Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

The effect of nonresponse adjustments on variance estimation

by David Haziza,Katherine Jenny Thompson and Wesley Yung

June 2010

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2010
Vol. 36, No. 1, pp. 35-43
Statistics Canada, Catalogue No. 12-001-X

The effect of nonresponse adjustments on variance estimation

David Haziza, Katherine Jenny Thompson and Wesley Yung [1]

Abstract

Many surveys employ weight adjustment procedures to reduce nonresponse bias. These adjustments make use of available auxiliary data. This paper addresses the issue of jackknife variance estimation for estimators that have been adjusted for nonresponse. Using the reverse approach for variance estimation proposed by Fay (1991) and Shao and Steel (1999), we study the effect of not re-calculating the nonresponse weight adjustment within each jackknife replicate. We show that the resulting 'shortcut' jackknife variance estimator tends to overestimate the true variance of point estimators in the case of several weight adjustment procedures used in practice. These theoretical results are confirmed through a simulation study where we compare the shortcut jackknife variance estimator with the full jackknife variance estimator obtained by re-calculating the nonresponse weight adjustment within each jackknife replicate.

Key Words: Calibration; Nonresponse adjustment; Unit nonresponse; Jackknife variance estimator; Linearization variance estimator.

1. Introduction

Unit nonresponse, which occurs when, for a sample unit, all the survey variables are missing or when not enough usable information is available, is unavoidable in surveys. To address this, the nonrespondents are deleted from the data file and the survey weights of the respondents are adjusted to compensate for the deletions. The primary objective of a weight adjustment procedure is to reduce the nonresponse bias, which is introduced when respondents and nonrespondents are different with respect to the survey variables. Key to achieving an efficient bias reduction is the use of powerful auxiliary information available for both respondents and nonrespondents.

In this paper, we consider jackknife variance estimation in the presence of unit nonresponse. This variance estimation method is widely used in practice because of its theoretical properties and computational ease. In contrast to Taylor linearization procedures, the jackknife method does not require a separate derivation for each parameter of interest nor the second-order inclusion probabilities that may be difficult to obtain in complex surveys. When using a jackknife variance estimator in the context of nonresponse, there is some question of whether or not the nonresponse adjustment needs to be replicated (*e.g.*, Valliant 2004). In this paper, we consider two jackknife variance estimators: (i) a *full* jackknife variance estimator which recalculates the nonresponse adjustment factor within each jackknife replicate and (ii) a *shortcut* jackknife variance estimator, which does not. The shortcut jackknife variance estimator is convenient in practice but its theoretical properties were not, to our knowledge, fully studied in the literature. Production reasons tend to drive the usage of a shortcut jackknife variance estimator, since the full jackknife variance estimator in the context of stratified sampling can be quite time-consuming and computer resource-intensive, especially when a survey utilizes a large number of weighting cells. Some recent studies conducted at the U.S. Census Bureau (Thompson 2005 and Ozcoskun, Thompson and Williams 2005) found negligible differences between variance estimates obtained using a fully replicated weight adjustment procedure and those obtained using a "shortcut" procedure with stratified jackknife, delete-a-group jackknife, and modified half sample variance estimators.

Two types of adjustment procedures are commonly used in practice. The first, called *nonresponse propensity weighting* (NPW), consists of first modeling the response propensities and using the inverse of the estimated propensities as the weighting adjustment. The estimated response propensities are typically obtained by fitting a parametric model (*e.g.*, logistic regression model) or by fitting a nonparametric model; *e.g.*, Da Silva and Opsomer (2006). A special case of NPW, which is very popular in practice, consists of first dividing the respondents and nonrespondents into weighting classes and adjusting the design weights of respondents by the inverse of the response rate within each class. These classes are formed on the basis of auxiliary information recorded for all units in the sample; see, for example, Eltinge and Yansaneh (1997) and Little (1986). The second type of adjustment procedures, called *nonresponse calibration weighting* (NCW) can be seen as an extension of the calibration approach (Deville and Särndal 1992) adapted to the context of unit nonresponse. The reader is referred to Särndal and Lundström (2005), Kott (2006) and Brick and Montaquila (2008) for a comprehensive overview of NPW and NWC. In some

1. David Haziza, Département de mathématiques et de statistique, Université de Montréal, Montréal, H3C 3J7, Canada. E-mail: David.haziza@umontreal.ca; Katherine Jenny Thompson, U.S. Census Bureau, Washington, DC 20233. E-mail: Katherine.J.Thompson@census.gov; Wesley Yung, Statistics Canada, Ottawa, Ontario, K1A 0T6. E-mail: wesley.yung@statcan.gc.ca.

situations, NPW and NCW lead to the same estimator; for example, the count-adjusted estimator presented below (see expression (1.4)). In this paper, we focus on NCW. The problem of variance estimation in the context of NPW has been recently studied by Kim and Kim (2007).

Consider a finite population U of size N. The objective is to estimate the population total $Y = \sum_{i \in U} y_i$, of a variable of interest y. Suppose that a random sample s of size n is selected from U according to a given design $p(s)$. In the case of complete data, a basic estimator of Y is the well-known expansion estimator given by

$$\hat{Y}_\pi = \sum_{i \in s} d_i y_i \tag{1.1}$$

where $d_i = 1/\pi_i$ denotes the design weight attached to unit i and $\pi_i = P(i \in s)$ denotes its first-order probability of inclusion in the sample. In the presence of unit nonresponse, only a subset of s is observed, and so the computation of $\hat{Y}_\pi$ in (1.1) is not possible.

To define a nonresponse adjusted estimator of Y, we assume that a vector of auxiliary variables $\mathbf{x}$ is available for all the sampled units (respondents and nonrespondents) so that the vector of estimated totals, $\hat{\mathbf{X}}_\pi = \sum_{i \in s} d_i \mathbf{x}_i$, is available. We also assume that a vector of instrumental variables $\mathbf{z}$, of the same dimension as $\mathbf{x}$, is available for the respondents. Let r_i be a response indicator attached to unit i such that $r_i = 1$ if unit i is a responding unit and $r_i = 0$, otherwise. To estimate Y, we consider calibration estimators of the form

$$\hat{Y}_{\text{CAL}} = \sum_{i \in s} w_i r_i y_i, \tag{1.2}$$

where $w_i = d_i g_i$ and g_i is a nonresponse weighting adjustment factor attached to unit i and given by

$$g_i = 1 + (\hat{\mathbf{X}}_\pi - \hat{\mathbf{X}}_r)' \hat{\mathbf{T}}_r^{-1} \mathbf{z}_i, \tag{1.3}$$

where $\hat{\mathbf{X}}_r = \sum_{i \in s} d_i r_i \mathbf{x}_i$ and $\hat{\mathbf{T}}_r = \sum_{i \in s} d_i r_i \mathbf{z}_i \mathbf{x}_i'$. When $\mathbf{z}_i = \mathbf{x}_i/\nu_i$, where ν_i is a known constant, then the estimator (1.3) is identical to the *InfoS* estimator given in Särndal and Lundström (2005, equation 7.15). The properties of the estimator (1.2) were studied by Deville (2002), Sautory (2003), Särndal and Lundström (2005) and Kott (2006), among others.

In this paper, the properties (*e.g.*, bias and variance) of $\hat{Y}_{\text{CAL}}$ are studied using the nonresponse model (NM) approach, under which inference is made with respect to the joint distribution induced by the sampling design and the nonresponse mechanism, $q(\mathbf{r} \mid \mathbf{I})$, where $\mathbf{I} = (I_1, ..., I_N)'$ is the vector of sample selection indicators such that $I_i = 1$ if unit i is selected in the sample and $I_i = 0$, otherwise and $\mathbf{r} = (r_1, ..., r_N)'$ is the vector of response indicators. Let $p_i = P(r_i = 1 \mid \mathbf{I}, I_i = 1)$ be the response probability for unit i. We assume that $p_i > 0$ for all i and that the units respond independently of one another; that is, $p_{ij} = P(r_i = 1, r_j = 1 \mid \mathbf{I}, I_i = 1, I_j = 1, i \neq j) = p_i p_j$.

The estimator $\hat{Y}_{\text{CAL}}$ is asymptotically unbiased for the true total Y if (i) $p_i^{-1} = 1 + \boldsymbol{\lambda}' \mathbf{z}_i$ for all $i \in U$, where $\boldsymbol{\lambda}$ is a vector of unknown constants or (ii) $y_i = \mathbf{x}_i' \boldsymbol{\beta}$ for all $i \in U$, where $\boldsymbol{\beta}$ is a vector of constants; see Särndal and Lundström (2005, chapter 9.5). If the condition (i) is satisfied, the point estimator $\hat{Y}_{\text{CAL}}$ is asymptotically unbiased for Y regardless of the variable of interest y being estimated. Also, it follows from (ii) that $\hat{Y}_{\text{CAL}}$ has a small bias if the residuals $E_i = y_i - \mathbf{x}_i' \mathbf{B}$, are small, where $\mathbf{B} = (\sum_{i \in U} \mathbf{z}_i \mathbf{x}_i')^{-1} \sum_{i \in U} \mathbf{z}_i y_i$. Therefore, the bias of the estimator $\hat{Y}_{\text{CAL}}$ is small if the vector $\mathbf{x}$ explains the variable of interest y. In the case of several variables of interest, note that the vector $\mathbf{x}$ may explain a given variable of interest well but may not be related to all, in which case some estimates could be potentially biased. We assume that $\hat{Y}_{\text{CAL}}$ is asymptotically unbiased for Y, so that the bias of the estimators under consideration is not an issue in the reminder of the paper.

We consider three special cases of (1.2) that are of interest in practice (see also Kalton and Flores-Cervantes 2003). First, let $\boldsymbol{\delta} = (\delta_{i1}, ..., \delta_{ic}, ..., \delta_{iC})'$ be a C-vector of weighting class indicators attached to unit i such that $\delta_{ic} = 1$ if unit i belongs to class c and $\delta_{ic} = 0$, otherwise for $c = 1, ..., C$. If $\mathbf{x}_i = \mathbf{z}_i = \boldsymbol{\delta}_i$, the adjustment factor g_i given by (1.3) reduces to $g_i = \hat{N}_c / \hat{N}_{rc}\, \delta_{ic}$, where $\hat{N}_c = \sum_{i \in s} d_i \delta_{ic}$ and $\hat{N}_{rc} = \sum_{i \in s} d_i r_i \delta_{ic}$. That is, the nonresponse weighting adjustment factor for a weighting cell is calculated as the sample-weighted number of sampled units in the weighting cell divided by the sample-weighted number of responding units in the weighting cell. We refer to this weight adjustment procedure as the *count adjustment* procedure. It follows that the estimator (1.2) reduces to the count adjusted estimator

$$\hat{Y}_{\text{count}} = \sum_{c=1}^{C} \frac{\hat{N}_c}{\hat{N}_{rc}} \hat{Y}_{rc}, \tag{1.4}$$

where

$$\hat{Y}_{rc} = \sum_{i \in s} d_i r_i \delta_{ic} y_i.$$

The second special case of (1.2) assumes that a continuous variable x is available for all the sampled units. Let $\mathbf{x}_i = (\delta_{i1} x_i, ..., \delta_{ic} x_i, ..., \delta_{iC} x_i)'$ and $\mathbf{z}_i = \boldsymbol{\delta}_i$. In this case, the adjustment factor g_i given by (1.3) reduces to $g_i = \hat{X}_c / \hat{X}_{rc}\, \delta_{ic}$ if unit i belongs to class c, where $\hat{X}_c = \sum_{i \in s} d_i \delta_{ic} x_i$ and $\hat{X}_{rc} = \sum_{i \in s} d_i r_i \delta_{ic} x_i$. Here, the nonresponse weighting adjustment factor for a weighting class c is the sum of the sample-weighted auxiliary data for units in the weighting cell divided by the sum of the

sample-weighted auxiliary data for all responding units in the weighting cell. We refer to this weight adjustment procedure as the *ratio adjustment* procedure. The estimator (1.2) reduces to the ratio adjusted estimator

$$\hat{Y}_{\text{ratio}} = \sum_{c=1}^{C} \frac{\hat{X}_c}{\hat{X}_{rc}} \hat{Y}_{rc}. \qquad (1.5)$$

Note that the count adjusted estimator (1.4) is a special case of the ratio adjusted estimator when $x_i = 1$ for all the sampled population units.

Finally, if $\mathbf{x}_i = \mathbf{z}_i = (\delta_{i1}, ..., \delta_{ic}, ..., \delta_{iC}, \delta_{i1} x_i, ..., \delta_{ic} x_i, ..., \delta_{iC} x_i)'$, we obtain another special case of (1.2). In this case, the adjustment factor g_i given by (1.3) reduces to

$$g_i = \hat{N}_c \left[1 + (\bar{x}_c - \bar{x}_{rc}) \frac{(x_i - \bar{x}_{rc})}{\sum_{i \in s} r_i \delta_{ic} (x_i - \bar{x}_{rc})^2} \right],$$

if unit i belongs to class c, where $\bar{x}_c = \hat{X}_c / \hat{N}_c$ and $\bar{x}_{rc} = \hat{X}_{rc} / \hat{N}_{rc}$. We refer to this weight adjustment procedure as the *simple linear regression adjustment* procedure. The estimator (1.2) reduces to the simple linear regression adjusted estimator

$$\hat{Y}_{slreg} = \sum_{c=1}^{C} \hat{N}_c [\hat{Y}_{rc} + (\bar{x}_c - \bar{x}_{rc}) \hat{B}_{rc}], \qquad (1.6)$$

where

$$\hat{B}_{rc} = \frac{\sum_{i \in s} d_i r_i \delta_{ci} (x_i - \bar{x}_{rc})(y_i - \bar{y}_{rc})}{\sum_{i \in s} d_i r_i \delta_{ic} (x_i - \bar{x}_{rc})^2}.$$

The estimators (1.4)-(1.6) use some form of weighting adjustment within classes. All of them are asymptotically unbiased for Y if the units have equal response probabilities within classes (*i.e.*, a uniform nonresponse mechanism within classes). This condition is a special case of condition (i) discussed above.

In this paper, we show that the shortcut jackknife variance estimator that treats the adjustment factors as fixed, tends to overestimate the true variance of $\hat{Y}_{\text{CAL}}$, at least in some simple cases. We build on earlier research by Thompson and Yung (2006) who derived expressions of the linearization version for both the full and shortcut jackknife variance estimators and evaluated these expressions empirically using data from the Annual Capital Expenditures Survey (ACES), conducted at the U.S. Census Bureau. In the context of NPW, it is interesting to note that Kim and Kim (2007) showed that treating the estimated response probabilities as fixed leads to an overestimation of the true variance when the sampling weights are not used in estimating these probabilities. Beaumont (2005) obtained similar results in the context of imputation when the response probabilities are estimated using a logistic regression model.

In Section 2, we discuss the full and shortcut jackknife variance estimators and show that the shortcut estimator is asymptotically biased. The severity of this bias is evaluated for two commonly used sample designs in Section 3. Section 4 presents the results of a simulation study comparing the full and shortcut jackknife variance estimators. We conclude in Section 5 with some general observations.

2. Jackknife variance estimation

Traditionally, variance estimation in the context of nonresponse has been performed using the two-phase framework, which consists of viewing nonresponse as a second-phase of selection. Instead, we consider the reverse framework that was proposed by Fay (1991) and further developed by Shao and Steel (1999). This framework provides a theoretical basis for studying the properties of jackknife variance estimators and can be described as follows: first, applying the nonresponse mechanism, the population U is randomly divided into a population of respondents U_r and a population of nonrespondents U_m. Then, given (U_r, U_m), the random sample s is selected according to the chosen sampling design. The total variance of $\hat{Y}_{\text{CAL}}$ can be expressed as

$$V(\hat{Y}_{\text{CAL}}) = E_q V_p (\hat{Y}_{\text{CAL}} \mid \mathbf{r}) + V_q E_p (\hat{Y}_{\text{CAL}} \mid \mathbf{r}), \qquad (2.1)$$

where $E_p(.)$ and $V_p(.)$ denote the expectation and the variance with respect to the sampling design and $E_q(.)$ and $V_q(.)$ denote the expectation and variance with respect to the nonresponse mechanism, $q(\mathbf{r} \mid \mathbf{I})$.

In this section, we focus on stratified simple random sampling, which is the design typically used in business surveys. With this sample design, the population U is partitioned into L strata $U_1, ..., U_L$ of size $N_1, ..., N_L$, respectively. A simple random sample without replacement s_h, of size n_h, is selected from stratum h, $h = 1, ..., L$. Each within-stratum sample is selected independently, and we assume that $n_h \geq 2$ for all h. In this context, the design weight of unit i in stratum h is $d_{hi} = N_h / n_h$. A full jackknife variance estimator of $\hat{Y}_{\text{CAL}}$, under stratified simple random sampling, is obtained as follows:

(i) remove unit (gj) from the sample, $g = 1, ..., L$; $j = 1, ..., n_h$;

(ii) adjust the design weights d_{hi} to obtain the jackknife weights $d_{hi(gj)}$, where $d_{hi(gj)}$ is given by

$$d_{hi(gj)} = \begin{cases} 0 & \text{if } (hi) = (gj) \\ \dfrac{n_g}{n_g - 1} d_{gi} & \text{if } h = g,\ i \neq j \\ d_{hi} & \text{otherwise} \end{cases}$$

(iii) compute the estimator $\hat{Y}_{\mathrm{CAL}(gj)}$ in the same way as $\hat{Y}_{\mathrm{CAL}}$ with the jackknife weights $d_{hi(gj)}$ instead of the design weights d_{hi}; that is, $\hat{Y}_{\mathrm{CAL}(gj)} = \sum_{(hi)\in s} w_{hi(gj)} r_{hi} y_{hi}$, where $w_{hi(gj)} = d_{hi(gj)} g_{hi(gj)}$ with $g_{hi(gj)} = 1 + (\hat{\mathbf{X}}_{\pi(gj)} - \hat{\mathbf{X}}_{r(gj)})' \hat{\mathbf{T}}_{r(gj)}^{-1} \mathbf{z}_{hi}$, $\hat{\mathbf{X}}_{\pi(gj)} = \sum_{i\in s} d_{hi(gj)} \mathbf{x}_{hi}$, $\hat{\mathbf{X}}_{r(gj)} = \sum_{(hi)\in s} d_{hi(gj)} r_{hi} \mathbf{x}_{hi}$ and $\hat{\mathbf{T}}_{r(gj)} = \sum_{(hi)\in s} d_{hi(gj)} r_{hi} \mathbf{z}_{hi} \mathbf{x}_{hi}'$.

(iv) replace the unit deleted in step (i) back into the sample;

(v) repeat steps (i)-(iv) for all (gj) units, $g = 1, ..., L$; $j = 1, ..., n_h$.

Note that the nonresponse adjustment factors g_{hi} are recalculated in each replicate. This leads to the full jackknife variance estimator

$$v_{JF} = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j\in s_h} (\hat{Y}_{\mathrm{CAL}(gj)} - \hat{Y}_{\mathrm{CAL}})^2. \qquad (2.2)$$

The variance estimator v_{JF} is an estimator of the first term on the right hand side of (2.1), $E_q V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$. This term represents the design variance that we would have obtained if the responding units were selected using stratified simple random sampling with replacement, or equivalently, if the stratum sampling fractions, (n_h / N_h) are negligible. In other words, the full jackknife variance estimator (2.2) is an estimator of the sampling variance conditional on the vector of response indicators $\mathbf{r}$. Therefore, v_{JF} is asymptotically unbiased and consistent for $E_q V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$ under stratified simple random sampling with replacement sampling regardless of the validity of the underlying assumptions. Note that since v_{JF} is an estimator of a sampling variance, it can be readily obtained using software designed for complete-data jackknife variance estimation. In other words, no specialized software is needed. Also, note that the second term on the right hand side of (2.1), $V_q E_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$, is not accounted for. Thus, the full jackknife variance estimator does not track the second term in (2.1). However, the contribution of this term to the total variance is negligible if the stratum sampling fractions, n_h / N_h, are negligible. As a result, v_{JF} is asymptotically unbiased and consistent for the total variance, $V(\hat{Y}_{\mathrm{CAL}})$. That is, $E_{pq}(v_{JF}) \approx V(\hat{Y}_{\mathrm{CAL}})$. Since the goal of the research is to compare the full and shortcut jackknife estimators, in the remainder of the paper, we assume that the stratum sampling fractions are negligible and focus on estimates of totals, so that we can omit the estimation of the second term in (2.1). We note that even if the second term is not negligible, our comparisons are valid as both the full jackknife and shortcut estimators would underestimate the total variance by the same term.

A shortcut jackknife variance estimator of $\hat{Y}_{\mathrm{CAL}}$ is given by

$$v_{JS} = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j\in s_h} (\hat{Y}^*_{\mathrm{CAL}(gj)} - \hat{Y}_{\mathrm{CAL}})^2, \qquad (2.3)$$

where $\hat{Y}^*_{\mathrm{CAL}(gj)} = \sum_{(hi)\in s} d_{hi(gj)} g_{hi} r_{hi} y_{hi}$. Note that the nonresponse weighting adjustment factors g_{hi} are not recalculated in each jackknife replicate. In other words, the factors g_{hi} are treated as constants, which is inappropriate since they depend on the sample and the set of respondents. Therefore, we have $E_{pq}(v_{JS}) \neq V(\hat{Y}_{\mathrm{CAL}})$, in general, and the shortcut variance estimator, v_{JS}, is biased.

To study the magnitude of the bias of v_{JS}, we consider the difference of the two jackknife variance estimators, $D = v_{JS} - v_{JF}$. Since the variance estimator v_{JF} is an asymptotically unbiased estimator of the term $V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$, it is asymptotically equivalent to a variance estimator obtained using a first-order Taylor expansion. The resulting variance estimator, denoted by $\tilde{v}_{JF}$, is the linearization jackknife variance estimator studied by Yung and Rao (2000). Similarly, the shortcut jackknife variance estimator v_{JS} is asymptotically equivalent to a variance estimator of $V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$ obtained by treating the nonresponse weighting adjustment factors g_{hi} as constants. We denote this variance estimator by $\tilde{v}_{JS}$. The quantity D can thus be approximated by $\tilde{D} = \tilde{v}_{JS} - \tilde{v}_{JF}$. For this approximation to be valid, we assume the number of respondents to be large.

Noting that $\mathrm{Bias}(v_{JF}) = E_{pq}(v_{JF}) - V(\hat{Y}_{\mathrm{CAL}}) \approx 0$, it follows that the bias of v_{JS}, $\mathrm{Bias}(v_{JS}) = E_{pq}(v_{JS}) - V(\hat{Y}_{\mathrm{CAL}})$, can be approximated by $E_{pq}(D) \approx E_{pq}(\tilde{D})$. Let $v(y)$ denote the variance estimator of the complete data estimator (1.1). Using a first-order Taylor expansion, it can be shown that an estimator of $V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$ is given by

$$\tilde{v}_{JF} = v(\hat{\xi}) \qquad (2.4)$$

where

$$\hat{\xi}_{hi} = \mathbf{x}_{hi}' \hat{\mathbf{B}}_r + g_{hi} r_{hi} e_{hi},$$

with $e_{hi} = (y_{hi} - \mathbf{x}_{hi}' \hat{\mathbf{B}}_r)$ and $\hat{\mathbf{B}}_r = \hat{\mathbf{T}}_r^{-1} \sum_{(hi)\in s} d_{hi} r_{hi} \mathbf{z}_{hi} y_{hi}$. On the other hand, treating the g_{hi}'s as constants implies that $\hat{Y}_{\mathrm{CAL}}$ is linear in the design weights d_{hi}. It follows that $\tilde{v}_{JS}$ is given by

$$\tilde{v}_{JS} = v(\psi), \qquad (2.5)$$

where $\psi_{hi} = g_{hi} r_{hi} y_{hi}$.

For example, for either a fixed size or a random size sampling design, a possible variance estimator is

$$\tilde{v}_{JF} = \sum_{i\in s} \sum_{j\in s} \Delta_{ij} \hat{\xi}_i \hat{\xi}_j,$$

where $\Delta_{ij} = (\pi_{ij} - \pi_i \pi_j) / \pi_{ij} \pi_i \pi_j$ and π_{ij} is the second-order inclusion probability of units i and j. Note that $\pi_{ii} = \pi_i$. Similarly, we have

$$\tilde{v}_{JS} = \sum_{i\in s}\sum_{j\in s}\Delta_{ij}\,\psi_i\,\psi_j.$$

3. Bias of v_{JS} in some special cases

3.1 Simple random sampling without replacement

In this section, we assume that the sample s has been selected according to simple random sampling without replacement. We also assume that the sampling fraction n/N is negligible and that the number of respondents r is large. Finally, we assume a single weighting class. Although the above situation is not realistic in practice, it provides some insight into the asymptotic bias of v_{JS}.

In the case of the ratio adjusted estimator (1.5), we can show that $\tilde{D}$ is approximately given by

$$\tilde{D} = \frac{N^2}{r}\left(1-\frac{r}{n}\right)\left\{\left(\frac{\bar{x}}{\bar{x}_r}\right)^2 (s_{yr}^2 - s_{er}^2) + 2\left(\frac{\bar{x}}{\bar{x}_r}\right)\hat{R}_r\left[\left(\frac{\bar{x}}{\bar{x}_r}\right)-1\right]\frac{s_{exr}}{n} + \hat{R}_r^2\left[\left(\frac{\bar{x}}{\bar{x}_r}\right)^2 s_{xr}^2 - s_x^2\right] + \left(\frac{\bar{x}}{\bar{x}_r}\right)^2 \bar{y}_r^2\right\}, \quad (3.1)$$

where $(\bar{x}_r, \bar{y}_r) = 1/r\sum_{i\in s} r_i (x_i, y_i)$ denote the mean of the respondents for variable x and y respectively and r is the number of respondents, $\hat{R}_r = \bar{y}_r/\bar{x}_r$, $s_{xr}^2 = 1/(r-1)\sum_{i\in s} r_i (x_i - \bar{x}_r)^2$, $s_x^2 = 1/(n-1)\sum_{i\in s}(x_i - \bar{x})^2$ with $\bar{x} = 1/n\sum_{i\in s} x_i$, $s_{er}^2 = 1/(r-1)\sum_{i\in s} r_i (y_i - \hat{R}_r x_i)^2$ and $s_{exr} = 1/(r-1)\sum_{i\in s} r_i (y_i - \hat{R}_r x_i) x_i$. If we further assume that all units have equal response probabilities (*i.e.*, a uniform response mechanism), we have $\bar{x}/\bar{x}_r \xrightarrow{p} 1$ and $s_{xr}^2/s_x^2 \xrightarrow{p} 1$. In this case, the asymptotic bias of v_{JS} is given by

$$\text{Bias}(v_{JS}) \approx E_{pq}(\tilde{D}) \approx \frac{N^2}{E_{pq}(r)}\left(1 - E_{pq}\left(\frac{r}{n}\right)\right) S_y^2\left(\frac{1}{\text{CV}(y)^2} + 2\frac{\text{CV}(x)}{\text{CV}(y)}\rho_{xy} - \frac{\text{CV}(x)^2}{\text{CV}(y)^2}\right), \quad (3.2)$$

where $\text{CV}(x) = S_x/\bar{X}$ and $\text{CV}(y) = S_y/\bar{Y}$ denote the population coefficients of variation for variables x and y, respectively with $S_y^2 = 1/(N-1)\sum_{i\in U}(y_i - \bar{Y})^2$ and $\bar{Y} = 1/N\sum_{i\in U} y_i$, S_x^2 and $\bar{X}$ are defined similarly, and ρ_{xy} denotes the finite population coefficient of correlation for variables x and y. From (3.2), it follows that the asymptotic bias of v_{JS} is nonnegative if and only if

$$B_0 < \frac{\bar{Y}}{2}\left(\frac{1+\text{CV}(x)^2}{\text{CV}(x)^2}\right), \quad (3.3)$$

provided $0 < E_{pq}(r/n) < 1$, where $B_0 = \bar{Y} - B_1\bar{X}$ is the finite population intercept of the least squares line when regressing y on x with

$$B_1 = \frac{\sum_{i\in U}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i\in U}(x_i - \bar{X})^2}.$$

From (3.2), it is clear that the bias of v_{JS} increases if (i) the expected response rate $E_{pq}(r/n)$ decreases; (ii) ρ_{xy} increases; (iii) $\text{CV}(y)$ decreases; or (iv) $\text{CV}(x)$ increases. Also, it follows from (3.3) that v_{JS} overestimates the true variance when the intercept B_0 is not too large. Table 1 illustrates the relationship between $\text{CV}(x)$ and the condition in (3.3). For example, when $\text{CV}(x) = 0$, v_{JS} always overestimates the true variance since, in this case, the condition (3.3) reduces to $B_0 < \infty$, which is always satisfied. This result is not surprising because when $\text{CV}(x) = 0$, the x-values are all equal and the ratio adjusted estimator (1.5) is identical to the count adjusted estimator (1.4). As we discuss below, v_{JS} always overestimates the true variance in this case. When $\text{CV}(x)$ is large (*e.g.*, $\text{CV}(x) = 2$), v_{JS} overestimates the true variance if and only if $B_0 < 0.625\bar{Y}$ The latter condition is satisfied if the intercept is not "too far" from the origin. Therefore, if the relationship between y and x goes through the origin (*i.e.*, if the ratio model holds), the shortcut variance estimator will overestimate the true variance. However, if the ratio adjusted estimator is used when the ratio model does not hold, such as when $B_0 \geq 0.625\bar{Y}$, the shortcut variance estimator v_{JS} will underestimate the true variance. In conclusion, we can expect v_{JS} to overestimate the true variance when a ratio adjustment procedure is used unless the ratio model is highly misspecified for the data at hand, which could happen, for example, if the variables y and x are negatively correlated.

Table 1
Relationship between $\text{CV}(x)$ and the condition in (3.3)

$\text{CV}(x)$	$\frac{\bar{Y}}{2}\left(\frac{1+\text{CV}(x)^2}{\text{CV}(x)^2}\right)$
0	∞
0.1	$50.5\,\bar{Y}$
0.5	$2.5\,\bar{Y}$
1	$2\,\bar{Y}$
1.5	$0.722\,\bar{Y}$
2	$0.625\,\bar{Y}$

Turning to the count adjusted estimator (1.4), we let $x_i = 1$ for all i in (3.1) and obtain

$$\tilde{D} = \frac{N^2}{r}\left(1-\frac{r}{n}\right)\bar{y}_r^2. \quad (3.4)$$

It follows from (3.4) that the relative bias of v_{JS}, $\mathrm{RB}(v_{JS}) = \mathrm{Bias}(v_{JS})/V(\hat{Y}_{\mathrm{CAL}})$, can be approximated by by $E_{pq}(R\tilde{D})$ where $R\tilde{D} = \tilde{D}/\tilde{v}_{JF}$. Under a uniform nonresponse mechanism, straightforward algebra leads to

$$\mathrm{RB}(v_{JS}) \approx E_{pq}(R\tilde{D}) \approx \left(1 - E_{pq}\left(\frac{r}{n}\right)\right)\frac{1}{\mathrm{CV}(y)^2}. \quad (3.5)$$

The expression (3.5) shows that, in the case of the count adjusted estimator (1.4), v_{JS} always overestimates the true variance. The magnitude of the overestimation increases as the expected response rate $E_{pq}(r/n)$ decreases or when $\mathrm{CV}(y)$ decreases. For example, if the expected response rate is equal to 70% and $\mathrm{CV}(y) = 1$, we have $E_{pq}(R\tilde{D}) = 1.3$ so the shortcut jackknife variance estimator, v_{JS}, is on average 30% larger than the true variance of $\hat{Y}_{\mathrm{CAL}}$. On the other hand, if the response rate is equal to 70% and $\mathrm{CV}(y) = 0.5$, we have $E_{pq}(R\tilde{D}) = 5.3$, in which case the overestimation is considerable.

Finally, we turn to the case of the simple linear regression adjusted estimator (1.6). Under a uniform nonresponse mechanism, it can be shown that the asymptotic bias of v_{JS} is given by

$$\begin{aligned}\mathrm{Bias}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \frac{N^2}{E_{pq}(r)}\left(1 - E_{pq}\left(\frac{r}{n}\right)\right) \\ &\quad S_y^2\left(\frac{1}{\mathrm{CV}(y)^2} + \rho_{xy}^2\right) \geq 0. \quad (3.6)\end{aligned}$$

From (3.6), it follows that v_{JS} always overestimates the true variance in the case of the simple linear regression adjusted estimator (1.6). The bias (3.6) increases if (i) the expected response rate decreases; (ii) ρ_{xy}^2 increases; or (iii) $\mathrm{CV}(y)$ decreases.

3.2 Stratified simple random sampling: Weighting classes are identical to strata

In this section, we assume that the weighting classes coincide with the original design strata. This situation is not uncommon in practice, especially in business surveys. If the strata are such that the units within stratum have approximately equal response propensities (*i.e.*, uniform response within stratum), expressions for the bias of v_{JS} are readily obtained from expressions (3.2), (3.4) and (3.6).

For the ratio adjusted estimator, expression (3.2) can be readily extended to the case of stratified simple random sampling to obtain

$$\begin{aligned}\mathrm{Bias}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \sum_{h=1}^{L}\frac{N_h^2}{E_{pq}(r_h)}\left(1 - E_{pq}\left(\frac{r_h}{n_h}\right)\right) \\ &\quad S_{yh}^2\left(\frac{1}{\mathrm{CV}_h(y)^2} + 2\frac{\mathrm{CV}_h(x)}{\mathrm{CV}_h(y)}\rho_{hxy} - \frac{\mathrm{CV}_h(x)^2}{\mathrm{CV}_h(y)^2}\right), \quad (3.7)\end{aligned}$$

where the quantities r_h, $\mathrm{CV}_h(x)$, $\mathrm{CV}_h(y)$, S_{yh}^2 and ρ_{hxy} correspond to r, $\mathrm{CV}(x)$, $\mathrm{CV}(y)$, S_y^2 and ρ_{xy} computed in each stratum.

For the count adjusted estimator, expression (3.4) can be readily extended to the case of stratified simple random sampling to obtain

$$\begin{aligned}\mathrm{Bias}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \sum_{h=1}^{L}\frac{N_h^2}{E_{pq}(r_h)}\left(1 - E_{pq}\left(\frac{r_h}{n_h}\right)\right)\frac{S_{yh}^2}{\mathrm{CV}_h(y)^2}. \quad (3.8)\end{aligned}$$

Finally, for the simple linear regression adjusted estimator, expression (3.6) can be readily extended to the case of stratified simple random sampling to obtain

$$\begin{aligned}\mathrm{Bias}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \sum_{h=1}^{L}\frac{N_h^2}{E_{pq}(r_h)}\left(1 - E_{pq}\left(\frac{r_h}{n_h}\right)\right) \\ &\quad S_{yh}^2\left(\frac{1}{\mathrm{CV}_h(y)^2} + \rho_{hxy}^2\right). \quad (3.9)\end{aligned}$$

From the expressions (3.7)-(3.9), it follows that the use of the shortcut jackknife variance estimator requires some caution. Indeed, even if the bias of the shortcut jackknife variance estimator is small in each stratum, they might sum up to a considerable bias at the population level if the biases are in the same direction.

4. Simulation study

A simulation study was performed to compare the statistical properties of the shortcut and the full jackknife variance estimators under varying conditions. Five different stratified populations of 30,000 units each with two variables were generated. First, the *x*-values were generated from a Gamma distribution with parameters α and λ. Then given the *x*-values, the *y*-values were generated according to the following model:

$$y_{hi} = \beta_0 + \beta_1 x_{hi} + \varepsilon_{hi},$$

where $\varepsilon_{hi} \sim N(0, \sigma_{\varepsilon h}^2)$. The variance and $\sigma_{\varepsilon h}^2$ was set such that the coefficient of correlation (denoted ρ_{xy}) between x_{hi} and y_{hi} is equal to 0.7 in all the populations. Each population was stratified into three strata, each with 10,000 units. The parameters of the simulated populations appear in Table 2.

Population 1 fits the ratio model very well with an intercept of zero in all strata. Population 2 has a non-negligible intercept term in all three strata. Population 3 is a mix of populations 1 and 2, where the ratio model fits well for strata 2 and 3 but not for stratum 1. Population 4 is

similar to population 1 except units in strata 1 and 2 have a 70% chance of reporting a zero. This population is intended to mimic the situation of the Annual Capital Expenditures Survey (ACES) of the U.S. Census Bureau, which provided the motivation for this research. The ACES employs a shortcut jackknife variance estimator that, empirically, has been shown to be close to the full jackknife variance estimates. Its population is characterized with many zeros for capital expenditures in the majority of sampled small and medium businesses, with the majority of the reported expenditures being provided by large businesses. Population 5 was generated to show that the shortcut estimator for the ratio adjusted estimator can actually have a negative bias when the ratio model is misspecified (demonstrated in expression (3.3) for a simple random sample). For this population, the intercept term is highly significant in all strata.

Table 2
Population parameters

Population	β_0 (Within Stratum)			β_1 (Within Stratum)			α	λ	CV(x)	CV(y)
	1	2	3	1	2	3				
1	0	0	0	2	4	6	4	5	50%	76%
2	120	240	360	2	4	6	4	5	50%	44%
3	120	0	0	2	4	6	4	5	50%	51%
4	0	0	0	2	4	6	4	5	50%	134%
5	50	200	300	0.5	1	2	4	5	200%	63%

From each population, 5,000 stratified simple random samples of size 300 (100 units per stratum) were drawn. In each sample, nonresponse was generated using a uniform response mechanism within each stratum with probabilities of response equal to 60% in stratum 1, 70% in stratum 2 and 90% in stratum 3. This response pattern is not uncommon in business surveys where more follow-up is performed for the medium and large size units (strata 2 and 3).

In each sample, both the count adjusted and the ratio adjusted estimators, given respectively by (1.4) and (1.5), were calculated using the strata as weighting classes. The variance of the point estimators was estimated by v_{JF} and v_{JS}, given respectively by (2.2) and (2.3). As a measure of the bias of a variance estimator v, we used the Monte Carlo percent relative bias given by

$$\mathrm{RB}_{\mathrm{MC}}(v) = \frac{1}{5{,}000}\sum_{t=1}^{5{,}000}\frac{v^{(t)} - \mathrm{MSE}_{\mathrm{MC}}(\hat{Y}_{\mathrm{CAL}})}{\mathrm{MSE}_{\mathrm{MC}}(\hat{Y}_{\mathrm{CAL}})}\times 100,$$

where $v^{(t)}$ is the variance estimate obtained from the t^{th} sample, and $\mathrm{MSE}_{\mathrm{MC}}(\hat{Y}_{\mathrm{CAL}})$ is the Monte Carlo Mean Squared Error (MSE) defined by

$$\mathrm{MSE}_{\mathrm{MC}}(\hat{Y}_{\mathrm{CAL}}) = \frac{1}{50{,}000}\sum_{t=1}^{50{,}000}(\hat{Y}_{\mathrm{CAL}}^{(t)} - Y)^2,$$

where $\hat{Y}_{\mathrm{CAL}}^{(t)}$ is the (ratio or count adjusted) estimate of Y for the t^{th} sample. Table 3 shows the Monte Carlo percent relative bias for both the count adjusted and the ratio adjusted estimators.

Table 3
Monte Carlo percent relative bias for the shortcut and full jackknife variance estimators

Population	Count adjusted estimator		Ratio adjusted estimator	
	$\mathrm{RB}_{\mathrm{MC}}(v_{JS})$	$\mathrm{RB}_{\mathrm{MC}}(v_{JF})$	$\mathrm{RB}_{\mathrm{MC}}(v_{JS})$	$\mathrm{RB}_{\mathrm{MC}}(v_{JF})$
1	57.3%	1.1%	80.5%	-0.3%
2	877.1%	0.4%	364.7%	0.5%
3	220.7%	0.6%	185.9%	-0.2%
4	21.6%	0.6%	29.1%	1.4%
5	266.4%	0.2%	-67.2%	5.0%

As expected, the shortcut estimator overestimates the Monte Carlo MSE for the count adjusted estimator for all populations. The overestimation varies from approximately 20% in population 4 to over 800% in population 2. From expression (3.8), we see that the bias of v_{JS} depends on the response rate and $\bar{y}_h^2$. Population 2 has a large intercept term which increases $\mathrm{CV}_h(y)$ in all strata, which in turn increases the bias of v_{JS}. Population 3 is similar to population 2 except only the first stratum has a large intercept term. As expected, the bias of v_{JS} in this population is between those of populations 1 and 2. Population 4 is the one generated to mimic the ACES population with some units' values replaced by zero in strata 1 and 2. The Monte Carlo relative bias of 21.6% is, for the most part, coming from the third stratum where no units have been replaced with zero (this can be seen using expression (3.8)). In comparison, for all five populations the full jackknife variance estimator is tracking the Monte Carlo MSE very well with absolute relative biases less than 1.1%.

Turning to the ratio adjusted estimator, we see that the full jackknife variance estimator again tracks the Monte Carlo MSE relatively well for all populations with absolute relative biases less than 5%. The shortcut estimator, on the other hand, has relative biases varying from -67% to 364%. Looking at expression (3.7), we see that for a fixed response rate the bias depends on the $\mathrm{CV}_h(y)$, $\mathrm{CV}_h(x)$ and ρ_{hxy}. Due to the large intercept terms in the second population, $\bar{y}_h$ are large and the corresponding $\mathrm{CV}_h(y)$ are smaller than in the other populations. Thus, the last term in expression (3.7) is quite large and the resulting relative bias of v_{JS} is also large. This is also seen for population 3 except to a lesser extent since only the first stratum has an intercept term. The opposite effect is seen in population 4, where the introduction of zeros has significantly increased $\mathrm{CV}_h(y)$ which has in turn reduced the Monte Carle percent relative bias of the shortcut estimator.

Additional simulations were performed using the some of the populations described in Table 2 but with varying response rates. The results are not presented here as they were as expected. That is, the bias of the shortcut estimator deceased as the response rate increased (with all the other parameters remaining fixed). The full jackknife estimator continued to track the Monte Carlo MSE very well.

5. Conclusion

In this paper, we evaluated both theoretically and empirically a shortcut jackknife variance estimator that does not re-calculate the nonresponse adjustment factors within each jackknife replicate, specifically considering three different nonresponse weighting adjustment procedures. We showed in the context of stratified simple random sampling that the shortcut jackknife variance estimator tends to overestimate the true variance of the estimators. In the context of the ratio adjustment procedure, however, the shortcut jackknife variance estimator may underestimate the true variance if the ratio model is not appropriate for the data at hand.

One justification for the use of a shortcut procedure in a replicate variance estimation method is to save time and computing resources. If these are truly issues and the program has consistently high unit response rates in all weighting cells, then while there are clearly theoretical advantages to replicating the weight adjustment procedure, there may be little or no practical advantage. Having said that, the conditions for "practical" equivalence between the full and shortcut procedure variance estimators are extremely restrictive, and we have demonstrated that small changes in underlying data conditions can easily violate these conditions. If computational concerns with a full jackknife are truly an issue, then the authors recommend the linearization jackknife variance estimation approach which has the same asymptotic properties as the full jackknife, but is computationally quick and computer overhead "free" (in terms of replicate storage). See Thompson and Yung (2006) for expressions for the linearization jackknife variance estimator for both the count and ratio adjusted estimators. Given these viable alternatives, we recommend against the use of a shortcut procedure variance estimator.

Acknowledgements

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau. The authors would like to thank the Associate Editor, two anonymous referees, Samson Adeshiyan, Patrick Cantwell, Carol Caldwell, Michael Hidiroglou, Rita Petroni, Mark Sands, and Jun Shao for their useful comments on earlier versions of this paper. Work of David Haziza was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

References

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society B*, 67, 445-458.

Brick, M.J., and Montaquila, J.M. (2009). Nonresponse and weighting. In the *Handbook of Statistics*, *Sample Surveys*: *Design*, *Methods and Applications*, (Eds., C.R. Rao and D. Pfeffermann), 29A, 163-185.

Da Silva, D.N., and Opsomer, J.D. (2006). A kernel smoothing method to adjust for unit nonresponse in sample surveys. *Canadian Journal of Statistics*, 34, 563-579.

Deville, J.-C. (2002). La correction de la non-réponse par calage généralisé. *Actes des Journées de Méthodologie Statistique*, Insee.

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

Eltinge, J.L., and Yansaneh, I.S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology*, 23, 33-40.

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

Kalton, G., and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.

Kim, J.K., and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35, 501-514.

Kott, P. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 133-142.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

Ozcoskun, L., Thompson, K.J. and Williams, Q. (2005). Investigation of balanced repeated replication (BRR) variance estimation for the Survey of Residential Alterations and Repairs (SORAR). *Proceedings of the Federal Committee on Statistical Methodology*, Office of Management and Budget.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.

Sautory, O. (2003). CALMAR 2: A new version of the CALMAR calibration adjustment program. *Proceedings*: *Symposium 2003*, *Challenges in Survey Taking for the Next Decade*, Ottawa, Canada.

Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 93, 254-265.

Thompson, K.J. (2005). An empirical investigation into the effects of replicate reweighting on variance estimates for the annual capital expenditures survey. *Proceedings of the Federal Committee on Statistical Methodology*, Office of Management and Budget.

Thompson, K.J., and Yung, W. (2006). To Replicate (A weight adjustment procedure) or not to replicate? An analysis of the variance estimation effects of a shortcut procedure using the stratified jackknife. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 3772-3779.

Valliant, R. (2004). The effect of multiple weighting steps on variance estimation. *Journal of Official Statistics*, 20, 1-18.

Yung, W., and Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, 95, 903-915.