Biometrika (2008), **95**, 3, *pp*. 539–553 doi: 10.1093/biomet/asn028

Printed in Great Britain

A new approach to weighting and inference in sample surveys

BY JEAN-FRANÇOIS BEAUMONT

Statistics Canada, Tunney's Pasture, R.H. Coats Building, 16th floor, Ottawa, Ontario, K1A 0T6, Canada

jean-francois.beaumont@statcan.ca

SUMMARY

The validity of design-based inference is not dependent on any model assumption. However, it is well known that estimators derived through design-based theory may be inefficient for the estimation of population totals when the design weights are weakly related to the variables of interest and have widely dispersed values. We propose estimators that have the potential to improve the efficiency of any estimator derived under the design-based theory. Our main focus is limited to the improvement of the Horvitz–Thompson estimator, but we also discuss the extension to calibration estimators. The new estimators are obtained by smoothing design or calibration weights using an appropriate model. Our approach to inference requires the modelling of only one variable, the weight, and it leads to a single set of smoothed weights in multipurpose surveys. This is to be contrasted with other model-based approaches, such as the prediction approach, in which it is necessary to postulate and validate a model for each variable of interest leading potentially to variable-specific sets of weights. Our proposed approach is first justified theoretically and then evaluated through a simulation study.

Some key words: Extreme weight; Generalized design-based inference; Horvitz–Thompson estimator; Model-based inference; Multipurpose survey; Smoothed estimator; Smoothed weight.

1. INTRODUCTION

The validity of design-based inference, sometimes called randomization-based inference, is not dependent on any model assumption. However, it is well known that estimators derived through design-based theory may be inefficient for the estimation of population totals when the design weights are weakly related to the variables of interest and have widely dispersed values. For the Horvitz–Thompson estimator, this fact was already recognized by Rao (1966) and Basu (1971), if not even earlier.

In the design-based framework, the above problem can be approached by reducing the variability observed in the design weights through some function (Beaumont & Alavi, 2004). For instance, winsorizing the largest design weights is common in practice (Potter, 1990; Elliott & Little, 2000). These weight modification methods usually require making a compromise between minimizing the design bias and the design variance of estimators by a suitable choice of a tuning constant. Unfortunately, this choice is not so straightforward in multipurpose surveys, since an appropriate tuning constant for one variable of interest may be inappropriate for another.

In this paper, we propose estimators that have the potential to improve the efficiency of any estimator derived under the design-based theory without requiring the specification of a tuning constant. Gains in efficiency are achieved at the expense of introducing a model for the survey weights. The approach to inference that we consider is thus a type of model-based inference.

Nevertheless, its basic principles remain close to those of design-based inference in the sense that it is conditional on the population values of the variables of interest and not conditional on the sample inclusion indicators. Such an approach to inference is called generalized design-based, as what is known as design-based inference in the literature is just a special case. Although the proposed approach is, to our knowledge new, the idea of modelling survey weights is not. It has been employed for the more efficient estimation of model parameters in Pfeffermann & Sverchkov (1999) and for the prediction of nonsampled units in Sverchkov & Pfeffermann (2004), who considered the estimation of finite population parameters.

The approach that we consider requires the modelling of only one variable, the survey weight, and it leads to a single set of smoothed weights in multipurpose surveys. This is to be contrasted with other model-based approaches, such as that proposed by Elliott & Little (2000) or the prediction approach of Royall (1970), where it is necessary to postulate and validate a model for each variable of interest, which may then lead to variable-specific sets of weights. Although these model-based approaches may be appealing from the viewpoint of statistical efficiency, they may be practically inconvenient when there are many variables of interest.

2. PRELIMINARIES AND THE BASIC PROBLEM

We consider the problem of estimating the vector of finite population totals $T_y = \sum_{k \in U} y_k$, where y_k is the vector of variables of interest for population unit k and U is the finite population of size N. We denote by Y the N-row matrix containing y_k' in its kth row. We assume that a sample s of size n is selected, from the finite population U, according to a probability sampling design $p(I \mid Z, Y) = p(I \mid Z)$, where Z is an N-row matrix containing z_k' in its kth row, z_k is the vector of design variables for population unit k and $I' = (i_1, \ldots, i_N)$ is a vector of sample inclusion indicators; that is, $i_k = 1$ if population unit k is selected in the sample s, and $i_k = 0$ otherwise.

We define generalized design-based inference as any inference that is conditional on Y but not on I. The reason for conditioning on Y is to avoid modelling the multiple variables of interest. In § 5, we will consider inference with respect to the distribution $F_{I,Z \mid Y}$, which is a special case of generalized design-based inference. Design- or randomization-based inference is another important special case in which inference is made with respect to the conditional distribution $F_{I \mid Z,Y}$. With this type of inference, only I is viewed as being random. In the design-based theory, the natural estimator of T_y is the Horvitz–Thompson estimator $\hat{T}_y^{\mathrm{HT}} = \sum_{k \in s} w_k y_k$, where $w_k = 1/\pi_k$ is the design weight of unit k and $\pi_k = E(i_k \mid Z, Y) = E(i_k \mid Z)$ is its selection probability, which is assumed to be strictly greater than 0. The Horvitz–Thompson estimator is design-unbiased, $E(\hat{T}_y^{\mathrm{HT}} \mid Z, Y) = T_y$.

A sampling design can usually be implicitly, or even explicitly, justified by a model for the conditional distribution $F_{Y \mid Z}$. If this model has high predictive power in the sense that the design variables z, and potentially the design weight variable w, are strongly related to the variables of interest y, then the sampling design is expected to lead to an efficient Horvitz–Thompson estimator $\hat{T}_y^{\mathrm{HT}}$ of T_y. For instance, a vector z strongly related to y can be used to construct strata that are very homogeneous with respect to y, which makes the Horvitz–Thompson estimator efficient. However, such a vector of design variables may not be available. Also, design variables are usually chosen by making compromises in multipurpose surveys, since design variables that are strongly related to one variable of interest are not necessarily related to another. Moreover, practical considerations often play a major role in the choice of a sampling design, especially in household surveys. For the above reasons, the vector of design variables z and the design weight variable w are often only weakly related to the variables of interest y. This may lead to quite

an inefficient Horvitz–Thompson estimator, especially when the design weights w_k have widely dispersed values, which is not uncommon in practice.

3. The smoothed Horvitz–Thompson estimator

To deal with the inefficiency of the Horvitz–Thompson estimator, we first consider the smoothed random variable

$$\tilde{T}_y^{\mathrm{SHT}} = E(\hat{T}_y^{\mathrm{HT}} \mid I, Y) = E\left(\sum\nolimits_{k\in U} w_k i_k y_k \mid I, Y\right) = \sum\nolimits_{k\in s} \tilde{w}_k y_k,$$

where $\tilde{w}_k = E(w_k \mid I, Y)$ is a smoothed weight for unit $k \in s$. The basic idea underlying $\tilde{T}_y^{\mathrm{SHT}}$ is to reduce the variability of the design weights, or, in other words, remove their noise by taking their conditional expectation. The reason for conditioning on I, when taking the expectation of the Horvitz–Thompson estimator, is to ensure that $\tilde{T}_y^{\mathrm{SHT}}$ does not involve the unknown nonsample portion of Y. This can be seen if we note that $E(w_k i_k y_k \mid I, Y) = E(w_k \mid I, Y) i_k y_k = 0$ for nonsample population units $k \in U - s$. By removing conditioning on I, we would have had $E(w_k i_k y_k \mid Y) = E(i_k w_k \mid Y) y_k = y_k$ for both sample and nonsample units.

The smoothed random variable $\tilde{T}_y^{\mathrm{SHT}}$ is not an estimator, since it depends on the unknown smoothed weights $\tilde{w}_k$, for $k \in s$. To solve this problem, we suggest modelling the design weights w_k, which allows us to obtain an estimator $\hat{w}_k$ of $\tilde{w}_k = E(w_k \mid I, Y)$. We can then construct the smoothed Horvitz–Thompson estimator $\hat{T}_y^{\mathrm{SHT}} = \sum_{k\in s} \hat{w}_k y_k$. An obvious unbiased estimator of $\tilde{w}_k = E(w_k \mid I, Y)$ is $\hat{w}_k = w_k$, which leads to the Horvitz–Thompson estimator. The unbiasedness property of $\hat{w}_k = w_k$ does not require specification of a model for the design weights; this estimator is thus robust in that sense. However, it is also inefficient as it uses only a single observation to estimate $\tilde{w}_k$. More efficient methods of estimating $\tilde{w}_k$ are discussed in § 4, while some theoretical properties of the smoothed Horvitz–Thompson estimator are given in § 5.

4. Estimation of the smoothed weight $\tilde{w}_k$

The smoothed Horvitz–Thompson estimator requires estimation of the smoothed weight $\tilde{w}_k = E(w_k \mid I, Y)$ only for sample units $k \in s$. We assume that, for $k \in s$, $\tilde{w}_k = g_s(y_k)$, where the subscript s indicates that the function $g_s(y_k)$ could vary from one sample to another. To simplify the notation, we do not use subscripts s when describing Models 1 and 2 below.

There may be a certain number of models that can be appropriate for modelling real survey data. We propose two such models below, but do not claim that they are appropriate in every practical scenario; they are given only to illustrate the theory and to give examples of possible useful models. For instance, we could consider the following linear model, which we call Model 1, $w_k = h_k'\beta + v_k^{1/2}\varepsilon_k$, for $k \in s$, where ε_k given I and Y are random errors independently and identically distributed with $E(\varepsilon_k \mid I, Y) = 0$ and $\mathrm{var}(\varepsilon_k \mid I, Y) = \sigma^2$, β and σ^2 are unknown model parameters and the vector h_k as well as $v_k > 0$ are known functions of y_k. With Model 1, the smoothed weight is given by $\tilde{w}_k = h_k'\beta$ and is estimated by $\hat{w}_k = h_k'\hat{\beta}$, where

$$\hat{\beta} = \left(\sum\nolimits_{k\in s} \frac{h_k h_k'}{v_k}\right)^{-1} \sum\nolimits_{k\in s} \frac{h_k}{v_k} w_k \tag{1}$$

is an estimator of β obtained using the generalized least-squares method. The design weights must not be used as weights to obtain $\hat{\beta}$, for example, using the weighted generalized least-squares method as in Binder (1983), because we are interested in modelling the conditional relationship

between the design weight and the variables of interest only for sample units, so that Model 1 is only applicable to sample units $k \in s$. As a result, classical model selection and validation techniques can be used to determine an appropriate model.

Two extreme special cases of Model 1 are of interest. First, if the design weights are independent of the variables of interest, conditionally on the sample s, Model 1 can be written as $w_k = \beta + \varepsilon_k$, i.e. $h_k = 1$ and $v_k = 1$, and the estimated smoothed weight reduces to the average of the design weights: $\hat{w}_k = \hat{N}/n$, where $\hat{N} = \sum_{k\in s} w_k$. This is the ultimate smoothing in which the variability of the design weights is entirely removed. This solution is quite intuitive, since the design weights provide no information about the variables of interest when the above model is true. This model leads to the ultimate smoothed Horvitz–Thompson estimator, $\hat{T}_y^{\text{SHT-U}} = \hat{N} \sum_{k\in s} y_k/n$.

The second special case occurs when the variables of interest are perfect predictors of the design weight, if, for instance, z is included in y so that Model 1 has no random error term and can be written as $w_k = h_k'\beta$. In this case, $\hat{w}_k = \tilde{w}_k = w_k$ and the smoothed Horvitz–Thompson estimator is exactly equal to the Horvitz–Thompson estimator with no smoothing at all and thus with no efficiency gain. In practice, we may expect it to lie somewhere in between these two extreme special cases. On the one hand, gains in efficiency should usually be achieved by using a model with a small number of parameters. On the other hand, it is important not to exclude from the model y-variables that are predictive of the design weight to avoid large biases. Classical tools may thus be quite useful for determining variables that should enter into the model, so that the bias is kept small without unduly sacrificing efficiency by avoiding the inclusion in the model of too many unnecessary variables.

The estimated smoothed weights $\hat{w}_k = h_k'\hat{\beta}$ can be obtained alternatively by finding the weights ω_k, for $k \in s$, that minimize $\sum_{k\in s} (\omega_k - 1)^2/v_k$ subject to the constraints

$$\sum_{k\in s} \omega_k \frac{h_k}{v_k} = \sum_{k\in s} w_k \frac{h_k}{v_k}, \tag{2}$$

provided that there exists a vector δ such that $h_k'\delta = 1$; that is, an intercept is implicitly or explicitly included in h_k. The proof that the above minimization leads to $\omega_k = h_k'\hat{\beta}$ is straightforward and is thus omitted. This is a minimization problem very similar to that considered in a prediction-model-based framework (Chambers, 1996). The only difference is that we perform an internal calibration, i.e. calibration is on a vector of design-based estimators, the Horvitz–Thompson estimators $\sum_{k\in s} w_k h_k/v_k$, whereas calibration is on a vector of known external benchmarks in a prediction-model-based framework.

Model 1 is one of the models used by Pfeffermann & Sverchkov (1999) in their empirical study. Beaumont & Rivest (2008) considered an analysis-of-variance model, which is a special case of Model 1, to deal with the problem of stratum jumpers in a typical business survey. Their empirical study illustrates that design weights can be successfully modelled in practice, resulting in a smoothed estimator that can be significantly more efficient than its unsmoothed version. Nevertheless, Model 1 may not always hold for real data. One issue with this model is that it may lead to estimated smoothed weights $\hat{w}_k$ smaller than 1, which does not make sense, given that $w_k \geqslant 1$. A solution to this problem is to use a model which ensures that $\tilde{w}_k = g_s(y_k) \geqslant 1$. One possible model that satisfies the constraint is $w_k = 1 + \exp(h_k'\beta + v_k^{1/2}\varepsilon_k)$, for $k \in s$, where h_k, β, v_k and ε_k are defined as in Model 1. This model, which we shall call Model 2, is slightly different from the exponential model used by Pfeffermann & Sverchkov (1999), who considered a model with additive errors for which $g_s(y_k) > 0$. The smoothed weight is thus

$$\tilde{w}_k = E(w_k \mid I, Y) = 1 + \exp(h_k'\beta)\, E\{\exp(v_k^{1/2}\varepsilon_k) \mid I, Y\}, \quad k \in s.$$

Replacing the expectation in the previous equation by an average over the sample, we obtain the following approximation for the smoothed weight $\tilde{w}_k$ of a sample unit $k \in s$:

$$\tilde{w}_k^a(\beta) = 1 + \exp(h_k'\beta) \sum\nolimits_{l \in s} \frac{\exp\{v_k^{1/2}\varepsilon_l(\beta)\}}{n} \doteq \tilde{w}_k, \tag{3}$$

where the random error $\varepsilon_l(\beta) \equiv \varepsilon_l = \{\log(w_l - 1) - h_l'\beta\}/v_l^{1/2}$ is viewed as a function of β and is obtained from Model 2. Since the model errors ε_k are identically distributed, we have $E\{\tilde{w}_k^a(\beta) \mid I, Y\} = \tilde{w}_k$. To obtain an estimator $\hat{w}_k$ of the smoothed weight $\tilde{w}_k$, it simply suffices to estimate β in equation (3), which can be achieved by using the generalized least-squares method and by noting that $\log(w_k - 1) = h_k'\beta + v_k^{1/2}\varepsilon_k$. We thus have

$$\hat{\beta} = \left(\sum\nolimits_{k \in s} \frac{h_k h_k'}{v_k}\right)^{-1} \sum\nolimits_{k \in s} \frac{h_k}{v_k} \log(w_k - 1)$$

and $\hat{w}_k = \tilde{w}_k^a(\hat{\beta})$. Again, we note that the design weights must not be used as weights to obtain $\hat{\beta}$. Also, if $h_k = 1$ and $v_k = 1$, equation (3) reduces to $\tilde{w}_k^a(\beta) = \hat{N}/n$ and the smoothed Horvitz–Thompson estimator still reduces to $\hat{T}_y^{\text{SHT-U}}$.

The above models were given as examples and may not hold in some contexts. In particular, the assumption that the model errors ε_k are independent conditional on I and Y may not be valid with some sampling designs. Conditions for achieving asymptotic independence of the model errors ε_k under various sampling designs have been given by Pfeffermann et. al (1998). One assumption underlying their result in our context is that the model errors are independent conditional on Y alone. Whether or not this condition is approximately satisfied may depend on the sampling design. With multi-stage sampling designs, one may argue that this condition does not always hold. Nothing in our theory precludes modelling this dependence or enhancing the proposed models in any way, if it is believed to be important. For instance, a referee pointed out that each stage of sampling leads to its own set of weights in multi-stage sampling designs and suggested that each set of weights be modelled separately as the independence may hold within each stage. Our estimator $\hat{\beta}$ remains model-unbiased for β, i.e. $E(\hat{\beta} \mid I, Y) = \beta$, even when the independence assumption does not hold so that our smoothed estimator remains valid, although possibly less efficient, under this type of model failure. In this context, it may be useful to consider a robust variance estimator, such as our design-based estimator (11), which does not depend on the validity of the independence assumption.

5. Theoretical properties of the smoothed Horvitz–Thompson estimator

Unlike the Horvitz–Thompson estimator, the smoothed Horvitz–Thompson estimator is not necessarily design-unbiased for every Z and Y. To evaluate properties of this smoothed estimator and to make inferences, we thus remove conditioning on Z and use the distribution $F_{I,Z \mid Y}$; i.e. we consider the joint distribution induced by the sampling design and the model.

We first evaluate some properties of the Horvitz–Thompson estimator under the distribution $F_{I,Z \mid Y}$ before considering in turn those of $\tilde{T}_y^{\text{SHT}}$ and $\hat{T}_y^{\text{SHT}}$. We note that the Horvitz–Thompson estimator remains unbiased and we thus have

$$E(\lambda'\hat{T}_y^{\text{HT}} \mid Y) = \lambda' E\{E(\hat{T}_y^{\text{HT}} \mid Z, Y) \mid Y\} = \lambda' T_y, \tag{4}$$

for any vector λ of constants. Moreover, we assume an asymptotic set-up similar in spirit to that in Isaki & Fuller (1982), for instance, with a sequence of units containing values of the variables of interest and defining a sequence of nested populations of increasing size. However, in contrast

to Isaki & Fuller (1982), not only are samples randomly selected according to a sequence of sampling designs with increasing sample size, but also design variables are randomly generated. Under such an asymptotic set-up, we assume that, as both n and N increase to infinity, the following condition holds.

Condition 1. $E\{\text{var}(\lambda'\hat{T}_y^{\text{HT}} \mid Z, Y)|Y\} = O(N^2/n)$.

Using Condition 1 and result (4), we obtain the following additional result, the proof of which is straightforward:

$$\text{var}(\lambda'\hat{T}_y^{\text{HT}} \mid Y) = O(N^2/n). \tag{5}$$

The Horvitz–Thompson estimator is consistent since, from results (4) and (5), we have $\lambda'\hat{T}_y^{\text{HT}} - \lambda'T_y = O_p(N/n^{1/2})$. We now turn to the properties of $\tilde{T}_y^{\text{SHT}}$. Using result (4) and the Rao–Blackwell theorem, we have

$$E(\lambda'\tilde{T}_y^{\text{SHT}} \mid Y) = \lambda'T_y, \quad \text{var}(\lambda'\tilde{T}_y^{\text{SHT}} \mid Y) \leqslant \text{var}(\lambda'\hat{T}_y^{\text{HT}} \mid Y). \tag{6}$$

This means that $\tilde{T}_y^{\text{SHT}}$ is unbiased and is not less efficient than $\hat{T}_y^{\text{HT}}$ for estimating T_y. Using results (5) and (6), we have $\lambda'\tilde{T}_y^{\text{SHT}} - \lambda'T_y = O_p(N/n^{1/2})$, and thus $\tilde{T}_y^{\text{SHT}}$ is also consistent.

In practice, we expect that the smoothed Horvitz–Thompson estimator $\hat{T}_y^{\text{SHT}}$ inherits properties of $\tilde{T}_y^{\text{SHT}}$. For instance, if Model 1 holds then we have

$$E(\lambda'\hat{T}_y^{\text{SHT}} \mid Y) = \lambda'T_y, \quad \text{var}(\lambda'\hat{T}_y^{\text{SHT}} \mid Y) \leqslant \text{var}(\lambda'\hat{T}_y^{\text{HT}} \mid Y). \tag{7}$$

The proof of (7) is given in the Appendix. A parallel can be drawn with the result in (6). It means that the smoothed Horvitz–Thompson estimator is unbiased and not less efficient than the Horvitz–Thompson estimator under Model 1. Note that the vector of constraints (2) gives the condition under which $\lambda'\hat{T}_y^{\text{SHT}} = \lambda'\hat{T}_y^{\text{HT}}$ for any given sample. Thus, if Model 1 holds, weight smoothing leads to gains in efficiency only for variables $\lambda'y$ that are not implicitly or explicitly included in the constraints (2). Finally, using (5) and (7), we have $\lambda'\hat{T}_y^{\text{SHT}} - \lambda'T_y = O_p(N/n^{1/2})$, which shows the consistency of the smoothed Horvitz–Thompson estimator.

The result in (7) has been obtained under the assumption that Model 1 holds. Under nonlinear models, provided some additional regularity conditions are satisfied, we may still expect to obtain a similar result, although it will generally hold only asymptotically.

6. VARIANCE ESTIMATION

Under the proposed generalized design-based approach to inference, the total variance of $\lambda'\hat{T}_y^{\text{SHT}}$ can be approximated as

$$\begin{aligned}\text{var}(\lambda'\hat{T}_y^{\text{SHT}} \mid Y) \;\doteq\; & E\{\text{var}(\lambda'\hat{T}_y^{\text{HT}} \mid Z, Y)|Y\} \\ & + E[\{\text{var}(\lambda'\hat{T}_y^{\text{SHT}} \mid I, Y) - \text{var}(\lambda'\hat{T}_y^{\text{HT}} \mid I, Y)\}|Y],\end{aligned} \tag{8}$$

whose proof is given in the Appendix. It requires the validity of the following condition.

Condition 2. $E(\lambda'\hat{T}_y^{\text{SHT}} \mid I, Y) = \lambda'\tilde{T}_y^{\text{SHT}} + o_p(N/n^{1/2})$.

Under Condition 2, we write $E(\lambda'\hat{T}_y^{\text{SHT}} \mid I, Y) \doteq \lambda'\tilde{T}_y^{\text{SHT}}$ since it was shown in § 5 that $\lambda'\tilde{T}_y^{\text{SHT}} - \lambda'T_y = O_p(N/n^{1/2})$. This condition is satisfied if $\lambda'y$ is bounded and $E(\hat{w}_k \mid I, Y) = \tilde{w}_k +$

$(N/n)o_p(n^{-1/2})$. For Model 1, we can replace '$\doteq$' by '='; that is, we have $E(\hat{w}_k \mid I, Y) = \tilde{w}_k$ and $E(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid I, Y) = \lambda' \tilde{T}_y^{\mathrm{SHT}}$ for this model.

If the variances on the right-hand side of equation (8) were known, one could estimate $\mathrm{var}(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid Y)$ unbiasedly by simply omitting the expectations in (8). Since these variances are unknown, we suggest the following estimator of the total variance $\mathrm{var}(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid Y)$,

$$\hat{V}(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid Y) = \hat{V}(\lambda' \hat{T}_y^{\mathrm{HT}} \mid Z, Y) + \{\hat{V}(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid I, Y) - \hat{V}(\lambda' \hat{T}_y^{\mathrm{HT}} \mid I, Y)\}, \tag{9}$$

where $\hat{V}(\lambda' \hat{T}_y^{\mathrm{HT}} \mid Z, Y)$ is a consistent estimator of $\mathrm{var}(\lambda' \hat{T}_y^{\mathrm{HT}} \mid Z, Y)$ under the sampling design, while $\hat{V}(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid I, Y)$ and $\hat{V}(\lambda' \hat{T}_y^{\mathrm{HT}} \mid I, Y)$ are consistent estimators of $\mathrm{var}(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid I, Y)$ and $\mathrm{var}(\lambda' \hat{T}_y^{\mathrm{HT}} \mid I, Y)$, respectively, under the model for the design weights.

Standard design-based variance estimators can be used to estimate the first term on the right-hand side of (8). The estimation of the second term must take into account the model for the design weights. For instance, if Model 1 holds, it is obvious from (A5) in the Appendix that $\mathrm{var}(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid I, Y) - \mathrm{var}(\lambda' \hat{T}_y^{\mathrm{HT}} \mid I, Y)$ can be estimated by

$$\hat{V}(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid I, Y) - \hat{V}(\lambda' \hat{T}_y^{\mathrm{HT}} \mid I, Y) = -\hat{\sigma}^2 \sum\nolimits_{k \in s} v_k \left(\lambda' y_k - \frac{h_k'}{v_k} \hat{\Omega}\right)^2,$$

where $\hat{\sigma}^2$ is a consistent estimator of σ^2 under the model and

$$\hat{\Omega} = \left(\sum\nolimits_{k \in s} \frac{h_k h_k'}{v_k}\right)^{-1} \sum\nolimits_{k \in s} h_k (\lambda' y_k).$$

For nonlinear models, Taylor linearization or the bootstrap technique are natural choices for the estimation of $\mathrm{var}(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid I, Y)$ and even $\mathrm{var}(\lambda' \hat{T}_y^{\mathrm{HT}} \mid I, Y)$. Since these variances are not taken with respect to the sampling design, classical bootstrap methods can be used to generate bootstrap design weights and estimate the variance of $\lambda' \hat{T}_y^{\mathrm{SHT}}$ and $\lambda' \hat{T}_y^{\mathrm{HT}}$. For instance, if we use Model 2, bootstrap design weights could be generated as follows.

Step 1. Generate the bootstrap model error ε_k^* by randomly selecting a unit $l \in s$ and letting $\varepsilon_k^* = \varepsilon_l(\hat{\beta})$. Alternatively, ε_k^* is generated using a parametric model. For instance, ε_k^* could be generated using the normal distribution $N(0, \hat{\sigma}^2)$, where $\hat{\sigma}^2$ is a consistent estimator of σ^2 under Model 2. This process is then repeated independently for each unit $k \in s$.

Step 2. Obtain bootstrap design weights $w_k^* = 1 + \exp(h_k' \hat{\beta} + v_k^{1/2} \varepsilon_k^*)$ for each $k \in s$.

Step 3. Repeat Steps 1 and 2 a large number of times, R say, to obtain R sets of bootstrap design weights.

The Horvitz–Thompson and smoothed Horvitz–Thompson estimates can then be computed for each set of bootstrap design weights and standard bootstrap variance estimates can simply be obtained from the variability among these estimates. In the case of the smoothed Horvitz–Thompson estimator, the estimation of β must be repeated for each set of bootstrap design weights in order to obtain sets of bootstrap smoothed weights. Ideally, the procedure used to select explanatory variables in the model is also repeated.

Although variance estimator (9) was always positive in our simulation study described in § 7, it seems difficult to prove this. Its validity depends on the general validity of the model for the design weights and, in particular, on the independence assumption for the model errors. If it is desired to avoid this dependence for variance estimation, there is an alternative. It is obtained by

first writing the total mean-squared error of $\lambda' \hat{T}_y^{\mathrm{SHT}}$ as

$$E\{(\lambda' \hat{T}_y^{\mathrm{SHT}} - \lambda' T_y)^2 | Y\} = E[E\{(\lambda' \hat{T}_y^{\mathrm{SHT}} - \lambda' T_y)^2 | Z, Y\} | Y]$$
$$= E[\{\mathrm{var}(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid Z, Y) + B^2(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid Z, Y)\} | Y], \qquad (10)$$

where $B(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid Z, Y) = E(\lambda' \hat{T}_y^{\mathrm{SHT}} - \lambda' T_y \mid Z, Y)$ is the design bias of $\lambda' \hat{T}_y^{\mathrm{SHT}}$. Equation (10) could be estimated unbiasedly by omitting the expectation on the right-hand side of the second equality if the design variance and the squared design bias were known. Since they are unknown, they must be estimated. The variance $\mathrm{var}(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid Z, Y)$ can be estimated by $\hat{V}(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid Z, Y)$ using standard design-based variance estimation techniques. Since $\lambda' \hat{T}_y^{\mathrm{SHT}}$ may have a complicated form, the bootstrap technique (Rao & Wu, 1988) is a natural candidate.

The squared design bias in (10) could be estimated by $(\lambda' \hat{T}_y^{\mathrm{SHT}} - \lambda' \hat{T}_y^{\mathrm{HT}})^2$. However, this is a biased estimator of $B^2(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid Z, Y)$. We consider the following estimator, which is constructed similarly to the squared bias estimator proposed by Gwet & Rivest (1992) in the context of outlier-robust estimation

$$\hat{B}^2 = \max\{0, (\lambda' \hat{T}_y^{\mathrm{SHT}} - \lambda' \hat{T}_y^{\mathrm{HT}})^2 - \hat{V}(\lambda' \hat{T}_y^{\mathrm{SHT}} - \lambda' \hat{T}_y^{\mathrm{HT}} \mid Z, Y)\},$$

where $\hat{V}(\lambda' \hat{T}_y^{\mathrm{SHT}} - \lambda' \hat{T}_y^{\mathrm{HT}} \mid Z, Y)$ is a consistent variance estimator for the design variance $\mathrm{var}(\lambda' \hat{T}_y^{\mathrm{SHT}} - \lambda' \hat{T}_y^{\mathrm{HT}} \mid Z, Y)$. Again, the bootstrap technique is a natural candidate for estimating this variance.

To achieve more stability when estimating (10), it may be desirable to ensure that the resulting mean-squared error estimator is not greater than $\hat{V}(\lambda' \hat{T}_y^{\mathrm{HT}} \mid Z, Y)$, since we expect the smoothed Horvitz–Thompson estimator to be in general not less efficient than the Horvitz–Thompson estimator unless the postulated model for the design weights is not satisfactory. This leads to our proposed design mean-squared error estimator of (10),

$$\mathrm{MSE_D}(\lambda' \hat{T}_y^{\mathrm{SHT}}) = \min\{\hat{V}(\lambda' \hat{T}_y^{\mathrm{SHT}} \mid Z, Y) + \hat{B}^2, \hat{V}(\lambda' \hat{T}_y^{\mathrm{HT}} \mid Z, Y)\}. \qquad (11)$$

7. SIMULATION STUDY

7·1. *Description of the simulation experiment*

We conducted a simulation study to evaluate the performance of the smoothed Horvitz–Thompson estimator when the assumed model for the design weights is misspecified. First, a population U of 50 000 units was generated. One design variable z_k was drawn independently for each population unit k from an exponential distribution with mean 30, to which we added 0·5. This means that the smallest value z_k can take is 0·5. Then we generated three variables of interest according to the simple linear regression model

$$y_k^{(i)} = \beta_0 + \beta^{(i)} z_k + \varepsilon_{yk}^{(i)} \quad (i = 1, 2, 3), \qquad (12)$$

where $\beta_0 = 30$ and $\varepsilon_{yk}^{(i)}$, for $k \in U$, are independent normal random variables with mean zero and variance 2000. The correlation $\rho_{yz}^{(i)}$ between $y^{(i)}$ and z under model (12) depends on $\beta^{(i)}$. The constants $\beta^{(1)}$, $\beta^{(2)}$ and $\beta^{(3)}$ were thus chosen to yield $\rho_{yz}^{(1)} = 0$, $\rho_{yz}^{(2)} = 0{\cdot}01^{1/2}$ and $\rho_{yz}^{(3)} = 0{\cdot}8^{1/2}$, respectively. Finally, 10 000 samples of size 500 were selected from the population with selection probabilities proportional to z using Sampford's method, which is implemented in the procedure SURVEYSELECT of SAS. This is a design-based simulation experiment.

Table 1. *Relative bias and relative efficiency of the Horvitz–Thompson and smoothed Horvitz–Thompson estimators*

Estimators	Variable $y^{(1)}$		Variable $y^{(2)}$		Variable $y^{(3)}$	
	RB (%)	RE (%)	RB (%)	RE (%)	RB (%)	RE (%)
HT	0·03	100	0·15	100	−0·06	100
SHT-U	−0·82	45·19	12·05	143·02	73·33	43301·07
SHT-1	−9·07	76·54	−5·72	58·51	8·34	686·36
SHT-5	−6·09	64·65	−4·37	57·59	0·18	83·59

RB, relative bias; RE, relative efficiency; HT, Horvitz–Thompson.

Four estimators were of primary interest in the simulation study; Horvitz–Thompson, HT, SHT-U, SHT-1 and SHT-5. The last three estimators are smoothed Horvitz–Thompson estimators obtained using different versions of Model 2 with $v_k = 1$; results for Model 1 are not reported here, since Model 2 fitted the data better and gave better results. For the SHT-U estimator, we used $h_k = 1$, which leads to the ultimate smoothing discussed earlier. The SHT-1 estimator uses $h_k = y_k$ with $y_k' = (y_k^{(1)}, y_k^{(2)}, y_k^{(3)})$ and the SHT-5 estimator uses $h_k' = (y_k', (y_k^{*2})', (y_k^{*3})', (y_k^{*4})', (y_k^{*5})')$, where y_k^{*j}, for $j = 2, 3, 4, 5$, is similar to the vector y_k, except that each component of y_k is raised to the power j. This polynomial model of order 5 brings some robustness if the model underlying the SHT-1 estimator does not hold. For the SHT-1 and SHT-5 estimators, a stepwise regression was performed for each selected sample in order to choose the important variables to be included in the model. None of the above assumed models is in perfect agreement with the true model in this experiment. This allows us to evaluate the proposed approach in a realistic context.

Two measures were estimated from the 10 000 samples for each estimator; namely the relative bias as a percentage, RB, and the relative efficiency as a percentage, RE. The values of RB and RE for an estimator $\hat{T}_y^{(i)}$ of $T_y^{(i)} = \sum_{k \in U} y_k^{(i)}$ are

$$\mathrm{RB}\big(\hat{T}_y^{(i)}\big) = \frac{100 E\big\{\big(\hat{T}_y^{(i)} - T_y^{(i)}\big) \mid Z, Y\big\}}{T_y^{(i)}}, \quad \mathrm{RE}\big(\hat{T}_y^{(i)}\big) = 100 \frac{E\big\{\big(\hat{T}_y^{(i)} - T_y^{(i)}\big)^2 \mid Z, Y\big\}}{E\big\{\big(\hat{T}_y^{(i),\mathrm{HT}} - T_y^{(i)}\big)^2 \mid Z, Y\big\}},$$

where $\hat{T}_y^{(i),\mathrm{HT}}$ is the Horvitz–Thompson estimator of $T_y^{(i)}$. Both measures were approximated by replacing the expectations in the previous two equations by averages over the 10 000 selected samples.

7·2. *Simulation results*

Table 1 contains simulation results. As expected from the theory, the SHT-U estimator is essentially unbiased and most efficient for variable $y^{(1)}$. This is not surprising, since this variable does not depend on z, so that the SHT-U model holds if we only consider this variable. However, the relative bias of the SHT-U estimator is not negligible for the other variables and it is the least efficient, mainly because of its bias. This is true even for variable $y^{(2)}$, which is very weakly correlated with z.

For variable $y^{(1)}$, the estimators SHT-1 and SHT-5 are both more efficient than the HT estimator and less efficient than the SHT-U estimator. They both have a small but nonnegligible bias. It is somewhat difficult to explain this bias since, for this variable, we would expect a relative bias close to the relative bias of the SHT-U estimator. For variable $y^{(2)}$, the estimators SHT-1 and SHT-5 are more efficient than both the HT and SHT-U estimators and less biased than the SHT-U estimator. This is an indication that their underlying model better fits the data than does the SHT-U model, even if the correlation between z and $y^{(2)}$ is weak. The improvement of the SHT-5 estimator over

the SHT-1 estimator is quite small for this variable. For variable $y^{(3)}$, the SHT-1 estimator performed poorly in terms of efficiency, although better than the SHT-U estimator, even though its relative bias is not too large. The SHT-5 estimator corrects the deficiencies of the SHT-1 estimator and is the most efficient of all the estimators. In § 7·3, it will become clear by analyzing in greater depth a particular sample why the SHT-1 estimator performed so badly and why the SHT-5 estimator did not have these problems. Overall, the SHT-5 estimator performed better than the SHT-1 estimator in terms of both relative bias and relative efficiency for all three variables. It also performed better than the HT estimator in terms of relative efficiency for all variables. The SHT-5 estimator was more efficient than the Horvitz–Thompson estimator even for variable $y^{(3)}$, which is strongly correlated with the design variable.

7·3. *Analysis of a particular sample*

The bad performance of the SHT-1 estimator for variable $y^{(3)}$ may appear disappointing at first glance. However, we illustrate in this section that, by a careful analysis, one is able to detect easily the deficiencies of this estimator in a particular sample and favour the SHT-5 estimator. We analyzed a few samples but only report results for the first selected sample, since the main conclusions were the same for all the samples analyzed.

We examined graphs of residuals $\varepsilon_k(\hat{\beta})$ for the SHT-U, SHT-1 and SHT-5 models, plotted against $y^{(1)}$, $y^{(2)}$ and $y^{(3)}$. Some of these graphs are shown in Fig. 1. The solid curves have been obtained using the nonparametric regression procedure TPSPLINE of SAS, which is based on penalized least-squares estimation.

Figures 1(a) and (b) show that there is a clear association between the residuals and variable $y^{(3)}$ for the SHT-U and SHT-1 models; this explains the inefficiency of the SHT-U and SHT-1 estimators observed in Table 1 and their large bias for this variable. The situation is much better for the SHT-5 model, although Fig. 1(c) shows that there seems to be a problem of heteroscedasticity. Nevertheless, this model misspecification is not important enough to make the SHT-5 estimator inefficient for variable $y^{(3)}$. The nonnegligible bias and the inefficiency of the SHT-U estimator for variable $y^{(2)}$ is not that clear from Fig. 1(d). It seems to be more difficult to detect this type of slight model misspecification from graphs alone, so that other diagnostics and model selection tools, such as stepwise regression, are necessary. There is no anomalous pattern in the remaining plots, not shown, which correspond to cases where the smoothed Horvitz–Thompson estimators were more efficient than the Horvitz–Thompson estimator.

7·4. *Simulation results for the comparison of mean-squared error estimators*

In this section, we compare the model-dependent variance estimator (9) to the design mean-squared error estimator (11). The former is obtained using the parametric bootstrap method described in § 6. In the latter, the Rao–Wu bootstrap method (Rao & Wu, 1988; Rao et al. 1992) is used to estimate design variances with the assumption that sampling was conducted with replacement. This assumption is reasonable, given the small sampling fraction in the simulation study. The Rao–Wu bootstrap method is also used to estimate the design variance of the Horvitz–Thompson estimator. It is implemented as in Rao et al. (1992) by selecting $n - 1$ units with replacement for each bootstrap replicate. It is worth mentioning that, for both estimators (9) and (11), the stepwise selection procedure was repeated for each of the 500 bootstrap replicates.

For computer-time considerations, only 1000 samples were used for the comparison. Three different measures were estimated using these samples for each mean-squared error estimator: the relative bias, RB, as a percentage; the coverage rate, CR, as a percentage of confidence intervals with 95% nominal confidence level obtained using the normal approximation; and the average length, AL, of these confidence intervals, in thousands. The relative bias of a mean-squared error

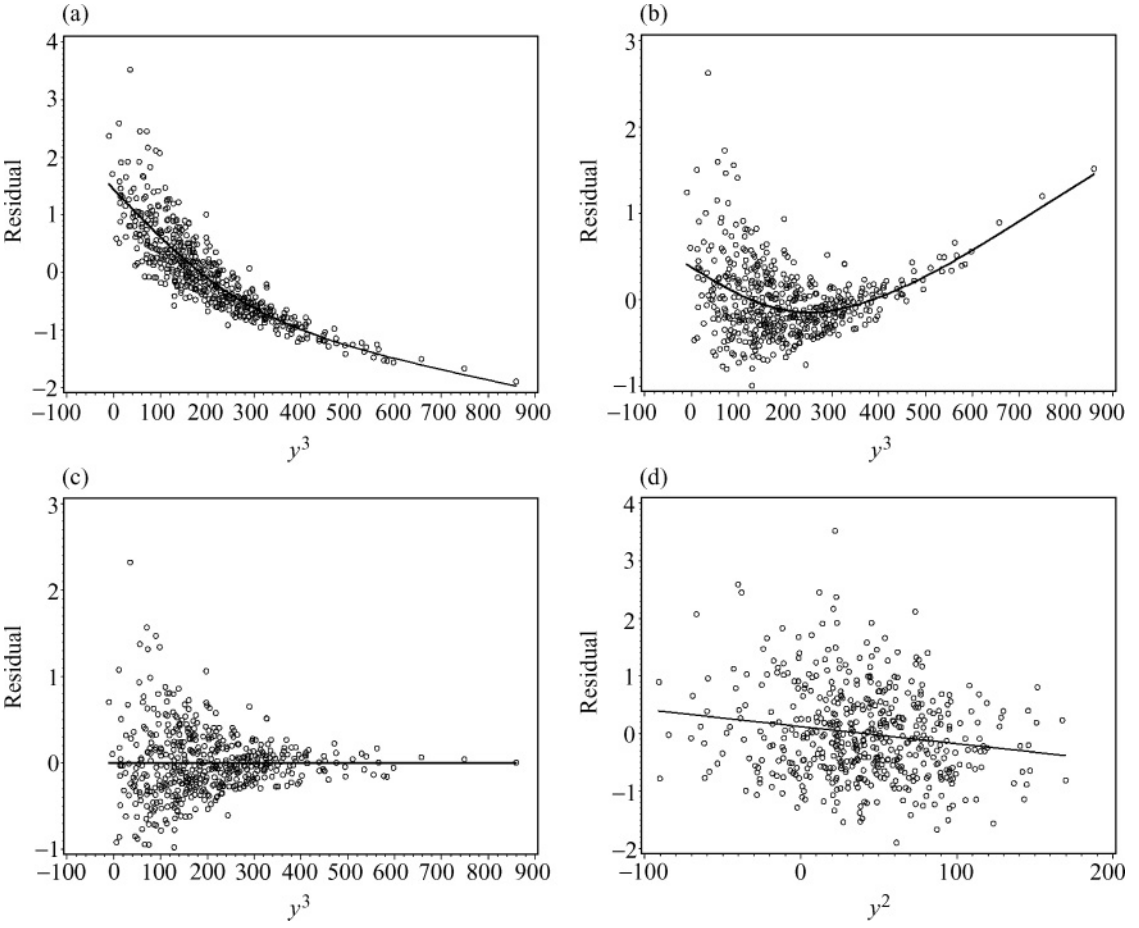

Fig. 1. Plots of residuals for (a) the SHT-U model against $y^{(3)}$, (b) the SHT-1 model against $y^{(3)}$, (c) the SHT-5 model against $y^{(3)}$ and (d) the SHT-U model against $y^{(2)}$. The solid curves are nonparametric regression fits.

Table 2. *Comparison of variance estimator* (9) *and mean-squared error estimator* (11)

Variance estimators	Variable $y^{(1)}$			Variable $y^{(2)}$			Variable $y^{(3)}$		
	RB (%)	CR (%)	AL	RB (%)	CR (%)	AL	RB (%)	CR (%)	AL
HT: Rao–Wu	−4·95	93·7	816	6·16	94·8	825	4·42	93·6	809
SHT-5: (9)	54·71	94·0	804	76·40	94·9	813	25·08	95·7	806
SHT-5: (11)	11·64	89·8	689	19·36	91·6	680	5·26	94·5	747

RB, relative bias; CR, coverage rate of confidence intervals; AL, average length of confidence intervals; HT, Horvitz–Thompson.

estimator is defined similarly to the relative bias of an estimator of a population total given in § 7·1, with the true design mean-squared error estimated using 1000 samples. Results are given in Table 2. The variance estimator (9) and mean-squared error (11) were computed only for the SHT-5 estimator, since it is the least biased of the three smoothed Horvitz–Thompson estimators evaluated. Also, the SHT-U and SHT-1 estimators would not be considered in practice if it were desired to use a single set of smoothed weights for estimating totals of all three variables. Results for the Horvitz–Thompson estimator are also presented.

From Table 2, we see that variance estimator (9) is significantly positively biased because of incorrect specification of the model. As a result, it has only a slight advantage in terms of average length of confidence intervals when compared to the design-based confidence intervals associated with the Horvitz–Thompson estimator. However, it leads to the best coverage rates, which are very close to 95%. In this simulation study, the incorrect specification of the model did not lead to underestimation of the true mean-squared error, so that the resulting confidence intervals using variance estimator (9) did not have coverage rates far below 95%. As expected, the design mean-squared error estimator (11) was significantly less biased than (9) but still positively so. This led to average lengths of confidence intervals that are much smaller than the corresponding intervals associated with the Horvitz–Thompson estimator. However, the coverage rate of these intervals is slightly below 95% for variables $y^{(1)}$ and $y^{(2)}$ because of the bias of the SHT-5 estimator for these variables, which was noted in Table 1.

7·5. *Simulation results discussion*

Our simulation results illustrate that it is possible to obtain inferences more efficient than design-based inferences by using our proposed generalized design-based approach. If the mean-squared error estimator (11) is chosen, the validity of the model for the design weights is required to obtain valid inferences but only to the extent that the bias of the smoothed Horvitz–Thompson estimator is kept small. The use of variance estimator (9) requires a more appropriate specification of the model underlying the smoothed Horvitz–Thompson estimator. Therefore, the mean-squared error estimator (11) may be preferred to the variance estimator (9) in practice, since it is more robust to model misspecifications. However, if the model is properly specified, the latter may lead to more precise variance estimates than the former so that no estimator can be claimed to be always better than the other.

One could argue that, in the context of this simulation study, a better alternative to the Horvitz–Thompson estimator would have been a model-assisted estimator such as a generalized regression estimator or, at least, the Hajek estimator $\hat{T}_y^{\mathrm{H}} = (N/\hat{N})\sum_{k\in s} w_k y_k$. Although this may be true, depending on which model-assisted estimator is used, we would like to emphasize again that the ideas developed in this paper are not restricted to the Horvitz–Thompson estimator and can be applied to any design-based estimator. Therefore, we could also smooth a generalized regression estimator to make it more efficient. This is discussed in § 8 for calibration estimators, which contain generalized regression estimators as a special case. The main point illustrated in this simulation study is that it is possible to obtain estimators that are more efficient than design-based estimators by smoothing.

8. SMOOTHED CALIBRATION ESTIMATORS

In the design-based framework, a vector x of calibration variables is often used at the estimation stage of a survey to construct a calibration estimator $\hat{T}_y^{\mathrm{CAL}} = \sum_{k\in s} w_k^c y_k$ of T_y (Deville & Särndal, 1992). The calibration weights w_k^c minimize a distance between the calibration and design weights, $D = \sum_{k\in s} d_k(w_k^c, w_k)$, for some function $d_k(\cdot,\cdot)$, subject to the calibration constraint $\sum_{k\in s} w_k^c x_k = T_x$. The quantity $T_x = \sum_{k\in U} x_k$ is the vector of known benchmarks associated with x. We denote by X the N-row matrix containing x_k' in its kth row.

In this context, we still define generalized design-based inference as any inference that is conditional on Y but not on I. Here, design-based inference is the special case where inference is made with respect to the conditional distribution $F_{I\mid Z,X,Y}$. The calibration estimator $\hat{T}_y^{\mathrm{CAL}}$ is consistent and asymptotically unbiased for T_y under this distribution (Deville & Särndal, 1992).

By analogy with § 3, we consider the smoothed random variable $\tilde{T}_y^{\mathrm{SCAL}} = E(\hat{T}_y^{\mathrm{CAL}} \mid I, Y) = \sum_{k\in s} \tilde{w}_k^c y_k$, where $\tilde{w}_k^c = E(w_k^c \mid I, Y)$ is a smoothed calibration weight for unit k. Then we obtain

a consistent estimator $\hat{w}_k^c$ of $\tilde{w}_k^c$ by modelling the calibration weights w_k^c and we construct the smoothed calibration estimator $\hat{T}_y^{\mathrm{SCAL}} = \sum_{k\in s} \hat{w}_k^c y_k$. To evaluate properties of this estimator and to make inferences, we remove conditioning on Z and X, and use the distribution $F_{I,Z,X\mid Y}$. Results similar to those in § 5 can be obtained when calibration is used. In particular, the smoothed calibration estimator $\hat{T}_y^{\mathrm{SCAL}}$ is not less efficient than the calibration estimator $\hat{T}_y^{\mathrm{CAL}}$ under a linear model for the calibration weights; that is, $\mathrm{var}(\lambda'\hat{T}_y^{\mathrm{SCAL}} \mid Y) \leqslant \mathrm{var}(\lambda'\hat{T}_y^{\mathrm{CAL}} \mid Y)$. This can be proven following a development very similar to the proof of equation (7) given in the Appendix.

An appropriate model for the calibration weights may perhaps be more difficult to find than an appropriate model for the design weights. Also, the smoothed calibration estimator $\hat{T}_y^{\mathrm{SCAL}}$ does not necessarily satisfy the calibration constraint $\sum_{k\in s} \hat{w}_k^c x_k = T_x$, which may be annoying for some users. To deal with these issues, one option is to use the smoothed calibration estimator, $\hat{T}_y^{\mathrm{SCAL}^*} = \sum_{k\in s} \hat{w}_k^{c*} y_k$. It is obtained by minimizing the distance $D^* = \sum_{k\in s} d_k(\hat{w}_k^{c*}, \hat{w}_k^*)$ subject to the calibration constraint $\sum_{k\in s} \hat{w}_k^{c*} x_k = T_x$. The smoothed weight $\hat{w}_k^*$ is an estimate of $\tilde{w}_k^* = E(w_k \mid I, X, Y)$ obtained using a model for the design weights w_k. With $\hat{T}_y^{\mathrm{SCAL}^*}$, we simply consider x as additional variables of interest and we smooth the design weights, as in § 3, before applying calibration. To evaluate properties of this estimator and to make inferences, we thus use the distribution $F_{I,Z\mid X,Y}$. The estimator $\hat{T}_y^{\mathrm{SCAL}^*}$ is expected to be less efficient than $\hat{T}_y^{\mathrm{SCAL}}$, especially when many design variables are included in x. However, it respects the calibration constraint. Further research is needed on the properties of $\hat{T}_y^{\mathrm{SCAL}^*}$.

9. Discussion

We have proposed an approach to weighting that extracts the useful portion of design-based estimation weights by removing their noise through an appropriate model. The main disadvantage of the resulting smoothed estimators in comparison with their corresponding design-based estimators is that the validity of a model is required to obtain valid inferences. Diagnostic tools for assessing the validity of the model, such as plots of residuals, may thus be quite useful.

An additional diagnostic would be a statistic for testing the null hypothesis that the conditional bias under the model is zero, i.e. a statistic for testing $H_0 : E(\lambda'\hat{T}_y^{\mathrm{SHT}} \mid I, Y) - \lambda'\tilde{T}_y^{\mathrm{SHT}} = 0$. This null hypothesis can be written equivalently as $H_0 : E\{(\lambda'\hat{T}_y^{\mathrm{SHT}} - \lambda'\hat{T}_y^{\mathrm{HT}}) \mid I, Y\} = 0$. A significant model bias is an indication that the model does not hold. It is not essential for the estimator to be exactly unbiased to achieve gains in efficiency. The goal of this diagnostic, as for other model diagnostic tools, is simply to avoid highly biased estimators that may result in inefficient smoothed estimators.

Although a suitable model can sometimes be found (Beaumont & Rivest, 2008), this may not always be possible. Nonparametric methods for the estimation of smoothed weights could thus be useful to avoid relying on a misspecified model without, one hopes, sacrificing too much efficiency. More investigation is needed about the use of such methods in this context. The good performance of the SHT-5 estimator in the simulation study is encouraging and makes nonparametric methods look promising.

Acknowledgement

I wish to thank three referees for feedback which helped in improving the quality of this paper, Ray Chambers from the University of Wollongong for useful and constructive discussions, comments and suggestions, and J.N.K. Rao from Carleton University, Claude Girard and Eric Rancourt from Statistics Canada and Danny Pfeffermann from the University of Southampton, for their constructive comments.

APPENDIX

Proofs

Proof of (7). Under Model 1, $\hat{w}_k = h_k'\hat{\beta}$ with $\hat{\beta}$ given in equation (1). It is straightforward to show that $E(\hat{w}_k \mid I, Y) = \tilde{w}_k$. It follows that $E(\lambda'\hat{T}_y^{\mathrm{SHT}} \mid I, Y) = \lambda'\tilde{T}_y^{\mathrm{SHT}}$ and that $E(\lambda'\hat{T}_y^{\mathrm{SHT}} \mid Y) = \lambda' T_y$, which shows the first part of the result. Also, we have

$$\begin{aligned}\operatorname{var}\left(\lambda'\hat{T}_y^{\mathrm{HT}} \mid Y\right) &= \operatorname{var}\left\{E\left(\lambda'\hat{T}_y^{\mathrm{HT}} \mid I, Y\right)|Y\right\} + E\left\{\operatorname{var}\left(\lambda'\hat{T}_y^{\mathrm{HT}} \mid I, Y\right)\middle|Y\right\} \\ &= \operatorname{var}\left(\lambda'\tilde{T}_y^{\mathrm{SHT}} \mid Y\right) + E\left\{\sigma^2 \sum\nolimits_{k\in s} v_k(\lambda' y_k)^2 \middle| Y\right\}, \qquad \text{(A1)}\end{aligned}$$

$$\begin{aligned}\operatorname{var}\left(\lambda'\hat{T}_y^{\mathrm{SHT}} \mid Y\right) &= \operatorname{var}\left\{E\left(\lambda'\hat{T}_y^{\mathrm{SHT}} \mid I, Y\right)\middle|Y\right\} + E\left\{\operatorname{var}\left(\lambda'\hat{T}_y^{\mathrm{SHT}} \mid I, Y\right)\middle|Y\right\} \\ &= \operatorname{var}\left(\lambda'\tilde{T}_y^{\mathrm{SHT}} \mid Y\right) + E\left\{\sigma^2 \sum\nolimits_{k\in s} h_k'(\lambda' y_k)\left(\sum\nolimits_{k\in s}\frac{h_k h_k'}{v_k}\right)^{-1}\sum\nolimits_{k\in s} h_k(\lambda' y_k) \mid Y\right\}. \qquad \text{(A2)}\end{aligned}$$

As a result, $\operatorname{var}(\lambda'\hat{T}_y^{\mathrm{SHT}} \mid Y) \leqslant \operatorname{var}(\lambda'\hat{T}_y^{\mathrm{HT}} \mid Y)$ since, for all s,

$$\sum\nolimits_{k\in s} h_k'(\lambda' y_k)\left(\sum\nolimits_{k\in s}\frac{h_k h_k'}{v_k}\right)^{-1}\sum\nolimits_{k\in s} h_k(\lambda' y_k) \leqslant \sum\nolimits_{k\in s} v_k(\lambda' y_k)^2. \qquad \text{(A3)}$$

The validity of (A3) follows directly from the application of a vector form of the Cauchy–Schwarz inequality by noting that it can be written as

$$\sum\nolimits_{k\in s} a_k b_k'\left(\sum\nolimits_{k\in s} b_k b_k'\right)^{-1}\sum\nolimits_{k\in s} a_k b_k \leqslant \sum\nolimits_{k\in s} a_k^2,$$

where $a_k = v_k^{1/2}(\lambda' y_k)$ and $b_k = h_k / v_k^{1/2}$. This shows the second part of the result.

The inequality $\operatorname{var}(\lambda'\hat{T}_y^{\mathrm{SHT}} \mid Y) \leqslant \operatorname{var}(\lambda'\hat{T}_y^{\mathrm{HT}} \mid Y)$ can also be shown to hold by noting from (A1) and (A2) that it is verified if $\operatorname{var}(\lambda'\hat{T}_y^{\mathrm{SHT}} \mid I, Y) \leqslant \operatorname{var}(\lambda'\hat{T}_y^{\mathrm{HT}} \mid I, Y)$ for all s. From (A1) and (A2) the latter can be verified since we have, after some straightforward algebra, that

$$\operatorname{var}\left(\lambda'\hat{T}_y^{\mathrm{SHT}} \mid I, Y\right) - \operatorname{var}\left(\lambda'\hat{T}_y^{\mathrm{HT}} \mid I, Y\right) = -\sigma^2 \sum\nolimits_{k\in s} v_k\left(\lambda' y_k - \frac{h_k'}{v_k}\hat{\Omega}\right)^2 \leqslant 0,$$

where

$$\hat{\Omega} = \left(\sum\nolimits_{k\in s}\frac{h_k h_k'}{v_k}\right)^{-1}\sum\nolimits_{k\in s} h_k(\lambda' y_k).$$

□

Proof of (8). Under Condition 2, we have $\operatorname{var}\{E(\lambda'\hat{T}_y^{\mathrm{SHT}} \mid I, Y) \mid Y\} \doteq \operatorname{var}(\lambda'\tilde{T}_y^{\mathrm{SHT}} \mid Y)$. As a result, the total variance of $\lambda'\hat{T}_y^{\mathrm{SHT}}$ can be approximated by

$$\operatorname{var}\left(\lambda'\hat{T}_y^{\mathrm{SHT}} \mid Y\right) \doteq \operatorname{var}\left(\lambda'\tilde{T}_y^{\mathrm{SHT}} \mid Y\right) + E\left\{\operatorname{var}\left(\lambda'\hat{T}_y^{\mathrm{SHT}} \mid I, Y\right)\middle|Y\right\}. \qquad \text{(A4)}$$

We also note that

$$\begin{aligned}\operatorname{var}\left(\lambda'\hat{T}_y^{\mathrm{HT}} \mid Y\right) &= \operatorname{var}\left(\lambda'\tilde{T}_y^{\mathrm{SHT}} \mid Y\right) + E\left\{\operatorname{var}\left(\lambda'\hat{T}_y^{\mathrm{HT}} \mid I, Y\right)\middle|Y\right\} \\ &= E\left\{\operatorname{var}\left(\lambda'\hat{T}_y^{\mathrm{HT}} \mid Z, Y\right)\middle|Y\right\}. \qquad \text{(A5)}\end{aligned}$$

The second equation of (A5) follows because $\operatorname{var}\{E(\lambda'\hat{T}_y^{\mathrm{HT}} \mid Z, Y) \mid Y\} = 0$. From (A5), we thus have

$$\operatorname{var}\left(\lambda'\tilde{T}_y^{\mathrm{SHT}} \mid Y\right) = E\left\{\operatorname{var}\left(\lambda'\hat{T}_y^{\mathrm{HT}} \mid Z, Y\right)\middle|Y\right\} - E\left\{\operatorname{var}\left(\lambda'\hat{T}_y^{\mathrm{HT}} \mid I, Y\right)\middle|Y\right\}. \qquad \text{(A6)}$$

Equation (8) is obtained by replacing $\text{var}(\lambda' \tilde{T}_y^{\text{SHT}} \mid Y)$ in (A4) by the right-hand side of equation (A6). □

References

Basu, D. (1971). An essay on the logical foundations of survey sampling, part 1. In *Foundations of Statistical Inference*, Ed. V. P. Godambe and D. A. Sprott, pp. 203–33. Toronto: Holt, Rinehart and Winston.

Beaumont, J.-F. & Alavi, A. (2004). Robust generalized regression estimation. *Survey Methodol.* **30**, 195–208.

Beaumont, J.-F. & Rivest, L.-P. (2008). Dealing with outliers in survey data. In *Handbook of Statistics, Vol.* 29*, Chapter* 11*, Sample Surveys: Theory, Methods and Inference*, Ed. D. Pfeffermann and C. R. Rao, to appear. Amsterdam: Elsevier BV.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int. Statist. Rev.* **51**, 279–92.

Chambers, R. L. (1996). Robust case-weighting for multipurpose establishment surveys. *J. Offic. Statist.* **12**, 3–32.

Deville, J.-C. & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *J. Am. Statist. Assoc.* **87**, 376–82.

Elliott, M. R. & Little, R. J. A. (2000). Model-based alternatives to trimming survey weights. *J. Offic. Statist.* **16**, 191–209.

Gwet, J.-P. & Rivest, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *J. Am. Statist. Assoc.* **87**, 1174–82.

Isaki, C. T. & Fuller, W. A. (1982). Survey design under the regression superpopulation model. *J. Am. Statist. Assoc.* **77**, 89–96.

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

Pfeffermann, D. & Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya* B **61**, 166–86.

Potter, F. (1990). A study of procedures to identify and trim extreme sampling weights. In *Proc. Survey Res. Meth. Sect.*, pp. 225–30. Alexandria, VA: American Statistical Association.

Rao, J. N. K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhya* A **28**, 47–60.

Rao, J. N. K. & Wu, C. F. J. (1988). Resampling inference with complex survey data. *J. Am. Statist. Assoc.* **83**, 231–41.

Rao, J. N. K., Wu, C. F. J. & Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodol.* **18**, 209–17.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* **57**, 377–87.

Sverchkov, M. & Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodol.* **30**, 79–92.

[*Received May* 2006. *Revised January* 2008]