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

Katherine Jenny Thompson[1] and Wesley Yung
U.S. Census Bureau
Statistics Canada

Abstract

Many surveys employ weight adjustment procedures to compensate for unit non-response. This paper frames such procedures as second-phase sampling adjustments (where the second phase refers to a Bernoulli sample of respondents from sampled units). We compare the effect of replicating the unit non-response weight adjustment procedure on stratified jackknife variance estimates to those obtained using the corresponding "shortcut procedure" variance estimates on the same data. Two weight adjustment procedures are examined.

Key words: unit non-response, Bernoulli sample, linearization variance estimator

1 Introduction

Many surveys employ weight adjustment procedures to compensate for unit non-response. Such weight adjustment procedures can be viewed as bias corrections accounting for underestimation or as second-phase sampling adjustments (where the second phase refers to a random sample of respondents from the initial sample). The latter view couches the issue of unit non-response as a variance estimation problem. If we assume a uniform response mechanism in each weighting class *p*, then the second phase sample can be viewed as a Bernoulli sample with a **random** sample size of respondents -- each with response probability π^p -- from the first-phase sampled units in each weighting class (Särndal, C., Swensson, B., and Wretman, J., 1992, pp. 62-65 and Kott, 1994). This two-phase sample interpretation is equally valid with a missing-at-random (MAR) model, which assumes probability of unit non-response depends on an auxiliary variable or set of auxiliary variables, not the characteristic(s) of interest. In the two-phase framework, we compare the effect of replicating the unit non-response weight adjustment procedure on stratified jackknife variance estimates to those obtained using the corresponding "shortcut procedure" stratified jackknife variance estimates on the same data.

Shortcut procedure variance estimates use replicate weights constructed from the **full-sample** unit non-response adjusted weights instead of repeating the weighting procedure in each replicate. This approach was originally presented (not recommended) in Wolter (1985, pp. 82-84). Note that the shortcut procedure variance estimator is **not** a naïve variance estimator, which treats imputed values as though they were reported values and which has been repeatedly shown to underestimate the true variance. The naïve variance estimator replaces the missing item responses with imputed values, yielding a dataset with no (visible) item non-response, and the replicate values do not contain any missing values. Our replicate assignment procedures assign **sample** units to replicates. Consequently, these replicate estimates contain both responding and non-responding units, regardless of the replicate reweighting procedure.

We consider two different weight adjustment procedures, both of which are adjustment-to-sample models described in Kalton and Flores-Cervantes (2003), i.e., all sampling weights in a weighting class *p* are multiplied by a factor derived from data corresponding to sample units. The first procedure (the **ratio adjustment** procedure) controls the respondent estimates to full-sample estimates of an auxiliary variable. The second procedure (the **count adjustment)** yields estimates that are controlled to sample counts; when weighted counts (population estimates) are used, then the count adjustment procedure is the "quasi-randomization" estimator (Oh and Scheuren, 1983); when weighting cells are sample strata, then the count procedure is the inverse unweighted response rate adjustment recommended by Vartivarian and Little (2002).

2 The Jackknife Linearization Estimator Given a Two-Phase Sample with Bernoulli Sampling at the Second Phase

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expression on statistical and methodological issues are those of the authors and do not necessarily reflect those of the U.S. Census Bureau or Statistics Canada.

Suppose we have a stratified SRS-WOR design with n_h sampled units in the h^{th} stratum (h = 1, 2, …, L) . Let y_{hi} (i=1, …, n_h) be the value of the variable of interest for the i^{th} unit in the h^{th} stratum (obtained from survey respondents) and let x_{hi} be the corresponding auxiliary variable (available for all **sampled** units). For simplicity, assume that sampling fractions in all strata are negligible, and that there are at least two sampled units in all strata. In addition, we assume that there exist P disjoint weighting classes, which can cut across the design strata. With unit non-response, we divide the sampled units, s, into respondents, s_r, and non-respondents, s_{nr}. The following derivations assume that each weighting class contains at least two respondents, although in practice, surveys generally collapse weighting cells with insufficient respondents.

2.1 Ratio Adjustment Procedure

Since x_{hi} is known for all sampled units in s, it can be used in the unit non-response weight adjustment. The non-response weight adjustment for the p^{th} weighting class is

$$d_1^P = \frac{\sum_s w_{hi}\, x_{hi}\, \delta_{hi}^p}{\sum_s w_{hi}\, x_{hi}\, \delta_{hi}^p I_{hi}} = \frac{\hat{X}^p}{\hat{X}_r^p}$$

where $\sum_s$ denotes the summation over all sampled units, I_{hi} is the response indicator variable (I_{hi} = 1 if (hi) $\in s_r$ and =0 otherwise), and δ_{hi}^p is the weighting class indicator variable ($\delta_{hi}^p = 1$ if (hi) $\in$ weighting class p and =0 otherwise), and w_{hi} is the design weight (= N_h/n_h for a stratified SRS-WOR sample). The non-response adjusted weight in each weighting cell p is then given by $\tilde{w}_{hi}^p = d_1^p w_{hi}$ and the non-response adjusted estimator is

$$\hat{Y} = \sum_p \sum_s \tilde{w}_{hi}^p y_{hi} I_{hi} \delta_{hi}^p = \sum_p \frac{\hat{X}^p}{\hat{X}_r^p} \hat{Y}_r^p,$$

where $\hat{Y}_r^p$ is the respondent-based estimate of characteristic Y. This expression is exactly equal to the ratio imputation estimator presented in Deville and Särndal (1994).

We obtain the variance of $\hat{Y}$ by decomposing it within weighting cell p (see Cochran, 1977, p. 343)

$$\hat{Y}^p - Y^p = \left(\frac{\hat{Y}_r^P}{\hat{X}_r^P} X^p - Y^p \right) + \frac{\hat{Y}_r^P}{\hat{X}_r^P} (\hat{X}^p - X^p)$$

$$= \frac{X^p}{\hat{X}_r^p} \left[\hat{Y}_r^p - \left(\frac{Y^p}{X^p} \right) \hat{X}_r^p \right] + \left[\frac{\hat{Y}_r^p}{\hat{X}_r^p} (\hat{X}^p - X^p) \right] = B_p + A_p$$

and thus,

$$\hat{Y} - Y = \sum_p \left(B_p + A_p \right) = \mathbf{B} + \mathbf{A}.$$

The first term (**B**) corresponds to the ordinary error of a ratio estimate; the second term (**A**) is the error contribution of the auxiliary variable from the 1st to 2nd phase sample. Consequently, the mean squared error of the non-response adjusted estimator is given by

$$E(\hat{Y} - Y)^2 = E\left[\sum_p B_p + A_p \right]^2 = E_1 E_2 \left[\sum_p \sum_{s^p} \left(B_p + A_p \right) \right]^2$$

where E_1 is the conditional expectation at the first stage of sampling (stratified SRS-WOR) and E_2 is the conditional expectation at the second stage of sampling (Bernoulli). It is difficult to derive a closed-form expression for this statistic, since sampling strata may not be entirely confined within weighting cells. The B_p component contains two sources of variability from the respondent sample: the variable of interest and the auxiliary variable, x.

If the weighting cells consist entirely of one or more strata, then the MSE expression can be further simplified. In this case, the weighting cell estimates are independent, and the expectations can be taken within the weighting cells. Weighting cells that comprise more than one stratum will yield second phase combined ratio estimates; weighting cells that equal strata will yield second phase separate ratio estimates. In the latter case, $MSE(\hat{Y}) = \sum_p MSE(\hat{Y}^p)$. Under an ignorable response mechanism, the MSE is minimized when the all units within the same weighing cell have (1) the same probability of response (response propensity) or (2) the same mean for the characteristic of interest. Satisfaction of <u>both</u> conditions (i.e., same response propensity and mean within weighting cell) results in substantial MSE reduction in both the bias and variance terms when the count adjustment procedure is applied to correct for unit non-response, in a model-based or model-assisted perspective (Little and Vartivarian, 2005).

Business surveys often use the sampling strata as weighting classes. When the stratification procedure depends on some sampling unit measure of size, for example, then the model assumption is that all units within a stratum have the same probability of response (e.g., larger units are more likely to respond or be recontacted for response than smaller units). The second model is assumed when all survey variables of interest are highly correlated with auxiliary stratification variables.

As shown below, the fully reweighted stratified jackknife variance estimator implicitly simulates all error components of the two-phase sample MSE$\left(\hat{Y}\right)$. To see this, we first define the stratified jackknife weights when the j[th] unit in the g[th] stratum is removed as

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

where $n_g \geq 2$ for all strata g. The non-response adjustment when the (gj)[th] unit is deleted is then

$$d_{1(gj)}^{p} = \frac{\sum_s w_{hi(gj)}\, x_{hi}\, \delta_{hi}^{p}}{\sum_s w_{hi(gj)}\, x_{hi}\, \delta_{hi}^{p} I_{hi}} = \frac{\hat{X}_{(gj)}^{p}}{\hat{X}_{r(gj)}^{p}}$$

The non-response adjusted jackknife weight is $\tilde{w}_{hi(gj)}^{p} = d_{1(gj)}^{p}\, w_{hi(gj)}$. The replicate estimate when the $(gj)^{\text{th}}$ unit is deleted is then

$$\hat{Y}_{(gj)}^{(1)} = \sum_p \sum_s \tilde{w}_{hi(gj)}^{p}\, y_{hi}\, I_{hi}\, \delta_{hi}^{p} = \sum_p \frac{\hat{X}_{(gj)}^{p}}{\hat{X}_{r(gj)}^{p}} \hat{Y}_{r(gj)}^{p},$$

and the jackknife variance estimator of $\hat{Y}$ is

$$v_J\left(\hat{Y}\right) = \sum_g \frac{n_g - 1}{n_g} \sum_j \left(\hat{Y}_{(gj)}^{(1)} - \hat{Y}\right)^2$$

$$= \sum_g \frac{n_g - 1}{n_g} \sum_j \left[\sum_p \left(\frac{\hat{X}_{(gj)}^{p}}{\hat{X}_{r(gj)}^{p}} \hat{Y}_{r(gj)}^{p} - \frac{\hat{X}^{p}}{\hat{X}_{r}^{p}} \hat{Y}_{r}^{p} \right) \right]^2 .$$

To obtain a linearization type variance estimator, note that from a Taylor Series expansion we have (ignoring second order and higher derivatives)

$$\frac{\hat{X}_{(gj)}^{p}}{\hat{X}_{r(gj)}^{p}} \hat{Y}_{r(gj)}^{p} - \frac{\hat{X}^{p}}{\hat{X}_{r}^{p}} \hat{Y}_{r}^{p}$$

$$\approx \frac{\hat{X}^{p}}{\hat{X}_{r}^{p}} \left[\left(\hat{Y}_{r(gj)}^{p} - \hat{Y}_{r}^{p} \right) - \frac{\hat{Y}_{r}^{p}}{\hat{X}_{r}^{p}} \left(\hat{X}_{r(gj)}^{p} - \hat{X}_{r}^{p} \right) \right] + \frac{\hat{Y}_{r}^{p}}{\hat{X}_{r}^{p}} \left(\hat{X}_{(gj)}^{p} - \hat{X}^{p} \right)$$

$$= \hat{\mathbf{B}}_{\mathbf{p}} + \hat{\mathbf{A}}_{\mathbf{p}} = (\hat{\mathbf{B}}_{\mathbf{pY}} - \hat{\mathbf{B}}_{\mathbf{pX}}) + \hat{\mathbf{A}}_{\mathbf{p}}$$

which yields the reweighted jackknife linearization variance estimator

$$v_{JL}^{R}(\hat{Y}) = \sum_g \frac{n_g}{n_g - 1} \sum_j \left[\sum_p \left(\bar{e}_{rg}^{p} - \tilde{w}_{gj} e_{gj} I_{gj} \delta_{gj}^{p} \right) + \sum_p \frac{\hat{Y}_{r}^{p}}{\hat{X}_{r}^{p}} \left(\bar{x}_{g}^{p} - \tilde{w}_{gj} x_{gj} I_{gj} \delta_{gj}^{p} \right) \right]^2$$

where $e_{gj} = y_{gj} - (\hat{Y}_r^{p} / \hat{X}_r^{p}) x_{gj}$, $\tilde{w}_{gj} = (\hat{X}^{p} / \hat{X}_r^{p}) w_{gj}$, $\bar{x}_g^{p} = (1/n_g) \sum_j w_{gj} x_{gj} \delta_{gj}^{p}$, and

$$\bar{e}_g^{p} = (1/n_g) \sum_j w_{gj} e_{gj} I_{gj} \delta_{gj}^{p} .$$

(Derivation available upon demand from the authors). It is not a coincidence that the bracketed term of $v_{JL}^{R}(\hat{Y})$ has exactly the same form as the MSE expression of a two-phase ratio estimator. Moreover, the first term in this linearization estimate – the $\hat{\mathbf{B}}_{\text{p}}$ term – contains two explicit error components: one for the characteristic of interest and one for the auxiliary variable (both within the response set). Our derived linearization estimator is slightly different from the comparable estimator derived in both Rao and Sitter (1995) and Shao and Steele (1999), following from different assumptions about the second phase sample of respondents.

Turning to the shortcut jackknife variance estimator, where the non-response adjustment is **not** recalculated in each replicate, we express the non-response adjusted jackknife weight as $\tilde{w}_{hi(gj)}^{p} = d_1^{p}\, w_{hi(gj)}$. The replicate estimate is given as

$$\hat{Y}_{(gj)}^{(2)} = \sum_p \sum_s \tilde{w}_{hi(gj)}^{p}\, y_{hi}\, I_{hi}\, \delta_{hi}^{p} = \sum_p \frac{\hat{X}^{p}}{\hat{X}_{r}^{p}} \hat{Y}_{r(gj)}^{p}$$

and the shortcut jackknife variance estimator is

$$v_{JL}^{S} = \sum_g \frac{n_g - 1}{n_g} \sum_j \left[\sum_p \frac{\hat{X}^{p}}{\hat{X}_{r}^{p}} \left(\hat{Y}_{r(gj)}^{p} - \hat{Y}_{r}^{p} \right) \right]^2$$

$$= \sum_g \frac{n_g - 1}{n_g} \sum_j \left[\sum_p \hat{B}_{pX} \right]^2 .$$

Comparing the fully reweighted and shortcut linearization estimators, we see that the difference between the reweighted and the shortcut jackknife variance estimators is the term

$$\sum_p \frac{\hat{Y}_{r}^{p}}{\hat{X}_{r}^{p}} \left[\left(\hat{X}_{(gj)}^{p} - \hat{X}^{p} \right) - \frac{\hat{X}^{p}}{\hat{X}_{r}^{p}} \left(\hat{X}_{r(gj)}^{p} - \hat{X}_{r}^{p} \right) \right]$$

$$= \sum_p \left(\hat{\mathbf{A}}_{\text{p}} - \hat{\mathbf{B}}_{\text{pX}} \right)$$

$$= \sum_p \frac{\hat{Y}_{r}^{p}}{\hat{X}_{r}^{p}} \left(\hat{X}_{(gj)}^{p} - \frac{\hat{X}^{p}}{\hat{X}_{r}^{p}} \hat{X}_{r(gj)}^{p} \right) = \sum_p \hat{\mathbf{A}}''_{p} = \hat{\mathbf{A}}''$$

within the squared term. The shortcut variance estimator is a **partial** approximation of the ordinary ratio estimate error: it does not include the error contribution of the auxiliary variable to the ratio estimate or its covariance with the characteristic. Thus, the $\hat{\mathbf{A}}''$ term approximates the difference between the error contribution of the auxiliary variable from 1st (sample units) to 2nd phase (respondent units) and the error contribution from the auxiliary variable in the ordinary ratio estimate, both missing from the shortcut jackknife variance estimator. In a given weighting cell, the $\hat{\mathbf{A}}''_p$ term should be small when the unit response rate is high (say greater than 70%) since $(\hat{X}^p / \hat{X}_r^p)\hat{X}_{r(gj)}^p$ is a non-response adjusted estimator of $\hat{X}^p$ when the $(gj)^{\text{th}}$ term is deleted. Finally

$$v_{JL}^R(\hat{Y}) - v_{JL}^S(\hat{Y}) = \sum_g \frac{n_g - 1}{n_g} \sum_j \left(\left(\hat{\mathbf{A}}''\right)^2 + \mathbf{2}\hat{\mathbf{A}}''\hat{\mathbf{B}}_Y \right)$$

where

$$\begin{aligned}\hat{\mathbf{B}}_Y &= \sum_p \frac{\hat{X}^p}{\hat{X}_r^p}\left(\hat{Y}_{r(gj)}^p - \hat{Y}_r^p\right) \\ &= \sum_p \frac{n_g}{n_g - 1}\left(\frac{1}{n_g}\sum_i \tilde{w}_{gi}\, y_{gi}\, I_{gi}\, \delta_{gi}^p - \tilde{w}_{gj}\, y_{gj}\, I_{gj}\, \delta_{gj}^p\right).\end{aligned}$$

With a stratified SRS sampling plan where the weighting classes correspond to strata, we can show that $\hat{Y}_r^p = (N_p r_p / n_p)\bar{y}_{pr}$, $\hat{X}_r^p = (N_p r_p / n_p)\bar{x}_{pr}$, $(\hat{Y}_r^p / \hat{X}_r^p) = (\bar{y}_{pr} / \bar{x}_{pr})$ and $\hat{X}^p / \hat{X}_r^p = (n_p \bar{x}_p / r_p \bar{x}_{pr})$ where $r_p = \sum_{s \in p} I_{hi}$, $\bar{y}_{pr} = (r_p)^{-1}\sum_i y_{pi} I_{pi}$, $\bar{x}_{pr} = (r_p)^{-1}\sum_i x_{pi} I_{pi}$ and $\bar{x}_p = (n_p)^{-1}\sum_i x_{pi}$. Using these expressions, we have

$$\hat{A}''_p = \frac{\bar{y}_{pr}}{\bar{x}_{pr}} \frac{n_p}{n_p - 1} \frac{N_p}{n_p} x_{pj}\left(\frac{n_p \bar{x}_p}{r_p \bar{x}_{pr}} I_{pj} - 1\right) \text{ and}$$

$$\begin{aligned}\left(\hat{\mathbf{A}}''\right)^2 &= \sum_g \frac{n_g - 1}{n_g} \sum_j \hat{A}''^2_g \\ &= \sum_g \frac{n_g}{n_g - 1}\left(\frac{\bar{y}_{gr}}{\bar{x}_{gr}}\right)^2 \left(\frac{N_g}{n_g}\right)^2 \sum_j x_{gj}^2 \left(\frac{n_g \bar{x}_g}{r_g \bar{x}_{gr}} I_{gj} - 1\right)^2.\end{aligned}$$

Turning to the second error component term, we have

$$\hat{B}_{pY} = \frac{n_p \bar{x}_p}{r_p \bar{x}_{pr}} \frac{n_p}{n_p - 1} \frac{N_p}{n_p}\left(\frac{r_p}{n_p}\bar{y}_{pr} - y_{pj} I_{pj}\right)$$

and

$$2\hat{\mathbf{A}}''\hat{\mathbf{B}}_Y = -2\sum_g \frac{n_g}{n_g - 1} \frac{\bar{y}_{gr}}{\bar{x}_{gr}} \frac{n_g \bar{x}_g}{r_g \bar{x}_{gr}} \left(\frac{N_g}{n_g}\right)^2 \sum_j x_{gj} y_{gj} I_{gj}\left(\frac{n_g \bar{x}_g}{r_g \bar{x}_{gr}} - 1\right)$$

When $\bar{x}_g / \bar{x}_{gr} \approx 1$, the weighing cell's contribution to the $\mathbf{2}\hat{\mathbf{A}}''_p\hat{\mathbf{B}}_{pY}$ term is positive or zero. When this condition holds for all strata – true under a MAR response mechanism -- then the cumulative contribution of $\mathbf{2}\hat{\mathbf{A}}''_p\hat{\mathbf{B}}_{pY}$ is negative or zero, reducing the estimated variance over the shortcut procedure. We can interpret the $\left(\hat{\mathbf{A}}''\right)^2$ term as the estimated cumulative MSE contribution of the auxiliary variable, and the $\mathbf{2}\hat{\mathbf{A}}''\hat{\mathbf{B}}_Y$ term as the cumulative covariance contribution. If the auxiliary variable has high positive correlation with the characteristic of interest and low relative-variance, then the shortcut procedure will **overestimate** the variance.

2.2 Count Adjustment Procedure

A commonly used unit non-response adjustment procedure controls the estimated respondent population to the population total estimates within weighting cell, setting x_{hi} =1 for all sample units. The non-response adjustment for the p^{th} weighting class is

$$d_2^p = \frac{\sum_s w_{hi}\, \delta_{hi}^p}{\sum_s w_{hi}\, I_{hi}\, \delta_{hi}^p} = \hat{N}^p / \hat{N}_r^p,$$

the non-response adjusted weight is $\tilde{w}_{hi}^p = d_2^p\, w_{hi}$, and the estimator of the total is

$$\hat{Y} = \sum_p \sum_s \tilde{w}_{hi}^p\, y_{hi}\, I_{hi}\, \delta_{hi}^p = \sum_p \frac{\hat{N}^p}{\hat{N}_r^p}\hat{Y}_r^p.$$

The reweighted jackknife variance estimator is

$$v_J^R\left(\hat{Y}\right) = \sum_g \frac{n_g - 1}{n_g}\sum_j\left(\sum_p\left[\frac{\hat{N}_{(gj)}^p}{\hat{N}_{r(gj)}^p}\hat{Y}_{r(gj)}^p - \frac{\hat{N}^p}{\hat{N}_r^p}\hat{Y}_r^p\right]\right)^2.$$

If the weighting classes correspond to strata, then $\hat{Y} = \sum_p \sum_s \frac{N_p}{n_p}\frac{n_p}{r_p} y_{hi} I_{hi} = \sum_p N_p \bar{y}_{pr}$, the Horvitz-Thompson estimator based on respondents within strata. The count adjustment procedure is designed to account for both estimator bias and variance given a MAR response mechanism by judiciously forming weighting cells (Little and Vartivarian, 2005). This estimator is mathematically equivalent to the mean imputation estimator presented in Deville and Särndal (1994).

Using the jackknife replicate weights from Section 2.1, the count procedure non-response adjusted replicate weights when the jth unit in the gth stratum is removed are

$$\tilde{w}_{hi(gj)} = \begin{cases} 0 & \text{if (hi)} = \text{(gj)} \\ \dfrac{N_g}{n_g}\dfrac{n_g}{n_g-1}\dfrac{n_g-1}{r_{(gj)}} = \dfrac{N_g}{r_{(gj)}} & \text{if h} = \text{g, i} \neq \text{j} \\ w_{hi} & \text{otherwise} \end{cases}$$

where $r_{(gj)}$ are the number of respondent units in the jackknife replicate. These replicate weights have exactly the same form as the full-sample adjusted weights, and the replicate estimates are also Horvitz-Thompson estimates based on respondents.

The reweighted jackknife linearization variance estimator is given as

$$v_{JL}^R(\hat{Y}) = \sum_g \frac{n_g-1}{n_g}\sum_j\left[\sum_p \frac{\hat{N}^p}{\hat{N}_r^p}\left\{\frac{1}{n_g}\sum_i w_{gi}I_{gi}\delta_{gi}^p\left(y_{gi}-\frac{\hat{Y}_r^p}{\hat{N}_r^p}\right) - w_{gj}I_{gj}\delta_{gj}^p\left(y_{gj}-\frac{\hat{Y}_r^p}{\hat{N}_r^p}\right)\right\} + \sum_p \frac{\hat{Y}_r^p}{\hat{N}_r^p}\left(\frac{1}{n_g}\sum_i w_{gi}\delta_{gi}^p - w_{gj}\delta_{gj}^p\right)\right]^2$$

(Derivation available upon demand from the authors), and the shortcut jackknife linearization variance estimator is given by

$$v_{JL}^S(\hat{Y}_{ACE2}) = \sum_g \frac{n_g-1}{n_g}\sum_j\left[\sum_p \frac{\hat{N}^p}{\hat{N}_r^p}\left(\frac{1}{n_g}\sum_i w_{gi}y_{gi}I_{gi}\delta_{gi}^p - w_{gj}y_{gj}I_{gj}\delta_{gj}^p\right)\right]^2.$$

The difference between $v_J^R(\hat{Y})$ and $v_J^S(\hat{Y})$ is approximately

$$\frac{\hat{Y}_r^p}{\hat{N}_r^p}\left(\left(\hat{N}_{(gj)}^p - \hat{N}^p\right) - \frac{\hat{N}^p}{\hat{N}_r^p}\left(\hat{N}_{r(gj)}^p - \hat{N}_r^p\right)\right)$$

$$= \hat{\mathbf{A}}_{\mathbf{p}} - \hat{\mathbf{B}}_{\mathbf{pX}} = \hat{\mathbf{A}}_{\mathbf{p}}''.$$

When the weighting cells correspond to the sample strata, then $\hat{N}_r^p = (N_p r_p / n_p)$, $(\hat{Y}_r^p / \hat{N}_r^p) = \bar{y}_{pr}$ and $\hat{N}^p / \hat{N}_r^p = n_p / r_p$. Substituting we get,

$$\hat{A}_p'' = \bar{y}_{pr}\frac{n_p}{n_p-1}\frac{N_p}{n_p}\left(\frac{n_p}{r_p}I_{pj}-1\right)$$

and

$$\left(\hat{\mathbf{A}}''\right)^2 = \sum_g \frac{n_g-1}{n_g}\sum_j\left(\hat{\mathbf{A}}''\right)^2 = \sum_g \bar{y}_{gr}^2\frac{n_g}{n_g-1}\frac{N_g^2}{n_g r_g}(n_g - r_g).$$

Turning to the remaining error term, we have

$$\hat{B}_{pY} = \frac{n_p}{r_p}\frac{n_p}{n_p-1}\frac{N_p}{n_p}\left(\frac{r_p}{n_p}\bar{y}_{pr} - y_{pj}I_{pj}\right)$$ and

$2\hat{\mathbf{A}}''\hat{\mathbf{B}}_Y - 2\sum_g \frac{n_g-1}{n_g}\sum_j\left(\hat{\mathbf{A}}''\right)^2$. Thus,

$$v_{JL}^R(\hat{Y}) - v_{JL}^S(\hat{Y}) = \sum_g \frac{n_g-1}{n_g}\sum_j\left(\left(\hat{\mathbf{A}}''\right)^2 + 2\hat{\mathbf{A}}''\hat{\mathbf{B}}_{\mathbf{Y}}\right)$$

$$= -\sum_g \frac{n_g-1}{n_g}\sum_j\left(\hat{A}''\right)^2 = -\left(\hat{\mathbf{A}}''\right)^2$$

That is, the shortcut procedure variance estimator is **always** larger than the reweighted estimator. As $r_g \to n_g$ the contribution from the g^{th} group to this term approaches zero, i.e., the higher the response rate, the closer the procedure is to the usual textbook variance estimator. As with the ratio adjustment procedure, the relative contribution of the missing $\left(\hat{\mathbf{A}}''\right)^2$ term to the total MSE depends on the unit response rate and the magnitude of the characteristic within each stratum.

3 Empirical Results

Section 2 addresses the question of what are the variance estimation effects of using a shortcut procedure compared to a reweighted procedure, demonstrating overestimation with the count adjustment procedure and expected overestimation with the ratio adjustment procedure. The question then becomes what is the degree of overestimation and what factors might cause the overestimation to be severe enough to bias survey conclusions.

To evaluate these questions empirically, we used survey data from the U.S. Census Bureau's Annual Capital Expenditures Survey (ACES). The ACES survey is a mail-out/mail-back that collects data about the nature and level of capital expenditures in non-farm businesses operating within the United States. Respondents report capital expenditures for the calendar year in all subsidiaries and divisions for all operations within the United States. ACES respondents report total capital expenditures, broken down by type (expenditures on Structures and expenditures on Equipment). Hereafter, we refer to these characteristics as Total, Structures, and Equipment. Tables 1 through 3 compare standard error estimates of capital expenditures statistics from three years' of ACES data: the first two data sets (from survey years 2002 and 2003) are the full collection of final tabulated ACES data and the third data set (survey year 2003) contains a mid-survey collection of data.

The ACES universe contains two sub-populations: employer companies and non-employer companies. Different forms are mailed to sample units depending on whether they are employer (ACE-1) companies or non-employer (ACE-2) companies. New ACE-1 and ACE-2 samples are selected each year, both with stratified SRS-WOR designs. The ACE-1 sample comprises approximately seventy-five percent of the ACES sample (roughly 45,000 companies selected per year for ACE-1, and 15,000 selected per year for ACE-2). The ACE-survey strata are defined by four company size class categories within industry (denoted 2A through 2D, ranked from largest to smallest within industry), with approximately 500 non-certainty strata each year. Sampling fractions in the large-size class-within-industry strata (2A) can be fairly high (approximately 55% of the sample in these strata are sampled at rates between 1 and 2); sampling fractions in the other three size class within-industry strata are usually less than 0.20 and sampling weights range from 5 to 1000, depending on industry and size-class strata. The ACE-2 component is much less highly stratified, with between a total of six to eight size-class strata used each year, and sampling fractions less than 0.01 in all strata. Because the response rates in certainty strata are generally close to 100%, we exclude certainty units from the variance estimates discussed below. We do not otherwise incorporate fpc's into our calculations. The ACE-1 component has a fairly high expected unit response rate (approximately 80% in most strata), whereas the corresponding rate for ACE-2 tends to be somewhat lower (ranging from 60 to 80%).

The ACE-1 component uses the **ratio adjustment** procedure with administrative data payroll as auxiliary variable to account for unit non-response, whereas the ACE-2 component uses the **count procedure**. In almost all cases, ACES uses design strata as weighting cells: under complete unit non-response in an ACE-1 industry's certainty stratum or in the large company (2A) stratum, the two strata are combined into one weighting cell (within the sample industry). Presently, there is no collapsing procedure in place for complete non-response in the three remaining ACE-1 (within-industry) non-certainty strata. In general, stratum collapsing for weight adjustment is a very rare occurrence and is hereafter ignored in this paper. To assess the effect of unit non-response weight adjustment procedure on the ACE-1 standard errors, we compute standard errors from ACE-1 data using both weight adjustment procedures (ratio and count). Since payroll data are not available for the ACE-2 component, we only present results using the count adjustment for that sub-population.

Capital expenditures data are fairly atypical business data, in that they often are characterized by low year-to-year correlation for the same reporting unit: for example, a company that spends a large amount of capital expenditures on structural (building) improvements one year is unlikely to invest much in structural improvements in the following year. Moreover, a reported value of zero for an expenditures item is quite legitimate and is often the response value for most items reported by a small company.

Table 1 presents standard error estimates for reweighted stratified jackknife (SJR) standard errors for the ACE-1 and ACE-2 data sets, along with the jackknife linearization standard errors obtained from reweighted procedures (LinR) and shortcut procedures (LinS). The shortcut procedure stratified jackknife and shortcut linearized jackknife standard errors are equivalent.

The reweighted jackknife and linearized reweighted jackknife standard error estimates are all within 1-percent of each other. Regardless of weight adjustment procedure, the linearized **shortcut** standard errors are larger than corresponding linearized reweighted standard errors. The degree of "overestimation" is, however, quite small: in most cases, the shortcut procedure standard errors are less than 2-percent larger than their fully reweighted procedure counterparts.

Table 1: Comparison of Fully Reweighted and Shortcut Procedure Standard Errors (in Millions)

			Ratio Adjustment			Count Adjustment		
			SJR	LinR	LinS	SJR	LinR	LinS
2001	ACE-1	Total	8.4	8.3	8.4	8.5	8.4	8.5
		Structures	6.1	6.1	6.1	6.2	6.2	6.2
		Equipment	4.6	4.6	4.7	4.7	4.6	4.7
	ACE-2	Total				3.0	3.0	3.0
		Structures				1.9	1.9	1.9
		Equipment				2.0	2.0	2.0
2002	ACE-1	Total	9.5	9.5	9.6	9.8	9.7	9.8
		Structures	8.3	8.3	8.3	8.7	8.5	8.5
		Equipment	4.2	4.2	4.2	4.3	4.2	4.3
	ACE-2	Total				5.7	5.7	5.8
		Structures				4.9	4.9	4.9
		Equipment				2.4	2.4	2.4
2003	ACE-1	Total	19.9	19.5	19.9	21.5	21.2	21.4
		Structures	43.0	42.6	43.1	44.2	44.0	44.4
		Equipment	17.4	17.1	17.4	19.0	18.9	18.8
	ACE-2	Total				15.8	15.8	15.8
		Structures				25.0	25.0	25.0
		Equipment				89.0	89.0	8.9

For the ratio adjustment procedure estimates, the difference between shortcut and reweighted procedure standard error estimates is the sum of two separate terms, $\left(\hat{\mathbf{A}}''\right)^2$ (the auxiliary variable MSE contribution from both sampling phases) and $2\hat{\mathbf{A}}''\hat{\mathbf{B}}_Y$. Table 2 presents the linearized reweighted variance estimates (LinR_V) along with these estimated components. Notice that the relative contribution of the $\left(\hat{\mathbf{A}}''\right)^2$ term is very small compared to the $2\hat{\mathbf{A}}''\hat{\mathbf{B}}_Y$ covariance term and to the total estimated variance. The relative magnitude of the $\left(\hat{\mathbf{A}}''\right)^2$ term indicates a "canceling" effect of the auxiliary variable error contributions from both sample phases, i.e., the error contribution from the 1st to 2nd phase is only **slightly** larger than the auxiliary variable error contribution to the ordinary ratio estimate. This is not unreasonable, given the consistently high unit response rates in the ACE-1 strata. The correlation between payroll and capital expenditures data is quite low for small companies, accounting for the proximity of corresponding reweighted and shortcut variance estimates.

Table 2: Variance Components of Fully Reweighted Linearized Jackknife with Ratio Adjustment (ACE-1) (x 10^{12})

	Item	**LinR_V**	$\left(\hat{\mathbf{A}}''\right)^2$	$2\hat{\mathbf{A}}''\hat{\mathbf{B}}_Y$	$\left(\hat{\mathbf{A}}''\right)^2+2\hat{\mathbf{A}}''\hat{\mathbf{B}}_Y$
2001	Total	69.7	2.3	-4.7	-2.3
	Structures	37.5	0..7	-1.5	-0.8
	Equipment	21.5	1.0	-1.9	-0.9
2002	Total	89.8	1.7	-3.8	-2.1
	Structures	68.4	0.68	-1.6	-0.9
	Equipment	17.3	0.6	-1.3	-0.6
2003	Total	381	9.3	-23.2	-14
	Structures	18.1	0.6	-1.1	-0.4
	Equipment	293	4.7	-13.3	-8.6

Table 3 presents the linearized reweighted variance estimates using the count adjustment procedure from the ACE-1 and ACE-2 data and the $\left(\hat{\mathbf{A}}''\right)^2$ component. Again, we see canceling in the auxiliary variable error contributions, with the ordinary ratio estimate error increase being offset by an approximately equally large error component due to the random sample size (1st to 2nd stage error component).

Table 3: Variance Components of Fully Reweighted Linearized Jackknife with Count Adjustment (x 10^{12})

		Item	**LinR_V**	$\left(\hat{\mathbf{A}}''\right)^2$
2001	ACE-1	Total	71.0	1.6
		Structures	38.2	0.5
		Equipment	21.5	0.7
	ACE-2	Total	9.0	0.1
		Structures	3.7	0.01
		Equipment	4.0	0.08
2002	ACE-1	Total	94.5	1.5
		Structures	72.2	0.6
		Equipment	18.0	0.5
	ACE-2	Total	32.8	0.7
		Structures	24.1	0.1
		Equipment	5.7	0.1
2003	ACE-1	Total	451	8.6
		Structures	19.3	0.4
		Equipment	349	5.5
	ACE-2	Total	248	1.6
		Structures	6.2	0.01
		Equipment	79.1	0.6

The ACES data are characterized by high unit response rates and high reported zero rates. The results presented in Tables 2 and 3 are completely in line with a MAR model assumption: that is, there is a very small contribution to the overall error from unit non-response. The variance estimation results alone are not, however, sufficient for assuming ignorable non-response, since our derivations assume a uniform or MAR response mechanism.

4 Conclusion

This paper presents research undertaken to investigate the variance estimation effects of **not** replicating a unit non-response weight adjustment procedure. We assume that respondents comprise a Bernoulli sample of sampled units within a weighting class, so that the realized sample has a random sample size and this random element is reflected in the fully reweighted jackknife variance estimates.

Given an ignorable response mechanism and a two-phase sample design (stratified SRS-WOR at 1st phase, Bernoulli at 2nd phase), we show that using a shortcut procedure yields overly large MSE estimates with the count procedure adjustment and generally overestimates the MSE with a ratio adjustment procedure. The degree of overestimation is, however, a function of the weighting cell sample size, the weighting cell respondent rate, and in the case of the ratio adjustment procedure, the covariance between the characteristic of interest and the auxiliary variable. In a highly stratified survey with varying survey weights and varying unit response rates, it is difficult to predict at what point the cumulative effect of the shortcut procedure overestimation will become severe enough to affect confidence interval coverage. In fact, it is likely that corresponding fully reweighted and shortcut variance estimates will often be comparable; the two MSE components from the auxiliary variable will nearly cancel, and the majority of the variance will derive from the respondent sample ratio estimate.

The 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 fully replicating the weight adjustment procedure, there may be little or no practical advantage. Having said that, our applications demonstrated excellent approximations to stratified jackknife variance estimates with our linearized jackknife variance estimators, both of which are computationally quick and computer overhead "free" (in terms of replicate storage). Given these viable alternatives, it is difficult to justify the use of a shortcut procedure variance estimator -- replicated or linearized – over a fully replicated procedure variance, at least in the case of weighting adjustment for unit non-response.

In conclusion, note that our findings rely on several assumptions. In Section 2, we assume that 2nd order and higher derivatives are negligible for the linearization. Moreover, these derivations rely on a Bernoulli sample of respondents. We make statements about the expected "overestimation" with a shortcut procedure stratified jackknife ratio estimator and a ratio adjustment procedure for unit non-response, assuming a positive correlation between characteristic(s) and auxiliary variables. The empirical results in Section 3 provide some support for the first two assumptions. More general statements about the variance estimation effects with a ratio adjustment procedure or the degree of variance overestimation using a shortcut procedure and a count adjustment are not possible without a controlled study; and in fact, the next steps of our research will explore these issues via a simulation study. Finally, the results presented in this paper are applicable to data sets with missing-completely-at-random (uniform) or missing-at-random (MAR) response mechanisms. However, the view of unit non-response as a second random stage of sampling is not necessarily realistic for a voluntary survey. It is equally likely that several non-respondents are **fixed** in the population, so that the unit non-response is an estimation-bias problem. A limitation of our results is they do not apply to survey data that has a nonignorable response mechanism. And of course, without conducting a survey of nonresponding units, it is impossible to assess the validity of the model assumptions used.

Acknowledgements

The authors thank Samson Adeshiyan, Michael Hidiroglou, Alphonso Mason, Douglas Olson, and Rita Petroni for their careful review of earlier versions of this manuscript.

References

Cochran, W. (1977). Sampling Techniques, 3rd Edition. New York: Wiley.

Deville, J.C. and Särndal, C. (1994). Variance Estimation for the Regression Imputed Horvitz-Thompson Estimator. *JOS*, **4**, pp. 381-394.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting Methods. *JOS*, **19**, pp. 81-97.

Kott, P. (1994). A Note on Handling Nonresponse in Sample Surveys. *Journal of the American Statistical Association*, **89**, pp. 693 –696.

Little, R.J.A. and Rubin, D.R. (2002). Statistical Analysis with Missing Data. New York: John Wiley and Sons.

Little, R.J.A. and Vartivarian, S.(2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*, **31**, pp. 161-168.

Oh, H.L. and Scheuren, F.J. (1983). Weighting Adjustment of Unit Nonresponse. Incomplete Data in Sample Surveys (Vol.20). New York: Academic Press, 143-184.

Rao, J.N.K. and Sitter, R.R. (1995). Variance Estimation Under Two-Phase Sampling With Application to Imputation for Missing Data. *Biometrika*, **82**, pp. 453-460.

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

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**, pp. 254-265.

Vartivarian, S. and Little, R.J. (2002). On the Formation of Weighting Adjustment Cells for Unit Non-response. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, pp. 3553-3558.

Wolter, K. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.