Linearization Variance Estimation and Allocation for Two-phase Sampling under Mass Imputation

A. Demnati and J.N.K. Rao

Statistics Canada Carleton University
Business Survey Methods Division School of Mathematics and Statistics
Ottawa, Canada, K1A 0T6 Ottawa, Canada, K1S 5B6
Abdellatif.Demnati@statcan.gc.ca JRao@math.carleton.ca

Abstract

We consider two-phase sampling in which values of a variable of interest are observed only in the second-phase sub-sample. Values for the first-phase units not sampled in the second-phase are mass imputed, using values from an administrative file when available and regression imputation otherwise. Such two-phase sampling methods are often used in annual business surveys to reduce survey costs and respondent burden, assuming that collecting values from administrative sources is much cheaper than obtaining values through questionnaires from sampled units. We study both naïve and design-consistent estimators for a total or mean under the above set-up. We also obtain associated variance estimators using a unified approach proposed by Demnati and Rao (2004). Simulation results on the finite sample performance of the estimators and associated variance estimators are also presented, using substitution or ratio mass imputation. We also study the case of missing sub-sample values and develop estimators of the total and associated variance estimators. Sample allocation issues are also studied.

KEY WORDS: Regression imputation; Sample sizes determination; Substitution method; Two-phase sampling.

1. Introduction

At Statistics Canada, more than sixty annual business surveys are carried out using a holistic design termed as the Unified Enterprise Survey (UES). One of the objectives of the UES is to reduce survey cost and response burden by replacing some questionnaires with values from administrative data. We assume that obtaining values from administrative data is much cheaper than obtaining values through questionnaires. A random sample of size \( m \) units is first selected. Then a sub-sample of size \( n \) is selected and surveyed. The remaining \( m - n \) units as well as non-respondents are imputed using administrative files. Composite imputation involving two or more different methods is also often used; for example, the values from administrative file (e.g., tax file), when available, and regression imputation otherwise.

Demnati and Rao (2008) considered the case of simulated census data generated from a probability sample by imputing for the non-sampled units and sample non-respondents using auxiliary variables. They studied the estimation of a finite population total and other parameters. In this paper, we extended their work to two-phase sampling. For variance estimation, we have used the unified approach proposed by Demnati and Rao (2004). We have studied naïve estimators based on the simulated first-phase data as well as design-consistent estimators of a population total. In section 2, we consider the case of complete response and report the results of a simulation study. The case of missing data is studied in section 3. In section 4, we consider stratified simple random sampling in both phases and obtain “optimal” first-phase and second-phase strata sample sizes, \( m_1 \) and \( n_1 \), that minimize the cost subject to constraints on the variances of estimators for one or more characteristics of interest.

2. Complete Response
2.1 Imputed Estimators

We first consider the case where complete responses are obtained from the second-phase subsample. Imputation is performed on all the first-phase units not sampled in the second-phase. Here, the imputed values $\hat{y}^*_i$ are given by $\hat{y}^*_i = (1 - I_i)\hat{y}_i + I_i\hat{y}_i^*$ with $\hat{y}_i^* = t_i$ and $\hat{y}_{2a}^*$ based on an imputation model, where $t_i$ is the value from an administrative file and the constant $I_i$ is the missing $t_i$ indicator.

A naïve estimator of the finite population total $Y = \sum y_i$ is given by

$$\hat{Y}^{(N)} = \sum d_i^{(1)}[a_i^{(2)}y_i + (1 - a_i^{(2)})\hat{y}_i^*],$$

(2.1)

where $d_i^{(1)} = a_i^{(1)}/\pi_i^{(1)}$, $a_i^{(1)}$ is the first-phase sample membership indicator variable, $\pi_i^{(1)} = E_\pi(a_i^{(1)})$, $E_\pi$ denotes design expectation, and $a_i^{(2)}$ is the conditional second-phase sample membership indicator variable.

Suppose the imputation model on the responses $y_i$ is specified by a generalized linear model with mean $E_\pi(y_i) = \mu_i(\beta) = h(x_i^T\beta)$, where $x_i$ is a $p \times 1$ vector of explanatory variables for the model mean, $h()$ is a “link” function and $E_\pi$ denotes model expectation. For example, the choice $h(a) = a$ gives a linear regression model, and $h(a) = e^a/(1+e^a)$ gives a logistic regression model for binary responses $y_i$. The imputed values $\hat{y}_{2a}^*$ are given by $\hat{y}_{2a}^* = \mu_i(\hat{\beta}^{(N)})$, where the naïve estimator $\hat{\beta}^{(N)}$ is obtained as the solution to the estimating equations

$$\hat{I}_\beta(\beta) = \sum a_i I_i \Phi_i(\beta)(y_i - \mu_i(\beta)) = 0,$$

(2.2)

$a_i = a_i^{(1)}a_i^{(2)}$ and $\Phi_i(\beta) = [\partial \mu_i(\beta)/\partial \beta][Var_\pi(y_i)]^{-1}$. For example, $\Phi_i(\beta) = x_i$ for linear and logistic regression models.

The finite population parameter $\theta_s$ induced by the estimator $\hat{Y}^{(N)}$ is given by

$$\theta_s = E_\pi(\hat{Y}^{(N)}) \approx \sum \pi_i^{(2)}y_i + \sum (1 - \pi_i^{(2)})\hat{y}_i^*,$$

(2.3)

where $\hat{y}_i^* = E_\pi(\hat{Y}_i^*)$. It is clear from (2.3) that $\theta_s$ depend on the selection probabilities. For example, the use of $\hat{Y}^{(N)}$ under simple random sampling at both stages induces a mixture of two totals as finite population parameter: $\theta_s = f^{(2)}\sum y_i + (1 - f^{(2)})\sum \hat{y}_i^*$, where $f^{(2)}$ is the conditional sampling fraction for the second-phase sample.

The sampling bias induced by the estimator $\hat{Y}^{(N)}$ in estimating the finite population total $Y$ is given by

$$\theta_s - Y \approx -\sum (1 - \pi_i^{(2)})(y_i - \hat{y}_i) \equiv \hat{B}.$$

(2.4)

In order to remove the sampling bias, one may first estimate the bias (2.4) by

$$\hat{B} = -\sum d_i(1 - \pi_i^{(2)})(y_i - \hat{y}_i^*),$$

(2.5)

and then adjust $\hat{Y}^{(N)}$ to get the bias-adjusted estimator

$$\hat{Y} = \sum d_i y_i + \sum (d_i^{(1)} - d_i)\hat{y}_i^*,$$

(2.6)

with $\hat{y}_{2a}^* = \mu_i(\hat{\beta})$, the estimator $\hat{\beta}$ is obtained as solution to the estimating equations

$$\hat{I}_\beta(\beta) = \sum a_i I_i \Phi_i(\beta)(y_i - \mu_i(\beta)) = 0,$$

(2.7)

where $d_i = a_i^{(1)}/\pi_i$ and $\pi_i = E_\pi(a_i)$. The design-based estimator $\hat{Y}$ is approximately unbiased for $Y$: $E_\pi(\hat{Y}) \approx Y$.

We may also write $\hat{Y}$ as

$$\hat{Y} = \sum d_i^{(1)}[a_i^{(2)}y_i + (1 - a_i^{(2)})\hat{y}_i^*].$$

2.2 Variance Estimation
We suppose first that the parameter of interest is \( \theta_k = E(Y_{ik}^*) \). Let \( d_k = (d_{k1}, d_{k2})' \), where \( d_{k1} = d_k^0 \), and \( d_{k2} = d_k' \). The Demnati–Rao (DR) variance estimator (Demnati and Rao, 2004) is simply given by
\[
\mathcal{G}_{dr}(\hat{Y}^{(s)}) = \mathcal{G}(z),
\]
where \( \mathcal{G}(u) \), in operation notation, is the variance estimator of the linear combination \( \hat{U} = \sum u_i d_i \), \( u_i = (u_{i1}, u_{i2})' \) a vector of constants, \( z_i = \partial f(A_i) / \partial b_i \mid_{A_i = A_i} \), \( f(A_i) = \hat{Y}^{(s)} \), \( A_i \) is a \( 2 \times N \) matrix with \( k^m \) column \( d_i \), and \( A_i \) is a \( 2 \times N \) matrix of arbitrary real numbers with \( k^m \) column \( b_i = (b_i^0, b_i^1)' \). The design based variance estimator \( \mathcal{G}(u) \) of the total \( \hat{U} \) is given by
\[
\mathcal{G}(u) = \sum_i \sum_k u'_i \text{cov}_p(d_i, d_i')u_i,
\]
with
\[
\text{cov}_p(d_i, d_i') = \text{diag}(d_i) \begin{bmatrix} 1 - \omega^{(s)}_{ii} & 1 - \omega^{(s)}_{ik} \\ 1 - \omega^{(s)}_{ik} & 1 - \omega^{(s)}_{kk} \end{bmatrix} \text{diag}(d_i),
\]
where \( \omega^{(s)}_{ii} = \pi^{(s)}_{i1} / \pi^{(s)}_i \), \( \omega^{(s)}_{ik} = \pi^{(s)}_i / \pi^{(s)}_i \), and \( (\pi^{(s)}_i, \pi^{(s)}_i) \) denote respectively the first and two phase joint inclusion probabilities. Substituting (2.10) into (2.9) we get
\[
\mathcal{G}(u) = \mathcal{G}^{(s)}(u_i, u_1) + 2c_i(u_i, u_1) + \mathcal{G}(u_i, u_2),
\]
with
\[
\mathcal{G}^{(s)}(x, y) = \sum \sum d_i^3 d_i^3 (1 - \omega^{(s)}_{ii}) x_i y_i,
\]
\[
c_i(x, y) = \sum \sum d_i^3 d_i^3 (1 - \omega^{(s)}_{ii}) x_i y_i,
\]
and
\[
\mathcal{G}(x, y) = \sum d_i^3 d_i^3 (1 - \omega^{(s)}_{ii}) x_i y_i.
\]
We may write \( c_i(x, y) \) as
\[
c_i(x, y) = \sum d_i^3 d_i^3 (1 - \omega^{(s)}_{ii}) x_i y_i = \mathcal{G}^{(s)}(x, y),
\]
where \( v_i = d_i^2(y_i) \) and \( d_i^2 = d_i^2 / \pi^{(s)}_i \) is the conditional second-phase weight. It is seen from (2.15) that \( c_i(x, y) \) is computed from the first-phase sampling variance estimator.

It remains to evaluate \( z_i \). We have \( f(A_i) = \sum b_i \pi^{(s)}_i y_i + \sum (b_i^1 - b_i \pi^{(s)}_i) \hat{Y}^{(s)}(b) \) with \( \hat{Y}^{(s)}(b) = (1 - I_i) \tau_i + I_i \mu_i(\hat{\beta}^{(s)}(b)) \) and \( \hat{I}^{(s)}(b) = \sum b_i \pi^{(s)}_i I_i \Phi_i(\hat{\beta}(b))(y_i - \mu_i(\beta(b))) = 0 \), where \( b \) is the second column of the matrix \( A_i \). Hence,
\[
z_i = \partial f(A_i) / \partial b_i \mid_{A_i = A_i} = \begin{bmatrix} \hat{Y}^{(s)}_i \\ \hat{p}^{(s)}_i \end{bmatrix},
\]
with
\[
\hat{p}^{(s)}_i = \pi^{(s)}_i - \pi^{(s)}_i \hat{J}^{(s)}(\beta) \hat{p}^{(s)}(\beta) \hat{J}^{(s)}(\beta)' \Phi_i(\hat{\beta}) I_i,
\]
where \( \hat{J}^{(s)}(\beta) = -\partial \hat{Y}^{(s)}(\beta) / \partial \beta \) and \( \hat{p}^{(s)}(\beta) = -\partial \hat{p}^{(s)}(\beta) / \partial \beta \).

Under the linear model
\[
\mu_i = x_i \beta \quad \text{and} \quad \text{Var}_n(y_i) = \sigma^2 / c_i,
\]
for some specified constants \( c_i \), \( \Phi_i(\beta) = x_i c_i \), \( \beta \) is estimated by \( \hat{\beta}^{(s)} = \hat{Q}^{(s)} \sum a_i I_i c_i x_i y_i \) with \( \hat{Q}^{(s)} = \sum a_i I_i c_i x_i y_i \). We have \( \hat{J}^{(s)}(\beta) = \sum d_i^3 (1 - a_i^2) I_i x_i \), \( \hat{p}^{(s)}(\beta) = \hat{Q}^{(s)} \), and
\( g_i^{(N)} = \pi_i^{(2N)} + \pi_i \sum d_i^{(N)} (1 - a_i^{(N)}) l, x_i^{(N)} \hat{Q}_i^{(N)} x_i c_i l_i. \)  

A variance estimator of the bias-adjusted estimator \( \hat{Y} \) is given by (2.9) with \( u_i \) replaced by 

\[
    z_i = \partial f(A_i) / \partial b_i \bigg|_{b_i = \hat{b}_i} = \begin{bmatrix} \hat{y}_i \n \hat{y}_i \end{bmatrix} \left( \hat{g}_i (y_i - \hat{y}_i) \right),
\]

where 

\[
    \hat{g}_i = 1 - \hat{J}_b (\beta) (\hat{J}_b (\hat{\beta}))^{\dagger} \Phi_i (\hat{\beta}) l_i,
\]

\( f(A_i) = \hat{Y}, \quad \hat{J}_b (\beta) = -\partial \hat{Y} / \partial \beta, \) and \( \hat{J}_b (\hat{\beta}) = -\hat{J}_b (\beta) / \partial \beta. \) Under the linear model, we have \( \hat{y}_i = x_i^{\dagger} \hat{\beta}, \)

\( \hat{\beta} = \hat{Q}^{\dagger} \sum d_i l_i c_i x_i y_i \) with \( \hat{Q} = \sum d_i l_i c_i x_i x_i^{\dagger}, \) and 

\[
    g_i = 1 + \sum (d_i^{(N)} - d_i) l, x_i^{\dagger} \hat{Q}^{\dagger} x_i c_i l_i. \]

### 2.3 Simulation Study

We conducted a small simulation study to examine the performances of the estimators \( \hat{Y} \) and \( \hat{\theta} \) and associated variance estimators, when substitution imputation or ratio imputation is used. The more general case of values from administrative file when available and regression imputation otherwise is not studied in this section. We first generated a finite population \( \{ y_1, \ldots, y_N \} \), with \( y_i = (y_{i1}, y_{i2}, y_{i3}) \), of size \( N = 393 \) from the following models: 

\( y_{i1} = x_i + x_i^{1/2} \epsilon_i, \quad y_{i2} = 1.2 x_i + x_i^{1/2} \epsilon_i, \) and \( y_{i3} = 5 + 1.2 x_i + x_i^{1/2} \epsilon_i, \) where \( \epsilon_i \) are independent observations generated from \( \mathcal{N}(0,1) \), and the fixed \( x_i \) are the “number of beds” for the Hospitals population studied in Valliant et al. (2000, p.424-427). We stratified the population into two strata with 272 units \( k \) having \( x_i \leq 350 \) in stratum 1 and 121 units \( k \) with \( x_i > 350 \) in stratum 2. We selected \( R = 5,000 \) stratified two-phase simple random samples of sizes \( m_1 = m_2 = 50 \) and \( n_1 = n_2 = 15 \). Our vector parameter of interest is the finite population total \( \theta_s = (Y_1, Y_2, Y_3)^{\dagger}. \) Values for the first-phase units not sampled in the second phase are mass imputed. For each variable \( l, l = 1,2,3, \) four estimators are considered. Two estimators used the substitution imputation method: the naïve estimator \( \hat{\theta}_s^{(N)} = \sum d_i^{(N)} (1 - a_i) x_i \), and the corresponding design-based estimator \( \hat{\theta}_d^{(N)} = \sum d_i y_i + \sum d_i^{(N)} - d_i x_i, \) \( l = 1,2,3. \) The other two estimators used the ratio imputation method: the naïve estimator \( \hat{\theta}_r^{(N)} = \sum d_i^{(N)} (1 - a_i) x_i \hat{\beta}_i^{(N)} \), and the associated design-based estimator \( \hat{\theta}_r = \sum d_i y_i + \sum d_i^{(N)} - d_i x_i \hat{\beta}_i, \) where \( \hat{\beta}_i^{(N)} = \sum a_i y_i / \sum a_i x_i \) and \( \hat{\beta}_i = \sum d_i y_i / \sum d_i x_i \). The estimator \( \hat{\theta}_r \) reduces to the two-phase ratio estimator: 

\( \hat{y}_s = \hat{Y}(\hat{X}^{(N)}) / \hat{X}. \)

Let \( \hat{\theta} \) denote an estimator of a population total \( \theta_s \) and \( \mathcal{H}(\hat{\theta}) \) be the associated variance estimator. We calculated the simulated relative bias of \( \hat{\theta} \) and \( \mathcal{H}(\hat{\theta}) \) as \( \text{RB}(\hat{\theta}) = (\hat{\theta} - \theta_s) / \theta_s \) and \( \text{RB}(\mathcal{H}(\hat{\theta})) = \{ \mathcal{H}(\hat{\theta}) - \text{MSE}(\hat{\theta}) \} / \text{MSE}(\hat{\theta}) \), where \( \hat{\theta} = R^{-1} \sum_{r=1}^{R} \hat{\theta}_r \) is the mean of the estimates \( \hat{\theta}_r \) from the simulated samples \( r = 1, \ldots, R \) and \( \text{MSE}(\hat{\theta}) = R^{-1} \sum_{r=1}^{R} (\hat{\theta}_r - \theta_s)^2 \) is the simulated mean squared error (MSE). We calculated \( \text{RB}(\hat{\theta}), \text{RB}(\mathcal{H}(\hat{\theta})) \) and \( \text{MSE}(\hat{\theta}) \) for each component of the vector \( (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)^{\dagger} \) with \( \hat{\theta}_i = (\hat{\theta}_i^{(N)}, \hat{\theta}_d^{(N)}, \hat{\theta}_r^{(N)}, \hat{\theta}_d). \)

It is clear from Table 1 that the substitution naïve estimator \( \hat{\theta}_s^{(N)} \) can lead to large relative bias and loss in efficiency when the model generating the variable of interest is different from the model generating the imputed values. For example, \( \hat{\theta}_s^{(N)} \) performs poorly for the variables 2 and 3. On the other hand, the design based estimator \( \hat{\theta} \) performs well regardless of the population model and imputation method. The variance estimator \( \hat{\theta}_d \) for \( \hat{\theta}_s^{(N)} \) leads to serious
underestimation of MSE for variables 2 and 3. On the other hand, variance estimator \( \mathcal{Q}_n \) for \( \dot{\theta} \) performs well with small \( RB \). The naïve estimator \( \hat{\theta}_n^{(N)} \) under ratio imputation is more robust to model deviations than \( \hat{\theta}_n^{(N)} \).

### 3. Missing Responses

In this section, we consider the case of missing second-phase responses, \( y \). Imputation is performed on all the first-phase missing responses. The resulting first-phase sample is complete. An estimator of the finite population total \( Y = \sum y_k \) is given by

\[
\hat{Y}(\tau) = \sum d_i \hat{\tau}_{o_i} y_k + \sum (d_i^{(i)} - d_i \hat{\tau}_{o_i}) \hat{y}_k^*,
\]

for specified \( \tau_{o_i} \), where \( o_i = 1 \) if \( y_k \) is observed, and \( o_i = 0 \) otherwise. Here, the imputed value \( \hat{y}_k^* \) is given by

\[
\hat{y}_k^* = \mu_\tau(\hat{\beta})
\]

where the estimator \( \hat{\beta} \) is obtained as solution to the estimating equations

\[
\hat{I}_p(\beta) = \sum d_i o_i \tau_{p,k} l_i \Phi_i(\beta)(y_k - \mu_k(\beta)) = 0,
\]

for specified \( \tau_{p,k} \).

In the case of complete response, the choice of \( \tau_{o,k} = \tau_{p,k} = 1 \) gives the design-based estimator \( \hat{Y} = \sum d_i y_k + \sum (d_i^{(i)} - d_i) \hat{y}_k^* \), and the choice of \( \tau_{o,k} = \tau_{p,k} = z_{k}^{(20)} \) leads to the naïve estimator \( \hat{Y}^{(N)} = \sum d_i^{(i)} x_k + (1 - a^{(20)}) \hat{y}_k^* \), studied in section 2. In the case of incomplete response, the choice of \( \tau_{o,k} = \tau_{p,k} = \xi_k^{(-1)} \) gives the design-response based estimator \( \hat{Y}_k = \sum d_i (o_i / \xi_k^*) y_k + \sum (d_i^{(i)} - d_i \xi_k^*) \hat{y}_k^* \), where \( \xi_k^* = E_x(o_i) \) and \( E_x \) denotes model expectation under response mechanism.

We suppose that the parameter of interest is \( \theta = E_p E_x(\hat{Y}(\tau)) \). Let \( d_{ik} = (d_{i1}, d_{i2}, d_{i3})^T \), where \( d_{in} = d_i^{(n)}, d_{2i} = d_i \) and \( d_{3i} = d_i o_i \). Then the Demnati–Rao (DR) variance estimator (Demnati and Rao, 2007) is simply given by

\[
\mathcal{Q}_{ra}(\hat{Y}(\tau)) = \mathcal{Q}(z),
\]

where \( \mathcal{Q}(u) \), in operation notation, is the variance estimator of the linear combination \( \hat{U} = \sum u_i^T d_i \), \( u = (u_{1i}, u_{2i}, u_{3i})^T \) a vector of constants. \( z_i = \partial f(A_i)/\partial b_i \big|_{A_i = a_i}, f(A_i) = \hat{Y}(\tau) \), \( A_i \) is a \( 3 \times N \) matrix with \( k \)-th column \( d_i \), and \( A_i \) is a \( 3 \times N \) matrix of arbitrary real numbers with \( k \)-th column \( b_i = (b_{i1}, b_{i2}, b_{i3})^T \). We have,

\[
z_i = \partial f(A_i)/\partial b_i \big|_{A_i = a_i} = \begin{bmatrix} \hat{y}_k^* \\ 0 \\ \{g_i(\tau)(y_k - \hat{y}_k^*)} \end{bmatrix}
\]

with

\[
g_i(\tau) = \tau_{o,k} - \tau_{p,k} \hat{J}_p(\hat{\beta})^{T} \Phi_x(\hat{\beta}) l_i
\]

where \( \hat{J}_p(\beta) = -\partial \hat{Y}(\tau)/\partial \beta \), and \( \hat{J}_p(\beta) = -\partial \hat{I}_p(\beta)/\partial \beta \). Under linear model with \( \tau_{o,k} = \tau_{p,k} = \tau_k \), we have

\[
\hat{p} = \hat{Q}^T \{ \sum d_i \tau_{o} l_i c_i x_i y_i \}, \hat{Q} = \sum d_i \tau_{o} l_i c_i x_i x_i^T, \text{ and } g_i(\tau) = \tau_{o} [1 + \sum (d_i^{(i)} - d_i \tau_{o}) l_i x_i^T x_i^T l_i].
\]

It remains to evaluate \( \mathcal{Q}(u) \). We have

\[
\mathcal{Q}(u) = \mathcal{Q}_1(u) + \mathcal{Q}_2(u),
\]
where \( u_{1s} = u_{is} \) and \( u_{2s} = (u_{is} u_{2s} + o_s u_{is})^T \).

The design based variance estimator \( \hat{\sigma}(u) \) is given by (2.9) with \( u \) replaced by \( u_{is} \).

The response estimated variance \( \hat{\sigma}(u) \), assuming that the order of expectation can be interchanged so that
\[ E_P E_r = E_r E_P, \]
given by
\[ \hat{\sigma}(u) = \sum d_i d_i^T \omega_s \omega_o (\hat{\varepsilon}_o - \hat{\varepsilon}_o^T) \hat{\sigma}_o u_i, \] (3.8)
where \( \hat{\varepsilon}_o = \hat{E}_o(o_s), \hat{\varepsilon}_o^T = \hat{E}_o(o_o) \), and \( \hat{E}_r \) denotes the estimator of model expectation under a model on \( o_s \). A variance estimator of \( \hat{Y}(T) \) is then given by (3.7) with \( u \) replaced by \( z_i \) given by (3.4). Note that \( \theta = Y \) if \( \hat{Y}(T) \) is unbiased for \( Y \) under the assumed response mechanism and the sampling design.

4. Determination of “Optimal” Sample Sizes

4.1 Complete Response

Let \( y_i \) be the value obtained using a questionnaire, and \( t_i \) or \( x_i \) be the value obtained from administrative sources. We assume that obtaining values from administrative sources is much cheaper than obtaining values through questionnaire. We consider the case where a stratified first-phase sample of size \( m = (m_1, \ldots, m_h, \ldots, m_s) \) of \( t_i \) or \( x_i \) is obtained, and then from each first-phase sample \( s_{hi} \), a sub-sample \( s_{h} \) of size \( n_h \) is interviewed and the response \( y_i \) is recorded. Our interest is to find the “optimal” sample sizes \( m = (m_1, \ldots, m_h, \ldots, m_s) \) and \( n = (n_1, \ldots, n_h, \ldots, n_s) \) for estimating the finite population parameter \( \theta \) that minimize the cost subject to constraints on the variances and the sample sizes.

Consider first the general form of an estimator in the complete response case:
\[ \hat{U} = \sum d_{i}^T u_{1} + \sum d_{i}^T u_{2} = \sum u_i^T d, \] (4.1)
where \( u_i = (u_{1i}, u_{2i})^T \) and \( d_i = (d_{i1}, d_{i2})^T \). If the conditional second-phase inclusion probabilities are constant, then
\[ \text{Cov}_r(d_{i}, d_{i}) = \begin{pmatrix} (1 - \omega_{i1})/\omega_{i1} & (1 - \omega_{i1})/\omega_{i1} \\ (1 - \omega_{i2})/\omega_{i2} & (1 - \omega_{i2})/\omega_{i2} \end{pmatrix}. \] (4.2)

Even when the conditional second-phase inclusion probabilities are random the right side of equation (4.2) is often used as an approximation to \( \text{Cov}_r(d_{i}, d_{i}) \). Using (4.2), the sampling variance of the estimated total \( \hat{U} \) is given by
\[ \text{Var}_r(u) = \sum \sum \text{Cov}_r(d_{i}, d_{i}) u_i. \] (4.3)

Substituting (4.2) into (4.3), we get
\[ \text{Var}_r(u) = \sum \sum [(1 - \omega_{i1})/\omega_{i1}] u_i u_i \]
\[ + \sum \sum [(1 - \omega_{i2})/\omega_{i2}] u_i u_i \]
\[ + 2 \sum \sum [(1 - \omega_{i1})/\omega_{i1}] u_i u_i. \] (4.4)

Under stratified simple random sampling (STSSRS) at both stages: \( s_{hi} \) is a simple random sample of size \( m_h \) from stratum \( h \) and \( s_{hi} \) is simple random sample of size \( n_h \) from \( s_{hi} \), we have \( \pi_{ih} = m_h / N_h, \pi_{ih} = n_h / N_h, \) and for \( k \neq t, \pi_{ih} = m_k (m_k - 1) / N_k (N_k - 1) \), and \( \pi_{ik} = n_k (n_k - 1) / N_k (N_k - 1) \). We get from (4.4),
\[ \text{Var}_r(u) = \sum i N_i \pi_i^2 (1 - m_i / N_i) (1 - n_i / N_i) S_{u_{1i} u_{1i}} + 2S_{u_{1i} u_{2i}} + \sum i N_i^2 \pi_i^2 (1 - m_i / N_i) S_{u_{1i} u_{2i}}. \] (4.5)
where \( S_{xy} = \sum J_{y} (x_i - \overline{x}_i)(y_i - \overline{y}_i)/(N_h - 1) \), \( \overline{x}_i = \sum J_{y} x_i / N_h \) and \( J_{y} \) is the stratum membership indicator variable for element \( k \).

4.2 Optimal Sample Sizes: STSRS

Consider the complete response case with \( p \) characteristics of interest \( y_1, \ldots, y_p \), and assume that the estimator \( \hat{U} \) is unbiased under the sampling design. Then, under STSRS, it follows from (4.5) that we can express \( Var_p(\hat{U}_j) \), for the \( j \)th variable \( y_j \), as

\[
Var_p(\hat{U}_j) = v_{j0} + \sum_j (v_{j0} / m_k + v_{j2} / n_k), \quad j = 1, \ldots, p 
\]

(4.6)

where \( v_{j0} = -\sum S_{h} S_{h0y}, \ v_{j1} = N_h^2 (S_{h0y} + 2S_{h0y2}), \ v_{j2} = N_h^4 S_{h0y2}, \ u_{jh} = u_{j1k} + u_{j2k} \). We determine the optimal \( m_k \) and \( n_k \) by mathematical programming methods such that the cost

\[
C = c_o + \sum (c_s m_k + c_a n_k) 
\]

(4.7)

is minimized subject to constraint on the variances

\[
Var_p(\hat{U}_j) \leq V_j, \quad j = 1, \ldots, p, 
\]

(4.8)

and constraint on samples sizes and population sizes

\[
2 \leq n_k \leq N_h, \quad h = 1, \ldots, H, 
\]

(4.9)

where \( c_o \) is the overhead cost, \( c_{s} \) and \( c_{a} \) are the cost per element in stratum \( h \) for the first-phase and second-phase sample respectively, and \( V_j \) is the specified tolerance. For example, one could specify upper limit, \( \phi_j \), on the coefficient of variation of \( \hat{\theta}_j \) so that \( V_j = (\phi_j, \hat{E}_j \hat{\theta}_j)^2 \). We assume that \( c_{a1} < c_{a2} \).

We consider the bias-adjusted estimator \( \hat{Y}_j = \sum d_j y_{j\beta} + \sum (d_j^{1-\alpha} - d_j) \hat{y}_{j\beta} \) of the total \( Y_j \), given by (2.6) with \( \hat{\beta} \) obtained as solution to the estimating equations

\[
\hat{I}_{\beta}(\beta) = \sum I_{\beta} \Phi_{\beta}(\beta)(y_{j\beta} - \mu_{j\beta}(\beta)) = 0. 
\]

In this case

\[
u_{\beta} = \begin{cases} y_{j\beta} \\ y_{j\beta} - \hat{y}_{j\beta} \\ u_{j1k} \\ u_{j2k} \end{cases} 
\]

(4.10)

with \( y_{j\beta} = (1 - I_{\beta}) y_j + I_{\beta} \mu_{\beta}(\beta) \), and \( \beta_{j\beta} \) is the solution to \( I_{\beta}(\beta) = \sum I_{\beta} \Phi_{\beta}(\beta)(y_{j\beta} - \mu_{j\beta}(\beta)) = 0 \).

4.3 Empirical Study

We considered the case of a single characteristic, \( y \), and generated a finite population \{\( y_1, \ldots, y_N \)\} of size \( N = 393 \) from the model

\[
y_i = \mu_i + x^\tau \varepsilon_i, 
\]

with independent errors \( \varepsilon_i \) generated from \( N(0,1) \), \( \mu_i = E_{x_i}(y_i) = \alpha + \beta x_i \) and specified constants \( (\alpha, \beta, \sigma) \), where the fixed \( x_i \) are the “number of beds” in hospital \( k \) for the hospitals population. We stratified the population into two strata with 272 units \( k \) having \( x_i \leq 350 \) in stratum 1 and 121 units \( k \) with \( x_i > 350 \) in stratum 2. For the cost, we set \( c_o = 0 \), \( c_{s2} = c_{s} = 1 \), and two different costs for \( c_{a1} = c_{a} : 0.1 \) and 0.5. We set \( \phi = 0.05 \) for the tolerances. Two estimators are considered \( \hat{Y} = \sum d_i y_i + \sum (d_i^{1-\alpha} - d_i) x_i \) and \( \hat{\beta} = \sum d_i y_i / \sum d_i x_i \). We have \( u_{\beta} = (x_i, y_i - x_i)^\tau \) for \( \hat{Y} \) and \( u_{\beta} = (x_i \beta, y_i - x_i \beta)^\tau \) for \( \hat{\beta} \), where \( \beta_{\beta} = \sum y_i / \sum x_i \).
We repeated the optimization process for different values of $\sigma(0,...,1.5)$ keeping $(\alpha, \beta) = (0,1)$ and $(0,1.2)$. Figures 1 and 2 provide the minimum cost for $\beta = 1.0$ under the two-phase design for $c_i = 0.5$ and $c_i = 0.1$ respectively. The results under substitution and ratio imputation are similar. For comparison, the minimum cost under stratified single phase sampling, using the unbiased Horvitz-Thompson (HT) estimator are also presented. It is seen from Figure 1 that when the cost ratio $c_2/c_1$ is small ($c_2/c_1 = 2$), minimum cost under two-phase is smaller than under one-phase sampling only for small values of $\sigma$ ($\leq 2$). On the other hand, Figure 2 shows that when the cost ratio is large ($c_2/c_1 = 10$), minimum cost is smaller under two-phase sampling for a much wider range of $\sigma$ ($\leq 10$). Figures 3 and 4 report the minimum cost for the case of $\beta = 1.2$ under substitution and ratio imputation for $c_i = 0.5$ and $c_i = 0.1$ respectively.

### 4.4 Missing Responses

In the case of missing responses, $y$, in the second-phase sample, the variance of the estimator of general form 

$$\hat{U} = \sum d_i^{(1)} u_{ix} + \sum d_i^{(2)} v_{ix}$$

is given by

$$Var(\hat{U}) = V'_i(u_i) + V'_j(u_j),$$

with

$$V'_i(u_i) = \sum \left[ \sum \left(1 - \omega_i^{(i)} \right)/\omega_i^{(i)} \right] u_{ix} u_{ix},$$

$$+ \sum \left[ \sum \left(1 - \omega_i^{(j)} \right)/\omega_i^{(j)} \right] \bar{\xi}_x u_{ix} u_{ix},$$

$$+ 2 \sum \left[ \sum \left(1 - \omega_i^{(j)} \right)/\omega_i^{(j)} \right] \bar{\xi}_x u_{ix} u_{ix},$$

(4.11)

and

$$V'_j(u_j) = \sum \left( \bar{\xi}_x - \bar{\xi}_x \right) u_{ix} u_{ix}. $$

(4.12)

We assume independent response mechanism. Under stratified simple random sampling at both stages, we have

$$V'_i(u_i) = \sum h_i^{(1)} N_h^{(1)} / n_h \left[ \sum S_{hv_h}^{(1)} + 2 \sum S_{hv_h}^{(2)} \right],$$

$$+ \sum h_i^{(2)} N_h^{(2)} / n_h \left[ \sum S_{hv_h}^{(2)} + 2 \sum S_{hv_h}^{(3)} \right],$$

$$+ \sum J_{hv_h} \bar{\xi}_x (1 - \bar{\xi}_x) u_{ix} u_{ix} u_{ix},$$

(4.13)

where $\bar{x}_i = x_i \bar{\xi}_x$, $u_i = u_{ix} + u_{ix}$, and $u_i = u_{ix} + u_{ix}$. It follows from (4.11), (4.12) and (4.13), that we can express $Var(\hat{U})$ as (4.6) with

$$v_{ki} = N_i^{(1)} \left[ S_{hv_h}^{(1)} + 2 S_{hv_h}^{(2)} \right],$$

$$v_{ki} = N_i^{(2)} \left[ S_{hv_h}^{(2)} + 2 S_{hv_h}^{(3)} \right],$$

$$v_{ki} = - \sum J_{hv_h} \bar{\xi}_x (1 - \bar{\xi}_x) u_{ix} u_{ix},$$

and

$$\text{References}$$


Table 1: Simulation Results

<table>
<thead>
<tr>
<th>Imputation Method</th>
<th>Estimator</th>
<th>Variable</th>
<th>$RB(\hat{\theta})$</th>
<th>$MSE(\hat{Y}^{(V)})/MSE(\hat{Y})$</th>
<th>$RB(\vartheta(\hat{\theta}))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substitution</td>
<td>Naïve</td>
<td>1</td>
<td>0.0001</td>
<td>94</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>-1.181</td>
<td>9.85</td>
<td>-0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>-1.228</td>
<td>10.66</td>
<td>-0.93</td>
</tr>
<tr>
<td></td>
<td>Design-based</td>
<td>1</td>
<td>0.0004</td>
<td>1</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0.0003</td>
<td>1</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>0.0005</td>
<td>1</td>
<td>-0.001</td>
</tr>
<tr>
<td>Ratio</td>
<td>Naïve</td>
<td>1</td>
<td>-0.0004</td>
<td>99</td>
<td>-0.010</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0.0004</td>
<td>99</td>
<td>-0.011</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>-0.0025</td>
<td>99</td>
<td>-0.006</td>
</tr>
<tr>
<td></td>
<td>Design-based</td>
<td>1</td>
<td>0.0004</td>
<td>1</td>
<td>-0.011</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0.0003</td>
<td>1</td>
<td>-0.010</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>0.0006</td>
<td>1</td>
<td>-0.002</td>
</tr>
</tbody>
</table>

Figure 1: minimum cost for $(\beta = 1.0, c_1 = 0.5)$

Figure 2: minimum cost for $(\beta = 1.0, c_1 = 0.1)$
Figure 3: minimum cost for \((\beta = 1.2, c_1 = 0.5)\)

Figure 4: minimum cost for \((\beta = 1.2, c_1 = 0.1)\)