

INFERENCE FOR REGRESSION COEFFICIENTS UNDER IMPUTATION FOR MISSING DATA

David Haziza and J.N.K. Rao¹

ABSTRACT

A population mean can be estimated unbiasedly under regression imputation for missing data and uniform response or ignorable response, but the imputed estimator of a census regression coefficient is generally biased. Following the approach of Skinner and Rao (1999), we obtain a bias-adjusted estimator of a regression coefficient under an arbitrary design. We derive consistent estimators for the variance-covariance matrix using a method introduced by Fay (1991) in which the usual sample-response path is reversed.

KEY WORDS: Census regression coefficient; Design-based framework; Ignorable response; Model-based framework; Regression imputation; Uniform response.

RÉSUMÉ

Il est en général possible d'obtenir un estimateur sans biais pour une moyenne sous un mécanisme de réponse uniforme ou ignorable quand l'imputation par régression a été utilisée. Ceci n'est cependant pas le cas quand il s'agit d'un coefficient de régression. Grâce à une correction, nous obtenons un estimateur sans biais pour les coefficients de régression en empruntant l'approche développée par Skinner et Rao (1999) dans le cas d'un plan de sondage arbitraire. De plus, en utilisant une méthode proposée par Fay (1991), qui suggère de renverser l'ordre habituel de plan de sondage-réponse, nous développons des estimateurs de la matrice de variance-covariance convergents.

MOTS CLÉS: Approche basée sur le plan de sondage; approche basée sur un modèle; coefficient de régression; imputation par régression; réponse ignorable; réponse uniforme.

1. INTRODUCTION

Marginal item imputation, for missing data, often gives unbiased estimators of marginal population means under uniform or ignorable response. In this paper, we are interested in making inference, under imputation, on census regression coefficients

$$\mathbf{B}_N = \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^N \mathbf{x}_i y_i \quad (1)$$

from a sample with missing y -values where \mathbf{x} is a p -vector of x -variables, y is the response variable and N is the population size (Binder, 1983). A particular case of (1) is when \mathbf{x}_i is a vector of domain indicators in which case \mathbf{B}_N is the vector of domain means. Because \mathbf{B}_N involves the product term $\sum \mathbf{x}_i y_i$, marginal imputation of missing y may not lead to unbiased estimators under uniform response, as shown in Haziza and Rao (2000) for the case of

domain means. Following Skinner and Rao (1999), we propose a bias-adjusted estimator of \mathbf{B}_N and derive consistent estimators for its variance-covariance matrix.

To estimate the variance, we use Fay's (1991) response-sampling approach:

Population \rightarrow census with
 nonrespondents \rightarrow sample with nonrespondents.

In this case (cf., Shao and Steel, 1999), $E(\hat{\boldsymbol{\theta}}) = E_r E_p(\hat{\boldsymbol{\theta}})$ and

$$V(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = E_r V_p(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + V_r E_p(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \quad (2)$$

where $\boldsymbol{\theta}$ denotes an arbitrary multi-dimensional parameter and $\hat{\boldsymbol{\theta}}$ denotes its estimator based on the observed and imputed data, $E_p(\cdot)$ and $V_p(\cdot)$ denote respectively the expectation and the variance with respect to the sampling design and $E_r(\cdot)$ and $V_r(\cdot)$

¹ David Haziza, Household Survey Methods Division, 16-R R.H. Coats Building, Statistics Canada, Ottawa, Ontario, K1A-0T6, david.haziza@a.statcan.ca and J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S-5B6.

denote respectively the expectation and the variance with respect to the assumed response mechanism. In the model-based framework, we replace $E_r(\cdot)$ and $V_r(\cdot)$ by $\tilde{E}_m(\cdot) = E_r E_m(\cdot)$ and $\tilde{V}_m(\cdot) = E_r V_m(\cdot) + V_r E_m(\cdot)$ respectively, where $E_m(\cdot)$ and $V_m(\cdot)$ denote respectively the expectation and the variance with respect to the imputation model.

An estimator of the overall variance-covariance matrix $V(\hat{\theta} - \theta)$ in (2) is given by $\mathbf{v}_t = \mathbf{v}_1 + \mathbf{v}_2$, where \mathbf{v}_1 is an estimator of $V_p(\hat{\theta} - \theta)$ conditional on the response indicators, and \mathbf{v}_2 is an estimator of $V_r E_p(\hat{\theta} - \theta)$. Finding \mathbf{v}_1 does not depend on the response mechanism or the assumed model, and thus \mathbf{v}_1 is valid under either the design-based framework or the model-based framework. Also, when the sampling fraction is negligible, the second component in (2) is negligible relative to the first component. We can then omit the derivation of \mathbf{v}_2 which, as noted by Shao and Steel (1999), may be quite tedious. However, the estimator $\hat{\theta}$ itself may be biased if the assumed response mechanism assumed model does not hold.

2. FRAMEWORK AND ASSUMPTIONS

We assume that \mathbf{x}_i is known for all the units in the sample. Suppose a random sample, s , of size n is selected according to some design $p(s)$ from the population P . Let s_r be the sample of respondents to y of size r and let s_m be the sample of nonrespondents to y of size m ; $r + m = n$.

We consider two distinct frameworks: the design-based framework and the model-based framework. Under the design-based framework, we assume a uniform mechanism within cells:

Assumption DB: Within an imputation cell, the response probability for a given variable of interest is a constant and the responses statuses for different units are independent.

Under the model-based framework, we assume:

Assumption MB: 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 an imputation

model. We consider linear regression imputation in which case the imputation model is given by

$$\begin{aligned} E_m(y_i) &= \mathbf{z}'_i \boldsymbol{\eta}, \quad V_m(y_i) = \sigma_i^2 = \sigma^2 \mathbf{z}'_i \boldsymbol{\lambda}, \\ \text{Cov}_m(y_i, y_j) &= 0 \text{ if } i \neq j \end{aligned} \quad (3)$$

where $\boldsymbol{\eta}$ is q -vector of unknown parameters, \mathbf{z}_i is a q -vector of auxiliary variables available for all $i \in s$, $\boldsymbol{\lambda}$ is q -vector of specified constants, σ^2 is an unknown parameter. The restriction $\sigma_i^2 = \sigma^2 \mathbf{z}'_i \boldsymbol{\lambda}$ does not severely restrict the range of imputation models. A particular case of (3) leads to the ratio imputation model given by

$$\begin{aligned} E_m(y_i) &= \eta z_i, \quad V_m(y_i) = \sigma^2 z_i, \\ \text{Cov}_m(y_i, y_j) &= 0 \text{ if } i \neq j \end{aligned} \quad (4)$$

where η and z_i are the scalar versions of $\boldsymbol{\eta}$ and \mathbf{z}_i respectively. For simplicity, we consider the case of a single imputation cell.

3. MODEL-BASED FRAMEWORK

3.1 Regression Imputation

Regression imputation under model (3) uses the predicted value $y_i^* = \mathbf{z}'_i \hat{\boldsymbol{\eta}}_r$ for missing y_i , where $\hat{\boldsymbol{\eta}}_r$ is the weighted least squares estimator of $\boldsymbol{\eta}$ given by $\hat{\boldsymbol{\eta}}_r = \left(\sum_{s_r} w_i \mathbf{z}_i \mathbf{z}'_i / \sigma_i^2 \right)^{-1} \sum_{s_r} w_i \mathbf{z}_i y_i / \sigma_i^2$ and w_i is the survey weight attached to sample unit i . Using the y_i^* 's, an imputed estimator of \mathbf{B}_N is given by

$$\hat{\mathbf{B}}_1 = \left(\sum_s w_i \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left[\sum_{s_r} w_i \mathbf{x}_i y_i + \sum_{s_m} w_i \mathbf{x}_i y_i^* \right]. \quad (5)$$

The estimator $\hat{\mathbf{B}}_1$ is approximately design-model unbiased under ignorable response and model (3), i.e., $E_p E_m(\hat{\mathbf{B}}_1 - \mathbf{B}_N) \approx 0$.

3.2 Variance estimation

We first write the imputed estimator $\hat{\mathbf{B}}_1$ as

$$\hat{\mathbf{B}}_1 = \hat{\mathbf{T}}^{-1} \left[\hat{\mathbf{Y}}_{ax} + (\hat{\mathbf{Z}}_x - \hat{\mathbf{Z}}_{ax}) \hat{\mathbf{V}}_a^{-1} \hat{\mathbf{t}}_a \right] = \varphi(\hat{\mathbf{U}})$$

where φ is a smooth function of totals, $\hat{\mathbf{U}} = (\hat{\mathbf{T}}, \hat{\mathbf{Y}}_{ax}, \hat{\mathbf{Z}}_x, \hat{\mathbf{Z}}_{ax}, \hat{\mathbf{V}}_a^{-1}, \hat{\mathbf{t}}_a)^\top$, with $\hat{\mathbf{T}} = \sum_s w_i \mathbf{x}_i \mathbf{x}'_i$,

$$\begin{aligned}\hat{Y}_{ax} &= \sum_s w_i a_i x_i y_i, \hat{Z}_x = \sum_s w_i x_i z_i', \hat{Z}_{ax} = \sum_s w_i a_i x_i z_i', \\ \hat{V}_a &= \sum_s w_i a_i z_i z_i' / \sigma_i^2, \hat{t}_a = \sum_s w_i a_i z_i y_i / \sigma_i^2\end{aligned}$$

and a_i is a response indicator such that $a_i = 1$ if $i \in s_r$ and $a_i = 0$ if $i \notin s_r$. It follows from (2) that the variance-covariance matrix $V(\hat{\mathbf{B}}_1)$ of $\hat{\mathbf{B}}_1$ can be estimated by $\mathbf{v}_1 = \mathbf{v}_1 + \mathbf{v}_2$, where \mathbf{v}_1 is an estimator of $V_p(\hat{\mathbf{B}}_1 - \mathbf{B})$ conditional on the a_i 's, and \mathbf{v}_2 is an estimator of $V_r E_p(\hat{\mathbf{B}}_1 - \mathbf{B})$. Denote the estimator of the variance-covariance matrix of $\hat{\mathbf{Y}} = \sum_s w_i y_i$ as $v(y_i)$. One can show, using Taylor linearization, that \mathbf{v}_1 reduces to:

$$\mathbf{v}_1 = \mathbf{v}(\hat{\xi}_i), \quad (6)$$

where

$$\hat{\xi}_i = \hat{\mathbf{T}}^{-1}(\hat{\xi}_{1i} - \mathbf{x}_i \mathbf{x}_i' \hat{\mathbf{B}}_1),$$

with

$$\begin{aligned}\hat{\xi}_{1i} &= a_i x_i y_i + (1 - a_i) x_i z_i' \hat{\eta}_r + \\ &\frac{a_i}{\sigma_i^2} (\hat{Z}_x - \hat{Z}_{ax}) \hat{V}_a^{-1} z_i (y_i - z_i' \hat{\eta}_r)\end{aligned}$$

To obtain \mathbf{v}_2 , let us first write $E_p(\hat{\mathbf{B}}_1 - \mathbf{B}_N) \approx \phi(E_p(\hat{\mathbf{U}})) - \mathbf{B}_N = \phi(\tilde{\mathbf{U}})$, ϕ is a smooth function of totals, and

$$\tilde{\mathbf{U}} = (\mathbf{T}, \mathbf{Y}_{ax}, \mathbf{Z}_x, \mathbf{Z}_{ax}, \mathbf{V}_a, \mathbf{t}_a)'$$

with

$$\begin{aligned}\mathbf{T} &= \sum_p x_i x_i', \mathbf{Y}_{ax} = \sum_p a_i x_i y_i, \mathbf{Z}_x = \sum_p x_i z_i', \mathbf{Z}_{ax} = \sum_p a_i x_i z_i', \\ \mathbf{V}_a &= \sum_p a_i z_i z_i' / \sigma_i^2 \text{ and } \mathbf{t}_a = \sum_p a_i z_i y_i / \sigma_i^2. \text{ Note that } \phi(\tilde{\mathbf{U}}) \text{ can be written as } \phi(\tilde{\mathbf{U}}) = \sum_p \mathbf{c}_i y_i \text{ where}\end{aligned}$$

$$\mathbf{c}_i = \mathbf{T}^{-1} \left[(\mathbf{Z}_x - \mathbf{Z}_{ax}) \mathbf{V}_a^{-1} \frac{a_i}{\sigma_i^2} z_i \right] - (1 - a_i) \mathbf{x}_i.$$

Now,

$$\begin{aligned}\tilde{V}_m E_p(\hat{\mathbf{B}}_1 - \mathbf{B}_N) &= \tilde{V}_m(\phi(\tilde{\mathbf{U}})) \\ &= E_r V_m[\phi(\tilde{\mathbf{U}})] + V_r E_m[\phi(\tilde{\mathbf{U}})].\end{aligned} \quad (7)$$

The second term in the right-hand side of (7) is zero because $E_p E_m(\hat{\mathbf{B}}_1 - \mathbf{B}_N) = 0$. Also,

$$E_r V_m(\phi(\tilde{\mathbf{U}})) = \sigma^2 \sum_p E_r(\mathbf{c}_i \mathbf{c}_i') z_i' \lambda \quad (8)$$

The component \mathbf{v}_2 , obtained by substituting estimators for the unknown quantities in (8), is given by

$$\mathbf{v}_2 = \sigma^2 \sum_s w_i \mathbf{c}_i \mathbf{c}_i' z_i' \lambda \quad (9)$$

where

$$\hat{\sigma}^2 = \frac{\sum_s w_i a_i (y_i - z_i' \hat{\eta}_r)^2}{\sum_s w_i a_i}$$

is a model-consistent estimator of σ^2 and $\hat{\mathbf{c}}_i = \hat{\mathbf{T}}^{-1} \left[(\hat{Z}_x - \hat{Z}_{ax}) \hat{V}_a^{-1} \frac{a_i}{\hat{\sigma}^2 z_i' \lambda} z_i \right] - (1 - a_i) \mathbf{x}_i$. The sum of (6) and (9) gives the variance estimator \mathbf{v}_1 .

4. DESIGN-BASED FRAMEWORK

In this section, we consider two imputed estimators of the regression coefficient \mathbf{B}_N under assumption DB: the unadjusted estimator and the adjusted estimator. Since it is difficult to justify regression imputation in the design-based framework, we confine ourselves to the case of ratio imputation.

4.1 The unadjusted estimator

Under ratio imputation, the bias of the unadjusted estimator (5) under uniform response is given by

$$\text{Bias}(\hat{\mathbf{B}}_1) = (1 - p) \left(\frac{\bar{Y}}{\bar{Z}} \mathbf{B}_Z - \mathbf{B}_N \right) \quad (10)$$

where $p = P(i \in s_r)$ is the probability of response,

$$(\bar{Y}, \bar{Z}) = \frac{1}{N} \sum_p (y_i, z_i) \mathbf{B}_Z = \mathbf{T}^{-1} \mathbf{X}_Z \quad \text{with}$$

$\mathbf{T} = \sum x_i x_i'$, and $\mathbf{X}_Z = \sum x_i z_i$. The bias will be zero in the full response case (i.e. $p = 1$) or if

$\mathbf{B}_N = \frac{\bar{Y}}{\bar{Z}} \mathbf{B}_Z$. It is interesting to note that in the case

of a domain mean \bar{Y}_d , the bias given in (10) is 0 if $R = R_d$ where $R = \frac{\bar{Y}}{\bar{Z}}$ is the overall population

ratio, $R_d = \frac{\bar{Y}_d}{\bar{Z}_d}$ is the domain population ratio, with

$$(\bar{Y}_d, \bar{Z}_d) = \frac{\sum_p x_i (y_i, z_i)}{\sum_p x_i}. \quad \text{Note that the unadjusted}$$

estimator $\hat{\mathbf{B}}_1$ is design-model unbiased provided the ratio imputation model (4) is true.

4.2 The adjusted estimator

Following Rao and Skinner (1999), a simple bias-adjusted estimator for the regression coefficient \mathbf{B}_N , is given by

$$\hat{\mathbf{B}}_1^a = \hat{p}^{-1} \hat{\mathbf{B}}_1 + (1 - \hat{p}^{-1}) \hat{\mathbf{B}}_z \frac{\bar{y}_l}{\bar{z}} \quad (11)$$

where \hat{p} is an estimate of p , $\bar{y}_l = \frac{1}{\sum_s w_i} (\sum_{s_r} w_i y_i + \sum_{s_m} w_i y_i^*) = \frac{\bar{y}_r}{\bar{z}}$ is an approximately design-unbiased imputed estimator, under uniform response, of the overall mean \bar{Y} , $\hat{\mathbf{B}}_z = \hat{\mathbf{T}}^{-1} \hat{\mathbf{X}}_z$ with $\hat{\mathbf{T}} = \sum_s w_i \mathbf{x}_i \mathbf{x}_i'$, $\hat{\mathbf{X}}_z = \sum_s w_i \mathbf{x}_i z_i$ and $\bar{z} = \sum_s w_i z_i$. It is easy to show that the adjusted estimator $\hat{\mathbf{B}}_1^a$ is approximately design-unbiased for the regression coefficient \mathbf{B}_N under uniform response. Also, note that the adjusted estimator $\hat{\mathbf{B}}_1^a$ is design-model unbiased provided the imputation model (4) is true. Hence, unlike the unadjusted estimator, the adjusted estimator (11) is robust in the sense of validity under both frameworks.

4.3 Variance estimator

In this section, we obtain consistent variance estimators for the adjusted estimator under both the design-based and model-based frameworks. Once again, we use the approach of Fay (1991). The derivation involves tedious but straightforward algebra. First, we express the adjusted estimator (11) as

$$\hat{\mathbf{B}}_1^a = \frac{\hat{N}}{\hat{I}_a} \hat{\mathbf{T}}^{-1} [\hat{\mathbf{Y}}_{\text{xa}} + (\hat{\mathbf{Z}}_x + \hat{\mathbf{Z}}_{\text{ax}}) \hat{R}_a] + \left(1 - \frac{\hat{N}}{\hat{I}_a}\right) \hat{\mathbf{B}}_z \frac{1}{\hat{Z}} [\hat{Y}_a + (\hat{Z} - \hat{Z}_a) \hat{R}_a]$$

where

$$\hat{N} = \sum_s w_i, \hat{I}_a = \sum_s w_i a_i, \hat{\mathbf{T}} = \sum_s w_i \mathbf{x}_i \mathbf{x}_i',$$

$$\hat{\mathbf{Y}}_{\text{xa}} = \sum_s w_i a_i \mathbf{x}_i y_i, \hat{\mathbf{Z}}_x = \sum_s w_i \mathbf{x}_i z_i,$$

$$\hat{\mathbf{Z}}_{\text{ax}} = \sum_s w_i a_i \mathbf{x}_i z_i', \hat{\mathbf{B}}_z = \hat{\mathbf{T}}^{-1} \hat{\mathbf{Z}}_z,$$

$$\hat{Z} = \sum_s w_i z_i, \hat{Z}_a = \sum_s w_i a_i z_i, \hat{Y}_a = \sum_s w_i a_i y_i$$

and $\hat{R}_a = \frac{\hat{Y}_a}{\hat{Z}_a}$. Then, one can show, using Taylor

linearization, that \mathbf{v}_1 reduces to:

$$\mathbf{v}_1 = \mathbf{v}(\hat{\xi}_i), \quad (12)$$

where

$$\hat{\xi}_i = \frac{1}{\hat{I}_a} (1 - a_i \hat{p}^{-1}) [\hat{\mathbf{B}}_1 - \hat{\mathbf{B}}_z \hat{R}_a] + \hat{p}^{-1} \hat{\mathbf{T}}^{-1} [\hat{\xi}_{1i} - \mathbf{x}_i \mathbf{x}_i' \hat{\mathbf{B}}_1] + (1 - \hat{p}^{-1}) \hat{\mathbf{T}}^{-1} [\hat{\xi}_{4i} - \mathbf{x}_i \mathbf{x}_i' \hat{\mathbf{B}}_z \hat{R}_a]$$

with

$$\hat{\xi}_{1i} = a_i \mathbf{x}_i y_i + (1 - a_i) \hat{R}_a \mathbf{x}_i z_i + (\hat{\mathbf{Z}}_x - \hat{\mathbf{Z}}_{\text{ax}}) \frac{1}{\hat{Z}_a} a_i (y_i - \hat{R}_a z_i),$$

$$\hat{\xi}_{4i} = \mathbf{x}_i \hat{R}_a z_i + \hat{\mathbf{Z}}_x \hat{\xi}_{3i},$$

$$\hat{\xi}_{3i} = \frac{1}{\hat{Z}} \left(\hat{\xi}_{2i} - \frac{\bar{y}_l}{\bar{z}} z_i \right),$$

and

$$\hat{\xi}_{2i} = a_i y_i + (1 - a_i) \hat{R}_a z_i + \frac{\hat{Z} - \hat{Z}_a}{\hat{Z}_a} a_i (y_i - \hat{R}_a z_i).$$

Note that \mathbf{v}_1 is valid under both frameworks. Denote the second component of the variance estimator under the design-based and the model-based frameworks by $\mathbf{v}_{2_{DB}}$ and $\mathbf{v}_{2_{MB}}$ respectively. We obtain

$$\mathbf{v}_{2_{DB}} = \hat{p} (1 - \hat{p}) \sum_s w_i a_i \hat{\mathbf{c}}_{1i} \hat{\mathbf{c}}_{1i}' \quad (13)$$

where

$$\hat{\mathbf{c}}_{1i} = \hat{p}^{-1} [\hat{\mathbf{T}}^{-1} \hat{\xi}_{5i} - \hat{\mathbf{B}}_1] + \hat{\mathbf{B}}_z \left[(1 - \hat{p}^{-1}) \tilde{\xi}_{6i} + \hat{p}^{-1} \frac{\bar{y}_l}{\bar{z}} \right],$$

$$\tilde{\xi}_{5i} = \left(\mathbf{x}_i + \frac{\hat{\mathbf{Z}}_x - \hat{\mathbf{Z}}_{\text{ax}}}{\hat{Z}_a} \right) (y_i - \hat{R}_a z_i)$$

and $\tilde{\xi}_{6i} = \frac{1}{\hat{Z}_a} (y_i - \hat{R}_a z_i)$. Also,

$$\mathbf{v}_{2_{MB}} = \hat{\sigma}^2 \sum_s w_i \hat{\mathbf{c}}_{2i} \hat{\mathbf{c}}_{2i}' \quad (14)$$

where

$$\hat{\mathbf{c}}_{2i} = \hat{p}^{-1} \hat{\mathbf{T}}^{-1} a_i \left[\mathbf{x}_i + \frac{\hat{\mathbf{Z}}_x - \hat{\mathbf{Z}}_{\text{ax}}}{\hat{Z}_a} \right] +$$

$$(1 - \hat{p}^{-1}) \hat{\mathbf{T}}^{-1} \frac{1}{\hat{Z}_a} a_i - \hat{\mathbf{T}}^{-1} \mathbf{x}_i$$

and $\hat{\sigma}^2 = \frac{\sum_s w_i a_i (y_i - \hat{R}_a z_i)^2}{\sum_s w_i a_i z_i}$ is a model-consistent

and asymptotically model-unbiased estimator of σ^2 . The variance estimator is then $\mathbf{v}_1 + \mathbf{v}_{2_{DB}}$ under the design-based framework, and $\mathbf{v}_1 + \mathbf{v}_{2_{MB}}$ under the model-based framework.

5. SIMULATION STUDY

We conducted a small simulation study using a population of size $N=2000$ (Lohr 1999, Appendix C, p. 441) containing two domains (males and females) of sizes $N_d=1000, d=1, 2$. The variable of interest, y , is the variable *Height* in cm. The goal is to estimate the vector of domain means $\mathbf{B}_N = (\bar{Y}_1, \bar{Y}_2)$. Note that this parameter is a particular case of (1) with $\mathbf{x}_i = (x_{1i}, x_{2i})$ and x_{1i} and x_{2i} are the domain indicators for domain 1 and domain 2 respectively. We drew $R=5000$ repeated simple random samples without replacement, s , of size varying between 50 and 120. Nonresponse was generated according to an uniform response mechanism. The response rates were set at 0.7 and 0.8. For simplicity, we used mean imputation.

To measure the relative bias of an estimator $\hat{\theta}$, we used $B_{rel}(\hat{\theta}) = \text{Bias}(\hat{\theta}) / \text{s.e.}(\hat{\theta})$. Table 1 reports the simulation results of relative bias of the unadjusted and adjusted estimators. It is clear from Table 1 that the relative bias of the unadjusted estimator is substantial while that of the adjusted estimator is negligible, as expected. Also, it might be of interest to test for the equality of domain means or to construct confidence intervals for the difference of domain means. Suppose we are interested in constructing confidence intervals for the difference $\bar{Y}_1 - \bar{Y}_2$. We may then construct an interval based on the difference of unadjusted estimates, given by $(\bar{y}_{1l} - \bar{y}_{2l}) \pm [v_1(\bar{y}_{1l} - \bar{y}_{2l})]^{1/2}$, or on the difference of adjusted estimates, given by $(\bar{y}_{1l}^a - \bar{y}_{2l}^a) \pm [v_1(\bar{y}_{1l}^a - \bar{y}_{2l}^a)]^{1/2}$. Table 2 reports the coverage probability of the two confidence intervals with a 95% nominal rate. It is clear from Table 2 that one should use a confidence interval based on the difference of adjusted estimators. Note that the difference of the adjusted estimates $\bar{y}_{2l}^a - \bar{y}_{1l}^a$ is $\hat{p}^{-1}(\bar{y}_{2l} - \bar{y}_{1l})$ so the pivotal quantity t used for the adjusted case reduces to

$$t \approx \frac{p^{-1}(\bar{y}_{1l} - \bar{y}_{2l}) - (\bar{Y}_1 - \bar{Y}_2)}{\sqrt{v_1 [p^{-1}(\bar{y}_{2l} - \bar{y}_{1l})]}} \quad (15)$$

From (15), it follows that the confidence intervals based on the difference of the adjusted estimates will have approximately the same coverage as the interval based on the unadjusted estimates when $\bar{Y}_1 \approx \bar{Y}_2$ or when $p \approx 1$, i.e., small nonresponse rate.

Table 1: Relative bias of the unadjusted and adjusted estimators of $\mathbf{B}_N = (\bar{Y}_1, \bar{Y}_2)$

p	n	$B_{rel}(\hat{\mathbf{B}}_1)$	$B_{rel}(\hat{\mathbf{B}}_1^a)$
0.7	50	(1.11, -0.91)	(0.024, -0.031)
	80	(1.40, -1.15)	(0.024, -0.031)
	12	(1.77, -1.40)	(0.038, -0.010)
	0		
0.8	50	(0.78, -0.61)	(-0.0036, -0.0061)
	80	(1.01, 0.78)	(0.027, -0.0009)
	12	(1.23, -1.00)	(0.0022, -0.0030)
	0		

Table 2: Coverage probability of the confidence intervals based on unadjusted and adjusted estimators

p = 0.8	unadjusted	adjusted
50	0.763	0.942
80	0.699	0.948
120	0.593	0.949

Finally, note that both the unadjusted and adjusted estimates possess the additive property for totals; that is, the sum of unadjusted (adjusted) estimators for the total in each domain is equal to the overall imputed estimator for the overall population total.

6. CONCLUSION

This article focused on estimating census regression coefficients under imputation for missing data. We have proposed a simple adjusted estimator, which, in our view, is attractive since its use can be justified under both the design-based and the model-based frameworks. We also derived consistent variance estimators using an approach proposed by Fay (1991) and developed by Shao and Steel (1999). Extension the case of stratified multistage sampling is under investigation. We are also investigating the extension to general functions $z(x, y)$ such as regression and correlation coefficients and cell proportions in two-way tables, for which both x and y may be missing.

REFERENCES

- Binder, D. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review*, 51, pp. 279-292.
- 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.
- Haziza, D. and Rao, J.N.K. (2000). Inference for Domain Means under Imputation for Missing Data. *Proceedings of the SSC Annual Meeting, Survey Methods Section*, pp. 197-201.
- Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press.
- 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.
- Skinner, C.J. and Rao, J.N.K. (1999). Jackknife Variance for Multivariate Statistics under Hot-deck Imputation from Common Donors. *Journal of Statistical Planning and Inference*, in press.