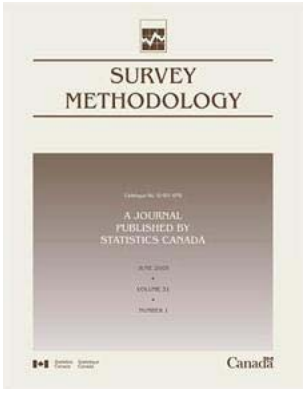
Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Cold deck and ratio imputation



by Jun Shao

June 2000

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2000
Vol. 26, No. 1, pp. 79-85
Statistics Canada, Catalogue No. 12-001

Cold deck and ratio imputation

Jun Shao [1]

Abstract

Imputation is a common procedure to compensate for nonresponse in survey problems. Using auxiliary data, imputation may produce estimators that are more efficient than the one constructed by ignoring nonrespondents and re-weighting. We study and compare the mean squared errors of survey estimators based on data imputed using three different imputation techniques: the commonly used ratio imputation method and two cold deck imputation methods that are frequently adopted in economic area surveys conducted by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics. A cold deck method imputes a nonrespondent of an item by reported values from anything other than reported values for the same item in the current data set (*e.g.*, values from a covariate and/or from a previous survey). Although sometimes a cold deck imputation method makes use of more auxiliary data than the other imputation methods, it is not always better in terms of the mean squared errors of the resulting survey estimators. In a simple case we compare explicitly the mean squared errors and discuss situations under which one method is better than the other two. In general cases we propose to compare mean squared errors empirically based on some consistent estimates of mean squared errors. Estimation of mean squared errors of survey estimators in the presence of imputed data is itself an important problem in surveys. A numerical example related to the Transportation Annual Survey is presented for illustration.

Key Words: Complex survey; Mean squared error; Nonresponse; Simple random sample; Variance estimation.

1. Introduction

Imputation is one of the most common procedures to compensate for nonresponse in survey problems. In addition to many practical reasons for imputation, imputation using auxiliary data may produce estimators that are more efficient than the one constructed by ignoring non-respondents and re-weighting. Suppose that we have a sample $\mathbf{s}$ selected from a finite population P consisting of some units represented by $i = 1, \ldots, M$, and that we observe $\{y_i, i \in \mathbf{r}\}$ (respondents), $\mathbf{r} \subset \mathbf{s}$. Suppose also that we have auxiliary data x_i's observed for all $i \in \mathbf{s}$ and $x_i > 0$. The commonly used ratio imputation method (see, for example, Kalton and Kasprzyk 1986) imputes non-respondents as follows. First, we create K imputation cells $P_k, P_1 \cup P_2 \cup \ldots \cup P_K = P$, according to a categorical auxiliary variable (which is observed for every $i \in \mathbf{s}$ and is typically different from x) such that for every k, the following model is assumed to hold:

$$y_i = \beta_k x_i + x_i^{1/2} e_i, \quad i \in P_k, \qquad (1)$$

$$\mathrm{P}(a_i = 1 | y_i, x_i) = \mathrm{P}(a_i = 1 | x_i),$$

where β_k is an unknown parameter, e_i is independent of x_i with $E(e_i) = 0$ and unknown $V(e_i) = \sigma_k^2 > 0$, a_i is the indicator of whether y_i is a respondent, and (a_i, x_i)'s are independent. Then, within imputation cell k, a nonrespondent y_i is imputed by $\hat{\beta}_k x_i$, where

$$\hat{\beta}_k = \sum_{i \in \mathbf{r_k}} w_i y_i \Big/ \sum_{i \in \mathbf{r_k}} w_i x_i \qquad (2)$$

is the best linear unbiased estimator of β_k under model (1), $\mathbf{r}_k$ is $\mathbf{r}$ restricted to the k^{th} imputation cell, and w_i is the survey weight associated with the i^{th} sampled unit. Note that model (1) consists of regression model between y_i and x_i (with no intercept and with error variance proportional to x_i) and a response model which assumes that the response mechanism is independent of y_i's, given x_i's. This response mechanism is termed as missing at random by Rubin (1976) or unconfounded response mechanism by Lee, Rancourt and Särndal (1994). Based on the imputed data set, the Horvitz-Thompson (HT) estimator of Y, the population total of y_i's, is

$$\hat{Y}_R = \sum_k \left(\sum_{i \in \mathbf{r_k}} w_i y_i + \sum_{i \in \mathbf{s_k} - \mathbf{r_k}} w_i \hat{\beta}_k x_i \right), \qquad (3)$$

where $\mathbf{s_K}$ is $\mathbf{s}$ restricted to the k^{th} imputation cell. The HT estimator of Y obtained by ignoring nonrespondents and re-weighting within each imputation cell is

$$\hat{Y}_W = \sum_k \sum_{i \in \mathbf{r_k}} \tilde{w}_{ik} y_i, \quad \tilde{w}_{ik} = w_i \left(\sum_{i \in \mathbf{s_k}} w_i \Big/ \sum_{i \in \mathbf{r_k}} w_i \right). \qquad (4)$$

It can be seen that if $x_i \equiv 1$, then the estimators in (3) and (4) are the same. Both estimators are unbiased if model (1) holds. (Throughout this paper, the bias and variance are with respect to model (1) and repeated sampling, unless otherwise specified.) Under model (1), however, $\hat{Y}_R$ is more efficient than $\hat{Y}_W$ if the size of $\mathbf{r}$ is substantially smaller than the size of $\mathbf{s}$. Even if the regression model in (1) does not hold, $\hat{Y}_R$ may still be more efficient than $\hat{Y}_W$ in terms of their mean squared errors with respect to repeated sampling (Cochran 1977, Chapter 6) when the response

1. Jun Shao, Department of Statistics, University of Wisconsin, Madison, WI53706 U.S.A. E-mail: shao@stat.wisc.edu.

probability is a constant in any given imputation cell (which ensures that $\hat{Y}_R$ and $\hat{Y}_W$ are approximately unbiased with respect to repeated sampling).

The purpose of this note is to compare the efficiency of $\hat{Y}_R$ with other estimators of Y based on data with nonrespondents imputed by using a method called cold deck. A cold deck method imputes a nonrepondents of y-variable by reported values from anything other than y-values (*e.g.*, values from a covariate and/or from a previous survey). Cold deck imputation is opposite to hot deck imputation in which a nonrespondent is imputed by a respondent from the same variable in the current survey. The ratio imputation method uses both reported y-values and auxiliary data and is sometimes called a "warm deck" method. The simplest cold deck imputes a nonrespondent y_i, $i \in \mathbf{s} - \mathbf{r}$, by x_i and the resulting HT estimator of Y is

$$\hat{Y}_C = \sum_{i\in\mathbf{r}} w_i y_i + \sum_{i\in\mathbf{s}-\mathbf{r}} w_i x_i. \tag{5}$$

The use of this simple cold deck is motivated by the fact that under model (1), β_k's are close to 1 in many survey problems, especially when x_i's are y-values from a previous survey. When some β_k's are not equal to 1, $\hat{Y}_C$ in (5) has a bias which does not vanish even if $\mathbf{s} = P$ (*i.e.*, the sample is a census). However, having a small bias may be paid off by lowering the variance so that the overall mean squared error $\text{mse}(\hat{Y}_C) = E(\hat{Y}_C - Y)^2$ may still be smaller than the mean squared error $\text{mse}(\hat{Y}_R) = E(\hat{Y}_R - Y)^2 = V(\hat{Y}_R - Y)$, where E and V denote the expectation and variance under model (1) and repeated sampling. More details can be found in section 2. The simple cold deck may be improved by another cold deck method, the cold deck-ratio method, which imputes a nonrespondent y_i by $x_i \tilde{y}_i / \tilde{x}_i$, where $\tilde{y}_i$ and $\tilde{x}_i$ are reported values from a previous survey. The corresponding HT estimator of Y is

$$\hat{Y}_{C-\mathbf{R}} = \sum_{i\in\mathbf{r}} w_i y_i + \sum_{i\in\mathbf{s}-\mathbf{r}} w_i x_i \tilde{y}_i / \tilde{x}_i. \tag{6}$$

The estimator in (6) is unbiased if model (1) holds for $\tilde{y}_i$ and $\tilde{x}_i$ (*i.e.*, $\tilde{y}_i = \beta_k \tilde{x}_i + \tilde{x}_i^{1/2} \tilde{e}_i$) with the same β_k as the one for y_i and x_i. These two cold deck methods are widely used in economic area surveys conducted by the U.S. Census Bureau (King and Kornbau 1994) and the U.S. Bureau of Labor Statistics (Butani, Harter and Wolter 1998). Applying cold deck imputation methods does not require knowing the imputation cells, although model (1) is assumed to ensure the unbiasedness of $\hat{Y}_C$ and $\hat{Y}_{C-\mathbf{R}}$.

Although the cold deck-ratio method makes use of more auxiliary data, it is not always better than the simple cold deck or the ratio imputation method. In section 2 we compare explicitly the mean squared errors of $\hat{Y}_R$, $\hat{Y}_C$ and $\hat{Y}_{C-\mathbf{R}}$ in a special case where the sample $\mathbf{s}$ is a simple random sample (SRS) and the response probability is a constant. Situations under which one method is better than the others are discussed. If the sampling design or the response mechanism is complex, then it is not easy to compare the mean squared errors explicitly. One may, however, estimate the mean squared errors of $\hat{Y}_R$, $\hat{Y}_C$ and $\hat{Y}_{C-\mathbf{R}}$ and make an empirical comparison. Variance or mean squared error estimation is itself an important problem, since it is common to report variance or mean squared error estimates along with the estimated totals. These are discussed in section 3.

Our results can also be applied to the problem related to two-pahse sampling or double sampling, which is often employed when it is cheap to take a large sample $\{x_i, i \in \mathbf{s}\}$ and expensive to obtain y-values so that a subsample $\{y_i, i \in \mathbf{r}\}$ is taken in the second-phase, $\mathbf{r} \subset \mathbf{s}$.

A numerical example is discussed in section 4 using data from the Transportation Annual Survey conducted by the U.S. Census Bureau.

2. SRS with uniform response

To illustrate the idea, we start with the simplest case where $\mathbf{s}$ is an SRS (without replacement from P but the sampling fraction is negligible); there is only one imputation cell so that we can drop the subscript k for imputation cell; and the response probability is a constant $p > 0$ (uniform response mechanism).

In this case $w_i = N/n$, where n is the size of the sample $\mathbf{s}$ and N is the size of the population P. Since $n/N \approx 0$ is assumed,

$$\text{mse}(\hat{Y}_R) \approx \frac{N^2}{n}\left(\frac{\sigma^2 \mu_x}{p} + \beta^2 v_x\right) \tag{7}$$

for large n, where $\mu_x = E(x_i)$ and $v_x = V(x_i)$ and, throughout the paper, $A \approx B$ means that A is equal to B up to a term which is relatively negligible compared to A and B as all samples sizes in imputation cells increase to infinity. A more detailed derivation of result (7) is given in the Appendix. For $\hat{Y}_W$ in (4), it is easy to see that $\tilde{w}_i = N/r$, where $\mathbf{r}$ is the size of r, and $\hat{Y}_W$ is unbiased. Then

$$\text{mse}(\hat{Y}_W) = V(\hat{Y}_W - Y) \approx V(\hat{Y}_W) = \frac{N^2}{n}\left(\frac{\sigma^2 \mu_x}{p} + \frac{\beta^2 v_x}{p}\right).$$

Hence $\hat{Y}_R$ is more efficient than $\hat{Y}_W$ unless $p = 1$ and $\beta^2 v_x = 0$. The gain in using $\hat{Y}_R$ is proportional to β^2 and v_x, both are measures of usefulness of the auxiliary variable x in explaining y through model (1).

For the simple cold deck,

$$\hat{Y}_C = \frac{N}{n}\left(\sum_{i\in\mathbf{r}} y_i + \sum_{i\in\mathbf{s}-\mathbf{r}} x_i\right) = \frac{N}{n}\left(\sum_{i\in\mathbf{r}} x_i^{1/2} e_i + \beta \sum_{i\in\mathbf{r}} x_i + \sum_{i\in\mathbf{s}-\mathbf{r}} x_i\right),$$

where e_i's are defined in (1). Consequently,

$$V(\hat{Y}_C) = \frac{N^2}{n}\{\sigma^2 p \mu_x + (\beta^2 p + 1 - p) v_x + (\beta - 1)^2 p(1-p)\mu_x^2\} \tag{8}$$

(see the Appendix). The bias of $\hat{Y}_C$ is

$$E(\hat{Y}_C - Y) = N\mu_x(1-p)(1-\beta)$$

and, hence,

$$\begin{aligned}\text{mse}(\hat{Y}_C) &= V(\hat{Y}_C - Y) + [E(\hat{Y}_C - Y)]^2 \\ &\approx V(\hat{Y}_C) + [E(\hat{Y}_C - Y)]^2 \\ &= \frac{N^2}{n}\{\sigma^2 p\mu_x + (\beta^2 p + 1 - p)v_x \\ &\quad + (\beta-1)^2(1-p)[p + n(1-p)]\mu_x^2\}. \quad (9)\end{aligned}$$

Comparing (7) and (9), we obtain the following conclusions.

1. When $p = 1$ (no response), $\text{mse}(\hat{Y}_C) = \text{mse}(\hat{Y}_R)$.
2. When $p < 1$ and $\beta = 1$ (y and x have the same mean), $\text{mse}(\hat{Y}_C) < \text{mse}(\hat{Y}_R)$.
3. When $p < 1$ and $\beta \neq 1$, $\text{mse}(\hat{Y}_C) \leq \text{mse}(\hat{Y}_R)$ if and only if

$$(\beta-1)^2[p + n(1-p)]\mu_x + (1-\beta^2)v_x/\mu_x - \sigma^2(p+1)/p \leq 0. \quad (10)$$

Assume that $\mu_x > 0$. In most economic surveys, the relative variance v_x/μ_x^2 is smaller than $p + n(1-p)$. Hence the left hand side of (10) is a quadratic function of β with a positive coefficient in the β^2 term and, therefore, the simple cold deck is better when β is in the interval with limits

$$\frac{[p+n(1-p)]\mu_x \pm \sqrt{v_x^2/\mu_x^2 + \{[p+n(1-p)]\mu_x - v_x/\mu_x\}\sigma^2(p+1)/p}}{[p+n(1-p)]\mu_x - v_x/\mu_x}.$$

This interval contains 1 since (10) holds if $\beta = 1$. Note that $[p + n(1-p)]\mu_x$ increases to infinity as n increases to infinity. Hence the interval of β's for which the simple cold deck is better shrinks to a single point ($\beta = 1$) as $n \to \infty$.

We now consider the cold deck-ratio. Assume that $\tilde{y}_i = \beta\tilde{x}_i + \tilde{x}_i^{1/2}\tilde{e}_i$, $E(\tilde{e}_i) = 0$, $V(\tilde{e}_i) = \sigma^2$, and that $\tilde{e}_i$, e_i, and $(x_i, \tilde{x}_i)$ are mutually independent. Let $z_i = x_i\tilde{y}_i/\tilde{x}_i$ and $\epsilon_i = y_i - z_i = x_i^{1/2}e_i - \tilde{e}_i x_i/\tilde{x}_i^{1/2}$. Then $E(\hat{Y}_{C-\mathbf{R}} - Y) = 0$ and

$$\text{mse}(\hat{Y}_{C-\mathbf{R}}) = \frac{N^2}{n}\{\sigma^2 p\mu_x + \beta^2 v_x + \sigma^2(1-p)\gamma_x\}, \quad (11)$$

where $\gamma_x = E(x_i^2/\tilde{x}_i)$ (see the Appendix). By (7) and (11),

$$\text{mse}(\hat{Y}_R) - \text{mse}(\hat{Y}_{C-\mathbf{R}}) = \frac{N^2\sigma^2(1-p)}{n}\left\{\left(\frac{1}{p}+1\right)\mu_x - \gamma_x\right\} \quad (12)$$

and, hence, the cold deck-ratio is better than the ratio imputation method if and only if $1/p + 1 \geq \gamma_x/\mu_x$. Note that $\gamma_x \geq \mu_x$ and γ_x is close to μ_x if x_i and $\tilde{x}_i$ are highly and positively related, in which case cold deck-ratio imputation can be much better than ratio imputation.

The comparison between the simple cold deck and the cold deck-ratio is the same as that between the simple cold deck and the ratio imputation method. One only needs to replace $(p+1)/p$ in the third term of the left hand side of (10) by γ_x/μ_x.

The parameters β, σ, μ_x, v_x and γ_x have to be estimated in order to compare the efficiencies of $\hat{Y}_R$, $\hat{Y}_C$ and $\hat{Y}_{C-\mathbf{R}}$. Instead, we can directly compare estimated mean squared errors of $\hat{Y}_R$, $\hat{Y}_C$ and $\hat{Y}_{C-\mathbf{R}}$. This is discussed next.

3. Stratified sampling with unconfounded response

We consider the following stratified sampling design adopted by many U.S. government survey agencies: the finite population P is stratified into H strata with N_h units in the h^{th} stratum; $n_h \geq 2$ units are selected without replacement from stratum h, according to some probability sampling plan; and the units are selected independently across the strata.

The survey weights w_i's are constructed so that if all y_i's are observed, the HT estimator $\sum_{i\in\mathbf{s}} w_i y_i$ is unbiased for Y under repeated sampling.

We assume model (1). The response probability is no longer a constant, but independent of the y-value. For the cold deck-ratio, we also assume that within the k^{th} imputation cell, $\tilde{y}_i = \beta_k\tilde{x}_i + \tilde{x}_i^{1/2}\tilde{e}_i$, $E(\tilde{e}_i) = 0$, $V(\tilde{e}_i) = \tilde{\sigma}_k^2$ and e_i, $\tilde{e}_i$, $(x_i, \tilde{x}_i)$ are mutually independent.

Explicit results for the mean squared errors such as (7), (9) and (11) are not easy to obtain. We may, however, make empirical comparisons of the efficiencies of $\hat{Y}_R$, $\hat{Y}_C$ and $\hat{Y}_{C-\mathbf{R}}$, based on their estimated mean squared errors. Estimation of the mean squared errors of $\hat{Y}_R$, $\hat{Y}_C$ and $\hat{Y}_{C-\mathbf{R}}$, is in fact an important part of the sampling theory. It is well known that for imputed data sets, the naive method that applies the standard variance estimation formulas by treating imputed nonrespondents as observed data leads to underestimation. When no correct method (for estimating the mean squared error) is available, the naive method is used in many survey agencies.

We now derive estimators for $V(\hat{Y})$ or $\text{mse}(\hat{Y})$ that are correct under model (1), where $\hat{Y}$ denotes $\hat{Y}_R$, $\hat{Y}_C$ or $\hat{Y}_{C-\mathbf{R}}$.

Let E_m and V_m be the expectation and variance with respect to model (1) and let E_s and V_s be the expectation and variance with respect to repeated sampling (conditional on the model and response). Then

$$V(\hat{Y} - Y) = E_m[V_s(\hat{Y})] + V_m[E_s(\hat{Y}) - Y]. \quad (13)$$

We first consider $E_m[V_s(\hat{Y})]$, the first variance component in (13). It suffices to obtain an estimator of $V_s(\hat{Y})$, conditional on $\{y_i, x_i, a_i, i \in P\}$ (and $\{\tilde{y}_i, \tilde{x}_i, i \in P\}$ for cold deck-ratio), where a_i is the response indicator for y_i.

The estimation of $V_s(\hat{Y}_C)$ and $V_s(\hat{Y}_{C-\mathbf{R}})$ is simple (which is an advantage of using a cold deck method). Let

$$v_1 = \sum_h \left(1 - \frac{n_h}{N_h}\right) \frac{n_h}{n_h - 1} \sum_{i \in \mathbf{s(h)}} \left(w_i t_i - \frac{1}{n_h} \sum_{i \in \mathbf{s(h)}} w_i t_i \right)^2 \quad (14)$$

be the standard variance estimator for $\sum_{i \in \mathbf{s}} w_i t_i$ when $\{t_i, i \in \mathbf{s}\}$ is treated as an observed sample (from $\{t_i, i \in P\}$), where $\mathbf{s(h)}$ is $\mathbf{s}$ restricted to stratum h. Then $V_s(\hat{Y}_C)$ can be estimated by using (14) with $t_i = a_i y_i + (1 - a_i) x_i$ and $V_s(\hat{Y}_{C-\mathbf{R}})$ can be estimated by using (14) with $t_i = a_i y_i + (1 - a_i) x_i \tilde{y}_i / \tilde{x}_i$.

The estimation of $V_s(\hat{Y}_R)$ is slightly more complicated but similar. Assume that in each imputation cell, the number of sampled units is large and the response probabilities are bounded away from 0. Note that

$$\begin{aligned} \hat{Y}_R &= \sum_k \left[\left(\sum_{i \in \mathbf{s_k}} w_i x_i \Big/ \sum_{i \in \mathbf{r_k}} w_i x_i \right) \right. \\ &\qquad \left. \times \sum_{i \in \mathbf{r_k}} w_i (y_i - \beta_k x_i) + \beta_k \sum_{i \in \mathbf{s_k}} w_i x_i \right] \\ &\approx \sum_k \left[\zeta_k \sum_{i \in \mathbf{s_k}} w_i a_i (y_i - \beta_k x_i) + \beta_k \sum_{i \in \mathbf{s_k}} w_i x_i \right] \\ &= \sum_{i \in \mathbf{s}} w_i \left[\zeta_i a_i (y_i - \beta_i x_i) + \beta_i x_i \right], \end{aligned}$$

where $\zeta_k = E(\sum_{i \in \mathbf{s_k}} w_i x_i) / E(\sum_{i \in \mathbf{r_k}} w_i x_i)$ and $\zeta_i = \zeta_k$ and $\beta_i = \beta_k$ for $i \in \mathbf{s_k}$. After estimating β_k by $\hat{\beta}_k$ and ζ_k by $\hat{\zeta}_k = \sum_{i \in \mathbf{s_k}} w_i x_i / \sum_{i \in \mathbf{r_k}} w_i x_i$, we estimate $V_s(\hat{Y}_R)$ by using (14) with $t_i = \hat{\zeta}_i a_i (y_i - \hat{\beta}_i x_i) + \hat{\beta}_i x_i$, where $\hat{\zeta}_i = \hat{\zeta}_k$ and $\hat{\beta}_i = \hat{\beta}_k$ for $i \in \mathbf{s_k}$.

Before we discuss the estimation of $V_m[E_s(\hat{Y}) - Y]$, the second variance component in (13), it should be noted that $V_m[E_s(\hat{Y}) - Y] / E_m[V_s(\hat{Y})] = O(n/N)$. This is because the variance of $E_s(\hat{Y}) - Y$ (if it is nonzero) is typically of the order N, whereas the order of $V_s(\hat{Y})$ is typically N^2/n and thus the order of $E_m[V_s(\hat{Y})]$ is N^2/n under some regularity conditions. Hence, in theory, it is not necessary to estimate $V_m[E_s(\hat{Y}) - Y]$ if the sampling fraction n/N is negligible. However, the constant in $O(n/N)$ is unknown and, hence, one may still want to estimate $V_m[E_s(\hat{Y}) - Y]$ in applications even when n/N is small.

We now consider the estimation of the second variance component in (13). For $\hat{Y}_C$,

$$\begin{aligned} E_s(\hat{Y}_C) - Y &= \sum_{i \in P} [a_i y_i + (1 - a_i) x_i] \\ &\qquad - \sum_{i \in P} y_i = - \sum_{i \in P} (1 - a_i)(y_i - x_i). \end{aligned}$$

Then, under model (1),

$$\begin{aligned} V_m[E_s(\hat{Y}_C) - Y] &= E_m \left[\sum_k \sigma_k^2 \sum_{i \in P_k} (1 - a_i) x_i \right] \\ &\qquad + V_m \left[\sum_{i \in P} (1 - a_i)(\beta_i - 1) x_i \right]. \end{aligned}$$

If we estimate σ_k^2 by

$$\hat{\sigma}_k^2 = \sum_{i \in \mathbf{s_k}} a_i w_i (y_i - \hat{\beta}_k x_i)^2 \Big/ \sum_{i \in \mathbf{s_k}} a_i w_i x_i,$$

then an estimator of $V_m[E_s(\hat{Y}_C) - Y]$ is

$$\begin{aligned} v_{2C} &= \sum_k \hat{\sigma}_k^2 \sum_{i \in \mathbf{s_k}} (1 - a_i) w_i x_i \\ &\qquad + \sum_h \frac{N_h}{n_h - 1} \sum_{i \in \mathbf{s(h)}} \left(u_i - \frac{1}{n_h} \sum_{i \in \mathbf{s(h)}} u_i \right)^2, \end{aligned} \quad (15)$$

where $u_i = (1 - a_i)(\hat{\beta}_i - 1) x_i$ and $\hat{\beta}_i = \hat{\beta}_k$ for $i \in \mathbf{s_k}$.

For $\hat{Y}_{C-\mathbf{R}}$,

$$E_s(\hat{Y}_{C-\mathbf{R}}) - Y = -\sum_{i \in P} (1 - a_i)(y_i - x_i \tilde{y}_i / \tilde{x}_i)$$

and

$$V_m[E_s(\hat{Y}_{C-\mathbf{R}}) - Y] = E_m \left[\sum_k \sigma_k^2 \sum_{i \in P_k} (1 - a_i) x_i + \sum_k \hat{\sigma}_k^2 \sum_{i \in P_k} (1 - a_i) x_i^2 / \tilde{x}_i \right].$$

Hence $V_m[E_s(\hat{Y}_{C-\mathbf{R}}) - Y]$ can be estimated by

$$v_{2C-\mathbf{R}} = \sum_k \left[\hat{\sigma}_k^2 \sum_{i \in \mathbf{s_k}} (1 - a_i) w_i x_i + \hat{\tilde{\sigma}}_k^2 \sum_{i \in \mathbf{s_k}} (1 - a_i) w_i x_i^2 / \tilde{x}_i \right], \quad (16)$$

where

$$\hat{\tilde{\sigma}}_k^2 = \sum_{i \in \mathbf{s_k}} w_i (\tilde{y}_i - \hat{\tilde{\beta}}_k \tilde{x}_i)^2 \Big/ \sum_{i \in \mathbf{s_k}} w_i \tilde{x}_i$$

and

$$\hat{\tilde{\beta}}_k = \sum_{i \in \mathbf{s_k}} w_i \tilde{y}_i \Big/ \sum_{i \in \mathbf{s_k}} w_i \tilde{x}_i.$$

For $\hat{Y}_R$,

$$E_s(\hat{Y}_R) - Y \approx \sum_k \left[\left(\sum_{i \in P_k} x_i \Big/ \sum_{i \in P_k} a_i x_i \right) \sum_{i \in P_k} a_i y_i - \sum_{i \in P_k} y_i \right]$$

and from Taylor's expansion,

$$V_m[E_s(\hat{Y}_R) - Y] \approx E_m \left\{ \sum_k \sigma_k^2 \left[\sum_{i \in P_k} x_i \sum_{i \in P_k} (1 - a_i) x_i \right] \Big/ \sum_{i \in P_k} a_i x_i \right\}.$$

It can be estimated by

$$v_{2R} = \sum_k \hat{\sigma}_k^2 \left[\sum_{i \in \mathbf{s_k}} w_i x_i \sum_{i \in \mathbf{s_k}} (1 - a_i) w_i x_i \right] \Big/ \sum_{i \in \mathbf{s_k}} a_i w_i x_i. \quad (17)$$

Finally, $\hat{Y}_R$ and $\hat{Y}_{C-\mathbf{R}}$ are unbiased buy $\hat{Y}_C$ has a bias

$$\sum_k (1 - \beta_k) E_m \left[\sum_{i \in P_k} (1 - a_i) x_i \right],$$

which can be estimated by

$$\sum_k (1-\hat{\beta}_k)\sum_{i\in \mathbf{s_k}} (1-a_i)\, w_i\, x_i.$$

Thus, we obtain the following estimated mean squared errors: $\mathrm{mse}(\hat{Y}_R)$ can be estimated by

$$\widehat{\mathrm{mse}}(\hat{Y}_R) = v_{1R} + v_{2R},$$

where v_{1R} is obtained using (14) with $t_i = \hat{\zeta}_i\, a_i\,(y_i - \hat{\beta}_i\, x_i) + \hat{\beta}_i\, x_i$, $\hat{\zeta}_i = \hat{\zeta}_k$ and $\hat{\beta}_i = \hat{\beta}_k$ for $i \in \mathbf{s_k}$, and v_{2R} is given by (17); $\mathrm{mse}(\hat{Y}_C)$ by

$$\widehat{\mathrm{mse}}(\hat{Y}_C) = v_{1C} + v_{2C} + \left[\sum_k (1-\hat{\beta}_k) \sum_{i\in \mathbf{s_k}-\mathbf{r_k}} w_i\, x_i\right]^2,$$

where v_{1C} is obtained by using (14) with $t_i = a_i\, y_i + (1-a_i)\, x_i$ and v_{2C} is given by (15); and $\mathrm{mse}(\hat{Y}_{C-\mathbf{R}})$ can be estimated by

$$\widehat{\mathrm{mse}}(\hat{Y}_{C-\mathbf{R}}) = v_{1C-\mathbf{R}} + v_{2C-\mathbf{R}},$$

where $v_{1C-\mathbf{R}}$ is obtained by using (14) with $t_i = a_i\, y_i + (1-a_i)\, x_i\, \tilde{y}_i / \tilde{x}_i$ and $v_{2C-\mathbf{R}}$ is given by (16).

Under model (1) and the asymptotic settings in Krewski and Rao (1981), Rao and Shao (1992) or Valliant (1993), the derived mean squared error estimator are asymptotically unbiased and consistent as all sample sizes in imputation cell increase to infinity.

For cold deck or cold deck-ratio imputation, the first term (v_{1C} or $v_{1C-\mathbf{R}}$) in the estimated mean squared error is the same as the one obtained by applying a standard formula (such as (14)) and treating imputed nonrespondents as observed data. For ratio imputation, applying (14) and treating imputed nonrespondents as observed data produces the following estimator of $\mathrm{mse}(\hat{Y}_R)$:

$$\tilde{v}_{1R} = \sum_h \left(1 - \frac{n_h}{N_h}\right) \frac{n_h}{n_h - 1} \sum_{i\in \mathbf{s(h)}} \left(w_i\, z_i - \frac{1}{n_h} \sum_{i\in \mathbf{s(h)}} w_i\, z_i\right)^2 \quad (18)$$

with $z_i = a_i\, y_i + (1-a_i)\hat{\beta}_i\, x_i$, which is different from the first term v_{1R} in our estimator $\widehat{\mathrm{mse}}(\hat{Y}_R)$ and, hence, is not asymptotically valid even if n/N is negligible.

4. An example

We consider an example using a data set from the Transportation Annual Survey (TAS) conducted by U.S. Census Bureau.

The TAS is a survey of firms with one or more establishments that are primarily engaged in providing commercial motor freight transportation or public warehousing services in U.S. A stratified simple random sample is selected without replacement from employers contained in the Census Bureau's Standard Statistical Establishment List. The strata, which are also the imputation classes in this example, are constructed according to company's size within each industry.

There are various variables in this survey. We consider the estimation of the population totals of the current year annual revenue (y) in four industries. The variable y has nonrespondents. Three covariates without nonrespondents are considered: the current year annual payroll, the previous year annual revenue, and the previous year annual payroll. The sample size, response size for y, and the sampling weight in each stratum and industry are given in Table 1.

Table 1
Sample sizes, response sizes, and sampling weights across industries and strata

Industry	Stratum	Sample Size	Response Size	Sampling Weight
1	0	31	24	1.00
	1	14	6	12.43
	2	11	7	8.91
	3	10	4	6.10
	4	11	6	5.73
	5	16	12	2.70
	6	18	13	2.17
2	0	86	82	1.00
	1	8	2	32.91
	2	13	10	9.85
	3	11	9	10.82
	4	12	10	6.08
	5	13	10	3.60
3	0	38	30	1.00
	1	14	9	87.91
	2	11	8	67.39
	3	13	10	44.48
	4	14	13	25.28
	5	16	13	15.57
	6	18	12	9.80
	7	15	11	6.23
	8	15	14	4.68
	9	40	33	2.13
4	0	28	23	1.00
	1	7	5	32.14
	2	13	6	16.75
	3	10	7	12.90
	4	14	12	7.00
	5	13	9	6.18
	6	11	7	4.70
	7	17	12	3.31
	8	19	14	1.89
	9	22	16	1.82

First, we use the previous year annual revenue as the covariate x in simple cold deck imputation and ratio imputation. The current year annual payroll and the previous year annual payroll are used as $\tilde{y}$ and $\tilde{x}$, respectively. For four industries and three imputation methods, Table 2 lists the estimated totals, the proposed estimated MSE's for the estimated totals, the naive estimated MSE's for the estimated totals (obtained by treating imputed values as

observed data), and the MSE ratios (the proposed estimated MSE over the naive estimated MSE). Note that the proposed estimated MSE is the sum of v_1 and v_2 for the ratio and cold deck-ratio methods or the sum of v_1, v_2, and the squared estimated bias for the simple cold deck method. Values of v_1 and v_2 are also included in the table

Table 2
Estimated totals and MSE's when x = the previous year annual revenue, $\tilde{y}$ = the current year payroll, and $\tilde{x}$ = the previous year annual payroll

Industry	Estimate	Method: Cold Deck	Cold Deck-Ratio	Ratio
1	Total	5.31×10^9	5.19×10^9	5.42×10^9
	v_1	7.73×10^{14}	8.46×10^{14}	2.60×10^{15}
	v_2	1.39×10^{15}	2.50×10^{15}	1.81×10^{15}
	Proposed MSE	2.30×10^{15}	3.34×10^{15}	4.40×10^{15}
	Naive MSE	7.73×10^{14}	8.46×10^{14}	2.46×10^{15}
	MSE Ratio	2.97	3.95	1.79
2	Total	1.66×10^{10}	1.63×10^{10}	1.67×10^{10}
	v_1	4.00×10^{15}	4.19×10^{15}	5.57×10^{16}
	v_2	6.03×10^{15}	2.88×10^{16}	6.54×10^{15}
	Proposed MSE	1.02×10^{16}	3.30×10^{16}	6.23×10^{16}
	Naive MSE	4.00×10^{15}	4.19×10^{15}	5.58×10^{16}
	MSE Ratio	2.54	7.87	1.12
3	Total	3.54×10^{10}	3.53×10^{10}	3.59×10^{10}
	v_1	1.32×10^{16}	1.80×10^{16}	1.94×10^{17}
	v_2	5.44×10^{16}	8.62×10^{16}	6.77×10^{16}
	Proposed MSE	6.97×10^{16}	1.04×10^{17}	2.62×10^{17}
	Naive MSE	1.32×10^{16}	1.80×10^{16}	1.87×10^{17}
	MSE Ratio	5.27	5.80	1.40
4	Total	1.27×10^{10}	1.22×10^{10}	1.30×10^{10}
	v_1	2.11×10^{16}	2.14×10^{16}	5.13×10^{15}
	v_2	3.91×10^{15}	8.26×10^{15}	5.06×10^{15}
	Proposed MSE	2.59×10^{16}	2.97×10^{16}	1.02×10^{16}
	Naive MSE	2.11×10^{16}	2.14×10^{16}	5.06×10^{15}
	MSE Ratio	1.23	1.39	2.01

Next, to see the effect of using a wrong covariate in using the simple cold deck method, we repeat the previous computations using the current year annual payroll as the covariate x, and the previous year annual revenue and payroll as $\tilde{y}$ and $\tilde{x}$, respectively. The results are reported in Table 3.

The following is a summary of the results in Tables 2 and 3.

1. The simple cold deck method depends heavily on the choice of the covariate x. When x is the previous year annual revenue (Table 2), the difference among the estimated totals provided by three methods is negligible; in terms of the estimated MSE, the simple cold deck method is the best. However, when x is the current year annual payroll (Table 3), the estimates from the simple cold deck is obviously too low; in terms of the estimated MSE, the simple cold deck method is the worst, because of its large bias (shown in Table 3).

Table 3
Estimated totals and MSE's when x = the current year annual payroll, $\tilde{y}$ = the previous year annual revenue, and $\tilde{x}$ = the previous year annual payroll

Industry	Estimate	Method: Cold Deck	Cold Deck-Ratio	Ratio
1	Total	4.49×10^9	5.19×10^9	5.39×10^9
	Bias	-8.99×10^8		
	v_1	8.10×10^{14}	8.46×10^{14}	2.85×10^{15}
	v_2	1.38×10^{15}	2.64×10^{15}	1.75×10^{15}
	Proposed MSE	1.03×10^{16}	3.49×10^{15}	4.60×10^{15}
	Naive MSE	8.10×10^{14}	8.46×10^{14}	2.55×10^{15}
	MSE Ratio	12.68	4.12	1.81
2	Total	1.59×10^{10}	1.63×10^{10}	1.71×10^{10}
	Bias	-1.21×10^9		
	v_1	4.36×10^{15}	4.19×10^{15}	5.74×10^{16}
	v_2	8.20×10^{15}	1.48×10^{16}	8.95×10^{15}
	Proposed MSE	2.73×10^{16}	1.90×10^{16}	6.64×10^{16}
	Naive MSE	4.36×10^{15}	4.19×10^{15}	5.62×10^{16}
	MSE Ratio	6.25	4.54	1.18
3	Total	3.10×10^{10}	3.53×10^{10}	3.47×10^{10}
	Bias	-3.62×10^9		
	v_1	1.25×10^{16}	1.80×10^{16}	2.30×10^{17}
	v_2	4.56×10^{16}	9.25×10^{16}	5.41×10^{16}
	Proposed MSE	1.89×10^{17}	1.10×10^{17}	2.84×10^{17}
	Naive MSE	1.25×10^{16}	1.80×10^{16}	1.83×10^{17}
	MSE Ratio	15.13	6.15	1.56
4	Total	1.06×10^{10}	1.22×10^{10}	1.20×10^{10}
	Bias	-1.35×10^9		
	v_1	1.93×10^{16}	2.14×10^{16}	5.84×10^{15}
	v_2	2.67×10^{15}	4.62×10^{15}	3.07×10^{15}
	Proposed MSE	4.03×10^{16}	2.60×10^{16}	8.92×10^{15}
	Naive MSE	1.93×10^{16}	2.14×10^{16}	8.92×10^{15}
	MSE Ratio	2.09	1.22	1.72

2. There is no definite conclusion on the relative performance (in terms of the estimated MSE) of the ratio imputation method and the cold deck-ratio method. In this example, the cold deck-ratio is better for industries 1-3, whereas the ratio imputation method is better for industry 4. Some scatter plots of the data (not shown) indicate that the correlation between x and $\tilde{x}$ in industries 1-3 is higher than that in industry 4, which might be the reason for the difference in relative performance of the two imputation methods. See also the discussion after formula (12)

3. The naive estimated MSE's are much lower than the proposed estimated MSE's and are too optimistic. For example, in Table 3, the naive MSE's for the simple cold deck method are always smaller than those for the cold deck-ratio method, although we know that the simple cold deck does not work well in this case. In this example, v_2 / v_1 is not small because of some large sampling fractions. Since the naive estimated MSE is either equal to v_1 (for the cold deck imputation

methods) or not very different from v_1 (for ratio imputation), the underestimation in using the naive estimated MSE is mainly due to treating imputed values as observed values in strata with large sampling fractions (and ignoring the bias of the simple cold deck estimators in the case of Table 3).

Acknowledgement

The author would like to thank referees for helpful comments and suggestions. The first draft of this paper was finished at the U.S. Census Bureau and the U.S. Bureau of Labor Statistics when the author was an ASA/NSF Senior Research Fellow. The research was also supported by National Science Foundation Grants DMS-9504425 and DMS-9803112 and National Security Agency Grant MDA904-99-1-0032.

Appendix

1. **Proof of (7):** When $n/N \approx 0$, $V(\hat{Y}_R - Y) \approx V(\hat{Y}_R)$. Then (7) follows from

$$V(\hat{Y}_R) = \frac{N^2}{n^2}\left\{\sigma^2 E\left[\left(\sum_{i\in\mathbf{s}} x_i\right)^2 \Big/ \left(\sum_{i\in\mathbf{r}} x_i\right)\right] + \beta^2 V\left(\sum_{i\in\mathbf{s}} x_i\right)\right\}$$

$$\approx \frac{N^2}{n}\left(\frac{\sigma^2 \mu_x}{p} + \beta^2 v_x\right)$$

for large n, where the last approximate equality follows from the fact that conditioned on x_i's, $E(\sum_{i\in\mathbf{r}} x_i) = p\sum_{i\in\mathbf{s}} x_i$.

2. **Proof of (9):** Under model (1),

$$V(\hat{Y}_C) = \frac{N^2}{n^2}\left\{V\left(\sum_{i\in\mathbf{r}} x_i^{1/2} e_i\right) + V\left(\beta\sum_{i\in\mathbf{r}} x_i + \sum_{i\in s-\mathbf{r}} x_i\right)\right\}$$

$$= \frac{N^2}{n^2}\left\{\sigma^2 p\mu_x + \beta^2 V\left(\sum_{i\in\mathbf{r}} x_i\right) + V\left(\sum_{i\in s-\mathbf{r}} x_i\right) + 2\beta\,\mathrm{Cov}\left(\sum_{i\in\mathbf{r}} x_i, \sum_{i\in s-\mathbf{r}} x_i\right)\right\}$$

$$= \frac{N^2}{n}\{\sigma^2 p\mu_x + \beta^2[pv_x + p(1-p)\mu_x^2] + (1-p)(v_x + p\mu_x^2) - 2\beta p(1-p)\mu_x^2\}$$

$$= \frac{N^2}{n}\{\sigma^2 p\mu_x + (\beta^2(p+1-p)v_x + (\beta-1)^2 p(1-p)\mu_x^2\}.$$

3. **Proof of (11):** Under the assumed conditions on (y_i, x_i) and $(\tilde{y}_i, \tilde{x}_i)$,

$$\mathrm{mse}(\hat{Y}_{C-\mathbf{R}}) = \frac{N^2}{n^2} V\left(\sum_{i\in\mathbf{r}} y_i + \sum_{i\in s-\mathbf{r}} z_i\right)$$

$$= \frac{N^2}{n^2} V\left(\sum_{i\in\mathbf{r}} \in_i + \sum_{i\in\mathbf{s}} z_i\right)$$

$$= \frac{N^2}{n^2}\left\{V\left(\sum_{i\in\mathbf{r}} \in_i\right) + V\left(\sum_{i\in\mathbf{s}} z_i\right) + 2\mathrm{Cov}\left(\sum_{i\in\mathbf{r}} \in_i, \sum_{i\in\mathbf{s}} z_i\right)\right\}$$

$$= \frac{N^2}{n}\{\sigma^2 p(\mu_x + \gamma_x) + (\beta^2 v_x + \sigma^2\gamma_x) - 2\sigma^2 p\gamma_x\}$$

$$= \frac{N^2}{n}\{\sigma^2 p\mu_x + \beta^2 v_x + \sigma^2(1-p)\gamma_x\}.$$

References

Butani, S., Harter, R. and Wolter, K. (1998). Estimation procedures for the Bureau of Labor Statistics current employment statistics program. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Cochran, W.G. (1977). *Sampling Techniques*. Third Edition. New York: John Wiley & Sons, Inc.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing data. *Survey Methodology*, 12. 1-16.

King, C., and Kornbau, M. (1994). Inventory of economic area statistical practices. ESMD Report Series 9401, Bureau of the Census, Washington D.C.

Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

Lee, H., Rancourt, E. and Särndal, C.-E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

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

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of American Statistical Association*, 88, 89-96.