Model-Assisted Approach to Inference for Totals in Cluster Sampling under Imputation for Missing Data

D. Haziza, Statistics Canada
J.N.K. Rao, Carleton University
D. Haziza, Statistics Canada, 16-R R.H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A 0T6

Keywords: Intracluster correlation, Model-assisted framework, Uniform response, Variance Estimation

1 Introduction

Many surveys use imputation to handle item nonresponse as a way to patch up the sample. However, it is a common practice to treat the imputed values as if they are true values, and then compute the variance estimates using standard formulas. This can lead to serious underestimation of the true variance of the estimates when the proportion of missing values is not small. Extensive literature exists on the model-assisted approach to inference for population totals and means under imputation for missing data; see for example Särndal (1992), Deville and Särndal (1994), Rancourt, Särndal and Lee (1994) and Shao and Steel (1999).

Under the model-assisted framework, the following assumption holds:
Assumption MA: Within an imputation cell the response mechanism is ignorable or unconfounded in the sense that whether or not a unit responds does not depend on the variable being imputed but may depend on the covariates used for imputation. Imputation is performed according to a model which Särndal (1992) calls "imputation model". For regression imputation, the imputation model is given by

$$E_m(y_i) = \mathbf{z_i'}\boldsymbol{\beta}, V_m(y_i) = \sigma_i^2 = \sigma^2\mathbf{z_i'}\boldsymbol{\lambda}, \quad Cov_m(y_i, y_j) = 0 \text{ if } i \neq j, \tag{1}$$

where $\boldsymbol{\beta}$ is a q-vector of unknown parameters, $\mathbf{z_i}$ is a q-vector of auxiliary variables available for all the sample units, $\boldsymbol{\lambda}$ is a q-vector of specified constants, σ^2 is an unknown parameter and E_m, V_m, and Cov_m denote respectively the expectation, the variance and the covariance operators with respect to the imputation model. The restriction $\sigma_i^2 = \sigma^2\mathbf{z_i'}\boldsymbol{\lambda}$ does not severely restrict the range of imputation models. For simplicity, we consider the case of a single imputation cell but extension to the case of multiple imputation cells is straightforward. Typically, a linear model of the form (1) with fixed effects and a diagonal variance-covariance matrix is used as an imputation model. However, if the existing "model-assisted methods" are applied to the case of single-stage or two-stage cluster sampling, the estimated variances of the imputed estimators may be biased since they do not take account of the intracluster correlation. A nested error linear regression model may provide a more appropriate theoretical framework for cluster sampling. In this paper, using such models as imputation models, we make inference on a population total and derive consistent variance estimators of the imputed estimator using a method introduced by Fay (1991) and developed by Shao and Steel (1999).

Traditionally, researchers have used the following sample-response path (sometimes called the two-phase approach) for variance estimation:

Population $\longrightarrow$ complete sample $\longrightarrow$ sample with nonrespondents.

In this case,

$$E\left(\hat{\theta}\right) = E_pE_r\left(\hat{\theta}\right), \tag{2}$$

$$V\left(\hat{\theta}-\theta\right) = E_p\left[V_r\left(\hat{\theta}-\theta\right)\right] + V_p\left[E_r\left(\hat{\theta}-\theta\right)\right], \tag{3}$$

where θ denotes an arbitrary parameter and $\hat{\theta}$ denotes its estimator based on the observed and

imputed data, $E_p(.)$ and $V_p(.)$ denote respectively the expectation and the variance with respect to the sampling design and $E_r(.)$ and $V_r(.)$ denote respectively the expectation and the variance with respect to the response mechanism. Note that in the case of two-stage sampling, $E_p(.) = E_1E_2(.)$ and $V_p(.) = E_1V_2(.) + V_1E_2(.)$ where E_1, E_2, V_1, V_2 denote the expectation and the variance operators with respect to the first and second stage respectively.

Fay(1991) proposed to use a different path which reverses the order of sampling and response (we will call it the reverse approach) that can be depicted as:

Population$\longrightarrow$ census with nonrespondents $\longrightarrow$ sample with nonrespondents.

In this case, (see Shao and Steel, 1999)

$$E\left(\hat{\theta}\right) = E_rE_p\left(\hat{\theta}\right), \tag{4}$$

$$V\left(\hat{\theta}-\theta\right) = E_r\left[V_p\left(\hat{\theta}-\theta\right)\right] + V_r\left[E_p\left(\hat{\theta}-\theta\right)\right]. \tag{5}$$

In the model assisted framework, we replace $E_r(.)$ and $V_r(.)$ by $\widetilde{E}_m(.) = E_rE_m(.)$ and $\widetilde{V}_m(.) = E_rV_m(.) + V_rE_m(.)$ respectively, where $E_m(.)$ and $V_m(.)$ denote respectively the expectation and the variance with respect to the imputation model. An estimator of the overall variance $V\left(\hat{\theta}-\theta\right)$ in (5) is given by $v_t = v_1 + v_2$, where v_1 is an estimator of $V_p\left(\hat{\theta}-\theta\right)$ conditional on the response indicators, and v_2 is an estimator of $V_rE_p\left(\hat{\theta}-\theta\right)$. One can show that under uniform response (see section 5), the order of $\frac{V_rE_p\left(\hat{\theta}-\theta\right)}{E_rV_p\left(\hat{\theta}\right)}$ is $O(\frac{n}{K})$ where n is the number of first-stage units (PSU's or clusters) in the sample and K is the number of ultimate units (or elements) in the population. Typically, $n << K$ and hence the second component in (5) is negligible relative to the first component. We can then omit the derivation of v_2 which could be quite tedious (Shao and Steel, 1999). However, under a "beta-binomial" response mechanism, one can show (see section 5) that the order of $\frac{\widetilde{V}_mE_p\left(\hat{\theta}-\theta\right)}{\widetilde{E}_mV_p\left(\hat{\theta}\right)}$ is $O(\frac{n}{N})$ where N is the number of first-stage units in the population and n is the number of PSU's selected at the first stage from the N clusters. The first-stage sampling fraction $\frac{n}{N}$ may not be negligible and hence the computation of v_2 must be performed.

We set out our basic framework and assumptions in section 2. In section 3, we present the imputed estimator of a population total and show that it is design-model unbiased. In section 4, using the method of Fay (1991), we derive a consistent estimator for the variance of the imputed estimator. In section 5, we investigate the underestimation of the variance of the imputed estimator occurring when one does not use the appropriate imputation model. The paper concludes with section 6, which summarizes the results and puts forth some suggestions.

2 Framework and Assumptions

Let U be a population consisting of N clusters, where N is known. Let U_i be the i^{th} cluster of size M_i, $i = 1, ..., N$ We have $\bigcup_{i=1}^{N} U_i = U$. For simplicity, we assume $M_i = M$ for all $i \in U$. We then have $K = MN$ ultimate units in the population. Let y be the variable of interest, y_{ij} be the value of y for the j^{th} element in the i^{th} cluster, $i = 1, ..., N; j = 1, ..., M$ and Y_i be the cluster total for the i^{th} cluster. The objective is to estimate the population total $Y = \sum_{i=1}^{N} Y_i = \sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}$ when imputation has been used to compensate for nonresponse. At the first stage, suppose a random sample of clusters, s, of size n is selected according to some design $p_1(s)$ from the population of clusters. At the second stage, suppose that from each cluster sampled at the first stage, a random sample of element, s_i, of size m_i $(i = 1, ..., n)$ is selected according to some design $p_2(s)$. Let s_{r_i} be the sample of respondents in the i^{th} cluster, of size r_i, and s_{o_i} be the sample of nonrespondents in the i^{th} cluster, of size o_i; $r_i + o_i = m_i$.

In this paper we consider mean imputation, in which case the imputation model under single-stage or two-stage cluster sampling, is the well known one-way ANOVA model with random effects given by

$$m : y_{ij} = \beta + \alpha_i + \epsilon_{ij}, \tag{6}$$

where β is the general mean, α_i is i-th cluster random effect and ϵ_{ij} is the residual error. We assume that
(i) $E_m(\alpha_i) = E_m(\epsilon_{ij}) = 0$,
(ii)$Cov_m(\epsilon_{ij}, \epsilon_{i'j'}) = 0$ except for $i = i'$ and $j = j'$,
$Cov_m(\alpha_i, \alpha_{i'}) = 0\ \forall i \neq i'$,
$Cov_m(\alpha_i, \epsilon_{i'j'}) = 0\ \forall i, i'$ and j',
(iii)$V_m(\alpha_i) = \sigma_\alpha^2\ \forall i$,
$V_m(\epsilon_{ij}) = \sigma_\epsilon^2\ \forall i, j$.
From (i)-(iii), we get

$$Cov_m\left(y_{ij}, y_{i'j'}\right) = \begin{cases} \sigma_\alpha^2 & \text{if } i = i' \text{ and } j \neq j' \\ \sigma_\alpha^2/\rho & \text{if } i = i' \text{ and } j = j' \\ 0 & \text{if } i \neq i' \end{cases}$$

where $\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2}$ is the intracluster correlation coefficient. We assume missing at random (MAR) mechanism so that the model holds for the sample respondents.

3 Point estimation

In this section, we present an imputed estimator of a total Y and show that it is design-model unbiased. Mean imputation under model (6) uses the predicted value $y_{ij}^* = \hat{\beta}_r$ for missing y_{ij}, where $\hat{\beta}_r$ is the weighted mean of respondents given by $\hat{\beta}_r = \frac{\sum_{i\in s}\sum_{j\in s_{r_i}} w_{ij}y_{ij}}{\sum_{i\in s}\sum_{j\in s_{r_i}} w_{ij}}$ and w_{ij} is the survey weight attached to unit j in cluster i. Using the y_{ij}^*'s, an imputed estimator of Y, denoted by $\hat{Y}_I$, is given by

$$\hat{Y}_I = \sum_{i\in s}\sum_{j\in s_i} w_{ij}\tilde{y}_{ij}, \tag{7}$$

where $\tilde{y}_{ij} = y_{ij}$ if $j \in s_{r_i}$ and $\tilde{y}_{ij} = y_{ij}^*$if $j \in s_{o_i}$. Since $E_m(Y) = K\beta$ and $E_m(y_{ij}^*) = \beta$, it follows that $E_m(\hat{Y}_I - Y) = 0$ and hence $E_pE_m(\hat{Y}_I - Y) = 0$. That is $\hat{Y}_I$ is design-model unbiased for Y, under the imputation model (6) .

As a second alternative, one may wish to use the empirical best linear unbiased predictor (EBLUP) of $\beta+\alpha_i$ as imputed values y_{ij}^* for $j \in s_{o_i}$. Under model (6), the best linear unbiased predictor of $\beta + \alpha_i$ is given by

$$\begin{aligned}\text{BLUP}(\beta + \alpha_i) &= \text{BLUE}\left(\beta\right) \\ &+\frac{r_i\sigma_\alpha^2}{r_i\sigma_\alpha^2 + \sigma_\epsilon^2}\left[\bar{y}_{ri} - \text{BLUE}(\beta)\right]\end{aligned} \tag{8}$$

where BLUE(β) denotes the best linear estimator of β under model (6), given by $\text{BLUE}(\beta) = \frac{\sum_s \frac{r_i\bar{y}_{r_i}}{r_i\sigma_\alpha^2+\sigma_\epsilon^2}}{\sum_s \frac{r_i}{r_i\sigma_\alpha^2+\sigma_\epsilon^2}}$ and $\bar{y}_{ri} = \frac{\sum_{s_{r_i}} w_{ij}y_{ij}}{\sum_{s_{r_i}} w_{ij}}$. The EBLUP of $\beta+\alpha_i$ is then obtained by estimating unknown quantities in (8). In what follows, we use the weighted mean of respondents $\hat{\beta}_r$, as imputed values.

4 Variance estimation

To estimate the variance of the imputed estimator (7), we use the reverse approach of Fay (1991). To apply the reverse approach, we first express $\hat{Y}_I$ as $\hat{Y}_I = \hat{T}\hat{R}_a$, where $\hat{R}_a = \frac{\hat{Y}_a}{\hat{T}_a}$, with $\hat{Y}_a = \sum_{i\in s}\sum_{j\in s_i} w_{ij}a_{ij}y_{ij}$ and $\hat{T}_a = \sum_{i\in s}\sum_{j\in s_i} w_{ij}a_{ij}$, $\hat{T} = \sum_{i\in s}\sum_{j\in s_i} w_{ij}$, $a_{ij} = 1$ if unit j in cluster i belongs to s_{r_i} and $a_{ij} = 0$ otherwise .It follows from (5) that the variance $V(\hat{Y}_I)$ of $\hat{Y}_I$ can be estimated by $v_t = v_1 + v_2$, where v_1 is an estimator of $V_p(\hat{Y}_I) = E_1V_2(\hat{Y}_I) + V_1E_2(\hat{Y}_I)$, conditional on the a_{ij}'s and v_2 is an estimator of $\widetilde{V}_m\left[E_p\left(\hat{Y}_I - Y\right)\right] = \widetilde{V}_m\left[E_1E_2\left(\hat{Y}_I - Y\right)\right]$. Denote the estimator of the variance of $\hat{Y} = \sum_{i\in s}\sum_{j\in s_i} w_{ij}y_{ij}$ based on the full sample as $v(y_{ij})$. Then, one can show, using Taylor linearization, that v_1 reduces to

$$v_1 = v(\hat{\xi}_{ij}), \tag{9}$$

where

$$\hat{\xi}_{ij} = a_{ij}y_{ij} + (1 - a_{ij})\,\hat{R}_a + \frac{\left(\hat{T} - \hat{T}_a\right)}{\hat{T}_a}a_{ij}\left(y_{ij} - \hat{R}_a\right).$$

To obtain v_2, note that $E_p\left(\hat{Y}_I - Y\right) = \sum_{i\in U}\sum_{j\in U_i} c_{ij}y_{ij}$, where $c_{ij} = K\frac{a_{ij}}{T_a} - 1$ with $T_a = \sum_{i\in U} T_{ai}$ and $T_{ai} = \sum_{j\in U_i} a_{ij}$. Now,

$$\begin{aligned}\widetilde{V}_mE_p\left(\hat{Y}_I - Y\right) &= E_rV_mE_p\left(\hat{Y}_I - Y\right) \\ &+V_rE_mE_p\left(\hat{Y}_I - Y\right).\end{aligned} \tag{10}$$

The second term on the right hand side of (10) is 0 because $E_mE_p\left(\hat{Y}_I - Y\right) = 0$. Also, it is easy to show that

$$\begin{aligned}E_rV_mE_p\left(\hat{Y}_I - Y\right) &\approx -K(M\sigma_\alpha^2 + \sigma_\epsilon^2) \\ &+\left(\frac{K}{E_r(T_a)}\right)^2\sigma_\alpha^2\sum_{i\in U}T_{ai}^2 + \frac{K^2}{E_r(T_a)}\sigma_\epsilon^2.\end{aligned} \tag{11}$$

The component v_2 is then obtained by substituting estimators for the unknown quantities in (11). Hence, v_2 is given by

$$\begin{aligned}v_2 &\approx -K(M\hat{\sigma}_\alpha^2 + \hat{\sigma}_\epsilon^2) \\ &+\left(\frac{K}{\hat{T}_a}\right)^2\hat{\sigma}_\alpha^2\sum_{i\in s}\left[\hat{T}_{ai}^2 - \hat{V}_2\left(\hat{T}_{ai}\right)\right] \\ &+\frac{K^2}{\hat{T}_a}\hat{\sigma}_\epsilon^2,\end{aligned} \tag{12}$$

where $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\epsilon^2$ are estimates of σ_α^2 and σ_ϵ^2 respectively that can be obtained using available methods such as ML, REML or MINQUE methods. Note that since $V_2\left(\hat{T}_{ai}\right) = E_2\left(\hat{T}_{ai}^2\right) - E_2\left(\hat{T}_{ai}\right)^2$, an estimate of T_{ai}^2 is given by $\hat{T}_{ai}^2 - \hat{V}_2\left(\hat{T}_{ai}\right)$, where $\hat{V}_2\left(\hat{T}_{ai}\right)$ is an estimate of $V_2\left(\hat{T}_{ai}\right)$, $\hat{T}_{ai} = \sum_{s_i} w_{j|i} a_{ij}$ and $w_{j|i}$ is the survey weight of unit j at the second stage given that cluster i has been selected in the first stage sample. The sum of (9) and (12) gives v_t.

5 Comparisons

In this section, we compare the magnitude of the second term $\widetilde{V}_m E_p\left(\hat{Y}_I - Y\right)$ under the "correct" model (6) and the "wrong" model (1), assuming single-stage cluster sampling. Note that under mean imputation, model (1) reduces to

$$\begin{aligned} E_m(y_i) = \beta, V_m(y_i) = \sigma^2, \\ Cov_m(y_i, y_j) = 0 \text{ if } i \neq j. \end{aligned} \tag{13}$$

Under this model, it is easy to show that $\widetilde{V}_m E_p\left(\hat{Y}_I - Y\right)$ is given by

$$\widetilde{V}_m E_p\left(\hat{Y}_I - Y\right) = \sigma^2 K\left(\frac{K}{E_r T_a} - 1\right). \tag{14}$$

We are interested in the ratio $R = \frac{\widetilde{V}_{m.correct} E_p\left(\hat{Y}_I - Y\right)}{\widetilde{V}_{m.wrong} E_p\left(\hat{Y}_I - Y\right)}$ where $\widetilde{V}_{m.correct} E_p\left(\hat{Y}_I - Y\right)$ is given in (11) and $\widetilde{V}_{m.wrong} E_p\left(\hat{Y}_I - Y\right)$ is given in (14). To determine this ratio, we need to specify a response mechanism. In this paper, we consider two distinct response mechanisms: the uniform response mechanism and the "beta-binomial" response mechanism.

5.1 The uniform response mechanism

In this section, we assume that the response mechanism is uniform; that is, we assume that $P(a_{ij} = 1) = p_1\ \forall i, j$ and that the a_{ij}'s are independent random variables, so that $E_r(a_{ij}) = p_1$. It is then easy to show that under this response mechanism, $\widetilde{V}_{m.correct} E_p\left(\hat{Y}_I - Y\right)$ given in (11) and $\widetilde{V}_{m.wrong} E_p\left(\hat{Y}_I - Y\right)$ given in (14) reduce respectively to

$$\widetilde{V}_{m.correct} E_p\left(\hat{Y}_I - Y\right) \approx p_1^{-1}(1-p_1) K\left(\sigma_\epsilon^2 + \sigma_\alpha^2\right), \tag{15}$$

and

$$\widetilde{V}_{m.wrong} E_p\left(\hat{Y}_I - Y\right) \approx p_1^{-1}(1-p_1) K \sigma^2. \tag{16}$$

Noting that σ^2 in (13) is equal to $\sigma_\epsilon^2 + \sigma_\alpha^2$, it is easy to see that $R = 1$ which suggests that the uniform response assumption "washes off" the effect of the model. Hence, under uniform response, the choice of the model is irrelvant. Note that the order for both (15) and (16) is $O(K)$. Also, note that if there is no nonresponse (i.e., $p_1 = 1$), both (15) and (16) are equal to 0, as expected.

5.2 The "beta-binomial" response mechanism

Uniform response is a strong assumption that may not be tenable in the case of cluster sampling. A more realistic response mechanism may be a beta-binomial mechanism. We assume that $E_r(a_{ij}) = p_1$ and $E_r(a_{ij} a_{ij'}) = p_2$. If ρ_p represents the correlation between the response indicator between two units in the same cluster, one can show that p_2 can be expressed as $p_2 = \rho_p p_1 + (1-\rho_p) p_1^2$. It is then easy to show that under this mechanism, $\widetilde{V}_{m.wrong} E_p\left(\hat{Y}_I - Y\right)$ is given by (16) and $\widetilde{V}_{m.correct} E_p\left(\hat{Y}_I - Y\right)$ is given by

$$\begin{aligned} \widetilde{V}_{m.correct} E_p\left(\hat{Y}_I - Y\right) \approx -K\left(M\sigma_\alpha^2 + \sigma_\epsilon^2\right) \\ + K(M-1)\sigma_\alpha^2 \frac{p_2}{p_1^2} + \frac{K}{p_1}\left(\sigma_\alpha^2 + \sigma_\epsilon^2\right). \end{aligned} \tag{17}$$

One can then show that the ratio R is given by

$$R = 1 + (M-1)\rho\rho_p, \tag{18}$$

where ρ is the intracluster correlation coefficient. Note first that under uniform response $p_2 = p_1^2$ (or $\rho_p = 0$), (18) reduces to 1. Second, note also that the order of (17) is $O(KM)$ as opposed to (16) which is only $O(K)$. Third, note that the ratio R increases as M, ρ and ρ_p increase. Finally, note that under full response $p_1 = p_2 = 1$, (17) is equal to 0, as expected. Table 1 gives the magnitude of R with $M = 20$ for different values of ρ and ρ_p.

Table 1
Magnitude of R for different values of ρ and ρ_p for $M = 20$

$\rho\backslash\rho_p$	0.1	0.3	0.5	0.7	0.9
0.1	1.19	1.57	1.95	2.33	2.71
0.3	1.57	2.71	3.85	4.99	6.13
0.5	1.95	3.85	5.75	7.65	9.55
0.7	2.33	4.99	7.65	10.31	12.97
0.9	2.71	6.13	9.55	12.97	16.39

5.3 Choice of the model

Results in Table 1 shows that choosing the wrong model may lead to appreciable underestimation of the second component $\widetilde{V}_m E_p\left(\hat{Y}_I - Y\right)$, especially for large M, which might suggest that, in the case of single-stage or two-stage cluster sampling, one should use model (6) and not model (13). It should be noted however that as M increases, ρ and ρ_p typically decrease.

Under uniform response, the first component $\widetilde{E}_m V_p\left(\hat{Y}_I - Y\right)$ is given by

$$\widetilde{E}_m V_p\left(\hat{Y}_I - Y\right) \approx NK\left(\frac{1}{n} - \frac{1}{N}\right) \times \left[p_1^{-1}\left(\sigma_\alpha^2 + \sigma_\epsilon^2\right) + (M-1)\sigma_\alpha^2\right]. \quad (19)$$

The order of the first component in (19) is thus $O\left(\frac{K^2}{n}\right)$. We have noted in section 5.1 that, under uniform response, the order of the second component $\widetilde{V}_m E_p\left(\hat{Y}_I - Y\right)$ is $O(K)$ so the order of $\frac{\widetilde{V}_m E_p\left(\hat{Y}_I - Y\right)}{\widetilde{E}_m V_p\left(\hat{Y}_I - Y\right)}$ is $O(\frac{n}{K})$. Typically, $n << K$ in cluster sampling so the second component is negligible relative to the first component. In this case, the computation of v_2 may be omitted. This may, however, be not true for the beta-binomial response mechanism. Under this response mechanism, one can show that the first component $\widetilde{E}_m V_p\left(\hat{Y}_I - Y\right)$ is given by

$$\widetilde{E}_m V_p\left(\hat{Y}_I - Y\right) \approx \frac{K^2}{A} N\left(\frac{1}{n} - \frac{1}{N}\right) \times \left(\left[\left(\sigma_\alpha^2 + \sigma_\epsilon^2\right) p_1 + (M-1)\sigma_\alpha^2 p_2\left(1 + \frac{p_1 - p_2}{A}\right)\right] + \left[\frac{K-M}{N-1}\left(p_2 - p_1^2\right) \times \left(\frac{\sigma_\alpha^2 + \sigma_\epsilon^2}{K p_1} + \frac{(M-1)\sigma_\alpha^2 p_2}{A}\right)\right]\right), \quad (20)$$

where $A = M\left(p_2 - p_1^2\right) + K p_1^2$. First, note that under uniform response $p_2 = p_1^2$, (20) reduces to (19), assuming $Kp_1 + (1 - p_1) \approx Kp_1$. Second, note that the order (18) is $O\left(\frac{K^2}{n}\right)$. Third, note that under full response $p_1 = p_2 = 1$, (20) reduces to $\widetilde{E}_m V_p\left(\hat{Y}_I - Y\right) \approx NK\left(\frac{1}{n} - \frac{1}{N}\right)\left(M\sigma_\alpha^2 + \sigma_\epsilon^2\right)$, as expected . Finally, in section 5.2, we have noted that under the beta-binomial response mechanism, the second component $\widetilde{V}_m E_p\left(\hat{Y}_I - Y\right)$ was of order $O(KM)$ and therefore the order of $\frac{\widetilde{V}_m E_p\left(\hat{Y}_I - Y\right)}{\widetilde{E}_m V_p\left(\hat{Y}_I - Y\right)}$ is $O(\frac{n}{N})$. The second component is hence negligible only if the first-stage sampling fraction is negligible which might not be the case in practice. Indeed, it not unusual, in cluster sampling, to have first-stage sampling fractions as high as 0.5. In this case, computation of the second component must be performed and the choice of the imputation model may then become relevant. We now investigate the magnitude of the underestimation of the total variance occuring when using the wrong model. A measure of the underestimation, denoted by Q, is defined as

$$Q = 1 - \frac{\widetilde{E}_{m.correct} V_p\left(\hat{Y}_I - Y\right) + \widetilde{V}_{m.wrong} E_p\left(\hat{Y}_I - Y\right)}{\widetilde{E}_{m.correct} V_p\left(\hat{Y}_I - Y\right) + \widetilde{V}_{m.correct} E_p\left(\hat{Y}_I - Y\right)},$$

where $\widetilde{E}_{m.correct} V_p\left(\hat{\theta} - \theta\right)$ is given by (20). Note that the first component of the total variance is computed under the correct model since it is valid regardless of the model and/or the response mechanism. Tables 2, 3 and 4 show the magnitude of the underestimation for different values of ρ, ρ_p, with $M = 20$, $p_1 = 0.7$ and $\frac{n}{N} = 0.1, 0.25$ and 0.5. The following conclusions may the be drawn: For modest and small first-stage sampling fractions, the underestimation is small to negligible; in this case, the choice of the model may not be relevant. As the first-stage sampling fraction $\frac{n}{N}$ increases, the underestimation increases for a given ρ and may be quite substantial (e.g., $\frac{n}{N} = 0.1$, $\rho = 0.3 \Longrightarrow 0.36\% \leq Q \leq 2.54\%$ and for $\frac{n}{N} = 0.5$, $\rho = 0.3 \Longrightarrow 3.03\% \leq Q \leq 18.40\%$). For a given $\frac{n}{N}$ and ρ, Q increases as ρ_p increases; (e.g., $\frac{n}{N} = 0.5$, $\rho = 0.5 \Longrightarrow 3.34\% \leq Q \leq 19.61\%$). This indicates that departure from uniform response leads to underestimation of the total variance. For a given $\frac{n}{N}$ and ρ_p, Q increases as ρ increases; (e.g., $\frac{n}{N} = 0.5$, $\rho_p = 0.5 \Longrightarrow 8.90\% \leq Q \leq 13.93\%$). Hence, correct specification of the model becomes critical when $\frac{n}{N}$, ρ and ρ_p increase. Finally, we investigate the relative magnitude of the underestimation, Q', occuring when the computation of the second component is not performed and given by

$$Q' = 1 - \frac{\widetilde{E}_{m.correct} V_p \left(\hat{Y}_I - Y\right)}{\widetilde{E}_{m.correct} V_p \left(\hat{Y}_I - Y\right) + \widetilde{V}_{m.correct} E_p \left(\hat{Y}_I - Y\right)}.$$

Tables 5, 6 and 7 shows the magnitude of Q' for different values of ρ, ρ_p, with $M = 20$, $p_1 = 0.7$ and $\frac{n}{N} = 0.1, 0.25$ and 0.0.5. These tables show that for a given ρ and ρ_p, Q' increases as $\frac{n}{N}$ increases. Also, for large $\frac{n}{N}$ and given ρ_p, Q' decreases as ρ increases because, as ρ increases, the numerator of Q' increases faster than the second component in the denominator of Q' (Tables 6 and 7).

Table 2
Magnitude of $Q \times 100\%$ with $\frac{n}{N} = 0.1$

$\rho \backslash \rho_p$	0.1	0.3	0.5	0.7	0.9
0.1	0.26	0.74	1.18	1.57	1.94
0.3	0.36	1.01	1.58	2.09	2.54
0.5	0.39	1.09	1.70	2.23	2.70
0.7	0.41	1.13	1.76	2.30	2.78
0.9	0.42	1.16	1.79	2.34	2.82

Table 3
Magnitude of $Q \times 100\%$ with $\frac{n}{N} = 0.25$

$\rho \backslash \rho_p$	0.1	0.3	0.5	0.7	0.9
0.1	0.75	2.14	3.37	4.48	5.49
0.3	1.07	2.95	4.57	5.96	7.18
0.5	1.16	3.20	4.92	6.38	7.65
0.7	1.21	3.32	5.08	6.58	7.87
0.9	1.24	3.40	5.18	6.70	8.00

Table 4
Magnitude of $Q \times 100\%$ with $\frac{n}{N} = 0.5$

$\rho \backslash \rho_p$	0.1	0.3	0.5	0.7	0.9
0.1	2.07	5.75	8.90	11.63	14.03
0.3	3.03	8.11	12.21	15.57	18.40
0.5	3.34	8.84	13.20	16.70	19.61
0.7	3.49	9.19	13.66	17.24	20.18
0.9	3.58	9.40	13.93	17.55	20.52

Table 5
Magnitude of $Q' \times 100\%$ with $\frac{n}{N} = 0.1$

$\rho \backslash \rho_p$	0.1	0.3	0.5	0.7	0.9
0.1	1.64	2.05	2.43	2.77	3.09
0.3	1.01	1.72	2.15	2.62	3.04
0.5	1.82	1.49	2.07	2.58	3.03
0.7	1.72	1.43	2.03	2.56	3.02
0.9	1.64	1.39	2.01	2.55	3.02

Table 6
Magnitude of $Q' \times 100\%$ with $\frac{n}{N} = 0.25$

$\rho \backslash \rho_p$	0.1	0.3	0.5	0.7	0.9
0.1	4.74	5.90	6.93	7.86	8.70
0.3	2.95	4.69	6.17	7.46	8.58
0.5	2.39	4.32	5.95	7.34	8.55
0.7	2.12	4.15	5.85	7.29	8.53
0.9	1.96	4.05	5.78	7.26	8.52

Table 7
Magnitude of $Q' \times 100\%$ with $\frac{n}{N} = 0.5$

$\rho \backslash \rho_p$	0.1	0.3	0.5	0.7	0.9
0.1	13.00	15.84	18.27	20.38	22.24
0.3	8.35	12.86	16.49	19.48	21.98
0.5	6.86	11.95	15.96	19.21	21.90
0.7	6.12	11.50	15.71	19.09	21.87
0.9	5.68	11.24	15.56	19.02	21.85

6 Summary and Conclusion

In this article, we have discussed variance estimation in cluster sampling under imputation for missing data. For simplicity, we have considered mean imputation but extension to ratio imputation has also been investigated. We have proposed to use nested error linear regression models as imputation models in order to take account of the intracluster correlation. Using a method developed by Shao and Steel (1999), we have derived consistent estimators for the variance of the imputed estimator. In the case of single-stage cluster sampling, we have compared the effect of using a linear model with fixed effects and a diagonal variance-covariance matrix as an imputation model instead of a random effect model; we have shown that the choice of the model becomes relevant under the beta-binomial response mechanism when the first-stage sampling fraction is large.

References

[1] Deville, J. C. and Särndal, C. E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.

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

[3] Rancourt, E., Särndal, C. E. and Lee, H. (1994). Estimation of the variance in presence of nearest neighbor imputation. *Survey Methodology*, 20, pp.137-147.

[4] Särndal, C. E. (1992). Bias Corrections for Survey Estimate from Data with Ratio Imputed Values for Confounded Nonresponse*Survey Methodology*, 18, 241-252.

[5] Shao, J. and Steel, P. (1999). Variance Estimation for Survey Data With composite Imputation and Nonnegligible Sampling Fractions. *Journal of the American Statistical Association*, 94, 254-265.