Sample Correlation Coefficients Based on Survey Data Under Regression Imputation

Jun SHAO and Hansheng WANG

Regression imputation is commonly used to compensate for item nonresponse when auxiliary data are available. It is common practice to compute survey estimators by treating imputed values as observed data and using the standard unbiased (or nearly unbiased) estimation formulas designed for the case of no nonresponse. Although the commonly used regression imputation method preserves unbiasedness for population marginal totals (i.e., survey estimators computed from imputed data are still nearly unbiased), it does not preserve unbiasedness for population correlation coefficients. A joint regression imputation method is proposed that preserves unbiasedness for marginal totals, second moments, and correlation coefficients. Some simulation results show that the joint regression imputation method produces not only sample correlation coefficients that are nearly unbiased, but also estimates that are more stable than those produced by marginal nonrandom regression imputation when correlation coefficients are in a certain range. Variance estimation for sample correlation coefficients under joint regression imputation is also studied, using a jackknife method that takes imputation into account.

KEY WORDS: Jackknife; Joint imputation; Marginal imputation; Random imputation; Variance estimation.

1. INTRODUCTION

Most surveys have nonresponse. *Item nonresponse* occurs when a sampled unit fails to provide information on some items (variables) in the survey. Imputation is commonly applied to compensate for item nonresponse. In addition to some practical reasons for imputation (Kalton and Kasprzyk 1986), imputation using auxiliary data may produce survey estimators that are more efficient than the one obtained by ignoring nonrespondents and reweighting. It is a common practice to compute survey estimators by treating imputed values as observed data and using the standard unbiased (or nearly unbiased) estimation formulas designed for the case of no nonresponse (e.g., standard survey estimators for population item totals and correlation coefficients among items are sample weighted totals and sample correlation coefficients). Therefore, we should select an imputation method that preserves the unbiasedness in the sense that the survey (point) estimators computed from imputed data using standard formulas are still unbiased or nearly unbiased. Throughout this article, an imputation method is called *unbiased* for estimating a given set of parameters if it preserves the unbiasedness. As we discuss later, the unbiasedness of an imputation method often depends on the type of parameters to be estimated.

For item nonresponse, imputing the whole vector of items whenever a sampled unit has at least one missing item is rarely used, because it discards many observed data. *Marginal imputation*, also called *item imputation*, is often used in practice; that is, items in a survey are imputed separately and the relationship among items is ignored. Marginal imputation is often unbiased for estimating functions of marginal population item totals (or means). For parameters measuring relationship among items such as the population correlation coefficients among items, marginal imputation may not be unbiased, because the relationship is not preserved during marginal imputation.

The main purpose of this article is to study (a) the unbiasedness of a popular imputation method, the *regression imputation* method (see, e.g., Deville and Särndal 1994), for estimating not only marginal totals, but also correlation coefficients among items, and (b) the variance estimation problem for an unbiased regression imputation method. The latter is an important issue because it is well known that, even for an imputation method that produces unbiased or nearly unbiased point estimators, treating imputed values as observed data and applying standard variance estimation formulas designed for the case of no nonresponse may produce seriously biased variance estimators.

In Section 2 we introduce the commonly used marginal regression imputation method and a *joint regression imputation* method, which is an extension of that of Srivastava and Carter (1986). The joint regression imputation method is shown to be unbiased for estimating marginal totals as well as correlation coefficients. We devote Section 3 to the variance estimation problem for sample correlation coefficients under joint regression imputation. We propose a jackknife method that takes imputation into account and produces asymptotically unbiased and consistent variance estimators. We present some simulation results in Section 4 to examine the finite-sample performance of joint regression imputation and the jackknife variance estimator. Simulation results show that the joint regression imputation method produces sample correlation coefficients that are not only nearly unbiased, but also more stable than those produced by marginal nonrandom regression imputation when correlation coefficients are in a certain range. Simulation results also show that the proposed jackknife variance estimator performs well in terms of the bias in variance estimation and the coverage probability of confidence intervals using the jackknife variance estimators.

2. REGRESSION IMPUTATION

Let $\mathcal{P}$ be a finite population containing M units and let $\mathcal{S}$ be a sample from $\mathcal{P}$ obtained according to the following commonly used stratified multistage sampling plan. The population $\mathcal{P}$ is stratified into H strata with N_h clusters in the hth stratum. In the first-stage sampling, $n_h \geq 2$ clusters are selected

Jun Shao is Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706 (E-mail: *shao@stat.wisc.edu*). Hansheng Wang is Principle Statistician, StatPlus, Inc., Yardley, PA 19067. The authors thank the referees for helpful comments and suggestions. The research of Jun Shao was partially supported by National Science Foundation grants DMS-9803112 and DMS-0102223 and National Security Agency grant MDA 904-99-1-0032.

Journal of the American Statistical Association
June 2002, Vol. 97, No. 458, Theory and Methods

without replacement from stratum h according to some probability sampling plan, and the clusters are selected independently across the strata. A second-stage sample, a third-stage sample, and so on may be taken from each sampled cluster, using some sampling plan independently across the sampled clusters. Associated with the ith unit in $\mathcal{P}$ is a vector $(y_i, z_i, \dots)$ of items of interest and a vector $\mathbf{x}_i$ of auxiliary variables (which is observed for all sampled units). For $i \in \mathcal{S}$, survey weights w_i are constructed so that when there is no nonresponse,

$$E_s\left(\sum_{i\in\mathcal{S}} w_i\psi_i\right) = \sum_{i\in\mathcal{P}} \psi_i,$$

for any set of values $\{\psi_i : i \in \mathcal{P}\}$, where E_s is the expectation with respect to the randomness from sampling $\mathcal{S}$.

2.1 Marginal Regression Imputation

As for any other statistical method, the validity of an imputation method relies on some assumptions and/or models. The regression imputation method is based on the following model assumption. The population $\mathcal{P}$ under consideration can be divided into subpopulations (called imputation classes) $\mathcal{P}_k$, $k = 1, \dots, K$, such that for an item y with nonresponse,

$$y_i = \beta_k'\mathbf{x}_i + v_{ki}^{1/2}\epsilon_i, \qquad i \in \mathcal{P}_k, \tag{1}$$

and

$$P(a_i = 1|y_i, \mathbf{x}_i) = P(a_i = 1|\mathbf{x}_i), \qquad i \in \mathcal{P}_k, \tag{2}$$

where β_k is a parameter vector and β_k' is its transpose; $v_{ki} = v_k(\mathbf{x}_i)$ with a known function $v_k(\cdot) > 0$; ϵ_i is a random error (independent of $\mathbf{x}_i$) with mean 0 and unknown variance $\sigma^2_{\epsilon,k} > 0$; $\beta_k, \sigma_{\epsilon,k}$, and $v_k(\cdot)$ may be different in different imputation classes or for different items; and a_i is the indicator of whether y_i is a respondent (i.e., $a_i = 1$ if y_i is observed and $a_i = 0$ if y_i is a nonrespondent). Note that (1) is a linear regression model with heteroscedastic errors and (2) means that given $\mathbf{x}$, the response probability for item y is independent of y. The creation of K imputation classes allows us to establish a reasonable model such as (1) in the imputation class $\mathcal{P}_k$ that may not hold for the whole population $\mathcal{P}$. Imputation classes are usually formed by using an auxiliary variable without nonresponse; for example, in many business surveys, imputation classes are strata or unions of strata.

Under (1)–(2), the nonrandom regression imputation method imputes a nonrespondent y_i in $\mathcal{S}_k = \mathcal{S} \cap \mathcal{P}_k$ by

$$y_i^* = \hat{\beta}_k'\mathbf{x}_i,$$

where

$$\hat{\beta}_k = \left(\sum_{i\in\mathcal{S}_k} w_i a_i v_{ki}^{-1}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i\in\mathcal{S}_k} w_i a_i v_{ki}^{-1}\mathbf{x}_i y_i \tag{3}$$

is the best linear unbiased estimator of β_k under (1) based on respondents in $\mathcal{S}_k$.

Once nonrespondents are imputed, survey estimators are computed by treating imputed data as observations and using the same formulas as in the case of no nonresponse. For the (marginal) total of item y, $Y = \sum_{i\in\mathcal{P}} y_i$, its standard unbiased estimator in the case of no nonresponse is $\widehat{Y} = \sum_{i\in\mathcal{S}} w_i y_i$. Hence its estimator based on imputed data is

$$\widehat{Y}^* = \sum_{i\in\mathcal{S}} w_i a_i y_i + \sum_{i\in\mathcal{S}} w_i(1 - a_i)y_i^*. \tag{4}$$

Let E_m denote the expectation under model (1)–(2) and let E_s be the expectation under sampling, given item values generated by model (1)–(2). Then $E_m E_s(\widehat{Y}^*) \approx E_m(Y)$, where $\approx$ is the equality up to a negligible term under the asymptotic setting described in Appendix A.1. That is, $\widehat{Y}^*$ is asymptotically unbiased for Y.

In this article we consider the estimation of the correlation coefficient between two items, say y and z, defined by

$$\rho = \left(\sum_{i\in\mathcal{P}} y_i z_i - YZ/M\right) \Bigg/ \left[\left(\sum_{i\in\mathcal{P}} y_i^2 - Y^2/M\right)\left(\sum_{i\in\mathcal{P}} z_i^2 - Z^2/M\right)\right]^{1/2},$$

where $Z = \sum_{i\in\mathcal{P}} z_i$. When there is no nonresponse, a standard estimator of ρ is the sample correlation coefficient

$$\hat{\rho} = \left(\sum_{i\in\mathcal{S}} w_i y_i z_i - \widehat{Y}\widehat{Z}/\widehat{M}\right) \Bigg/ \left[\left(\sum_{i\in\mathcal{S}} w_i y_i^2 - \widehat{Y}^2/\widehat{M}\right)\left(\sum_{i\in\mathcal{S}} w_i z_i^2 - \widehat{Z}^2/\widehat{M}\right)\right]^{1/2},$$

where $\widehat{Y}$ is as defined previously, $\widehat{M} = \sum_{i\in\mathcal{S}} w_i$ (M may be unknown in many surveys), and $\widehat{Z} = \sum_{i\in\mathcal{S}} w_i z_i$. Note that $\hat{\rho}$ is not exactly unbiased. Because $\widehat{Y}, \widehat{Z}, \sum_{i\in\mathcal{S}} w_i y_i z_i, \sum_{i\in\mathcal{S}} w_i y_i^2$, and $\sum_{i\in\mathcal{S}} w_i z_i^2$ are unbiased for $Y, Z, \sum_{i\in\mathcal{P}} y_i z_i, \sum_{i\in\mathcal{P}} y_i^2$, and $\sum_{i\in\mathcal{P}} z_i^2$, $\hat{\rho}$ is asymptotically unbiased under the asymptotic setting given in Appendix A.1.

Consider now the situation in which both y and z have nonrespondents. A model for item z analogous to (1)–(2) is

$$z_i = \gamma_k'\mathbf{x}_i + u_{ki}^{1/2}\eta_i, \qquad i \in \mathcal{P}_k \tag{5}$$

and

$$P(b_i = 1|z_i, \mathbf{x}_i) = P(b_i = 1|\mathbf{x}_i), \qquad i \in \mathcal{P}_k, \tag{6}$$

where γ_k is unknown, $u_{ki} = u_k(\mathbf{x}_i)$ with a known function $u_k(\cdot) > 0$, and η_i is a random error (independent of $\mathbf{x}_i$) with mean 0 and unknown variance $\sigma^2_{\eta,k} > 0$. Note that ϵ_i in (1) and η_i in (5) are generally dependent. Similarly, a_i in (2) and b_i in (6) are generally dependent.

Nonrandom regression imputed value of a missing z_i in $\mathcal{S}_k$ is $z_i^* = \hat{\gamma}_k'\mathbf{x}_i$, where $\hat{\gamma}_k$ is an analog of $\hat{\beta}_k$ for item z. Suppose that nonrespondents for y and z are imputed marginally (separately). For simplicity, let $y_i^* = y_i$ if y_i is observed and $y_i^* =$ the imputed value if y_i is a nonrespondent, and let z_i^* be defined similarly. Then the sample correlation coefficient based on the imputed data is

$$\hat{\rho}^* = \left(\sum_{i\in\mathcal{S}} w_i y_i^* z_i^* - \widehat{Y}^*\widehat{Z}^*/\widehat{M}\right) \Bigg/ \left[\left(\sum_{i\in\mathcal{S}} w_i y_i^{*2} - \widehat{Y}^*/\widehat{M}\right)\left(\sum_{i\in\mathcal{S}} w_i z_i^{*2} - \widehat{Z}^*/\widehat{M}\right)\right]^{1/2}, \tag{7}$$

where $\widehat{Y}^*$ is defined in (4) and $\widehat{Z}^*$ is similarly defined.

For marginal imputation, the previous discussion can be extended to the case of more than two items in a straightforward way.

It can be shown that $\sum_{i\in\mathcal{S}} w_i y_i^{*2}$ and $\sum_{i\in\mathcal{S}} w_i y_i^* z_i^*$ have asymptotic biases

$$-E_m\left(\sum_k \sum_{i\in\mathcal{P}_k} (1-a_i) v_{ki} \sigma^2_{\epsilon,k}\right)$$

and

$$-E_m\left(\sum_k \sum_{i\in\mathcal{P}_k} (1-a_i b_i) v_{ki}^{1/2} u_{ki}^{1/2} \sigma_{\epsilon,\eta,k}\right), \tag{8}$$

where $\sigma_{\epsilon,\eta,k} = E_m(\epsilon_i \eta_i)$ for $i \in \mathcal{P}_k$ (details can be found in a technical report). These biases are of a nonnegligible order. Consequently, $\hat{\rho}^*$ in (7) based on marginal nonrandom regression imputation is not unbiased for the correlation coefficient ρ.

We now consider the *random regression imputation* method, which adds a random term to the regression imputed nonrespondent; that is, a nonrespondent y_i in $\mathcal{S}_k$ is imputed by

$$y_i^* = \hat{\beta}_k' \mathbf{x}_i + v_{ki}^{1/2} \epsilon_i^*, \tag{9}$$

where, given the observed data, the ϵ_i^*'s are independent with mean 0 and variance

$$\hat{\sigma}^2_{\epsilon,k} = \sum_{i\in\mathcal{S}_k} w_i a_i v_{ki}^{-1} (y_i - \hat{\beta}_k' \mathbf{x}_i)^2 \Big/ \sum_{i\in\mathcal{S}_k} w_i a_i$$

(a consistent and asymptotically unbiased estimator of $\sigma^2_{\epsilon,k}$). Adding random terms is common in imputation (Rubin and Schenker 1986; Srivastava and Carter 1986; Rubin 1987) and confidentiality editing (Kim 1986).

The random regression imputation method imputes a nonrespondent z_i in $\mathcal{S}_k$ by

$$z_i^* = \hat{\gamma}_k' \mathbf{x}_i + u_{ki}^{1/2} \eta_i^*, \tag{10}$$

where, given the observed data, the η_i^*'s are independent with mean 0 and variance

$$\hat{\sigma}^2_{\eta,k} = \sum_{i\in\mathcal{S}_k} w_i b_i u_{ki}^{-1} (z_i - \hat{\gamma}_k' \mathbf{x}_i)^2 \Big/ \sum_{i\in\mathcal{S}_k} w_i b_i.$$

For marginal imputation, items y and z are imputed according to (9) and (10) separately, and the ϵ_i^*'s and η_i^*'s are generated independently.

It can be shown that $E_m E_s E_*(\sum_{i\in\mathcal{S}} w_i y_i^*) \approx E_m(Y)$ and $E_m E_s E_*(\sum_{i\in\mathcal{S}} w_i y_i^{*2}) \approx E_m(\sum_{i\in\mathcal{P}} y_i^2)$, where E_* is the expectation with respect to the random term in the imputation. This means that the random regression imputation method is unbiased for the marginal first and second moments. However, as an estimator of the cross-product moment, $\sum_{i\in\mathcal{S}} w_i y_i^* z_i^*$ has an asymptotic bias given by (8). Thus $\hat{\rho}^*$ based on marginal random regression imputation is asymptotically biased, unless $\sigma_{\epsilon,\eta,k} = 0$ for all k (i.e., y_i and z_i are conditionally uncorrelated, given $\mathbf{x}_i$).

2.2 Joint Regression Imputation

To obtain an unbiased imputation method for the correlation coefficient, some type of joint imputation is required. Under parametric models, there exist efficient imputation methods (see, e.g., Rubin and Schenker 1986; Little and Rubin 1987; Schafer 1997). For complex survey data, however, it is often hard to find a suitable parametric model. Note that the regression imputation method discussed in this article is not a parametric method, because no assumption is imposed on the form of the distribution of ϵ_i and η_i.

Consider first the case in which only y and z are the items of interest (i.e., the bivariate case). We propose a joint regression imputation method that is the same as the random regression imputation method described in Section 2.1 [i.e., missing y_i's and z_j's are imputed according to (9) and (10)], except that the random terms ϵ_i^* and η_j^* are generated according to the following scheme:

1. In the kth imputation class, when $a_i = 0$ but $b_i = 1$ (y_i is missing but z_i is observed),

$$\epsilon_i^* = \frac{\hat{\sigma}_{\epsilon,\eta,k}}{u_{ki}^{1/2} \hat{\sigma}^2_{\eta,k}} (z_i - \hat{\gamma}_k' \mathbf{x}_i) + \tilde{\epsilon}_i^*, \tag{11}$$

where, given the observed data, the $\tilde{\epsilon}_i^*$'s are independent random variables with mean 0 and variance $\hat{\sigma}^2_{\epsilon,k} - \hat{\sigma}^2_{\epsilon,\eta,k}/\hat{\sigma}^2_{\eta,k}$,

$$\begin{pmatrix} \hat{\sigma}^2_{\epsilon,k} & \hat{\sigma}_{\epsilon,\eta,k} \\ \hat{\sigma}_{\epsilon,\eta,k} & \hat{\sigma}^2_{\eta,k} \end{pmatrix} = \sum_{i\in\mathcal{S}_k} w_i a_i b_i \begin{pmatrix} r^2_{y,ki} & r_{y,ki} r_{z,ki} \\ r_{y,ki} r_{z,ki} & r^2_{z,ki} \end{pmatrix} \Big/ \sum_{i\in\mathcal{S}_k} w_i a_i b_i, \tag{12}$$

$r_{y,ki} = v_{ki}^{-1/2}(y_i - \hat{\beta}_k' \mathbf{x}_i)$, and $r_{z,ki} = u_{ki}^{-1/2}(z_i - \hat{\gamma}_k' \mathbf{x}_i)$.

2. In the kth imputation class, when $a_i = 1$ but $b_i = 0$,

$$\eta_i^* = \frac{\hat{\sigma}_{\epsilon,\eta,k}}{v_{ki}^{1/2} \hat{\sigma}^2_{\epsilon,k}} (y_i - \hat{\beta}_k' \mathbf{x}_i) + \tilde{\eta}_i^*, \tag{13}$$

where, given the observed data, the $\tilde{\eta}_i^*$'s are independent random variables with mean 0 and variance $\hat{\sigma}^2_{\eta,k} - \hat{\sigma}^2_{\epsilon,\eta,k}/\hat{\sigma}^2_{\epsilon,k}$.

3. In the kth imputation class, when both $a_i = 0$ and $b_i = 0$, the (ϵ_i^*, η_i^*)'s are independently (given the observed data) distributed with mean 0 and covariance matrix given in (12).

Note that there are two major differences between joint regression imputation and marginal random regression imputation:

1. When both y_i and z_i are nonrespondents, ϵ_i^* and η_i^* are independent in marginal random regression imputation, whereas for joint regression imputation, ϵ_i^* and η_i^* have covariance given by (12), which is a consistent estimator of the covariance between ϵ_i and η_i.

2. When exactly one of y_i and z_i is missing, ϵ_i^* or η_i^* is generated according to (11) or (13) to ensure that the imputed pair of data preserves the same correlation as that of the original pair. Conditional on the observed data, ϵ_i^* or η_i^* in (11) or (13) does not have mean 0, whereas ϵ_i^* or η_i^* in marginal random regression imputation has mean 0.

Consider now the multivariate case with $q \geq 3$ items and the estimation of correlation coefficients between all pairs of q items. Let $\mathbf{t}_i = (y_i, z_i, \dots)'$ be the q vector of item values from unit i. Model (1)–(2) and (5)–(6) should be replaced by

$$\mathbf{t}_i = B_k' \mathbf{x}_i + V_{ki}^{1/2} \mathbf{e}_i, \qquad i \in \mathcal{P}_k \tag{14}$$

and

$$P(\mathbf{a}_i = a | \mathbf{t}_i, \mathbf{x}_i) = P(\mathbf{a}_i = a | \mathbf{x}_i), \qquad i \in \mathcal{P}_k, \tag{15}$$

where $B_k = (\beta_k, \gamma_k, \dots)$ is a matrix of regression parameters; V_{ki} is a diagonal matrix whose diagonal elements are $v_{ki} = v_k(\mathbf{x}_i)$ and $u_{ki} = u_k(\mathbf{x}_i), \dots$, with known positive functions $v_k, u_k, \dots$; $\mathbf{e}_i$ is a random error (independent of $\mathbf{x}_i$) with mean 0 and unknown covariance matrix Σ_k; and $\mathbf{a}_i$ is the vector of response indicators for $\mathbf{t}_i$. Let $\widehat{B}_k = (\hat{\beta}_k, \hat{\gamma}_k, \dots)$ with $\hat{\beta}_k$ computed according to (3) and $\hat{\gamma}_k, \dots$, computed similarly, and let $\widehat{\Sigma}_k$ be a positive definite consistent estimator of Σ_k. Let $\nu \subset \{1, \dots, q\}$ be the set of indices corresponding to missing components in q items and let ν^c be the complement of ν. For each $\mathbf{t}_i$, let $\mathbf{t}_{\nu i}$ be the subvector of $\mathbf{t}_i$ containing all nonrespondents and let $\mathbf{t}_{\nu^c i}$ be the subvector containing all respondents. For any $q \times q$ matrix A and two subsets ν_1 and ν_2 of $\{1, \dots, q\}$, let $A^{(\nu_1, \nu_2)}$ be the submatrix containing rows of A indexed by the integers in ν_1 and columns of A indexed by the integers in ν_2. Define $V_{ki}^{(\nu,\nu)} = V_{\nu ki}$, $\widehat{\Sigma}_k^{(\nu,\nu)} = \widehat{\Sigma}_{\nu k}$, and $\widehat{\Sigma}_k^{(\nu_1,\nu_2)} = \widehat{\Sigma}_{\nu_1 \nu_2 k}$. The joint regression imputation method imputes $\mathbf{t}_{\nu i}$ by

$$\mathbf{t}_{\nu i}^* = \widehat{B}_{\nu k}' \mathbf{x}_i + V_{\nu ki}^{1/2} \left[\widehat{\Sigma}_{\nu \nu^c k} \widehat{\Sigma}_{\nu^c k}^{-1} V_{\nu^c ki}^{-1/2} (\mathbf{t}_{\nu^c i} - \widehat{B}_{\nu^c k}' \mathbf{x}_i) + \tilde{\mathbf{e}}_i^* \right], \tag{16}$$

where, given the observed data, the $\tilde{\mathbf{e}}_i^*$'s are independent random vectors with mean 0 and covariance matrix $\widehat{\Sigma}_{\nu k} - \widehat{\Sigma}_{\nu \nu^c k} \widehat{\Sigma}_{\nu^c k}^{-1} \widehat{\Sigma}_{\nu \nu^c k}'$ ($\widehat{\Sigma}_{\nu^c k}^{-1}$ is defined to be 0 if $\nu = \{1, \dots, q\}$). One choice of a positive definite consistent estimator of Σ_k is

$$\widehat{\Sigma}_k = \sum_{i \in \mathcal{R}_k} w_i V_{ki}^{-1/2} (\mathbf{t}_i - \widehat{B}_k' \mathbf{x}_i)(\mathbf{t}_i - \widehat{B}_k' \mathbf{x}_i)' V_{ki}^{-1/2} \Big/ \sum_{i \in \mathcal{R}_k} w_i, \tag{17}$$

where $\mathcal{R}_k$ is the set of i's in imputation class k for which $\mathbf{t}_i$ has no missing components.

The proposed joint regression imputation is an extension of formulas (6)–(7) of Srivastava and Carter (1986), who considered the situation of no auxiliary $\mathbf{x}$, normally distributed $(y_i, z_i, \dots)$'s, and uniform nonresponse (i.e., missing completely at random). Also, our method of generating random terms incorporates the survey weights w_i's so that our results are applicable to complex surveys. Srivastava and Carter (1986) proposed drawing random terms from residuals computed using respondents for all items, whereas our method is flexible for implementation, because random terms can be generated from any distribution with the covariance matrix given by $\widehat{\Sigma}_k$. If parameters other than functions of the first- and second-order moments are considered (e.g., quantiles), then random terms should be generated from residuals as follows. For a given imputation class k, let $\mathcal{R}_{k\nu}$ be the set of i's with missing components indexed by ν. For each $j \in \mathcal{R}_{k\nu}$, $\tilde{\mathbf{e}}_j^*$ is generated according to

$$P^*(\tilde{\mathbf{e}}_j^* = \mathbf{r}_{\nu ki} - \bar{\mathbf{r}}_{\nu k}) = w_i \Big/ \sum_{i \in \mathcal{R}_k} w_i, \qquad i \in \mathcal{R}_k,$$

where $\mathcal{R}_k$ is the set of i's in imputation class k for which $\mathbf{t}_i$ has no missing components,

$$\mathbf{r}_{\nu ki} = V_{\nu ki}^{-1/2} (\mathbf{t}_{\nu i} - \widehat{B}_{\nu k}' \mathbf{x}_i) - \widehat{\Sigma}_{\nu \nu^c k} \widehat{\Sigma}_{\nu^c k}^{-1} V_{\nu^c ki}^{-1/2} (\mathbf{t}_{\nu^c i} - \widehat{B}_{\nu^c k}' \mathbf{x}_i),$$

and $\bar{\mathbf{r}}_{\nu k} = \sum_{i \in \mathcal{R}_k} w_i \mathbf{r}_{\nu ki} / \sum_{i \in \mathcal{R}_k} w_i$.

It is shown in Appendix A.2 that this joint regression imputation method is unbiased for estimating marginal totals and correlation coefficients among items.

3. VARIANCE ESTIMATION

Once a nearly unbiased survey estimator is obtained, the next important step in sample surveys is to find a nearly unbiased and consistent variance estimator for assessing the variability of the survey estimator. It is well known that if we treat imputed values as observed data and apply standard variance estimation formulas for the case of no nonresponse, then we may seriously underestimate the true variances.

It is possible to obtain a nearly unbiased and consistent variance estimator for the sample correlation coefficient $\hat{\rho}^*$ using linearization (i.e., Taylor expansion) and substitution. However, under joint regression imputation, linearization for $\hat{\rho}^*$ involves very messy derivations, especially for the multivariate case. Instead, we use the following adjusted jackknife method, which is an extension of the method of Rao and Shao (1992) for estimating the variance of $\widehat{Y}^*$ under marginal random imputation. Some other resampling methods have been given by Srivastava (1997).

Assume that the first-stage sampling fraction $\sum_{h=1}^H n_h / \sum_{h=1}^H N_h$ is negligible, although n_h/N_h may not be negligible for some h's. In the case of no nonresponse, the jackknife method works as follows. Let $\mathbf{X} = \{\mathbf{t}_i, \mathbf{x}_i, w_i : i \in \mathcal{S}\}$ denote the dataset including covariates and survey weights, and let $\mathbf{X}_{hj} = \{\mathbf{t}_i, \mathbf{x}_i, w_i^{(hj)} : i \in \mathcal{S}\}$ be the (h, j)th pseudoreplicate after deleting the first-stage cluster j in stratum h and suitably adjusting the survey weights, where

$$w_i^{(hj)} = \begin{cases} 0 & \text{if } i \text{ is in cluster } j \\ \dfrac{n_h w_i}{n_h - 1} & \text{if } i \text{ is not in cluster } j \text{ but in stratum } h \\ w_i & \text{if } i \text{ is not in stratum } h. \end{cases}$$

Let $\hat{\theta} = \hat{\theta}(\mathbf{X})$ be a survey estimator. The jackknife variance estimator for $\hat{\theta}(\mathbf{X})$ is

$$v_{\text{Jack}} = \sum_{h=1}^H \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} \left[\hat{\theta}(\mathbf{X}_{hj}) - \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{\theta}(\mathbf{X}_{hi}) \right]^2. \tag{18}$$

Note that the $\hat{\theta}(\mathbf{X}_{hj})$'s can be computed repeatedly using a program similar to that for computing $\hat{\theta}(\mathbf{X})$.

When there are imputed nonrespondents, formula (18) has to be modified to take imputation into account. Let $\mathbf{X}^*$ and $\mathbf{X}_{hj}^*$ be the same as $\mathbf{X}$ and $\mathbf{X}_{hj}$ but with imputed nonrespondents, and let $\widehat{B}_k^{(hj)}$ and $\widehat{\Sigma}_k^{(hj)}$ be $\widehat{B}_k$ and $\widehat{\Sigma}_k$, with w_i's replaced by $w_i^{(hj)}$'s; that is, $\widehat{B}_k^{(hj)}$ and $\widehat{\Sigma}_k^{(hj)}$ are $\widehat{B}_k$ and $\widehat{\Sigma}_k$ recalculated using observed data in $\mathbf{X}_{hj}$. We consider the following adjustment. For each imputed vector $\mathbf{t}_{\nu i}^*$ in $\mathbf{X}_{hj}^*$ [see (16)], we reimpute $\mathbf{t}_{\nu i}^*$ using the same imputation method but the observed data in $\mathbf{X}_{hj}$. More precisely, for each (h, j), the reimputed vector is the same as $\mathbf{t}_{\nu i}^*$ in (16) but with $\widehat{B}_k$ and $\widehat{\Sigma}_k$ replaced by $\widehat{B}_k^{(hj)}$ and $\widehat{\Sigma}_k^{(hj)}$ and with $\tilde{\mathbf{e}}_i^*$ replaced by $(\widehat{\Sigma}_{ek}^{(hj)} \widehat{\Sigma}_{ek}^{-1})^{1/2} \tilde{\mathbf{e}}_i^*$, where

$\widehat{\Sigma}_{ek} = \widehat{\Sigma}_{\nu k} - \widehat{\Sigma}_{\nu\nu^c k}\widehat{\Sigma}^{-1}_{\nu^c k}\widehat{\Sigma}'_{\nu\nu^c k}$ and $\widehat{\Sigma}^{(hj)}_{ek}$ is $\widehat{\Sigma}_{ek}$ recalculated using $w_i^{(hj)}$'s instead of w_i's. In the bivariate case, for example, if $a_i = 0$ and $b_i = 1$, then y_i^* is reimputed by

$$(\hat{\beta}_k^{(hj)})'\mathbf{x}_i + \frac{v_{ki}^{1/2}\hat{\sigma}^{(hj)}_{\epsilon,\eta,k}}{u_{ki}^{1/2}(\hat{\sigma}^{(hj)}_{\eta,k})^2}(z_i - (\hat{\gamma}_k^{(hj)})'\mathbf{x}_i) + \left(\frac{v_{ki}[(\hat{\sigma}^{(hj)}_{\epsilon,k})^2 - (\hat{\sigma}^{(hj)}_{\epsilon,\eta,k})^2/(\hat{\sigma}^{(hj)}_{\eta,k})^2]}{\hat{\sigma}^2_{\epsilon,k} - \hat{\sigma}^2_{\epsilon,\eta,k}/\hat{\sigma}^2_{\eta,k}}\right)^{1/2}\tilde{\epsilon}_i^*,$$

where $\hat{\beta}_k^{(hj)}, \hat{\gamma}_k^{(hj)}, \hat{\sigma}^{(hj)}_{\epsilon,k}, \hat{\sigma}^{(hj)}_{\eta,k}$, and $\hat{\sigma}^{(hj)}_{\epsilon,\eta,k}$ are the same as $\hat{\beta}_k, \hat{\gamma}_k, \hat{\sigma}_{\epsilon,k}, \hat{\sigma}_{\eta,k}$, and $\hat{\sigma}_{\epsilon,\eta,k}$, except that the w_i's are replaced by $w_i^{(hj)}$'s in their definitions. Note that not only are $\widehat{B}_k$ and $\widehat{\Sigma}_k$ changed to $\widehat{B}_k^{(hj)}$ and $\widehat{\Sigma}_k^{(hj)}$ in the reimputation, but also the random term $\tilde{\mathbf{e}}_i^*$ is changed to $(\widehat{\Sigma}^{(hj)}_{ek}\widehat{\Sigma}^{-1}_{ek})^{1/2}\tilde{\mathbf{e}}_i^*$, which has imputation variance $\widehat{\Sigma}^{(hj)}_{ek}$, although we do not generate any new random vectors in reimputation.

Let $\mathbf{X}_{hj}^{*\mathrm{RI}}$ be the reimputed $\mathbf{X}_{hj}^*$. The adjusted jackknife variance estimator $v_{\mathrm{jack}}^{\mathrm{adj}}$ for $\hat{\theta}(\mathbf{X}^*)$ is then obtained using (18) with $\mathbf{X}_{hj}$ replaced by $\mathbf{X}_{hj}^{*\mathrm{RI}}$.

To show that $v_{\mathrm{jack}}^{\mathrm{adj}}$ is asymptotically unbiased and consistent for the asymptotic variance of $\widehat{Y}^*$ or $\hat{\rho}^*$, we consider for simplicity the special case of bivariate $\mathbf{t}_i = (y_i, z_i)'$. Note that both $\widehat{Y}^*$ or $\hat{\rho}^*$ are differentiable functions of weighted averages of the form $\sum_{i\in\mathcal{S}_k} w_i\psi_i$. For example, $\widehat{Y}^* = \widehat{Y}(\mathbf{X}^*)$ is a differentiable function of $\hat{\beta}_k, \hat{\sigma}_{\epsilon,\eta,k}, \hat{\sigma}^2_{\epsilon,k}, \hat{\sigma}^2_{\eta,k}$ (each of which is a function of weighted averages),

$$\sum_{i\in\mathcal{S}_k} w_i a_i y_i, \quad \sum_{i\in\mathcal{S}_k} w_i(1-a_i)b_i\mathbf{x}_i, \quad \sum_{i\in\mathcal{S}_k} w_i(1-a_i)b_i v_{ki}^{1/2}u_{ki}^{-1/2}z_i,$$

$$\sum_{i\in\mathcal{S}_k} w_i(1-a_i)b_i v_{ki}^{1/2}u_{ki}^{-1/2}\mathbf{x}_i, \quad \sum_{i\in\mathcal{S}_k} w_i(1-a_i)(1-b_i)\mathbf{x}_i,$$

$$\sum_{i\in\mathcal{S}_k} w_i(1-a_i)b_i v_{ki}^{1/2}\delta_i, \quad \text{and} \quad \sum_{i\in\mathcal{S}_k} w_i(1-a_i)(1-b_i)v_{ki}^{1/2}\tau_i,$$

where $\delta_i = \tilde{\epsilon}_i^*/(\hat{\sigma}^2_{\epsilon,k} - \hat{\sigma}^2_{\epsilon,\eta,k}/\hat{\sigma}^2_{\eta,k})^{1/2}$ and $\tau_i = \epsilon_i^*/\hat{\sigma}_{\epsilon,k}$ are random variables with mean 0 and variance 1. (Because $\widehat{Y}^*$ is a nonlinear function of so many terms and $\hat{\rho}^*$ is a nonlinear function of even more terms, variance estimation by linearization/Taylor expansion involves very messy derivations.) Note that $\hat{\theta}(\mathbf{X}_{hj}^{*\mathrm{RI}})$ is the same function of the same weighted averages, except that the w_i's are replaced by $w_i^{(hj)}$'s. From the theory of jackknife (see, e.g., Krewski and Rao 1981), the jackknife variance estimator $v_{\mathrm{jack}}^{\mathrm{adj}}$ is asymptotically unbiased and consistent (under the asymptotic setting in Appendix A.1). Note that the same conclusion cannot be reached if we use $\hat{\theta}(\mathbf{X}_{hj}^*)$ in (18).

Although the same adjusted jackknife method can be applied to nonrandom regression imputation and marginal random regression imputation to obtain consistent variance estimators for $\widehat{Y}^*$, jackknife variance estimators for $\hat{\rho}^*$ under marginal regression imputation are meaningless when $\hat{\rho}^*$ is not asymptotically unbiased.

4. SIMULATION RESULTS

We conducted a simulation study to study the finite-sample performance of (a) the sample correlation coefficient $\hat{\rho}^*$ under joint regression imputation, (b) the proposed jackknife variance estimator $v_{\mathrm{jack}}^{\mathrm{adj}}$ for $\hat{\rho}^*$, and (c) the confidence interval (for the true correlation coefficient) using $\hat{\rho}^*$ and $v_{\mathrm{jack}}^{\mathrm{adj}}$.

4.1 Bivariate Case With Normal Errors

We consider a one-stage stratified simple random sampling design used in the Transportation Annual Survey conducted by the U.S. Census Bureau (U.S. Census Bureau 1987). There are 33 strata divided into 4 imputation classes according to business type. Bivariate data (y_i, z_i)'s are generated according to (1) and (5) with a univariate x and $v_{ki}(x) = u_{ki}(x) = x$. Stratum sample sizes, survey weights, and model parameter values are listed in Table 1. The x values are independently generated from gamma distributions with means and variances given by those in Table 1. The error terms ϵ_i and η_i are independently generated according to

$$\epsilon_i = \kappa\zeta_i + \delta_i \quad \text{and} \quad \eta_i = \kappa\zeta_i + \tau_i, \tag{19}$$

where ζ_i, δ_i, and τ_i are independently distributed as $N(0, 1)$ and $\kappa \geq 0$ is a parameter. As κ increases, the value of the correlation coefficient increases. When $\kappa = 0$, y_i and z_i are conditionally independent, given x_i.

Table 1. Sample Size, Survey Weight, and Model Parameters Across Imputation Classes and Strata

Imputation class	Stratum	Sample size	Survey weight	Model parameters			
				$E(x)$	$\sqrt{var(x)}$	β	γ
1	1	14	12.43	2.0	.1	1.0	.6
	2	11	8.91	3.0	.1	1.0	.6
	3	10	6.10	4.0	.1	1.0	.6
	4	11	5.73	5.0	.1	1.0	.6
	5	16	2.70	6.0	.1	1.0	.6
	6	18	2.17	7.0	.1	1.0	.6
	7	31	1.00	8.0	.1	1.0	.6
2	1	8	32.91	2.5	.2	.5	.7
	2	13	9.85	3.5	.2	.5	.7
	3	11	10.82	4.5	.2	.5	.7
	4	12	6.08	5.5	.2	.5	.7
	5	13	3.60	6.5	.2	.5	.7
	6	86	1.00	7.5	.2	.5	.7
3	1	14	87.91	3.0	.1	1.1	1.1
	2	11	67.39	4.0	.1	1.1	1.1
	3	13	44.48	5.0	.1	1.1	1.1
	4	14	25.28	6.0	.1	1.1	1.1
	5	16	15.57	7.0	.1	1.1	1.1
	6	18	9.80	8.0	.1	1.1	1.1
	7	15	6.23	9.0	.1	1.1	1.1
	8	15	4.68	10.0	.1	1.1	1.1
	9	40	2.13	11.0	.1	1.1	1.1
	10	38	1.00	12.0	.1	1.1	1.1
4	1	7	32.14	3.5	.1	1.0	1.0
	2	13	16.75	4.5	.1	1.0	1.0
	3	10	12.90	5.5	.1	1.0	1.0
	4	14	7.00	6.5	.1	1.0	1.0
	5	13	6.18	7.5	.1	1.0	1.0
	6	11	4.70	8.5	.1	1.0	1.0
	7	17	3.31	9.5	.1	1.0	1.0
	8	19	1.89	10.5	.1	1.0	1.0
	9	22	1.82	11.5	.1	1.0	1.0
	10	28	1.00	12.5	.1	1.0	1.0

Given the generated data, respondents are generated according to (2) and (6), with independent a_i's and b_i's (although our results in Sections 2 and 3 hold even when a_i and b_i are related),

$$P(a_i = 1|x_i) = \frac{e^{.1+.05x_i}}{1+e^{.1+.05x_i}}$$

and

$$P(b_i = 1|x_i) = \frac{e^{.2+.04x_i}}{1+e^{.2+.04x_i}}.$$

The average response rates, $E[P(a_i = 1|x_i)]$ and $E[P(b_i = 1|x_i)]$, are approximately 61.75%.

The random terms in imputation are generated from a normal distribution with covariance matrix given by (12).

We consider $\hat{\rho}^*$ based on the joint regression imputation and $\hat{\rho}^*$ based on the marginal nonrandom regression imputation. The sample correlation coefficient $\hat{\rho}$ based on data without nonresponse is also included. Table 2 lists the mean and standard deviation of three sample correlation coefficients for some values of κ and ρ, computed based on 500 simulations. The results in Table 2 can be summarized as follows:

1. The mean of the sample correlation coefficient $\hat{\rho}^*$ under joint regression imputation is very close to the true ρ. This supports our theory of the asymptotic unbiasedness of the joint regression imputation method. The standard deviation of $\hat{\rho}^*$ under joint regression imputation is higher than that of $\hat{\rho}$, indicating the price paid for nonresponse and imputation.

2. The sample correlation coefficient $\hat{\rho}^*$ under marginal nonrandom regression imputation is biased, except for the case where $\kappa = 1$ (biases canceled out). Comparing standard deviations of $\hat{\rho}^*$ for marginal nonrandom imputation and joint regression imputation, we find that marginal nonrandom imputation has lower standard deviation when $\kappa \leq 1$, whereas joint regression imputation has lower standard deviation when $\kappa \geq 1.2$. The variability in marginal nonrandom imputation increases with the value of κ.

The simulation average of $\sqrt{v_{\text{jack}}^{\text{adj}}}$ as an estimator of the standard deviation of $\hat{\rho}^*$ under joint regression imputation is given in Table 3 (the case of normal errors), which shows that the proposed jackknife estimator has a negligible bias in all cases. The average of the naive jackknife standard deviation estimator obtained by treating imputed values as observed data [i.e., using $\hat{\theta}(\mathbf{X}_{hj}^*)$ instead of $\hat{\theta}(\mathbf{X}_{hj}^{*\text{RI}})$ in (18)] was .0317 in the case of $\kappa = 1$, which resulted in a very large negative bias as expected.

A large sample $100(1-\alpha)\%$ confidence interval for the true correlation coefficient ρ has the limits $\hat{\rho}^* \pm z_\alpha \sqrt{v_{\text{jack}}^{\text{adj}}}$, where z_α is the $1-\alpha$ quantile of the standard normal distribution. It is well known that confidence intervals based on the Fisher Z scale $h(\rho) = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$ perform better in this problem. A large sample $100(1-\alpha)\%$ confidence interval for $h(\rho)$ has the limits $h(\hat{\rho}^*) \pm z_\alpha \sqrt{v_{\text{jack}}^{\text{adj}}}$, where $v_{\text{jack}}^{\text{adj}}$ is the proposed jackknife variance estimator for $h(\hat{\rho}^*)$. The simulation average of the coverage probabilities and the lengths of 95% confidence intervals (under the raw scale and Fisher Z scale) are reported in Table 3 (the case of normal errors). In terms of the coverage probability, it is clear that the intervals under the Fisher Z scale perform better.

Rubin and Schenker (1986) proposed a multiple imputation method that imputes nonrespondents using an approximate Bayesian bootstrap (ABB) imputation. In our simulation, ABB imputation is carried out by generating bootstrap samples from the centered residuals. Multiple imputation imputes nonrespondents independently $m \geq 2$ times to obtain imputed datasets $\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_m$. For a survey estimator $\hat{\theta}$, multiple imputation uses

$$\tilde{\theta}_m = \frac{1}{m}\sum_{l=1}^{m} \hat{\theta}(\tilde{\mathbf{X}}_l)$$

as the point estimator and

$$v_m = \widehat{W} + \frac{m+1}{m}\widehat{B}$$

Table 2. Averages of 500 Simulations of Sample Correlation Coefficients and Standard Deviations (SD) of Sample Correlation Coefficients (bivariate case)

		Without nonresponse		Marginal nonrandom regression imputation		Joint regression imputation	
κ	ρ	$\hat{\rho}$	$SD(\hat{\rho})$	$\hat{\rho}^*$	$SD(\hat{\rho}^*)$	$\hat{\rho}^*$	$SD(\hat{\rho}^*)$
0	.5491	.5521	.0351	.6618	.0494	.5462	.0693
.2	.5581	.5563	.0362	.6620	.0498	.5596	.0685
.4	.5777	.5779	.0378	.6656	.0478	.5795	.0691
.6	.6091	.6090	.0351	.6697	.0483	.6076	.0671
.8	.6462	.6467	.0332	.6796	.0485	.6456	.0626
1.0	.6850	.6834	.0314	.6854	.0524	.6872	.0590
1.2	.7220	.7223	.0283	.6917	.0569	.7233	.0549
1.4	.7559	.7576	.0251	.6990	.0652	.7561	.0487
1.6	.7863	.7858	.0230	.7065	.0725	.7819	.0443
1.8	.8117	.8130	.0210	.7081	.0797	.8140	.0387
2.0	.8352	.8345	.0192	.7155	.0848	.8350	.0350
2.4	.8711	.8709	.0145	.7238	.0998	.8714	.0280
2.8	.8975	.8976	.0120	.7286	.1027	.8971	.0226
3.2	.9172	.9166	.0104	.7322	.1146	.9165	.0184
3.6	.9320	.9311	.0087	.7442	.1225	.9325	.0153
4.0	.9433	.9432	.0070	.7520	.1247	.9426	.0127

as the variance estimator for $\tilde{\theta}_m$, where

$$\widehat{W} = \frac{1}{m}\sum_{l=1}^{m} v(\widetilde{\mathbf{X}}_l),$$

$v(\mathbf{X})$ is a variance estimator for $\hat{\theta}(\mathbf{X})$ in the case of no nonresponse [but $v(\widetilde{\mathbf{X}}_l)$ underestimates the variance of $\hat{\theta}(\widetilde{\mathbf{X}}_l)$] and

$$\widehat{B} = \frac{1}{m-1}\sum_{l=1}^{m}(\hat{\theta}(\widetilde{\mathbf{X}}_l) - \tilde{\theta}_m)^2.$$

A $100(1-\alpha)\%$ confidence interval based on multiple imputation has the limits $\tilde{\theta}_m \pm t_\alpha(\eta)\sqrt{v_m}$, where $t_\alpha(\eta)$ is the $1-\alpha$ quantile of the t distribution with η degrees of freedom and $\eta = (m-1)[1+m\widehat{W}/(m+1)\widehat{B}]^2$.

Standard deviation of $\tilde{\rho}_m$, $\sqrt{v_m}$, coverage probability and length of confidence interval based on ABB multiple imputation with $m = 3$ are evaluated by simulation, and the results are given in Table 3 (the case of normal errors). The results indicate that v_m underestimates, but the bias becomes smaller as κ (or ρ) increases. Raw scale confidence intervals based on multiple imputation generally have higher coverage probabilities than those based on $\hat{\rho}^*$ and the jackknife, but they also have longer average lengths. Under multiple imputation, use of the Fisher Z scale seems unnecessary.

4.2 Bivariate Case With Nonnormal Errors

To examine the effect of normality of the errors in (19), we repeat the simulation given in Table 3 with nonnormal errors. ζ_i is distributed as an exponential random variable with probability density $e^{-(x+1)}$, $x \geq -1$, δ_i and τ_i are distributed as a mixture of normal $.4N(0, .9) + .6N(0, 3.2/3)$, and ζ_i, δ_i, and τ_i are independent. Consequently, the distributions of the errors ϵ_i and η_i are skewed. The variables ζ_i, δ_i, and τ_i still have mean 0 and variance 1, so that the correlation coefficient between y_i and z_i is the same as that in Section 4.1. The other settings of the simulation are also the same as those in Section 4.1. In particular, the random terms in imputation are still generated from the normal distribution with covariance matrix given by (12).

The simulation results are included in Table 3 (the case of nonnormal errors). It can be seen that the performance of the proposed method is similar to that in the case of normal errors. The standard deviations of $\hat{\rho}^*$ and $\tilde{\rho}_m$ are increased, so are

Table 3. Averages of 500 Simulations of the Standard Deviation (SD) of Sample Correlation Coefficients, SD Estimators, and the Coverage Probability (CP) and Length of 95% Confidence Intervals (bivariate case)

	Proposed method						Multiple imputation ($m = 3$)					
			Raw scale		Fisher Z scale				Raw scale		Fisher Z scale	
κ	$SD(\hat{\rho}^*)$	$\sqrt{v_{jack}^{adj}}$	CP	Length	CP	Length	$SD(\tilde{\rho}_m)$	$\sqrt{v_m}$	CP	Length	CP	Length
Case of normal errors												
0	.0693	.0693	.944	.272	.946	.393	.0582	.0503	.938	.290	.942	.419
.2	.0685	.0672	.924	.263	.934	.389	.0573	.0499	.940	.288	.946	.420
.4	.0692	.0678	.914	.266	.928	.406	.0568	.0494	.937	.287	.941	.434
.6	.0671	.0660	.918	.259	.932	.416	.0551	.0480	.937	.281	.943	.451
.8	.0626	.0624	.908	.245	.926	.426	.0527	.0459	.940	.269	.944	.465
1.0	.0590	.0579	.926	.227	.946	.436	.0479	.0430	.943	.253	.947	.479
1.2	.0534	.0534	.918	.209	.930	.445	.0433	.0396	.950	.234	.956	.491
1.4	.0487	.0477	.906	.187	.920	.442	.0397	.0359	.948	.211	.952	.496
1.6	.0443	.0440	.914	.172	.926	.448	.0345	.0327	.953	.193	.960	.507
1.8	.0387	.0387	.922	.152	.934	.453	.0306	.0294	.953	.172	.960	.509
2.0	.0350	.0345	.912	.135	.930	.451	.0269	.0263	.960	.154	.965	.511
2.4	.0280	.0268	.920	.105	.942	.441	.0215	.0212	.960	.124	.968	.515
2.8	.0226	.0220	.926	.086	.952	.448	.0171	.0173	.962	.101	.970	.518
3.2	.0184	.0181	.912	.071	.938	.448	.0137	.0142	.962	.083	.970	.526
3.6	.0153	.0145	.916	.057	.940	.440	.0112	.0118	.968	.069	.974	.526
4.0	.0127	.0125	.924	.049	.940	.443	.0097	.0100	.961	.059	.968	.531
Case of nonnormal errors												
0	.0716	.0700	.940	.274	.948	.397	.0597	.0511	.935	.292	.938	.422
.2	.0680	.0702	.934	.275	.944	.405	.0596	.0512	.941	.296	.946	.432
.4	.0687	.0693	.930	.272	.934	.417	.0598	.0506	.936	.294	.942	.446
.6	.0697	.0709	.928	.277	.940	.449	.0593	.0495	.931	.288	.936	.461
.8	.0687	.0666	.926	.261	.934	.459	.0585	.0479	.932	.279	.936	.482
1.0	.0685	.0662	.918	.260	.926	.491	.0557	.0459	.932	.267	.935	.502
1.2	.0649	.0609	.922	.239	.928	.492	.0517	.0431	.937	.249	.940	.517
1.4	.0617	.0576	.918	.226	.924	.517	.0477	.0403	.937	.231	.939	.534
1.6	.0539	.0519	.942	.204	.930	.524	.0435	.0374	.944	.213	.947	.551
1.8	.0478	.0461	.930	.181	.940	.528	.0392	.0345	.947	.195	.949	.566
2.0	.0445	.0416	.920	.163	.940	.531	.0350	.0316	.952	.177	.951	.575
2.4	.0352	.0347	.934	.136	.944	.543	.0287	.0263	.955	.144	.956	.588
2.8	.0288	.0282	.932	.111	.938	.541	.0234	.0221	.959	.120	.959	.605
3.2	.0247	.0234	.920	.088	.934	.545	.0192	.0185	.959	.099	.959	.613
3.6	.0206	.0194	.946	.076	.938	.559	.0161	.0156	.964	.083	.963	.620
4.0	.0161	.0162	.950	.064	.952	.556	.0135	.0134	.964	.072	.964	.637

NOTE: $\hat{\rho}^*$ denotes the sample correlation coefficient based on joint regression imputation; v_{jack}^{adj}, the proposed jackknife variance estimator; ρ_m, the sample correlation coefficient based on multiple imputation; and v_m, the variance estimator based on multiple imputation.

Table 4. Averages of 500 Simulations of Sample Correlation Coefficients, Standard Deviation (SD) of Sample Correlation Coefficients, SD Estimators, and the Coverage Probability (CP) of 95% Confidence Intervals (multivariate case)

p, q	ρ	$\hat{\rho}^*$	$SD(\hat{\rho}^*)$	$\sqrt{v_{jack}^{adj}}$	CP
1, 2	.5121	.5039	.0701	.0740	.950
1, 3	.5085	.5022	.0679	.0724	.958
1, 4	.4981	.4964	.0649	.0701	.948
1, 5	.4844	.4818	.0646	.0689	.942
1, 6	.4669	.4686	.0645	.0666	.966
1, 7	.4505	.4548	.0630	.0656	.952
1, 8	.4323	.4392	.0628	.0647	.936
1, 9	.4152	.4216	.0605	.0637	.944
1, 10	.3992	.4056	.0588	.0626	.946
2, 3	.5482	.5413	.0665	.0678	.944
2, 4	.5556	.5419	.0635	.0671	.950
2, 5	.5483	.5438	.0616	.0640	.944
2, 6	.5405	.5344	.0565	.0630	.972
2, 7	.5321	.5329	.0576	.0601	.942
2, 8	.5201	.5234	.0541	.0590	.954
2, 9	.5107	.5138	.0547	.0571	.946
2, 10	.4996	.4986	.0540	.0572	.960
3, 4	.5906	.5797	.0587	.0636	.940
3, 5	.6002	.5878	.0567	.0598	.956
3, 6	.6029	.5912	.0545	.0575	.970
3, 7	.6001	.5941	.0516	.0555	.956
3, 8	.5964	.5908	.0500	.0542	.960
3, 9	.5906	.5839	.0479	.0527	.962
3, 10	.5833	.5782	.0488	.0515	.942
4, 5	.6384	.6230	.0557	.0574	.950
4, 6	.6490	.6381	.0501	.0537	.954
4, 7	.6549	.6478	.0483	.0508	.954
4, 8	.6550	.6492	.0465	.0499	.968
4, 9	.6558	.6466	.0467	.0478	.954
4, 10	.6522	.6455	.0432	.0465	.958
5, 6	.6834	.6704	.0473	.0502	.976
5, 7	.6950	.6860	.0439	.0471	.966
5, 8	.7012	.6901	.0442	.0456	.948
5, 9	.7051	.6964	.0423	.0431	.944
5, 10	.7062	.6967	.0415	.0418	.970
6, 7	.7247	.7142	.0414	.0432	.968
6, 8	.7361	.7264	.0386	.0411	.968
6, 9	.7424	.7324	.0365	.0390	.962
6, 10	.7472	.7382	.0347	.0374	.976
7, 8	.7607	.7526	.0356	.0376	.948
7, 9	.7716	.7653	.0315	.0350	.962
7, 10	.7779	.7709	.0322	.0334	.964
8, 9	.7921	.7840	.0306	.0322	.952
8, 10	.8019	.7946	.0299	.0302	.956
9, 10	.8188	.8152	.0246	.0276	.968

NOTE: p, q represent the case of estimating the correlation coefficient ρ between variables p and q, $1 \le p \le q \le 10$; $\hat{\rho}^*$, sample correlation coefficient based on joint regression imputation; and v_{jack}^{adj}, the proposed jackknife variance estimator.

the lengths of confidence intervals, which indicates the efficiency loss of the proposed method (and the multiple imputation method) due to skewness of the errors. On the other hand, the coverage probabilities of the confidence intervals are almost the same as those in the case of normal error, which confirms the asymptotic unbiasedness and consistency of the proposed estimators.

4.3 Multivariate Case With Normal Errors

Finally, we consider a simulation for the multivariate case where $q = 10$–dimensional $\mathbf{t}_i$'s are generated according to (14) with $B_k' = (.1, .2, \ldots, 1.0)$ and $V_{ki} = x_i I_q$, where I_q is the identity matrix of order q. The sampling design and the distributions of x_i's are still given in Table 1. The distribution of $\mathbf{e}_i$ is multivariate normal with mean 0 and covariance matrix $I_q + 1_q 1_q'$, where 1_q is the column vector of 1s. The nonrespondents are generated with constant probability. The probability that $\mathbf{t}_i$ has no missing component is .5, and given that $\mathbf{t}_i$ has at least one missing component, the components of $\mathbf{t}_i$ are missing independently with probability .5. The random terms in imputation are generated from a normal distribution with covariance matrix given by (17).

For each pair of components of $\mathbf{t}$, the true value of the correlation coefficient ρ, simulation averages of $\hat{\rho}^*$ (the sample correlation coefficient based on joint regression imputation), the standard deviation of $\hat{\rho}^*$, the jackknife variance estimator $v_{\text{jack}}^{\text{adj}}$ for $\hat{\rho}^*$, and the coverage probability of 95% confidence intervals based on $\hat{\rho}^*$ and $v_{\text{jack}}^{\text{adj}}$ are given in Table 4. The results indicate that the performances of $\hat{\rho}^*$ and $v_{\text{jack}}^{\text{adj}}$ are good, although their biases are larger than those in the bivariate case (the same sample sizes are used in both cases). However, because the jackknife estimator overestimates in this case, the coverage probabilities of the confidence intervals are larger than 95% in many cases.

APPENDIX

A.1

The asymptotic results in this article are based on the following asymptotic framework (see, e.g., Krewski and Rao 1981; Bickel and Freedman 1984; Valliant 1993). The population $\mathcal{P}$ is assumed to be a member of a sequence of populations indexed by l, but l is suppressed for simplicity of notation. As $l \to \infty$, $n \to \infty$, $n/N \to 0$, and $\max_{h,j} \sum_{i \in \mathcal{S}(h,j)} w_i/M = O(n^{-1})$, where $n = \sum_h n_h$, $N = \sum_h N_h$, M is the total number of ultimate units in $\mathcal{P}$, and $\mathcal{S}(h, j)$ is the set of indices of sampled units in stratum h and cluster j. Also, as $l \to \infty$,

$$\sum_{h=1}^{H} \sum_{j=1}^{n_h} E\|\psi_{hj} - E(\psi_{hj})\|^{2+\delta} = O(n^{-(1+\delta)})$$

for some $\delta > 0$, where $\|\|$ is the usual vector norm and $\psi_{hj} = \sum_{i \in \mathcal{S}(h,j)} w_i \psi_i / M$, with ψ_i being any component of the matrix $\mathbf{t}_i \mathbf{t}_i'$ or $\mathbf{x}_i \mathbf{x}_i'$. Finally, as $l \to \infty$, $0 < \liminf[n\text{var}(\widehat{Y}/M)]$ and $0 < \liminf[n\text{var}(\hat{\rho})]$.

Models (14) and (15) are assumed. Also, for every $\mathbf{x}$ in the support of the distribution of $\mathbf{x}_i$, it is assumed that $P(\mathbf{a} = a|\mathbf{x}) > 0$.

A.2

Here we show that the estimated total $\widehat{Y}^*$ in (4) and the sample correlation coefficient $\hat{\rho}^*$ in (7) based on the joint regression imputation described in Section 2 are asymptotically unbiased, under the asymptotic setting in Section A.1. We consider only the bivariate case. First,

$$\begin{aligned} E_m E_s E_*(\widehat{Y}^*) &= E_m E_s \Bigg(\sum_{i \in \mathcal{S}} w_i a_i y_i + \sum_k \sum_{i \in \mathcal{S}_k} w_i (1 - a_i) \\ &\qquad \times \left[\hat{\beta}_k' \mathbf{x}_i + \frac{\hat{\sigma}_{\epsilon,\eta,k}}{u_{ki}^{1/2} \hat{\sigma}_{\eta,k}^2} (z_i - \hat{\gamma}_k' \mathbf{x}_i) \right] \Bigg) \\ &\approx E_m \left(\sum_{i \in \mathcal{P}} a_i y_i + \sum_k \sum_{i \in \mathcal{P}_k} (1 - a_i) \left(\beta_k' \mathbf{x}_i + \frac{\sigma_{\epsilon,\eta,k}}{\sigma_{\eta,k}^2} \eta_i \right) \right) \\ &= E_m \left(\sum_{i \in \mathcal{P}} y_i \right). \end{aligned}$$

Next,

$$\begin{aligned}
&E_m E_s E_* \left(\sum_{i\in\mathcal{S}} w_i (1-a_i) b_i y_i^* z_i \right) \\
&\quad = E_m E_s \left(\sum_k \sum_{i\in\mathcal{S}_k} w_i (1-a_i) b_i \hat{\beta}_k' \mathbf{x}_i z_i \right. \\
&\qquad \left. + \sum_k \sum_{i\in\mathcal{S}_k} w_i (1-a_i) b_i \frac{v_{ki}^{1/2} \hat{\sigma}_{\epsilon,\eta,k}}{u_{ki}^{1/2} \hat{\sigma}_{\eta,k}^2} (z_i - \hat{\gamma}_k' \mathbf{x}_i) z_i \right) \\
&\quad \approx E_m \left(\sum_k \sum_{i\in\mathcal{P}_k} (1-a_i) b_i \left(\beta_k' \mathbf{x}_i z_i + \frac{v_{ki}^{1/2} \sigma_{\epsilon,\eta,k}}{\sigma_{\eta,k}^2} \eta_i z_i \right) \right) \\
&\quad = E_m \left(\sum_k \sum_{i\in\mathcal{P}_k} (1-a_i) b_i (\beta_k' \mathbf{x}_i \gamma_k' \mathbf{x}_i + v_{ki}^{1/2} \sigma_{\epsilon,\eta,k}) \right) \\
&\quad = E_m \left(\sum_k \sum_{i\in\mathcal{P}_k} (1-a_i) b_i y_i z_i \right).
\end{aligned}$$

Similarly,

$$E_m E_s E_* \left(\sum_{i\in\mathcal{S}} w_i a_i (1-b_i) y_i z_i^* \right) \approx E_m \left(\sum_k \sum_{i\in\mathcal{P}_k} a_i (1-b_i) y_i z_i \right)$$

and

$$\begin{aligned}
&E_m E_x E_* \left(\sum_{i\in\mathcal{S}} w_i (1-a_i)(1-b_i) y_i^* z_i^* \right) \\
&\qquad \approx E_m \left(\sum_k \sum_{i\in\mathcal{P}_k} (1-a_i)(1-b_i) y_i z_i \right).
\end{aligned}$$

Thus

$$E_m E_s e_* \left(\sum_{i\in\mathcal{S}} w_i y_i^* z_i^* \right) \approx E_m \left(\sum_{i\in\mathcal{P}} y_i z_i \right).$$

Finally, note that

$$E_*(\epsilon_i^{*2}) = \frac{\hat{\sigma}_{\epsilon,\eta,k}^2}{u_{ki} \hat{\sigma}_{\eta,k}^4} (z_i - \hat{\gamma}_k' \mathbf{x}_i)^2 + \hat{\sigma}_{\epsilon,k}^2 - \frac{\hat{\sigma}_{\epsilon,\eta,k}^2}{\hat{\sigma}_{\eta,k}^2}.$$

Then

$$\begin{aligned}
&E_m E_s E_* \left(\sum_{i\in\mathcal{S}} w_i y_i^{*2} \right) \\
&\quad = E_m E_s \left(\sum_{i\in\mathcal{S}} w_i a_i y_i^2 + \sum_k \sum_{i\in\mathcal{S}_k} w_i (1-a_i) [(\hat{\beta}_k' \mathbf{x}_i)^2 + v_{ki} \hat{\sigma}_{\epsilon,k}^2] \right. \\
&\qquad \left. + \sum_k \sum_{i\in\mathcal{S}_k} w_i (1-a_i) \left[\frac{\hat{\sigma}_{\epsilon,\eta,k}^2}{u_{ki} \hat{\sigma}_{\eta,k}^4} (z_i - \hat{\gamma}_k' \mathbf{x}_i)^2 - \frac{\hat{\sigma}_{\epsilon,\eta,k}^2}{\hat{\sigma}_{\eta,k}^2} \right] \right) \\
&\quad \approx E_m \left(\sum_{i\in\mathcal{P}} a_i y_i^2 + \sum_k \sum_{i\in\mathcal{P}_k} (1-a_i) [(\beta_k' \mathbf{x}_i)^2 + v_{ki} \sigma_{\epsilon,k}^2] \right. \\
&\qquad \left. + \sum_k \sum_{i\in\mathcal{P}_k} (1-a_i) \left[\frac{\sigma_{\epsilon,\eta,k}^2}{\sigma_{\eta,k}^4} \eta_i^2 - \frac{\sigma_{\epsilon,\eta,k}^2}{\sigma_{\eta,k}^2} \right] \right) \\
&\quad = E_m \left(\sum_{i\in\mathcal{P}} a_i y_i^2 + \sum_k \sum_{i\in\mathcal{P}_k} (1-a_i) [(\beta_k' \mathbf{x}_i)^2 + v_{ki} \sigma_{\epsilon,k}^2] \right) \\
&\quad = E_m \left(\sum_{i\in\mathcal{P}} y_i^2 \right).
\end{aligned}$$

A.3

As we argued in Section 3, $\hat{\rho}^*$ is a function of weighted averages. Furthermore, this function is a differentiable function. Hence, under the asymptotic setting in Appendix A.1, asymptotic unbiasedness, consistency, and asymptotic normality of $\hat{\rho}^*$ and consistency of the adjusted jackknife variance estimator for $\hat{\rho}^*$ can be proved using standard arguments of Krewski and Rao (1981), Bickel and Freedman (1984), and Valliant (1993).

[Received November 1999. Revised August 2001.]

REFERENCES

Bickel, P. J., and Freedman, D. A. (1984), "Asymptotic Normality and the Bootstrap in Stratified Sampling," *The Annals of Statistics*, 12, 470–482.

Deville, J. C., and Särndal, C. E. (1994), "Variance Estimation for the Regression Imputed Horvitz–Thompson Estimator," *Journal of Official Statistics*, 10, 381–394.

Kalton, G., and Kasprzyk, D. (1986), "The Treatment of Missing Data," *Survey Methodology*, 12, 1–16.

Kim, J. J. (1986), "A Method for Limiting Disclosure in Micodata Based on Random Noise and Transformation," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 303–308.

Krewski, D., and Rao, J. N. K. (1981), "Inference From Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods," *The Annals of Statistics*, 9, 1010–1019.

Little, R. J., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data*, New York: Wiley.

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. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Rubin, D. B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 366–374.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman and Hall.

Srivastava, M. S. (1997), "Resampling Methods for Imputing Missing Observation in Regression Models, Where the Jackknife as Well as the Bootstrap Estimate of the Variance is Given for the Imputed Estimator in a Regression Model," Technical Report #9707, University of Toronto, Dept. of Statistics.

Srivastava, M. S., and Carter, E. M. (1986), "The Maximum Likelihood Method for Nonresponse in Sample Surveys," *Survey Methodology*, 12, 61–72.

U.S. Census Bureau (1987), "Noncertainty Sample Specification," BSR-87 Action Memo D.06, U.S. Census Bureau.

Valliant, R. (1993), "Poststratification and Conditional Variance Estimation," *Journal of the American Statistical Association*, 88, 89–96.