_1 follows a log-normal distribution with mean and variance both set to one. The second regressor x_2 follows a Poisson distribution with mean 5. The target y-variable in each setting is generated as the absolute value of a certain function of x_1 and x_2 plus a regression error.

Table 1: Simulation results of HT, GREG and LOO-RB-GREG estimator, by two different sampling designs. Monte Carlo errors of bias estimates in parentheses.

	Simple Random Sampling			Probability Proportional to x_2		
Estimator	Bias (MC Error)		RMSE	Bias (MC Error)		RMSE
HT	0.08	(0.19)	1.91	0.12	(0.20)	2.02
GREG	-0.09	(0.13)	1.29	0.03	(0.16)	1.60
LOO-RB-GREG	0.10	(0.14)	1.39	0.16	(0.16)	1.63
Variance by jackknife	*0.52*	*(0.10)*	*1.09*	*4.73*	*(1.02)*	*11.23*

We start with a setting where the GREG estimator should have a negligible or very small bias. Let the population size be 200, and let the target survey y-variable be the absolute value of $1.5x_1 + x_2 + \epsilon$, where ϵ follows a normal distribution with zero mean and variance that is a quarter of the variance of x_1. Let the sample size be 20. Two sampling designs are used: SRS, or conditional Poisson sampling with probabilities proportional to x_2 as the size variable. The results are given in Table 1.

It can be seen that under both sampling designs, GREG and LOO-RB-GREG have essentially the same efficiency, and both outperform HT estimation. Recall that the bias

of the GREG estimator is negligible in this scenario because of the underlying linear population model. Clearly, the jackknife variance estimator (7), which is derived as a direct analogy to the IID-sample situation, needs to be modified for unequal probability sampling designs such as the conditional Poisson sampling here.

Table 2: Simulation results of HT, GREG and LOO-RB-GREG estimator under SRS, but two different population models. Monte Carlo errors of bias estimates in parentheses.

	$V(y) \propto x_1$, $n = 5$			Non-linear, $V(y) \propto \sqrt{x_1}$, $n = 20$		
Estimator	Bias (MC Error)		RMSE	Bias (MC Error)		RMSE
HT	-0.46	(0.50)	5.03	0.95	(1.26)	12.57
GREG	-0.82	(0.41)	4.16	-2.41	(0.51)	5.62
LOO-RB-GREG	-0.68	(0.77)	7.69	0.68	(0.86)	8.62
Variance by jackknife	*-13.22*	*(15.75)*	*157.31*	*-7.36*	*(10.50)*	*104.76*

Consider now two potentially problematic settings. First, we introduce heteroscedasticity by make the variance of the y-variable proportional to x_1, while reducing the sample size at the same time, where $n = 5$ (from $N = 100$). The results are given in the left part of Table 2. The LOO-RB-GREG is the least efficient estimator here: the heteroscedasticity setting increases the variance of $\widehat{Y}_1$ based on each subsample, whereas the small sample size implies RB averaging over only 5 subsample estimates (instead of 20 above). The RMSE of the jackknife variance estimator is much bigger for similar reasons.

Next, reverting to $(N, n) = (200, 20)$, we generate the target y-variable non-linearly as the absolute value of $0.5x_1 + 0.25x_1^2 + x_2 + \epsilon$, where ϵ follows a normal distribution with zero mean and variance proportional to $\sqrt{x_1}$. The results under SRS are given in the right part of Table 2. The GREG estimator has now a relatively large bias, which is removed by the LOO-RB-GREG estimator. However, the unbiased learning estimator loses efficiency compare to GREG in terms of the MSE, although it is still much better than the HT estimator. The performance of the variance estimator is similar as before.

These results illustrate the basic pros and cons of delete-one RB-GREG vs. standard GREG estimation. On the one hand, the GREG estimator may suffer from non-negligible bias, e.g. because one applies the assisting linear model in a routine manner without conducting careful model diagnostics as one should, whereas the unbiased learning approach avoids the bias by definition. On the other hand, delete-one subsampling may suffer from loss of efficiency given heteroscedastic observations in very small samples.

4.2 Simulations with real data

The population consists of a sample of about 17000 small and medium-sized enterprises from the Spanish Structural Business Survey (SSBS). As the target variables we consider three survey variables collected in the SSBS: Turnover, Total personnel expenses and Total procurements of goods and services. Seventeen variables from the administrative corporate income tax data are imported as the regressors. One of them is turnover, although

for many enterprises the turnover from tax data will be different to the turnover from SSBS by definition; in addtion the two observed values may differ because of registration delays or other operational reasons. The estimators to be considered are: HT, GREG1 with one regressor (turnover), GREG17 with the seventeen regressors (as main effects), LOO-RB-GREG1 with one regressor (turnover), and LOO-RB random forest (RF) with seventeen features. When only one regressor is used, RF is not a good option to be included here. Jackknife variance estimation is applied to the two SRB estimators.

4.2.1 Turnover

This case is interesting because turnover (from tax data) is one of the regressors. We consider SRS and stratified SRS designs. For the latter, three strata are created by the number of employees, which is a commonly used stratification variable in SBS, although the actual designs always have many other complicating details in practice. The stratum sample sizes are allocated proportionally to the stratum population sizes. The total sample size is 10% of the population under both the designs. The simulation results are given in Table 3, similarly as before and suitably scaled for presentation.

Table 3: Simulation results for Turnover estimation by different estimators under two sampling designs. Monte Carlo errors of bias estimates in parentheses.

	SRS			Stratified SRS		
Estimator	Bias (MC Error)		RMSE	Bias (MC Error)		RMSE
HT	-0.22	(0.47)	4.65	-0.36	(0.32)	3.20
GREG1	0.01	(0.26)	2.56	0.30	(0.22)	2.16
LOO-RB-GREG1	-0.21	(0.27)	2.68	0.05	(0.23)	2.31
Variance LOO-RB-GREG1	*-1.02*	*(0.32)*	*3.37*	*0.60*	*(0.32)*	*3.27*
GREG17	0.42	(0.22)	2.30	0.79	(0.75)	7.47
LOO-RB-RF	-0.19	(0.14)	1.37	-0.10	(0.14)	1.38
Variance LOO-RB-RF	*0.09*	*(0.03)*	*0.34*	*-0.04*	*(0.04)*	*0.38*

The LOO-RB-RF (by random forest) is the most efficient estimator under SRS. It is more efficient than the GREG17 estimator, because RF yields a better prediction model than simple linear regression using all the regressors as main effects. In fact, the GREG17 estimator introduces a small bias compared to the GREG1 estimator, and is only more efficient by a small margin. The LOO-RB-GREG1 estimator has about the same efficiency as the GREG1 estimator. Compared to the simple simulation results earlier, heteroscedastic variance does not cause loss of efficiency to the LOO-RB method here, because the sample size is large enough. The jackknife variance estimator has no statistically significant bias for the LOO-RB-RF estimator, but it has a small negative bias for the LOO-RB-GREG1 estimator.

Next, under the stratified SRS design, the LOO-RB-RF is again the most efficient estimator. Its RMSE is about the same as under SRS, which is not surprising given

proportional allocation of stratum sample sizes, because RF is able to account for the design size variable using the auxiliary information in the 17 regressors. In contrast, the GREG17 estimator actually loses efficiency and does not behave well here, which again illustrates that applying the GREG estimator without appropriate attention to model diagnostics can be counter-productive in practice. The relative performance of the simple GREG1 estimator and its unbiased counterpart LOO-RG-GREG1 is similar as under SRS, and both are slightly more efficient under the stratified design. The jackknife variance estimators perform similarly as under the SRS design.

4.2.2 Other target variables

Simulation results for the other two target variables under the stratified SRS design are given in Table 4.

Table 4: Simulation results for Total personal expenses and Total procurements under stratified SRS design. Monte Carlo errors of bias estimates in parentheses.

	Total personal expenses			Total procurements		
Estimator	Bias (MC Error)		RMSE	Bias (MC Error)		RMSE
HT	0.03	(0.04)	0.58	-0.05	(0.29)	2.88
GREG1	0.09	(0.04)	0.60	-0.07	(0.21)	2.08
LOO-RB-GREG1	0.06	(0.04)	0.60	-0.20	(0.21)	2.14
Variance LOO-RB-GREG1	*0.02*	*(0.01)*	*0.03*	*0.03*	*(0.14)*	*1.37*
GREG17	0.01	(0.02)	0.33	0.50	(0.18)	1.90
LOO-RB-RF	0.01	(0.02)	0.32	-0.07	(0.11)	1.11
Variance LOO-RB-RF	*0.03*	*(0.00)*	*0.07*	*-0.21*	*(0.03)*	*0.34*

When it comes to Total personal expenses, turnover from the tax data is not a good regressor at all, such that the simple GREG1 estimator yields no improvement over the HT estimator. Similarly with the LOO-RB-GREG1 estimator, which performs similarly as the GREG1 estimator, as can be expected. Meanwhile, both the GREG17 and LOO-RB-RF estimators are noticeably better. This suggests that the other regressors can be linearly related to this target variable, and the RF model is flexible enough to automatically capture this linear regression relationship here. The jackknife variance estimators exhibit no bias for the LOO-RB methods in this case.

Turning to the results for Total procurements of goods and services, the LOO-RB-RF estimator is again by far the most efficient of all. The GREG1 and LOO-RB-GREG1 estimators similarly improve on the HT estimator, where turnover from tax data is a reasonable regressor for this variable. The GREG17 estimator is more efficient than the simple GREG1 estimator by a small margin, albeit at the cost of introducing a small bias that is statistically significant. In contrast, the LOO-RB-RF estimator provides a much greater gain of efficiency while remaining design-unbiased. The jackknife variance estimator is essentially unbiased for the LOO-RB-GREG1 estimator, but it has a small negative bias for the LOO-RB-RF estimator.

4.2.3 Conclusions

The following conclusions seem warranted based on the simulation results above.

In situations where simple GREG estimation (with few regressors) have little bias to start with, i.e. when the simple linear regression model is a reasonable statistical model, the proposed unbiased learning approach is unlikely to offer appreciable improvement in practice. More advanced learning techniques cannot be of much help, *without* a supply of additional useful features. Nevertheless, while GREG estimation (or other model-assisted design-consistent methods) can possibly suffer from non-negligible bias in a given situation, because the assisting model is never entirely correct and the sample size may not be large enough for design consistency to fully take care of the bias, the unbiased learning approach (if applicable) can avoid the bias automatically.

More importantly, provided richer auxiliary information, the proposed unbiased SRB learning approach can yield large gains. On the one hand, it allows one to make use of modern ML techniques that can potentially lead to much more flexible and powerful prediction models, without demanding the same kind of effort that is often necessary for building complex parametric models. On the other hand, the theory for design-unbiased statistical learning developed in this paper ensures the resulting ML-assisted estimator is valid for descriptive inference, so that the ML-prediction model can help to generate *valid and efficient* estimation at the aggregated level, wihtout requiring the model to be entirely correct at the individual level, because the prediction errors in the sample are extrapolated to the population of interest based on the known pq-sampling design.

5 Summary remarks

Amalgamating classic ideas of Statistical Science and Machine Learning, we developed an ML-assisted SRB approach for pq-design-unbiased statistical learning in survey sampling. It allows one to generally achieve design-unbiased model-assisted estimation based on probability sampling from the population of interest. The freedom to adopt modern as well as emerging powerful algorithmic ML-prediction models should enable one to make more efficient use of the rich auxiliary information whenever it is available.

Some topics for future research can be noted immediately. In this paper, we have only developed a jackknife method that is applicable under SRS or stratified SRS, which does not incorporate any sophisticated adjustment for the finite-population sampling fraction. It is envisaged that more efficient and accurate variance estimation methods will be discovered by future research, possibly tailored to the specific ML models.

Next, as mentioned earlier, it is an open question at this stage how to construct the efficient subsampling scheme $q(s_1|s)$, including the choice $n_1 = |s_1|$. A related issue is the sampling design. While the inclusion probabilities in s_1 can be derived from $q(s_1|s)$ and π_i, the inclusion probabilities in s_2 may not be exactly known under a general sampling

design $p(s)$. Three directions may be investigated.

First, granted the ease of making greater use of the auxiliary information at the estimation stage, through flexible non-parametric ML models, one may adopt simple robust designs such as stratified SRS in practice, instead of complex unequal-probability designs where it can be hard to fine-tune the inclusion probabilities for multivariate surveys.

Second, in case a complex design is used, one may estimate the required inclusion probabilities by MC simulation (Fattorini, 2006). The resulting SRB method is no longer exactly unbiased as long as the number of MC samples is finite. However, this bias only depends on the computational resource, which is avoidable in principle, unlike the bias of a finite-sample non-parametric model-assisted estimator following the standard recipe, because the model is never entirely correct and the estimator may not be consistent.

Finally, in this paper, we have assumed the pq-design approach, because it fits naturally with the current practice of survey sampling, where the sampling design $p(s)$ is already implemented, so that only the subsampling scheme $q(s_1|s)$ is left to one's own device at the stage of estimation. However, by construction, the combined randomisation distribution induced by (p, q) is the same as that induced by (p_1, p_2), for any $s_1 \cup s_2 = s$ and $s_1 \cap s_2 = \emptyset$. It is worth investigating whether a direct approach to the design of (p_1, p_2) may offer certain advantages.

References

[1] Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Ann. Math. Statist.*, 18: 105-110.

[2] Bousquet, O. and Elisseeff, A. (2002). Stability and generalization, *J. Mach. Learning Res.*, 2:499-526.

[3] Breidt, F.J. and Opsomer, J.D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statist. Scien.*, 32:190-205.

[4] Brieman, L. (1996a). Heuristics of instability and stabilization in model selection. *Ann. Statist.*, 24:2350-2383.

[5] Breiman, L. (1996b). Bagging predictors. *Mach. Learn.*, 26:123-140.

[6] Cassel, C. M., Särndal, C.-E. and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63:615-620.

[7] Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, 87:376-382.

[8] Gordon, L. and Olshen, R. (1978). Asymptotically Efficient Solutions to the Classification Problem. *Ann. Statist.*, 6:515-533.

[9] Gordon, L. and Olshen, R. (1980). Consistent Nonparametric Regression From Recursive Partitioning Schemes. *J. Mult. Ana.*, 10:611-627.

[10] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, 47:663-685.

[11] McConville, K.S., Breidt, F.J., Lee, T.C.M. and Moisen, G.C. (2017). Model-assisted survey regression estimation with the LASSO. *J. Surv. Stat. Methodol.*, 5 (2): 131-158.

[12] McConville, K. S. and Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scand. J. Stat*, 46: 389-413.

[13] Mukherjee, S., Niyogi, P., Poggio, T. and Rifkin, R. (2006). Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv. Comp. Math.*, 25:161-193.

[14] Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81-91.

[15] Särndal, C.-E. (2010). The calibration approach in survey theory and practice. *Surv. Methodol.*, 33:99-119.

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

[17] Toth, D. and Eltinge, J. L. (2011). Building consistent regression trees from complex sample data. *J. Amer. Statist. Assoc.*, 106:1626-1636.

[18] Tsymbal, A. (2004). The problem of concept drift: definitions and related work. *Comp. Scien.*, 106 (2), 58.

[19] Tukey, J.W. (1958). Bias and confidence in not quite large samples (abstract). *Ann. Math. Statist.*, 29:614.

A Proof of Theorem 1

Proof: We have $V_p(\widehat{Y}_M^*) = E_1\big(V_2(\widehat{Y}_{1M}|s_1)\big) - E_p\big(V_q(\widehat{Y}_{1M}|s)\big)$ as by (3), where

$$\begin{aligned}\widehat{Y}_{1M}(s_j) &= \sum_{k\in s_j} y_k + \sum_{k\notin s_j} \mu(x_k, s_j) + (N-n+1)z_j(s_j)\\ &= \sum_{k\in s} y_k + \sum_{k\notin s} \mu(x_k, s_j) + (N-n)z_j(s_j)\end{aligned}$$

and $z_k(s_j) = y_k - \mu(x_k, s_j)$ for any k and delete-j sample s_j. Under SRS, we have

$$V_2\big(\widehat{Y}_{1M}(s_j)|s_j\big) = (N-n+1)\sum_{k\in s_j^c}\big(z_k(s_j) - \bar{Z}_{s_j^c}(s_j)\big)^2$$

where $\bar{Z}_{s_j^c}(s_j) = \sum_{k\in s_j^c} z_k(s_j)/(N-n)$ and $s_j^c = U \setminus s_j$. By (14), we have

$$z_k(s_j) - \bar{Z}_{s_j^c}(s_j) = [1+o_p(1)]\big(z_k(s_{ij}) - \bar{Z}_{s_j^c}(s_{ij})\big)$$

for any $i \in s_j$, where $\bar{Z}_{s_j^c}(s_{ij}) = \sum_{k\in s_j^c} z_k(s_{ij})/(N-n)$, and, averaged over all $i \in s_j$,

$$\begin{aligned}V_2\big(\widehat{Y}_{1M}(s_j)|s_j\big) &= [1+o_p(1)](N-n+1)(N-n)\{\sum_{i\in s_j}\frac{1}{n-1}\widehat{V}_{s_j^c}\}\\ \widehat{V}_{s_j^c} &= \frac{1}{n_{s_j^c}-1}\sum_{k\in s_j^c}\big(z_k(s_{ij}) - \bar{Z}_{s_j^c}(s_{ij})\big)^2\end{aligned}$$

where $n_{s_j^c} = |s_j^c| = N-n+1$. One can consider $\widehat{V}_{s_j^c}$ as an unbiased estimator of

$$\tau(s_{ij}) = \frac{1}{N_{s_{ij}^c}-1}\sum_{k\in s_{ij}^c}\big(z_k(s_{ij}) - \bar{Z}_{s_{ij}^c}(s_{ij})\big)^2$$

where $N_{s_{ij}^c} = |s_{ij}^c| = N-n+2$, i.e. the population variance of $z_k(s_{ij})$ in s_{ij}^c, based on SRS sample s_j^c from s_{ij}^c conditional on s_{ij}, since $s_{ij}^c = s_j^c \cup \{i\} = U \setminus s_{ij}$, such that

$$E_1\big(\sum_{i\in s_j}\frac{1}{n-1}\widehat{V}_{s_j^c}\big) = E_{s_{ij}}\big(E_i(\widehat{V}_{s_j^c}|s_{ij})\big) = E_{s_{ij}}\big(\tau(s_{ij})\big)\ .$$

Given $y_k - \mu(x_k, s) = O(1)$ for any s and $k \in U$, we obtain

$$E_1\big(V_2(\widehat{Y}_{1M}|s_1)\big) = [1+o(1)](N-n+1)(N-n)E_{s_{ij}}\big(\tau(s_{ij})\big)\ .$$

Next, for $V_q(\widehat{Y}_{1M}|s)$, we notice that, by $\mu(x_k, s_j) - \mu(x_k, s) \xrightarrow{P} 0$ in (14),

$$\widehat{Y}_{1M}(s_j) = [1+o_p(1)]\big[\sum_{k\in s} y_k + \sum_{k\notin s}\mu(x_k, s) + (N-n)z_j(s_j)\big]$$

where $V_q\big(z_j(s_j)|s\big) = \sum_{i\in s}\big(z_i(s_i) - \bar{z}(s)\big)^2/n$, for $\bar{z}(s) = \sum_{j\in s} z_j(s_j)/n$. In Sen-Yates-Grundy type expression using pairwise differences, we can write

$$\begin{aligned} V_q\big(z_j(s_j)|s\big) &= \frac{n-1}{n}\{\big(\frac{n(n-1)}{2}\big)^{-1}\sum_{i<j\in s}\frac{1}{2}\big(z_i(s_i) - z_j(s_j)\big)^2\} \\ &= \frac{n-1}{n}\{\big(\frac{n(n-1)}{2}\big)^{-1}\sum_{i<j\in s}\widehat{V}_{ij}\}[1+o_p(1)] \end{aligned}$$

where $\widehat{V}_{ij} = \frac{1}{2}\big(z_i(s_{ij}) - z_j(s_{ij})\big)^2$, by $\mu(x_k, s_{ij}) - \mu(x_k, s_j) \overset{P}{\to} 0$ in (14). One can consider $\widehat{V}_{ij}$ as an unbiased estimator of $\tau(s_{ij})$, i.e. the population variance of $z_k(s_{ij})$ in s_{ij}^c, based on SRS sample $s_2' = \{i, j\}$ from s_{ij}^c conditional on $s_1' = s_{ij}$. Moreover, one can view the expression in last brackets $\{\}$ above as $E_{q'}(\widehat{V}_{ij}|s)$ with respect to $q'(s_{ij}|s)$, such that

$$E_p\big(\{\cdot\}\big) = E_p\big(E_{q'}(\widehat{V}_{ij}|s)\big) = E_{s_1'}\big(E_{s_2'}(\widehat{V}_{ij}|s_{ij})\big) = E_{s_1'}\big(\tau(s_{ij})\big) \equiv E_{s_{ij}}\big(\tau(s_{ij})\big)\ .$$

Given $y_k - \mu(x_k, s) = O(1)$ for any s and $k \in U$, we obtain

$$E_p\big(V_q(\widehat{Y}_{1M}|s)\big) = [1+o(1)](N-n)^2\frac{n-1}{n}E_{s_{ij}}\big(\tau(s_{ij})\big)\ .$$

Finally, the result follows from $E_1\big(V_2(\widehat{Y}_{1M}|s_1)\big)$ and $E_p\big(V_q(\widehat{Y}_{1M}|s)\big)$ above, since

$$V_p(\widehat{\bar{Y}}^*_M) = (1-\frac{n}{N})\frac{1}{n}E_{s_{ij}}\big(\tau(s_{ij})\big) + o(1)\ . \quad \square$$