International Statistical Review

International Statistical Review (2019), 87, 2, 347–367 doi:10.1111/insr.12313

Linearisation for Variance Estimation by Means of Sampling Indicators: Application to Non-response

Audrey-Anne Vallée and Yves Tillé

Institute of Statistics, University of Neuchâtel, Avenue de Bellevaux 51, Neuchâtel 2000, Switzerland
E-mail: audrey-anne.vallee@unine.ch

Summary

In the presence of non-response, it is usual to impute missing values, to reweight responding units and to adapt the estimators accordingly. The computation of the precision of the estimators becomes rapidly complex; it must take into account the sampling design, the treatment and the refinement of the estimators. In the absence of non-response, it is possible to linearise estimators with respect to the sampling indicators to compute explicit variance estimators. In this paper, we extend this linearisation method to deal with non-response. It becomes particularly straightforward to compute explicit variance estimators. Some known results are revisited in a simpler way than the usual developments, and new results for complex estimators are proposed. A simulation study evaluates the proposed methodology.

Key words: calibration; imputation; response indicator; reverse approach; reweighting.

1 Introduction

In survey sampling, variance estimation is a delicate matter. It depends on the sampling design, the complexity of the parameter, the non-response, the imputation and the reweighting processes. An important part of the literature on survey sampling is devoted to this problem. Resampling methods like jackknife (Quenouille, 1949) or bootstrap (Efron, 1979) were adapted to complex estimators (see, among others, Rao & Shao, 1992; Booth *et al.*, 1994; Shao & Sitter, 1996; Berger & Skinner, 2005; Antal & Tillé, 2011; Beaumont & Patak, 2012). These methods may be demanding computationnally, and explicit variance estimators may be preferred. Taylor linearisation have been used to approximate variance estimators explicitly. In the complete case, when all the values of a survey are observed, Woodruff (1971) and Binder (1996) proposed to linearise estimators with respect to estimated totals, while Demnati & Rao (2004) and Demnati & Rao (2010) linearised with respect to the sampling weights. Binder (1983); Binder & Patak (1994) and Deville (1999) proposed a method using estimating equations and influence functions. Graf (2011) returned to the classical definition of Taylor linearisation, presented for instance in Stuart and Ord (1994, Chapter 10) to ensure that a variance estimator can be calculated. She linearised the estimators with respect to the sampling indicators, which are dummy variables indicating if each unit is selected in the sample. All these methods are of common practice, and they usually lead to similar results. For instance, Särndal (1982) and Deville &

Särndal (1992) proposed variance estimators for calibrated total estimators. Berger (2008) and Langel & Tillé (2013) compared variance estimators for the Gini index.

In the presence of non-response, Kott (2006) linearises with respect to the vector of parameters $\boldsymbol{\lambda}$ in the calibration weights. Kim & Kim (2007) and Kim & Rao (2009) used estimating equations and linearised with respect to parameters in the non-response model and the imputation model. When the non-response treatment and the estimator are complex, for example, when the estimator is not linear in the variable of interest, it might still be challenging to find a variance estimator with those linearisation methods.

In this paper, the linearisation method of Graf (2011) is extended to deal with non-response. The estimators are linearised with respect to the response indicators and the variable of interest. It is then possible to compute variance estimators even if the parameter and the non-response treatment are complex. In Section 2, the method of Graf (2011) in the complete case is reviewed. Her results on calibrated totals are reproduced and extended to any calibrated estimator. Two frameworks used for inference with non-response are introduced in Section 3. The linearisation method is extended to deal with variance estimation in the case of inference based on a non-response model in Section 4. The method is applied to revisit the results of Kott (2006) in a simpler manner and to a general class of calibrated estimators for which there was no solution with Kott's method. In Section 5, the linearisation method dealing with inference based on an imputation model is presented and applied to any imputed estimator. Some variance estimators are evaluated through simulations in Section 6.

2 Linearisation in the Complete Case

In order to compute variance estimators, Graf (2011) proposed to linearise estimates with respect to the sampling indicators. The linearisation method is detailed in this section.

2.1 Graf's Method

Consider a finite population U of size N. We are interested in estimating a parameter $\theta = \theta(\mathbf{y})$, where $\mathbf{y} = (y_1 \cdots y_k \cdots y_N)^\top$ and y_k is the value of the variable of interest y of unit k. A sample s of size n is randomly selected in U by means of a sampling design $p(s)$ such that $p(s) \geq 0$ for all $s \subset U$ and $\sum_{s\subset U} p(s) = 1$. Define $\mathbf{a} = (a_1 \cdots a_k \cdots a_N)^\top$, the vector of sampling indicators, where a_k is 1 if unit k is selected in the sample and 0 otherwise. The first order inclusion probability of unit k is π_k, and $\mathrm{E}_p(\mathbf{a}) = \boldsymbol{\pi} = (\pi_1 \cdots \pi_k \cdots \pi_N)^\top$, where $\mathrm{E}_p(.)$ denotes the expectation with respect to the sampling design.

Consider $\widehat{\theta} = \widehat{\theta}(\mathbf{y}, \mathbf{a})$, an estimator of $\theta = \theta(\mathbf{y})$, such that $\widehat{\theta}(\mathbf{y}, \mathbf{a})$ is twice differentiable with respect to a_ℓ, $\ell = 1, \ldots, N$. Graf (2011) proposed two ways to linearise $\widehat{\theta}$ with respect to the sampling indicators: in the neighbourhood of the population parameter and in the neighbourhood of the estimator. In the first case,

$$\widehat{\theta} = \widehat{\theta}(\mathbf{y}, \boldsymbol{\pi}) + \sum_{\ell \in U} \tilde{z}_\ell (a_\ell - \pi_\ell) + R(\tau_1), \tag{1}$$

where the linearisation variable is

$$\tilde{z}_\ell = \left. \frac{\partial \widehat{\theta}}{\partial a_\ell} \right|_{\mathbf{a}=\boldsymbol{\pi}},$$

and the remainder is

$$R(\tau_1) = \frac{1}{2}\sum_{k\in U}\sum_{\ell\in U} \left.\frac{\partial^2\widehat{\theta}}{\partial a_k \partial a_\ell}\right|_{\mathbf{a}=\tau_1\mathbf{a}+(1-\tau_1)\boldsymbol{\pi}} (a_k - \pi_k)(a_\ell - \pi_\ell),$$

for some $\tau_1 \in (0, 1)$ (see, for instance, Edwards, 1994, p. 133). In the second case,

$$\widehat{\theta}(\mathbf{y}, \boldsymbol{\pi}) = \widehat{\theta} + \sum_{\ell\in U} z_\ell(\pi_\ell - a_\ell) + R(\tau_2), \tag{2}$$

where $\tau_2 \in (0, 1)$, the linearisation variable is

$$z_\ell = \left.\frac{\partial\widehat{\theta}(\mathbf{y}, \boldsymbol{\pi})}{\partial \pi_\ell}\right|_{\boldsymbol{\pi}=\mathbf{a}}.$$

The approximations are used to estimate the variance of $\widehat{\theta}$. Suppose that

$$\theta^{-1}R(\tau_1) = O_p(h_n), \tag{3}$$

$$\theta^{-1}R(\tau_2) = O_p(\eta_n), \tag{4}$$

where h_n and η_n are two sequences of real numbers such that $\lim_{n\to\infty} h_n = \lim_{n\to\infty} \eta_n = 0$, when the sample size n and the population size N are large. If (3) holds, expression (1) can be used to approximate the variance of $\widehat{\theta}$ by using $\tilde{z}_\ell$ in the Horvitz-Thompson variance,

$$\mathrm{V}(\widehat{\theta}) \approx \sum_{k\in U}\sum_{\ell\in U}(\pi_{k\ell} - \pi_k\pi_\ell)\tilde{z}_k\tilde{z}_\ell,$$

where $\pi_{k\ell}$ is the probability that units k and ℓ are both selected in the sample (Horvitz & Thompson, 1952). The value $\tilde{z}_\ell$ is not necessarily known for $\ell \in U$; it is estimated by some $\widehat{z}_\ell$ and the variance is estimated by

$$\widehat{\mathrm{V}}(\widehat{\theta}) = \sum_{k\in U}\sum_{\ell\in U} a_k a_\ell \frac{\pi_{k\ell} - \pi_k\pi_\ell}{\pi_{k\ell}}\widehat{z}_k\widehat{z}_\ell. \tag{5}$$

Considering Equations 1–4, a natural estimator is $\widehat{z}_\ell = z_\ell$ (Graf, 2011).

Remark 1. *In common cases, $h_n = \eta_n = n^{-1}$, but this is not a general rule; for instance, in quantile estimation, $h_n \geq n^{-1}$.*

Remark 2. *Define $\widehat{\theta}_{approx,1} = \widehat{\theta}(\mathbf{y}, \boldsymbol{\pi}) + \sum_{\ell\in U} \tilde{z}_\ell(a_\ell - \pi_\ell)$. If $\widehat{\theta}(\mathbf{y}, \boldsymbol{\pi}) = \theta$, then $\mathrm{E}_p(\widehat{\theta} - \widehat{\theta}_{approx,1}) = \mathrm{E}_p(\widehat{\theta} - \theta)$ is the design bias of the estimator.*

Remark 3. *Define $\widehat{\theta}_{approx,2} = \widehat{\theta}(\mathbf{y}, \boldsymbol{\pi}) + \sum_{\ell\in U} z_\ell(a_\ell - \pi_\ell)$. If Equations (3) and (4) hold, $\widehat{\theta}_{approx,1}$ and $\widehat{\theta}_{approx,2}$ are consistent estimators of $\widehat{\theta}$. Equations (3) and (4) must be evaluated on a case-by-case basis.*

Example 1. *Consider the population total of the variable of interest, $Y = \sum_{k\in U} y_k$, and its estimator $\widehat{Y} = N\widehat{Y}/\widehat{N}$, where $\widehat{Y} = \sum_{k\in U} a_k d_k y_k$, $\widehat{N} = \sum_{k\in U} a_k d_k$ and $d_k = \pi_k^{-1}$. Then $\tilde{z}_\ell = d_\ell(y_\ell - Y/N)$. A natural estimator of $\tilde{z}_\ell$ would be $\widehat{z}_\ell = d_\ell(y_\ell - \widehat{Y}/\widehat{N})$, but using (2), we find $z_\ell = d_\ell N(y_\ell - \widehat{Y}/\widehat{N})/\widehat{N}$.*

Example 2. *Consider the ratio* $R_{yx} = Y/X$, *where* $X = \sum_{k\in U} x_k$, *which is estimated by* $\widehat{R}_{yx} = \widehat{Y}/\widehat{X}$, *where* $\widehat{X} = \sum_{k\in U} a_k d_k x_k$. *Then*

$$z_\ell = d_\ell \frac{y_\ell - x_\ell \widehat{R}_{yx}}{\widehat{X}},\ \tilde{z}_\ell = d_\ell \frac{y_\ell - x_\ell R_{yx}}{X} \text{ and } \frac{\partial^2 \widehat{R}_{yx}}{\partial a_k \partial a_\ell} = d_k d_\ell \frac{2\widehat{R}_{yx} x_k x_\ell - x_\ell y_k - x_k y_\ell}{\widehat{X}^2}.$$

Suppose that $X^{-1}(\widehat{Y} - Y) = O_p(n^{-1/2})$ *and* $X^{-1}(\widehat{X} - X) = O_p(n^{-1/2})$, *then*

$$R(\tau) = \frac{1}{\widehat{X}_\tau^2}\left[\widehat{R}_{yx\tau}(\widehat{X} - X)^2 - (\widehat{X} - X)(\widehat{Y} - Y)\right] = O_p\left(\frac{1}{n}\right),$$

where $\widehat{R}_{yx\tau} = \widehat{X}_\tau^{-1}\widehat{Y}_\tau$, $\widehat{X}_\tau = \tau\widehat{X} + (1-\tau)X$ *for* $\tau \in (0,1)$. *Expressions (3) and (4) are* $O_p(n^{-1})$.

Example 3. *Let the population geometric mean be*

$$g = \left(\prod_{k\in U} y_k\right)^{1/N}, \tag{6}$$

with $y_k > 0$. *It is estimated by*

$$\widehat{g} = \prod_{k\in s} y_k^{1/(\widehat{N}\pi_k)} = \exp\left(\frac{1}{\widehat{N}}\sum_{k\in U}\frac{a_k}{\pi_k}\log y_k\right).$$

The linearisation variables are $z_\ell = \widehat{N}^{-1}\widehat{g} d_\ell \log(y_\ell/\widehat{g})$ *and* $\tilde{z}_\ell = N^{-1} g d_\ell \log(y_\ell/g)$. *Then,*

$$\frac{\partial^2 \widehat{g}}{\partial a_k a_\ell} = d_k d_\ell \frac{\widehat{g}}{\widehat{N}^2}\left[-\log\frac{y_k}{\widehat{g}} + \log\frac{y_k}{\widehat{g}}\log\frac{y_\ell}{\widehat{g}} - \log\frac{y_\ell}{\widehat{g}}\right]$$

and

$$R = \frac{\widehat{g}_\tau}{2\widehat{N}_\tau^2}\left[\left(\widehat{L}_\tau - L_\tau\right)^2 - 2\left(\widehat{L}_\tau - L_\tau\right)\left(\widehat{N} - N\right)\right],$$
$$\widehat{L}_\tau = \sum_{k\in U} a_k d_k \log\frac{y_k}{\widehat{g}_\tau}, \quad L_\tau = \sum_{k\in U}\log\frac{y_k}{\widehat{g}_\tau}$$

and $\widehat{g}_\tau$ *is* $\widehat{g}$ *where* $\boldsymbol{a}$ *is replaced by* $\tau\boldsymbol{a} + (1-\tau)\boldsymbol{\pi}$.

2.2 Calibrated Total Estimator

In calibrated estimation, the original sampling weight $d_k = \pi_k^{-1}$ is modified for $k \in s$ (Deville & Särndal, 1992; Deville *et al.*, 1993). Consider the vector $\mathbf{x}_k$ containing the values taken by p auxiliary variables on unit k. The new weights are such that

$$\sum_{k\in U} a_k w_k \mathbf{x}_k = \sum_{k\in U}\mathbf{x}_k, \tag{7}$$

where w_k is the calibrated weight of unit k. The weights are defined by

$$w_k = d_k F_k(\mathbf{x}_k^\top \boldsymbol{\lambda}). \tag{8}$$

The calibration function $F_k(.)$ is strictly increasing, $F_k(0) = 1$ and $F_k'(0) = q_k$, where $F_k'(.)$ is the derivative of $F_k(.)$ and q_k is a tuning parameter. The vector $\boldsymbol{\lambda}$ contains the Lagrange multiplier (see Deville & Särndal 1992). The calibrated total estimator of a variable y is

$$\widehat{Y}_c = \sum_{k \in U} a_k d_k F_k(\mathbf{x}_k^\top \boldsymbol{\lambda}) y_k = \sum_{k \in U} a_k w_k y_k.$$

Graf (2015) linearised the calibrated estimator to estimate its variance (see also Appendix A).

Proposition 1. *The linearisation variable of $\widehat{Y}_c$ obtained when linearising with respect to a_ℓ is $z_\ell = w_\ell e_\ell$, where $e_\ell = y_\ell - \mathbf{x}_\ell^\top \widehat{\mathbf{B}}_{y|x}$, $\widehat{\mathbf{B}}_{y|x} = \widehat{\mathbf{T}}_{xx}^{-1} \widehat{\mathbf{t}}_{xy}$,*

$$\widehat{\mathbf{T}}_{xx} = \sum_{k \in U} a_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) \mathbf{x}_k \mathbf{x}_k^\top, \quad \widehat{\mathbf{t}}_{xy} = \sum_{k \in U} a_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) \mathbf{x}_k y_k. \tag{9}$$

The linearisation variables are weighted residuals of the regression of y by the auxiliary variables, where the regression coefficients are weighted by $d_\ell F_\ell'(\mathbf{x}_\ell^\top \boldsymbol{\lambda})$. Deville & Särndal (1992) stated that the calibrated estimator is asymptotically equivalent to the generalised regression (GREG) estimator and their regression coefficients are weighted by $w_k q_k$ instead of $d_\ell F_\ell'(\mathbf{x}_\ell^\top \boldsymbol{\lambda})$. Demnati & Rao (2004) obtained the linearisation variable of Proposition 1. The difference is that the proposed methodology targets the sampling indicators, without the sampling weights. This difference is important in the proposed extension to non-response.

Example 4. *In the GREG estimator, the calibration function is $F_k(u) = 1 + q_k u$ and the regression coefficients in Proposition 1 are weighted by $d_\ell F_\ell'(\mathbf{x}_\ell^\top \boldsymbol{\lambda}) = d_\ell q_\ell$. Then $z_\ell = d_\ell g_\ell e_\ell$, where $g_\ell = w_\ell / d_\ell$ is called g-weight. Särndal (1982) advocated for the use of the g-weights when estimating the variance of the GREG estimator.*

Example 5. *Consider the linear truncated function $F_k(u) = \max(L, \min(1 + q_k u, H))$, where $L < 1 < H$ are respectively a lower and an upper bound for the g-weights* (Deville & Särndal, 1992). *In this case, the derivative of $F_k(u)$ is null when u is outside the interval $[(L-1)/q_k, (H-1)/q_k]$. This means that the regression coefficients of Proposition 1 are computed in the set of units that have weights strictly between the bounds.*

2.3 Calibration in a Complex Estimator

Consider $\mathbf{d} = (d_1 \cdots d_k \cdots d_N)^\top$, where $d_k = \pi_k^{-1}$ is the sampling weight of unit k, and suppose that an estimator of θ is $\widehat{\theta}_d = \widehat{\theta}(\mathbf{y}, \mathbf{a}, \mathbf{d})$. A vector of calibrated weights $\mathbf{w} = (w_1 \cdots w_k \cdots w_N)^\top$, where w_k is defined in (8), can be used to estimate θ in the calibrated estimator $\widehat{\theta}_w = \widehat{\theta}(\mathbf{y}, \mathbf{a}, \mathbf{w})$. The linearisation method is used to compute the variance estimator of any calibrated estimator in Proposition 2, which is proved in Appendix B.

Condition 1. *Assume that, in $\widehat{\theta}_d$, a_k is always multiplied by its associated weight d_k. Then the linearisation variable of $\widehat{\theta}_d$ is $z_\ell^d = d_\ell h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{d})$, where $h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{d}) = d_\ell \partial \widehat{\theta}_d / \partial a_\ell$. This implies that $h_{\ell 2}(\mathbf{y}, \mathbf{a}, \mathbf{d}) = \partial \widehat{\theta}_d / \partial d_\ell = a_\ell h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{d})$.*

Proposition 2. *If Condition 1 holds, the linearisation variable of* $\widehat{\theta}_w$ *obtained when linearising with respect to* a_ℓ *is* $z_\ell^w = w_\ell v_\ell^w$*, where*

$$v_\ell^w = h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w}) - \mathbf{x}_\ell^\top \widehat{\mathbf{B}}_{h1|x}, \tag{10}$$
$$\widehat{\mathbf{B}}_{h1|x} = \widehat{\mathbf{T}}_{xx}^{-1} \sum_{k \in U} a_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) \mathbf{x}_k h_{k1}(\mathbf{y}, \mathbf{a}, \mathbf{w}),$$

and $\widehat{\mathbf{T}}_{xx}$ *is defined in Equation (9).*

Condition 1 assumes that w_k and a_k appear in $\widehat{\theta}_w$ only as a product $w_k a_k$, which is common in practice. In this case, $h_{k2}(.) = a_k h_{k1}(.)$ and expression (10) is the residual of the regression of h_{k1} by $\mathbf{x}_k$. The computation of the linearisation variable can then be carried out in two steps. First, the linearisation variable $h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w})$ is computed naïvely, as if the vector $\mathbf{w}$ was not depending on $\mathbf{a}$. Next, the residual v_ℓ^w is calculated with the regression coefficients in $\widehat{\mathbf{B}}_{h1|x}$. The linearisation method is also applicable if Condition 1 does not hold, but the calculation of z_ℓ^w is more complicated, and it does not result in Proposition 2.

Example 6. *Consider the calibrated estimator of the Gini index*

$$\widehat{G} = \frac{1}{2\widehat{N}_c \widehat{Y}_c} \sum_{i \in U} \sum_{k \in U} a_i a_k w_i w_k |y_i - y_k|,$$

where $\widehat{N}_c = \sum_{k \in U} a_k w_k$*. If the weights* w_k *are considered as fixed, Langel & Tillé (2013) showed that* $z_\ell = w_\ell h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w})$*, where*

$$h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w}) = \frac{1}{\widehat{N}_c \widehat{Y}_c} \left[\widehat{Y}_c - \widehat{N}_c y_\ell + 2\widehat{N}_\ell (y_\ell - \widehat{\widehat{Y}}_\ell) - \widehat{G}(\widehat{Y}_c + \widehat{N}_c y_\ell) \right],$$
$$\widehat{N}_\ell = \sum_{i \in U} a_i w_i \mathrm{I}_{y_i \le y_\ell}, \widehat{\widehat{Y}}_\ell = \frac{1}{\widehat{N}_\ell} \sum_{i \in U} a_i w_i y_i \mathrm{I}_{\widehat{N}_i \le \widehat{N}_\ell}.$$

The linearisation variable considering that the weights are calibrated on random totals is $z_\ell^w = w_\ell \left[h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w}) - \mathbf{x}_\ell^\top \widehat{\mathbf{B}}_{h1|x} \right]$.

3 Dealing with Non-response

Two types of non-response are distinguished in survey sampling. The first one is unit non-response, in which all variables are missing for some units. Usually, it is addressed by reweighting; each sampled respondent k receives a new weight w_k. The second type is item non-response, in which some (but not all) items are missing. In this case, the missing values can be imputed, and y_k is replaced by $\tilde{y}_k = R_k y_k + (1 - R_k) y_k^*$, where y_k^* is the imputed value of non-respondent k. The response indicator R_k is 1 if unit k is a respondent and 0 otherwise.

For the purpose of inference, three sources of randomness are now considered. As in the complete case, the sampling indicator is random and generated by the sampling design. The response indicator is a random variable motivated by the non-response model. The variable of interest is modelled by a random superpopulation model, also called an imputation model. Those sources of randomness lead to two approaches to inference: the non-response model approach and the imputation model approach.

The variance can be decomposed in two ways. The first one is the two-phase approach, in which a sample is selected in the population and then respondents are chosen in the sample. Rao (1990) and Rao & Sitter (1995) discussed this approach under the imputation model context and Särndal (1992) under the non-response model context. The second decomposition is the reverse approach, in which the respondents are identified in the population and then a sample containing respondents and non-respondents is chosen in the population (Fay, 1991). The respondent sample is seen as a strongly invariant two-phase design (Beaumont & Haziza, 2016). This means that the selection of the respondents in the second phase is independent of the first phase. Shao & Steel (1999) discussed the reverse approach under the two inference contexts. This approach is often favoured when estimating variances on account of simplifications due to the non-response mechanism. The linearisation method is extended to compute variances with the reverse approach and inference based on a non-response model in Section 4 and with inference based on an imputation model in Section 5. In the interest of brevity, the two-phase approach is omitted, but it could be treated similarly.

4 Inference Based on a Non-response Model

In the non-response model approach, the vector of sampling indicators **a** and the vector of response indicators $\mathbf{R} = (R_1 \cdots R_k \cdots R_N)^\top$ are random, while the variable of interest is seen as fixed. A non-response model is assumed; components of **R** have independent bernoulli distributions with parameter p_k, for $k = 1, \ldots, N$, where p_k is the probability that unit k responds. This probability is usually unknown and estimated by $\widehat{p}_k$.

Suppose that θ has an estimator $\widehat{\theta}_{NR}$, which is treated for non-response. The reverse variance of estimator $\widehat{\theta}_{NR}$ is

$$\mathrm{V}(\widehat{\theta}_{NR}) = \mathrm{E}_q \mathrm{V}_p(\widehat{\theta}_{NR}) + \mathrm{V}_q \mathrm{E}_p(\widehat{\theta}_{NR}), \tag{11}$$

where $\mathrm{E}_q(.)$ and $\mathrm{V}_q(.)$ are respectively the expectation and the variance with respect to the non-response model.

4.1 Linearisation for Inference Based on a Non-response Model

The linearisation method in Section 2 is extended, in Methodology 1, to approximate variances in the case of inference based on a non-response model. Estimator $\widehat{\theta}_{NR}$ is written $\widehat{\theta}_{NR} = \widehat{\theta}_{NR}(\mathbf{y}, \mathbf{a}, \mathbf{R})$ and it is linearised with respect to a_ℓ and R_ℓ.

Methodology 1 Linearisation method to estimate Equation (11), the reverse variance with inference based on a non-response model

---

1: The approximation of the parameter with respect to the sampling indicators is,

$$\widehat{\theta}_{NR} \approx \widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) + \sum_{\ell \in U} \tilde{z}_\ell^a (a_\ell - \pi_\ell), \tag{12}$$

where

$$\tilde{z}_k^a = z_k^a \Big|_{\mathbf{a}=\boldsymbol{\pi}}, \quad z_k^a = \frac{\partial \widehat{\theta}_{NR}}{\partial a_\ell}.$$

2: The first component of the term on the right-hand side of (11) becomes

$$\mathrm{V}_1 = \mathrm{E}_q \mathrm{V}_p(\widehat{\theta}_{NR}) \approx \mathrm{E}_q \left[\sum_{k\in U} \sum_{\ell\in U} (\pi_{k\ell} - \pi_k \pi_\ell) \tilde{z}_k^a \tilde{z}_\ell^a \right]$$

and is estimated by

$$\widehat{\mathrm{V}}_1 = \sum_{k\in U} a_k \sum_{\ell\in U} a_\ell \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_{k\ell}} z_k^a z_\ell^a. \tag{13}$$

3: The second component of the expression on the right-hand side of (11) is approximated using (12), $\mathrm{V}_2 = \mathrm{V}_q \mathrm{E}_p(\widehat{\theta}_{NR}) \approx \mathrm{V}_q \left[\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) \right]$.

4: V_2 is approximated by linearising $\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R})$ with respect to the response indicators,

$$\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) \approx \widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{p}) + \sum_{\ell\in U} \tilde{z}_\ell^{aR} (R_\ell - p_\ell),$$

where $\mathbf{p} = (p_1 \cdots p_k \cdots p_N)^\top$ and

$$\tilde{z}_\ell^{aR} = z_\ell^{aR} \Big|_{\mathbf{R}=\mathbf{p}}, \quad z_\ell^{aR} = \frac{\partial \widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R})}{\partial R_\ell}.$$

5: Expression V_2 becomes

$$\mathrm{V}_2 \approx \mathrm{V}_q \left[\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) \right] \approx \sum_{k\in U} \sum_{\ell\in U} (p_{k\ell} - p_k p_\ell) \tilde{z}_k^{aR} \tilde{z}_\ell^{aR} = \sum_{k\in U} p_k (1 - p_k) (\tilde{z}_k^{aR})^2,$$

where $p_{k\ell}$ is the joint response probability of units k and ℓ. The respondent set is seen as a Poisson sample, $p_{k\ell} = p_k p_\ell$ if $k \neq \ell$, $p_{kk} = p_k$, and the variance reduces to a single sum. Expression V_2 is estimated by

$$\widehat{\mathrm{V}}_2 = \sum_{k\in U} a_k R_k \frac{(1 - \widehat{p}_k)}{\pi_k} (\hat{z}_k^{aR})^2, \tag{14}$$

where $\widehat{p}_k$ and $\hat{z}_k^{aR}$ are estimators of p_k and $\tilde{z}_k^{aR}$ respectively.

Remark 4. *The estimation of* V_2 *is based on the approximation of* $\mathrm{E}_p(\widehat{\theta}_{NR})$*, which may introduce a bias. As seen in Section 2, this bias depends on the remainder.*

Remark 5. *Beaumont et al. (2015) discuss situations in which it is possible to ignore some terms of the variance. The simplifications can greatly shorten the proposed methodology.*

With this methodology, variances can be estimated even in complicated cases. Linearising with respect to the sampling and response indicators guaranties the ability to solve the expectations and the variances. This extension to non-response was not necessarily possible with other linearisation methods. Kott (2006) proposed a one-step variance approximation method, which is applicable only if a_k and R_k are present in the estimated parameter as a product. The quantity $a_k R_k$ becomes a single indicator $\delta_k = a_k R_k$, which is 1 if unit k is selected in the sample and is a respondent, 0 otherwise. Similarly to Methodology 1, the estimated parameter may be linearised with respect to the variable δ_k to obtain a variance estimator in one single step. In Section 4.2, the proposed methodology is applied to revisit the results of Kott (2006). In Section 4.3, the linearisation method is used to find variance estimators of calibrated estimators that are not accessible with Kott's method.

4.2 Revisiting Kott's Method for Calibration on Totals

Consider $\mathbf{x}_k$, a vector of auxiliary variables that is completely known for unit k such that $a_k R_k = 1$, and consider $\sum_{k \in U} \mathbf{x}_k$, known population totals. The total Y is estimated in the presence of unit non-response by

$$\widehat{Y}_R = \sum_{k \in U} a_k R_k w_k y_k,$$

where the calibration weights are such that $w_k = d_k F_k(\mathbf{x}_k^\top \boldsymbol{\lambda})$ and

$$\sum_{k \in U} a_k R_k w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k. \tag{15}$$

Indicators a_k and R_k appear only as a product $\delta_k = a_k R_k$. Kott (2006) supposes that the set of respondents is selected according to a Poisson sampling design and that the estimated response probability of unit k is $\hat{p}_k = 1/F_k(\mathbf{x}_k^\top \boldsymbol{\lambda})$. Reweighting thus consists in assuming an underlying non-response model. For instance, the choice of the calibration function $F(u) = 1 + \exp u$ ensures that $0 < \hat{p}_k < 1$ and then the underlying model is logistic.

The variance estimator is obtained by considering the indicator δ_k associated to probability $\pi_k^* = \mathrm{E}(\delta_k) = \pi_k p_k$. The joint inclusion probability of unit k and ℓ in the respondent sample is $\pi_{k\ell}^* = \pi_{k\ell} p_k p_\ell$ if $k \neq \ell$ and $\pi_{kk}^* = \pi_k p_k$. Kott (2006) proposed to expand $F_k(\mathbf{x}_k^\top \boldsymbol{\lambda})$ around $\boldsymbol{\lambda}$, which is an extension of Binder (1983) and Demnati & Rao (2004). This leads to

$$\mathrm{V}(\widehat{Y}_R) = \sum_{k \in U} \left(\frac{1}{\pi_k^*} - \frac{1}{\pi_k} \right) e_k^2 + \sum_{k \in U} \sum_{\ell \in U} (\pi_{k\ell} - \pi_k \pi_\ell) \frac{e_k}{\pi_k} \frac{e_\ell}{\pi_\ell},$$

where

$$e_k = y_k - \mathbf{x}_k^\top \left[\sum_{k \in U} F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) p_k \mathbf{x}_k \mathbf{x}_k^\top \right]^{-1} \sum_{k \in U} F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) p_k \mathbf{x}_k y_k.$$

The proposed variance estimator is

$$\widehat{\mathrm{V}}(\widehat{Y}_R) = \sum_{k \in U} a_k R_k (w_k^2 \pi_k - w_k) v_k^2 + \sum_{k \in U} a_k R_k \sum_{\ell \in U} a_\ell R_\ell \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_{k\ell}} w_k v_k w_\ell v_\ell, \tag{16}$$

where $v_k = y_k - \mathbf{x}_k^\top \widehat{\mathbf{B}}_{y|x}, \widehat{\mathbf{B}}_{y|x} = \widehat{\mathbf{T}}_{xx}^{-1} \widehat{\mathbf{t}}_{xy}$,

$$\widehat{\mathbf{T}}_{xx} = \sum_{k \in U} a_k R_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) \mathbf{x}_k \mathbf{x}_k^\top, \quad \widehat{\mathbf{t}}_{xy} = \sum_{k \in U} a_k R_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) \mathbf{x}_k y_k.$$

The derivation of this result is relatively convoluted. The proposed linearisation method can be used to find kott's result in a straightforward manner. In Proposition 3, the estimator is linearised with respect to the indicator δ_k in one single step (see Appendix C).

Proposition 3. *The linearisation variable of $\widehat{Y}_R$ obtained when linearising with respect to $\delta_k = a_k R_k$ is $w_\ell e_\ell$, where $e_\ell = y_\ell - \mathbf{x}_\ell^\top \widehat{\mathbf{B}}_{y|x}$.*

4.3 Simultaneous Calibration and Reverse Approach

In practical applications, auxiliary information can be available at different levels, for instance, at the population level and at the sample level (see, among others, Dupont, 1995; Hidiroglou & Särndal, 1998; Estevao & Särndal, 2002). The set of respondents may be simultaneously calibrated on these two levels. Consider a vector of auxiliary variables, which is decomposed into two parts, that is, $\mathbf{x}_k = (\mathbf{x}_k^{\bullet\top} \;\; \mathbf{x}_k^{\circ\top})^\top$. Vector $\mathbf{x}_k$ is known for all units in the sample. The vector of totals $\mathbf{X}^\bullet = \sum_{k\in U}\mathbf{x}_k^\bullet$ is assumed to be known, while vector $\mathbf{X}^\circ = \sum_{k\in U}\mathbf{x}_k^\circ$ is unknown.

The reweighted total estimator of variable y is

$$\widehat{\theta}_{NR}(\mathbf{y},\mathbf{a},\mathbf{R}) = \sum_{k\in U} a_k R_k w_{k2} y_k,$$

where the weights are obtained by solving the following two systems of equations,

$$\sum_{k\in U} a_k R_k w_{k2}\mathbf{x}_k = \begin{pmatrix} \mathbf{X}^\bullet \\ \sum_{k\in U} a_k w_{k1}\mathbf{x}_k^\circ \end{pmatrix} \tag{17}$$

and

$$\sum_{k\in U} a_k w_{k1}\mathbf{x}_k^\bullet = \mathbf{X}^\bullet.$$

The weights are defined by $w_{k1} = d_k F_{k1}(\mathbf{x}_k^{\bullet\top}\boldsymbol{\lambda}_1)$ and $w_{k2} = d_k F_{k2}(\mathbf{x}_k^\top\boldsymbol{\lambda}_2)$, where $F_{k1}(.)$ and $F_{k2}(.)$ are two calibration functions. The response probabilities can be estimated by $\widehat{p}_k = 1/F_{k2}(\mathbf{x}_k^\top\boldsymbol{\lambda}_2)$. It is interesting to consider $F_{2k}(.) = 1 + \exp(.)$ to ensure $0 < \widehat{p}_k < 1$.

The proposed methodology leads to a variance estimator even if the calibration is on population or sample totals. That means that it is possible to find a variance estimator if the indicator a_k is not always accompanied by the indicator R_k, $k = 1,\ldots,N$, which is not the case of Kott's method. Following Methodology 1, two linearisation variables are needed: in proposition 4, the linearisation variable z_ℓ^a is computed, and in Proposition 5, the linearisation variable z_ℓ^{aR} is computed (see Appendix D).

Proposition 4. *The linearisation variable of* $\widehat{\theta}_{NR}(\mathbf{y},\mathbf{a},\mathbf{R})$ *when linearising with respect to* a_ℓ *is*

$$z_\ell^a = R_\ell w_{\ell 2} e_\ell + \begin{pmatrix} \mathbf{0} \\ w_{\ell 1}\mathbf{e}_\ell^\circ \end{pmatrix}^\top \widehat{\mathbf{B}}_{y|x},$$

where $e_\ell = y_\ell - \mathbf{x}_\ell^\top\widehat{\mathbf{B}}_{y|x}$, $\mathbf{e}_\ell^\circ = \mathbf{x}_\ell^\circ - \mathbf{x}_\ell^{\bullet\top}\widehat{\mathbf{B}}_{x^\circ|x^\bullet}$, $\mathbf{0}$ *is a null vector of the same length as* $\mathbf{x}_k^\bullet$,

$$\widehat{\mathbf{B}}_{x^\circ|x^\bullet} = \widehat{\mathbf{T}}_{x^\bullet x^\bullet}^{-1}\widehat{\mathbf{T}}_{x^\bullet x^\circ}, \qquad \widehat{\mathbf{B}}_{y|x} = \widehat{\mathbf{T}}_{xx}^{-1}\widehat{\mathbf{t}}_{xy},$$

$$\widehat{\mathbf{T}}_{x^\bullet x^\bullet} = \sum_{k\in U} a_k d_k F'_{k1}(\mathbf{x}_k^{\bullet\top}\boldsymbol{\lambda}_1)\mathbf{x}_k^\bullet\mathbf{x}_k^{\bullet\top}, \qquad \widehat{\mathbf{T}}_{x^\bullet x^\circ} = \sum_{k\in U} a_k d_k F'_{k1}(\mathbf{x}_k^{\bullet\top}\boldsymbol{\lambda}_1)\mathbf{x}_k^\bullet\mathbf{x}_k^{\circ\top},$$

$$\widehat{\mathbf{T}}_{xx} = \sum_{k\in U} a_k R_k d_k F'_{k2}(\mathbf{x}_k^\top\boldsymbol{\lambda}_2)\mathbf{x}_k\mathbf{x}_k^\top, \quad \widehat{\mathbf{t}}_{xy} = \sum_{k\in U} a_k R_k d_k F'_{k2}(\mathbf{x}_k^\top\boldsymbol{\lambda}_2)\mathbf{x}_k y_k.$$

Proposition 5. *The linearisation variable of* $\widehat{\theta}_{NR}(\boldsymbol{y}, \boldsymbol{\pi}, \mathbf{R})$ *when linearising with respect to* R_ℓ *is* $z_\ell^{aR} = F_{\ell 2}(\mathbf{x}_\ell^\top \boldsymbol{\lambda}_2^\pi) e_\ell^\pi$*, where* $\boldsymbol{\lambda}_2^\pi$ *and* e_ℓ^π *are* $\boldsymbol{\lambda}_2$ *and* e_ℓ*, respectively, when* a_k *is replaced by* π_k.

The linearisation variable z_ℓ^a in Proposition 4 can be integrated to estimator $\widehat{\mathrm{V}}_1$ in (13). The linearisation variable z_ℓ^{aR} can be estimated by $\hat{z}_\ell^{aR} = F_{\ell 2}(\mathbf{x}^\top \boldsymbol{\lambda}_2) e_\ell$, and it can be plugged in estimator $\widehat{\mathrm{V}}_2$ given in (14). When the vector of auxiliary variables of unit k is $\mathbf{x}_k = \mathbf{x}_k^\bullet$, this variance estimator reduces to the one obtained by Kott (2006).

5 Inference Based on an Imputation Model

In the imputation model approach, vectors **a** and **R** and a model for the y-variable are seen as random. The non-response mechanism is only assumed to be ignorable; it does not depend on the variable of interest when auxiliary information is taken into account. Some assumptions on the imputation model are made. For instance, assume that the model of y_k is $m : \; y_k = \mathbf{x}_k \boldsymbol{\beta} + \varepsilon_k$, where $\mathbf{x}_k$ is a vector of auxiliary variables of length p and $\boldsymbol{\beta}$ is a vector of regression coefficients. The random variable is ε_k, an error term such that $\mathrm{E}_m(\varepsilon_k) = 0$, $\mathrm{E}_m(\varepsilon_k^2) = \sigma^2$ and $\mathrm{E}_m(\varepsilon_k \varepsilon_j) = 0$ if $k \neq j$. Notations $\mathrm{E}_m(.)$ and $\mathrm{Var}_m(.)$ hold respectively for expectation and variance with respect to the imputation model. Considering $\tilde{\theta}_{NR} = \mathrm{E}_p(\widehat{\theta}_{NR})$, the reverse variance is

$$\begin{aligned}\mathrm{V}(\widehat{\theta}_{NR} - \theta) &= \mathrm{E}_q \mathrm{E}_m \mathrm{E}_p \left(\widehat{\theta}_{NR} - \tilde{\theta}_{NR} + \tilde{\theta}_{NR} - \theta\right)^2 \\ &= \mathrm{E}_q \mathrm{E}_m \mathrm{V}_p(\widehat{\theta}_{NR} - \tilde{\theta}_{NR}) + \mathrm{E}_q \mathrm{V}_m \mathrm{E}_p(\tilde{\theta}_{NR} - \theta) \\ &\quad + 2\mathrm{E}_q \mathrm{E}_m \mathrm{E}_p \left[(\widehat{\theta}_{NR} - \tilde{\theta}_{NR})(\tilde{\theta}_{NR} - \theta)\right] \\ &= \mathrm{E}_q \mathrm{E}_m \mathrm{V}_p(\widehat{\theta}_{NR}) + \mathrm{E}_q \mathrm{V}_m \mathrm{E}_p(\widehat{\theta}_{NR} - \theta). \end{aligned} \tag{18}$$

5.1 Linearisation for Inference Based on an Imputation Model

The linearisation method in Section 2 is extended, in Methodology 2, to approximate variances in the case of inference based on an imputation model. Estimator $\widehat{\theta}_{NR} = \widehat{\theta}_{NR}(\mathbf{y}, \mathbf{a}, \mathbf{R})$ is linearised with respect to a_ℓ and y_ℓ.

Methodology 2 Linearisation method to estimate Equation (18), the reverse variance with inference based on an imputation model

---

1: The approximation of the parameter with respect to the sampling indicators is calculated in (12).

2: The first component of the term on the right-hand side of (18) becomes

$$\mathrm{V}_1 = \mathrm{E}_q \mathrm{E}_m \mathrm{V}_p(\widehat{\theta}_{NR}) \approx \mathrm{E}_q \mathrm{E}_m \left[\sum_{k \in U} \sum_{\ell \in U} (\pi_{k\ell} - \pi_k \pi_\ell) \tilde{z}_k^a \tilde{z}_\ell^a\right]$$

and is estimated by Expression (13).

3: The second component of the expression on the right-hand side of (18) is approximated using (12), $\mathrm{V}_2 = \mathrm{E}_q \mathrm{V}_m \mathrm{E}_p \left[\widehat{\theta}_{NR} - \theta(\mathbf{y})\right] \approx \mathrm{E}_q \mathrm{V}_m \left[\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) - \theta(\mathbf{y})\right]$.

4: Variance V_2 is approximated by linearising $\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) - \theta(\mathbf{y})$ with respect to the variable of interest,

$$\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) - \theta(\mathbf{y}) \approx \widehat{\theta}_{NR}\left[\mathrm{E}_m(\mathbf{y}), \boldsymbol{\pi}, \mathbf{R}\right] - \theta\left[\mathrm{E}_m(\mathbf{y})\right] + \sum_{\ell \in U} \tilde{z}_\ell^{ay}\left[y_\ell - \mathrm{E}_m(y_\ell)\right],$$

where

$$\tilde{z}_\ell^{ay} = z_\ell^{ay}\Big|_{\mathbf{y}=\mathrm{E}_m(\mathbf{y})} - z_\ell^{y}\Big|_{\mathbf{y}=\mathrm{E}_m(\mathbf{y})}, \quad z_\ell^{ay} = \frac{\partial \widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R})}{\partial y_\ell}, \quad z_\ell^{y} = \frac{\partial \theta(\mathbf{y})}{\partial y_\ell}.$$

5: Expression V_2 becomes

$$\mathrm{V}_2 \approx \mathrm{E}_q \mathrm{V}_m\left[\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) - \theta(\mathbf{y})\right] \approx \sigma^2 \sum_{k \in U} (\tilde{z}_k^{ay})^2,$$

where σ^2 is the variance of y_ℓ. Expression V_2 is estimated by

$$\widehat{\mathrm{V}}_2 = \widehat{\sigma}^2 \sum_{k \in U} a_k \frac{(\hat{z}_k^{ay})^2}{\pi_k}, \tag{19}$$

where $\widehat{\sigma}^2$ and $\hat{z}_k^{ay}$ are estimators of σ^2 and $z_k^{ay} - z_\ell^{y}$, respectively.

This methodology is used to find an explicit variance estimator for imputed complex estimators in Section 5.2.

5.2 Imputation in a Complex Estimator

Item non-response can be treated by imputation. Consider a variable $\mathbf{y}$ with missing values imputed by a regression imputation method. Parameter $\theta = \theta(\mathbf{y})$ is estimated by $\widehat{\theta}_{NR} = \widehat{\theta}_{NR}(\tilde{\mathbf{y}}, \mathbf{a}, \mathbf{R})$, where $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_N)^\top$, $\tilde{y}_k = R_k y_k + (1 - R_k) y_k^*$, $y_k^* = \mathbf{x}_k^\top \widehat{\mathbf{B}}_{y|x}$, $\widehat{\mathbf{B}}_{y|x} = \widehat{\mathbf{T}}_{xx}^{-1} \widehat{\mathbf{t}}_{xy}$,

$$\widehat{\mathbf{T}}_{xx} = \sum_{k \in U} a_k R_k \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\pi_k}, \quad \widehat{\mathbf{t}}_{xy} = \sum_{k \in U} a_k R_k \frac{\mathbf{x}_k y_k}{\pi_k}$$

and $\mathbf{x}_k$ is a vector of auxiliary variables available for all of the units in the sample.

The linearisation method is used to estimate the variance (18) of this imputed estimator. Following Methodology 2, the linearisation variable z_ℓ^a is computed in Proposition 6 and the linearisation variable z_ℓ^{ay} is computed in Proposition 7 (see Appendix E for more details).

Proposition 6. *The linearisation variable of* $\widehat{\theta}_{NR} = \widehat{\theta}_{NR}(\tilde{\mathbf{y}}, \mathbf{a}, \mathbf{R})$ *when linearising with respect to* a_ℓ *is*

$$z_\ell^a = \frac{h_{\ell 1}(\tilde{\mathbf{y}}, \mathbf{a}, \mathbf{R})}{\pi_\ell} + \frac{R_\ell e_\ell}{\pi_\ell} \mathbf{x}_\ell^\top \widehat{\mathbf{T}}_{xx}^{-1} \sum_{k \in U} a_k (1 - R_k) \mathbf{x}_k h_{k2}(\tilde{\mathbf{y}}, \mathbf{a}, \mathbf{R}),$$

where $e_\ell = y_\ell - \mathbf{x}_\ell \widehat{\mathbf{B}}_{y|x}$,

$$\frac{h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{R})}{\pi_\ell} = \frac{\partial \widehat{\theta}_{NR}(\mathbf{y}, \mathbf{a}, \mathbf{R})}{\partial a_\ell}, \quad h_{\ell 2}(\mathbf{y}, \mathbf{a}, \mathbf{R}) = \frac{\partial \widehat{\theta}_{NR}(\mathbf{y}, \mathbf{a}, \mathbf{R})}{\partial y_\ell}.$$

Proposition 7. *The linearisation variable of* $\widehat{\theta}_{NR}(\tilde{\mathbf{y}}^{\pi}, \boldsymbol{\pi}, \mathbf{R})$*, where* $\tilde{\mathbf{y}}^{\pi}$ *is the vector* $\tilde{\mathbf{y}}$ *with* $\mathbf{a} = \boldsymbol{\pi}$*, when linearising with respect to* y_ℓ *is*

$$z_\ell^{ay} = R_\ell h_{\ell 2}(\tilde{\mathbf{y}}^{\pi}, \boldsymbol{\pi}, \mathbf{R}) + R_\ell \mathbf{x}_\ell^\top \left(\sum_{k \in U} R_k \mathbf{x}_k^\top \mathbf{x}_k \right)^{-1} \sum_{k \in U} (1 - R_k) \mathbf{x}_k h_{k2}(\tilde{\mathbf{y}}^{\pi}, \boldsymbol{\pi}, \mathbf{R}).$$

The linearisation variable z_ℓ^a in Proposition 6 is integrated to estimator $\widehat{\mathrm{V}}_1$ in (13). The linearisation variable $z_\ell^{ay} - z_\ell^y$ is estimated and plugged in estimator $\widehat{\mathrm{V}}_2$ given in (19).

Remark 6. *When the parameter of interest is the population total, the variance estimator in Kim & Rao (2009) coincides with the results in this section. However, as the authors point out, their linearisation method needs to be adjusted if the parameter of interest is not linear in the* y_k*.*

Example 7. *The geometric mean in (6) can be estimated with the imputed estimator*

$$\widehat{g}_I = \prod_{k \in s} \tilde{y}_k^{d_k / \widehat{N}} = \exp \frac{1}{\widehat{N}} \sum_{k \in U} a_k d_k \log \tilde{y}_k.$$

Using $d_\ell h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{R}) = \widehat{N}^{-1} \widehat{g} d_\ell \log(y_\ell / \widehat{g})$ *and* $h_{\ell 2}(\mathbf{y}, \mathbf{a}, \mathbf{R}) = \widehat{g} a_\ell d_\ell / (\widehat{N} y_\ell)$*, the linearisation variable with respect to* a_ℓ *is*

$$z_\ell^a = \frac{\widehat{g}_I d_\ell}{\widehat{N}} \left[\log\left(\frac{\tilde{y}_\ell}{\widehat{g}_I} \right) + R_\ell e_\ell \mathbf{x}_\ell^\top \widehat{\mathbf{T}}_{xx}^{-1} \sum_{k \in U} a_k d_k (1 - R_k) \frac{\mathbf{x}_k}{y_k^*} \right].$$

The estimator of $z_\ell^{ay} - z_\ell^y$*, the linearisation variable with respect to* y_ℓ *is*

$$\begin{aligned} \widehat{z}_\ell^{ay} &= \frac{\widehat{g}_I R_\ell}{\widehat{N}} \left[\frac{1}{y_\ell} + \mathbf{x}_\ell^\top \widehat{\mathbf{T}}_{xx}^{-1} \sum_{k \in U} a_k d_k (1 - R_k) \frac{\mathbf{x}_k}{y_k^*} \right] - \frac{\widehat{g}_I}{\widehat{N} \tilde{y}_\ell} \\ &= \frac{\widehat{g}_I}{\widehat{N}} \left[R_\ell \mathbf{x}_\ell^\top \widehat{\mathbf{T}}_{xx}^{-1} \sum_{k \in U} a_k d_k (1 - R_k) \frac{\mathbf{x}_k}{y_k^*} - (1 - R_\ell) \frac{1}{y_\ell^*} \right]. \end{aligned}$$

Example 8. *Consider the estimator of the Gini index with values imputed by regression imputation, such that* $\tilde{y}_k \neq \tilde{y}_i$*, for* $k \neq i$*,*

$$\widehat{G}_I = \frac{1}{2 \widehat{N} \widehat{Y}_I} \sum_{k \in U} a_k d_k \sum_{i \in U} a_i d_i |\tilde{y}_i - \tilde{y}_k|,$$

where $\widehat{Y}_I = \sum_{k \in U} a_k d_k \tilde{y}_k$*. Using*

$$\begin{aligned} h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{R}) &= \frac{1}{\widehat{N} \widehat{Y}} \left[\widehat{Y} - \widehat{N} y_\ell + 2 \widehat{N}_\ell (y_\ell - \widehat{\widehat{Y}}_\ell) - \widehat{G}(\widehat{Y} + \widehat{N} y_\ell) \right], \\ h_{\ell 2}(\mathbf{y}, \mathbf{a}, \mathbf{R}) &= \frac{a_\ell d_\ell}{\widehat{N} \widehat{Y}} \left(\sum_{k \in U} a_k d_k \frac{y_\ell - y_k}{|y_\ell - y_k|} - \widehat{G} \widehat{N} \right) \end{aligned}$$

and $\widehat{N}_\ell$ *and* $\widehat{\widehat{Y}}_\ell$ *defined in Example 6, the linearisation variable with respect to* a_ℓ *is*

$$z_\ell^a = \frac{d_\ell}{\widehat{N}\widehat{Y}_I}\left[\widehat{Y}_I - \widehat{N}\tilde{y}_\ell + 2\widehat{N}_\ell(\tilde{y}_\ell - \widehat{\widehat{Y}}_{I\ell}) - \widehat{G}_I(\widehat{Y}_I + \widehat{N}\tilde{y}_\ell)\right] + R_\ell d_\ell e_\ell \mathbf{x}_\ell^\top \widehat{\mathbf{T}}_{xx}^{-1}\sum_{k\in U} a_k(1-R_k)\mathbf{x}_k \frac{d_k}{2\widehat{N}\widehat{Y}_I}(2\widehat{N}_k - \widehat{N} - 2\widehat{N}\widehat{G}_I + d_k)$$

and the estimator of $z_\ell^{ay} - z_\ell^y$ *is*

$$\widehat{z}_\ell^{ay} = R_\ell \mathbf{x}_\ell^\top \widehat{\mathbf{T}}_{xx}^{-1}\sum_{k\in U} a_k(1-R_k)d_k\frac{\mathbf{x}_k}{\widehat{N}\widehat{Y}_I}\left(\sum_{k\in U} a_k d_k \frac{\tilde{y}_\ell - \tilde{y}_k}{|\tilde{y}_\ell - \tilde{y}_k|} - \widehat{G}_I\widehat{N}\right) - \frac{(1-R_\ell)}{\widehat{N}\widehat{Y}_I}\left(\sum_{k\in U} a_k d_k \frac{\tilde{y}_\ell - \tilde{y}_k}{|\tilde{y}_\ell - \tilde{y}_k|} - \widehat{G}_I\widehat{N}\right).$$

6 Simulation Study

Some variance estimators obtained in this paper were evaluated in two simulated data sets and in a real data set. The accuracy of the linearisation method was compared with the one of a resampling method.

6.1 Simulated Data

Two populations of $N = 500$ and $N = 1000$ units were generated. A vector of auxiliary variables was generated for each unit, $\mathbf{x}_k = (x_{1k} \quad x_{2k} \quad x_{3k})^\top$, where $x_{1k} = 1$, x_{2k} and x_{3k} have a Gamma distribution with expectation 100 and 200 and variance 2000 and 4000, respectively. The variable of interest was generated according to the model $m:\ y_k = 20 + 3x_{2k} + 2x_{3k} + \varepsilon_k$, where ε_k was normally distributed with mean 0 and variance 9. This model is adequate to estimate a geometric mean and to use regression imputation in this study. The correlation between the y-variable and x_2 was 0.73, and the one between y and x_3 was 0.68.

6.2 Real Data

The data set '`Ilocos`' is available in the `R` package '`ineq`' (Zeileis, 2014). The data originally come from the Philippines' National Statistics Office. More especially, they are from the 1997 Family and Income and Expenditure Survey and the 1998 Annual Poverty Indicators Survey. The data set contains the income and seven auxiliary variables of $N = 632$ households. In this simulation study, the natural logarithm of the total income of the household was used as the variable of interest. The vector of auxiliary variables for unit k was $\mathbf{x}_k = (x_{1k} \quad x_{2k} \quad x_{3k})^\top$, where $x_{1k} = 1$, x_{2k} was the family size and x_{3k} was a dummy variable indicating if the household is rural or urban. The correlation between the y-variable and x_2 is 0.25 and the one between y and x_3 is 0.28.

6.3 Simulation Framework

For each population, four samples were selected with a simple random sampling design without replacement so that the sampling rates were $n/N = 0.1, 0.2, 0.3, 0.4$. In each sample, a set of respondents was selected with Poisson sampling, where the unit probability of response was a logistic function of x_2 and the expected response rate was approximately 0.70. The following estimators were computed:

- in the complete sample:
  - the Gini index $\widehat{G}$, with calibrated weights and its variance, using the linearisation variable in Example 6 and the calibration vector $\mathbf{x}_k = (x_{1k} \quad x_{2k} \quad x_{3k})^\top$,
  - the geometric mean $\widehat{g}$ and its variance, using Example 3,
- in the sample with non-response:
  - the total $\widehat{Y}_I$, calibrated on population and sample totals, and its variance, using $\mathbf{x}_k^\bullet = (x_{1k} \quad x_{2k})^\top$, $\mathbf{x}_k^\circ = x_{3k}$ and results of Section 4.3,
  - the Gini index $\widehat{G}_I$, with regression imputation and its variance, as in Example 8,
  - the geometric mean $\widehat{g}_I$, with regression imputation and its variance, as in Example 7.

Each variance estimator was compared with a bootstrap variance estimator (Efron, 1979; Shao & Sitter, 1996), with the following procedure to estimate the variance of an estimator $\widehat{\theta}$ (Haziza, 2009):

1. A bootstrap sample s^* of size $n^* = n - 1$ of the original sample s is selected with simple random sampling with replacement.
2. Consider R_k^*, the response indicator of unit k in s^*. The missing values are imputed with regression imputation, where the regression coefficients are computed using s^*.
3. The estimator $\widehat{\theta}^*$ is computed using s^* and the new imputed data.
4. Steps 1–3 are repeated $B = 1000$ times.
5. The bootstrap variance estimator of $\widehat{\theta}$ is

$$\widehat{\mathrm{V}}_B(\widehat{\theta}) = \frac{N-n}{N}\frac{1}{1-B}\sum_{b=1}^{B}\left(\widehat{\theta}^*_{(b)} - \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^*_{(b)}\right)^2,$$

where $\widehat{\theta}^*_{(b)}$ is the value of $\widehat{\theta}^*$ at the replication b, $b = 1, \ldots, B$.

The sample selection was repeated $R = 10\,000$ times. For each population and sampling rate, the Monte Carlo relative bias and relative root mean squared error of each variance estimator were calculated. Consider $\widehat{\theta}$, an estimator of the parameter θ, and $\widehat{\mathrm{V}}$, an estimator of $\mathrm{V}(\widehat{\theta})$. The Monte Carlo relative bias of $\widehat{V}$ is

$$\mathrm{RB}_{MC}(\widehat{\mathrm{V}}) = \frac{\mathrm{E}_{MC}(\widehat{\mathrm{V}}) - \mathrm{V}_{MC}(\widehat{\theta})}{\mathrm{V}_{MC}(\widehat{\theta})} \times 100,$$

where

$$\mathrm{E}_{MC}(\widehat{\mathrm{V}}) = \frac{1}{R}\sum_{r=1}^{R}\widehat{\mathrm{V}}^{(r)}, \quad \mathrm{V}_{MC}(\widehat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}\left[\widehat{\theta}^{(r)} - \mathrm{E}_{MC}(\widehat{\theta})\right]^2$$

and $\widehat{\mathrm{V}}^{(r)}$ and $\widehat{\theta}^{(r)}$ are the values of estimators $\widehat{\mathrm{V}}$ and $\widehat{\theta}$ at simulation r, $r = 1, \ldots, R$. The Monte Carlo relative root mean squared error of $\widehat{V}$ is,

$$\mathrm{RRMSE}_{MC}(\widehat{\mathrm{V}}) = \frac{\sqrt{\mathrm{E}_{MC}\left\{\left[\widehat{\mathrm{V}} - \mathrm{V}_{MC}(\widehat{\theta})\right]^2\right\}}}{\mathrm{V}_{MC}(\widehat{\theta})} \times 100.$$

6.4 Simulation Results

Tables 1–3 present the results of the simulated data with $N = 500$, with $N = 1000$ and of the data set ‘Ilocos’, respectively. The first five rows of the tables correspond to the variance estimators obtained with the linearisation method (Lin), and the last five rows correspond to the bootstrap variance estimators (Boot). The relative biases and relative root mean squared errors (MSE), between brackets, are presented for each sampling rate in columns.

Consider the results of the linearisation method (Lin) in Tables 1–3. Except for the Gini index variance estimator, all the relative biases are smaller than 5.00% in absolute terms. Globally, the sample size and the population size do not seem to have a large impact on the relative biases, but as the sample size and the population size increase, the relative root MSE seems to decrease. The variance of the Gini index have relative biases smaller than 12.00% in absolute terms, and it tends to decrease as n and N increase. This difference in terms of relative biases could be explained by the instability of the variance estimators of dispersion parameters. As a matter of fact, the relative root mean squared errors of the Gini variance estimators are greater than the others.

In the simulated data sets, Tables 1 and 2, the relative biases of the bootstrap variance estimators (Boot) are smaller than 5.00% in absolute terms. Except for the Gini index with smaller population and sample sizes, the bootstrap method is comparable with the linearisation one, in terms of relative biases and relative root MSE. In the real data set, Table 3, the relative biases of the bootstrap estimators in the presence of non-response tend to increase, up to 21.19% in absolute terms, when the sample size increases. Indeed, in the presence of imputed data, this bootstrap variance estimator is known to be reliable when the sampling rate is small (Shao & Sitter, 1996).

Table 1. *Relative biases (RB) and relative root mean squared errors (RRMSE) of the variance estimators in the simulated population of $N = 500$ units.*

		$n = 50$		$n = 100$		$n = 150$		$n = 200$	
$\widehat{\theta}$	Method	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE
$\widehat{G}$	Lin	−11.35	(34.31)	−6.10	(23.79)	−3.49	(17.66)	−5.13	(14.35)
$\widehat{g}$	Lin	0.76	(20.97)	0.71	(14.16)	−1.03	(10.63)	−2.18	(8.69)
$\widehat{Y}_I$	Lin	−3.77	(21.55)	−1.42	(14.33)	−1.14	(10.82)	−0.27	(8.67)
$\widehat{G}_I$	Lin	1.23	(32.19)	−0.62	(21.19)	0.63	(16.39)	−1.50	(13.00)
$\widehat{g}_I$	Lin	0.73	(20.94)	0.72	(14.15)	−0.99	(10.62)	−2.17	(8.68)
$\widehat{G}$	Boot	2.85	(39.89)	−0.03	(25.52)	0.36	(18.79)	−2.35	(14.84)
$\widehat{g}$	Boot	0.77	(21.34)	0.69	(14.88)	−1.08	(11.59)	−2.21	(9.76)
$\widehat{Y}_I$	Boot	−1.26	(22.24)	−0.28	(15.18)	−0.48	(11.74)	0.31	(9.79)
$\widehat{G}_I$	Boot	−0.07	(31.41)	−1.39	(21.31)	0.08	(16.82)	−1.88	(13.67)
$\widehat{g}_I$	Boot	0.74	(21.31)	0.70	(14.87)	−1.05	(11.58)	−2.21	(9.75)

Table 2. *Relative biases (RB) and relative root mean squared errors (RRMSE) of the variance estimators in the simulated population of $N = 1000$ units.*

		$n = 100$		$n = 200$		$n = 300$		$n = 400$	
$\widehat{\theta}$	Method	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE
$\widehat{G}$	Lin	−6.11	(24.13)	−2.49	(15.48)	−2.49	(11.75)	−1.15	(9.15)
$\widehat{g}$	Lin	0.55	(13.70)	3.67	(10.00)	−0.62	(6.87)	1.27	(5.72)
$\widehat{Y}_I$	Lin	−2.42	(15.45)	0.46	(10.21)	−2.74	(8.01)	0.10	(6.19)
$\widehat{G}_I$	Lin	−1.48	(21.67)	0.20	(14.66)	−0.16	(11.25)	0.50	(9.02)
$\widehat{g}_I$	Lin	0.57	(13.70)	3.69	(10.01)	−0.61	(6.87)	1.27	(5.72)
$\widehat{G}$	Boot	0.50	(26.33)	0.68	(16.79)	−0.53	(12.6)	0.41	(10.37)
$\widehat{g}$	Boot	0.67	(14.46)	3.70	(10.99)	−0.53	(8.27)	1.25	(7.31)
$\widehat{Y}_I$	Boot	−1.32	(16.1)	1.00	(11.28)	−2.35	(9.10)	0.38	(7.69)
$\widehat{G}_I$	Boot	−2.37	(21.63)	−0.21	(15.16)	−0.54	(11.94)	0.30	(10.01)
$\widehat{g}_I$	Boot	0.69	(14.46)	3.71	(10.99)	−0.53	(8.27)	1.25	(7.30)

Table 3. *Relative biases (RB) and relative root mean squared errors (RRMSE) of the variance estimators in the dataset Ilocos of $N = 632$ units.*

		$n = 63$		$n = 126$		$n = 190$		$n = 253$	
$\widehat{\theta}$	Method	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE
$\widehat{G}$	Lin	−5.31	(30.92)	−2.32	(20.46)	−1.81	(15.54)	−2.43	(12.37)
$\widehat{g}$	Lin	0.80	(17.24)	−0.50	(11.29)	1.30	(8.89)	1.12	(7.11)
$\widehat{Y}_I$	Lin	−4.98	(23.90)	−3.28	(16.02)	−1.21	(12.65)	−1.27	(10.83)
$\widehat{G}_I$	Lin	1.07	(37.53)	−1.78	(24.54)	−6.34	(19.21)	−9.25	(17.54)
$\widehat{g}_I$	Lin	−3.43	(22.48)	−2.71	15.43)	0.35	(12.45)	−0.22	(10.51)
$\widehat{G}$	Boot	−2.10	(31.10)	−1.09	(20.79)	−1.03	(16.07)	−1.82	(13.02)
$\widehat{g}$	Boot	0.76	(17.78)	−0.51	(12.15)	1.23	(9.94)	1.11	(8.45)
$\widehat{Y}_I$	Boot	−3.66	(24.39)	−8.21	(17.16)	−11.32	(16.13)	−16.88	(19.21)
$\widehat{G}_I$	Boot	−2.03	(36.15)	−7.42	(24.51)	−14.79	(22.29)	−21.19	(24.89)
$\widehat{g}_I$	Boot	−3.98	(23.00)	−8.16	(16.81)	−10.00	(15.24)	−15.75	(18.20)

More generally, the objectives of the linearisation method are to obtain explicit variance estimators and to have a small computation cost. In other words, the variance estimators should have a comparable precision but a lower computational cost than a resampling method. In this simulation study, the precision of both methods is broadly comparable. The main exceptions concern the Gini index in function of the sampling rate and the non-response.

7 Discussion

An important aspect of the proposed linearisation method is that it is usable in any circumstance, if the derivatives of the parameter exist and the remainder is small. This is not necessarily the case of other linearisation methods. As seen in Sections 4 and 5, our linearisation method gives similar results as in Kott (2006) or in Kim & Rao (2009). Nevertheless, the derivation of the linearised variable is simpler and faster.

The interest of our method is that it enables to linearise the variables in such a way that all the sources of variability and the statistical treatments are taken into account: the random sample, the non-response, the calibration, the reweighting for non-response and the imputation.

We considered the case of deterministic imputation in Section 5. The extension of the linearisation method to deal with random imputation should be relatively straightforward. However, the application to the case of nearest neighbour imputation does not seem direct.

Acknowledgements

We would like to thank the associate editor and the referees for their scrupulous reading. Their reports led to major improvements of the manuscript.

References

Antal, E. & Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *J. Amer. Statist. Assoc.*, **106**, 534–543.

Beaumont, J. -F., Béliveau, A. & Haziza, D. (2015). Clarifying some aspects of variance estimation in two-phase sampling. *J. Survey Stat. Methodology*, 3524–542.

Beaumont, J. -F. & Haziza, D. (2016). A note on the concept of invariance in two-phase sampling designs. *Survey Methodology*, **42**, 319–323.

Beaumont, J. -F. & Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *Int. Stat. Rev.*, **80**, 127–148.

Berger, Y. G. (2008). A note on asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient. *J. Off. Stat.*, **24**, 541–555.

Berger, Y. G. & Skinner, C. J. (2005). A jackknife variance estimator for unequal probability sampling. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **B67**, 79–89.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex survey. *Int. Stat. Rev.*, **51**, 279–292.

Binder, D. A. (1996). Linearization methods for single phase and two-phase samples: a cookbook approach. *Survey Methodology*, **22**, 17–22.

Binder, D. A. & Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *J. Amer. Statist. Assoc.*, **89**, 1035–1043.

Booth, J. G., Butler, R. W. & Hall, P. (1994). Bootstrap methods for finite populations. *J. Amer. Statist. Assoc.*, **89**, 1282–1289.

Demnati, A. & Rao, J. N. K. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology*, **30**, 17–34.

Demnati, A. & Rao, J. N. K. (2010). Linearization variance estimators for model parameters from complex survey data. *Survey Methodology*, **36**, 193–201.

Deville, J. -C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, **25**, 193–204.

Deville, J. -C. & Särndal, C. -E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, **87**, 376–382.

Deville, J. -C., Särndal, C. -E. & Sautory, O. (1993). Generalized raking procedure in survey sampling. *J. Amer. Statist. Assoc.*, **88**, 1013–1020.

Dupont, F. (1995). Alternative adjustments where there are several levels of auxiliary information. *Survey Methodology*, **21**, 125–135.

Edwards, C. H. (1994). *Advanced Calculus of Several Variables.* Massachusetts, United States: Courier Corporation.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**, 1–26.

Estevao, V. M. & Särndal, C. -E. (2002). The ten cases of auxiliary information for calibration in twophase sampling. *Journal of Official Statistics*, **18**, 233–255.

Fay, R. E. (1991). A design-based perspective on missing data variance. In *Proceedings of the 1991 Annual Research Conference*, pp. 429–440. Census Bureau, U.S.

Graf, M. (2011). Use of survey weights for the analysis of compositional data. In *Compositional Data Analysis, Theory and Applications*, Eds. V. Pawlowsky-Glahn & A. Buccianti, pp. 114-127. Chichester: Wiley.

Graf, M. (2015). A Simplified Approach to Linearization Variance for Surveys. Technical Report, Institut De Statistique, Université de Neuchâtel. Neuchâtel.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In *Handbook of Statistics, Volume 29, Sample Surveys: Design, methods and applications*, Eds. D. Pfeffermann & C.R. Rao, pp. 215–246. Amsterdam: Elsevier.

Hidiroglou, M. A. & Särndal, C. -E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, **24**, 11–20.
Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, **47**, 663–685.
Kim, J. K. & Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canad. J. Statist.*, **35**, 501–514.
Kim, J. K. & Rao, J. N. K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, **96**, 917–932.
Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, **32**, 133–142.
Langel, M. & Tillé, Y. (2013). Variance estimation of the Gini index: revisiting a result several times published. *J. Roy. Statist. Soc. Ser. A*, **A176**, 521–540.
Quenouille, M. H. (1949). Problems in plane sampling. *Ann. Math. Stat.*, **20**, 355–375.
Rao, J. N. K. (1990). Variance Estimation under Imputation for Missing Data. Technical Report, Statistics Canada. Ottawa.
Rao, J. N. K. & Shao, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika*, **79**, 811–822.
Rao, J. N. K. & Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, **82**, 453–460.
Särndal, C. -E. (1982). Implication of survey design for generalized regression estimation of linear functions. *J. Statist. Plann. Inference*, **7**, 155–170.
Särndal, C. -E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, **18**, 241–252.
Shao, J. & Sitter, R. R. (1996). Bootstrap for imputed survey data. *J. Amer. Statist. Assoc.*, **91**, 1278–1288.
Shao, J. & Steel, P. (1999). Variance estimation for survey data with composite imputation and nonneglible sampling fractions. *J. Amer. Statist. Assoc.*, **94**, 254–265.
Stuart, A. & Ord, J. K. (1994). *Kendall's Advanced Theory of Statistics*, Vol. 1. London (UK): Edward Arnold.
Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *J. Amer. Statist. Assoc.*, **66**, 411–414.
Zeileis, A. (2014). *ineq: Measuring Inequality, Concentration, and Poverty*. R package version 0.2-13.

[*Received April 2018, accepted January 2019*]

Appendix A: Proof of Proposition 1

Proof. First, let derive constraint (7) with respect to indicator a_ℓ and obtain

$$d_\ell F_\ell(\mathbf{x}_\ell^\top \boldsymbol{\lambda})\mathbf{x}_\ell + \sum_{k\in U} a_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda})\mathbf{x}_k \mathbf{x}_k^\top \frac{\partial \boldsymbol{\lambda}}{\partial a_\ell} = 0.$$

This leads to

$$\frac{\partial \boldsymbol{\lambda}}{\partial a_\ell} = -\widehat{\mathbf{T}}_{xx}^{-1} d_\ell F_\ell(\mathbf{x}_\ell^\top \boldsymbol{\lambda})\mathbf{x}_\ell = -\widehat{\mathbf{T}}_{xx}^{-1} w_\ell \mathbf{x}_\ell. \tag{A1}$$

The partial derivative of the calibration estimator is then given by

$$\frac{\partial \widehat{Y}}{\partial a_\ell} = w_\ell y_\ell - \sum_{k\in U} a_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) y_k \mathbf{x}_k^\top \widehat{\mathbf{T}}_{xx}^{-1} w_\ell \mathbf{x}_\ell = w_\ell e_\ell. \tag{A2}$$

Appendix B: Proof of Proposition 2

Proof. Noting that vector $\mathbf{w}$ depends on $\mathbf{a}$, the linearisation variable is $z_\ell^w = w_\ell v_\ell^w$, with

$$v_\ell^w = h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w}) + \sum_{k\in U} a_k h_{k2}(\mathbf{y}, \mathbf{a}, \mathbf{w}) \frac{\partial w_k}{\partial a_\ell}.$$

The derivative of w_k is obtained by differentiating Equation (7) with respect to a_ℓ:

$$w_\ell \mathbf{x}_\ell + \sum_{k\in U} a_k d_k \mathbf{x}_k \mathbf{x}_k^\top F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) \frac{\partial \boldsymbol{\lambda}}{\partial a_\ell} = \mathbf{0}.$$

Thus

$$\frac{\partial \boldsymbol{\lambda}}{\partial a_\ell} = -w_\ell \widehat{\mathbf{T}}_{xx}^{-1} \mathbf{x}_\ell,$$
$$\frac{\partial w_k}{\partial a_\ell} = d_k \frac{\partial F_k(\mathbf{x}_k^\top \boldsymbol{\lambda})}{\partial a_\ell} = -d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) \mathbf{x}_k^\top \widehat{\mathbf{T}}_{xx}^{-1} w_\ell \mathbf{x}_\ell$$

and

$$\begin{aligned} v_\ell^w &= h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w}) - \mathbf{x}_\ell^\top \widehat{\mathbf{T}}_{xx}^{-1} \sum_{k\in U} a_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) \mathbf{x}_k h_{k2}(\mathbf{y}, \mathbf{a}, \mathbf{w}) \\ &= h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w}) - \mathbf{x}_\ell^\top \widehat{\mathbf{B}}_{h1|x}. \end{aligned}$$

Appendix C: Proof of Proposition 3

Proof. The partial derivative of equation (15) with respect to δ_ℓ is

$$d_\ell F_\ell(\mathbf{x}_\ell^\top \boldsymbol{\lambda}) \mathbf{x}_\ell + \sum_{k\in U} \delta_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) \mathbf{x}_k \mathbf{x}_k^\top \frac{\partial \boldsymbol{\lambda}}{\partial \delta_\ell} = 0.$$

This implies that $\partial \boldsymbol{\lambda}/\partial \delta_\ell = -\widehat{\mathbf{T}}_{xx}^{-1} w_\ell \mathbf{x}_\ell$. The linearisation variable of $\widehat{Y}_R$ is

$$z_\ell = \frac{\partial \widehat{Y}_R}{\partial \delta_\ell} = d_\ell F_\ell(\mathbf{x}_\ell^\top \boldsymbol{\lambda}) y_\ell - \sum_{k\in U} \delta_k d_k F_k(\mathbf{x}_k^\top \boldsymbol{\lambda}) y_k \mathbf{x}_k^\top \widehat{\mathbf{T}}_{xx}^{-1} w_\ell \mathbf{x}_\ell = w_\ell e_\ell.$$

Appendix D: Proofs of Propositions 4 and 5

Proof. By considering Proposition 1, the linearisation variable with respect to a_ℓ of $\sum_{k\in U} a_k w_{k1} \mathbf{x}_k^\circ$ is $\mathbf{z}_\ell^\circ = w_{\ell 1} \mathbf{e}_\ell^\circ$. The partial derivative of the calibration constraint (17) with respect to a_ℓ is

$$R_\ell w_{\ell 2} \mathbf{x}_\ell + \widehat{\mathbf{T}}_{xx} \frac{\partial \boldsymbol{\lambda}_2}{\partial a_\ell} = \begin{pmatrix} \mathbf{0} \\ w_{\ell 1} \mathbf{e}_\ell^\circ \end{pmatrix},$$

which leads to

$$\frac{\partial \boldsymbol{\lambda}_2}{\partial a_\ell} = \widehat{\mathbf{T}}_{xx}^{-1} \left[\begin{pmatrix} \mathbf{0} \\ w_{\ell 1} \mathbf{e}_\ell^\circ \end{pmatrix} - R_\ell w_{\ell 2} \mathbf{x}_\ell \right].$$

The partial derivative of the estimator with respect to a_ℓ is

$$\frac{\partial\widehat{\theta}_{NR}(\mathbf{y},\mathbf{a},\mathbf{R})}{\partial a_\ell} = R_\ell w_{\ell 2} y_\ell + \widehat{\mathbf{t}}_{xy}^\top \frac{\partial \boldsymbol{\lambda}_2}{\partial a_\ell} = R_\ell w_{\ell 2} y_\ell + \left[\begin{pmatrix} \mathbf{0} \\ w_{\ell 1}\mathbf{e}_\ell^\circ \end{pmatrix} - R_\ell w_{\ell 2}\mathbf{x}_\ell\right]^\top \widehat{\mathbf{B}}_{y|x}.$$

Proof. When $\mathbf{a} = \boldsymbol{\pi}$, the estimator becomes

$$\widehat{\theta}_{NR}(\mathbf{y},\boldsymbol{\pi},\mathbf{R}) = \sum_{k\in U} \pi_k R_k w_{k2}^\pi y_k,$$

where the weights are obtained by solving

$$\sum_{k\in U} \pi_k R_k w_{k2}^\pi \mathbf{x}_k = \begin{pmatrix} \mathbf{X}^\bullet \\ \sum_{k\in U} \pi_k w_{k1}^\pi \mathbf{x}_k^\circ \end{pmatrix},$$

and

$$\sum_{k\in U} \pi_k w_{k1}^\pi \mathbf{x}_k^\bullet = \mathbf{X}^\bullet.$$

The partial derivative of $\widehat{\theta}_{NR}(\mathbf{y},\boldsymbol{\pi},\mathbf{R})$ with respect to R_ℓ is

$$\frac{\partial\widehat{\theta}_{NR}(\mathbf{y},\boldsymbol{\pi},\mathbf{R})}{\partial R_\ell} = \pi_\ell w_{2\ell}^\pi y_\ell + \sum_{k\in U} R_k F_{k2}'(\mathbf{x}_k^\top \boldsymbol{\lambda}_2^\pi)\mathbf{x}_k^\top \frac{\partial \boldsymbol{\lambda}_2^\pi}{\partial R_\ell} y_k = F_{\ell 2}(\mathbf{x}_\ell \boldsymbol{\lambda}_2^\pi) e_\ell^\pi,$$

where $\partial\boldsymbol{\lambda}_2^\pi/\partial R_\ell$ is the solution to

$$\frac{\partial \sum_{k\in U} \pi_k R_k w_{k2}^\pi \mathbf{x}_k}{\partial R_\ell} = \pi_\ell w_{2\ell}^\pi \mathbf{x}_\ell + \sum_{k\in U} R_k F_{k2}'(\mathbf{x}_k^\top \boldsymbol{\lambda}_2^\pi)\mathbf{x}_k \mathbf{x}_k^\top \frac{\partial \boldsymbol{\lambda}_2^\pi}{\partial R_\ell} = \mathbf{0}.$$

Appendix E: Proof of Proposition 6

Proof. The partial derivative of $\widehat{\theta}_{NR}$ with respect to a_ℓ is

$$z_\ell^a = \frac{\partial\widehat{\theta}_{NR}}{\partial a_\ell} = \frac{h_{\ell 1}(\tilde{\mathbf{y}},\mathbf{a},\mathbf{R})}{\pi_\ell} + \sum_{k\in U} a_k(1-R_k) h_{k2}(\tilde{\mathbf{y}},\mathbf{a},\mathbf{R}) \frac{\partial y_k^*}{\partial a_\ell}.$$

In the case of regression imputation,

$$\frac{\partial y_k^*}{\partial a_\ell} = \frac{R_\ell e_\ell \mathbf{x}_\ell^\top}{\pi_\ell} \widehat{\mathbf{T}}_{xx}^{-1} \mathbf{x}_k.$$