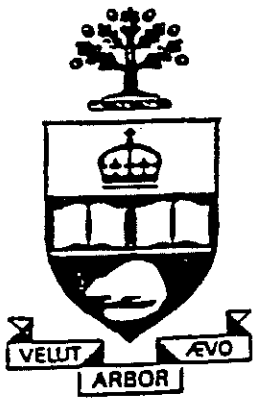

Resampling Methods for Imputing Missing Observations in Regression Models

by

M.S. Srivastava
Department of Statistics
University of Toronto

Technical Report No. 9707, March 12, (1997)

TECHNICAL REPORT SERIES

University of Toronto

Department of Statistics

Resampling Methods for Imputing Missing Observations in Regression Models

by M. S. Srivastava

Department of Statistics University of Toronto, Toronto, Ontario, Canada
M5S 3G3

email: srivasta@utstat.toronto.edu

This paper was supported by Natural Sciences and Engineering Research Council of Canada

1991 AMS Subject Classification: 62JOS, 62DOS, 62E20

Key Words and Phrases: Bootstrapping, Jackknifing, Regression Model, Missing Observations, Single and Multiple Imputation

Abstract

The problem of imputing missing observations in a regression model is considered. It is assumed that the observations are missing at random and all the observations on the auxiliary or independent variables are available. Estimates based on singly and multiply imputed values are given. Jackknife as well as bootstrap estimates of the variance of the estimator of the regression parameters are given. For a singly imputed estimator of the regression parameters, a non-parametric bootstrap distribution is given. For multiply imputed estimators, asymptotic distributions are given to obtain confidence intervals. It is shown that only the records of the respondents are needed to obtain the estimates of the parameters and the variance and distribution of these estimators.

1 Introduction

In most surveys with nonresponse data, it is a common practice to replace the missing values by some kind of imputed values. For random nonresponse, the most commonly used method to impute the missing values is to take a simple random sample with replacement from the observed data. The data is then completed with these imputed values and estimates of the population parameters are obtained from the completed data set treating the imputed values as the true values. However, when the variance of these estimators are estimated from the completed data set by the usual formula, it often gives an underestimate as it fails to reflect the extra variability due to imputation. Rao and Shao (1992) provided an adjusted jackknife estimator of the variance overcoming this shortcoming.

Rubin (1978) has proposed multiple imputation since it provides a distribution of the missing values. Rubin and Schenker (1986) have shown that multiple imputing provides a better coverage for the population mean than the singly imputed method. The main objection raised against multiply imputing is that it is expensive and difficult to maintain the records of so many completed data sets. Fay (1993) has argued against it on some technical grounds.

In a recent report, Srivastava (1997) has proposed multiply imputation methods which require only the record of the respondents. In the same paper, for a singly imputed data, Jackknife estimates of the variance are also given which do not require any adjustment. In addition bootstrap estimators of the variance as well as non-parametric bootstrap distribution are also given to obtain confidence intervals for the mean. However, the above study considered only simple random sampling. In this paper, we extend these results to the regression model where all the observations on the auxiliary or independent variables are available. Following Srivastava and Carter (1986), it is shown that only the records of the respondents are needed for all the results given in this paper. The organization of the paper is as follows.

In Section 2, an estimate of the population mean is given for a simgly imputed data. A Jackknife estimate of the variance of this estimator is given in Section 3 and a bootstrap estimate is given in Section 4. The nonparametric bootstrap distribution is given in Section 5 which can be used to obtain confidence intervals for the population mean. Multiple imputation methods are given in Sections 6 and 7. The paper concludes in Section 8.

2 Estimation Based On Single Imputation

Consider the following regression model

$$E(y_i|x_i) = \boldsymbol{\beta}'\mathbf{x}_i,\ \text{var}(y_i|x_i) = \sigma^2,\ \text{and}\ \ \text{cov}(y_i y_j) = 0,\ i \neq j \tag{1}$$

$i, j = 1, \ldots, N$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ is the ith observation on the p auxiliary variables, $x_1, \ldots, x_p$ and $\boldsymbol{\beta}$ is an unknown p-vector of regression parameters. We shall assume that N is large and thus no finite population correction is needed in the following ananlysis. We take a simple random sample of size n in which n_1 subjects respond and n_0 subjects do not respond on the item y. However, we have all the observations on the auxiliary variables $x_1, \ldots, x_p$. Thus we shall have

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & n_{np} \end{pmatrix} = \begin{pmatrix} X_1 \\ X_0 \end{pmatrix} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)' \tag{2}$$

where X_1 is an $n_1 \times p$ matrix that corresponds to the observations on the auxiliary variables for the respondents $y_1, \ldots, y_{n_1}$, and X_0 is the $n_0 \times p$ matrix of observations for the nonrespondents. We shall assume that $n_i > p,\ i = 0, 1$, and for simplicity of presentation we shall also assume that $x_{11} = \cdots = x_{n1} = 1$, that is, there is a constant term in the model. Let $\mathbf{y}_1 = (y_1, \ldots, y_{n_1})'$ denote the vector of respondents. The least squares estimate of $\boldsymbol{\beta}$ based on respondents only will be denoted by $\mathbf{b}_1$ and is given by

$$\mathbf{b}_1 = (X_1'X_1)^{-1}X_1'\mathbf{y}_1. \tag{3}$$

Let

$$\boldsymbol{\epsilon} = \left(\frac{n_1}{n_1 - p}\right)^{\frac{1}{2}} (\mathbf{y}_1 - X_1\mathbf{b}_1) = (\epsilon_1, \ldots, \epsilon_{n_1}).' \tag{4}$$

We take a random sample of size n_0 with replacement from $\epsilon_1, \ldots, \epsilon_{n_1}$. We shall denote this random sample by $\epsilon_1^*, \ldots, \epsilon_{n_0}^*$, and define the n_0 imputed values of y's by

$$\mathbf{y}_0^* = X_0\mathbf{b}_1 + \boldsymbol{\epsilon}_0^* \tag{5}$$

where

$$\boldsymbol{\epsilon}_0^* = (\epsilon_1^*, \ldots, \epsilon_{n_0}^*)'. \tag{6}$$

We define the imputed estimate of $\boldsymbol{\beta}$ by

$$\mathbf{b}_I^* = (X'X)^{-1}X' \begin{pmatrix} \hat{\mathbf{y}}_1 \\ \mathbf{y}_0^* \end{pmatrix} = \mathbf{b}_1 + (X'X)^{-1}X_0'\boldsymbol{\epsilon}_0^*, \tag{7}$$

where $\hat{\mathbf{y}}_1 = X_1\mathbf{b}_1$. Since

$$E_*(\epsilon_i^*) = 0,\ \text{and}\ E(\epsilon_i^{*2}) = S_1^2 = \frac{1}{n_1 - p}\sum_{i=1}^{n_1}(y_i - \mathbf{x}_i'\mathbf{b}_1)^2, \tag{8}$$

we find that

$$E_*(\mathbf{b}_i^*) = \mathbf{b}_1,$$

4 Bootstrap Estimate of the Covariance

We define ϵ_i as in Section 2, eq (4), namely

$$\epsilon_i = \left(\frac{n_1}{n_1 - p}\right)^{\frac{1}{2}} (y_i - \mathbf{x}_i'\mathbf{b}_1),\ i = 1, \ldots, n_1.$$

We draw a sample of size n with replacement from $\epsilon_1, \ldots, \epsilon_{n_1}$. And we replicate it m times. We shall denote the first n_1 observations of the ith replicate by ϵ^*_{ij}, $j = 1, \ldots, n_1$, $i = 1, \ldots, m$, and the remaining $n_0 = n - n_1$ observations by ϵ^{**}_{ij}, $j = 1, \ldots, n_0$, $i = 1, \ldots, m$. Define

$$\mathbf{y}^*_{1i} = X_1\mathbf{b}_1 + \boldsymbol{\epsilon}^*_{1i},\ i = 1, \ldots, m$$

as the n_1 vector of imputed values for the observed values $y_1, \ldots, y_{n_1}$, where

$$\mathbf{y}^*_{1i} = (y^*_{1i}, \ldots, y^*_{n_1 i})' \text{ and } \boldsymbol{\epsilon}^*_{1i} = (\epsilon^*_{1i}, \ldots, \epsilon^*_{n_1 i}).'$$

Then the imputed estimate of $\mathbf{b}_1$ for the ith replicate is given by

$$\begin{aligned} \mathbf{b}^*_{1i} &= (X_1'X_1)^{-1}X_1'\mathbf{y}^*_{1i} \\ &= \mathbf{b}_1 + (X_1'X_1)^{-1}X_1'\boldsymbol{\epsilon}^*_{1i},\ \ i = 1, \ldots, m. \end{aligned} \tag{14}$$

Hence, the predicted value of $\mathbf{y}^*_{1i}$ is given by

$$\begin{aligned} \hat{\mathbf{y}}^*_{1i} &= X_1\mathbf{b}^*_{1i} \\ &= H_1\mathbf{y}^*_{1i} \\ &= X_1\mathbf{b}_1 + H_1\boldsymbol{\epsilon}^*_{1i},\ i = 1, \ldots, m, \end{aligned}$$

where

$$H_1 = X_1(X_1'X_1)^{-1}X_1' \equiv X_1A_1,\ \ A_1 = (X_1'X_1)^{-1}X_1'.$$

Next, we impute the missing observations in the ith replicate by

$$\begin{aligned} \mathbf{y}^{**}_{0i} &= X_0\mathbf{b}^*_{1i} + \boldsymbol{\epsilon}^{**}_{0i} \\ &= X_0\mathbf{b}_1 + X_0A_1\boldsymbol{\epsilon}^*_{1i} + \boldsymbol{\epsilon}^{**}_{0i} \end{aligned} \tag{15}$$

where

$$\mathbf{y}^{**}_{0i} = (y^{**}_{1i}, \ldots, y^{**}_{n_0 i})' \text{ and } \boldsymbol{\epsilon}^{**}_{0i} = (\epsilon^{**}_{1i}, \ldots, \epsilon^{**}_{n_0 i})',$$

$i = 1, \ldots, m$. We define the imputed estimate of $\boldsymbol{\beta}$ in the ith replicate by

$$\begin{aligned} \mathbf{b}^{**}_{Ii} &= (X'X)^{-1}X'\begin{pmatrix} \hat{\mathbf{y}}^*_{1i} \\ \mathbf{y}^{**}_{0i} \end{pmatrix} \\ &= \mathbf{b}_1 + A_1\boldsymbol{\epsilon}^*_{1i} + (X'X)^{-1}X_0'\boldsymbol{\epsilon}^{**}_{0i}. \end{aligned} \tag{16}$$

Let

$$\mathbf{b}^{**}_{I\cdot} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{b}_{Ii}.$$

Then the bootstrap estimate of the cov($\mathbf{b}_I^*$) is given by

$$V_B^* = \frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{b}_{Ii} - \mathbf{b}_{I.})(\mathbf{b}_{Ii} - \mathbf{b}_{I.})'. \tag{17}$$

In terms of ϵ_{ij}^* and ϵ_{ij}^{**}, this is given by

$$\begin{aligned} V_B^* &= (X_1'X_1)^{-1}X_1'[\frac{1}{m-1}\sum_{i=1}^{m}(\epsilon_{1i}^* - \overline{\epsilon}_{1.}^*)(\epsilon_{1i}^* - \overline{\epsilon}_{1.}^*)']X_1(X_1'X_1)^{-1} \\ &\quad +(X'X)^{-1}X_0'[\frac{1}{m-1}\sum_{i=1}^{m}(\epsilon_{0i}^* - \overline{\epsilon}_{0.}^*)(\epsilon_{0i}^* - \overline{\epsilon}_{0.}^*)]X_0(X'X)^{-1}, \end{aligned}$$

where $\overline{\epsilon}_{1.}^* = n_1^{-1}\sum_{i=1}^{n_1}\epsilon_{1i}^*$ and $\overline{\epsilon}_{0.}^{**} = n_0^{-1}\sum_{i=1}^{n_0}\epsilon_{0i}^*$. It can be shown that

$$E_*(V_B^*) = [(X_1'X_1)^{-1} + (X'X)^{-1}X_0'X_0(X'X)^{-1}]S_1^2.$$

Hence the bootstrap estimate of the cov($\mathbf{b}_I^*$) is an unbiased estimator.

5 Bootstrap Distribution

In order to obtain tests and confidence intervals for the regression parameters $\boldsymbol{\beta}$ using the imputed estimator $\mathbf{b}_I^*$ in Section 2, we need its distribution. In this section, we propose a nonparametric bootstrap distribution for $\mathbf{b}_I^*$. In order to obtain this distribution we draw a random sample of size n with replacement from $\epsilon_1, \ldots, \epsilon_{n_1}$ defined in Section 2. We shall replicate it m times. The first n_1 observations in the ith replicate will be denoted by ϵ_{ij}^*, $j = 1, \ldots, n_1$, $i = 1, \ldots, m$ and the last $n_0 = n - n_1$ observations in the ith replicate will be denoted by ϵ_{ij}^{**}, $j = 1, \ldots, n_0$, $i = 1, \ldots, m$. Define $\mathbf{b}_{Ii}^{**}$ as

$$\mathbf{b}_{Ii}^{**} = \mathbf{b}_I^* + A_1\epsilon_{1i}^* + (X'X)^{-1}X_0'\epsilon_{0i}^{**}.$$

Let

$$\begin{aligned} h_{1ii} &= \mathbf{x}_i'(X_1'X_1)^{-1}\mathbf{x}_i, \quad i = 1, \ldots, n_1, \ n_1 > p, \\ h_{0ii} &= \mathbf{x}_i'(X_0'X_0)^{-1}\mathbf{x}_i, \quad i = 1, \ldots, n_0, \ n_0 > p, \\ (n-2)S_i^{**2} &= \sum_{j=1}^{n_1}(\epsilon_{ij}^* - \overline{\epsilon}_{i.}^*)^2 + \sum_{j=1}^{n_0}(\epsilon_{ij}^{**} - \overline{\epsilon}_{i.}^{**})^2, \end{aligned}$$

where

$$\overline{\epsilon}_{i.}^* = n_1^{-1}\sum_{j=1}^{n_1}\epsilon_{ij}^* \text{ and } \overline{\epsilon}_{i.}^{**} = n_0^{-1}\sum_{j=1}^{n_1}\epsilon_{ij}^{**}$$

We shall assume that

$$\max_{1\le i\le n_1} h_{1ii} \to 0 \text{ and } \max_{1\le i\le n_0} h_{0ii} \to 0.$$

Then from Srivastava (1971) it follows that given $y_1, \ldots, y_{n_1}$,

$$A_1\epsilon_{1i}^* \to N_p(\mathbf{0}, S_1^2(X_1'X_1)^{-1}), \quad A_1 = (X_1'X_1)^{-1}X_1', \quad i = 1, \ldots, m$$

and

$$(X'X)^{-1}X_0'\epsilon_{0i}^{**} \rightarrow N_p(\mathbf{0}, S_1^2(X'X)^{-1}X_0'X_0(X'X)^{-1}),\ i = 1, \ldots, m.$$

Hence

$$S_i^{**-1}[A_1\epsilon_{1i}^* + (X'X)^{-1}X_0'\epsilon_{0i}^{**}] \rightarrow N_p(\mathbf{0}, Q),\ \ i = 1, \ldots, m$$

where

$$Q = (X_1'X_1)^{-1} + (X'X)^{-1}X_0'X_0(X'X)^{-1}.$$

Thus

$$\frac{Q^{-\frac{1}{2}}[\mathbf{b}_{Ii}^{**} - \mathbf{b}_I^*]}{S_i^{**}} \rightarrow N_p(\mathbf{0}, I),\ i = 1, \ldots, m. \tag{18}$$

and

$$V_J^{*-\frac{1}{2}}[\mathbf{b}_I^* - \boldsymbol{\beta}] \rightarrow N_p(\mathbf{0}, I) \tag{19}$$

where $A^{\frac{1}{2}}$ is the unique positive definite and symmetric square root of the positive definite matrix A such that $A = A^{\frac{1}{2}}A^{\frac{1}{2}}$.

Hence, by replicating the pivotal quantity in (18) a large number of times ($m \rightarrow \infty$) a nonparametric bootstrap distribution of (19) can be obtained. This can be used for obtaining tests and confidence regions for any estimable function $C\boldsymbol{\beta}$.

6 Rubin-Schenker Type Estimator

In this section, we describe an estimator which is similar to the one obtained by Rubin and Schenker (1986) using the Asumptotic Bayesian Bootstrap (ABB) method. The method presented here, however, requires only the records of the respondents and the auxiliary variables. We impute the missing values as in (15) and define

$$\begin{aligned}\mathbf{b}_{Ii}^{**}(s) &= (X'X)^{-1}X'\begin{pmatrix}\hat{\mathbf{y}}_1\\ \mathbf{y}_{0i}^{**}\end{pmatrix}\\ &= \mathbf{b}_1 + (X'X)^{-1}X_0'X_0A_1\epsilon_{1i}^* + (X'X)^{-1}X_0'\epsilon_{0i}^{**},\ \ i = 1, \ldots, m.\end{aligned}$$

It can be shown that

$$\text{cov}(\mathbf{b}_{Ii}^{**}(s)) = \sigma^2(X_1'X_1)^{-1} + \sigma^2(X'X)^{-1}(X_0'X_0)(X_1'X_1)^{-1}.$$

Thus

$$\text{cov}(\mathbf{b}_{Ii}^{**(s)}) - \text{cov}(\mathbf{b}_I^*) \geq 0,$$

that is, the difference is at least positive semi-definite. Hence, the multiply imputed estimator of $\boldsymbol{\beta}$ in any replicate has a larger covariance than a singly imputed estimator. The average of all the m replicates is defined by

$$\begin{aligned}\mathbf{b}_{I\cdot}^{**}(s) &= \frac{1}{m}\sum_{i=1}^{m}\mathbf{b}_{Ii}^{**}(s)\\ &= \mathbf{b}_1 + (X'X)^{-1}X_0'X_0A_1\overline{\epsilon}_{1\cdot}^* + (X'X)^{-1}X_0'\overline{\epsilon}_{0\cdot}^{**},\end{aligned}$$

where

$$m\overline{\epsilon}_{1\cdot}^* = \sum_{i=1}^{m}\epsilon_{1i}^* \text{ and } m\overline{\epsilon}_{0\cdot}^{**} = \sum_{i=1}^{m}\epsilon_{0i}^{**}.$$

The covariance of $\mathbf{b}_{I\cdot}^{**}(s)$ is given by

$$\text{cov}(\mathbf{b}_{I\cdot}^{**}(s)) = \sigma^2[(X_1'X_1)^{-1} + \frac{1}{m}((X_1'X_1)^{-1} - (X'X)^{-1})] \to \sigma^2(X_1'X_1)^{-1} \text{ as } m \to \infty.$$

This follows from the following matrix manipulations:

$$\begin{aligned}
& (X'X)^{-1}(X_0'X_0)(X_1'X_1)^{-1}(X_0'X_0)(X'X)^{-1} + (X'X)^{-1}(X_0'X_0)(X'X)^{-1} \\
= \; & (X'X)^{-1}[(X_0'X_0)(X_1'X_1)^{-1} + I](X_0'X_0)(X'X)^{-1} \\
= \; & (X_1'X_1)^{-1}(X_0'X_0)(X'X)^{-1} \\
= \; & (X_1'X_1)^{-1}[X'X - X_1'X_1](X'X)^{-1} \\
= \; & (X_1'X_1)^{-1} - (X'X)^{-1}
\end{aligned}$$

We shall use this representation frequently in what follows.

This covariance can be estimated by

$$T^{**} = \frac{1}{m}\sum_{i=1}^{m} S_i^{**} + \frac{m+1}{m}B^{**},$$

where

$$\begin{aligned}
S_i^{**} \; = \; & (X'X)^{-1}(X_1'X_1)[\mathbf{b}_1 - \mathbf{b}_{Ii}^{**}(s)][\mathbf{b}_1 - \mathbf{b}_{Ii}^{**}(s)]' \\
& +(X'X)^{-1}(X_0'X_0)[\mathbf{b}_{0i}^{*} - \mathbf{b}_{Ii}^{**}(s)][\mathbf{b}_{0i}^{*} - \mathbf{b}_{Ii}^{**}(s)]',
\end{aligned}$$

and

$$(m-1)B^{**} = \sum_{i=1}^{m}(\mathbf{b}_{Ii}^{**}(s) - \bar{\mathbf{b}}_{I\cdot}^{**}(s))(\mathbf{b}_{Ii}^{**}(s) - \bar{\mathbf{b}}_{I\cdot}^{**}(s))',$$

where

$$\mathbf{b}_{0i}^{*} = (X_0'X_0)^{-1}X_0'\mathbf{y}_{0i}^{**}.$$

It can be shown that

$$\begin{aligned}
E_*(S_i^{**}) \; = \; & (X'X)^{-1}(X_1'X_1)[(X_1'X_1)^{-1} - (X'X)^{-1}]S_1^2 \\
& +(X'X)^{-1}(X_0'X_0)[(X_0'X_0)^{-1} - (X'X)^{-1}]S_1^2 \\
= \; & (X'X)^{-1}S_1^2,
\end{aligned}$$

and

$$E_*(B^{**}) = [(X_1'X_1)^{-1} - (X'X)^{-1}]S_1^2.$$

Hence

$$E(T^{**}) = ((X_1'X_1)^{-1} + \frac{1}{m}[(X_1'X_1)^{-1} - (X'X)^{-1}])\sigma^2.$$

Thus under Srivastava's (1971) condition that

$$\max_{1\le i\le n_1} h_{1ii} \to 0 \text{ and } max_{1\le i\le n_0} h_{0ii} \to 0,$$

it follows that as $n \to \infty$,

$$\mathbf{b}_{I\cdot}^{**}(s) \to N_p(\boldsymbol{\beta}, T^{**}).$$

Thus tests and confidence regions for any estimable linear function $C\boldsymbol{\beta}$ can be obtained from the above distribution. However, in using $\mathbf{b}_{I\cdot}^{**}(s)$ as an estimator of $\boldsymbol{\beta}$, we will be giving very little weight to the imputed values especially when m is large.

7 A New Multiply Imputed Estimator

In the previous section we proposed a multiply imputed estimator for the regression parameter, which is similar in spirit to the Asymptotic Bayesian Bootstrap estimator of Rubin and Schenker (1986). However, the covariance of this estimator in any replicate is larger than the corresponding covariance for a singly imputed estimator. In this section, we propose a new multiply imputed estimator which has the same covariance as a singly imputed estimator. In addition, it does not require double imputing, ie. it requires only imputing the missing values. We begin with the residuals $\epsilon_1, \ldots, \epsilon_{n_1}$ defined in (4). We draw a sample of size n_0 with replacement from these residuals. We also replicate it m times. These observations will be denoted by ϵ^*_{ij}, $j = 1, \ldots, n_0$, $i = 1, \ldots, m$. We impute the n_0 missing values by

$$\mathbf{y}^*_{0i} = X_0\mathbf{b}_1 + \boldsymbol{\epsilon}^*_{0i}, \quad i = 1, \ldots, m$$

where $\mathbf{y}^*_{0i} = (y^*_{1i}, \ldots, y^*_{n_0 i})'$ and $\boldsymbol{\epsilon}^*_{0i} = (\epsilon^*_{1i}, \ldots, \epsilon^*_{n_0 i})'$. We define the imputed estimator of $\boldsymbol{\beta}$ for the ith replicate by

$$\begin{aligned}
\mathbf{b}^*_{Ii}(s) &= (X'X)^{-1}X' \begin{pmatrix} \hat{\mathbf{y}}_1 \\ \mathbf{y}^*_{0i} \end{pmatrix} \\
&= (X'X)^{-1}X' \begin{pmatrix} X_1\mathbf{b}_1 \\ X_0\mathbf{b}_1 + \boldsymbol{\epsilon}^*_{0i} \end{pmatrix} \\
&= \mathbf{b}_1 + (X'X)^{-1}X_0'\boldsymbol{\epsilon}^*_{0i}, \quad i = 1, \ldots, m.
\end{aligned}$$

Its covariance is given by

$$\text{cov}(\mathbf{b}^*_{Ii}(s)) = \sigma^2(X_1'X_1)^{-1} + \sigma^2(X'X)^{-1}X_0'X_0(X'X)^{-1},$$

the same as $\text{cov}(\mathbf{b}^*_I)$ given in (10). The overall estimator based on all the m replicates is defined by

$$\begin{aligned}
\bar{\mathbf{b}}^*_{I\cdot}(s) &= \frac{1}{m}\sum_{i=1}^{m}\mathbf{b}^*_{Ii}(s) \\
&= \mathbf{b}_1 + (X'X)^{-1}X_0'\bar{\boldsymbol{\epsilon}}^*_{0\cdot}
\end{aligned}$$

where $\bar{\boldsymbol{\epsilon}}^*_{0\cdot} = m^{-1}\sum_{i=1}^{m}\boldsymbol{\epsilon}^*_{0i}$. Its covariance is given by

$$\begin{aligned}
\text{cov}(\mathbf{b}^*_{I\cdot}(s)) &= \sigma^2(X_1'X_1)^{-1} + \frac{1}{m}\sigma^2(X'X)^{-1}X_0'X_0(X'X)^{-1} \\
&\to \sigma^2(X_1'X_1)^{-1} \text{ as } m \to \infty.
\end{aligned}$$

To estimate this covariance, we define

$$\begin{aligned}
(m-1)B^* &= \sum_{i=1}^{m}(\mathbf{b}^*_{Ii}(s) - \bar{\mathbf{b}}^*_{I\cdot}(s))(\mathbf{b}^*_{Ii}(s) - \bar{\mathbf{b}}^*_{I\cdot}(s))' \\
&= \sum_{i=1}^{m}(X'X)^{-1}X_0'(\boldsymbol{\epsilon}^*_{0i} - \bar{\boldsymbol{\epsilon}}^*_{0\cdot})(\boldsymbol{\epsilon}^*_{0i} - \bar{\boldsymbol{\epsilon}}^*_{0\cdot})'X_0(X'X)^{-1}.
\end{aligned}$$

Clearly

$$E(B^*) = \sigma^2(X'X)^{-1}X_0'X_0(X'X)^{-1}.$$

Further, we define

$$(n-p)S_i^* = (X_1'X_1)^{-1}[(\mathbf{y}_1 - X_1\mathbf{b}_1)'(\mathbf{y}_1 - X_1\mathbf{b}_1) + (\mathbf{y}_{0i}^* - X_0\mathbf{b}_1)'(\mathbf{y}_{0i}^* - X_0\mathbf{b}_1)]$$

Then

$$E(S_i^*) = \sigma^2(X_1'X_1)^{-1}.$$

Hence

$$T^* = \frac{1}{m}\sum_{i=1}^{m} S_i^* + \frac{1}{m}\sum_{i=1}^{m} B_i^*$$

is an unbiased estimator of $\mathrm{cov}(\mathbf{b}_{I.}^*(s))$. It can be shown that

$$\mathbf{b}_{I.}^*(s) - \boldsymbol{\beta} \to N_p(\mathbf{0}, T^*)$$

as $n \to \infty$, $\max_{1\le i\le n_1} h_{1ii} \to 0$, and $\max_{1\le i\le n_0} h_{0ii} \to 0$.

Thus tests and confidence regions for any estimable linear function $C\boldsymbol{\beta}$ can be obtained from the above distribution.

8 Concluding Remarks

In this paper, several imputed estimators of the regression parameters are presented. Among them only the singly imputed estimator treats the imputed values almost as "one of the observed values" as it should be. The multiply imputed estimators gives almost no weight to the imputed values, especially when the number of replications m is sufficiently large. However, to obtain inference or confidence regions for $C\boldsymbol{\beta}$ it remains an open problem to choose between the bootstrap distribution described in Section 5 and the two asymptotic distributions given in Sections 6 and 7. Finally, if the regression model does not have a constant term, the mean of the residuals should be subtracted from each residual and the sampling should be done from these centered residuals.

9 References

1. R. E. Fay *Valid inferences from imputed survey data* In Proceedings of the Section on Survey Research Methods, Washington D. C. American Statistical Association (1993) p. 227-232
2. D. V. Hinkley *Jackknifing in unbalanced situations* Technometrics 19 (1977) 285-292.
3. J. N. K. Rao and J. Shao *Jacknife Variance estimation with survey data under hot deck imputation* Biometrika, 79, (1992) 811-822.
4. D. B. Rubin Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse In Proceedings of the section on Survey Research Methods, Washington, D. C. American Statistical Association (1978) pp. 20-34.
5. A. Sen and M. S. Srivastava, *Regression Analysis, Theory, Methods and Applications* Springer Verlag, New York (1990)
6. M. S. Srivastava On fixed-width confidence bounds for regression parameters Ann. Math. Statistics. 42 (1971) 1403-1411.

7. M. S. Srivastava *Resampling methods for imputing missing observations* Tech. Report No. 9706. University of Toronto (1997)

8. M. S. Srivastava *Discussion to the paper Jackknife and bootstrap and other resampling methods in regression analysis by C. F. J. Wu* Annals of Statistics. 14 (1986) 1331-1334.

9. Srivastava, M. S. and Carter E. M. *The maximum likelihood method for non-response in sample surveys* Survey Methodology, 12 (1986) 61-72.