EFFICIENT REPLICATION VARIANCE ESTIMATION FOR TWO-PHASE SAMPLING

Jae-Kwang Kim[1] Randy Sitter[2]*

ABSTRACT

Variance estimation for the regression estimator for a two-phase sample is investigated. A replication variance estimator with number of replicates equal to or slightly larger than the size of the second-phase sample is developed. In these cases, the proposed method is asymptotically equivalent to the full jackknife, but uses smaller number of replications.

KEY WORDS. *Double sampling; Jackknife; Regression estimator*

1. Introduction

It is common in surveys to use *two-phase* or *double* sampling when it is relatively inexpensive to draw a large first-phase sample for which a vector auxiliary variate, x, correlated with the characteristic of interest y, alone is observed. A second-phase subsample of the initial first-phase sample is then drawn and both y and x are measured. Various estimation strategies exist for combining the information from both phases of sampling to estimate characteristics of the population based on y or (y, x).

Jackknife variance estimators for two-phase sampling have been developed in Rao and Sitter (1995). Kim, Navarro and Fuller (2000) develop a jackknife variance estimator to handle what Kott and Stukel (1997) term the double expansion estimator. It should be noted that some of the ideas for these jackknife estimators are implicit in Rao and Shao (1992), though they consider the jackknife for random imputation.

A key feature of these full jackknife variance estimators is that replicates are formed for each unit in the first-phase sample. When the first-phase sample is very large, as is common, and in particular much larger than the second-phase sample, there are practical reasons why having so many replicates may be undesirable. When the final user is different than the data provider, it is common practice to include the set of replicate weights in the data set. Thus a large number of replicates in a large survey with many measured characteristics causes, what turns out to be unnecessary, computational and storage burdens on the end user.

Fuller (1998) recognizes this practical issue and proposes a creative solution whereby the required number of jackknife replicate weights can be reduced in some cases. In Section 2, we propose an alternative method of reducing the replicates. In Section 3, we highlight its application by considering the double expansion estimator when the second-phase strata are nested within the first-phase strata.

[1] Department of Statistics, Hankuk University of Foreign Studies, Yongin, Kyungki-Do 449-791, Korea.

[2] Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, V5A 1S6 CANADA.

2. Reducing the Number of Replicate Weights in the Jackknife for the Two-phase Regression Estimator

Let us consider estimation of the population total, $\boldsymbol{Y}$, of vector $\boldsymbol{y}$ from a two-phase sample. Let $\hat{\boldsymbol{X}}_1$ be an unbiased estimate of the population total, $\boldsymbol{X}$, of vector $\boldsymbol{x}$ constructed from the first-phase sample, A_1, $\hat{\boldsymbol{X}}_2$ be an unbiased estimate of the population total of $\boldsymbol{x}$ constructed from the second-phase sample, A_2, and $\hat{\boldsymbol{Y}}_2$ be an unbiased estimate of the population total of $\boldsymbol{y}$ constructed from A_2. Write

$$\hat{\boldsymbol{X}}_1 = \sum_{i \in A_1} w_i \boldsymbol{x}_i \quad \text{and} \quad \left(\hat{\boldsymbol{X}}_2, \hat{\boldsymbol{Y}}_2\right) = \sum_{i \in A_2} w_i w_{i2} \left(\boldsymbol{x}_i, \boldsymbol{y}_i\right). \tag{1}$$

The first-phase sampling weight, w_i, is often the inverse of the inclusion probability for the first-phase sampling. The second-phase sampling weight, w_{i2}, is often the inverse of the conditional selection probability for the second-phase sample given the first-phase sample.

For simplicity, consider a scalar y variable. The regression estimator of Y takes the form

$$\hat{Y}_{reg} = \hat{Y}_2 + (\hat{\boldsymbol{X}}_1 - \hat{\boldsymbol{X}}_2)' \hat{\boldsymbol{\beta}}_{(2)}, \tag{2}$$

where

$$\hat{\boldsymbol{\beta}}_{(2)} = \Big(\sum_{j \in A_2} w_j w_{j2} \boldsymbol{x}_j \boldsymbol{x}_j' \Big)^{-1} \sum_{j \in A_2} w_j w_{j2} \boldsymbol{x}_j y_j,$$

and $w_j w_{j2}$ are the two-phase weights used in (1). If we include 1 as the first component of $\boldsymbol{x}$, i.e. an intercept, then we can rewrite $\hat{Y}_{reg}$ as

$$\hat{Y}_{reg} = \hat{\boldsymbol{X}}_1' \hat{\boldsymbol{\beta}}_{(2)}. \tag{3}$$

Let us imagine we had the observed y on the entire first-phase sample. Then the full first-phase sample variance of $\hat{Y}_1 = \sum_{i \in A_1} w_i y_i$ can be estimated by a full jackknife method of the form

$$v_J(\hat{Y}_1) = \sum_{k \in A_1} c_k (\hat{Y}_1^{(k)} - \hat{Y}_1)^2, \tag{4}$$

where $\hat{Y}_1^{(k)} = \sum_{i \in A_1} w_i^{(k)} y_i$ and c_k is a factor associated with the sampling design. For example, in an unstratified setting commonly used replication weights are $w_i^{(k)} = (n-1)^{-1} n w_i$ for $k \neq i$ and $w_i^{(i)} = 0$, and the factor c_k is equal to $n^{-1}(n-1)$.

Let the two-phase sample variance of $\hat{Y}_2 = \sum_{i \in A_2} w_i w_{i2} y_i$ also be estimated by a full jackknife method of the form

$$v_J(\hat{Y}_2) = \sum_{k \in A_1} c_k (\hat{Y}_2^{(k)} - \hat{Y}_2)^2, \tag{5}$$

where $\hat{Y}_2^{(k)} = \sum_{i \in A_2} w_i^{(k)} w_{i2}^{(k)} y_i$ and $w_{i2}^{(k)}$ is the k-th replicate of the second phase weighting factor w_{i2}. (see Kim, Navarro, and Fuller, 2000).

Now, given the variance estimators for the first-phase estimator of the form (4) and for the two-phase direct estimator of the form (5), replication variance estimators are available for the two-phase regression estimators because the regression estimator is a smooth function of the direct estimators on the first-phase sample and the second-phase sample. Thus, one can define for $k \in A_1$,

$$\hat{Y}_{reg}^{(k)} = \sum_{i \in A_2} \alpha_i^{(k)} y_i = \hat{\boldsymbol{X}}_1^{(k)\prime} \hat{\boldsymbol{\beta}}_{(2)}^{(k)} \tag{6}$$

where $\hat{\boldsymbol{X}}_1^{(k)} = \sum_{i \in A_1} w_i^{(k)} \boldsymbol{x}_i$ and

$$\hat{\boldsymbol{\beta}}_{(2)}^{(k)} = \Big(\sum_{j \in A_2} w_j^{(k)} w_{j2}^{(k)} \boldsymbol{x}_j \boldsymbol{x}_j'\Big)^{-1} \sum_{j \in A_2} w_j^{(k)} w_{j2}^{(k)} \boldsymbol{x}_j y_j .$$

The full jackknife variance estimator would then take the form

$$v_J(\hat{Y}_{reg}) = \sum_{k \in A_1} c_k (\hat{Y}_{reg}^{(k)} - \hat{Y}_{reg})^2,$$

and would require the formation of n_1 sets of replicate weights, for n_2 records.

When the second-phase sample is much smaller than the first-phase sample, we may wish to reduce the total number of replicates. Having smaller number of replicates is particularly important in practice not only because of faster computation but also because of the smaller storage needed. When the final user is different from the data provider, it is a common practice to include the replication weights in the data set.

Fuller (1998) recognizes this problem and is able to reduce the number of required replicates in such cases by considering the regression estimator in two parts, namely as,

$$\hat{Y}_{reg} \doteq (\hat{Y}_2 - \hat{\boldsymbol{X}}_2' \boldsymbol{\beta}) + \hat{\boldsymbol{X}}_1' \boldsymbol{\beta},$$

and decomposing the variance into that corresponding to each term. He then shows that, if one has a replication method for estimating the variance of the first term, one can use a simple method to adjust it to add back the variance for the second term provided the covariance between the terms is negligible.

Instead, note that

$$\begin{aligned} v_J(\hat{Y}_{reg}) &= \sum_{k \in A_2} c_k (\hat{Y}_{reg}^{(k)} - \hat{Y}_{reg})^2 + \sum_{k \in A_1 \cap A_2^c} c_k (\hat{Y}_{reg}^{(k)} - \hat{Y}_{reg})^2 \\ &= v_{J,2} + v_{J,1-2} \end{aligned} \tag{7}$$

and consider the second term. It turns out to be possible in many situations to create fewer replicates than the number of elements in $A_1 \cap A_2^c$ to capture the second term of (7). To see this, rewrite

$$\begin{aligned} v_{J,1-2} &= \hat{\boldsymbol{\beta}}_{(2)}' \Big[\sum_{k \in A_1 \cap A_2^c} c_k (\hat{\boldsymbol{X}}_1^{(k)} - \hat{\boldsymbol{X}}_1)(\hat{\boldsymbol{X}}_1^{(k)} - \hat{\boldsymbol{X}}_1)' \Big] \hat{\boldsymbol{\beta}}_{(2)} + \sum_{k \in A_1 \cap A_2^c} c_k \left[\hat{\boldsymbol{X}}_1^{(k)\prime} (\hat{\boldsymbol{\beta}}_{(2)}^{(k)} - \hat{\boldsymbol{\beta}}_{(2)}) \right]^2 \\ &= \hat{\boldsymbol{\beta}}_{(2)}' \tilde{\boldsymbol{V}}_x \hat{\boldsymbol{\beta}}_{(2)} + V_2 . \end{aligned}$$

Then, in some common cases

$$\hat{\boldsymbol{\beta}}_{(2)}^{(hi)} \doteq \hat{\boldsymbol{\beta}}_{(2)} \tag{8}$$

for all units (hi) that belong to the first-phase sample but not the second, and thus $V_2 \doteq 0$.

If (8) holds, we can employ a tactic similar to that of Fuller (1998) but applied to this portion of v_J. That is, let $\boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_{n_2}$ be a set of m-dimensional vectors, where $m < n_2$ is the dimension of $\boldsymbol{x}$, and

$$\sum_{j=1}^{n_2} \boldsymbol{\delta}_j \boldsymbol{\delta}_j' = \tilde{\boldsymbol{V}}_x.$$

For example, let $\boldsymbol{\gamma}_j$ be the characteristic vectors of $\tilde{\boldsymbol{V}}_x$ and λ_j their corresponding roots. Then define $\boldsymbol{\delta}_j = \lambda_j^{1/2}\boldsymbol{\gamma}_j$ for $j = 1, ..., m$ and $\boldsymbol{\delta}_j = \mathbf{0}$ for $j = m+1, ..., n_2$.

Now obtain a set of $2n_2$ adjusted replicate weights, $\tilde{\alpha}_i^{(k)}$, for the set $k \in A_2$ such that

$$\tilde{Y}_{reg}^{(k)} = \sum_{i\in A_2} \tilde{\alpha}_i^{(k)} y_i = \hat{\boldsymbol{X}}_1^{(k)\prime} \hat{\boldsymbol{\beta}}_{(2)}^{(k)} + c_k^{-1/2} \boldsymbol{\delta}_k' \hat{\boldsymbol{\beta}}_{(2)}.$$

That is,

$$\tilde{\alpha}_i^{(k)} = \alpha_i^{(k)} + c_k^{-1/2} \boldsymbol{\delta}_k' \Big(\sum_{j\in A_2} w_j w_{j2} \boldsymbol{x}_j \boldsymbol{x}_j'\Big)^{-1} w_i w_{i2} \boldsymbol{x}_i, \tag{9}$$

for $k \in A_2$, where $\alpha_i^{(k)}$ is given in (6). If one repeats this entire process creating $\tilde{Y}_{reg2}^{(k)} = \hat{\boldsymbol{X}}_1^{(k)\prime} \hat{\boldsymbol{\beta}}_{(2)}^{(k)} - c_k^{-1/2} \boldsymbol{\delta}_k' \hat{\boldsymbol{\beta}}_{(2)}$ by subtracting $c_k^{-1/2}\boldsymbol{\delta}_k$ in the right hand side of (9), for $k \in A_2$, then we can define a new jackknife variance estimator as

$$\tilde{v}_J = \sum_{k\in A_2} c_k (\tilde{Y}_{reg}^{(k)} - \hat{Y}_{reg})^2 + \sum_{k\in A_2} c_k (\tilde{Y}_{reg2}^{(k)} - \hat{Y}_{reg})^2.$$

If so, it follows that $\tilde{v}_J \doteq v_J$, and even though only $2n_2$ replicates are needed the efficiency of the full jackknife variance estimator is retained.

3. Illustration via nested-strata two-phase estimator

In this section, we will illustrate the potential of the proposed method in a simple situation. Consider stratified simple random sampling, where n_h units are selected with equal probability without replacement from a population of size N_h, independently across H strata. Let y_{hi} be the value of the study variable of unit i in stratum h. Instead of observing the y_{hi}'s directly, assume that we observe $\boldsymbol{x}_{hi} = (x_{hi1}, x_{hi2}, \cdots, x_{hiG_h})$, where x_{hig} takes the value 1 if unit i in stratum h belongs to group g, and takes the value 0 otherwise. Each unit belongs to one and only one group. We call group g in stratum h sub-stratum (hg). There are $n_{hg} = \sum_{\{i:(hi)\in A_1\}} x_{hig}$ units in sub-stratum (hg).

For the second-phase sampling, we assume that $r_{hg} \geq 2$ elements are selected without replacement with equal probability independently across the sub-strata. From the selected elements, we observe y_{hig}, where the subscript g is used to emphasize that unit (hi) belongs to group g. Then, an unbiased estimator for the total of the y-variable is

$$\hat{Y}_2 = \sum_{(hi)\in A_2} \sum_{g=1}^{G_h} \frac{N_h}{n_h} \frac{n_{hg}}{r_{hg}} y_{hig}. \tag{1}$$

The first factor $n_h^{-1}N_h$ is the inverse of the inclusion probability for the first-phase sampling and the second factor $r_{hg}^{-1}n_{hg}$ is the inverse of the inclusion probability for the second-phase sampling. The variance of $\hat{Y}_2$ can be written as

$$\begin{aligned} Var\left(\hat{Y}_2\right) &= E\left\{\sum_{h=1}^{H}\left(\frac{N_h}{n_h}\right)^2(1-f_{1h})\frac{n_h}{n_h-1}\sum_{g=1}^{G_h}n_{hg}\left[\left(\bar{y}_{hg}-\bar{y}_h\right)^2+s_{hg}^2\right]\right\} \\ &+ E\left\{\sum_{h=1}^{H}\left(\frac{N_h}{n_h}\right)^2\sum_{g=1}^{G_h}\frac{n_{hg}^2}{r_{hg}}\left(1-\frac{r_{hg}}{n_{hg}}\right)s_{hg}^2\right\}, \qquad (2) \end{aligned}$$

where $f_{1h} = N_h^{-1}n_h$ is the first phase sampling rate, $s_{hg}^2=(n_{hg}-1)^{-1}\sum_{\{i:(hi)\in A_1\}}\left(y_{hig}-\bar{y}_{hg}\right)^2$ is the first-phase sample variance of sub-stratum (hg), $\bar{y}_{hg} = n_{hg}^{-1}\sum_{\{i:(hi)\in A_1\}}y_{hig}$ is the first-phase sample mean of sub-stratum (hg), $s_h^2 = (n_h-1)^{-1}\sum_{\{i:(hi)\in A_1\}}\left(y_{hi}-\bar{y}_h\right)^2$ is the first-phase sample variance of stratum h, and $\bar{y}_h = n_h^{-1}\sum_{g=1}^{G_h}n_{hg}\bar{y}_{hg}$ is the first-phase sample mean of stratum h.

A variance estimator can be easily derived from (2) by replacing $\bar{y}_{hg}$ and s_{hg}^2 by their estimates $\bar{y}_{hg2} = r_{hg}^{-1}\sum_{\{i:(hi)\in A_2\}}y_{hig}$ and $s_{hg2}^2 = (r_{hg}-1)^{-1}\sum_{\{i:(hi)\in A_2\}}\left(y_{hig}-\bar{y}_{2hg}\right)^2$, respectively. That is, ignoring the f_{1h} terms, a consistent variance estimator is

$$\hat{V} = \sum_{h=1}^{H}N_h^2n_h^{-2}\sum_{g=1}^{G_h}n_{hg}\left(\bar{y}_{hg2}-\bar{y}_{h2}\right)^2+\sum_{h=1}^{H}N_h^2n_h^{-2}\sum_{g=1}^{G_h}r_{hg}^{-1}n_{hg}^2s_{hg2}^2. \qquad (3)$$

Kim, Navarro and Fuller (2000) develop a jackknife variance estimator by successively deleting units from the entire first-phase sample and then adjusting the weights. The weights of the two-phase estimator in (1) are products of $w_{hi} = n_h^{-1}N_h$, the first-phase sampling weight, and $w_{hgi2} = r_{hg}^{-1}n_{hg}$, the second-phase sampling weight. The full jackknife replicate weights for w_{hi} and w_{hig2} are

$$w_{hi}^{(h'i')} = \begin{cases} 0 & \text{if } h=h', i=i' \\ \left(n_h-1\right)^{-1}n_hw_{hi} & \text{if } h=h', i\neq i' \\ w_{hi} & \text{if } h\neq h' \end{cases} \qquad (4)$$

and

$$w_{hgi2}^{(h'i')} = \begin{cases} 0 & \text{if } h=h', i=i' \\ (r_{hg}-1)^{-1}(n_{hg}-1) & \text{if } h=h', i\neq i', x_{h'i'g}=1, \text{ and } (h'i')\in A_2 \\ r_{hg}^{-1}(n_{hg}-1) & \text{if } h=h', i\neq i', x_{h'i'g}=1, \text{ and } (h'i')\notin A_2 \\ r_{hg}^{-1}n_{hg} & \text{otherwise,} \end{cases} \qquad (5)$$

respectively.

The full jackknife variance estimator of the form

$$\hat{V}_J = \sum_{(hi)\in A_1}\frac{(n_h-1)}{n_h}\left(\hat{Y}_2^{(h,i)}-\hat{Y}_2\right)^2,$$

where $\hat{Y}_2^{(h'i')} = \sum_{(hi)\in A_2}\sum_{g=1}^{G_h} w_{hi}^{(h'i')} w_{hgi2}^{(h'i')} y_{hig}$, is asymptotically equivalent to the variance estimator in (3), with total number of replicates $n = \sum_{h=1}^{H}\sum_{g=1}^{G_h} n_{hg}$ for r records. To apply the idea proposed in the previous section, note that

$$\hat{Y}_2^{(hi)} - \hat{Y}_2 = \begin{cases} \frac{N_h}{n_h-1}\left(\bar{y}_{h2}-\bar{y}_{hg2}\right) + \frac{N_h}{n_h-1}\frac{n_{hg}-1}{r_{hg}-1}\left(\bar{y}_{hg2}-y_{hig}\right) & \text{if } (hi)\in A_2 \\ \frac{N_h}{n_h-1}\left(\bar{y}_{h2}-\bar{y}_{hg2}\right) & \text{if } (hi)\notin A_2 \end{cases}$$

for unit (hi) in group g, which makes the decomposition of the full jackknife variance estimator as in (7) particularly simple,

$$\begin{aligned} \hat{V}_J &= v_{J,2} + v_{J,1-2} \\ &\doteq \sum_{(hi)\in A_2} \frac{n_h-1}{n_h}(\hat{Y}_2^{(hi)} - \hat{Y}_2)^2 + \sum_{h=1}^{H}\left(\frac{N_h}{n_h}\right)^2 \sum_{g=1}^{G_h}(n_{hg}-r_{hg})\left(\bar{y}_{hg2}-\bar{y}_{h2}\right)^2 . \end{aligned}$$

Thus, deleting a unit which is in the first-phase sample but not in the second-phase sample does not contribute to the method's capturing of the second component of (3).

Using the full jackknife method directly uses r replicates to calculate $v_{J,2}$ and $n-r$ replicates to calculate $v_{J,1-2}$. Our proposed method amounts to calculating $v_{J,2}$ from the full jackknife method using the same r replicates, but calculating $v_{J,1-2}$ using a smaller number of replicates. In this simple setting, it is quite easy to create replicates for $v_{J,1-2}$. To see this, note that we can write

$$v_{J,1-2} = \sum_{h=1}^{H}\sum_{g=1}^{G_h} c_{hg}\left(\hat{Y}_2^{(hg)} - \hat{Y}_2\right)^2 ,$$

where

$$\hat{Y}_2^{(hg)} = \hat{Y}_2 + c_{hg}^{-1/2}\left(n_{hg}-r_{hg}\right)^{1/2}\frac{N_h}{n_h}\left(\bar{y}_{hg2}-\bar{y}_{h2}\right)$$

for any $c_{hg} \geq (n_{hg}-r_{hg})$, where condition $c_{hg} \geq (n_{hg}-r_{hg})$ guarantees nonnegative replication weights for all records. Therefore, the total number of replicates is reduced to $r+G$, where the first r replicates are used to estimate $v_{J,2}$ and the last $G<r$ replicates are used to estimate $v_{J,1-2}$.

References

Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica* **8**, 1153-1164.

Kim, J.K., Navarro, A. and Fuller, W.A. (2000). Variance estimation for 2000 census coverage estimates. *Proc. ASA Section on Survey Research Methods*, 515-520.

Kott, P.S. and Stukel, D.M. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology* **23**, 81-89.

Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79**, 811-822.

Rao, J.N.K. and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, **82**, 453-460.