Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–15

DOI: 10.1177/0962280217715663
journals.sagepub.com/home/smm
SAGE

A robust imputation method for missing responses and covariates in sample selection models

Emmanuel O Ogundimu[1] and Gary S Collins[2]

Abstract

Sample selection arises when the outcome of interest is partially observed in a study. Although sophisticated statistical methods in the parametric and non-parametric framework have been proposed to solve this problem, it is yet unclear how to deal with selectively missing covariate data using simple multiple imputation techniques, especially in the absence of exclusion restrictions and deviation from normality. Motivated by the 2003–2004 NHANES data, where previous authors have studied the effect of socio-economic status on blood pressure with missing data on income variable, we proposed the use of a robust imputation technique based on the selection-t sample selection model. The imputation method, which is developed within the frequentist framework, is compared with competing alternatives in a simulation study. The results indicate that the robust alternative is not susceptible to the absence of exclusion restrictions – a property inherited from the parent selection-t model – and performs better than models based on the normal assumption even when the data is generated from the normal distribution. Applications to missing outcome and covariate data further corroborate the robustness properties of the proposed method. We implemented the proposed approach within the MICE environment in R Statistical Software.

Keywords

Student-*t* distribution, Heckman model, missing data, multiple imputation, robust method, MICE package

1 Introduction

Missing data are ubiquitous throughout the social, behavioral, and medical sciences. The incompleteness of a data set may lead to results that are different from those that would have been obtained had the data set been completely observed. Carpenter et al.[1] distinguished four different approaches to the analysis of missing data: analysis of only those subjects who complete the study; analysis of available data; use of a single or multiple imputation techniques to replace the missing observations with plausible values, then analyze the complete data set; and joint modelling of observed data and the missingness process. The choice of the method of analysis depends on the missing data mechanism as characterized by Rubin.[2] Data are missing completely at random (MCAR) when the probability of missing data on a variable is not related to other measured variables and is unrelated to the variable itself. In this case, a complete case analysis could be used. A less restrictive assumption than the MCAR is the missing at random (MAR) missing data assumption. This occurs when the probability of missing data for a variable is related to other measured variables in the model but not on the values of the variable itself. This assumption and the distinctiveness of the parameters in the observed data and missingness process allow the missingness process to be ignorable. Likelihood inference can be used under ignorability. Data are missing not at random (MNAR) when the probability of missing data on a variable is related to the values of the variable, even after adjusting for other variables. Indeed, the validity of inferences made under different statistical methods depends on the assumption made about the missingness process.

[1]Department of Mathematics, Northumbria University, Newcastle upon Tyne, UK
[2]Centre for Statistics in Medicine, University of Oxford, Oxford, UK

Corresponding author:
Emmanuel O Ogundimu, Department of Mathematics, Northumbria University, Newcastle upon Tyne NE1 8ST, UK.
Email: emmanuel.ogundimu@northumbria.ac.uk

In settings where covariates or covariates and outcomes are subject to selective missing, the use of joint modelling of the observed and the non-response process may not be straightforward. Multiple imputation (MI) is commonly used in such settings, where missing values are *filled in* (singly or multiply) to produce complete data. It has been suggested that the outcome variable be included in the imputation of missing covariates in order to preserve the relationships among variables.[3] This may be challenging for non-monotone missing data patterns. Incidentally, the use of FCS (fully conditional specification) algorithm can simplify the imputation process. This algorithm has been implemented in MICE (Multivariate imputation by chained equations) package in R and STATA. Details of Multiple imputation using MICE can be found in White et al.[4]

A common misunderstanding about MI is that it is restricted to MAR. While it is certainly true that imputation techniques commonly assume MAR, the theory of MI is completely general and also applies to MNAR.[5] For example, a pattern-mixture approach to sensitivity analysis, where missing values are imputed under a plausible MNAR scenario, has been proposed.[6] A tipping point approach or the so-called delta method adds a constant δ to the imputation to create a difference in the means of respondents and nonrespondents that is equal to δ[7]. This requires sensitivity analyses to obtain an appropriate mean difference which may not be tenable for observational studies with MNAR mechanism, especially the Heckman-type sample selection problem.

Sample selection arises when the outcome of interest can only be observed in a subset of the population under study. The data are MNAR because the observed data do not represent a random sample from the population, even after controlling for covariates. A model for selected sample was introduced by Heckman,[8] and several extensions in the parametric framework,[9–12] semi-parametric framework[13] and non-parametric framework[14] have been proposed. Earlier review of sample selection models can be found in Vella.[15] A unified approach to parametric multilevel sample selection model was proposed in Ogundimu and Hutton.[16]

The use of a sample selection modelling framework as an imputation model for MI has been suggested in the literature.[12,17] This approach was implemented for missing covariates data by Galimard et al.[18] and compared against competing methods. The author implemented the method using the moment-based two-step method because of the perceived computational complexities of the corresponding full information maximum likelihood (FIML) method. However, the two-step estimator and its corresponding FIML estimator have often been shown to be susceptible to collinearity in the absence of an exclusion restriction.[19] An exclusion restriction implies that there are variables in the selection equation that are absent in the outcome model or vice versa. This is to avoid multicollinearity as the inverse Mills ratio, which links the outcome and selection equations, in the two-step method can be linear over a wide range of its support. In the absence of an exclusion restriction, model identifiability relies on the non-linearity of the inverse Mills ratio.

In addition, the methods are not robust to outliers and deviations from the assumption of normality. The former led to the proposal of a model that is robust to outliers even in the design space[20] and the latter led to robust alternatives such as the selection-t model.[11] Variants of these proposals exist in the literature (see Lee[10]). In this paper, we focus on the use of selection-t model as the imputation model for missing outcome and covariate data in a Heckman-type missing data problem. We examine its performance in the absence of an exclusion restriction and other forms of model misspecifications. The method is compared against competing alternatives. We also apply our approach to the imputation of missing outcome data (Ambulatory Expenditure) and covariates in the NHANES (National Health and Nutrition Examination Survey) data sets. The method is implemented within the MICE environment in R statistical software.

2 Sample selection model and multiple imputation

In this section, we review the classical Heckman selection model[8,9] and describe the implementation of multiple imputation (frequentist) approach for these models.

2.1 Heckman section model

Let $Y_i^\star$ be the outcome variable of interest, assumed to be linearly related to covariates x_i through the standard multiple regression model

$$Y_i^\star = \beta' x_i + \sigma\varepsilon_{1i}, \quad i = 1, \ldots, N \tag{1}$$

Suppose the main model is supplemented by a selection (missingness) equation

$$S_i^\star = \gamma' w_i + \varepsilon_{2i}, \quad i = 1, \ldots, N \tag{2}$$

where β, γ and σ are unknown parameters; x_i and w_i, which can overlap, are fixed observed characteristics that may be subject to missingness; and $(\varepsilon_{1i}, \varepsilon_{2i})$ are random errors with means zero, variances one and correlation ρ. If we observe $S_i = I(S_i^\star > 0)$ and $Y_i = Y_i^\star S_i$ for $n = \sum_{i=1}^{N} S_i$ of N individuals, the sample $Y_i, i = 1, \ldots, n$ is a selection from the N individuals. The variance of $S_i^\star$ is fixed at 1, because we only observe the sign of $S^\star$, which is insufficient information to estimate its variance. Suppose further that the errors are correlated and follow a bivariate normal distribution, that is

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim \mathcal{N}_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\}$$

where $\rho \in (-1, 1)$ determines the correlation of $Y_i^\star$ and $S_i^\star$, and hence the nature and severity of the selection process. The selection framework factorizes the joint density of $Y^\star$ and $S^\star$ as

$$f(Y^\star, S^\star | x, w, \beta, \gamma) = f(Y^\star | x, \beta) f(S^\star | Y^\star, w, \gamma)$$

Thus

$$f(y|x, S^\star > 0) = \frac{f(y|x) P(S^\star > 0 | y, w)}{P(S^\star > 0 | w)} \tag{3}$$

The observed data, therefore, has a density given by

$$f(y|x, S = 1; \Theta) = \frac{1}{\sigma} \phi\left(\frac{y - \beta' x}{\sigma}\right) \Phi\left(\frac{\gamma' w + \rho\left(\frac{y - \beta' x}{\sigma}\right)}{\sqrt{1 - \rho^2}}\right) \Big/ \Phi(\gamma' w) \tag{4}$$

where $\Theta = (\beta, \sigma, \gamma, \rho)$. This density describes the distribution of the observed data. If the non-intercept terms in γ as well as ρ are 0 in equation (4), the data is MCAR, $\rho = 0$ implies the data is MAR while $\rho \neq 0$ means the missing data is MNAR. The complete density of the sample selection model is used to avoid bias in the estimator when $\rho \neq 0$. This density comprises of a conditional density defined in equation (4), and a discrete component given by $P(S = 1|w)$. The likelihood-based inference for sample selection model is based on

$$l(\Theta) = \sum_{i=1}^{n} S_i (\ln f(y_i | x_i, S_i = 1; \Theta)) + \sum_{i=1}^{n} S_i (\ln \Phi(\gamma' w_i)) + \sum_{i=1}^{n} (1 - S_i) \ln(\Phi(-\gamma' w_i)) \tag{5}$$

The conditional expectation of the observed data, often referred to as the two-step estimator, is given by

$$E(Y|x, S^\star > 0) = \beta' x + \sigma \rho \Lambda(\gamma' w) \tag{6}$$

where $\Lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the inverse Mills ratio. To use equation (6) in practice, a standard probit model for S provides an estimate of $\hat{\gamma}$. The quantity $\Lambda(\hat{\gamma}' w)$ is then taken as an additional covariate in equation (6), and the least squares coefficient of $\Lambda(\hat{\gamma}' w)$ gives an estimate of $\sigma\rho$.

2.2 Multiple imputation for Heckman-type MNAR missing data

2.2.1 *Galimard et al.'s[18] proposal*

Recall that the density $f(y|x, S^\star > 0)$ describes the observed data. Imputation of the missing component under MAR missing data mechanism is based on the assumption that the distribution of the observed data and the non-response process are the same. That is, $f(y|x, S^\star > 0) = f(y|x, S^\star \leq 0)$. This relationship does not hold under the MNAR assumption. The effect of a negative correlation between the outcome and the selection errors is equivalent

to a positive correlation, but selection if $S^{\star} \leq 0$. The imputation model for the missing values in $Y^{\star}$ is then written as

$$E(Y|x, S^{\star} \leq 0) = \beta' x - \sigma\rho\Lambda(-\gamma' w) \tag{7}$$

Equation (7) is equivalent to equation 6 in Galimard et al.[18] The authors *filled in* the missing data using the Heckman two-step approach with the equation of the form

$$Y_i^* = \beta'^* x - (\sigma\rho)^* \Lambda(-\gamma'^* w) + \eta^* \tag{8}$$

where $\eta^* \sim \mathcal{N}(0, \sigma_\eta^{2*})$ and σ_η^{2*}, β^*, $(\sigma\rho)^*$ are drawn using approximate proper imputation. Estimated values of these parameters are obtained from the two-step estimator.

Galimard et al.[18] acknowledged that the two-step method is less efficient than the ML/FIML estimator. With the advent of powerful computational tools, efficiency and accuracy cannot be traded for computational complexities. We therefore present two approaches based on the ML estimator, which will be used in subsequent analyses.

2.2.2 Maximum likelihood estimator approach

If $Y_{i,obs}$ and $Y_{i,mis}$ represent the observed and the missing parts of $Y^{\star}$, respectively, then the process of imputing values requires that missing values are drawn multiple (M) times from

$$Y_{i,mis}^{(k)} \sim p\big(Y_{i,mis}|Y_{i,obs}, x_i\big), \quad \text{with } k \in \{1, \ldots, M\} \tag{9}$$

where $p(\cdot)$ denotes the posterior predictive distribution. It can be difficult sometimes to draw from this distribution; therefore, iterative imputation approaches such as data augmentation[21] can be used. Whilst this approach is theoretically preferable, it can be computationally burdensome. We therefore propose two methods based on the approximation of the predictive distribution in a frequentist framework. In order to motivate these methods, we first note that sampling from equation (9) requires the true value of the parameter Θ, which is unknown in practice. As an alternative value, we consider $\hat{\Theta}$, the maximum likelihood estimator of Θ and, in order to formally account for the uncertainty on the true value of Θ, we consider sampling from the following conditional distribution

$$p\big(Y_{i,mis}|Y_{i,obs}, x_i\big) = \int p\Big(Y_{i,mis}|Y_{i,obs}, x_i, \hat{\Theta}\Big)\pi(\hat{\Theta})\mathrm{d}\hat{\Theta} \tag{10}$$

where $\pi(\hat{\Theta})$ represents the distribution of the MLE. This approach clearly takes into consideration the uncertainty about the parameters in the light of the data, which is summarized in $\pi(\hat{\Theta})$, and integrated out using the law of total probability. Given that the distribution $\pi(\hat{\Theta})$ is not available in closed form in equation (10), we consider two approaches for approximating this function and sequentially sampling from it. The first approach is based on the asymptotic normality of the maximum likelihood estimators Θ obtained from equation (5). We approximate the draws in equation (9) using

$$Y_{i,mis}^{(k)} \sim p\big(Y_{i,mis}|Y_{i,obs}, x_i, \tilde{\Theta}^{(k)}\big)$$

where $\tilde{\Theta}^{(k)} = (\tilde{\beta}^{(k)}, \tilde{\sigma}^{(k)}, \tilde{\gamma}^{(k)}, \tilde{\rho}^{(k)})$ are drawn from the asymptotic normal distribution of the maximum likelihood estimator, $\hat{\Theta}_{ML}$. If we denote the consistent estimator of the corresponding large-sample covariance matrix by $C(\hat{\Theta}_{ML})$, then

$$\tilde{\Theta}^{(k)} \sim \mathcal{N}\Big(\hat{\Theta}_{ML}, C(\hat{\Theta}_{ML})\Big)$$

$C(\hat{\Theta}_{ML})$ is obtained from the inversion of the observed information matrix of the FIML estimator.

The second approach is based on non-parametric bootstrap and the draws in equation (9) is approximated by

$$Y_{i,mis}^{(k)} \sim p\Big(Y_{i,mis}|Y_{i,obs}, x_i, \breve{\Theta}^{(k)}\Big)$$

where $\breve{\Theta}^{(k)} = (\breve{\beta}^{(k)}, \breve{\sigma}^{(k)}, \breve{\gamma}^{(k)}, \breve{\rho}^{(k)})$ are the maximum likelihood estimates based on a bootstrapped sample $B^{(k)}_{boot}$ of the original data set.

Now, for a given draw $\Theta^{(k)}$ based on the asymptotic or bootstrap approach, the imputation of $Y^{(k)}_{i,mis}$ is predicted from equation (7).

3 Robust alternatives using t-distribution

Suppose the error terms in equations (1) and (2) follow a bivariate t-distribution. That is

$$\begin{pmatrix}\varepsilon_{1i}\\ \varepsilon_{2i}\end{pmatrix} \sim t_2\left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}, \nu\right\}$$

where t_2 is the PDF of a bivariate t-distribution, ρ is the correlation parameter and ν is degrees of freedom. Marchenko et al.[11] proposed this approach as a robust alternative to the selection normal model of Heckman.[8] The conditional distribution, using equation (3), is given by

$$\begin{aligned} f(y|x, S=1; \Xi) = &\frac{1}{\sigma} t\left(\frac{y-\beta' x}{\sigma}; \nu\right) \\ &\times T\left\{\left(\frac{\gamma' w + \rho\left(\frac{y-\beta' x}{\sigma}\right)}{\sqrt{1-\rho^2}}\right)\left(\frac{\nu+1}{\nu+\left(\frac{y-\beta' x}{\sigma}\right)^2}\right)^{1/2}; \nu+1\right\} \Big/ T(\gamma' w; \nu) \end{aligned} \tag{11}$$

where $\Xi = (\beta, \sigma, \gamma, \rho, \nu)$, $t(\cdot; \nu)$ and $T(\cdot; \nu)$ are the PDF and CDF of a univariate Student's t-distribution with ν degrees-of-freedom. The likelihood function corresponding to equation (11) is written as

$$l(\Xi) = \sum_{i=1}^{n} S_i(\ln f(y_i|x_i, S_i=1; \Xi)) + \sum_{i=1}^{n} S_i(\ln T(\gamma' w_i; \nu)) + \sum_{i=1}^{n}(1-S_i)\ln(T(-\gamma' w_i; \nu)) \tag{12}$$

The conditional moment is given by

$$E(Y|x, S^\star > 0) = \beta' x + \sigma\rho\Lambda_\nu(\gamma' w), \quad \nu > 1 \tag{13}$$

where $\Lambda_\nu(k) = \frac{\nu+k^2}{\nu-1}\frac{t(k;\nu)}{T(k;\nu)}$. Equation (13) can be used in a similar way as equation (6). A robit (binary regression with t-distribution) for S provides estimate of $\hat{\gamma}$ and $\hat{\nu}$. The quantity $\Lambda_\nu(k)$ is taken as an additional covariate in equation (13) and the least squares estimate gives the value of $\sigma\rho$. The conditional variance is

$$\begin{aligned} \mathrm{var}(Y|x, S^\star > 0) = \sigma^2\Bigg[&\frac{\nu(1-\rho^2)}{(\nu-1)} + \Lambda_{\nu-2}(\gamma' w)\left\{\frac{1+\nu\rho^2-2\rho^2}{\nu-1}\right\} \\ &-\Lambda_\nu(\gamma' w)\left\{\frac{\gamma' w(1-\rho^2)}{(\nu-1)} + \gamma' w\rho^2 + \rho^2\Lambda_\nu(\gamma' w)\right\}\Bigg] \end{aligned} \tag{14}$$

where $\Lambda_{\nu-2}(k) = \frac{\nu}{\nu-2}\frac{T_1(k\sqrt{(\nu-2)/\nu}; \nu-2)}{T_1(k;\nu)}$, $\nu > 2$ and $\Lambda_\nu(k)$ is as defined in equation (13). Unlike in equation (6), where the estimates of both ρ and σ are obtained by equating the average value of the conditional variance to the observed residual variance of the OLS regression in the second stage, equation (13) does not allow for such simplification. Theoretically, the variance of t-distribution is related to its tailweight (ν) as can been seen in equation (14).

To minimize the burden of the estimation of ν in equation (12), we used a discretized version of the t-distribution (see Villa and Walker[22]) to estimate ν. Since it is hard to distinguish models in-between two integer degrees-of-freedom, and models with $\nu > 50$ are indistinguishable from the normal model (see Figure 1), we consider the set $\nu = \{2.5, 3, 3.5, \ldots, 100\}$. This approach is expected to produce accurate estimate of ν. Initial values for other parameters are obtained from their corresponding two-step estimator. These estimates are then used as the initial value in the likelihood function (12) to minimize the possibility of model convergence to a local maximum. We imputed the missing data using $E(Y|x, S^\star < 0) = \beta' x - \sigma\rho\Lambda_\nu(-\gamma' w)$.

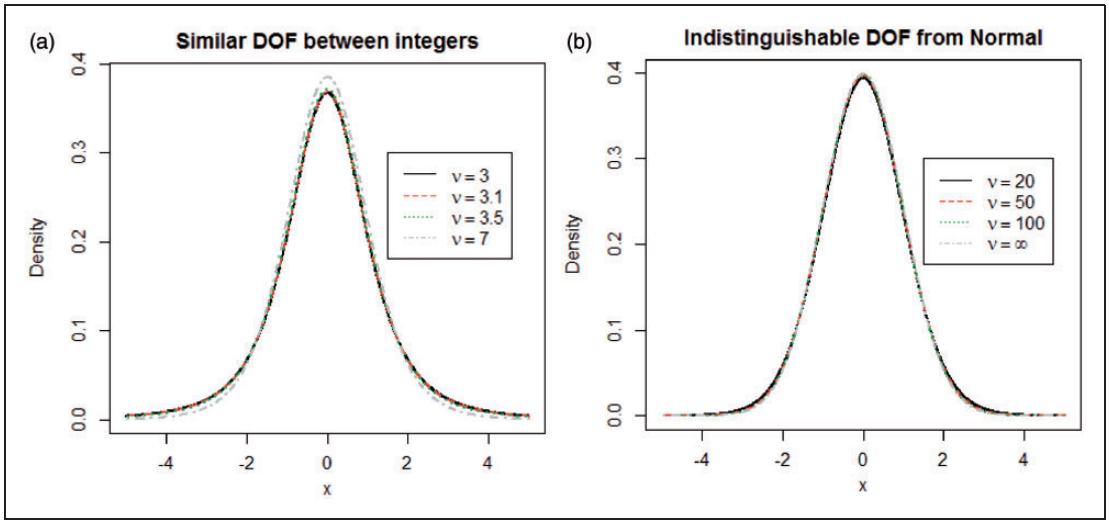

Figure 1. PDF of t distribution. DOF stands for degrees-of-freedom: (a) with varying ν showing difference between integers are indistinguishable. Only $\nu = 7$ differs from others; (b) with varying ν showing $\nu = 50$ can approximate the normal distribution.

4 Simulation study

We compare the performance of the proposed robust alternative method of imputation based on asymptotic (ST) and bootstrap (STB) with Galimard et al.'s[18] two-step imputation method (Tstep). We also included the Heckman full information maximum likelihood method with the asymptotic (SNM) and bootstrap (SNMB) imputation for *control* purposes. We first consider simulation settings where the outcome and selection models have bivariate-t error distribution. The outcome equation is $Y_i^\star = 0.5 + 1.5x_{1i} + x_{2i} + \varepsilon_{1i}$, where x_{1i} and $x_{2i} \overset{iid}{\sim} N(0,1)$ and $i = 1, \ldots, N = 1000$. The impact of exclusion restriction on the proposed method is evaluated with selection equations $S_i^\star = 1 + x_{1i} + 0.2x_{2i} + 1.5w_i + \varepsilon_{2i}$, $w_i \overset{iid}{\sim} N(0,1)$ for exclusion restriction, and $S_i^\star = 1 + x_{1i} + 0.2x_{2i} + \varepsilon_{2i}$ without an exclusion restriction. Hence, $\beta' = (0.5, 1.5, 1.0)$, and $\gamma' = (1, 1, 0.2, 1.5)$ and (1, 1, 0.2) for selection with and without the exclusion restriction, respectively. The covariates x_{1i}, x_{2i} and w_i are independent and are also independent of the error terms $\varepsilon_i' = (\varepsilon_{1i}, \varepsilon_{2i})$. The error terms are generated from bivariate t distribution with degree-of-freedom = 5 and $\rho = 0.5$. The covariance matrix is $\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}$, where $\sigma = 1$. We only observe values of $Y_i^\star$ when $S_i^\star > 0$. About 33% of values are not observed with this exclusion restriction, and about 28% without it.

Although it is, perhaps, more instructive to focus on the use of the robust imputation method on covariates that are subject to sample selection, we first examined the performance of the methods on the imputation of missing outcome data. The following scenarios are considered.

4.1 First scenario – missingness in outcome

(i) Impact of an exclusion restriction: we used $S_i^\star = 1 + x_{1i} + 0.2x_{2i} + 1.5w_i + \varepsilon_{2i}$ as the selection component of the model so that there is an additional variable w not present in the outcome model
(ii) Impact of the no exclusion restriction: we used $S_i^\star = 1 + x_{1i} + 0.2x_{2i} + \varepsilon_{2i}$ as the selection component of the model. That is, the covariates in the outcome and the selection equations overlap.
(iii) Impact of noise variables as exclusion restriction: we used the model in 2 with additional three noise variables. Noise variables are variables whose true regression coefficients are zero. These variables are common in prespecified models.
(iv) Impact of outliers: we generated the data from a selection model with Gaussian mixture errors of the form $(1-p)\mathcal{N}_2(\mathbf{0}, \Sigma) + p\mathcal{N}_2(\mathbf{0}, k\Sigma)$, where $k > 0$, p = 0.1 and $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. The selection equation with the exclusion restriction in 1 is used.

4.2 Second scenario – missingness in covariates

Simulated data were generated using the regression equation $Y_i = 0.5 + 1.5x_{1i} + x^{\star}_{2i} + \sigma\varepsilon_{3i}$, where $\varepsilon_{3i} \sim N(0,1)$, $x_{1i} \sim N(0,1)$ and $\sigma = 1$. The underlying regression equation for $x^{\star}_{2i} = 1 + x_{1i} + \varepsilon_{1i}$ with missingness process $S^{\star}_i = 0.4 + x_{1i} + \varepsilon_{2i}$. We generated the errors from a bivariate-t distribution with $\nu = 5$ as above. The observed version of $x^{\star}_{2i}$ has about 40% missing data. We consider $\rho = 0$ (MAR) and $\rho = 0.3$ and 0.5 (MNAR).

4.2.1 Simulation results

Figure 2 shows the boxplot of the parameter estimates from the imputed outcome data with or without the exclusion restriction from 1000 simulated data sets. None of the methods is biased under the exclusion restriction (Figure 2(a)). ST and STB methods are, however, closer to the true parameter and less variable than the methods based on the normal distribution (Tstep, SNM & SNMB). Interestingly, the performance of ST and STB methods in the absence of the exclusion restriction (Figure 2(b)) is superior to their counterparts under the exclusion restriction. This supported the argument of alleviation of the problem of collinearity adduced for the use of t distribution in sample selection framework by Marchenko and Genton.[11] The methods based on the normal distribution are not only biased but highly variable (i.e. imprecise). In particular, the parameter estimates of the outcome under the Tstep method ranges between -0.15 and 1.27 whereas the ST and STB estimates range between 0.22 and 0.87, and 0.23 and 0.86, respectively. Thus, Figure 2 implies the distributional misspecification does not bias the parameter estimates significantly when extra variable predictive of missingness is included in the selection equation of the normal imputation models. The parameter estimates are biased in the absence of an exclusion restriction.

The quest for variables to use for the exclusion restriction criteria to be fulfilled is a daunting exercise in practice. Sometimes, the use of noise variables in the selection equation can constitute a nuisance in the model estimation. Figure 3(a) shows that the inclusion of irrelevant (noise) variables in order to satisfy the exclusion restriction criteria does not make the associated problem go away under the normal models. The Tstep method, although unbiased is highly variable (similar variability as the absence of an exclusion criteria in Figure 2(b)) providing the most extreme parameter estimates of the response means. ST and STB methods yielded unbiased and low variability results.

The robustness of the proposed imputation method is examined against outliers induced as a result of mixture of normal distributions (Figure 3(b)). Clearly, both the normal and robust imputation is misspecified. This however, does not appear to bias the parameter estimates (although ST and STB methods are more concentrated near the true parameter). Again, both the ST and STB methods are robust to this misspecification as the estimates are unbiased and show lower variability than the normal models.

Table 1 shows the coverage of the 95% confidence interval for the parameter estimates using the five imputation methods. The nominal coverage level of 95% is satisfied by the methods for data generated under the exclusion

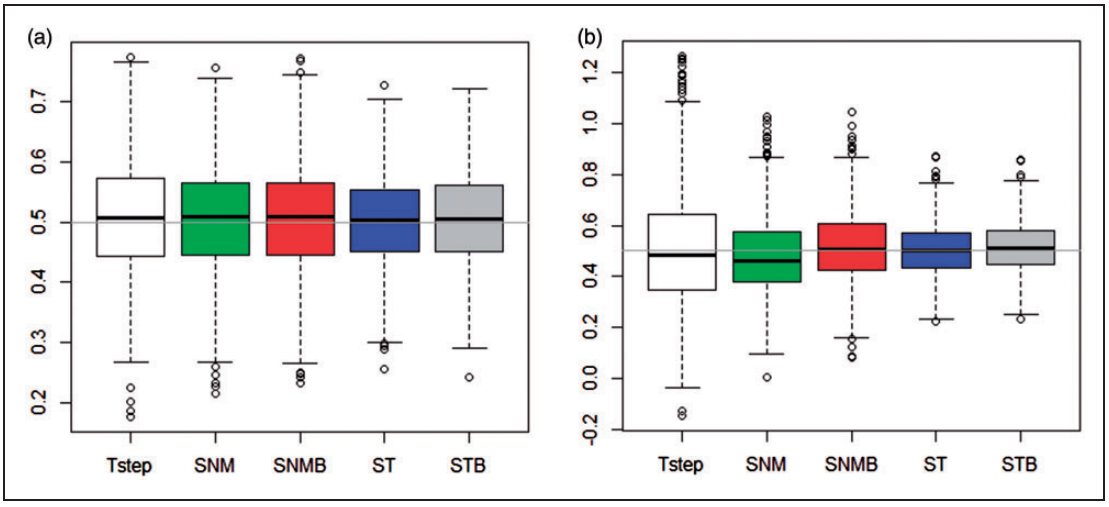

Figure 2. Boxplots for the imputed outcome data: (a) Exclusion restriction; (b) absence of exclusion restriction.

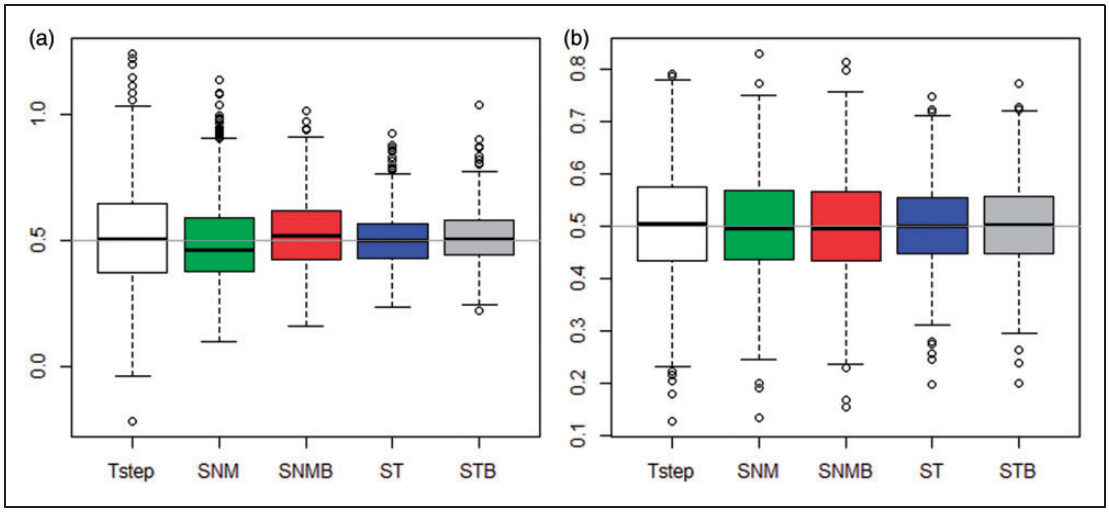

Figure 3. Boxplots for the imputed outcome data: (a) noise variables as an exclusion restriction; (b) mixture distribution.

Table 1. 95% Confidence interval coverage.

	Exclusion	No Exclusion	Noise variable	Mixture
Tstep	94.1	91.8	92.9	95.3
SNM	94.9	96.5	90.1	94.1
SNMB	94.9	93.6	92.2	94.5
ST	95.1	95.5	94.8	94.7
STB	95.3	95.8	96.0	96.2

Table 2. Simulation results for missing covariate data with coefficient of $x_2 = 1$.

	$\rho = 0$		$\rho = 0.3$		$\rho = 0.5$	
	Mean	Variance	Mean	Variance	Mean	Variance
Tstep	0.956	0.003	0.966	0.003	0.972	0.004
SNM	0.972	0.002	0.982	0.002	0.986	0.003
SNMB	0.974	0.002	0.983	0.002	0.984	0.003
ST	0.981	0.002	0.991	0.002	0.992	0.002
STB	0.980	0.002	0.989	0.002	0.989	0.002

$\rho = 0$ represents MAR.

restriction criteria and the mixture distribution. However, the Tstep method shows poor coverage in the absence of the exclusion restriction (91.8%) and noise variables (92.9%). Only the ST and STB methods exhibit satisfactory coverage under these conditions.

Table 2 shows the results of fitting a normal error regression model using the fully observed variable x_1 and partially observed x_2. The performance of the models improved as ρ increases. The results also supported the robustness of the ST model.

5 Empirical studies

We consider two data examples to illustrate the performance of the robust imputation method. The first data are the ambulatory expenditure from the 2001 Medical Expenditure Panel Survey analyzed by Cameron and Trivedi.[23] The data were also analyzed using selection-t model in Marchenko and Genton.[11] The second application involves

the imputation of the income variable in the 2003–2004 NHANES data. This data were analyzed in Little and Zhang,[24] where missing income data were assumed to be missing not at random. We focus on the Tstep, SNM and ST methods since the performance of the bootstrap imputation methods is not too different from the corresponding asymptotic imputation methods.

5.1 Ambulatory expenditure data

The data on ambulatory expenditure contains 3328 observations of which 526 (15.8%) of the outcome of interest (expenditure) is missing. Apart from expenditure, which is highly skewed, other explanatory variables such as age, gender, education status (educ), ethnicity (blhisp), number of chronic conditions (totchr), insurance status (ins) and income are available in the data. We use log expenditure (lambexp) as the outcome variable due to skewness in line with earlier proposals.[11,23] The outcome equation, which is usually the model of interest, contains $x = (1, age, female, educ, blhisp, totchr, ins)$ while the selection equation, $w = (x, income)$. Income is included for the exclusion restriction criteria although its use for this purpose is debatable.[11,23] We emphasize that an exclusion restriction is not a necessary condition for the consistency of the proposed imputation method.

Table 3 shows the results of the robust imputation method and the two alternatives based on the normal distribution. The result is consistent with equivalent results in Marchenko and Genton,[11] supporting the adequacy of the proposed method. Figure 4 shows the distribution of the residuals of the missing data model.

Table 3. Estimates from the Outcome model of the Ambulatory expenditure data after multiple imputation.

	Selection-t		Selection-normal		Two-step	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Outcome model						
(Intercept)	5.203	(4.711, 5.695)	5.122	(4.727, 5.517)	5.117	(4.329, 5.905)
age	0.207	(0.161, 0.252)	0.207	(0.165, 0.250)	0.206	(0.152, 0.261)
female	0.309	(0.182, 0.436)	0.341	(0.236, 0.445)	0.339	(0.182, 0.496)
educ	0.018	(−0.003, 0.039)	0.016	(−0.004, 0.037)	0.018	(−0.010, 0.045)
blhisp	−0.193	(−0.317, −0.069)	−0.218	(−0.344, −0.092)	−0.212	(−0.333, −0.091)
totchr	0.509	(0.422, 0.596)	0.533	(0.464, 0.601)	0.526	(0.405, 0.647)
ins	−0.054	(−0.152, 0.044)	−0.030	(−0.125, 0.065)	−0.030	(−0.137, 0.078)
σ	1.211	(1.144, 1.278)	1.283	(1.246, 1.319)	1.291	(1.187, 1.394)
ν	13.508	(6.214, 20.801)				

CI: Confidence interval.

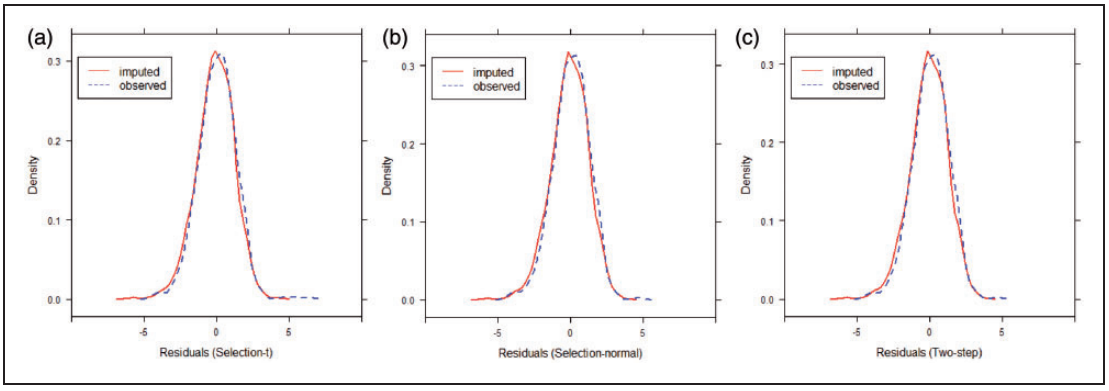

Figure 4. Distribution of residuals of the missing data model for the outcome data: (a) selection-t (ST); (b) selection normal (SNM); (c) two-step.

Table 4. Imputation of missing outcome in the Ambulatory expenditure data under the *MAR* assumption.

	Estimate	95% CI
(Intercept)	4.872	(44.536, 5.208)
age	0.219	(0.174, 0.265)
female	0.389	(0.291, 0.487)
educ	0.025	(0.006, 0.044)
blhisp	−0.244	(−0.350, −0.138)
totchr	0.569	(0.510, 0.627)
ins	−0.013	(−0.113, 0.087)
σ	1.181	(1.132, 1.231)
ν	15.709	(8.054, 23.364)

CI: Confidence interval.

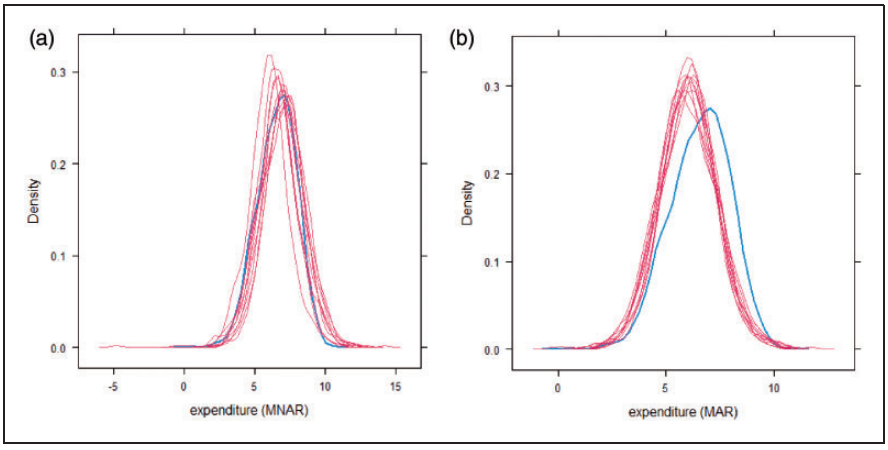

Figure 5. Kernel density estimates for the marginal distributions of the observed data (blue) and the 10 densities for each imputation for expenditure: (a) MNAR (selection-t); (b) MAR (regression with t-distributed errors).

Since the imputation model is correctly specified as an MNAR, the spread of the residuals of the observed and the imputed data is very similar.

We also imputed the data under the MAR assumption. That is, we imputed the data using the model specification for the outcome equation. The results are shown in Table 4. Previous analyses of the data posited that all the factors other than the insurance status (ins) are strong predictors of expenditure.[11] This was supported by the MNAR results in Table 3. The parameter estimates under the MAR assumption are generally larger in magnitude than their MNAR counterparts. Further, two variables (education and insurance status) are not predictors of expenditure under the MAR model. Figure 5 shows kernel density estimates of the imputed and observed expenditure data. There are discrepancies between the densities of the observed and imputed data under the MAR model whereas the densities under the MNAR model are similar.

5.2 NHANES data

The US National Health and Nutrition Examination Study (NHANES) is a survey data collected by the US National Center for Health Statistics. The survey data dates back to 1999, where individuals of all ages are interviewed in their home annually and complete the health examination component of the survey. The study variables include demographic variables (e.g. age and annual household income), physical measurements (e.g. BMI – body mass index), health variables (e.g. diabetes status), and lifestyle variables (e.g. smoking status).

We used NHANES 2003–2004 data to illustrate the methodology of using the robust imputation strategy proposed for the imputation of missing covariates (household income) in a model developed to study the effect of socio-economic status on systolic blood pressure (SBP). The data has been used in Little and Zhang[24] to

Table 5. Percentages of missing data in the NHANES, 2003–2004.

		SBP and BMI
Variable	Percentage of missing ($n = 9643$)	Without missing ($n = 6193$)
SBP (mm Hg)	34.94	0
Age (years)	0	0
Gender	0	0
BMI (kg/m^2)	9.91	0
Education (years)	17.23	0
Race	0	0
Income ($1000 per year)	24.41	25.43

SBP: systolic blood pressure; BMI: body mass index.

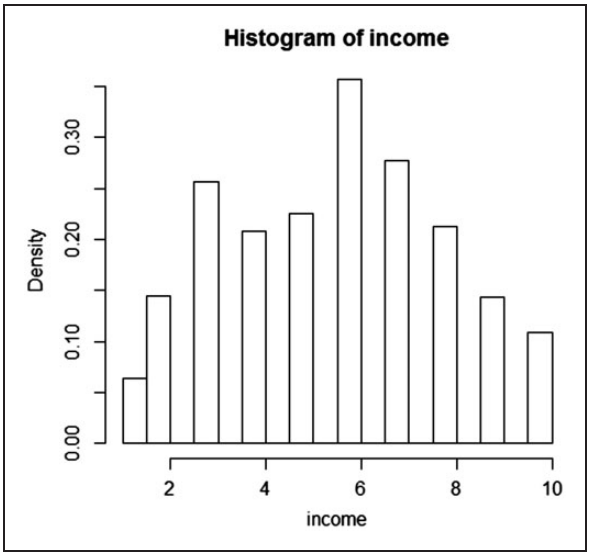

Figure 6. Histogram of ordinal income variable.

illustrate the method of subsample ignorable likelihood for MNAR missing covariates (Income) to study the same effect.[24] considered three covariates: age (in years), gender and BMI, and two socio-economic status variables: income and years of education. We used the same set of variables but added race as additional variable that can predict missingness in household income.

Table 5 shows the percentage of missing data in the variables selected for analysis. Age, gender, and race are fully observed, whereas SBP, BMI, and household income are subject to missing data. The data analyzed are reconstructed such that only income variable has missing data. That is, complete data on SBP and BMI are selected with corresponding measurements on the other variables. This resulted in income having 25.43% missing data and no missing data on other variables. In principle, there is no need for this as the MICE algorithm allows imputation of multiple missing variables under MAR and MNAR assumptions. We focus on income variable in order to evaluate the unalloyed effect of the proposed imputation strategy.

Household income ($1000 per year) was reported as a range of values in dollar (e.g. 0–4999, 5000–9999, etc.) and had 10 interval categories. Figure 6 shows that the ordinal categories of income can be approximated by a continuous distribution. This allows straightforward adaptation of the proposed method without the need for adjustments for ordinal data imputed as continuous data. Education is dichotomized into high school and above versus less than high school and race is treated as categorical variable with five levels.

Age, gender, education, and race are potential factors that can predict income. These factors are also known to lead to selective reporting of income. Therefore, the same set of variables is used in the selection and outcome equations (without exclusion restrictions) of the imputation model. Income was imputed using 10 imputations.

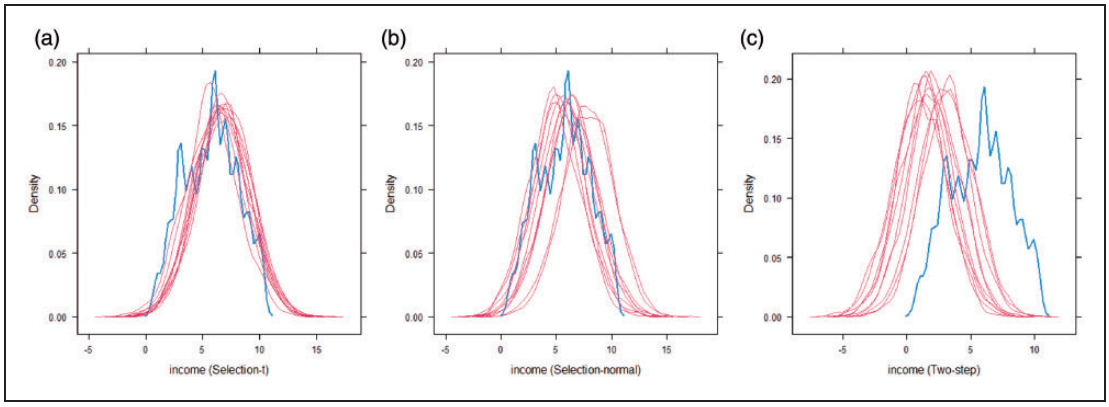

Figure 7. Kernel density estimates for the marginal distributions of the observed data (blue) and the 10 densities for each imputation for income (red): (a) Selection-t (ST); (b) selection normal (SNM); (c) two-step.

Table 6. Estimates of the effect of socio-economic status on Systolic blood pressure (NHANES, 2003–2004).

	Selection-t		Selection-normal		Two-step	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Outcome model						
(Intercept)	93.173	(91.483, 94.863)	92.676	(90.665, 94.688)	92.498	(90.694, 94.301)
Age	0.489	(0.470, 0.508)	0.558	(0.539, 0.577)	0.558	(0.540, 0.577)
Sex (male)	−4.362	(−5.022, −3.702)	−2.957	(−3.730, −2.184)	−2.953	(−3.725, −2.181)
Education	−2.981	(−3.728, −2.234)	−3.273	(−4.131, −2.416)	−3.283	(−4.127, −2.439)
BMI	0.432	(0.376, 0.488)	0.382	(0.318, 0.447)	0.382	(0.318, 0.447)
Income	0.005	(−0.147, 0.156)	0.007	(−0.179, 0.192)	0.045	(−0.102, 0.193)
σ	11.047	(10.689, 11.405)	15.462	(15.190, 15.735)	15.462	(15.190, 15.734)
N	3.782	(3.394, 4.171)				

CI: confidence interval; BMI: body mass index.

Figure 7 shows kernel density estimates of the imputed and observed income data. The plot based on the selection-t model produces densities of observed and imputed data that match up well. The densities based on the selection normal model approximately match up (two of the imputed data appear to be shifted away from the observed data). However, there are discrepancies between the densities of the observed and imputed data for the two-step method. A possible explanation for this is the anomalous behavior of the method in the absence of an exclusion restriction, which was also evident from the simulation studies in Section 4.

Table 6 shows the results of fitting regression models (Selection-t: linear regression with student-t errors; Selection-normal and Two-step: OLS regression) to the imputed data sets. The models showed that income is not significantly related to SBP. This observation is analogous to the effect of income in the ignorable likelihood method proposed for the same data in Little and Zhang.[24] The degrees of freedom (ν) in the selection-t model is significant (Estimate = 3.782, CI = [3.394, 4.171]).

6 Concluding remarks

This paper proposes the use of selection-t model developed by Marchenko and Genton[11] as a robust imputation alternative for missing outcome and covariates data in Heckman-type missing data problem. We have denoted the proposed method as *ST* and compared it with competing alternatives based on the Heckman's full information maximum likelihood (*SNM*) and the two-step (*Tstep*) method. Contrary to the common notion that MI is valid only under MAR, we have shown that correct specification of the imputation model under MNAR can result in unbiased parameter estimates and valid statistical inference. We have imputed partially observed data by drawing from their conditional distributions using the FCS algorithm. Our proposed imputation method is based on frequentist philosophy (approximate proper imputation) as

opposed to the Bayesian (proper) imputation method. The former is easier to use, less computationally intensive and works well in large samples.

Apart from the use of *ST* method to impute missing covariates data, we have shown its performance for missing outcome data. This was done for two reasons. First, we are able to show that the method performs equally well as its parent sample selection model. Second, the method lends itself naturally to various extensions of the traditional MI techniques (e.g. double robustness concepts can be easily integrated into the imputation model). Specifically, the method can be easily extended to other MNAR imputation models. For instance, instead of the use of imputation model $E(Y|x, S^{\star} < 0) = \beta' x - \sigma\rho\Lambda_{\nu}(-\gamma' w)$, which is similar to the jump to reference approach,[25,26] the imputation approach can also incorporate some form of pattern mixture-model. That is, the imputation model can be multiplied by a factor or offsets added based on subject matter knowledge.

Two simulation studies were conducted to assess the performance of the *ST* imputation method in missing outcome and covariates data. The method uniformly outperformed the *SNM* and *Tstep* methods in terms of bias and low variance. It attains the nominal coverage level when the missing outcome is imputed under four possible model misspecifications (absence of an exclusion restriction, noise variables, distributional misspecification and outliers). In particular, the *ST* method performs very well especially in the absence of an exclusion restriction, a problem which has bedeviled the sample selection modelling framework for some time. This attribute is inherited from the parent selection-t model. Basically, for the selection-t model, the inverse Mills ratio is mostly non-linear over a wide range of its support. This may also explain the shrinkage effect of the function on the noise variables in the selection process. We emphasize that the good performance of the *ST* method is not attributable to the data generation process. Figure 8 (Appendix) shows that the *ST* method still outperforms its competitors even when the data are generated from the normal distribution. The simulation based on covariates data also supported the superiority of the ST method over the *SNM* and *Tstep* methods.

We analyzed two sets of data – Ambulatory expenditure and the 2003–2004 NHANES data sets. The results from the former are comparable with previous analyses in Marchenko and Genton.[11] The method showed similar fit to the *SNM* and the *Tstep* methods. This is, perhaps, due to the exclusion restriction as a result of the omission of income from the outcome equation. However, the advantage of the robust method became pronounced in the imputation of missing covariate (income) in the NHANES data set. We have judged the adequacy of the robust MI method by comparing the distributions of the observed and imputed data. This approach is only valid for the imputation of data that are MAR. Theoretically, the purpose of a reasonably complex imputation model, such as the one proposed here, is to supply sufficient auxiliary variables in appropriate form to make MAR assumption more plausible. As can be seen from Figure 7, the densities of the observed and imputed data are satisfactorily close for the *ST* method than competing alternatives. This may be due to the absence of an exclusion restriction. It is noteworthy that the use of complete data on systolic blood pressure (SBP) in the NHANES data is for illustrative purposes only. Clearly, missing outcome and covariate data can be accommodated within the MICE imputation algorithm simultaneously.

Another strength of the proposed methodology is that we do not need to fix any value for ν (the degrees-of-freedom). This can be estimated from data, even when the data are approximately normally distributed. To prevent the likelihood function from possibly converging to a local maximum, we first searched for the values of ν between 2 and 100 that yielded the best fit for the data. This was set as the initial value for ν in the second stage of the maximization of the full log-likelihood function. This may be superfluous in many practical applications. Initial values for other model parameters were obtained from the corresponding two-step method.

Various extensions of the model proposed in Section 3 can be formulated. One such extension involves the development of a more flexible imputation model than the method introduced in Section 3 using copulas. The use of copulas as alternative modelling framework in selectively reported samples was suggested in Lee[10] and further expounded in Smith.[27] The fact that different copulas exhibit different dependence patterns offer additional flexibility in its use as imputation model in this setting. The method can readily be extended to impute missing covariates in multilevel sample selection settings.[16] We are currently investigating methodologies for obtaining unbiased imputation and variance estimators in this framework.

Finally, although the method we proposed is robust against certain misspecification, it has its limitations, some of which are inherited from the parent selection-t model. For example, the proposed method produced bias estimates when the missingness in a covariate depends on the value of the covariate but conditionally independent of the outcome. Apart from the use of parametric models to achieve robustness, semiparametric and nonparametric sample selection models can be adapted. Ultimately, the guiding principle of any imputation method should be based on the research questions and the use of appropriate sensitivity analysis. The code for the proposed method is available in the Supporting Information.

Acknowledgements

The authors would like to thank the editor and the anonymous referees for the constructive comments, which led to improvements in the manuscript. We thank Francisco J. Rubio for useful discussion.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Carpenter J, Pocock S and Lamm CJ. Coping with missing data in clinical trials: a model-based approach applied to asthma trials. *Stat Med* 2002; **21**: 1043–1066.
2. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581–592.
3. Moon KG, Donders R, Stijnen T, et al. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* 2006; **59**: 1092–1101.
4. White IR, Royston P and Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011; **30**: 377–399.
5. Van Burren S and Groothuis-Oudshoorn K. MICE: Multivariate imputation by chained equations in R. *J Stat Softw* 2011; **45**: 1–67.
6. Yuan Y. *Sensitivity analysis in multiple imputation for missing data*. In: *Proceedings of the SAS Global Forum 2014 Conference*. http://support.sas.com/resources/papers/proceedings14/SAS270-2014.pdf.
7. Van Burren S. *Flexible imputation of missing data*. Boca Raton, FL: Chapman & Hall, 2012, pp.88–89.
8. Heckman J. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Ann Econ Soc Meas* 1976; **5**: 475–492.
9. Heckman J. Sample selection bias as a specification error. *Econometrica* 1979; **47**: 153–161.
10. Lee L. Generalized econometric models with selectivity. *Econometrica* 1983; **51**: 507–512.
11. Marchenko YV and Genton MG. A Heckman Selection-t model. *J Am Stat Assoc* 2012; **107**: 304–317.
12. Ogundimu EO and Hutton JL. A sample selection model with Skew-normal distribution. *Scand J Stat* 2015; **43**: 172–190.
13. Ahn H and Powell JL. Semi-parametric estimation of censored selection models with a nonparametric selection mechanism. *J Econometrics* 1993; **58**: 3–29.
14. Das M, Newwey WK and Vella F. Non-parametric estimation of sample selection models. *Rev Econ Stud* 2003; **70**: 33–58.
15. Vella F. Estimating models with sample selection bias: a survey. *J Hum Res* 1998; **33**: 127–172.
16. Ogundimu EO and Hutton JL. A unified approach to multilevel sample selection model. *Commun Stat – Theory Meth* 2016; **45**: 2592–2611.
17. Copas JB and Li HG. Inference for non-random samples. *J R Statist Soc B* 1997; **59**: 55–95.
18. Galimard J, Chevret S, Protopopescu C, et al. A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Stat Med* 2016. DOI: 10.1002/sim.6902.
19. Leung SF and Yu S. Collinearity and two-step estimation of sample selection models: problems, origins and remedies. *Comput Econ* 2000; **15**: 173–199.
20. Zhelonkin M, Genton MG and Ronchetti E. Robust inference in sample selection models. *J R Statist Soc B* 2016; **78**: 805–827.
21. Tanner MA and Wong WH. The calculation of posterior distributions by data augmentation (with discussion). *J Am Stat Assoc* 1987; **82**: 528–550.
22. Villa C and Walker SG. Objective prior for the number of degrees of freedom of a t distribution. *Bayesian Anal* 2014; **9**: 197–220.
23. Cameron AC and Trivedi PK. *Microeconometrics using Stata*, Revised ed. College Station, TX: Stata Press, 2010.
24. Little RJ and Zhang N. Subsample ignorable likelihood for regression analysis with missing data. *J R Statist Soc C* 2011; **60**: 591–605.
25. Little R and Yau L. Intent-to-treat analysis for longitudinal studies with dropouts. *Biometrics* 1996; **52**: 1324–1333.
26. Akacha M and Ogundimu EO. Sensitivity analyses for partially observed recurrent event data. *Pharm Stat* 2015; **15**: 4–14.
27. Smith MD. Modelling sample selection using Archimedean copulas. *Econometrics J* 2003; **6**: 99–123.

Appendix I

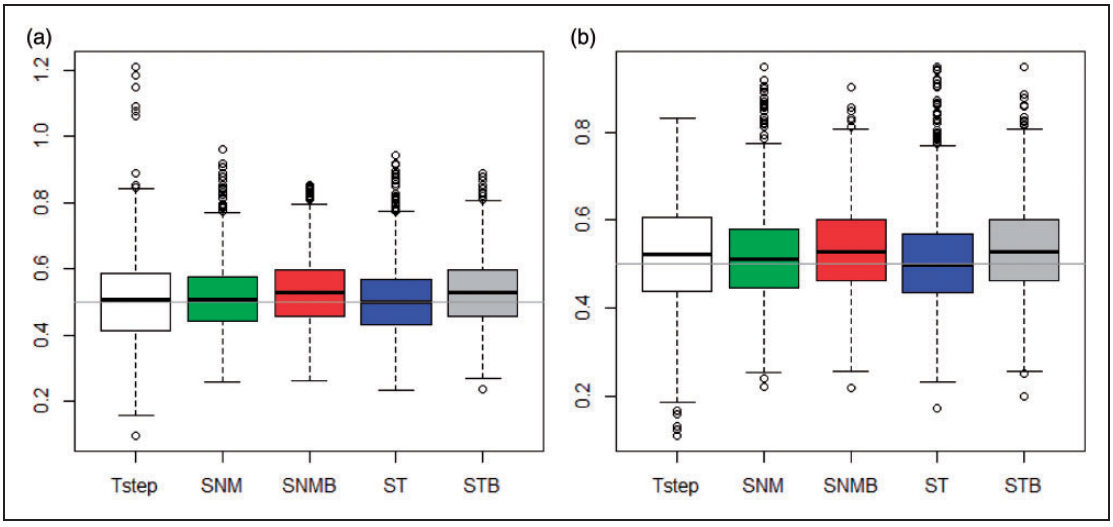

Figure 8. Boxplots for imputed outcome data with normally generated data: (a) absence of exclusion restriction; and (b) noise variables as exclusion restriction.