Statistics and Its Interface Volume 6 (2013) 361–368

Estimation and imputation in linear regression with missing values in both response and covariate

Jun Shao*

We consider linear regression with missing responses as well as missing covariate data. When the missing data mechanism is ignorable, we show that regression parameters and the response mean can be estimated using standard methods and treating imputed values as observed data. We also show that the same procedure results in biased and inconsistent estimators when missing response mechanism depends on covariates that also have missing values and thus is nonignorable. Efficient estimation and imputation under nonignorable missingness is a challenge problem. Under some conditions, we derive some asymptotically unbiased and consistent estimators via direct estimation or imputation. Some simulation results are presented to examine the finite sample performance of various estimators.

AMS 2000 subject classifications: Primary 62J05; secondary 62G20.
Keywords and phrases: Asymptotic unbiasedness and consistency, Imputation, Linear regression, Missing covariate data, Missing response data, Nonignorable missingness.

1. INTRODUCTION

Nonresponse or data missing with an appreciable rate exists in many applications in areas such as medicine, population health, economics, social sciences, and sample surveys. Let y be a response variable of interest and $\boldsymbol{x}$ be a covariate vector associated with y. We consider the regression between y and $\boldsymbol{x}$ and/or the estimation of the mean of y using $\boldsymbol{x}$ as an auxiliary variable. There is rich literature for the case where y has missing values but $\boldsymbol{x}$ is always observed (see, for example, Little and Rubin, 2002; Kim and Shao, 2013). In the situation where $\boldsymbol{x}$ has missing values but y is always observed, which is referred to as the problem of missing covariate values, a review given by Little (1992) summarized results obtained prior to 1992 but not much has been added since that time. The focus of the current paper is on the situation where both y and $\boldsymbol{x}$ have missing data, a problem often encounted nowadays but has not been thoroughly discussed in the literature.

*Partially supported by the NSF Grant DMS-1007454.

When $\boldsymbol{x}$ has missing data but y does not, what is often done in practice is that missing data in $\boldsymbol{x}$ are first imputed (using some appropriate method given in the literature) and then a further analysis on the association between y and $\boldsymbol{x}$ or estimation of the mean of y is carried out by treating imputed values in $\boldsymbol{x}$ as observed values. A conclusion in Little (1992) is that treating imputed covariate values as observed data for further analysis of y data is valid provided that

(i) the conditional expectation $E(y|\boldsymbol{x})$ is linear in $\boldsymbol{x}$;
(ii) the missing data mechanism for $\boldsymbol{x}$ is ignorable in the sense that it depends only on the observed part of $\boldsymbol{x}$;
(iii) imputation for missing values in $\boldsymbol{x}$ is done using expectations conditional on the observed part of $\boldsymbol{x}$.

However, the main finding of the current paper is that the conclusion is different when y also has missing values although (i)–(iii) still hold. If the missing data mechanism of y is also ignorable, which is a somewhat too strong assumption as we explained in Section 4, then treating imputed covariate values as observed data leads to nearly unbiased estimators of regression parameters and the mean of y; otherwise treating imputed covariate values as observed data results in biased estimators.

Details on notation and assumption on the model and the missing data mechanism that generates observations are described in Section 2. The results under the approach of treating imputed covariate values as observed data are given in Section 3. In Section 4, we consider some estimation methods valid in the presence of nonignorable missing data. Some simulation results are presented in Section 5. The last section contains some discussion.

2. NOTATION AND ASSUMPTION

Let $\boldsymbol{u}$ and $\boldsymbol{z}$ be two covariate vectors such that $\boldsymbol{u}$ has missing values but $\boldsymbol{z}$ is always observed. We denote $\boldsymbol{x} = (1, \boldsymbol{u}', \boldsymbol{z}')'$, where $\boldsymbol{t}'$ denotes the transpose of a column vector $\boldsymbol{t}$, as the entire covariate vector including a constant component. For the variable y of interest, we assume a linear model for the conditional expectation

$$E(y|\boldsymbol{x}) = \theta'\boldsymbol{x} = \alpha + \beta_u'\boldsymbol{u} + \beta_z'\boldsymbol{z}, \tag{1}$$

where $\theta = (\alpha, \beta_u', \beta_z')'$, α is an unknown parameter, and β_u and β_z are unknown parameter vectors with dimensions the same as $\boldsymbol{u}$ and $\boldsymbol{z}$, respectively. We assume that the conditional expectation

$$E(\boldsymbol{u}|\boldsymbol{z}) = \Gamma \boldsymbol{z} \tag{2}$$

also follows a linear model, where Γ is an unknown parameter matrix that has row dimension the same as $\boldsymbol{u}$ and column dimension the same as $\boldsymbol{z}$. Note that a vector of intercepts can be added to the right-hand side of (2), but we omit it since the discussion with the intercept vector is the same by applying some transformations of covariates. Furthermore, we assume that, conditioned on $\boldsymbol{z}$, components of $\boldsymbol{u}$ are independent.

When there are missing data, we use a to denote the indicator of whether y is observed ($a = 1$ if y is observed and $a = 0$ otherwise) and use $\boldsymbol{b}$ to denote the vector whose components are the indicators of whether components of $\boldsymbol{u}$ are observed. We assume that the missing data mechanism for $\boldsymbol{u}$ satisfies

$$p(\boldsymbol{b}|y, \boldsymbol{u}, \boldsymbol{z}) = p(\boldsymbol{b}|\boldsymbol{z}), \tag{3}$$

where $p(\cdot|\cdot)$ denotes conditional probability density. For the missing data mechanism of y, we assume that

$$p(a|y, \boldsymbol{b}, \boldsymbol{u}, \boldsymbol{z}) = p(a|\boldsymbol{u}, \boldsymbol{z}). \tag{4}$$

Note that (3) is stronger than the general ignorable missing data assumption, whereas (4) is nonignorable since $\boldsymbol{u}$ has missing values.

Note that assumptions (3)–(4) are nonparametric since we do not have any condition on the form of conditional densities; assumptions (1)–(2) are semi-parametric since the conditional expectations are parametric but the forms of the densities are unspecified.

Under (3)–(4),

$$p(\boldsymbol{b}|y, \boldsymbol{u}, \boldsymbol{z}, a) = \frac{p(a|y, \boldsymbol{u}, \boldsymbol{z}, \boldsymbol{b}) p(\boldsymbol{b}|y, \boldsymbol{u}, \boldsymbol{z})}{p(a|y, \boldsymbol{u}, \boldsymbol{z})} = p(\boldsymbol{b}|\boldsymbol{z}), \tag{5}$$

a result seems to be stronger than (3) but is actually equivalent to (3).

Throughout we assume that n subjects are sampled and that $(y_i, \boldsymbol{u}_i, \boldsymbol{z}_i, a_i, \boldsymbol{b}_i)$, $i = 1, \ldots, n$, are independent and identically distributed as $(y, \boldsymbol{u}, \boldsymbol{z}, a, \boldsymbol{b})$. For the ith sampled subject, y_i is observed if and only if $a_i = 1$, components of $\boldsymbol{u}_i$ are observed if and only if their corresponding components of $\boldsymbol{b}_i$ are equal to 1, and $\boldsymbol{z}_i$ is always observed. Based on the observed data, we are interested in the estimation of θ in (1) and the mean of y, $\mu = E(y)$.

3. TREATING IMPUTED COVARIATE VALUES AS OBSERVED DATA

Let u be a univariate component of $\boldsymbol{u}$ and γ' be the row of Γ in (2) corresponding to u, i.e., $E(u|\boldsymbol{z}) = \gamma' \boldsymbol{z}$. Also, let b be the component of $\boldsymbol{b}$ corresponding to the component u. Based on assumption (3), if u_i (the u-value of the ith sampled subject) is missing, it can be imputed by $\hat{\gamma}' \boldsymbol{z}_i$ with

$$\hat{\gamma} = \left(\sum_{i=1}^{n} b_i \boldsymbol{z}_i \boldsymbol{z}_i' \right)^{-1} \sum_{i=1}^{n} b_i u_i \boldsymbol{z}_i. \tag{6}$$

For missing values in $\boldsymbol{u}_i$, imputation is done component-wise. Let $\hat{\boldsymbol{u}}_i$ denote the vector $\boldsymbol{u}_i$ with missing components imputed according to this imputation procedure, and let

$$\hat{\boldsymbol{x}}_i = (1, \hat{\boldsymbol{u}}_i', \boldsymbol{z}_i')'.$$

Note that assumptions (1) and (3) correspond to requirements (i) and (ii) in the discussion in Section 1, respectively, while imputation as previously described is exactly the same as (iii) in Section 1 under model (2).

3.1 Estimation of θ

If we treat imputed covariate values as observed data, then we estimate $\theta = (\alpha, \beta_u', \beta_z')'$ by

$$\tilde{\theta} = \left(\sum_{i=1}^{n} a_i \hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i' \right)^{-1} \sum_{i=1}^{n} a_i y_i \hat{\boldsymbol{x}}_i. \tag{7}$$

Although each $\hat{\gamma}$ in (6) is asymptotically unbiased (as $n \to \infty$) under assumption (3) (ignorable missing u-values), $\tilde{\theta}$ in (7) is asymptotically biased under assumption (4) (non-ignorable missing y-values). This is because

$$\begin{aligned} E\left(\sum_{i=1}^{n} a_i y_i \hat{\boldsymbol{x}}_i \right) &= \sum_{i=1}^{n} E[E(a_i y_i \hat{\boldsymbol{x}}_i | \boldsymbol{u}_i, \boldsymbol{z}_i)] \\ &= \sum_{i=1}^{n} E[E(a_i y_i | \boldsymbol{u}_i, \boldsymbol{z}_i) \hat{\boldsymbol{x}}_i] \\ &= \sum_{i=1}^{n} E[E(a_i | \boldsymbol{u}_i, \boldsymbol{z}_i) E(y_i | \boldsymbol{u}_i, \boldsymbol{z}_i) \hat{\boldsymbol{x}}_i] \\ &= \sum_{i=1}^{n} E[E(a_i | \boldsymbol{u}_i, \boldsymbol{z}_i) \hat{\boldsymbol{x}}_i \boldsymbol{x}_i' \theta] \end{aligned}$$

whereas

$$\begin{aligned} E\left(\sum_{i=1}^{n} a_i \hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i' \right) &= \sum_{i=1}^{n} E[E(a_i \hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i' | \boldsymbol{u}_i, \boldsymbol{z}_i)] \\ &= \sum_{i=1}^{n} E[E(a_i | \boldsymbol{u}_i, \boldsymbol{z}_i) \hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i']. \end{aligned}$$

In general, $E(a_i|\boldsymbol{u}_i, \boldsymbol{z}_i)$ is a non-linear function of $(\boldsymbol{u}_i, \boldsymbol{z}_i)$ and, thus,

$$E[E(a_i|\boldsymbol{u}_i, \boldsymbol{z}_i) \hat{\boldsymbol{x}}_i \boldsymbol{x}_i'] \neq E[E(a_i|\boldsymbol{u}_i, \boldsymbol{z}_i) \hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i'],$$

which implies that $\tilde{\theta}$ is asymptotically biased.

It is interesting to see under what kind of additional assumption $\tilde{\theta}$ is asymptotically unbiased. Let $\boldsymbol{u}_i^o$ denote the observed components of $\boldsymbol{u}_i$. Suppose that we replace assumption (4) by a stronger assumption

$$p(a|y,\boldsymbol{b},\boldsymbol{u},\boldsymbol{z}) = p(a|\boldsymbol{u}^o,\boldsymbol{z}), \tag{8}$$

i.e., missing y values are ignorable. Then,

$$\begin{aligned} E[E(a_i|\boldsymbol{u}_i,\boldsymbol{z}_i)\hat{\boldsymbol{x}}_i\boldsymbol{x}_i'] &= E[E\{E(a_i|\boldsymbol{u}_i^o,\boldsymbol{z}_i)\hat{\boldsymbol{x}}_i\boldsymbol{x}_i'|\boldsymbol{u}_i^o,\boldsymbol{z}_i\}] \\ &= E[E(a_i|\boldsymbol{u}_i^o,\boldsymbol{z}_i)E(\hat{\boldsymbol{x}}_i\boldsymbol{x}_i'|\boldsymbol{u}_i^o,\boldsymbol{z}_i)] \\ &= E[E(a_i|\boldsymbol{u}_i^o,\boldsymbol{z}_i)\hat{\boldsymbol{x}}_iE(\boldsymbol{x}_i'|\boldsymbol{u}_i^o,\boldsymbol{z}_i)] \\ &= E[E(a_i|\boldsymbol{u}_i^o,\boldsymbol{z}_i)\hat{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i'], \end{aligned}$$

where $\tilde{\boldsymbol{x}}_i$ is the same as $\hat{\boldsymbol{x}}_i$ except that $\hat{\gamma}$ is replaced by γ. Since $\hat{\gamma}$ is asymptotically unbiased and consistent,

$$E[E(a_i|\boldsymbol{u}_i^o,\boldsymbol{z}_i)\hat{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i'] = E[E(a_i|\boldsymbol{u}_i,\boldsymbol{z}_i)\hat{\boldsymbol{x}}_i\hat{\boldsymbol{x}}_i'] + o(1),$$

where $o(1)$ is a term converges to 0 as $n \to \infty$. Hence, $\tilde{\theta}$ is asymptotically unbiased and consistent.

Thus, we reach the following conclusions under assumptions (1)–(4).

(a) If covariate missing values are imputed using (2) and (6) and if we treat imputed covariate values as observed data in the estimation of θ under (1), then $\tilde{\theta}$ in (7) is asymptotically biased and inconsistent. Hence, using $\tilde{\theta}$ for any further estimation or inference leads to invalid results.

(b) If we replace the nonignorable missing y data assumption (4) by the ignorable missing y data assumption (8), then treating imputed covariate values as observed data provides an asymptotically unbiased and consistent estimator $\tilde{\theta}$. Clearly, this conclusion includes the special case where y has no missing value, the conclusion in Little (1992).

3.2 Estimation of μ

If we treat imputed covariate values as observed data, then we estimate the mean of y, $\mu = E(y)$, by the sample mean of the imputed y data:

$$\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n}[a_iy_i + (1-a_i)\tilde{\theta}'\hat{\boldsymbol{x}}_i]. \tag{9}$$

Obviously, we cannot expect $\tilde{\mu}$ to be asymptotically unbiased and consistent when $\tilde{\theta}$ is biased and inconsistent. But even if we replace $\tilde{\theta}$ in (9) by a consistent estimator of θ, $\tilde{\mu}$ is still biased because of the nonignorable missing y data. To illustrate this, we consider the special case where $\boldsymbol{u} = u$ is univariate. Further, since $\hat{\gamma}$ is consistent and we assume that $\tilde{\theta}$ is replaced by a consistent estimator, we can replace $\tilde{\theta}$ and $\hat{\gamma}$ in (9) by their true values θ and γ in the following discussion. Then, $\tilde{\mu}$ becomes

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}[a_iy_i + (1-a_i)\theta'\hat{\boldsymbol{x}}_i] \\ &= \frac{1}{n}\sum_{i=1}^{n}[a_iy_i + (1-a_i)\{\alpha + b_i\beta_uu_i \\ &\quad +(1-b_i)\beta_u\gamma'\boldsymbol{z}_i + \beta_z'\boldsymbol{z}_i\}] \\ &= \frac{1}{n}\sum_{i=1}^{n}[a_iy_i + (1-a_i)(\alpha + \beta_uu_i + \beta_z'\boldsymbol{z}_i) \\ &\quad +(1-a_i)(1-b_i)\beta_u(\gamma'\boldsymbol{z}_i - u_i)] \\ &= \frac{1}{n}\sum_{i=1}^{n}[a_iy_i + (1-a_i)\theta'\boldsymbol{x}_i \\ &\quad +(1-a_i)(1-b_i)\beta_u(\gamma'\boldsymbol{z}_i - u_i)] \end{aligned}$$

Under (1) and (4),

$$\begin{aligned} &E[a_iy_i + (1-a_i)\theta'\boldsymbol{x}_i] \\ &= E[E\{a_iy_i + (1-a_i)\theta'\boldsymbol{x}_i|u_i,\boldsymbol{z}_i\}] \\ &= E[E(a_i|u_i,\boldsymbol{z}_i)E(y_i|u_i,\boldsymbol{z}_i) + E(1-a_i|u_i,\boldsymbol{z}_i)\theta'\boldsymbol{x}_i] \\ &= E[E(a_i|u_i,\boldsymbol{z}_i)\theta'\boldsymbol{x}_i + E(1-a_i|u_i,\boldsymbol{z}_i)\theta'\boldsymbol{x}_i] \\ &= E(\theta'\boldsymbol{x}_i) \\ &= E(y_i). \end{aligned}$$

Under (2) and (3),

$$\begin{aligned} &E[(1-a_i)(1-b_i)(\gamma'\boldsymbol{z}_i - u_i)] \\ &= E[E\{(1-a_i)(1-b_i)(\gamma'\boldsymbol{z}_i - u_i)|u_i,\boldsymbol{z}_i\}] \\ &= E[E\{(1-a_i)(1-b_i)|u_i,\boldsymbol{z}_i\}(\gamma'\boldsymbol{z}_i - u_i)] \\ &= E[E(1-a_i|u_i,\boldsymbol{z}_i)E(1-b_i|\boldsymbol{z}_i)(\gamma'\boldsymbol{z}_i - u_i)] \end{aligned}$$

which is not equal to 0 since $E(1-a_i|u_i,\boldsymbol{z}_i)$ is typically nonlinear in u_i.

The previous argument also helps us to establish a result similar to that in Section 3.1. That is, if assumption (8) holds, then

$$\begin{aligned} &E\{(1-a_i)(1-b_i)|u_i,\boldsymbol{z}_i\} \\ &= E\{(1-a_i)(1-b_i)|u_i,\boldsymbol{z}_i,b_i=0\}P(b_i=0|u_i,\boldsymbol{z}_i) \\ &= E(1-a_i|u_i,\boldsymbol{z}_i,b_i=0)P(b_i=0|u_i,\boldsymbol{z}_i) \\ &= E(1-a_i|\boldsymbol{z}_i,b_i=0)P(b_i=0|\boldsymbol{z}_i) \\ &= E\{(1-a_i)(1-b_i)|\boldsymbol{z}_i,b_i=0\}P(b_i=0|\boldsymbol{z}_i) \\ &= E\{(1-a_i)(1-b_i)|\boldsymbol{z}_i\}, \end{aligned}$$

where the second equality follows from assumption (3) and the fact that, under assumption (8), a_i is independent of u_i when u_i in unobserved ($b_i = 0$). Consequently,

$$
\begin{aligned}
&E[(1-a_i)(1-b_i)(\gamma'\boldsymbol{z}_i-u_i)]\\
&=E[E\{(1-a_i)(1-b_i)|u_i,\boldsymbol{z}_i\}(\gamma'\boldsymbol{z}_i-u_i)]\\
&=E[E\{(1-a_i)(1-b_i)|\boldsymbol{z}_i\}(\gamma'\boldsymbol{z}_i-u_i)]\\
&=E\Big(E[E\{(1-a_i)(1-b_i)|\boldsymbol{z}_i\}(\gamma'\boldsymbol{z}_i-u_i)|\boldsymbol{z}_i]\Big)\\
&=E\Big(E\{(1-a_i)(1-b_i)|\boldsymbol{z}_i\}E[(\gamma'\boldsymbol{z}_i-u_i)|\boldsymbol{z}_i]\Big)\\
&=0
\end{aligned}
$$

and $\tilde{\mu}$ in (9) is asymptotically unbiased and consistent. This result can be extended to the general multivariate $\boldsymbol{u}$. Thus, we obtain the following conclusion.

(c) Conclusions (a)–(b) in Section 3.1 hold with $\tilde{\theta}$ in (7) replaced by $\tilde{\mu}$ in (9), when we consider the estimation of the mean of y instead of θ.

4. METHODS FOR NONIGNORABLE MISSING Y VALUES

The results in the previous section show that, if missing y data are ignorable, i.e., assumption (8) holds, then asymptotically unbiased and consistent estimators of θ and μ can be obtained by first imputing covariate missing values and then treating imputed covariate values as observed data. How reasonable is the ignorable missing data assumption (8)? When $\boldsymbol{u}_i$ has missing components, why does the missingness of y_i depends on observed but not the unobserved components of $\boldsymbol{u}_i$?

On the other hand, missingness of y depends on all covariate vectors, such as assumption (4), is much more reasonable than assumption (8). The trouble is, under (4), it is a challenge problem to derive valid parameter estimators because missingness is nonignorable (i.e., $\boldsymbol{u}$ has missing values).

The same problem does not occur for missing $\boldsymbol{u}$ data, because if we view $\boldsymbol{u}$ as a response and $\boldsymbol{z}$ as its covariate, then (3) is missingness dependent on covariate $\boldsymbol{z}$ that has no missing value. Of course, missingness of y becomes ignorable if we assume that it depends on covariate $\boldsymbol{z}$ but not on covariate $\boldsymbol{u}$. This is, however, somewhat a too strong assumption.

In this section, we derive some estimation methods valid under assumptions (1)–(4).

4.1 Estimation of θ and μ

To estimate θ in (1), the easiest way is to fit a regression between y and $\boldsymbol{x}$ based on subjects with observed y_i and completely observed $\boldsymbol{x}_i$. This leads to the following estimator:

$$
\hat{\theta}=\left(\sum_{i=1}^{n}a_ic_i\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\sum_{i=1}^{n}a_ic_iy_i\boldsymbol{x}_i, \tag{10}
$$

where $c_i=1$ if all components of $\boldsymbol{u}_i$ are observed and $c_i=0$ otherwise. This estimator is asymptotically unbiased and consistent, because

$$
\begin{aligned}
&E\left(\sum_{i=1}^{n}a_ic_iy_i\boldsymbol{x}_i\right)\\
&=\sum_{i=1}^{n}E[E(a_ic_iy_i\boldsymbol{x}_i|\boldsymbol{u}_i,\boldsymbol{z}_i)]\\
&=\sum_{i=1}^{n}E[E(a_i|\boldsymbol{u}_i,\boldsymbol{z}_i)E(c_iy_i|\boldsymbol{u}_i,\boldsymbol{z}_i)\boldsymbol{x}_i]\\
&=\sum_{i=1}^{n}E[E(a_i|\boldsymbol{u}_i,\boldsymbol{z}_i)E(c_i|\boldsymbol{u}_i,\boldsymbol{z}_i)E(y_i|\boldsymbol{u}_i,\boldsymbol{z}_i)\boldsymbol{x}_i]\\
&=\sum_{i=1}^{n}E[E(a_i|\boldsymbol{u}_i,\boldsymbol{z}_i)E(c_i|\boldsymbol{z}_i)\boldsymbol{x}_i\boldsymbol{x}_i']\theta
\end{aligned}
$$

and

$$
\begin{aligned}
&E\left(\sum_{i=1}^{n}a_ic_i\boldsymbol{x}_i\boldsymbol{x}_i'\right)\\
&=\sum_{i=1}^{n}E[E(a_ic_i\boldsymbol{x}_i\boldsymbol{x}_i'|\boldsymbol{u}_i,\boldsymbol{z}_i)]\\
&=\sum_{i=1}^{n}E[E(a_ic_i|\boldsymbol{u}_i,\boldsymbol{z}_i)\boldsymbol{x}_i\boldsymbol{x}_i']\\
&=\sum_{i=1}^{n}E[E(a_i|\boldsymbol{u}_i,\boldsymbol{z}_i)E(c_i|\boldsymbol{z}_i)\boldsymbol{x}_i\boldsymbol{x}_i'].
\end{aligned}
$$

However, the estimator $\hat{\theta}$ in (10) may not be efficient especially when the dimension of $\boldsymbol{u}$ is not small, since observed data in $\boldsymbol{u}_i$ are not used whenever $\boldsymbol{u}_i$ has at least one missing component ($c_i=0$). Because of the nonignorable missing y value assumption (4), it is difficult to improve $\hat{\theta}$, unless some more assumptions are added.

Once θ is estimated, we can estimate μ by

$$
\hat{\mu}=\hat{\theta}'\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{x}}_i\right), \tag{11}
$$

which is based on the facts that

$$
\mu=E(y)=E[E(y|\boldsymbol{u},\boldsymbol{z})]=E(\theta'\boldsymbol{x})=\theta'E(\boldsymbol{x})
$$

and, in the estimation of $E(\boldsymbol{x})$, we can treat imputed covariate values as observed data since (3) is an ignorable missing data assumption.

4.2 Imputation

Imputation results in a "complete" data set, which is often adopted for practical reasons. A basic requirement on the imputation method is that, after missing values are imputed, the sample means of $\boldsymbol{u}$ and y based on imputed data (treating imputed values as observed data) are valid (asymptotically unbiased and consistent) estimators of $E(\boldsymbol{u})$ and $\mu=E(y)$, respectively. Missing $\boldsymbol{u}$ values can be imputed as

discussed in Section 3, and the resulting sample mean of $\boldsymbol{u}$ is valid as discussed in the end of Section 4.1. How to impute missing y values, however, is a challenge problem because of the nonignorable missingness (4).

To illustrate, we first consider the case of univariate $\boldsymbol{u} = u$ with a univariate indicator b of whether u is observed. The set of all sampled subjects can be divided into four subsets according to the values of a_i and b_i:

$$A_{k,l} = \{i : a_i = k, b_i = l\}, \qquad k = 0, 1,\ l = 0, 1.$$

No imputation is needed in $A_{1,1}$. Imputation in $A_{0,1}$ is simple: a missing y_i with an observed u_i is imputed by $\hat{\theta}'\boldsymbol{x}_i$. It seems that there is no need for imputation in $A_{1,0}$ since y_i is observed, but it is not simple to impute missing y values in $A_{0,0}$, as we explain as follows. It is shown in Section 4.3 that

(i) $E(y|\boldsymbol{z}, a = 1, b = 1) = E(y|\boldsymbol{z}, a = 1, b = 0) = E(y|\boldsymbol{z}, a = 1)$;
(ii) $E(y|\boldsymbol{z}, a = 0, b = 1) = E(y|\boldsymbol{z}, a = 0, b = 0) = E(y|\boldsymbol{z}, a = 0)$;
(iii) $E(y|\boldsymbol{z}, a = 0) \neq E(y|\boldsymbol{z}, a = 1)$;
(iv) $E(y|\boldsymbol{z}, a = 0)$ is nonparametric in general, and is nonlinear in $\boldsymbol{z}$ even when a parametric model on the missing data mechanism is imposed.

Since missing y values in $A_{0,0}$ should be imputed by an estimated conditional expectation $E(y|\boldsymbol{z}, a = 0, b = 0) = E(y|\boldsymbol{z}, a = 0)$, it is very difficult to impute when the conditional expectation is nonparametric or nonlinear.

We next show that this difficulty can be overcome, i.e., valid imputation can be made without fitting nonparametric regression if we give up the observed y values in set $A_{1,0}$. We still first illustrate the idea in the case of univariate u and b.

After giving up observed y values in $A_{1,0}$, we merge $A_{1,0}$ into $A_{0,0}$. More importantly, this changes the missing y data mechanism, i.e., $b_i = 0$ now implies $a_i = 0$. Consequently,

$$E(y|\boldsymbol{z}, a = 0, b = 0) = E(y|\boldsymbol{z}, b = 0).$$

But (3) implies

$$E(y|\boldsymbol{z}, b = 0) = E(y|\boldsymbol{z}, b = 1) = E(y|\boldsymbol{z}),$$

which is linear in $\boldsymbol{z}$. Thus, we can impute each y_i in $A_{1,0} \cup A_{0,0}$ by an estimated $E(y_i|\boldsymbol{z}_i)$. By (1)–(2),

$$E(y_i|\boldsymbol{z}_i) = \alpha + (\beta_u'\Gamma + \beta_z')\boldsymbol{z}_i = \xi'\tilde{\boldsymbol{z}}_i,$$

where $\xi = (\alpha, \beta_u'\Gamma + \beta_z')'$ and $\tilde{\boldsymbol{z}}_i = (1, \boldsymbol{z}_i')'$. Since ξ is a function of θ and Γ, it can be estimated by $\hat{\xi}$ with θ and Γ replaced by $\hat{\theta}$ and $\hat{\Gamma}$, respectively, where each row of $\hat{\Gamma}$ is in the form of (6). Each y_i in $A_{1,0} \cup A_{0,0}$ is then imputed by $\hat{\xi}'\tilde{\boldsymbol{z}}_i$.

After imputation, the sample mean of imputed y data,

$$\sum_{i=1}^{n}[a_i b_i y_i + (1 - a_i)b_i\hat{\theta}'\boldsymbol{x}_i + (1 - b_i)\hat{\xi}'\tilde{\boldsymbol{z}}_i],$$

is asymptotically unbiased and consistent for μ. We would like to emphasize that one should not include the observed y data in $A_{1,0}$ when calculating the sample mean, i.e.,

$$\sum_{i=1}^{n}[a_i b_i y_i + (1-a_i)b_i\hat{\theta}'\boldsymbol{x}_i + a_i(1-b_i)y_i + (1-a_i)(1-b_i)\hat{\xi}'\tilde{\boldsymbol{z}}_i]$$

is an asymptotically biased estimator of μ, for the reasons (i)–(iv) previously given.

We now describe this imputation procedure for a general multivariate $\boldsymbol{u}$ with dimension $q \geq 1$. First, we divide the set of all sampled subjects into 2^q subsets according to the vector values of $\boldsymbol{b}$, i.e.,

$$B_{k_1,\ldots,k_q} = \{i : \boldsymbol{b}_i' = (k_1, \ldots, k_q)\}, \quad k_j = 0, 1,\ j = 1, \ldots, q.$$

Note that these B subsets are similar to the previously defined A subsets, but the indicator a for y is not involved in the B subsets.

Let B_c denote the subset corresponding to $\boldsymbol{b}' = (1, \ldots, 1)$, i.e., B_c contains subjects with completed $\boldsymbol{u}_i$'s. A missing y_i in B_c can be imputed by $\hat{\theta}'\boldsymbol{x}_i$, as discussed previously. In any subset B that is not B_c, we give up observed y values in B. For a fixed B, we denote the corresponding $\boldsymbol{b}$ vector as $\boldsymbol{b}_{_B}$, the sub-vector containing observed components of $\boldsymbol{u}$ as $\boldsymbol{u}_o$, and the sub-vector containing missing components of $\boldsymbol{u}$ as $\boldsymbol{u}_m$. Note that $\boldsymbol{u}_o$ contains components of $\boldsymbol{u}$ corresponding to components of $\boldsymbol{b}_{_B}$ that are equal to 1 and $\boldsymbol{u}_m$ contains components of $\boldsymbol{u}$ corresponding to components of $\boldsymbol{b}_{_B}$ that are equal 0. In subset B, we may impute each y value by an estimated $E(y|\boldsymbol{u}_o, \boldsymbol{z}, \boldsymbol{b} = \boldsymbol{b}_{_B})$. We now show that this conditional expectation is equal to $E(y|\boldsymbol{u}_o, \boldsymbol{z}, \boldsymbol{b} = \mathbf{1})$, where $\mathbf{1}$ is the vector whose components are all equal to 1. Under (3),

$$\begin{aligned} p(\boldsymbol{u}_m|\boldsymbol{u}_o, \boldsymbol{z}, \boldsymbol{b}) &= \frac{p(\boldsymbol{u}_m, \boldsymbol{b}|\boldsymbol{u}_o, \boldsymbol{z})}{p(\boldsymbol{b}|\boldsymbol{u}_o, \boldsymbol{z})} \\ &= \frac{p(\boldsymbol{b}|\boldsymbol{u}_m, \boldsymbol{u}_o, \boldsymbol{z})p(\boldsymbol{u}_m|\boldsymbol{u}_o, \boldsymbol{z})}{p(\boldsymbol{b}|\boldsymbol{u}_o, \boldsymbol{z})} \\ &= p(\boldsymbol{u}_m|\boldsymbol{u}_o, \boldsymbol{z}). \end{aligned}$$

Consequently,

$$\begin{aligned} E(y|\boldsymbol{u}_o, \boldsymbol{z}, \boldsymbol{b}) &= \int E(y|\boldsymbol{u}_m, \boldsymbol{u}_o, \boldsymbol{z}, \boldsymbol{b})p(\boldsymbol{u}_m|\boldsymbol{u}_o, \boldsymbol{z}, \boldsymbol{b})d\boldsymbol{u}_m \\ &= \int E(y|\boldsymbol{u}_m, \boldsymbol{u}_o, \boldsymbol{z})p(\boldsymbol{u}_m|\boldsymbol{u}_o, \boldsymbol{z})d\boldsymbol{u}_m \\ &= E(y|\boldsymbol{u}_o, \boldsymbol{z}). \end{aligned}$$

Therefore,

$$E(y|\boldsymbol{u}_o, \boldsymbol{z}, \boldsymbol{b} = \boldsymbol{b}_{_B}) = E(y|\boldsymbol{u}_o, \boldsymbol{z}) = E(y|\boldsymbol{u}_o, \boldsymbol{z}, \boldsymbol{b} = \mathbf{1}).$$

If we impute each y in B using this conditional expectation, then we need more assumptions. From (1),

$$\begin{aligned} E(y|\boldsymbol{u}_o, \boldsymbol{z}) &= \alpha + \beta_u' E(\boldsymbol{u}|\boldsymbol{u}_o, \boldsymbol{z}) + \beta_z' \boldsymbol{z} \\ &= \alpha + \beta_{uo}' \boldsymbol{u}_o + \beta_{um}' E(\boldsymbol{u}_m|\boldsymbol{u}_o, \boldsymbol{z}) + \beta_z' \boldsymbol{z}, \end{aligned}$$

where β_{uo} and β_{um} are the sub-vectors of β_u corresponding to $\boldsymbol{u}_o$ and $\boldsymbol{u}_m$, respectively. Thus, an assumption on $E(\boldsymbol{u}_m|\boldsymbol{u}_o, \boldsymbol{z})$ is needed. For example, we may assume that $E(\boldsymbol{u}_m|\boldsymbol{u}_o, \boldsymbol{z})$ is linear in $\boldsymbol{u}_o$ and $\boldsymbol{z}$, which is true when $p(\boldsymbol{u}|\boldsymbol{z})$ is multivariate normal. That is,

$$E(\boldsymbol{u}_m|\boldsymbol{u}_o, \boldsymbol{z}) = \alpha_{_B} + \Psi_{_B} \boldsymbol{u}_o + \Gamma_{_B} \boldsymbol{z}, \tag{12}$$

where $\alpha_{_B}$ is an unknown vector and $\Psi_{_B}$ and $\Gamma_{_B}$ are unknown matrices, which are different in different B. Then, we may fit a linear regression based on model (12) using $\boldsymbol{u}$ and $\boldsymbol{z}$ data in the subset B_c. After we obtain estimators $\hat{\alpha}_{_B}$, $\hat{\Psi}_{_B}$, $\hat{\Gamma}_{_B}$ and estimators $\hat{\alpha}$, $\hat{\beta}_u$, $\hat{\beta}_z$ based on $\hat{\theta}$, each y_i in B is imputed by

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_{uo}' \boldsymbol{u}_{oi} + \hat{\beta}_{um}' (\hat{\alpha}_{_B} + \hat{\Psi}_{_B} \boldsymbol{u}_{oi} + \hat{\Gamma}_{_B} \boldsymbol{z}_i) + \hat{\beta}_z' \boldsymbol{z}_i.$$

Once imputation is completed, the mean μ can be estimated by the sample mean of y_i's with imputed y values treated as observed data, i.e.,

$$\sum_{i \in B_c} [a_i y_i + (1 - a_i) \hat{\theta}' \boldsymbol{x}_i] + \sum_B \sum_{i \in B} \hat{y}_i, \tag{13}$$

where $\sum_B$ is the sum over all possible subsets B's that are not B_c. This estimator is asymptotically unbiased and consistent. Again, one should not use any observed y_i to replace imputed value $\hat{y}_i$ in $B \neq B_c$, i.e., the biased estimator

$$\sum_{i \in B_c} [a_i y_i + (1 - a_i) \hat{\theta}' \boldsymbol{x}_i] + \sum_B \sum_{i \in B} [a_i y_i + (1 - a_i) \hat{y}_i]. \tag{14}$$

An empirical confirmation is given in Section 5, where it is shown by simulation that the sample mean in (13) is almost unbiased but the sample mean in (14) is seriously biased.

4.3 Proof of (i)–(iv) in Section 4.2

Under (3)–(4), result (5) implies that

$$p(y|\boldsymbol{z}, a, \boldsymbol{b}) = \frac{p(\boldsymbol{b}|y, \boldsymbol{z}, a) p(y|\boldsymbol{z}, a)}{p(\boldsymbol{b}|\boldsymbol{z}, a)} = p(y|\boldsymbol{z}, a). \tag{15}$$

These results hold for general multivariate $\boldsymbol{u}$ and $\boldsymbol{b}$ so that they also hold in the special case of univariate u and b. In particular, (i)–(ii) in Section 4.2 follow directly from (15).

To show (iii), we note that

$$\begin{aligned} E(y|\boldsymbol{z}, a = 0) &= \int E(y|u, \boldsymbol{z}, a = 0) p(u|\boldsymbol{z}, a = 0) du \\ &= \int E(y|u, \boldsymbol{z}) p(u|\boldsymbol{z}, a = 0) du \\ &= \int (\alpha + \beta_u u + \beta_z' \boldsymbol{z}) p(u|\boldsymbol{z}, a = 0) du \\ &= \alpha + \beta_z' \boldsymbol{z} + \beta_u \int u p(u|\boldsymbol{z}, a = 0) du \\ &= \alpha + \beta_z' \boldsymbol{z} + \frac{\beta_u \int u P(a = 0|u, \boldsymbol{z}) p(u|\boldsymbol{z}) du}{\int P(a = 0|u, \boldsymbol{z}) p(u|\boldsymbol{z}) du} \end{aligned}$$

Let

$$g(\boldsymbol{z}) = \int (u - \gamma' \boldsymbol{z}) P(a = 0|u, \boldsymbol{z}) p(u|\boldsymbol{z}) du$$

and

$$h(\boldsymbol{z}) = \int P(a = 0|u, \boldsymbol{z}) p(u|\boldsymbol{z}) du.$$

Then

$$E(y|\boldsymbol{z}, a = 0) = \alpha + (\beta_z' + \beta_u \gamma') \boldsymbol{z} + \beta_u \frac{g(\boldsymbol{z})}{h(\boldsymbol{z})}.$$

Similarly,

$$E(y|\boldsymbol{z}, a = 1) = \alpha + (\beta_z' + \beta_u \gamma') \boldsymbol{z} - \beta_u \frac{g(\boldsymbol{z})}{1 - h(\boldsymbol{z})}.$$

Therefore, $E(y|\boldsymbol{z}, a = 0) = E(y|\boldsymbol{z}, a = 1)$ if and only if $\beta_u g(\boldsymbol{z}) = 0$.

We now give a counter example in which $g(\boldsymbol{z}) \neq 0$, which is sufficient for showing $E(y|\boldsymbol{z}, a = 0) \neq E(y|\boldsymbol{z}, a = 1)$ in general. Assume that conditional on $\boldsymbol{z}$, $u \sim N(\gamma' \boldsymbol{z}, \sigma_u^2)$ and $P(a = 0|u, \boldsymbol{z}) = \Phi(\kappa u)$, where Φ is the standard normal distribution function and κ and σ_u^2 are some constants that are not 0. Using integration by parts, we obtain that

$$\begin{aligned} g(\boldsymbol{z}) &= \frac{1}{\sqrt{2\pi}\sigma_u} \int (u - \gamma' \boldsymbol{z}) \Phi(\kappa u) \exp\left\{ -\frac{(u - \gamma' \boldsymbol{z})^2}{2\sigma_U^2} \right\} du \\ &= \frac{\kappa \sigma_U^2}{2\pi \sigma_u} \int \exp\left\{ -\frac{\kappa^2 u^2}{2} \right\} \exp\left\{ -\frac{(u - \gamma' \boldsymbol{z})^2}{2\sigma_U^2} \right\} du \\ &= \frac{\kappa \sigma_U^2}{\sqrt{2\pi(1 + \kappa^2 \sigma_u^2)}} \exp\left\{ -\frac{(\kappa \gamma' \boldsymbol{z})^2}{2(1 + \kappa^2 \sigma_u^2)} \right\}. \end{aligned}$$

Clearly, $g(\boldsymbol{z}) \neq 0$.

It is clear from the form of $g(\boldsymbol{z})/h(\boldsymbol{z})$ that it is is nonparametric when either $P(a = 0|u, \boldsymbol{z})$ or $p(u|\boldsymbol{z})$ is nonparametric. Thus $E(y|\boldsymbol{z}, a = 0)$ is nonparametric when either $P(a = 0|u, \boldsymbol{z})$ or $p(u|\boldsymbol{z})$ is nonparametric. When both $P(a = 0|u, \boldsymbol{z})$ and $p(u|\boldsymbol{z})$ are parametric, the previous counter example shows that $g(\boldsymbol{z})/h(\boldsymbol{z})$ is nonlinear in $\boldsymbol{z}$ unless $\gamma = 0$, because

$$h(\boldsymbol{z}) = \Phi\left(\kappa \gamma' \boldsymbol{z} / \sqrt{1 + \kappa^2 \sigma_u^2} \right).$$

This shows (iv) in Section 4.2.

5. SIMULATION RESULTS

A simulation study was conducted to check finite sample performance of estimators of θ and μ discussed in the previous sections, under (1)–(4). These simulation results also provide an empirical confirmation for the results derived in the previous sections.

5.1 Simulation setting

The simulation setting is described as follows.

1. The covariate $\boldsymbol{z} = z$ is univariate and normally distributed with mean 2 and variance 1. All values of z are observed.
2. The covariate $\boldsymbol{u} = (u_1, u_2)'$ is 2-dimensional and, conditional on z, $\boldsymbol{u}$ has the bivariate normal distribution with $E(u_1|z) = z$, $E(u_2|z) = 0.5z$, and covariance matrix
$$\begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}.$$
3. Let $\boldsymbol{b} = (b_1, b_2)'$ be the vector of indicators of whether two components u_1 and u_2 are observed. Conditional on $\boldsymbol{u}$ and z, b_1 and b_2 are independent and
$$P(b_1 = 1|\boldsymbol{u}, z) = P(b_2 = 1|\boldsymbol{u}, z) = \frac{1}{1 + \exp(1 - 0.7z)}.$$
4. Conditional on $\boldsymbol{u}$ and z, y is normally distributed with mean
$$E(y|\boldsymbol{u}, z) = \alpha + \beta_{u1}u_1 + \beta_{u2}u_2 + \beta_z z$$
and variance 1. The true values of the parameters in the simulation are $\alpha = 1$, $\beta_{u1} = 4$, $\beta_{u2} = 6$, and $\beta_z = 3$.
5. Conditional on y, $\boldsymbol{u}$, z, and $\boldsymbol{b}$, the indicator a of whether y is observed follows
$$P(a = 1|y, \boldsymbol{u}, z, \boldsymbol{b}) = \frac{1}{1 + \exp(2 + u_1 - 3u_2 - z)}.$$

The missing data mechanisms were chosen so that the unconditional probabilities for b_1, b_2, and a were given in the following table.

b_1	b_2	a	Probability
0	0	0	0.10
1	0	0	0.09
0	1	0	0.09
1	1	0	0.12
0	0	1	0.10
1	0	1	0.13
0	1	1	0.13
1	1	1	0.24

5.2 Estimation of θ

The following three methods for estimating $\theta = (\alpha, \beta_{u1}, \beta_{u2}, \beta_z)'$ were considered in the simulation.

Method I. The estimator of θ based on model (1) and data without any missing value, which was used as a standard.
Method II. The estimator $\tilde{\theta}$ defined in (7).
Method III. The estimator $\hat{\theta}$ defined in (10).

The sample size considered was $n = 1,000$. Based on a simulation of 1,000 runs, the following table gives the simulation bias and standard error (SD) for estimators of θ based on the three methods. For comparison, true values of parameters to be estimated are included in the table.

Parameter	Method	Bias	SD
$\alpha = 1$	I	−0.003	0.002
	II	3.090	0.019
	III	0.003	0.007
$\beta_{u1} = 4$	I	−0.001	0.001
	II	0.578	0.006
	III	−0.002	0.002
$\beta_{u2} = 6$	I	0.000	0.001
	II	−0.425	0.007
	III	0.004	0.003
$\beta_z = 3$	I	0.003	0.001
	II	−1.194	0.008
	III	−0.000	0.003

Obviously, $\tilde{\theta}$ defined in (7) is seriously biased, which confirms our theoretical finding in Section 3.1, i.e., treating imputed covariate values as observed data in the estimation of θ leads to a bias. The estimator $\hat{\theta}$ in (10), which is shown to be asymptotically unbiased in Section 4.1, has a negligible bias although it is less efficient than the estimator based on full data.

5.3 Estimation of μ

The following six estimators of $\mu = E(y)$ (with true value 21) were considered.

Method I. The sample mean of y based on data without any missing value, which was used as a standard.
Method II. The estimator $\tilde{\mu}$ defined in (9).
Method IIa. The estimator defined in (9) but with $\tilde{\theta}$ in (7) replaced by $\hat{\theta}$ in (10).
Method III. The estimator $\hat{\mu}$ defined in (11).
Method IIIa. The sample mean of imputed y data given by formula (13) with imputation as described in Section 4.2.
Method IIIb. The sample mean given by formula (14).

Based on the same 1,000 simulation runs as in Section 5.2, the following table gives the simulation bias and standard error (SD) for estimators of μ based on the six methods.

Clearly, treating imputed covariate values as observed data (methods II and IIa) provides biased estimators of μ,

Parameter	Method	Bias	SD
$\mu = 21$	I	−0.015	0.013
	II	1.433	0.013
	IIa	0.653	0.013
	III	−0.003	0.014
	IIIa	−0.004	0.014
	IIIb	3.081	0.015

which confirms our theory in Section 3.2. Note that this is still true even when we use a valid estimator $\hat{\theta}$ to estimate θ. Although method IIa provides less biased estimation than method II, the bias for method IIa is still significant.

On the other hand, methods III and IIIa produce nearly unbiased estimators of μ, and they are comparable. In this case, the estimators by methods III and IIIa have almost the same efficiency as the estimator based on no missing data, probably because, although there are many missing y values, most information is recovered through imputation using covariates.

Finally, method IIIb provides a biased estimator with a bias even larger than those for method II. This confirms our discussion in Section 4.2, i.e., the imputation procedure is valid when observed y_i's with missing $\boldsymbol{u}$ covariate values are discarded so that a bias is created if we include these observed y_i's in the computation of the sample mean. It is possible to not discard any observed y_i's in the estimation of μ, but a much more complicated imputation procedure (which is based on nonparametric or nonlinear regression) has to be developed.

6. DISCUSSION

In the previous sections we focus on point estimation of θ or μ only. To assess statistical accuracy or make inference, we also need estimators of the variability of the asymptotically unbiased point estimators. In general, there are two ways to obtain variance estimators. The first one is based on theoretical derivation. We first establish the asymptotic normality of the asymptotically unbiased point estimators and then obtain variance estimators by substituting unknown quantities in the derived asymptotic variances with consistent estimators. The second method is to apply the bootstrap or other resampling methods by adding a re-imputation or adjustment step when point estimators are computed using imputed data. Details can be found in Shao and Sitter (1996), Shao (2001), or Kim and Shao (2013).

It can be shown that, for the estimation of μ, the estimator $\hat{\mu}$ in (11) is asymptotically more efficient than the estimator in (13), provided that all models are correct. In our simulation study, however, these two estimators are almost the same.

Although we focus on a univariate y, extensions to multivariate y can be made. If we want to estimate correlations among different y components, however, the imputation has to be carefully done in the presence of nonignorable missing values. Also, we may add random noises to imputed values for the purpose of estimating quantiles. Further research will be carried out on these topics.

ACKNOWLEDGMENTS

The author's research is partially supported by the NSF Grant DMS-1007454. The author would like to thank a referee for useful comments and Mr. Quefeng Li who provided help in the simulation study.

Received 3 February 2013

REFERENCES

Kim, J. K. and Shao, J. (2013). *Statistical Methods for Incomplete Data Analysis*, Chapman & Hall/CRC, New York.

Little, R. J. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, **87**, 1227–1237.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, New York. MR1925014

Shao, J. (2001). Replication methods for variance estimation in complex surveys with imputed data. *Survey Nonresponse*, (edited by R. Groves, D. Dillman, J. Eltinge, and R. Little), 303–314, Wiley and Sons, New York.

Shao, J. and Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, **91**, 1278–1288. MR1424624

Jun Shao
School of Finance and Statistics
East China Normal University
Shanghai 200241
China

Department of Statistics
University of Wisconsin
Madison WI 53706
USA
E-mail address: shao@stat.wisc.edu