SAMPLE CORRELATION COEFFICIENTS BASED ON SURVEY DATA
UNDER REGRESSION IMPUTATION

Jun Shao and Hansheng Wang
Department of Statistics
University of Wisconsin
Madison, WI 53706

Abstract

Regression imputation is commonly used to compensate for item nonresponse when auxiliary data are available. 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. An imputation method preserves unbiasedness if survey estimators computed from imputed data are still unbiased or nearly unbiased for the parameters of interest. We study whether regression imputation preserves unbiasedness for not only population marginal totals, but also population correlation coefficients. We find that the commonly used marginal nonrandom regression imputation method does not preserve unbiasedness for marginal second moments (which are parts of a correlation coefficient); a marginal random regression imputation method rectifies the bias in estimating second moments, but it still does not preserve unbiasedness for correlation coefficients; a joint random regression imputation method preserves unbiasedness for marginal totals, second moments, and correlation coefficients. Some simulation results show that the joint random regression imputation method produces not only nearly unbiased sample correlation coefficients, 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 random regression imputation is also studied, using a jackknife method that takes imputation into account.

Keywords. Marginal imputation, joint imputation, random imputation, jackknife, 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 re-weighting. 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 that are designed for the case of no nonresponse (for example, 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 can seen later, the unbiasedness of an imputation method often depends on the type of parameters to be estimated.

For item nonresponse, vector imputation, which imputes the whole vector of items whenever a sampled unit has at least one missing item, is rarely used because it discards many observed data and may be very inefficient. *Marginal imputation*, which is also called *item imputation*, is often used in practice; i.e., 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, since the relationship is not preserved during marginal imputation.

The main purpose of this article is to study

1. 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

2. the variance estimation problem for an unbiased regression imputation method, which
is another 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 focus on marginal regression imputation, which has not been fully studied when parameters other than marginal totals are of concern. We first show that although the commonly used marginal nonrandom regression imputation method is unbiased for the estimation of marginal totals (first moments), it is not unbiased for marginal second moments or correlation coefficients. A marginal random regression imputation method, which adds random terms to regression imputed values, is shown to be unbiased for marginal first and second moments, but is still not unbiased for correlation coefficients except for some special cases. The biases of estimators under marginal (random or nonrandom) regression imputation are explicitly given. In Section 3, we propose a joint random regression imputation method, which is an extension of that in Srivastava and Carter (1986), and show that it is unbiased for estimating marginal totals as well as correlation coefficients. Some simulation results are also presented to examine the finite sample performance of joint random regression imputation. Simulation results show that the joint random regression imputation method produces not only nearly unbiased sample correlation coefficients, but also estimates that are more stable than those produced by marginal nonrandom regression imputation when correlation coefficients are in a certain range. Section 4 is devoted to the variance estimation problem for sample correlation coefficients under joint random regression imputation. We propose a jackknife method that takes imputation into account. The resulting jackknife variance estimator is shown to be asymptotically unbiased and consistent and performs well in a simulation study.

2. RESULTS FOR MARGINAL 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 $h$th stratum. In the first stage sampling, $n_h \geq 2$ clusters are selected 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 \( i \)th unit in \( \mathcal{P} \) is a vector \( (y_i, z_i, ...) \) of items of interest and a vector \( x_i \) 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. Nonrandom Regression Imputation

Like 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 cells) \( \mathcal{P}_k, k = 1, ..., K \), such that for an item \( y \) with nonresponse,

\[
y_i = \beta'_k x_i + v^{1/2} \epsilon_i \quad i \in \mathcal{P}_k
\]

and

\[
P(a_i = 1|y_i, x_i) = P(a_i = 1|x_i) \quad i \in \mathcal{P}_k,
\]

where \( \beta_k \) is a parameter vector and \( \beta'_k \) is its transpose, \( v_{ki} = v_k(x_i) \) with a known function \( v_k(\cdot) > 0 \), \( \epsilon_i \) is a random error (independent of \( x_i \)) 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 cells 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 \( x \), the response probability for item \( y \) is independent of \( y \).

The creation of \( K \) imputation cells allows us to establish a reasonable model such as (1) in the imputation cell \( \mathcal{P}_k \) which may not hold for the whole population \( \mathcal{P} \). Imputation cells are usually formed by using another auxiliary variable without nonresponse; for example, in many business surveys, imputation cells are strata or unions of strata.

Under assumption (1)-(2), the nonrandom regression imputation method imputes a non-respondent \( y_i \) in \( \mathcal{S}_k = \mathcal{S} \cap \mathcal{P}_k \) by

\[
y^*_i = \hat{\beta}'_k x_i,
\]
where

\[ \hat{\beta}_k = \left( \sum_{i \in S_k} w_i a_i v_{k_i}^{-1} x_i \hat{\beta}_i \right)^{-1} \sum_{i \in S_k} w_i a_i v_{k_i}^{-1} x_i y_i \]  

(3)

is the best linear unbiased estimator of \( \beta_k \) under (1) based on respondents in \( S_k \).

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 example, if the (marginal) total of item \( y_i \), \( Y = \sum_{i \in P} y_i \), is the parameter of interest, then its standard unbiased estimator in the case of no nonresponse is \( \hat{Y} = \sum_{i \in S} w_i y_i \) and, thus, its estimator based on imputed data is

\[ \hat{Y}^* = \sum_{i \in S} w_i a_i y_i + \sum_{i \in S} w_i (1 - a_i) y_i^* \]
\[ = \sum_{i \in S} w_i a_i y_i + \sum_{k \in S_k} \sum_{i \in S_{k_i}} w_i (1 - a_i) \hat{\beta}_k' x_i. \]  

(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

\[ E_m E_s(\hat{Y}^*) \approx E_m \left( \sum_{i \in P} a_i y_i + \sum_{k \in P_k} \sum_{i \in P_{k_i}} (1 - a_i) \hat{\beta}_k' x_i \right) \]
\[ = E_m \left( \sum_{i \in P} a_i y_i + \sum_{k \in P_k} \sum_{i \in P_{k_i}} (1 - a_i) \hat{\beta}_k' x_i \right) \]
\[ = E_m \left( \sum_{i \in P} y_i \right) \]
\[ = E_m(Y), \]

where \( \approx \) is the equality up to a negligible term under the asymptotic setting described in the Appendix,

\[ \hat{\beta}_k = \left( \sum_{i \in P_k} a_i v_{k_i}^{-1} x_i \hat{\beta}_i \right)^{-1} \sum_{i \in P_k} a_i v_{k_i}^{-1} x_i y_i \]  

(5)

the second equality follows from \( E_m(\hat{\beta}_k | x_i) = \beta_k \) under (1)-(2), and the third equality follows from (1). Hence, \( \hat{Y}^* \) is asymptotically unbiased for \( Y \).

In this article we focus on the estimation of the correlation coefficient between two items, say \( y \) and \( z \), defined by

\[ \rho = \frac{\sum_{i \in P} (y_i - \bar{Y}) (z_i - \bar{Z})}{\sqrt{\left[ \sum_{i \in P} (y_i - \bar{Y})^2 \sum_{i \in P} (z_i - \bar{Z})^2 \right]^{1/2}}} \]
where $\hat{Y} = Y / M$, $\hat{Z} = Z / M$, and $Z = \sum_{i \in \mathcal{P}} z_i$. When there is no nonresponse, a standard estimator of $\rho$ is the sample correlation coefficient

$$
\hat{\rho} = \frac{\sum_{i \in S} w_i (y_i - \hat{Y}) (z_i - \hat{Z})}{\sqrt{\left( \sum_{i \in S} w_i (y_i - \hat{Y})^2 \sum_{i \in S} w_i (z_i - \hat{Z})^2 \right)}}
$$

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

Consider now the situation where both $y$ and $z$ have nonrespondents. A model for item $z$ analogous to (1)-(2) is

$$
z_i = \gamma_k x_i + u_{ki}^{1/2} \eta_i \quad i \in \mathcal{P}_k
$$

and

$$
P(b_i = 1 | z_i, x_i) = P(b_i = 1 | x_i) \quad i \in \mathcal{P}_k,
$$

where $\gamma_k$ is unknown, $u_{ki} = u_k(x_i)$ with a known function $u_k(\cdot) > 0$, and $\eta_i$ is a random error (independent of $x_i$) with mean 0 and unknown variance $\sigma_{\eta_k}^2 > 0$. Nonrandom regression imputed value of a missing $z_i$ in $\mathcal{S}_k$ is then

$$
z_i^* = \hat{\gamma}_k x_i,
$$

where $\hat{\gamma}_k$ is an analog of $\hat{\beta}_k$ for item $z$.

Note that $\epsilon_i$ in (1) and $\eta_i$ in (7) are generally dependent. Similarly, $a_i$ in (2) and $b_i$ in (8) are generally dependent.

As we discussed in Section 1, nonrespondents for $y$ and $z$ are often imputed marginally (separately). For simplicity, let $y_i^* = y_i$ if $y_i$ is observed and $y_i^* = \text{the imputed value if } y_i$ is a nonrespondent, and let $z_i^*$ be similarly defined. Then the sample correlation coefficient based on the imputed data is

$$
\hat{\rho}^* = \frac{\sum_{i \in \mathcal{S}} w_i (y_i^* - \hat{Y}) (z_i^* - \hat{Z})}{\sqrt{\left( \sum_{i \in \mathcal{S}} w_i (y_i^* - \hat{Y})^2 \sum_{i \in \mathcal{S}} w_i (z_i^* - \hat{Z})^2 \right)}}
$$

where $\hat{Y} = \sum_{i \in \mathcal{S}} w_i y_i^* / \sum_{i \in \mathcal{S}} w_i$, $\hat{Z} = \sum_{i \in \mathcal{S}} w_i z_i^* / \sum_{i \in \mathcal{S}} w_i$, and $\sum_{i \in \mathcal{S}} w_i = \sum_{i \in \mathcal{S}} w_i y_i^2 / \sum_{i \in \mathcal{S}} w_i$.
where \( \hat{Y}^* = \hat{Y}^*/\hat{M} \) and \( \hat{Z}^* \) is similarly defined.

For marginal imputation, the previous discussion can be extended to the case of more than two items in a straightforward way. To study properties of sample correlation coefficients in (9) we only need to focus on the bivariate case.

2.2. The Bias of \( \hat{\rho}^* \) Based on Marginal Nonrandom Regression Imputation

Is \( \hat{\rho}^* \) asymptotically unbiased for \( \rho \) in (6)? Since \( \hat{Y}^* \), \( \hat{Z}^* \), and \( \hat{M} \) are asymptotically unbiased for \( Y \), \( Z \), and \( M \), respectively, we only need to study the biases of \( \sum_{i \in S} w_i y_i^2 \), \( \sum_{i \in S} w_i z_i^2 \), and \( \sum_{i \in S} w_i y_i^2 z_i \), as estimators of \( \sum_{i \in \mathcal{P}} y_i^2 \), \( \sum_{i \in \mathcal{P}} z_i^2 \), and \( \sum_{i \in \mathcal{P}} y_i z_i \), respectively.

The bias of \( \sum_{i \in S} w_i y_i^2 \) is similar to the bias of \( \sum_{i \in S} w_i z_i^2 \). Thus, we only need to consider one of them. Under (1)-(2) and the asymptotic setting in the Appendix,

\[
E_m \mathbb{E}_s \left( \sum_{i \in S} w_i y_i^2 \right) = E_m \mathbb{E}_s \left( \sum_{i \in S} w_i a_i \beta_k \bar{x}_i^2 \right)
\]

\[
\approx E_m \left( \sum_{i \in \mathcal{P}} a_i \beta_k \bar{x}_i^2 \right) + \sum_{k} \sum_{i \in \mathcal{P}_k} (1 - a_i) \beta_k \bar{x}_i^2
\]

\[
\approx E_m \left( \sum_{k} \sum_{i \in \mathcal{P}_k} a_i [(\beta_k \bar{x}_i)^2 + v_{ki} \sigma_{e_k}^2] + \sum_{k} \sum_{i \in \mathcal{P}_k} (1 - a_i) \beta_k \bar{x}_i^2 \right),
\]

since \( E_m (\beta_k \bar{x}_i)^2 \approx (\beta_k \bar{x}_i)^2 \), where \( \beta_k \) is given by (5). Note that

\[
E_m \left( \sum_{i \in \mathcal{P}} y_i^2 \right) = E_m \left( \sum_{k} \sum_{i \in \mathcal{P}_k} [(\beta_k \bar{x}_i)^2 + v_{ki} \sigma_{e_k}^2] \right).
\]

Hence, asymptotically, \( \sum_{i \in S} w_i y_i^2 \) has a bias

\[-E_m \left( \sum_{k} \sum_{i \in \mathcal{P}_k} (1 - a_i) v_{ki} \sigma_{e_k}^2 \right),
\]

which is always negative and is not of a negligible order.

Note that \( \sum_{i \in S} w_i y_i^2 \) is the standard unbiased estimator for \( \sum_{i \in \mathcal{P}} y_i^2 \) when there is no nonresponse. The previous result shows that the nonrandom regression imputation method is not unbiased for the marginal population second moment \( \sum_{i \in \mathcal{P}} y_i^2 \).

We now show that marginal nonrandom regression is not unbiased for the cross-product moment \( \sum_{i \in \mathcal{P}} y_i z_i \) either. Note that

\[
E_m \mathbb{E}_s \left( \sum_{i \in \mathcal{P}} w_i y_i^2 z_i \right) = E_m \mathbb{E}_s \left( \sum_{i \in \mathcal{P}} w_i a_i \beta_k \bar{x}_i z_i \right) + \sum_{k} \sum_{i \in \mathcal{P}_k} w_i (1 - a_i) b_i \beta_k \bar{x}_i z_i.
\]
\begin{align*}
+ \sum_k \sum_{i \in S_k} w_i a_i (1 - b_i) y_i \gamma_k x_i \\
+ \sum_k \sum_{i \in S_k} w_i (1 - a_i) (1 - b_i) \beta_k^i x_i \gamma_k^i x_i \\
\approx E_m \left( \sum_{i \in P} a_i b_i y_i z_i + \sum_k \sum_{i \in P_k} (1 - a_i) b_i \beta_k^i x_i \gamma_k x_i \\
+ \sum_k \sum_{i \in P_k} a_i (1 - b_i) \beta_k^i x_i \gamma_k^i x_i \\
+ \sum_k \sum_{i \in P_k} (1 - a_i) (1 - b_i) \beta_k^i x_i \gamma_k^i x_i \right)
\end{align*}

On the other hand, under (1) and (7),

\[ E_m \left( \sum_{i \in P} y_i z_i \right) = E_m \left( \sum_{i \in P} a_i b_i y_i z_i + \sum_k \sum_{i \in P_k} (1 - a_i b_i) (\beta_k^i x_i \gamma_k x_i + v_{k_i}^{1/2} u_{k_i}^{1/2} \sigma_{e,\eta,k}) \right), \]

where \( \sigma_{e,\eta,k} = E_m (e_i \eta_i) \) for \( i \in P_k \). Thus, \( \sum_{i \in S} w_i y_i^i z_i^i \) has an asymptotic bias

\[ -E_m \left( \sum_k \sum_{i \in P_k} (1 - a_i b_i) v_{k_i}^{1/2} u_{k_i}^{1/2} \sigma_{e,\eta,k} \right). \]

This bias is non-positive if \( \sigma_{e,\eta,k} \) is non-negative, which is true in most of practical problems. The reason for this bias is that the correlation between \( y_i \) and \( z_i \) is not preserved by their imputed values when at least one of \( y_i \) and \( z_i \) is imputed.

Consequently, \( \hat{\rho}^i \) in (9) based on marginal nonrandom regression imputation is not unbiased for the correlation coefficient \( \rho \).

### 2.3. Marginal Random Regression Imputation

The result in the previous section shows that marginal nonrandom regression imputation produces an asymptotically biased sample correlation coefficient and the bias comes from two sources: the estimation of the marginal second moments and the estimation of the cross-product moment.

We first consider the estimation of the marginal second moment. Is there an imputation method which is close to nonrandom regression imputation but is unbiased for the marginal second moment? We consider the random regression imputation method which adds a random term to the regression imputed nonresponder; i.e., a nonresponder \( y_i \) in \( S_k \) is imputed
by
\[ y_i^* = \beta_k x_i + v_{k_i}^{1/2} \epsilon_i^*, \] (11)
where, given the observed data, \( \epsilon_i^* \)'s are independent with mean 0 and variance
\[ \hat{\sigma}_{e,k}^2 = \sum_{i \in S_k} w_i a_i v_{k_i}^{-1} (y_i - \hat{\beta}_k x_i)^2 / \sum_{i \in S_k} w_i a_i \] (12)
(a consistent and asymptotically unbiased estimator of \( \sigma_{e,k}^2 \)). Adding random terms is common in imputation (Rubin 1987) and in confidentiality editing (Kim 1986). One way to generate \( \epsilon_i^* \)'s with mean 0 and variance given by (12) is to draw \( \epsilon_i^* \)'s according to
\[ P^*(\epsilon_i^* = r_{kj} - \bar{r}_k) = w_i a_i / \sum_{i \in S_k} w_i a_i \quad i \in S_k, \]
where \( r_{kj} = (y_j - \hat{\beta}_k x_i) / v_{k_i}^{1/2} \) and \( \bar{r}_k = \sum_{j \in S_k} w_j a_j r_{kj} / \sum_{i \in S_k} w_i a_i \). However, \( \epsilon_i^* \)'s can also be generated from an arbitrary distribution with mean 0 and variance \( \hat{\sigma}_{e,k}^2 \) (within the \( k \)th imputation cell). This flexibility may be useful in the implementation of random regression imputation.

If nonrespondents are imputed according to (11), then
\[ E_m E_s E_*( \sum_{i \in S} w_i y_i^* ) = E_m E_s \left( \sum_{i \in S} w_i a_i y_i + \sum_{k} \sum_{i \in S_k} w_i (1 - a_i) \beta_k x_i \right) \approx E_m(Y) \]
and
\[ E_m E_s E_*( \sum_{i \in S} w_i y_i^* ) = E_m E_s E_*( \sum_{i \in S} w_i a_i y_i^2 + \sum_{k} \sum_{i \in S_k} w_i (1 - a_i)(\beta_k^2 x_i + v_{k_i}^{1/2} \epsilon_i^* )^2 ) \]
\[ = E_m E_s \left( \sum_{i \in S} w_i a_i y_i^2 + \sum_{k} \sum_{i \in S_k} w_i (1 - a_i)((\beta_k^2 x_i + v_{k_i} \hat{\sigma}_{e,k}^2)^2 ) \right) \]
\[ \approx E_m \left( \sum_{i \in P} a_i y_i^2 + \sum_{k} \sum_{i \in S_k} (1 - a_i)((\beta_k^2 x_i + v_{k_i} \hat{\sigma}_{e,k}^2)^2 ) \right) \]
\[ \approx E_m \left( \sum_{i \in P} y_i^2 \right), \]
where \( E_\cdot \) is the expectation with respect to the random term in imputation,
\[ \hat{\sigma}_{e,k}^2 = \sum_{i \in S_k} a_i v_{k_i}^{-1} (y_i - \hat{\beta}_k x_i)^2 / \sum_{i \in S_k} a_i \]
and the result follows from \( E_s(\epsilon_i^* x_i) = 0 \), \( E_s(\epsilon_i^2) = \hat{\sigma}_{e,k}^2 \), \( E_s(\hat{\epsilon}_{e,k}^2) \approx \sigma_{e,k}^2 \), and \( E_m(\hat{\epsilon}_{e,k}^2) \approx \sigma_{e,k}^2 \). This shows that the random regression imputation method is unbiased for the marginal first and second moments.
The random regression imputation method imputes a nonrespondent \( z_i \) in \( S_k \) by
\[
z_i^* = \hat{\gamma}_k' \mathbf{x}_i + u_{k,t}^{1/2} \eta_i^*,
\]
where, given the observed data, \( \eta_i^* \)'s are independent with mean 0 and variance
\[
\hat{\sigma}_{\eta,k}^2 = \frac{\sum_{i \in S_k} w_i b_i u_{k,t}^{-1} (z_i - \hat{\gamma}_k' \mathbf{x}_i)^2}{\sum_{i \in S_k} w_i b_i}.
\]
(14)

For marginal imputation, items \( y \) and \( z \) are imputed according to (11) and (13) separately and \( \epsilon_i^* \)'s and \( \eta_i^* \)'s are generated independently.

We now consider the unbiasedness of this marginal random regression imputation method for the cross-product moment \( \sum_{i \in \mathcal{P}} y_i z_i \). Note that
\[
\sum_{i \in \mathcal{S}} w_i y_i^* z_i^* = \sum_{i \in \mathcal{S}} w_i a_i b_i y_i z_i + \sum_{k} \sum_{i \in S_k} w_i (1 - a_i) b_i (\hat{\beta}_k^* \mathbf{x}_i + u_{k,t}^{1/2} \epsilon_i^*) z_i
\]
\[
+ \sum_{k} \sum_{i \in S_k} w_i a_i (1 - b_i) y_i (\hat{\gamma}_k' \mathbf{x}_i + u_{k,t}^{1/2} \eta_i^*)
\]
\[
+ \sum_{k} \sum_{i \in S_k} w_i (1 - a_i) (1 - b_i) (\hat{\beta}_k^* \mathbf{x}_i + u_{k,t}^{1/2} \epsilon_i^*) (\hat{\gamma}_k' \mathbf{x}_i + u_{k,t}^{1/2} \eta_i^*).
\]

Then
\[
E_m E_s E_*(\sum_{i \in \mathcal{S}} w_i y_i^* z_i^*) = E_m E_s \left( \sum_{i \in \mathcal{S}} w_i a_i b_i y_i z_i + \sum_{k} \sum_{i \in S_k} w_i (1 - a_i) b_i \hat{\beta}_k^* \mathbf{x}_i z_i \right)
\]
\[
+ \sum_{k} \sum_{i \in S_k} w_i a_i (1 - b_i) y_i \hat{\gamma}_k' \mathbf{x}_i
\]
\[
+ \sum_{k} \sum_{i \in S_k} w_i (1 - a_i) (1 - b_i) \hat{\beta}_k^* \mathbf{x}_i \hat{\gamma}_k' \mathbf{x}_i
\]
\[
\approx E_m \left( \sum_{i \in \mathcal{P}} a_i b_i y_i z_i + \sum_{k} \sum_{i \in \mathcal{P}_k} (1 - a_i) b_i \hat{\beta}_k^* \mathbf{x}_i \hat{\gamma}_k' \mathbf{x}_i \right)
\]
\[
+ \sum_{k} \sum_{i \in \mathcal{P}_k} a_i (1 - b_i) \hat{\beta}_k^* \mathbf{x}_i \hat{\gamma}_k' \mathbf{x}_i
\]
\[
+ \sum_{k} \sum_{i \in \mathcal{P}_k} (1 - a_i) (1 - b_i) \hat{\beta}_k^* \mathbf{x}_i \hat{\gamma}_k' \mathbf{x}_i,
\]
which is the same as that for nonrandom regression imputation. Thus, \( \sum_{i \in \mathcal{S}} w_i y_i^* z_i^* \) has an asymptotic bias given by (10). The reason for this bias is the same as that for the case of marginal nonrandom regression imputation.

When \( \sigma_{\epsilon,\eta,k} = 0 \) for all \( k \) (i.e., \( y_i \) and \( z_i \) are conditionally unrelated, given \( \mathbf{x}_i \)), the bias given in (10) is 0. Thus, \( \hat{\rho}^* \) based on marginal random regression imputation is asymptotically unbiased. However, \( \hat{\rho}^* \) based on marginal nonrandom regression imputation is still
asymptotically biased, because of the bias in estimating the marginal second moments in the denominator of $\rho$ in (6).

3. JOINT RANDOM REGRESSION IMPUTATION

We have shown that the marginal random regression imputation method is unbiased for marginal first and second moments, but not unbiased for the correlation coefficient, except for the special case of $\sigma_{\epsilon \eta, k} = 0$ for all $k$. To obtain an unbiased imputation method for the correlation coefficient in general, some type of joint imputation is required. The simplest joint imputation method is vector imputation that imputes the whole vector of items whenever a sampled unit has at least one missing item. But vector imputation is rarely used because it discards many observed data and may be very inefficient.

3.1. Bivariate Case

We first consider the bivariate case where only $y$ and $z$ are the items of interest. We propose a joint random regression imputation method that is the same as the random regression imputation method described in Section 2.3 (i.e., missing $y_i$'s and $z_i$'s are imputed according to (11) and (13)) except that the random terms $\epsilon_i^*$ and $\eta_i^*$ are generated according to the following scheme:

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

$$
\epsilon_i^* = \frac{\hat{\sigma}_{\epsilon \eta, k}}{\hat{\sigma}_{\eta, k}} (z_i - \hat{\gamma}_k x_i) + \tilde{\epsilon}_i^*,
$$

where, given the observed data, $\tilde{\epsilon}_i^*$'s are independent random variables with mean 0 and variance $\hat{\sigma}_{\epsilon, k}^2 - \hat{\sigma}_{\epsilon \eta, k}^2 / \hat{\sigma}_{\eta, k}^2$, $\hat{\sigma}_{\epsilon, k}^2$ and $\hat{\sigma}_{\eta, k}^2$ are given by (12) and (14), respectively, and

$$
\hat{\sigma}_{\epsilon \eta, k} = \sum_{i \in S_k} w_i a_i b_i v_i^{-1/2} u_i^{-1/2} (y_i - \hat{\beta}_k x_i)(z_i - \hat{\gamma}_k x_i) / \sum_{i \in S_k} w_i a_i b_i.
$$

2. In the $k$th imputation cell, when $a_i = 1$ but $b_i = 0$,

$$
\eta_i^* = \frac{\hat{\sigma}_{\epsilon \eta, k}}{\hat{\sigma}_{\epsilon, k}} (y_i - \hat{\beta}_k x_i) + \tilde{\eta}_i^*,
$$

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

11
3. In the $k$th imputation cell, when both $a_i = 0$ and $b_i = 0$, $(\epsilon_i^*, \eta_i^*)$’s are independently (given the observed data) distributed with mean 0 and covariance matrix
\[
\begin{pmatrix}
\hat{\sigma}_{\epsilon,k}^2 & \hat{\sigma}_{\epsilon,\eta,k}
\end{pmatrix}
\begin{pmatrix}
\hat{\sigma}_{\eta,k}^2
\end{pmatrix}.
\]

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

1. When both $y_i$ and $z_i$ are nonrespondents, $\epsilon_i^*$ and $\eta_i^*$ are independent in marginal random regression imputation whereas for joint random regression imputation, $\epsilon_i^*$ and $\eta_i^*$ have covariance given by (16) which is a consistent estimator of the covariance between $\epsilon_i$ and $\eta_i$.

2. When exactly one of $y_i$ and $z_i$ is missing, $\epsilon_i^*$ or $\eta_i^*$ is generated according to (15) or (17) to ensure that the imputed pair of data preserves the same correlation as that of the original pair. Conditional on the observed data, $\epsilon_i^*$ or $\eta_i^*$ in (15) or (17) does not have zero mean, whereas $\epsilon_i^*$ or $\eta_i^*$ in marginal random regression imputation has zero mean.

The proposed joint random regression imputation is an extension of formulas (6)-(7) in Srivastava and Carter (1986) who considered the situation of no auxiliary $x$, normally distributed $(y_i, z_i)$’s, and uniform nonresponse (missing completely at random). Also, Srivastava and Carter (1986) proposed to draw random terms from residuals computed using respondents for all items, whereas in our method random terms can be generated from any distribution with the variances and covariances given by (12), (14) and (16), which is not only flexible for implementation but also more efficient especially for the multivariate case considered in Section 3.2. Furthermore, our method of generating random terms incorporates the survey weights $w_i$’s so that our results are applicable to complex surveys.

We now show that the estimated total $\hat{Y}^*$ in (4) and the sample correlation coefficient $\hat{\rho}^*$ in (9) based on this joint random regression imputation are asymptotically unbiased. First,
\[
E_{m}E_{s}E_{*}(\hat{Y}^*) = E_{m}E_{s}\left(\sum_{i \in S} w_i a_i y_i + \sum_{k \in S_k} \sum_{i \in S_k} w_i (1 - a_i) [\hat{\beta}_k^* x_i + \frac{\hat{\sigma}_{\epsilon,\eta,k}}{\sqrt{u_{ki}} \hat{\sigma}_{\eta,k}} (z_i - \hat{\gamma}_k^* x_i)]\right)
\approx E_{m}\left(\sum_{i \in P} a_i y_i + \sum_{k \in P_k} \sum_{i \in P_k} (1 - a_i) (\beta_k^* x_i + \frac{\sigma_{\epsilon,\eta,k}}{\sigma_{\eta,k}} - \eta_i)\right)
= E_{m}\left(\sum_{i \in P} y_i\right).
\]
Next,
\[ E_{m}E_{s}E_{s}\left(\sum_{i \in S} w_{i}(1 - a_{i})b_{i}y_{i}^{*}z_{i}\right) = E_{m}E_{s}\left(\sum_{k \in S_{k}} \sum_{i \in S_{k}} w_{i}(1 - a_{i})b_{i}\beta_{k}^{*}\mathbf{x}_{i}z_{i}\right.\]
\[ \quad + \sum_{k \in S_{k}} \sum_{i \in S_{k}} w_{i}(1 - a_{i})b_{i}\frac{v_{ki}^{1/2}\sigma_{\epsilon,\eta,k}}{u_{ki}\sigma_{\eta,k}^{2}}(z_{i} - \hat{\gamma}_{k}\mathbf{x}_{i})z_{i}\bigg)\]
\[ \approx E_{m}\left(\sum_{k \in P_{k}} \sum_{i \in P_{k}} (1 - a_{i})b_{i}(\beta_{k}^{*}\mathbf{x}_{i}z_{i} + \frac{v_{ki}^{1/2}\sigma_{\epsilon,\eta,k}}{\sigma_{\eta,k}^{2}}\eta_{i}z_{i})\right)\]
\[ = E_{m}\left(\sum_{k \in P_{k}} \sum_{i \in P_{k}} (1 - a_{i})b_{i}\beta_{k}^{*}\mathbf{x}_{i} + \frac{v_{ki}^{1/2}\sigma_{\epsilon,\eta,k}}{\sigma_{\eta,k}^{2}}\eta_{i}\right)\]
\[ = E_{m}\left(\sum_{k \in P_{k}} \sum_{i \in P_{k}} (1 - a_{i})b_{i}y_{i}z_{i}\right).\]

Similarly,
\[ E_{m}E_{s}E_{s}\left(\sum_{i \in S} w_{i}a_{i}(1 - b_{i})y_{i}z_{i}^{*}\right) \approx E_{m}\left(\sum_{k \in P_{k}} a_{i}(1 - b_{i})y_{i}z_{i}\right)\]

and
\[ E_{m}E_{s}E_{s}\left(\sum_{i \in S} w_{i}(1 - a_{i})(1 - b_{i})y_{i}^{*}z_{i}^{*}\right) \approx E_{m}\left(\sum_{k \in P_{k}} (1 - a_{i})(1 - b_{i})y_{i}z_{i}\right).\]

Thus,
\[ E_{m}E_{s}E_{s}\left(\sum_{i \in S} w_{i}y_{i}z_{i}^{*}\right) \approx E_{m}\left(\sum_{i \in P} y_{i}z_{i}\right).\]

Finally, note that
\[ E_{s}(\epsilon_{i}^{2}) = \frac{\sigma_{\epsilon,\eta,k}^{2}}{u_{ki}\sigma_{\eta,k}^{4}}(z_{i} - \hat{\gamma}_{k}\mathbf{x}_{i})^{2} + \hat{\sigma}_{\epsilon,k}^{2} - \frac{\sigma_{\epsilon,\eta,k}^{2}}{\sigma_{\eta,k}^{2}}.\]

Then
\[ E_{m}E_{s}E_{s}\left(\sum_{i \in S} w_{i}y_{i}^{2}\right) = E_{m}E_{s}\left(\sum_{i \in S} w_{i}a_{i}y_{i}^{2} + \sum_{k \in S_{k}} \sum_{i \in S_{k}} w_{i}(1 - a_{i})[(\beta_{k}^{*}\mathbf{x}_{i})^{2} + v_{ki}\sigma_{\epsilon,k}^{2}]\right.\]
\[ \quad + \sum_{k \in S_{k}} \sum_{i \in S_{k}} w_{i}(1 - a_{i})\frac{\sigma_{\epsilon,\eta,k}^{2}}{u_{ki}\sigma_{\eta,k}^{2}}(z_{i} - \hat{\gamma}_{k}\mathbf{x}_{i})^{2} - \frac{\sigma_{\epsilon,\eta,k}^{2}}{\sigma_{\eta,k}^{2}}\bigg)\]
\[ \approx E_{m}\left(\sum_{i \in P} a_{i}y_{i}^{2} + \sum_{k \in P_{k}} \sum_{i \in P_{k}} (1 - a_{i})[(\beta_{k}^{*}\mathbf{x}_{i})^{2} + v_{ki}\sigma_{\epsilon,k}^{2}]\right.\]
\[ \quad + \sum_{k \in P_{k}} \sum_{i \in P_{k}} (1 - a_{i})\frac{\sigma_{\epsilon,\eta,k}^{2}\eta_{i}^{2}}{\sigma_{\eta,k}^{2}} - \frac{\sigma_{\epsilon,\eta,k}^{2}}{\sigma_{\eta,k}^{2}}\bigg)\]
\[ = E_{m}\left(\sum_{i \in P} a_{i}y_{i}^{2} + \sum_{k \in P_{k}} \sum_{i \in P_{k}} (1 - a_{i})[(\beta_{k}^{*}\mathbf{x}_{i})^{2} + v_{ki}\sigma_{\epsilon,k}^{2}]\right).\]
\[ E_m \left( \sum_{i \in P} y_i^2 \right) \, . \]

### 3.2. Multivariate Case

Consider the estimation of marginal totals of \( m \) items as well as correlation coefficients between all pairs of these \( m \) items. Let \( t_i = (y_i, z_i, \ldots)' \) be the \( m \)-vector of items of interest from unit \( i \). Model (1)-(2) and (7)-(8) should be replaced by the following one:

\[ t_i = B'_k x_i + V_{ki}^{1/2} e_i \quad i \in P_k \]

and

\[ P(a_i = a | t_i, x_i) = P(a_i = a | x_i) \quad i \in P_k, \]

where \( B_k = (\beta_k, \gamma_k, \ldots) \) is a matrix of regression parameters, \( V_{ki} \) is a diagonal matrix whose diagonal elements are \( v_{ki} = v_k(x_i), u_{ki} = u_k(x_i), \ldots \), with known positive functions \( v_k, u_k, \ldots \), \( e_i \) is a random error (independent of \( x_i \)) with mean 0 and unknown covariance matrix \( \Sigma_k \), and \( a_i \) is the vector of response indicators for \( t_i \). Let \( \hat{B}_k = (\hat{\beta}_k, \hat{\gamma}_k, \ldots) \) with \( \hat{\beta}_k \) computed according to (3) and \( \hat{\gamma}_k, \ldots \), computed similarly, and let \( \hat{\Sigma}_k \) be the estimator of \( \Sigma_k \) whose diagonal elements (variances) are computed according to (12) and (14) and off-diagonal elements (covariances) are computed according to (16). The joint random regression imputation method in this multivariate case can be described as follows. For a given sampled unit \( t_i \) with item nonresponse, let \( t_{0i} \) be the sub-vector of \( t_i \) containing all nonrespondents and \( t_{1i} \) be the sub-vector containing all respondents (the dimension of \( t_{1i} \) may be 0, in which case all items in \( t_i \) are nonrespondents). Without loss of generality we assume \( t'_{0i} = (t'_{0i}, t'_{1i}) \). Write

\[ \hat{B}_k = (\hat{B}_{0k}, \hat{B}_{1k}), \quad V_{ki} = \begin{pmatrix} V_{0ki} & 0 \\ 0 & V_{1ki} \end{pmatrix} \quad \text{and} \quad \hat{\Sigma}_k = \begin{pmatrix} \hat{\Sigma}_{0k} & \hat{\Sigma}_{01k} \\ \hat{\Sigma}_{10k} & \hat{\Sigma}_{1k} \end{pmatrix}, \]

where \( \hat{B}_{0k}, V_{0ki}, \) and \( \hat{\Sigma}_{0k} \) have the same dimension as \( t_{0i} \). Then \( t_{0i} \) is imputed by

\[ t_{0i}^* = \hat{B}'_{0k} x_i + V_{0ki}^{1/2} \left[ \hat{\Sigma}_{01k}^{-1} V_{1ki}^{-1/2} (t_{1i} - \hat{B}'_{1k} x_i) + \hat{e}_i^* \right], \]

where, given the observed data, \( \hat{e}_i^* \)'s are independent random vectors with mean 0 and covariance matrix \( \hat{\Sigma}_{0k} - \hat{\Sigma}_{01k} \hat{\Sigma}_{1k}^{-1} \hat{\Sigma}_{10k} \) (\( \hat{\Sigma}_{1k}^{-1} \) is defined to be 0 if the dimension of \( t_{1i} \) is 0).

Calculations similar to those in Section 3.1 show that this joint random regression imputation method is unbiased for estimating marginal totals and correlation coefficients among items.
Our method of generating random term $\hat{e}_i^*$ is better than generating $\hat{e}_i^*$ from residuals computed using respondents for all items when there are many items, because each element of $\hat{\Sigma}_k$ is computed using a data set substantially larger than the set of all respondents (each diagonal element of $\hat{\Sigma}_k$ is computed using respondents from one item and each off-diagonal element of $\hat{\Sigma}_k$ is computed using respondents for two items).

3.3. A Simulation Study

Some simulation results are presented here to study the finite sample performance of the sample correlation coefficient $\hat{\rho}^*$ under joint random regression imputation and $\hat{\rho}^*$ under marginal nonrandom regression imputation. The sample correlation coefficient $\hat{\rho}$ based on data without nonresponse is also included in the study as a standard.

In the simulation, the sampling design is 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 (7) with a univariate $x$ and $u_{kt}(x) = u_{kt}(x) = x$. Stratum sample sizes, survey weights, and model parameter values are listed in Table 1. The $x$'s values are independently generated from gamma distributions with means and variances given by those in Table 1. The error terms $e_i$ and $\eta_i$ are independently generated according to

$$
\begin{align*}
   e_i &= \kappa \zeta_i + \delta_i \\
   \eta_i &= \kappa \zeta_i + \tau_i,
\end{align*}
$$

where $\zeta_i$, $\delta_i$, and $\tau_i$ are independently distributed as $N(0,1)$ and $\kappa \geq 0$ is a parameter. As $\kappa$ increases, the value of the correlation coefficient increases. When $\kappa = 0$, $y_i$ and $z_i$ are conditionally unrelated, given $x_i$.

Given the generated data, respondents are generated according to (2) and (8), with independent $a_i$'s and $b_i$'s (although our results in Section 3.1 hold even when $a_i$ and $b_i$ are related),

$$
P(a_i = 1|x_i) = \frac{e^{0.1+0.05x_i}}{1 + e^{0.1+0.05x_i}}
$$

and

$$
P(b_i = 1|x_i) = \frac{e^{0.2+0.04x_i}}{1 + e^{0.2+0.04x_i}}.
$$

The average response rates, $E[P(a_i = 1|x_i)]$ and $E[P(b_i = 1|x_i)]$, are approximately 61.75%.
Table 2 lists the mean and standard deviation of three sample correlation coefficients for some values of $\kappa$, 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 random regression imputation is very close to the mean of the sample correlation coefficient $\hat{\rho}$, which is used as a standard. This supports our theory of the asymptotic unbiasedness of the joint random regression imputation method. The standard deviation of $\hat{\rho}^*$ under joint random 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 cancelled out). Comparing standard deviations of $\hat{\rho}^*$ for marginal nonrandom imputation and joint random imputation, we find that marginal nonrandom imputation has lower standard deviation when $\kappa \leq 1$, whereas joint random imputation has lower standard deviation when $\kappa \geq 1.2$. The variability in marginal nonrandom imputation increases with the value of $\kappa$.

4. 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. Srivastava and Carter (1986) did not study variance estimation. 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 (Taylor expansion) and substitution. However, under joint random 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 first proposed by Rao and Shao (1992) for estimating the variance of $\hat{\gamma}^*$ under marginal random hot deck imputation (no covariate $x$).

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 \( X = \{x_i, w_i : i \in S\} \) denote the dataset including covariates and survey weights and let \( X_{hj} = \{x_i, w_i^{(hj)} : i \in S\} \) be the \((h,j)\)th pseudoreplicate after deleting the first stage cluster \( j \) in stratum \( h \) and suitably adjusting the survey weights, where

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

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

\[
   v_{j\text{ack}} = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} \left[ \hat{\theta}(X_{hj}) - \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{\theta}(X_{hi}) \right]^2.
\]  

(19)

Note that the computation of \( \hat{\theta}(X_{hj}) \)'s can be done repeatedly using a program similar to that for computing \( \hat{\theta}(X) \).

When there are imputed nonrespondents, formula (19) has to be modified to take imputation into account. Let \( X^* \) and \( X_{hj}^* \) be the same as \( X \) and \( X_{hj} \) but with imputed nonrespondents, and let \( \hat{B}_k^{(hj)} \) and \( \hat{\Sigma}_k^{(hj)} \) be \( \hat{B}_k \) and \( \hat{\Sigma}_k \), respectively, with \( w_i \)'s replaced by \( w_i^{(hj)} \)'s (i.e., \( \hat{B}_k^{(hj)} \) and \( \hat{\Sigma}_k^{(hj)} \) are \( \hat{B}_k \) and \( \hat{\Sigma}_k \) re-calculated using observed data in \( X_{hj} \)). We consider the following adjustment. For each imputed vector \( t_{i}^{*} \) in \( X_{hj}^{*} \) (see (18)), we re-impute \( t_{i}^{*} \) using the same imputation method but the observed data in \( X_{hj} \). More precisely, for each \((h,j)\), the re-imputed vector is the same as \( t_{i}^{*} \) in (18) but with \( \hat{B}_k \) and \( \hat{\Sigma}_k \) replaced by \( \hat{B}_k^{(hj)} \) and \( \hat{\Sigma}_k^{(hj)} \), respectively, and with \( \hat{\varepsilon}_i^* \) replaced by \((\hat{\Sigma}_{ek}^{(hj)} \hat{\Sigma}_{ek}^{-1})^{1/2} \hat{\varepsilon}_i^*\), where \( \hat{\Sigma}_{ek} = \hat{\Sigma}_{0k} - \hat{\Sigma}_{01k} \hat{\Sigma}_{1k}^{-1} \hat{\Sigma}_{10k} \) and \( \hat{\Sigma}_{ek}^{(hj)} \) is \( \hat{\Sigma}_{ek} \) re-calculated using \( w_i^{(hj)} \)'s instead of \( w_i \)'s. In the bivariate case (Section 3.1), for example, if \( a_t = 0 \) and \( b_t = 1 \), then \( y_t^{*} \) is re-imputed by

\[
   (\hat{\beta}_k^{(hj)})'x_i + \frac{\hat{\gamma}_k^{(hj)} - \hat{\gamma}_k^{(hj)}}{\hat{\gamma}_k^{(hj)}} \hat{\gamma}_k^{(hj)} x_i + \left( \frac{\hat{\gamma}_k^{(hj)} - \hat{\gamma}_k^{(hj)}}{\hat{\gamma}_k^{(hj)}} \hat{\gamma}_k^{(hj)} x_i \right)^{-1} \hat{\varepsilon}_i^*,
\]

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

Let \( X_{hj}^{*\text{RI}} \) be the re-imputed \( X_{hj}^{*} \) and let \( \hat{\theta} = \hat{\theta} \) for \( \hat{\theta} = \hat{\theta} \) and \( \hat{\rho} = \hat{\rho} \).
The adjusted jackknife variance estimator $v_{\text{jack}}^{\text{Adj}}$ is then obtained using (19) with $X_{hj}$ replaced by $X_{hj}^{\text{RRI}}$.

To show that $v_{\text{jack}}^{\text{Adj}}$ is asymptotically unbiased and consistent for the asymptotic variance of $\hat{Y}^*$ or $\hat{\rho}^*$, let us consider for simplicity the special case of bivariate $t_i = (y_i, z_i)'$. Note that both $\hat{Y}^*$ and $\hat{\rho}^*$ are differentiable functions of weighted averages of the form $\sum_{i \in S_k} w_i y_i$. For example, $\hat{Y}^* = \hat{Y}(X^*)$ 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),

$$
\sum_{i \in S_k} w_i a_i y_i, \quad \sum_{i \in S_k} w_i (1 - a_i) b_i x_i, \quad \sum_{i \in S_k} w_i (1 - a_i) b_i v_{ki}^{1/2} u_{ki}^{-1/2} z_i,
$$

$$
\sum_{i \in S_k} w_i (1 - a_i) b_i v_{ki}^{1/2} u_{ki}^{-1/2} x_i, \quad \sum_{i \in S_k} w_i (1 - a_i) (1 - b_i) x_i,
$$

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

where $\delta_i = \epsilon_i^*/(\hat{\sigma}_{\epsilon, k}^2 - \hat{\sigma}_{\epsilon, \eta, k}^2/\hat{\sigma}_{\eta, k}^2)^{1/2}$ and $\tau_i = \epsilon_i^*/\hat{\sigma}_{\epsilon, k}$ are random variables with mean 0 and variance 1. (Since $\hat{Y}^*$ 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 $\hat{\theta}(X_{hj}^{\text{RRI}})$ is the same function of the same weighted averages except that weights $w_i$'s are replaced by $w_i^{(hj)}$'s. From the theory of jackknife (see, e.g., Krewski and Rao 1981), the jackknife variance estimator $v_{\text{jack}}^{\text{Adj}}$ is asymptotically unbiased and consistent (under the asymptotic setting in the Appendix). Note that the same conclusion cannot be reached if we use $\hat{\theta}(X_{hj}^{*})$ in formula (19).

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

We study the finite sample performance of the proposed $v_{\text{jack}}^{\text{Adj}}$ in the simulation setting described in Section 3.3 with $\kappa = 1$. The result on the mean and standard deviation of $\sqrt{v_{\text{jack}}^{\text{Adj}}}$ as an estimator of the standard deviation of $\hat{\rho}^*$ under joint random regression imputation is given in Table 3. The proposed jackknife estimator has a negligible bias and a reasonably small standard deviation. For comparison, we also include the results for the naive jackknife estimator obtained by treating imputed values as observed data (i.e., using $\hat{\theta}(X_{hj}^{*})$ instead of $\hat{\theta}(X_{hj}^{\text{RRI}})$ in (19)). As expected, the naive jackknife estimator has a very large negative bias.
APPENDIX

The asymptotic results in this article is based on the following asymptotic framework (see, for example, 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 $\nu$, but $\nu$ is suppressed for simplicity of notation. As $\nu \to \infty$, $n \to \infty$, $n/N \to 0$, and

$$\max_{h,j} \sum_{i \in S(h,j)} w_i/M = O(n^{-1})$$

where $n = \sum_h n_h$, $N = \sum_h N_h$, $M$ is the total number of ultimate units in $\mathcal{P}$, and $S(h,j)$ is the set of indices of sampled units in stratum $h$ and cluster $j$. Also, as $\nu \to \infty$,

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

for some $\delta > 0$, where $\| \|$ is the usual vector norm and $\psi_{hj} = \sum_{i \in S(h,j)} w_i \psi_i / M$ with $\psi_i$ being any component of the matrix $t_i t_i'$ or $x_i x_i'$. Finally, as $\nu \to \infty$ $0 < \lim inf[n \text{Var}(\hat{Y}/M)]$ and $0 < \lim inf[n \text{Var}(\hat{p})]$.

Under this asymptotic setting, various claimed results on approximations, asymptotic unbiasedness, and consistency can be proved using standard arguments in Krewski and Rao (1981), Bickel and Freedman (1984), and Valliant (1993).

REFERENCES


Table 1. Sample size, survey weight, and model parameters across imputation classes and strata

<table>
<thead>
<tr>
<th>Imputation Class</th>
<th>Stratum</th>
<th>Sample Size</th>
<th>Survey Weight</th>
<th>Model Parameters</th>
<th>( E(x) )</th>
<th>( \sqrt{Var(x)} )</th>
<th>( \beta )</th>
<th>( \gamma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>14</td>
<td>12.43</td>
<td></td>
<td>2.0</td>
<td>0.1</td>
<td>1.0</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>11</td>
<td>8.91</td>
<td></td>
<td>3.0</td>
<td>0.1</td>
<td>1.0</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>10</td>
<td>6.10</td>
<td></td>
<td>4.0</td>
<td>0.1</td>
<td>1.0</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>11</td>
<td>5.73</td>
<td></td>
<td>5.0</td>
<td>0.1</td>
<td>1.0</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>16</td>
<td>2.70</td>
<td></td>
<td>6.0</td>
<td>0.1</td>
<td>1.0</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>18</td>
<td>2.17</td>
<td></td>
<td>7.0</td>
<td>0.1</td>
<td>1.0</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>31</td>
<td>1.00</td>
<td></td>
<td>8.0</td>
<td>0.1</td>
<td>1.0</td>
<td>0.6</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>8</td>
<td>32.91</td>
<td></td>
<td>2.5</td>
<td>0.2</td>
<td>0.5</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>13</td>
<td>9.85</td>
<td></td>
<td>3.5</td>
<td>0.2</td>
<td>0.5</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11</td>
<td>10.82</td>
<td></td>
<td>4.5</td>
<td>0.2</td>
<td>0.5</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>12</td>
<td>6.08</td>
<td></td>
<td>5.5</td>
<td>0.2</td>
<td>0.5</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>13</td>
<td>3.60</td>
<td></td>
<td>6.5</td>
<td>0.2</td>
<td>0.5</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>86</td>
<td>1.00</td>
<td></td>
<td>7.5</td>
<td>0.2</td>
<td>0.5</td>
<td>0.7</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>14</td>
<td>87.91</td>
<td></td>
<td>3.0</td>
<td>0.1</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>11</td>
<td>67.39</td>
<td></td>
<td>4.0</td>
<td>0.1</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>13</td>
<td>44.48</td>
<td></td>
<td>5.0</td>
<td>0.1</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>14</td>
<td>25.28</td>
<td></td>
<td>6.0</td>
<td>0.1</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>16</td>
<td>15.57</td>
<td></td>
<td>7.0</td>
<td>0.1</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>18</td>
<td>9.80</td>
<td></td>
<td>8.0</td>
<td>0.1</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>15</td>
<td>6.23</td>
<td></td>
<td>9.0</td>
<td>0.1</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>15</td>
<td>4.68</td>
<td></td>
<td>10.0</td>
<td>0.1</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>40</td>
<td>2.13</td>
<td></td>
<td>11.0</td>
<td>0.1</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>38</td>
<td>1.00</td>
<td></td>
<td>12.0</td>
<td>0.1</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>7</td>
<td>32.14</td>
<td></td>
<td>3.5</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>13</td>
<td>16.75</td>
<td></td>
<td>4.5</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>10</td>
<td>12.90</td>
<td></td>
<td>5.5</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>14</td>
<td>7.00</td>
<td></td>
<td>6.5</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>13</td>
<td>6.18</td>
<td></td>
<td>7.5</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>11</td>
<td>4.70</td>
<td></td>
<td>8.5</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>17</td>
<td>3.31</td>
<td></td>
<td>9.5</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>19</td>
<td>1.89</td>
<td></td>
<td>10.5</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>22</td>
<td>1.82</td>
<td></td>
<td>11.5</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>28</td>
<td>1.00</td>
<td></td>
<td>12.5</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Table 2. Simulation averages of mean and standard deviation (SD) of sample correlation coefficients

<table>
<thead>
<tr>
<th>$\kappa$</th>
<th>Without nonresponse Mean</th>
<th>Without nonresponse SD</th>
<th>Marginal nonrandom regression imputation Mean</th>
<th>Marginal nonrandom regression imputation SD</th>
<th>Joint random regression imputation Mean</th>
<th>Joint random regression imputation SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.5521</td>
<td>0.0351</td>
<td>0.6618</td>
<td>0.0494</td>
<td>0.5562</td>
<td>0.0553</td>
</tr>
<tr>
<td>0.2</td>
<td>0.5563</td>
<td>0.0362</td>
<td>0.6620</td>
<td>0.0498</td>
<td>0.5604</td>
<td>0.0561</td>
</tr>
<tr>
<td>0.4</td>
<td>0.5779</td>
<td>0.0378</td>
<td>0.6656</td>
<td>0.0478</td>
<td>0.5783</td>
<td>0.0540</td>
</tr>
<tr>
<td>0.6</td>
<td>0.6090</td>
<td>0.0351</td>
<td>0.6697</td>
<td>0.0483</td>
<td>0.6010</td>
<td>0.0587</td>
</tr>
<tr>
<td>0.8</td>
<td>0.6467</td>
<td>0.0332</td>
<td>0.6796</td>
<td>0.0485</td>
<td>0.6364</td>
<td>0.0577</td>
</tr>
<tr>
<td>1.0</td>
<td>0.6834</td>
<td>0.0314</td>
<td>0.6854</td>
<td>0.0524</td>
<td>0.6687</td>
<td>0.0568</td>
</tr>
<tr>
<td>1.2</td>
<td>0.7223</td>
<td>0.0283</td>
<td>0.6917</td>
<td>0.0569</td>
<td>0.7084</td>
<td>0.0517</td>
</tr>
<tr>
<td>1.4</td>
<td>0.7576</td>
<td>0.0251</td>
<td>0.6990</td>
<td>0.0652</td>
<td>0.7404</td>
<td>0.0475</td>
</tr>
<tr>
<td>1.6</td>
<td>0.7858</td>
<td>0.0230</td>
<td>0.7065</td>
<td>0.0725</td>
<td>0.7696</td>
<td>0.0482</td>
</tr>
<tr>
<td>1.8</td>
<td>0.8130</td>
<td>0.0210</td>
<td>0.7081</td>
<td>0.0797</td>
<td>0.7947</td>
<td>0.0422</td>
</tr>
<tr>
<td>2.0</td>
<td>0.8345</td>
<td>0.0192</td>
<td>0.7155</td>
<td>0.0848</td>
<td>0.8190</td>
<td>0.0409</td>
</tr>
<tr>
<td>2.4</td>
<td>0.8709</td>
<td>0.0145</td>
<td>0.7238</td>
<td>0.0998</td>
<td>0.8523</td>
<td>0.0345</td>
</tr>
<tr>
<td>2.8</td>
<td>0.8976</td>
<td>0.0120</td>
<td>0.7286</td>
<td>0.1027</td>
<td>0.8821</td>
<td>0.0279</td>
</tr>
<tr>
<td>3.2</td>
<td>0.9166</td>
<td>0.0104</td>
<td>0.7322</td>
<td>0.1146</td>
<td>0.9013</td>
<td>0.0248</td>
</tr>
<tr>
<td>3.6</td>
<td>0.9311</td>
<td>0.0087</td>
<td>0.7442</td>
<td>0.1225</td>
<td>0.9167</td>
<td>0.0219</td>
</tr>
<tr>
<td>4.0</td>
<td>0.9432</td>
<td>0.0070</td>
<td>0.7520</td>
<td>0.1247</td>
<td>0.9295</td>
<td>0.0195</td>
</tr>
</tbody>
</table>

Table 3. Simulation averages of mean and standard deviation (SD) of jackknife estimators of the SD of $\hat{\rho}^*$ under joint random regression imputation

| True SD of $\hat{\rho}^*$ Mean | The naive jackknife Mean | 0.0568 | 0.0309 | 0.0060 | 0.0563 | 0.0155 |