View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by City Research Online

City Research Online

City, University of London Institutional Repository

Citation: Gomes, M., Radice, R., Camarena Brenes, J. and Marra, G. (2019). Copula selection models for non-Gaussian responses that are missing not at random. Statistics in Medicine, 38(3), pp. 480-496. doi: 10.1002/sim.7988

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: http://openaccess.city.ac.uk/id/eprint/20471/

Link to published version: http://dx.doi.org/10.1002/sim.7988

Copyright and reuse: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

Copula selection models for non-Gaussian responses that are missing not at random

Manuel Gomes* Rosalba Radice† Jose Camarena Brenes†

Giampiero Marra‡

Abstract

Missing not at random (MNAR) data poses key challenges for statistical inference because the model of interest is typically not identifiable without imposing further (e.g., distributional) assumptions. Sample selection models have been routinely used for handling MNAR by jointly modelling the outcome and selection variables assuming that these follow a bivariate normal distribution. Recent studies have advocated parametric selection model approaches, for example estimated by multiple imputation and maximum likelihood, that are more robust to departures from the normality assumption. However, the proposed methods have been mostly restricted to a specific joint distribution (e.g., bivariate t-distribution). This paper discusses a flexible copula-based selection approach (which accommodates a wide range of non-Gaussian outcome distributions and offers great flexibility in the choice of functional form specifications for both the outcome and selection equations) and proposes a flexible imputation procedure that generates plausible imputed values from the copula selection model. A simulation study characterises the relative performance of the copula model compared with the most commonly used selection models for estimating average treatment effects with MNAR data. We illustrate the methods in the REFLUX study, which evaluates the causal effect of laparoscopic surgery compared to usual medical management on long-term quality of life in patients with reflux disease. We provide software code for implementing the proposed copula framework using the `R` package `GJRM`.

*Department of Applied Health Research, University College London, London, UK.
†Department of Economics, Mathematics and Statistics, Birkbeck, London, UK.
‡Department of Statistical Science, University College London, London, UK.

Key Words: copula, joint model, missing not at random, multiple imputation, non-Gaussian outcome, selection model, simultaneous equation model.

1 Introduction

Missing data remains a major concern in clinical and epidemiological studies. In many settings, the chances of observing the data tend to be associated with the (underlying) unobserved values of the outcome of interest. For example, health-related quality of life outcomes are increasingly used for assessing the benefits of health care interventions (NICE, 2013). However, these outcomes are typically self-reported and a key concern is that the chances of completing the quality of life questionnaire are likely to be related to patient's true health status (after adjusting for the observed data) (Mason et al., 2017). In such cases, the data is said to be missing not at random (MNAR), and non-response must be modelled together with the substantive model for the observed data.

Selection models have been commonly used to handle MNAR data in clinical and epidemiological research (e.g., Sales et al., 2004; Del Bianco & Borgoni, 2006; Alva et al., 2014), by jointly modelling the outcome and missingness models typically assuming bivariate normality. Heckman (1974) was one of the first to propose such model (Heckman selection model) using a simultaneous equation approach, where the error terms were assumed to follow a bivariate Gaussian. At that time, to circumvent some of the difficulties associated with direct likelihood maximisation, Heckman proposed a 2-stage least-squares estimation procedure (Heckman, 1979). This involved combining a probit model for the probability of observing the outcome (1st stage) with a linear regression model for the outcome (2nd stage), which was a function of the estimates obtained in the 1st stage. Although this approach is somehow robust to deviations from normality (method of moments estimator), it relies crucially on the availability of exclusion restrictions, i.e. variables that predict missingness but are unrelated to the outcome of interest (Puhani, 2000).

An alternative approach to estimating selection models is the single-step full-information maximum likelihood (FIML) which jointly estimates the outcome and missing data equations. For example, Diggle and Kenward (1994) combined a marginal model for the outcome with a logistic regression for the missing data mechanism, allowing the latter to be a function of the unobserved

outcome. The study used the Nelder-Mead optimisation algorithm, although in practice such selection models may be more flexibly estimated by MCMC techniques in a Bayesian framework (Daniels & Hogan, 2008). Alternatively, selection models can also be implemented with multiple imputation (MI) (Galimard et al., 2016). Essentially, MI imputes a set of plausible values for each missing observation, which are drawn from the posterior distribution of the missing values conditional on the observed data. To handle MNAR, the imputed values can be obtained using a selection model, such as the Heckman model, to recognise that the missing data may be related to unobserved values. A recent study suggested that the FIML and MI approaches based on the Heckman model are sensitive to departures from the assumption of Gaussian outcomes (Gomes et al., 2017).

Recent studies have considered various generalisations to address possible deviations from normality. For example, the Heckman selection model has been extended to accommodate data with heavier tails by considering a bivariate t-distribution (Marchenko & Genton, 2012; Ding, 2014; Ogundimu & Collins, 2017), whereas Zhelonkin et al. (2015) introduced a procedure for robustifying the Heckman's two-step estimator by using M-estimators of Mallows' type for both steps. However, most of these approaches are restricted to a specific joint distribution for the selection and outcome processes, and the extension to non-Gaussian outcomes may not be easy. Semi-parametric (e.g., Lee, 2008; Chib et al., 2009; Newey, 2009) and non-parametric (e.g., Das et al., 2003; Chen & Zhou, 2010) selection models have been proposed, but these have not permeated practice as their implementation may be challenging in the regression context (Pigini, 2015).

This paper addresses some of these concerns by discussing a copula-based selection framework which accommodates a wide range of non-Gaussian distributions, not restricted to the exponential family. The link function for the copula's selection equation is allowed to be different from the classic probit, and the parameters of the copula and marginal distributions can be made dependent on several types of covariate effects (e.g., linear, non-linear, random and spatial effects). In principle, the approach allows for any parametric continuous marginal outcome distribution and link function for the selection equation, and several dependence structures between the margins as implied by copulae. While the copula approach is fully parametric, it is computationally more tractable than semi/non-parametric approaches, particularly in a regression context, and it still al-

lows the researcher to assess the sensitivity of results to different modelling assumptions. This article then introduces a flexible imputation procedure that generates plausible imputed values from the copula selection model. The proposed methods can be easily implemented using the `gjrm()` function in the `R` package `GJRM` (Marra & Radice, 2017b), and we provide software code to encourage the uptake of the methods (see the Supplementary Material online).

The plan for the remainder of the paper is as follows. In Section 2, we describe our motivating example. Section 3 describes the original Heckman selection model and Section 4 discusses the copula selection model framework and the proposed imputation procedure. Section 5 presents the design and results of a simulation study that evaluates the relative merits of the copula approach compared to commonly used selection models across different MNAR settings. Section 6 reports the results from applying the methods to our case-study, and Section 7 discusses the findings in light of previous methodological studies and provides directions for further research.

2 Motivating example

Gastro-Oseophageal Reflux Disease (GORD) develops when reflux of the stomach acid cause troublesome symptoms or complications which adversely affect patients' well-being. About 20-30% of adult 'Western' populations experience heartburn or reflux intermittently, and many of these patients are often treated with Proton Pump Inhibitors (PPIs) to suppress acid reflux. While PPIs are generally effective, there is the concern that long-term acid suppression with PPIs may be associated with increased risk of chronic hypergastrinaemia and gastric cancer. An alternative to long-term medication is to have laparoscopic surgery, which is a minimally invasive procedure but carries some risk of side effects. Our motivating example, the REFLUX study, compares these interventions for treating patients with GORD in the UK. The REFLUX study was a pragmatic randomised controlled trial, which included two components: a randomised component in which patients were randomised to surgery and medical management, and a non-randomised comparison (according to patient preferences) between a policy of offering early surgery and a policy of continued medical management; full details can be found in Grant et al. (2013). Our study focused on the latter design, because it better represented the target population of prime interest to

policy makers. The non-randomised preference arms included 261 patients for surgery and 192 in medical management. Self-reported quality of life was measured at baseline, 3 months and then annually up to 5 years, using the EQ-5D-3L (measure anchored on a scale that includes 0 - death, and 1 - perfect health) questionnaire (EuroQol, 1990). The primary outcome of interest was quality-adjusted life-years (QALYs), an outcome that combines both quality of life and survival, measured at 5 years.

A significant proportion of patients failed to complete the EQ-5D-3L questionnaires, resulting in missing 5-year QALYs for 55% (106 out of 192) of patients in the medical management group and 48% (125 out of 261) in the surgery group. Baseline covariate information was mostly complete. A key concern in this study was that the relative effectiveness of laparoscopic surgery vs medical management may be sensitive to alternative assumptions about the missing data. More specifically, trial coordinators suspected that patients in worse health in the control group lost interest in the study (because the intervention was not working for them), and hence were less likely to return quality of life questionnaires or answer the phone.

Table 1 provides a description of the main baseline covariates and their association with the missing data indicator. This included key prognostic factors of the health outcome, which had large imbalances between the intervention groups and were included in the substantive model, and a set of variables that were predictive of missingness but anticipated to be conditionally independent of the outcome, i.e. met the criteria for the 'exclusion restriction'. Missing data were strongly associated with some key prognostic factors such as age and activity score, but weakly associated with other the exclusion restrictions (views about medicine). This raises some pertinent questions for the choice of selection model approach; for example, whether the relative merits of alternative selection models differ according to the strength of association between the exclusion restriction and non-response.

In addition, QALYs are typically left-skewed and often include negative values (reflecting health states judged worse than death). This unusual distributional shape raises significant challenges to sample selection models, and existing approaches (which assume a bivariate normal or t-distribution) are unlikely to be plausible in this context. Figure 1 suggests that a Gumbel distribution, as defined in Rigby & Stasinopoulos (2005), provides a reasonable fit to QALYs.

Table 1: Descriptive statistics of the baseline characteristics and their correlation with non-response, and the quality-adjusted life year (QALY) outcome.

Variable	Medical management (N=192)	Surgery (N=261)	Standardised difference (%)	Correlation with non-response[#]
Baseline prognostic factors				
Male	111 (58%)	170 (65%)	15.1	0.06
Age	49.9 (11.8)	44.4 (12.0)	45.9	0.18***
BMI (kg/m2)	27.4 (4.1)	27.7 (3.9)	7.5	-0.05
REFLUX quality-of-life	76.1 (19.8)	55.9 (22.8)	94.7	0.06
Baseline EQ-5D-3L	0.75 (0.22)	0.68 (0.26)	27.5	0.06
Heart burn score	72.2 (20.9)	49.4 (24.2)	101	0.08
Gastro 1 symptom score	59.3 (22.2)	47.2 (21.1)	55.8	0.06
Gastro 2 symptom score	82.9 (17.5)	75.9 (21.7)	35.3	-0.01
Nausea symptom score	89.4 (13.5)	77.0 (19.7)	73.8	0.12
Activity score	86.6 (12.8)	74.5 (15.9)	83.4	0.02**
Previous hiatus hernia	73 (38%)	76 (29%)	18.9	-0.09**
Smoker	39 (20%)	71 (27%)	16.2	0.1
Asthma	36 (19%)	30 (11%)	20.4	-0.04
Duration of REFLUX symptoms (days)	45.9 (53.7)	55.9 (67.3)	16.4	-0.03
Employment status				0.01
Full-time	101 (53%)	171 (66%)	26.5	
Part-time	20 (10%)	35 (13%)	9.3	
School leaving age				0.10**
16 year or younger	107 (56%)	154 (59%)	6.6	
20 years or older	40 (21%)	44 (17%)	10.2	
Exclusion restriction variables				
General views about medicine				
Doctors use too many medicines	32 (17%)	61 (23%)	16.8	0.03
People should pause treatments	42 (22%)	76 (29%)	16.7	-0.01
Medicines are addictive	22 (11%)	39 (15%)	10.3	-0.07*
Natural remedies are safer	30 (16%)	38 (15%)	3	0.07
Medicines do more harm than good	4 (2%)	5 (2%)	1.1	-0.02
All medicines are poisons	13 (7%)	7 (3%)	19.3	-0.03
Doctors trust medicines too much	26 (14%)	51 (20%)	16.2	0.09**
Doctors should spend more time with patients	69 (36%)	94 (36%)	0.2	0.06
Outcome	(N=106)	(N=125)		
QALYs (5-year)	3.594 (0.83)	3.777 (0.94)		

Notes: Continuous covariates (and outcome) reported as Mean (SD) and binary covariates as N (%). Belief variables are dichotomised: 1 if patient agrees or strongly agrees with the statement, 0 otherwise.

[#] Pearson correlation coefficient between each variable and the binary missing data indicator. Statistical significance is based on the corresponding coefficients from the logistic model: *p<0.1, **p<0.05 ***p<0.001

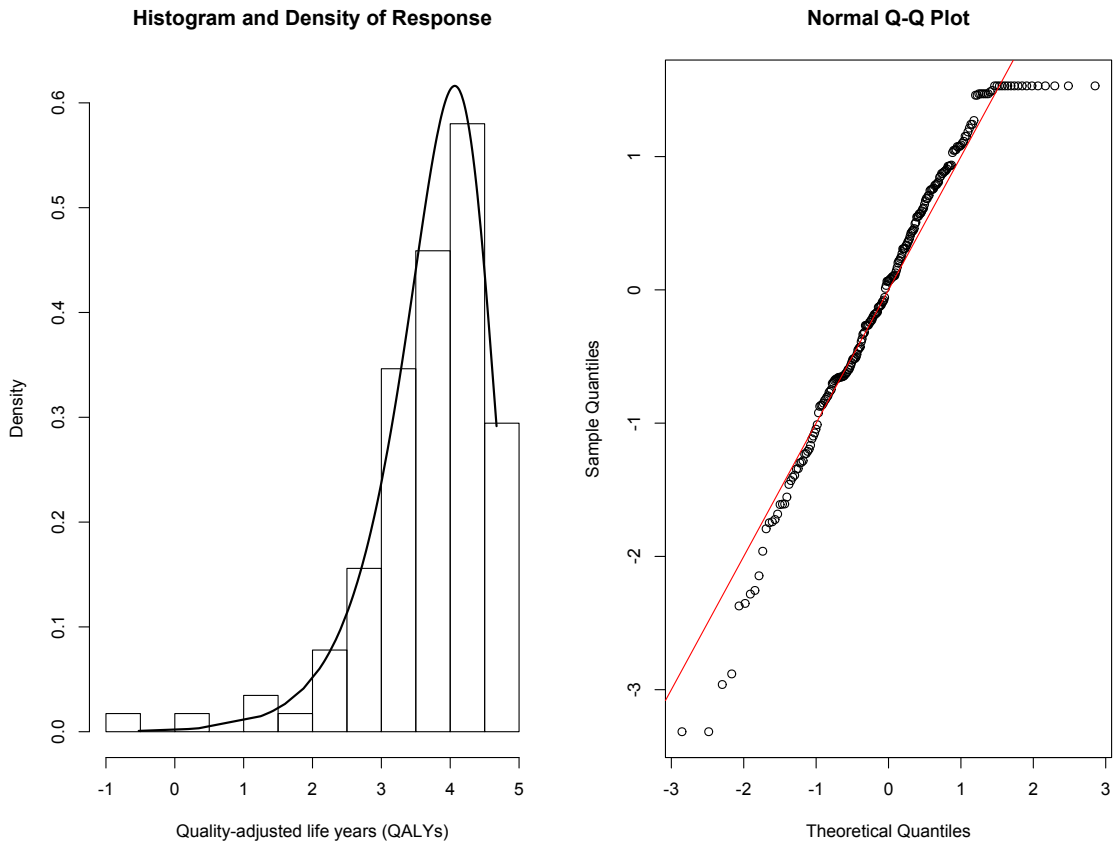

Figure 1: REFLUX study: Histogram, kernel density estimate and normal Q-Q plot of the 5-year quality-adjusted life years obtained from a Gumbel fit.

3 Classic sample selection model

In this section, we briefly introduce the classical sample selection model, assuming bivariate normality. In the sample selection problem, the outcome of interest is observed only for a restricted non-randomly selected sample of the population, i.e. data are MNAR. Using a utilitarian framework, we assume that Y_{2i}^* is a latent continuous random variable of primary interest, for $i = 1, \ldots, n$ where n denotes the sample size; and we represent *selection* using the pair (Y_{1i}, Y_{2i}) such that $Y_{i1} \in \{0, 1\}$ and $Y_{2i} = Y_{1i}Y_{2i}^*$. The missing data indicator, Y_{1i} is a Bernoulli random variable indicating whether or not the outcome is observed and Y_{1i}^* is the underlying latent continuous variable such that $Y_{1i} = \mathbf{1}(Y_{1i}^* > 0)$, where $\mathbf{1}(\cdot)$ is the indicator function taking value 1 if Y_{2i} is observed and 0 otherwise. In addition, let $\mathbf{X}_{2i}$ be the set of prognostic variables and $\mathbf{X}_{1i}$ the set of predictors of the probability of observing the outcome ($\mathbf{X}_{1i}$ should include $\mathbf{X}_{2i}$). Then we can write the classical selection model (Heckman, 1974) as:

$$\begin{aligned} Y_{1i}^* &= \mathbf{X}_{1i}^\mathsf{T}\boldsymbol{\beta}_1 + e_{1i} \\ Y_{2i} &= \mathbf{X}_{2i}^\mathsf{T}\boldsymbol{\beta}_2 + e_{2i} \end{aligned}, \qquad \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \theta\sigma_2 \\ & \sigma_2^2 \end{pmatrix} \right]. \tag{1}$$

For identification, the variance of the latent variable, σ_1^2, is fixed to 1. $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the vector of regression coefficients in the outcome (Y_2) and missingness (Y_1^*) models, respectively. The model allows for the dependence between the outcome and selection equations (MNAR mechanism) through correlation parameter θ.

The log-likelihood for the full sample is a combination of the likelihood function for the individuals for whom Y_2 is observed, $f(Y_{2i})P(Y_{1i}^* > 0|Y_{2i}, \mathbf{X}_{1i}, \mathbf{X}_{2i})$, and for the individuals for whom Y_2 is missing (marginal probability that $Y_1^* \leq 0$), $P(e_{1i} \leq -\mathbf{X}_{1i}^\mathsf{T}\boldsymbol{\beta}_1)$. That is,

$$\sum_{Y_1^* \leq 0} \log\left\{1 - \Phi\left(\mathbf{X}_{1i}^\mathsf{T}\boldsymbol{\beta}_1\right)\right\} + \sum_{Y_1^* > 0} \left[-\log(\sigma_2) + \log\left\{ \phi\left(\frac{Y_{2i} - \mathbf{X}_{2i}^\mathsf{T}\boldsymbol{\beta}_2}{\sigma_2} \right) \right\} + \log\left\{ \Phi\left(\frac{\mathbf{X}_{1i}^\mathsf{T}\boldsymbol{\beta}_1 + \frac{\theta}{\sigma_2}(Y_{2i} - \mathbf{X}_{2i}^\mathsf{T}\boldsymbol{\beta}_2)}{\sqrt{1-\theta^2}} \right) \right\} \right],$$

where ϕ is the standard normal density and Φ is the standard normal cumulative distribution function. Alternatively, Heckman (1979) proposed a two-step estimator derived from the conditional

expectation of the observed data, which can be expressed as follows:

$$E(Y_{2i}|\mathbf{X}_{1i}, Y_{1i}^* > 0) = \mathbf{X}_{2i}^{\mathsf{T}}\boldsymbol{\beta}_2 + \theta\sigma_2\lambda_i, \qquad \lambda_i = \frac{\phi(\mathbf{X}_{1i}^{\mathsf{T}}\boldsymbol{\beta}_1)}{\Phi(\mathbf{X}_{1i}^{\mathsf{T}}\boldsymbol{\beta}_1)}, \tag{2}$$

where λ is denominated as the inverse Mills ratio. To estimate the parameters of interest (β_2) the two-step estimation approach involves: 1) Regressing Y_1 on $\mathbf{X}_1$ (using a probit model) in the full sample to obtain $\hat{\boldsymbol{\beta}}_1$ and construct $\hat{\lambda}_i$; 2) Obtain an estimate of $\boldsymbol{\beta}_2$ from the following linear model (on the observed sample): $Y_{2i} = \mathbf{X}_{2i}^{\mathsf{T}}\boldsymbol{\beta}_2 + \beta_\lambda\hat{\lambda}_i + e_{2i}$.

4 Copula selection model

4.1 Likelihood

Let us define the cumulative distribution function (cdf) of Y_{mi}^* as $F_m\left(y_{mi}^*\right) = P\left(Y_{mi}^* \leq y_{mi}^*\right)$, for $m = 1, 2$, and the joint cdf of (Y_{1i}^*, Y_{2i}^*) as $F\left(y_{1i}^*, y_{2i}^*\right) = P\left(Y_{1i}^* \leq y_{1i}^*, Y_{2i}^* \leq y_{2i}^*\right)$. Let us also denote $y_{11}, \ldots, y_{1n}$ and $y_{21}, \ldots, y_{2n}$ as the sets of observations generated on Y_1 and Y_2, respectively. Given the observation rules described above and for a random sample of n observations the likelihood function for the sample selection model can be written as

$$L = \prod_{i=1}^{n} F_1\left(0\right)^{1-y_{1i}} \left\{ f_2\left(y_{2i}\right) - \frac{\partial F\left(0, y_{2i}\right)}{\partial y_{2i}} \right\}^{y_{1i}},$$

where $f_2\left(y_{2i}\right) = \partial F_2\left(y_{2i}\right)/\partial y_{2i}$. Note that in the above the presence of covariates and parameters have been suppressed for the sake of notational convenience.

4.2 Copula and marginal distributions

To simplify the notation, and without loss of generality, let us drop the observation index i. Also let $F(y_1^*, y_2^*|\boldsymbol{\delta})$ denote the joint cdf of Y_1^* and Y_2^* conditional on $\mathbf{X}$ (representing a generic set of covariates). It is possible to show that (Sklar, 1973)

$$F(y_1^*, y_2^*|\boldsymbol{\delta}) = C\left(F_1(y_1^*|\mu_1, \sigma_1, \nu_1), F_2(y_2^*|\mu_2, \sigma_2, \nu_2); \theta\right), \tag{3}$$

where $\boldsymbol{\delta} = (\mu_1, \sigma_1, \nu_1, \mu_2, \sigma_2, \nu_2, \theta)^\intercal$, $F_m(y_m^*|\mu_m, \sigma_m, \nu_m)$ is a conditional marginal cdf with distributional parameters μ_m, σ_m and ν_m, $C(\cdot, \cdot)$ is a uniquely defined two-place copula function which does not depend on the marginal cdfs, and θ is an association copula parameter representing the dependence between the two marginals. Our framework allows one to relate all marginal distribution and dependence parameters to additive predictors η's (which essentially contain regression coefficients and $\mathbf{X}$) via known monotonic link functions which ensure that the restrictions on the parameter spaces are maintained. This offers a lot of flexibility in the choice of functional form specifications for the selection and outcome equations, and we refer the reader to Marra & Radice (2017a) for further details and some examples of covariate effects that can be considered. The above result shows that a joint cdf can be conveniently expressed in terms of arbitrary univariate marginal cdfs and a function C that binds them together. The copulae (as well as rotated versions of these) implemented in `GJRM` are reported in Table 2 which also shows the relation between θ and the Kendall's τ coefficient which is a more interpretable measure of association that lies in the customary range $[-1, 1]$.

The marginal distributions of Y_1^* and Y_2^* are specified through parametric cdfs and densities denoted as $F_m(y_m^*|\mu_m, \sigma_m, \nu_m)$ and $f_m(y_m^*|\mu_m, \sigma_m, \nu_m)$, for $m = 1, 2$, where μ_m, σ_m and ν_m represent sometimes location, scale and shape (Rigby & Stasinopoulos, 2005). For Y_1^* we have considered the Gaussian, logistic and Gumbel distributions with $\mu_1 = 0$ and $\sigma_1 = 1$ (which yield `"probit"`, `"logit"` and `"cloglog"` link functions, respectively), whereas for Y_2^* we have considered the two and three parameter distributions described in Table 2 of Marra & Radice (2017a). These are the normal (`"N"`), log-normal (`"LN"`), Gumbel (`"GU"`), reverse Gumbel (`"rGU"`), logistic (`"LO"`), Weibull (`"WEI"`), inverse Gaussian (`"iG"`), gamma (`"GA"`), Dagum (`"DAGUM"`), Singh-Maddala (`"SM"`), beta (`"BE"`), and Fisk (`"FISK"`) distributions. Note that the adopted notation reflects the fact that we have considered two and three parameter distributions for the outcome equation. However, our framework can in principle accommodate distributions with any number of parameters. Argument `margins` of `gjrm()` in `GJRM` allows the user to employ the desired link function and outcome distribution and can be set to any of the values indicated above. For example, `margins = c("cloglog", "GU")`.

Copula	$C(p_1, p_2; \theta)$	Range of θ	Link	Kendall's τ
AMH (`"AMH"`)	$\frac{p_1 p_2}{1-\theta(1-p_1)(1-p_2)}$	$\theta \in [-1, 1]$	$\tanh^{-1}(\theta)$	$-\frac{2}{3\theta^2}\left\{\theta + (1-\theta)^2 \log(1-\theta)\right\} + 1$
Clayton (`"C0"`)	$\left(p_1^{-\theta} + p_2^{-\theta} - 1\right)^{-1/\theta}$	$\theta \in (0, \infty)$	$\log(\theta)$	$\frac{\theta}{\theta+2}$
FGM (`"FGM"`)	$p_1 p_2 \left\{1 + \theta(1-p_1)(1-p_2)\right\}$	$\theta \in [-1, 1]$	$\tanh^{-1}(\theta)$	$\frac{2}{9}\theta$
Frank (`"F"`)	$-\theta^{-1} \log\left\{1 + (\exp\left\{-\theta p_1\right\} - 1)(\exp\left\{-\theta p_2\right\} - 1)/(\exp\left\{-\theta\right\} - 1)\right\}$	$\theta \in \mathbb{R} \setminus \{0\}$	$-$	$1 - \frac{4}{\theta}\left[1 - D_1(\theta)\right]$
Hougaard (`"HO"`)	$\exp\left[-\left\{(-\log p_1)^{\frac{1}{\theta}} + (-\log p_2)^{\frac{1}{\theta}}\right\}^{\theta}\right]$	$\theta \in (0, 1)$	$\log\left(\frac{\theta}{1-\theta}\right)$	$1 - \theta$
Gaussian (`"N"`)	$\Phi_2\left(\Phi^{-1}(p_1), \Phi^{-1}(p_2); \theta\right)$	$\theta \in [-1, 1]$	$\tanh^{-1}(\theta)$	$\frac{2}{\pi}\arcsin(\theta)$
Gumbel (`"G0"`)	$\exp\left[-\left\{(-\log p_1)^{\theta} + (-\log p_2)^{\theta}\right\}^{1/\theta}\right]$	$\theta \in [1, \infty)$	$\log(\theta - 1)$	$1 - \frac{1}{\theta}$
Joe (`"J0"`)	$1 - \left\{(1-p_1)^{\theta} + (1-p_2)^{\theta} - (1-p_1)^{\theta}(1-p_2)^{\theta}\right\}^{1/\theta}$	$\theta \in (1, \infty)$	$\log(\theta - 1)$	$1 + \frac{4}{\theta^2} D_2(\theta)$
Plackett (`"PL"`)	$\left(Q - \sqrt{R}\right) / \left\{2(\theta - 1)\right\}$	$\theta \in (0, \infty)$	$\log(\theta)$	$-$
Student-t (`"T"`)	$t_{2,\zeta}\left(t_\zeta^{-1}(p_1), t_\zeta^{-1}(p_2); \zeta, \theta\right)$	$\theta \in [-1, 1]$	$\tanh^{-1}(\theta)$	$\frac{2}{\pi}\arcsin(\theta)$

Table 2: Definition of copulae implemented in `GJRM`, with corresponding parameter range of association parameter θ, link function of θ, and relation between Kendall's τ and θ. $\Phi_2(\cdot, \cdot; \theta)$ denotes the cumulative distribution function (cdf) of a standard bivariate normal distribution with correlation coefficient θ, and $\Phi(\cdot)$ the cdf of a univariate standard normal distribution. $t_{2,\zeta}(\cdot, \cdot; \zeta, \theta)$ indicates the cdf of a standard bivariate Student-t distribution with correlation θ and fixed $\zeta \in (2, \infty)$ degrees of freedom, and $t_\zeta(\cdot)$ denotes the cdf of a univariate Student-t distribution with ζ degrees of freedom. $D_1(\theta) = \frac{1}{\theta}\int_0^\theta \frac{t}{\exp(t)-1} dt$ is the Debye function and $D_2(\theta) = \int_0^1 t \log(t)(1-t)^{\frac{2(1-\theta)}{\theta}} dt$. Quantities Q and R are given by $1 + (\theta - 1)(p_1 + p_2)$ and $Q^2 - 4\theta(\theta - 1)p_1 p_2$, respectively. The Kendall's τ for `"PL"` is computed numerically as no analytical expression is available. Argument `BivD` of `gjrm()` in `GJRM` allows the user to employ the desired copula function and can be set to any of the values within brackets next to the copula names in the first column; for example, `BivD = "J0"`. For Clayton, Gumbel and Joe, the number after the capital letter indicates the degree of rotation required: the possible values are `0`, `90`, `180` and `270`.

4.3 Some estimation details and further considerations

The log-likelihood of the copula sample selection model can be written as

$$\ell(\boldsymbol{\delta}) = \sum_{i=1}^{n} (1 - y_{1i}) \log \left\{ F_1 \left(0 \right) \right\} + y_{1i} \log \left[f_2(y_{2i}|\mu_{2i}, \sigma_{2i}, \nu_{2i}) \left(1 - h_i \right) \right],$$

where

$$h_i = \frac{\partial C \left(F_1 \left(0 \right), F_2(y_{2i}|\mu_{2i}, \sigma_{2i}, \nu_{2i}) \right)}{\partial F_2(y_{2i}|\mu_{2i}, \sigma_{2i}, \nu_{2i})}. \tag{4}$$

The distributional parameters are defined as $\mu_1 = g_{\mu_1}^{-1}(\eta_{\mu_1 i})$, $\mu_{2i} = g_{\mu_2}^{-1}(\eta_{\mu_2 i})$, $\sigma_{2i} = g_{\sigma_2}^{-1}(\eta_{\sigma_2 i})$, $\nu_{2i} = g_{\nu_2}^{-1}(\eta_{\nu_2 i})$ and $\theta_i = g_{\theta}^{-1}(\eta_{\theta i})$, where the g's are link functions which ensure that the restrictions on the parameter spaces are maintained. Parameter vector $\boldsymbol{\delta}$ is made up of the distributional parameters which in turn contain $\boldsymbol{\beta}_{\mu_1}$, $\boldsymbol{\beta}_{\mu_2}$, $\boldsymbol{\beta}_{\sigma_2}$, $\boldsymbol{\beta}_{\nu_2}$ and $\boldsymbol{\beta}_{\theta}$ (the coefficient vectors associated with $\eta_{\mu_1 i}$, $\eta_{\mu_2 i}$, $\eta_{\sigma_2 i}$, $\eta_{\nu_2 i}$ and $\eta_{\theta i}$). Parameter estimation is achieved using an extended version of the efficient and stable trust region algorithm introduced by Marra & Radice (2017a). In particular, the algorithm uses the analytical score and Hessian of $\ell(\boldsymbol{\delta})$, which have been derived in a modular fashion. For instance, the score vector is made up of

$$\frac{\partial \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\beta}_{\mu_1}} = \sum_{i=1}^{n} \left\{ \frac{1 - y_{1i}}{F_1 \left(0 \right)} - \frac{y_{1i}}{1 - h_i} \frac{\partial h_i}{\partial F_1 \left(0 \right)} \right\} \frac{\partial F_1 \left(0 \right)}{\partial \eta_{\mu_1 i}} \mathbf{X}_{\mu_1 i},$$

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\beta}_{\mu_2}} = \sum_{i=1}^{n} \frac{y_{1i}}{f_2(y_{2i}|\mu_{2i}, \sigma_{2i}, \nu_{2i})} \Bigg\{ & (1 - h_i) \frac{\partial f_2(y_{2i}|\mu_{2i}, \sigma_{2i}, \nu_{2i})}{\partial \mu_{2i}} - \\ & f_2(y_{2i}|\mu_{2i}, \sigma_{2i}, \nu_{2i}) \frac{\partial h_i}{\partial F_2(y_{2i}|\mu_{2i}, \sigma_{2i}, \nu_{2i})} \frac{\partial F_2(y_{2i}|\mu_{2i}, \sigma_{2i}, \nu_{2i})}{\partial \mu_{2i}} \Bigg\} \frac{\partial \mu_{2i}}{\partial \eta_{\mu_2 i}} \mathbf{X}_{\mu_2 i}, \end{aligned} \tag{5}$$

$\partial \ell(\boldsymbol{\delta}) / \partial \boldsymbol{\beta}_{\sigma_2}$ and $\partial \ell(\boldsymbol{\delta}) / \partial \boldsymbol{\beta}_{\nu_2}$ (whose expressions are not reported here since they very similar to (5)), and

$$\frac{\partial \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\beta}_{\theta}} = \sum_{i=1}^{n} \left\{ \frac{y_{1i}}{h_i - 1} \frac{\partial h_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \eta_{\theta_i}} \right\} \mathbf{X}_{\theta i},$$

where $\mathbf{X}_{\mu_1 i}$, $\mathbf{X}_{\mu_2 i}$ and $\mathbf{X}_{\theta i}$ represent the covariate vectors associated with their respective additive predictors. Looking at equation (5), we see that there are two components which depend on the chosen copula, four terms that are dependent on the marginal distribution, and one derivative whose form will depend on the adopted link function between μ_{2i} and $\eta_{\mu_2 i}$. However, the main

structure of the equation will be unaffected by the specific choices made. This means that it will be easy to extend our algorithm to other copulae and marginal distributions not considered in this work, as long as their cdfs and probability density functions are known and their derivatives with respect to their parameters exist. As shown in Marra & Radice (2017a), if the equations of the joint model contain flexible functions of covariates then a penalty term is employed in estimation, in which case the approach becomes penalised likelihood-based.

The proposed approach offers a lot of flexibility for modelling MNAR data. For choosing a suitable copula function, link function and response distribution, as well as selecting covariates in the model's additive predictors if required, we recommend using the Akaike information criterion (AIC) and/or Bayesian information criterion (BIC), normal Q-Q plots of normalized quantile residuals and hypothesis testing (Marra & Radice, 2017a). Note that as far as the choice of link function is concerned, as pointed out in Section 4.2, we have considered the Gaussian, logistic and Gumbel distributions (with location and scale parameters equal to 0 and 1) which yield probit, logit and cloglog links. In this case, since the response is binary, residual analysis would not be informative unless residuals could be grouped in a meaningful way (e.g., Collett, 2002). We therefore recommend avoiding looking at residual plots for the selection equation and doing a sensitivity analysis using the links available.

Reliable point-wise 'confidence' intervals for any linear and non-linear function of the model coefficients are obtained using the posterior distribution $\boldsymbol{\delta} \dot{\sim} \mathcal{N}(\hat{\boldsymbol{\delta}}, -\hat{\boldsymbol{\mathcal{H}}}_p^{-1})$, where $\boldsymbol{\mathcal{H}}_p$ is the model's Hessian. The rationale for using this result post-estimation is provided in Marra & Wood (2012) and Marra & Radice (2017a), for instance. These papers show that using the above posterior distribution yields confidence intervals with better frequentist properties than those obtained using a frequentist approach itself. Other advantages of using the Bayesian result are that the distribution of non-linear functions of the model parameters can easily be obtained by posterior simulation and that the resulting distribution need not be symmetric. Note that if the model's additive predictors do not include terms which require penalization during model fitting (like smooth functions of continuous covariates), then the expressions for the frequentist and Bayesian covariance matrices are identical.

The proposed copula models can be easily fitted in R using the package `GJRM` (Marra &

Radice, 2017b). For instance,

```
fl <- list(y1 ~ x1 + x2, y2 ~ x1)
md <- gjrm(fl, Model = "BSS", margins = c("logit", "WEI"), BivD = "PL")
```

where `fl` is a list containing the selection and outcome equations, respectively, and `"BSS"` stands for bivariate sample selection model.

4.4 Multiple imputation

In this section we introduce a flexible imputation procedure that generates plausible imputed values from the proposed copula selection model for handling non-Gaussian (continuous) outcomes that are MNAR. The general idea of MI can be summarised in three steps: 1) generate more than one complete data set by filling each missing value with draws from the posterior distribution of the missing data given the observed data; 2) apply the model of interest to each of the imputed data sets; 3) combine the resulting parameter estimates of interest, for example average treatment effects and standard errors, obtained from the analyses of the different imputed data sets. The standard implementation of MI is valid under the MAR assumption. However, MI can also be used when the missing data mechanism is suspected to be MNAR (Rubin, 1987; Schafer, 1999). Recent studies have considered the MI approach for handling MNAR in the context of selection models, for example, assuming bivariate normality under the classical selection model (Galimard et al., 2016), and bivariate $t-$distribution (Ogundimu & Collins, 2017). The proposed MI approach allows us to model more flexibly the selection and outcome models in that several types of marginal and bivariate distributions can be employed when specifying the joint model.

Given the flexibility afforded by copulae, we considered a joint modelling approach to MI, although fully conditional specification could also be adopted; the equivalence between the two approaches are discussed elsewhere (Liu et al., 2014; Hughes et al., 2014). From a Bayesian perspective, we can consider the missing values as a set of additional (unknown) parameters and derive an imputation model that corresponds to the posterior predictive distribution of the missing

data given the observed data

$$f(y_2|y_1 = 0) = \int f(y_2|y_1 = 0, \boldsymbol{\delta})\ f(\boldsymbol{\delta}|y_1, y_2)\ d\boldsymbol{\delta}, \tag{6}$$

where $f(y_2|y_1 = 0, \boldsymbol{\delta})$ represents the conditional distribution of the missing values, and $f(\boldsymbol{\delta}|y_1, y_2)$ is the posterior distribution of $\boldsymbol{\delta}$. In order to obtain plausible imputed values from (6), we first draw $\tilde{\boldsymbol{\delta}}$ from $f(\boldsymbol{\delta}|y_1, y_2)$ and then draw $\tilde{y}$ from $f(y_2|y_1 = 0, \boldsymbol{\delta} = \tilde{\boldsymbol{\delta}})$. Little & Rubin (1987) suggested using the asymptotic distribution of the maximum likelihood estimates since it propagates the uncertainty in the estimated coefficients $\hat{\boldsymbol{\delta}}$, and are typically readily available. Here, we employ $\boldsymbol{\delta} \ \dot{\sim}\ \mathcal{N}(\hat{\boldsymbol{\delta}}, -\hat{\boldsymbol{\mathcal{H}}}_p^{-1})$ which, as mentioned in Section 4.3, can provide more reliable confidence intervals in our context. The conditional density of the missing outcomes can be derived from joint distribution (3). That is,

$$\begin{aligned} f(y_2|y_1 = 0; \boldsymbol{\delta}) &= \frac{\partial F(y_2|y_1 = 0; \boldsymbol{\delta})}{\partial y_2} = \frac{\partial}{\partial y_2}\left[\frac{F(0, y_2; \boldsymbol{\delta})}{F_1(0)}\right] \\ &= \frac{1}{F_1(0)} \frac{\partial F(0, y_2; \boldsymbol{\delta})}{\partial y_2} \\ &= \frac{1}{F_1(0)} \frac{\partial C(F_1(0), F_2(y_2|\mu_2, \sigma_2, \nu_2)}{\partial F_2(y_2|\mu_2, \sigma_2, \nu_2)} \frac{\partial F_2(y_2|\mu_2, \sigma_2, \nu_2)}{\partial y_2} \\ &= \frac{1}{F_1(0)}\ h\ f_2(y_2|\mu_2, \sigma_2, \nu_2). \end{aligned} \tag{7}$$

where h is defined in equation (4). We propose sampling from $f(y_2|y_1 = 0; \boldsymbol{\delta})$ using an acceptance/rejection approach (e.g., Robert & Casella, 2005). This method requires the use of a known distribution (also called instrumental) that has the same support as that of the target distribution we need to draw from. In our context, it is sensible to employ $f_2(y_2^*|\mu_2, \sigma_2, \nu_2)$ as the instrumental probability density function (pdf). The algorithm proceeds by drawing a candidate $\tilde{y}$ from the instrumental pdf and then accept it with probability proportional to $f(\tilde{y}|y_1 = 0; \boldsymbol{\delta})/f(\tilde{y}|\mu_2, \sigma_2, \nu_2)$. The constant of proportionality, given by $1/M$, corresponds to the probability of acceptance and M must satisfy the inequality $t(\tilde{y}) = f(\tilde{y}|y_1 = 0; \boldsymbol{\delta})/f(\tilde{y}|\mu_2, \sigma_2, \nu_2) \leq M$.

M is chosen by maximising $t(\tilde{y})$ using a trust-region algorithm which typically provides more accurate results than standard alternatives (e.g., Nocedal & Wright, 2006). To prevent the algorithm from searching the solution outside the domain of the objective function, we transform the

candidate $\tilde{y}$, using a differentiable and monotone function, such that its range is unbounded. For example, in the case of distributions with a positive support we use transformation $\breve{y} = \log(\tilde{y})$. The use of such algorithm requires computing first and second derivatives. These are given by

$$\frac{dt(\tilde{y})}{d\breve{y}} = \frac{1}{F_1(0)}\, h'\, f_2(\tilde{y}|\mu_2, \sigma_2, \nu_2)\frac{d\tilde{y}}{d\breve{y}}$$

and

$$\frac{d^2t(\tilde{y})}{d\breve{y}^2} = \frac{1}{F_1(0)}\left[h''\left(f_2(\tilde{y}|\mu_2, \sigma_2, \nu_2)\right)^2 + h'\frac{\partial f_2(\tilde{y}|\mu_2, \sigma_2, \nu_2)}{\partial \tilde{y}}\right]\left(\frac{d\tilde{y}}{d\breve{y}}\right)^2 + \frac{dt(\tilde{y})}{\tilde{y}}\frac{d^2\tilde{y}}{d\breve{y}^2},$$

where h' and h'' are the first and second derivative of h which are defined as

$$h' = \frac{\partial^2 C\left(F_1\left(0\right), F_2(\tilde{y}|\mu_2, \sigma_2, \nu_2)\right)}{\partial F_2(\tilde{y}|\mu_2, \sigma_2, \nu_2)^2} \text{ and } h'' = \frac{\partial h'}{\partial F_2(\tilde{y}|\mu_2, \sigma_2, \nu_2)}.$$

Having obtained all the components needed to sample from (7), the algorithm proposed for imputation reduces to two steps: i) draw $\tilde{\boldsymbol{\delta}}$ from $\mathcal{N}(\hat{\boldsymbol{\delta}}, -\hat{\boldsymbol{\mathcal{H}}}_p^{-1})$; ii) draw a candidate $\tilde{y}$ from $f(y_2|y_1 = 0, \boldsymbol{\delta} = \tilde{\boldsymbol{\delta}})$.

The last step in the MI process consists of combining the estimates obtained from the analysis of all imputed data sets, for example, by applying what is commonly known as Rubin's rules (Rubin, 1987). Let $\hat{\boldsymbol{\delta}}^{(1)}, \ldots, \hat{\boldsymbol{\delta}}^{(n_c)}$ denote the estimates for the n_c completed data sets with corresponding variance/covariance matrices represented by $\hat{\boldsymbol{V}}^{(1)}, \ldots, \hat{\boldsymbol{V}}^{(n_c)}$. The MI estimate is given by $\hat{\boldsymbol{\delta}}_{\text{MI}} = \frac{1}{n_c}\sum_{j=1}^{n_c}\hat{\boldsymbol{\delta}}^{(j)}$ (average across the multiple estimates), whereas the corresponding variance/covariance matrix is made of two components: the within-imputation variance $\hat{\boldsymbol{W}}_{\text{MI}} = \frac{1}{n_c}\sum_{j=1}^{n_c}\hat{\boldsymbol{V}}^{(j)}$ and the between-imputation variance $\hat{\boldsymbol{B}}_{\text{MI}} = \frac{1}{n_c-1}\sum_{j=1}^{n_c}(\hat{\boldsymbol{\delta}}^{(j)} - \hat{\boldsymbol{\delta}}_{\text{MI}})(\hat{\boldsymbol{\delta}}^{(j)} - \hat{\boldsymbol{\delta}}_{\text{MI}})^{\mathsf{T}}$. These two components are the combined to give $\hat{\boldsymbol{V}}_{\text{MI}} = \hat{\boldsymbol{W}}_{\text{MI}} + \left(1 + \frac{1}{n_c}\right)\hat{\boldsymbol{B}}_{\text{MI}}$, where the extra $n_c^{-1}\hat{\boldsymbol{B}}_{\text{MI}}$ term is included to adjust for the finite number of imputations, n_c.

When the model contains parametric and smooth components, we can use the Bayesian posterior distributions obtained from the models fitted to the complete data sets to create a global posterior distribution that can be employed to construct intervals (e.g., Daniels & Hogan, 2008). Specifically, for each model, we simulate a large number of draws from the corresponding pos-

terior distribution of the parameters and then combine all draws from all models to form a set of realisations from all posterior distributions. $100(1-\alpha)\%$ intervals for the parametric components can be obtained by computing the $\alpha/2$ and $1-\alpha/2$ quantiles of the set. Intervals for the smooth components or non-linear functions of the model parameters are calculated in a similar fashion.

Function `imputeSS(x, m)` in `GJRM`, where `x` is a fitted `gjrm` object and `m` is the number of sets of imputed values for the outcome of interest, implements the above imputation approach.

5 Monte Carlo simulations

5.1 Data generating process

Through a simulation study, we characterised the relative performance of the copula selection approach with and without MI compared to widely used sample selection models such as FIML and 2-step Heckman approach. The data generating process was informed by our motivating example and previous empirical studies (Sales et al., 2004; Imai, 2009; Alva et al., 2014) in order to reflect a wide range of MNAR settings typically seen in practice. For example, we considered scenarios based on different strengths of the MNAR mechanism, alternative proportions of missing data, and strength of the exclusion restrictions.

The data generating process assumed the following. We generated an exclusion restriction variable $X_1 \sim N(0,\ 1)$, a prognostic factor $X_2 \sim N(0.1-0.1X_1,\ 1)$, and the treatment indicator $P(T=1|X_1) = \Phi(0.2+0.3X_1)$. The missingness ($Y_1$) and outcome ($Y_2$) variables were generated from a joint distribution using copulae as described in Marra & Radice (2017a). The parameters of the marginals are specified as $\eta_{\mu_1} = 1 + 0.4T + 0.3X_2 + 0.5X_1$ and $\eta_{\mu_2} = 1 + 0.2T + 0.1X_2$, and θ governs the dependence between Y_1 and Y_2 (this induces MNAR). For the purposes of the simulations and to make the comparisons with existing methods more straightforward, we have considered throughout: i) a Gaussian copula (Table 2), ii) a Gamma distributed outcome (where the mean and skewness were simple functions of the usual shape and scale parameters), iii) a probit model for the selection, and iv) linear functional forms for both Y_2 on X_2, X_2 on X_1 and T on X_1.

By construction, X_1 is independent of Y_2 conditional on X_2 (i.e., it meets the condition for

exclusion restriction). The parameter of interest was the average treatment effect (ATE - obtained from coefficient of Y_2 on T; true value is 0.2) in the outcome model. We varied the following parameters: % of missing data (by varying the intercept in η_{μ_2}); skewness parameter, 0.5 and 2; strength of MNAR (by varying θ such that $cor(Y_1, Y_2) = 0.1, 0.25$, and 0.5); strength of the exclusion restriction (coefficient of η_{μ_1} on X_1 is varied such that $cor(Y_1, X_1) = 0.1$ and 0.4).

In addition, we have considered three sensitivity analyses which allowed for: i) potential misspecification of the 'true' outcome distribution (i.e., outcome data were simulated using a log-normal distribution, but used a gamma for the analysis model); ii) heavier tails for the selection model by simulating non-response (Y_1) from a *t*-distribution instead of normal; and iii) an alternative data generating process, where the joint model is generated using a marginal and a conditional distribution instead of copulae (further details are given in Appendix A of the Supplementary Material).

For each scenario, we employed 1000 simulated datasets (each dataset included 1000 individuals) and compared the following methods: Full-data (no missing data - 'benchmark' for the selection models), full-information maximum likelihood (FIML) assuming bivariate normality, two-step Heckman selection model, copula selection model with and without MI (we considered 20 imputations throughout although using a larger number did not lead to different conclusions and only increased the computing time). Performance was assessed according to bias, root mean squared error (rMSE), and confidence interval (CI) coverage for estimating treatment effects (β_1) in the following outcome model: $Y_{2i} = \beta_0 + \beta_1 T_i + \beta_2 X_{2i} + e_{2i}$. FIML and the two-step Heckman approach were implemented in the `sampleSelection` package in `R` (Toomet & Henningsen, 2008), whereas the copula selection framework used the `GJRM R` package.

5.2 Results

Table 3 reports bias, CI coverage and rMSE across scenarios with slightly skewed data (skewness factor is 0.5), and alternative strengths (θ) of MNAR. Even in such scenarios with small deviations from normality, ATE estimates using FIML were biased and CI coverage was 'poor' (below 0.9). The two-step Heckman model provided unbiased estimates and CI coverage close to nominal levels (95%) across most scenarios, although for some scenarios with large % missing data

and 'stronger' MNAR ($\theta = 0.25$ and 0.5) treatment effect estimates were slightly biased and less precise (larger rMSE). Copula selection models, estimated by either maximum likelihood or MI led to unbiased estimates, CI coverage close to nominal levels and the lowest rMSE across all scenarios. Figure 2 suggests that the differences in relative performance across alternative approaches were maintained in scenarios with highly skewed data (skewness factor was 2). In particular, the 2-step Heckman model appeared to be less robust to large departures from the bivariate normal assumption, and provided larger biases and less precise estimates compared to the copula selection models.

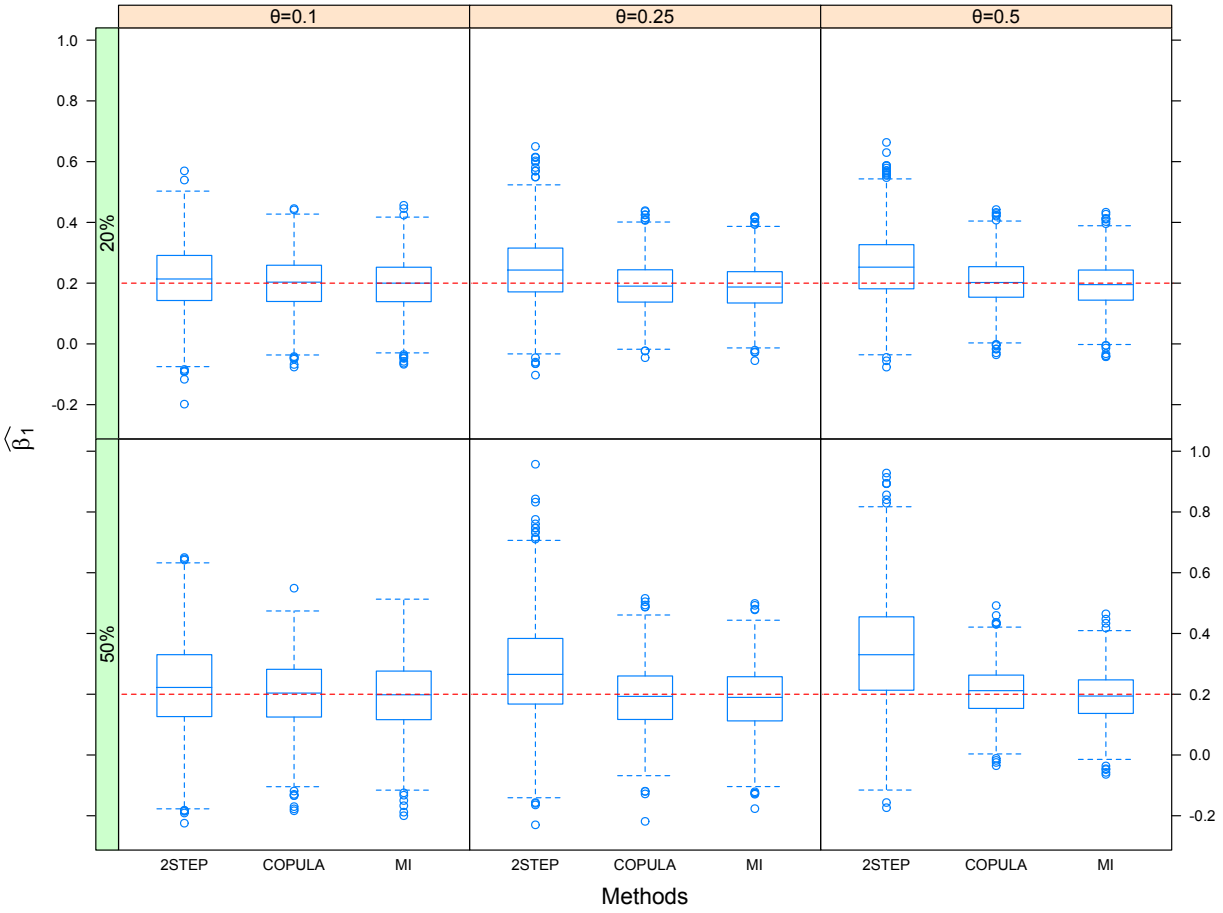

Figure 2: Estimated parameter of interest ($\widehat{ATE}$) according to method for scenarios with highly skewed data, increasing levels of θ and alternative % missing data. The boxplots show bias and variation, as median, quartiles and 1.5 times interquartile range for the estimated parameter across the 1000 replications. The dashed lines are the true values. 2STEP: 2-step Heckman model, COPULA: copula selection model using penalized ML; MI: copula selection model using MI.

Scenarios described thus far have assumed a 'strong' exclusion restriction ($cor(Y_1, X_1) = 0.4$).

Table 3: Percent bias, rMSE, and confidence interval coverage for treatment effect (true $\widehat{ATE}$ is 0.2) according to method for scenarios with slightly skewed data (skewness factor is 0.5).

% missing data	θ	Method	Bias (%)	Coverage	rMSE
20%	0.1	FULL DATA	0%	0.949	0.019
		FIML	21%	0.422	0.067
		2STEP	1%	0.956	0.027
		COPULA	1%	0.938	0.023
		MI	1%	0.934	0.023
	0.25	FULL DATA	0%	0.949	0.019
		FIML	23%	0.516	0.054
		2STEP	5%	0.943	0.029
		COPULA	1%	0.941	0.022
		MI	2%	0.935	0.023
	0.5	FULL DATA	0%	0.949	0.019
		FIML	13%	0.809	0.035
		2STEP	6%	0.947	0.030
		COPULA	1%	0.939	0.021
		MI	2%	0.38	0.021
50%	0.1	FULL DATA	0%	0.949	0.019
		FIML	12%	0.802	0.070
		2STEP	2%	0.962	0.035
		COPULA	1%	0.937	0.029
		MI	1%	0.929	0.028
	0.25	FULL DATA	0%	0.949	0.019
		FIML	27%	0.683	0.076
		2STEP	8%	0.946	0.042
		COPULA	1%	0.938	0.030
		MI	2%	0.915	0.030
	0.5	FULL DATA	0%	0.949	0.019
		FIML	19%	0.820	0.050
		2STEP	14%	0.915	0.049
		COPULA	2%	0.933	0.026
		MI	3%	0.912	0.027

Notes: FULL DATA: no missing data; FIML: Full-information maximum likelihood (assuming bivariate normality); 2STEP: 2-step Heckman model, COPULA: copula selection model using ML; MI: copula selection model using MI.

Figure 3 describes the relative performance of the different selection approaches with a 'weak' exclusion restriction, i.e. correlation between Y_1 and X_1 was 0.1 (as in the REFLUX study). As anticipated, estimates by the Heckman model are highly imprecise as the 2-step approach relies more heavily on the presence of a valid exclusion restriction. Both copula-based selection approaches perform relatively well across all scenarios; for example, rMSE from these methods is only slightly (around 10%) higher compared to the scenarios with 'strong' exclusion restriction (presented in Figure 2).

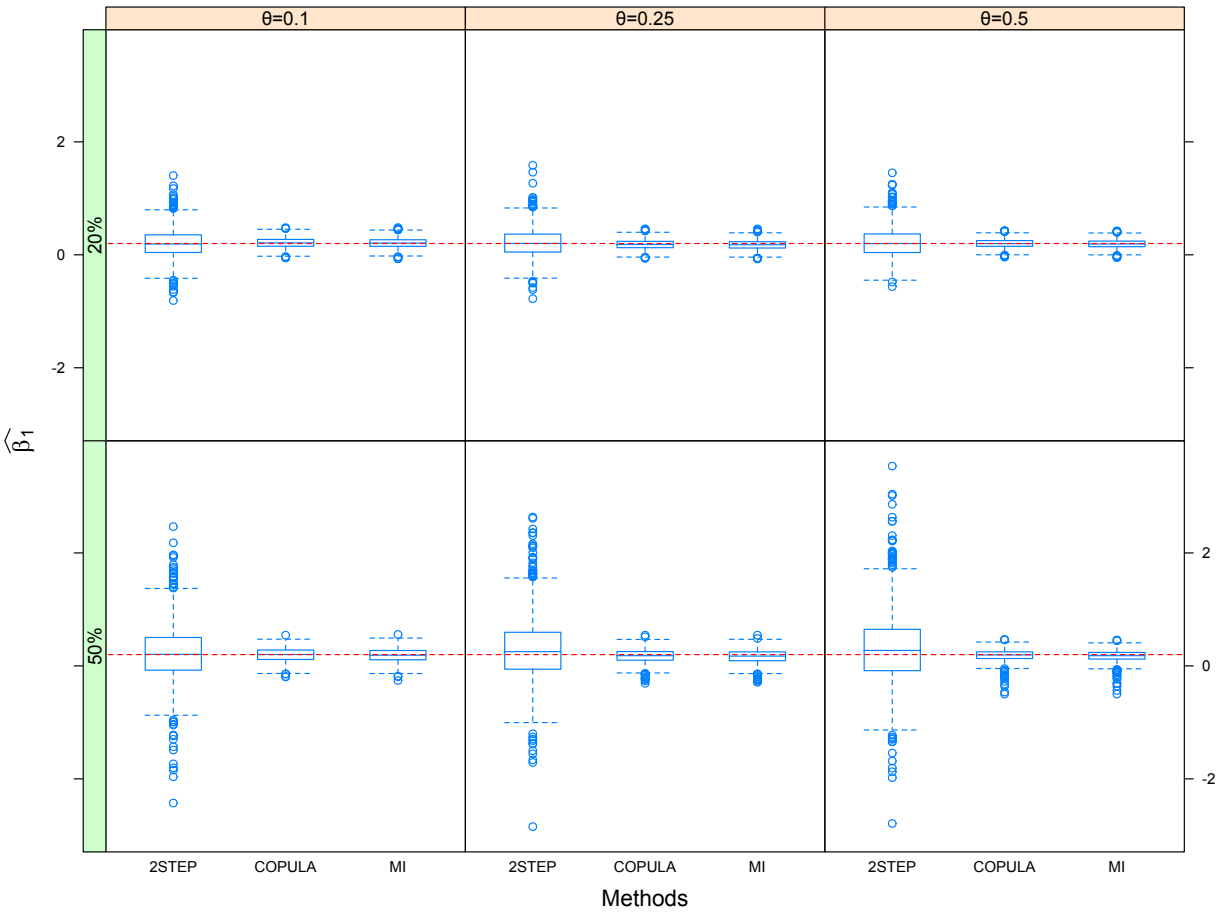

Figure 3: Estimated parameter of interest ($\widehat{ATE}$) according to method for scenarios with 'weak' exclusion restriction ($cor(Y_1, X_1) = 0.1$), increasing levels of θ and alternative % missing data. The dashed lines are the true values. 2STEP: 2-step Heckman model, COPULA: copula selection model using ML; MI: copula selection model using MI.

Figure 4 presents the results for settings where the copula model was slightly misspecified (the copula model assumes gamma when the 'true' outcome data was log-normal). Both copula selection model and MI still provided lower rMSE (results between these methods mostly overlap)

than the 2-step Heckman approach. The plot suggests that the higher the θ (the stronger the MNAR) the more precise estimates the copula approaches provide. Results for the remaining sensitivity analyses (alternative DGP and t-distributed non-response) are reported in Appendix B of the Supplementary Material.

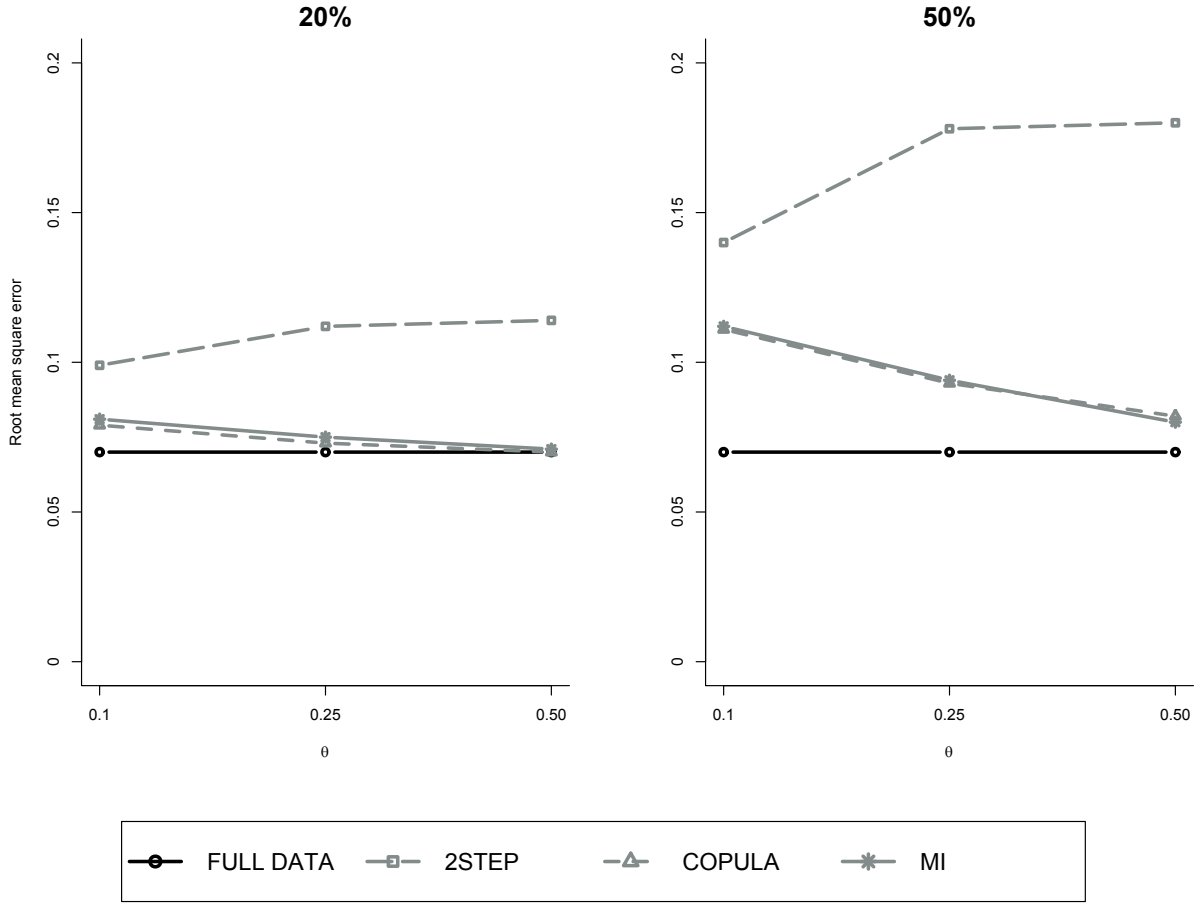

Figure 4: Root mean square error according to method for scenarios with misspecification of the copula model, increasing levels of θ and alternative % missing data. FULL DATA: no missing data; 2STEP: 2-step Heckman model, COPULA: copula selection model using ML; MI: copula selection model using MI.

6 Application to REFLUX data

When building the model we investigated the appropriateness of several copulae, link functions (for the selection equation) and outcome distributions. As for the selection equation, the `cloglog` link provided the best fit based on the AIC/BIC. Amongst the available copulae, the Plackett pro-

vided the lowest BIC/AIC. Post-estimation normal Q-Q plots suggested that the Gumbel distribution provided a good fit for the (observed) QALY endpoint (Appendix C, Supplementary Material).

The results from applying the various selection models to the REFLUX data are reported in Table 4. We followed the same regression adjustment used in the primary analysis of these data (Grant et al., 2013). The outcome linear model included key prognostic factors, such as age, gender, baseline HRQL (both REFLUX-specific and generic) and body mass index. The missing data model for each method included all the covariates used in the analysis model, other patient characteristics, such as education and employment status, reported in Table 1, and the variables that met the exclusion restriction (patient's views about medicine). The parameter of interest was the treatment effect of surgery on the QALY.

There was strong evidence that patients receiving laparoscopic surgery had better QALY five years after the intervention than those receiving standard medical care. This effect was somewhat larger for the complete cases, FIML and Heckman model compared to the copula selection model and MI. There was some evidence that five-year QALY was associated with baseline quality of life score, gender, heart burn and symptoms scores, although the strength and statistical significance of the regression coefficients for these prognostic factors differed across methods. The copula selection models led to substantially smaller standard errors for all regression coefficients, including treatment effect, compared to the other approaches (as suggested more generally in our simulation study). For example, the MI approach (which provided the most precise estimates) produced standard errors that were, on average, over 30% lower than those from 2-step Heckman or FIML.

7 Discussion

With missing not at random data, statistical inference fundamentally rests on untestable assumptions about the missing data mechanism. Selection models can make plausible assumptions regarding the missing data by allowing for departures from the standard missing-at-random assumption. However, the use of selection models for handling MNAR data requires that the analyst recognises the additional, untestable assumptions imposed by these models. Of particular concern (and the focus of this paper) is the distributional assumption considered in the analysis of non-Gaussian

Table 4: Regression coefficients and standard errors from applying alternative selection models to the REFLUX data. The parameter of interest is the average treatment effect.

	CCA (N=231)	FIML (N=453)	2STEP (N=231)	COPULA (N=453)	MI (N=453)
Treatment	0.433 (0.101)***	0.407 (0.121)***	0.434 (0.098)***	0.376 (0.074)***	0.383 (0.071)***
Age	-0.006 (0.004)	-0.020 (0.005)***	-0.007 (0.005)	-0.004 (0.003)	-0.004 (0.003)
Male	-0.201 (0.098)*	-0.285 (0.117)*	-0.212 (0.098)*	-0.119 (0.072)*	-0.128 (0.067)*
Baseline EQ-5D	2.176 (0.221)***	1.720 (0.247)***	2.171 (0.216)***	1.406 (0.177)***	1.437 (0.163)***
REFLUX score	-0.007 (0.004)*	-0.006 (0.004)	-0.007 (0.003)*	-0.003 (0.002)	-0.004 (0.002)*
BMI (kg/m2)	-0.008 (0.012)	-0.008 (0.013)	-0.006 (0.012)	-0.011 (0.010)	-0.011 (0.009)
Heart burn score	0.006 (0.003)*	0.003 (0.003)	0.006 (0.003)*	0.003 (0.002)*	0.004 (0.002)*
Symptom score 1	0.006 (0.002)*	0.005 (0.003)*	0.006 (0.002)*	0.005 (0.002)*	0.005 (0.002)*
Symptom score 2	-0.003 (0.002)	-0.001 (0.003)	-0.003 (0.002)	0.001 (0.002)	0.001 (0.002)
Nausea score	0.002 (0.003)	0.005 (0.004)	0.001 (0.004)	0.005 (0.002)*	0.005 (0.002)
Activity score	0.002 (0.003)	0.004 (0.004)	0.005 (0.004)	-0.004 (0.003)*	0.005 (0.003)
Intercept	2.080 (0.514)***	3.452 (0.564)***	2.193 (0.547)***	2.675 (0.419)***	2.666 (0.350)***

CCA: complete-case analysis; FIML: Full-information maximum likelihood (assuming bivariate normality); 2STEP: 2-step Heckman model, COPULA: copula selection model using ML; MI: copula selection model using MI. *p<0.05, **p<0.01 ***p<0.001.

outcome data such as patient-reported quality of life responses.

This paper discusses a flexible copula-based selection model, which can help make more plausible assumptions about the distribution of the data and offer flexibility about the choice of joint model for both the outcome and non-response. The modularity of the estimation approach allows for easy inclusion of potentially any parametric continuous marginal distribution, link function, and one-parameter copula function as long as the cdf and pdf are known and their derivatives with respect to their parameters exist. This article also proposes a flexible imputation procedure that generates plausible imputed values from the copula selection model. The methods can be readily implemented via `gjrm()` in the `R` package `GJRM` (`R` code is provided in Appendix C of the supplementary material).

Through simulations we have studied the performance of the copula selection models across a wide range of MNAR settings. The proposed approach provided lower bias and root mean square error compared to popular selection model approaches such as FIML and 2-step Heckman approach across all scenarios considered. Our method improves the *status quo* particularly in the absence of valid exclusion restrictions, which tend to be rare in medical and epidemiological studies. In addition, our simulation results suggested that the copula selection approach is somewhat robust to (some degree of) misspecification of the outcome and selection equations. Note that it may be difficult to simulate the potentially highly complex processes that likely underlie the relation between the missingness mechanism and outcome of interest. Therefore, in practical applications, it is not possible to determine with certainty how the model assumptions and/or lack of valid exclusion restrictions may affect the empirical results. Nevertheless, our findings suggest that the copula approach has merit in dealing with missingness not at random and we believe that the approach discussed in this paper is a useful addition to the statistical toolbox.

A major strength of the copula framework is that it can be easily embedded in sensitivity analysis to alternative missing data assumptions, as widely recommended by methodological guidelines for addressing MNAR data (Carpenter & Kenward, 2013; Sterne et al., 2009; Molenberghs et al., 2014). In fact, the true missing data mechanism (and data distribution) will always be unknown, and sensitivity analyses provide a helpful framework for assessing how conclusions may differ under different plausible assumptions. As well as allowing for alternative dependencies between the

outcome and selection (MNAR mechanisms), the copula approach enables the analyst to assess whether the study's conclusions are robust to key parametric assumptions such as the functional form of outcome and selection models, and data distribution. In addition, the fact that the copula approach can be implemented with MI, makes it particularly attractive for medical and epidemiological researchers given their familiarity with MI for handling missing data.

There are some aspects of the proposed modelling approach that have scope for improvement and provide direction for future research. Firstly, for comparative purposes we have assumed linear covariate effects throughout. However, incorporating penalised regression splines in the discussed copula framework to allow for various degrees of non-linear effects is straightforward and already implemented in our software (e.g., Marra & Radice, 2017a). Secondly, this study focused on handling missing outcome data. In many settings, there will be both missing outcomes and covariates. In such settings, multiple imputation approaches such as that proposed in this paper can be easily extended to accommodate the missingness both in covariates (typically assuming MAR) and outcomes (MNAR). Extending and embedding our proposed MI approach within popular MI software, such as the `R` package `mice` will further encourage the uptake of the copula methodology. Thirdly, another area which warrants further consideration is longitudinal data. Longitudinal settings typically poses further challenges to joint selection models as these often require additional parametric assumptions, for example about the longitudinal correlation structure. Finally our proposed approach can be extended to settings where decision problem requires joint inferences for more than one outcome (Filippou et al., 2017) or addressing hierarchical and/or spatial effects (Marra et al., 2017).

References

Alva, M., Gray, A., & et al (2014). The effect of diabetes complications on health-related quality of life: the importance of longitudinal data to address patient heterogeneity. *Health Econ*, 23(4), 487–500.

Carpenter, J. & Kenward, M. (2013). *Multiple Imputation and its Application*. Statistics in Practice. Wiley.

Chen, S. & Zhou, Y. (2010). Semiparametric and nonparametric estimation of sample selection models under symmetry. *Journal of Econometrics*, 157, 143–150.

Chib, S., Greenberg, E., & Jeliazkov, I. (2009). Estimation of semiparametric models in the presence of endogeneity and sample selection. *Journal of Computational and Graphical Statistic*, 18, 321–348.

Collett, D. (2002). *Modelling Binary Data*. London: Chapman & Hall/CRC Texts in Statistical Science.

Daniels, M. & Hogan, J. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Boca Raton, FL: Chapman and Hall CRC.

Das, M., Newey, W., & Vella, F. (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies*, 70, 33–58.

Del Bianco, P. & Borgoni, R. (2006). Handling dropout and clustering in longitudinal multicentre clinical trials. *Statistical Modelling*, 6(2), 141–157.

Diggle, P. & Kenward, M. G. (1994). Informative drop-out in longitudinal data-analysis. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 43(1), 49–93.

Ding, P. (2014). Bayesian robust inference of sample selection using selection-models. *Journal of Multivariate Analysis*, 124, 451–464.

EuroQol (1990). Euroqol-a new facility for the measurement of health-related quality of life. *Health Policy*, 16(3), 199–208.

Filippou, P., Radice, R., & Marra, G. (2017). Penalized likelihood estimation of a trivariate additive probit model. *Biostatistics*, 18, 569–585.

Galimard, J.-E., Chevret, S., Protopopescu, C., & Resche-Rigon, M. (2016). A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Statistics in Medicine*, 35(17), 2907–2920.

Gomes, M., Kenward, M., Grieve, R., & Carpenter, J. (2017). Estimating treatment effects under untestable assumptions with non-ignorable missing data. *Statistical Methods in Medical Research*, (under review).

Grant, A. M., Boachie, C., Cotton, S. C., Faria, R., & et al (2013). Clinical and economic evaluation of laparoscopic surgery compared with medical management for gastro-oesophageal reflux disease: 5-year follow-up of multicentre randomised trial (the reflux trial). *Health Technol Assess*, 17(22), 1–167.

Heckman, J. (1974). Shadow prices, market wages and labor supply. *Econometrica*, 42, 679–694.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–162.

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., & Sterne, J. A. C. (2014). Joint modelling rationale for chained equations. *Bmc Medical Research Methodology*, 14.

Imai, K. (2009). Statistical analysis of randomized experiments with non-ignorable missing binary outcomes: an application to a voting experiment. *Journal of the Royal Statistical Society Series C*, 58, 83–104.

Lee, D. S. (2008). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(11721), 1071–1102.

Little, R. J. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Liu, J. C., Gelman, A., Hill, J., Su, Y. S., & Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika*, 101(1), 155–173.

Marchenko, Y. V. & Genton, M. G. (2012). A heckman selection-t model. *Journal of the American Statistical Association*, 107(497), 304–317.

Marra, G. & Radice, R. (2017a). Bivariate copula additive models for location, scale and shape. *Computational Statistics & Data Analysis*, 112, 99–113.

Marra, G. & Radice, R. (2017b). *GJRM: Generalised Joint Regression Modelling*. R package version 0.1-4.

Marra, G., Radice, R., Bärnighausen, T., Wood, S. N., & McGovern, M. E. (2017). A simultaneous equation approach to estimating hiv prevalence with non-ignorable missing responses. *Journal of the American Statistical Association*, 112(518), 484–496.

Marra, G. & Wood, S. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39, 53–74.

Mason, A., Gomes, M., Grieve, R., Ulug, P., Powell, J., & Carpenter, J. (2017). Development of a practical approach to expert elicitation for randomised controlled trials with missing health outcomes: Application to the improve trial. *Clinical Trials*, (in press).

Molenberghs, G., Fitzmaurice, G. M., Kenward, M., Tsiatis, A. A., & Verbeke, G. (2014). *Handbook of missing data methodology*. Boca Raton, US: Chapman and Hall/CRC.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *Econometrics Journal*, 12, S217–S229.

NICE (2013). *Guide to the methods of technology appraisal*. National Institute for Health and Care Excelence, London, UK.

Nocedal, J. & Wright, S. J. (2006). *Numerical Optimization*. New York: Springer-Verlag.

Ogundimu, E. & Collins, G. (2017). A robust imputation method for missing responses and covariates in sample selection models. *Statistical Methods in Medical Research*, 26, 1–15.

Pigini, C. (2015). Bivariate non-normality in the sample selection model. *Journal of Econometric Methods*, 4, 123–144.

Puhani, P. A. (2000). The heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 14, 53–68.

Rigby, R. A. & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society, Series C*, 54, 507–554.

Robert, C. & Casella, G. (2005). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer New York.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Classics Library. Wiley.

Sales, A. E., Plomondon, M. E., Magid, D. J., Spertus, J. A., & Rumsfeld, J. S. (2004). Assessing response bias from missing quality of life data: the heckman method. *Health Qual Life Outcomes*, 2, 49.

Schafer, J. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1), 3–15.

Sklar, A. (1973). Random variables, joint distributions, and copulas. *Kybernetica*, 9, 449–460.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 339(b2393), (29 June 2009).

Toomet, O. & Henningsen, A. (2008). Sample selection models in r: Package sampleselection. *Journal of Statistical Software*, 27(7), 1–23.

Zhelonkin, M., Genton, M. G., & Ronchetti, E. (2015). Robust inference in sample selection models. *Journal of the Royal Statistical Society Series B*.