Published in final edited form as:
Canadian Journal of Statistics (2015), **43**(2): 240–264
DOI: 10.1002/cjs.11249

Multiple imputation for the analysis of incomplete compound variables

JIWEI ZHAO

Department of Biostatistics,

University of Buffalo, Buffalo, NY, 14214, USA

E-mail: zhaoj@buffalo.edu

RICHARD J. COOK

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

CHANGBAO WU

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

Summary

In many settings interest lies in modelling a compound variable defined as a function of two or more component variables. When one or more of the components are missing, the compound variable is not observed and a strategy for handling incomplete data is required. Analyses based on individuals with complete data are inefficient and yield potentially inconsistent estimators.We develop a multiple imputation strategy in this setting with an auxiliary model for imputing the compound variable directly, and one based on a multivariate imputation model for the component variables. Asymptotic properties of the imputation-based estimators are presented for the case in which the imputation model is correctly specified, and a shrinkage estimator is proposed to reduce the bias arising from misspecification of the imputation model. Finite sample properties of the various estimators are examined through simulations. An application to data from the Canadian Youth Smoking Survey involving a study of body mass index illustrates the approach.

Keywords: Asymptotic variance; compound variable; multiple imputation; relative efficiency; shrinkage estimator.

1 INTRODUCTION

In many areas of health research interest lies in a compound variable defined as a deterministic function of two or more component variables. Examples of compound variables include composite outcomes in clinical trials, indices based on instruments measuring health and well-being (Fries et al., 1981; Ware & Sherbourne, 1992; McHorney et al., 1993), or summary measures of response to interventions in laboratory studies (Stroncek & Rebulla, 2007). We focus here on the analysis of body mass index (BMI), the ratio of body weight in kilograms and squared height, where height is measured in meters (Hedley et al., 2004; Krebs et al., 2007).

Often one or more of the components of the compound variable are missing and when there are several component variables, the overall rate of missing data can be appreciable. In a recent study (Elton-Marshall et al., 2014) aiming to estimate the proportion of overweight and obese adolescents, BMI could not be computed for approximately one-third of the participating children. This difficulty arose due to missing height, missing weight, or missing height and weight. A simple and naive approach is to restrict attention to individuals with complete data, but this can be highly inefficient. We aim to develop a general framework for exploiting available information from individuals with partial data on the component variables.

General strategies for the analysis of incomplete data include a complete case analysis, a weighted complete case analysis, use of augmented inverse probability weighted estimating equations (Robins & Rotnitzky, 1992; Robins et al., 1994; Robins & Rotnitzky, 1995; Tsiatis, 2006), and multiple imputation (Rubin, 1978, 1987, 1996; Little & Rubin, 2002). To handle incomplete compound variables, imputation is appealing and our work makes use of the large sample theory of multiple imputation given in Wang & Robins (1998) and Robins & Wang (2000).

Multiple imputation, proposed by Rubin (1978) and refined in Rubin (1987), was developed in a Bayesian paradigm, but has been popularized in part due to attractive frequentist properties. The original multiple imputation approach (Rubin, 1987, 1996) involves sampling the parameters for the imputation model from its posterior distribution, but Wei & Tanner (1990) suggested that one simply use the maximum likelihood estimate of the parameters for the imputation model; the resulting estimator belongs to the "type B" class of estimators in Wang & Robins (1998). Wang & Robins (1998) and Robins & Wang (2000) rigorously established the large sample theory for "type B" estimators based on a frequentist multiple imputation procedure. Although the relative efficiency of "type B" to classical estimators decreases as the number of imputations $M \to \infty$, the "type B" estimator is always more efficient than the one based on Bayesian multiple imputation since the asymptotic variance of the latter introduces an additional source of variability by sampling the parameter from the posterior distribution at each imputation (Nielsen, 2003). Others who studied efficiency issues in multiple imputation include Lu, Jiang & Tsiatis (2010), who analyzed an incomplete dichotomized response through multiply imputing the underlying continuous longitudinal measurements. Shen & Chen (2013) considered model selection in the context of generalized estimating equations with multiply imputed longitudinal responses and concluded that the model selection criteria based on frequentist multiple imputation generally performed better than the analogous Bayesian procedure. See Tsiatis (2006, Chapter 14) and Kim & Shao (2013, Chapter 4) for an excellent account of the theory of multiple imputation.

Here we explore efficient use of available data by adopting a frequentist multiple imputation procedure. The work is motivated by the need to summarize BMI in youth and to make comparisons between aboriginal and non-aboriginal youth in terms of this outcome using data from the Canadian Youth Smoking Survey (Elton-Marshall et al., 2014). While BMI is a continuous variable, interest lies in the proportion of youth in the "overweight" and "obese" categories. We adopt a link based on the cumulative distribution of the continuous BMI (Agresti, 2010) for an ordinal response (normal/overweight/obese); the proposed model and method can accommodate covariate-specific cut

points for the three levels. We propose two frequentist imputation-based approaches for dealing with incomplete component variables: one based on imputing the continuous univariate BMI, and one based on a bivariate imputation model for the component variables of weight and height. We quantify the efficiency gain of the second approach when the imputation model is correctly specified and propose a shrinkage estimator (Chen, Chatterjee & Carroll, 2009) which partially corrects for estimation bias arising due to misspecification of the imputation model.

The remainder of the paper is organized as follows. In Section 2, we define an ordinal model for BMI. The frequentist imputation-based procedures are presented in Sections 3 and 4 along with the related asymptotic results. Simulation studies are described and the finite sample performances of the different estimators are reported in Section 5. An application to a recent study of obesity in adolescents is reported in Section 6 and concluding remarks are made in Section 7. Proofs and technical details are provided in the Appendix.

2 AN ORDINAL RESPONSE MODEL

2.1 NOTATION AND MODEL FORMULATION WITH COMPLETE DATA

BMI defined as an individual's weight (kg) divided by their height squared (m^2), is a useful measure of health and risk for many diseases (Hedley et al., 2004; Krebs et al., 2007). BMI data are often skewed and so the natural logarithmic transformation is routinely adopted (e.g. Lamon-Fava et al., 1996). In this case if $Y = \log \text{BMI}$, $Y_1 = \log \text{weight}$ where weight is measured in kilograms, $Y_2 = \log \text{height}^2 = 2 \log \text{height}$ where height is measured in meters, then $Y = Y_1 - Y_2$.

Consider the linear model

$$Y = \boldsymbol{X}^T \boldsymbol{\alpha} + \sigma W, \tag{1}$$

where Y represents a 1-1 transformation of BMI taking values on the whole real line, $\boldsymbol{X} = (1, X_1, \ldots, X_{p-1})^T$ is a $p \times 1$ covariate vector, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_{p-1})^T$ is the vector of respective regression coefficients, and W is an error term independent of $\boldsymbol{X}$ with $\mathrm{E}(W) = 0$ and finite variance. We let $\boldsymbol{\gamma} = (\boldsymbol{\alpha}^T, \sigma)^T$ where σ is a dispersion parameter.

Interest lies in classifying individuals into meaningfully different categories based on BMI. In adults, for example, individuals are designated as obese if $Y \geq \log 30$, overweight if $\log 25 \leq Y < \log 30$, and normal otherwise. In adolescents, the corresponding cut points are gender and age specific percentiles based on reference data and growth charts of the World Health Organization (Cole & Green, 1992; Onis et al., 2007). To cover both cases we let $-\infty = c_0 < c_1(\boldsymbol{X}) < \cdots < c_{K-1}(\boldsymbol{X}) < c_K = \infty$ denote cut points on the Y-scale and denote the kth interval as $\mathcal{C}_k(\boldsymbol{X}) = [c_{k-1}(\boldsymbol{X}), c_k(\boldsymbol{X}))$, $k = 1, \ldots, K$. The variable $Z(\boldsymbol{X}, Y) = \sum_{k=1}^{K} k \cdot I(Y \in \mathcal{C}_k(\boldsymbol{X}))$ records the interval in which BMI falls for an individual with covariate $\boldsymbol{X}$. If $\bar{\boldsymbol{X}}_k = (\boldsymbol{X}^T, c_k(\boldsymbol{X}))^T$, then

$$p_k(\boldsymbol{\theta}) = P(Z = k|\boldsymbol{X}; \boldsymbol{\theta}) = F(\bar{\boldsymbol{X}}_k^T \boldsymbol{\theta}) - F(\bar{\boldsymbol{X}}_{k-1}^T \boldsymbol{\theta}), \quad k = 1, \ldots, K\ , \tag{2}$$

where $\boldsymbol{\beta} = -\tau \boldsymbol{\alpha}$, $\tau = \sigma^{-1}$, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \tau)^T$, and $F(\cdot)$ is the standard cumulative distribution function of W with $F(\bar{\boldsymbol{X}}_0^T \boldsymbol{\theta}) = 0$ and $F(\bar{\boldsymbol{X}}_K^T \boldsymbol{\theta}) = 1$.

We write $Z_k = I(Z = k)$, $k = 1, 2, \ldots, K$ and let $\boldsymbol{Z} = (Z_2, \ldots, Z_K)^T$. If $\boldsymbol{p}(\boldsymbol{\theta}) = (p_2(\boldsymbol{\theta}), \ldots, p_K(\boldsymbol{\theta}))^T$, we let $\phi_k(\boldsymbol{\theta}) = \log(p_k(\boldsymbol{\theta})/p_1(\boldsymbol{\theta}))$, $k = 2, \ldots, K$, and $\boldsymbol{\phi}(\boldsymbol{\theta}) = (\phi_2(\boldsymbol{\theta}), \ldots, \phi_K(\boldsymbol{\theta}))^T$. We then have

$$p(Z|\boldsymbol{X}; \boldsymbol{\theta}) \;\; = \;\; \exp\left\{ \boldsymbol{Z}^T \boldsymbol{\phi}(\boldsymbol{\theta}) - \log\left[1 + \textstyle\sum_{k=2}^{K} \exp\{\phi_k(\boldsymbol{\theta})\}\right] \right\}.$$

If we introduce a subscript i to distinguish between different individuals, data from a sample of n i.i.d. individuals are denoted by $\{(Z_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$. The log-likelihood is then given by

$$l(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left[\boldsymbol{Z}_i^T \boldsymbol{\phi}_i(\boldsymbol{\theta}) - b\{\boldsymbol{\phi}_i(\boldsymbol{\theta})\} \right], \tag{3}$$

where for any $d \times 1$ vector $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)^T$, we define $b(\boldsymbol{\nu}) = \log\{1 + \sum_{i=1}^{d} \exp(\nu_i)\}$.

The maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{\text{full}}$ is obtained by solving the score equation

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{S}(Z_i|\boldsymbol{X}_i;\boldsymbol{\theta}) = \mathbf{0}\ , \tag{4}$$

where $\boldsymbol{S}(Z|\boldsymbol{X};\boldsymbol{\theta}) = \bar{\boldsymbol{X}} D(\boldsymbol{\theta}) V(\boldsymbol{\theta})(\boldsymbol{Z} - \boldsymbol{p}(\boldsymbol{\theta}))$ with components $\bar{\boldsymbol{X}} = (\bar{\boldsymbol{X}}_1, \ldots, \bar{\boldsymbol{X}}_{K-1})$, $D(\boldsymbol{\theta}) = D_1(\boldsymbol{\theta}) - D_2(\boldsymbol{\theta})$, $D_1(\boldsymbol{\theta})$ is a $(K-1) \times (K-1)$ matrix with $(j+1, j)$ element $F'(\bar{\boldsymbol{X}}_{j+1}^T\boldsymbol{\theta})$, $j = 1, \ldots, K-2$, $D_2(\boldsymbol{\theta})$ is a $(K-1) \times (K-1)$ matrix with (j, j) element $F'(\bar{\boldsymbol{X}}_j^T\boldsymbol{\theta})$, $j = 1, \ldots, K-1$; $V(\boldsymbol{\theta}) = V_1(\boldsymbol{\theta}) + (p_1(\boldsymbol{\theta}))^{-1}$, $V_1(\boldsymbol{\theta})$ is a $(K-1) \times (K-1)$ matrix with (j, j) element $(p_{j+1}(\boldsymbol{\theta}))^{-1}$, $j = 1, \ldots, K-1$. Under mild regularity conditions (Fahrmeir & Tutz, 2001),

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\text{full}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, I_c^{-1}(\boldsymbol{\theta}_0)),$$

where the information matrix $I_c(\boldsymbol{\theta}_0)$ can be estimated by $\sum_{i=1}^{n} \bar{\boldsymbol{X}}_i D_i(\widehat{\boldsymbol{\theta}}_{\text{full}}) V_i(\widehat{\boldsymbol{\theta}}_{\text{full}}) D_i^T(\widehat{\boldsymbol{\theta}}_{\text{full}}) \bar{\boldsymbol{X}}_i^T / n$.

2.2 Naive Methods for Dealing with Missing Components of Z

Let $R = 1$ if Z is observed and $R = 0$ otherwise. Throughout we assume the covariate $\boldsymbol{X}$ is fully observed and data are missing at random (MAR), so $P(R = 1|Z, \boldsymbol{X}) = P(R = 1|\boldsymbol{X}) = \pi_0(\boldsymbol{X})$. We also let $\pi_0 = P(R = 1) = \mathrm{E}\{\pi_0(\boldsymbol{X})\}$, $0 < \pi_0 < 1$.

The most common naive method of analysis is to restrict attention to individuals with complete data and to solve the complete-case score equation given by $\sum_{i=1}^{n} R_i \cdot \boldsymbol{S}(Z_i|\boldsymbol{X}_i;\boldsymbol{\theta})/n = \mathbf{0}$. If $\widehat{\boldsymbol{\theta}}_{\text{cc}}$ solves the complete-case score equation, then

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\text{cc}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, I_o^{-1}),$$

where $I_o = \mathrm{E}(R \cdot S^{\otimes 2})$ is the corresponding observed information. Note that $I_m = I_c - I_o$ denotes the lost information arising from the missing responses, and when π_0 is small, this can be appreciable.

To implement a frequentist MI procedure (Wang & Robins, 1998), we impute each missing Z_i by drawing a random sample $\tilde{Z}_i$ from $p(Z|\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_{\text{cc}})$ and calculate the MLE as if the data were complete. This is repeated M times and the final estimator $\widehat{\boldsymbol{\theta}}_{\text{mi0}}$ is the average of M estimators derived from the M imputed data sets. This is asymptotically equivalent to solving the following estimating equation (Wang & Robins, 1998):

$$\frac{1}{n}\sum_{i=1}^{n}\left[M^{-1}\sum_{m=1}^{M}\{R_i \cdot \boldsymbol{S}(Z_i|\boldsymbol{X}_i;\boldsymbol{\theta}) + (1 - R_i) \cdot \boldsymbol{S}(\tilde{Z}_i^m(\widehat{\boldsymbol{\theta}}_{\text{cc}})|\boldsymbol{X}_i;\boldsymbol{\theta})\}\right] = \mathbf{0}.$$

The asymptotic distribution of $\widehat{\boldsymbol{\theta}}_{\text{mi0}}$ is as follows:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\text{mi0}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, I_o^{-1} + M^{-1} I_c^{-1} I_m I_c^{-1}).$$

In the asymptotic variance of $\sqrt{n}\widehat{\boldsymbol{\theta}}_{\text{mi0}}$, the term I_o^{-1} is the asymptotic variance of the preliminary $\sqrt{n}\widehat{\boldsymbol{\theta}}_{\text{cc}}$; the term $M^{-1} I_c^{-1} I_m I_c^{-1}$ is attributable to the additional variability resulting from the finite number of imputations. Although the second term becomes negligible as $M \to \infty$, the asymptotic variance of $\widehat{\boldsymbol{\theta}}_{\text{mi0}}$ is a bit larger than $\widehat{\boldsymbol{\theta}}_{\text{cc}}$ due to the randomness in the imputation process.

Since $\widehat{\boldsymbol{\theta}}_{\text{cc}}$ is the MLE based on individuals with complete data, it is chosen as the preliminary estimator in the MI procedure for the sake of efficiency (Wang & Robins, 1998). Other choices for a preliminary asymptotically linear estimator will result in larger variance for the corresponding $\widehat{\boldsymbol{\theta}}_{\text{mi0}}$, based on Theorem 1 of Wang & Robins (1998). It is possible to gain efficiency if one generates (i) continuous BMI values or (ii) weight/height values by making additional parametric assumptions. We do so in the following two sections and quantify the consequent efficiency gain.

3 MULTIPLE IMPUTATION BASED ON THE COMPOUND VARIABLE Y

If the functional relationship between Z and Y is denoted generally by $Z = T_1(Y, \boldsymbol{X})$, one can impute a missing Z by $\tilde{Z} = T_1(\tilde{Y}, \boldsymbol{X})$, where the imputed value $\tilde{Y}$ is generated from $p(Y|\boldsymbol{X};\boldsymbol{\gamma})$, the density of Y given $\boldsymbol{X}$ in (1).

A preliminary estimate $\widehat{\boldsymbol{\gamma}}$ obtained from subjects with complete data is found by solving

$$\frac{1}{n}\sum_{i=1}^{n} R_i \cdot \boldsymbol{U}(Y_i|\boldsymbol{X}_i;\boldsymbol{\gamma}) = \boldsymbol{0}, \tag{5}$$

where $\boldsymbol{U}(Y|\boldsymbol{X};\boldsymbol{\gamma}) = \partial \log p(Y|\boldsymbol{X};\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}$. Under mild regularity conditions, $\widehat{\boldsymbol{\gamma}}$, the unique solution to (5), has an asymptotic linear representation:

$$\sqrt{n}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) = n^{-1/2}\sum_{i=1}^{n} Q(R_i, Y_i, \boldsymbol{X}_i) + o_p(1), \tag{6}$$

where $Q = \{E(R\boldsymbol{U}^{\otimes 2})\}^{-1}R\boldsymbol{U}$ and $\mathrm{Var}(Q) = \{E(R\boldsymbol{U}^{\otimes 2})\}^{-1}$.

The MI procedure MI1 proceeds as follows. Let $\widehat{\boldsymbol{\gamma}}$ be the preliminary estimator of $\boldsymbol{\gamma}$. For each m, we first impute each missing Y_i by $\tilde{Y}_i^m$, drawn from $p(Y|\boldsymbol{X}_i;\widehat{\boldsymbol{\gamma}})$, and then compute the corresponding Z_i by $\tilde{Z}_i^m(\widehat{\boldsymbol{\gamma}}) = T_1(\tilde{Y}_i^m(\widehat{\boldsymbol{\gamma}}), \boldsymbol{X}_i)$. The estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{mi1}}$ is then obtained by solving:

$$\frac{1}{n}\sum_{i=1}^{n}\bar{S}^i(\boldsymbol{\theta},\widehat{\boldsymbol{\gamma}}) = \frac{1}{n}\sum_{i=1}^{n}\left[M^{-1}\sum_{m=1}^{M}\{R_i\cdot\boldsymbol{S}(Z_i|\boldsymbol{X}_i;\boldsymbol{\theta}) + (1-R_i)\cdot\boldsymbol{S}(\tilde{Z}_i^m(\widehat{\boldsymbol{\gamma}})|\boldsymbol{X}_i;\boldsymbol{\theta})\}\right] = 0. \tag{7}$$

We have the following asymptotic result regarding $\widehat{\boldsymbol{\theta}}_{\mathrm{mi1}}$.

Theorem 1. *Under (6) and regularity conditions (A), (B), (C) given in the Appendix, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathrm{mi1}} - \boldsymbol{\theta}_0)$ is asymptotically normal with mean zero and variance $I_c^{-1}\Omega I_c^{-1}$, where*

$$\Omega = M^{-1}(I_c - I_o) + I_o + J_c\mathrm{Var}(Q)J_c^T - J_o\mathrm{Var}(Q)J_o^T,$$

and $J_c = \mathrm{E}(S\boldsymbol{U}^T)$, $J_o = \mathrm{E}(R\boldsymbol{S}\boldsymbol{U}^T)$, $\mathrm{Var}(Q) = \{\mathrm{E}(R\boldsymbol{U}^{\otimes 2})\}^{-1}$.

This result is fundamental to our efficiency comparisons and follows from Wang & Robins (1998). It can be seen that if we use (2) to do imputation instead of (1) (i.e. replace the score U by S), the estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{mi1}}$ becomes $\widehat{\boldsymbol{\theta}}_{\mathrm{mi0}}$. In general, when using (1) for imputation, $\widehat{\boldsymbol{\theta}}_{\mathrm{mi1}}$ is more efficient compared to $\widehat{\boldsymbol{\theta}}_{\mathrm{mi0}}$.

Proposition 1. *The estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{mi1}}$ is more efficient than $\widehat{\boldsymbol{\theta}}_{\mathrm{mi0}}$, i.e., $\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\mathrm{mi1}}) \le \mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\mathrm{mi0}})$.*

The essence of the proof is that $K(\mathrm{E}(R\boldsymbol{U}^{\otimes 2}))^{-1}K^T \le (\mathrm{E}(R\boldsymbol{S}^{\otimes 2}))^{-1}$, where $K = \partial\boldsymbol{\theta}/\partial\boldsymbol{\gamma}$ is evaluated at the true parameter value, i.e., the estimator of $\boldsymbol{\theta}$ derived from (1) is more efficient than from (2). This is intuitive since if $Z = T_1(Y, \boldsymbol{X})$, the σ-algebra $\sigma(Z) \subset \sigma(Y)$. The efficiency gain of $\widehat{\boldsymbol{\theta}}_{\mathrm{mi1}}$ is due to the component $J_c\mathrm{Var}(Q)J_c^T - J_o\mathrm{Var}(Q)J_o^T$ in Ω, whereas the counterpart in $\widehat{\boldsymbol{\theta}}_{\mathrm{mi0}}$ is $I_cI_o^{-1}I_c - I_o$. When the missing data mechanism is missing completely at random (MCAR), i.e., $\pi_0(\boldsymbol{X}) = \pi_0$, this difference (in absolute value) becomes $(\pi_0^{-1} - \pi_0)\{I_c - \mathrm{E}(SU^T)(\mathrm{E}(U^{\otimes 2}))^{-1}(\mathrm{E}(SU^T))^T\}$, which decreases when the missing proportion decreases.

The variance of $\widehat{\boldsymbol{\theta}}_{\mathrm{mi1}}$ can be estimated by plugging in the sample version for each component in $\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\mathrm{mi1}})$. As suggested in Wang & Robins (1998) and Robins & Wang (2000), I_c is estimated by $\widehat{I}_c = M^{-1}\sum_{m=1}^{M}\tilde{I}_c^m$ and J_c is estimated by $\widehat{J}_c = M^{-1}\sum_{m=1}^{M}\tilde{J}_c^m$, where $\tilde{I}_c^m$ and $\tilde{J}_c^m$ are calculated from the mth imputed data set.

Robins & Wang (2000) allow potential misspecification or incompatibility between the imputation model and the analysis model in which case the limiting value of $\widehat{\boldsymbol{\beta}}$, denoted by $\boldsymbol{\beta}^*$, may not equal $\boldsymbol{\beta}_0$. Although our imputation model (1) is different from the analysis model (2), model (1) is not misspecified if (2) is correctly specified, so the estimator $\widehat{\boldsymbol{\theta}}_{\rm mi1}$ is asymptotically unbiased in this case. More interestingly, $\widehat{\boldsymbol{\theta}}_{\rm mi1}$, using a different imputation model, is more efficient than $\widehat{\boldsymbol{\theta}}_{\rm mi0}$, an appealing feature to the investigators.

4 Multiple Imputation based on Components (Y_1, Y_2)

A second MI procedure, MI2, is motivated by the fact that Z is a compound variable based on (Y_1, Y_2), so we can write $Z = T_2(Y_1, Y_2, \boldsymbol{X})$. We define the component specific missing indicators $R_k = I(Y_k \text{ is observed }),\ k = 1, 2$. Given $\boldsymbol{X}$, we label the patterns of missing data as pattern 1 ($R_1 = 1, R_2 = 0$), pattern 2 ($R_1 = 0, R_2 = 1$) and pattern 3 ($R_1 = 0, R_2 = 0$). Under a MAR mechanism, we let $P(R_1 = r_1, R_2 = r_2|\boldsymbol{X})$ denote the joint probability of $(R_1, R_2)|\boldsymbol{X}$ and $\pi_0(\boldsymbol{X}) = P(R_1 = 1, R_2 = 1|\boldsymbol{X})$ be the probability of pattern 0 ($R_1 = 1, R_2 = 1$) given $\boldsymbol{X}$. Let $\pi_1(\boldsymbol{X})$, $\pi_2(\boldsymbol{X})$ and $\pi_3(\boldsymbol{X})$ be the conditional probabilities of the other three patterns.

Assuming a parametric model $p(Y_1, Y_2|\boldsymbol{X}; \boldsymbol{\eta})$ with true value $\boldsymbol{\eta}_0 \in \mathcal{R}^q$, $q > p + 1$, we let $\boldsymbol{V}(Y_1, Y_2|\boldsymbol{X}) = \partial \log p(Y_1, Y_2|\boldsymbol{X}; \boldsymbol{\eta})/\partial \boldsymbol{\eta}$, $\boldsymbol{V}_1(Y_2|Y_1, \boldsymbol{X}) = \partial \log p(Y_2|Y_1, \boldsymbol{X}; \boldsymbol{\eta})/\partial \boldsymbol{\eta}$, and $\boldsymbol{V}_2(Y_1|Y_2, \boldsymbol{X}) = \partial \log p(Y_1|Y_2, \boldsymbol{X}; \boldsymbol{\eta})/\partial \boldsymbol{\eta}$, and write the observed data score vector as $\boldsymbol{V}_0 = R\boldsymbol{V} + R_1(1 - R_2)\boldsymbol{V}_1 + (1 - R_1)R_2\boldsymbol{V}_2$. The main steps of MI2 are as follows:

B1. Fit the model $p(Y_1, Y_2|\boldsymbol{X}; \boldsymbol{\eta})$ using all available data, or, equivalently, solve the estimating equation using the observed data score function $\boldsymbol{V}_0$, to obtain the MLE $\widehat{\boldsymbol{\eta}}$. This step can be implemented by an expectation-maximization algorithm (Dempster et al., 1977).

B2. Impute the missing values by the following strategy:

- If $R_{1i} = 1, R_{2i} = 0$, impute the missing Y_{2i} by drawing a random sample from $p(Y_{2i}|Y_{1i}, \boldsymbol{X}_i; \widehat{\boldsymbol{\eta}})$;
- If $R_{1i} = 0, R_{2i} = 1$, impute the missing Y_{1i} by drawing a random sample from $p(Y_{1i}|Y_{2i}, \boldsymbol{X}_i; \widehat{\boldsymbol{\eta}})$;
- If $R_{1i} = 0, R_{2i} = 0$, impute the missing (Y_{1i}, Y_{2i}) by drawing a random sample from $p(Y_{1i}, Y_{2i}|\boldsymbol{X}_i; \widehat{\boldsymbol{\eta}})$.

B3. Repeat step B2 M times and define the estimator $\widehat{\boldsymbol{\theta}}_{\rm mi2}$ as the solution to equation (7) where $\bar{S}^i$ is replaced by

$$M^{-1} \sum_{m=1}^{M} \{R_i \cdot \boldsymbol{S}(Z_i|\boldsymbol{X}_i; \boldsymbol{\theta}) + (1 - R_i) \cdot \boldsymbol{S}(\tilde{Z}_i^m(\widehat{\boldsymbol{\eta}})|\boldsymbol{X}_i; \boldsymbol{\theta})\},$$

in which the second term includes three different patterns: $\tilde{Z}_i^m(\widehat{\boldsymbol{\eta}}) = T_2(Y_{1i}, \tilde{Y}_{2i}^m(\widehat{\boldsymbol{\eta}}), \boldsymbol{X}_i)$ if $R_1 = 1$ and $R_2 = 0$, $\tilde{Z}_i^m(\widehat{\boldsymbol{\eta}}) = T_2(\tilde{Y}_{1i}^m(\widehat{\boldsymbol{\eta}}), Y_{2i}, \boldsymbol{X}_i)$ if $R_1 = 0$ and $R_2 = 1$, and finally $\tilde{Z}_i^m(\widehat{\boldsymbol{\eta}}) = T_2(\tilde{Y}_{1i}^m(\widehat{\boldsymbol{\eta}}), \tilde{Y}_{2i}^m(\widehat{\boldsymbol{\eta}}), \boldsymbol{X}_i)$ if $R_1 = 0$ and $R_2 = 0$.

As in Theorem 1, the estimator $\widehat{\boldsymbol{\theta}}_{\rm mi2}$ is also $\sqrt{n}$-consistent, but we skip the details; the influence function of $\widehat{\boldsymbol{\theta}}_{\rm mi2}$ is given in the Appendix. The asymptotic efficiency results are as follows, where we write $\widehat{\boldsymbol{\theta}}_{\rm mi2}$ as $\widehat{\boldsymbol{\theta}}_{\rm mi2}(\pi_0, \pi_1, \pi_2)$ as necessary.

Proposition 2. *Under the assumption that the model $p(Y_1, Y_2|\boldsymbol{X}; \boldsymbol{\eta})$ is correctly specified, we have*

1. *The estimator $\widehat{\boldsymbol{\theta}}_{\rm mi2}$ is more efficient than $\widehat{\boldsymbol{\theta}}_{\rm mi1}$, hence, more efficient than $\widehat{\boldsymbol{\theta}}_{\rm mi0}$, i.e.,* $\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\rm mi2}) \leq \mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\rm mi1}) \leq \mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\rm mi0})$;

2. *The estimator* $\widehat{\boldsymbol{\theta}}_{\text{mi2}}(\pi_0, \pi_1 > 0, \pi_2 > 0)$ *is more efficient than* $\widehat{\boldsymbol{\theta}}_{\text{mi2}}(\pi_0, \pi_1 = 0, \pi_2 = 0)$*;*

3. *The estimator* $\widehat{\boldsymbol{\theta}}_{\text{mi2}}(\pi_0, \pi_1 = 0, \pi_2 = 0)$ *is more efficient than* $\widehat{\boldsymbol{\theta}}_{\text{mi1}}$.

There are three interesting observations from Proposition 2. First, since the estimator $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$ is built upon a more informative model $p(Y_1, Y_2|\boldsymbol{X}; \boldsymbol{\eta})$, it generally leads to more efficient estimation. Second, the partially observed height/weight data in patterns 1 and 2 yield a more efficient estimator of $\boldsymbol{\eta}$ and therefore can be used to improve efficiency. Third, even in the special case where $\pi_1 = \pi_2 = 0$, since the estimator of $\boldsymbol{\gamma}$ derived from the joint (Y_1, Y_2) model is more efficient than from the model for Y, the estimator $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$ is more efficient than $\widehat{\boldsymbol{\theta}}_{\text{mi1}}$.

Note that the unbiasedness of $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$ depends on correct specification of the bivariate (Y_1, Y_2) model so it is natural to ask what model for $p(Y_1, Y_2|\boldsymbol{X}; \boldsymbol{\eta})$ matches that of $p(Y|\boldsymbol{X}; \boldsymbol{\gamma})$ and $p(Z|\boldsymbol{X}; \boldsymbol{\theta})$. The case of the probit link is straightforward due to properties of the multivariate normal distribution. For example, if

$$Y_k = \boldsymbol{X}^T \boldsymbol{\alpha}_k + \sigma_k W_k, \quad k = 1, 2, \tag{8}$$

where $(W_1, W_2) \perp \boldsymbol{X}$ is bivariate normal with mean zero, variance one and correlation ρ, then $Y = \boldsymbol{X}^T \boldsymbol{\alpha} + \sigma W$, where $\boldsymbol{\alpha} = \boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2$, $\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$, and W is standard normal. The case of the logit link is more complicated; see the Appendix.

It is challenging to identify a correct bivariate model for (Y_1, Y_2). Moreover different (Y_1, Y_2) models may result in the same model for Y. Without strong rationale for a particular model, we suggest use of a bivariate normal model; its performance is investigated in our simulation studies of Section 5.

To provide some protection for misspecification of the imputation model, we next present a shrinkage estimator. Of the four estimators $\widehat{\boldsymbol{\theta}}_{\text{cc}}$, $\widehat{\boldsymbol{\theta}}_{\text{mi0}}$, $\widehat{\boldsymbol{\theta}}_{\text{mi1}}$ and $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$, the first three do not require specification of a model for $(Y_1, Y_2)|\boldsymbol{X}$, but $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$ does. The shrinkage estimator is created based on one of the first three estimators and $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$. Since $\widehat{\boldsymbol{\theta}}_{\text{mi1}}$ generally has better performance than $\widehat{\boldsymbol{\theta}}_{\text{cc}}$ and $\widehat{\boldsymbol{\theta}}_{\text{mi0}}$, the proposal in this subsection builds on $\widehat{\boldsymbol{\theta}}_{\text{mi1}}$ and $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$.

Let $\boldsymbol{\theta}^*$ denote the limiting value of $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$, which need not be $\boldsymbol{\theta}_0$ when the model for $(Y_1, Y_2)|\boldsymbol{X}$ is misspecified (Robins & Wang, 2000). From Theorem 1, both $\widehat{\boldsymbol{\theta}}_{\text{mi1}}$ and $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$ have asymptotic linear representations. For simplicity, we denote

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\text{mi1}} - \boldsymbol{\theta}_0) = n^{-1/2} \sum_{i=1}^{n} A_i + o_p(1), \text{ and } \sqrt{n}(\widehat{\boldsymbol{\theta}}_{\text{mi2}} - \boldsymbol{\theta}^*) = n^{-1/2} \sum_{i=1}^{n} B_i + o_p(1), \tag{9}$$

where A_i and B_i are defined quantities following Theorem 1. To protect against potential estimation bias of $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$, Chen, Chatterjee & Carroll (2009) suggested a shrinkage estimator of the form

$$\widehat{\boldsymbol{\theta}}_{\text{mis}} = \widehat{\boldsymbol{\theta}}_{\text{mi1}} + G(\widehat{\boldsymbol{\theta}}_{\text{mi2}} - \widehat{\boldsymbol{\theta}}_{\text{mi1}}),$$

where the shrinkage factor G is a $p \times p$ matrix which determines the extent of shrinkage from $\widehat{\boldsymbol{\theta}}_{\text{mi1}}$ towards $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$; two extreme cases are with $G = 0$ implying $\widehat{\boldsymbol{\theta}}_{\text{mis}} = \widehat{\boldsymbol{\theta}}_{\text{mi1}}$ and $G = I_p$ implying $\widehat{\boldsymbol{\theta}}_{\text{mis}} = \widehat{\boldsymbol{\theta}}_{\text{mi2}}$.

Let $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\theta}}_{\text{mi2}} - \widehat{\boldsymbol{\theta}}_{\text{mi1}}$ with jth element $\widehat{\delta}_j$ and $V_\delta = \text{Var}(\widehat{\boldsymbol{\delta}})$ with jth diagonal element v_{jj}, $j = 1, \ldots, p$. From the asymptotic linear representation, $V_\delta = n^{-1}\text{Var}(B - A)$. The matrix G is defined as a diagonal matrix with its jth element $G_{jj} = v_{jj}/(v_{jj} + \widehat{\delta}_j^2)$. From (9), $v_{jj} = O_p(n^{-1})$. If $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$, $\widehat{\delta}_j = O_p(n^{-1/2})$, so $G_{jj} = O_p(1)$ and $\widehat{\boldsymbol{\theta}}_{\text{mis}}$ receives weights from both $\widehat{\boldsymbol{\theta}}_{\text{mi1}}$ and $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$. If $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$, $\widehat{\delta}_j = O_p(1)$, so $G_{jj} \to 0$ and $\widehat{\boldsymbol{\theta}}_{\text{mis}} = \widehat{\boldsymbol{\theta}}_{\text{mi1}}$ asymptotically. Therefore $\widehat{\boldsymbol{\theta}}_{\text{mis}}$ is always asymptotically unbiased.

To derive the asymptotic variance of $\widehat{\boldsymbol{\theta}}_{\rm mis}$, we let $\Gamma = (I_p - G, G)$ denote a $p \times 2p$ matrix, and $\widehat{\boldsymbol{\theta}}_{\rm mi} = (\widehat{\boldsymbol{\theta}}^T_{\rm mi1}, \widehat{\boldsymbol{\theta}}^T_{\rm mi2})^T$. Since $\widehat{\boldsymbol{\theta}}_{\rm mis} = \Gamma\widehat{\boldsymbol{\theta}}_{\rm mi}$, $\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\rm mis}) = \Gamma\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\rm mi})\Gamma^T$. The estimation of Γ was discussed earlier, while an estimate of $\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\rm mi}) = n^{-1}\mathrm{Cov}\{(A^T, B^T)^T\}$ can be derived from the asymptotic linear representation (9).

5 Simulation Studies

We conducted two simulation studies to examine the finite sample performance of the proposed methods based on frequentist MI, and include comparisons to some naive methods. For the first simulation study, we generate bivariate (Y_1, Y_2) based on the model:

$$Y_k = \alpha_{k0} + \alpha_{k1}X + \sigma_k W_k, \;\; k = 1, 2, \tag{10}$$

where $(W_1, W_2) \perp X$, $X \sim N(1, 1)$ and (W_1, W_2) is bivariate normal with zero means, unit variance and correlation ρ. If $Y = Y_1 - Y_2$, then

$$Y = \alpha_0 + \alpha_1 X + \sigma W \ ,$$

where W is standard normal, $\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$, $\alpha_0 = \alpha_{10} - \alpha_{20}$ and $\alpha_1 = \alpha_{11} - \alpha_{21}$. We set $\alpha_{10} = -1$, $\alpha_{11} = 2$, $\alpha_{20} = 0$, $\alpha_{21} = 1$, $\sigma_1 = \sigma_2 = 1$, and $\rho = 0.6$ or 0, use common cut points $c_1 = 0$ and $c_2 = 1$, and let $Z_1 = I(Y \le c_1)$, $Z_2 = I(c_1 < Y \le c_2)$, and $Z_3 = I(c_2 < Y)$.

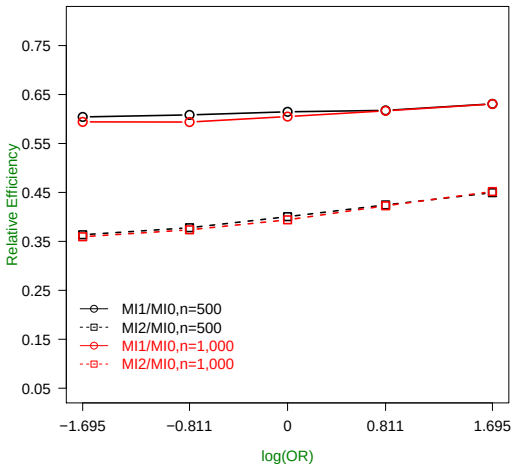
The missing indicators (R_1, R_2) are generated from a model for correlated binary data with margins $P(R_k = 1|X) = \mathrm{expit}(a_{k0} + a_{k1}X)$, $k = 1, 2$, and association $\log\Psi = b_0 + b_1 X$, where

$$\Psi = \frac{P(R_1 = 1, R_2 = 1|X)P(R_1 = 0, R_2 = 0|X)}{P(R_1 = 1, R_2 = 0|X)P(R_1 = 0, R_2 = 1|X)}$$

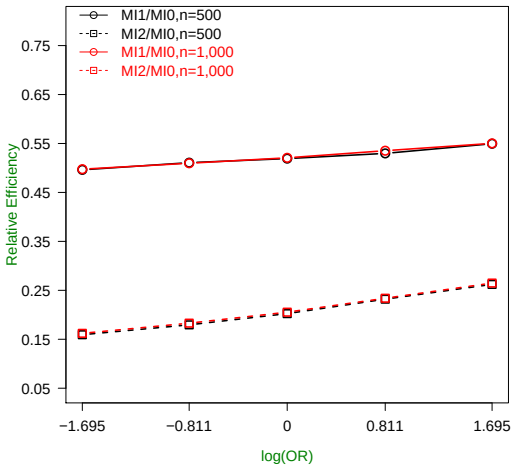
characterizes the association between the missing data indicators for the two components given X. For simplicity, we take $a_{10} = a_{20} = 0$, $a_{11} = a_{21} = 0, 0.421, 1.025$, and 2.779, resulting in marginal probabilities of 0.50, 0.60, 0.70 and 0.80, respectively. We take $b_1 = 0$, and $b_0 = -1.695, -0.811, 0, 0.811, 1.695$ giving odds ratios $\Psi = 0.18, 0.44, 1, 2.25, 5.45$, respectively.

We consider situations with sample sizes of $n = 500$ and $n = 1,000$ with $M = 10$ imputations, and report results based on 3,000 replications. The naive estimator $\widehat{\boldsymbol{\theta}}_{\rm mi0}$, the first proposed estimator $\widehat{\boldsymbol{\theta}}_{\rm mi1}$ and the second proposed estimator $\widehat{\boldsymbol{\theta}}_{\rm mi2}$ are compared in terms of relative efficiency through $\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\rm mi1})/\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\rm mi0})$ and $\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\rm mi2})/\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\rm mi0})$. In Figure 1, $P(R_k = 1), k = 1, 2$ are approximately 50% and the odds ratio Ψ is indicated on the horizontal axis. Here we see that $\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\rm mi1})/\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\rm mi0})$ is less than one with greatest efficiency gains when $\log\Psi < 0$. This finite sample evidence is consistent with our expectation based on the theory in Proposition 1; a similar conclusion is obtained for $\widehat{\boldsymbol{\theta}}_{\rm mi2}$, which is more efficient than $\widehat{\boldsymbol{\theta}}_{\rm mi1}$. Figure 2 shows the relative efficiency of $\widehat{\boldsymbol{\theta}}_{\rm mi1}$ or $\widehat{\boldsymbol{\theta}}_{\rm mi2}$ when the odds ratio is fixed at 1 and show that as $P(R_k = 1)$ increases the efficiency gains decrease. For the coefficient of covariate X, the effect of the correlation ρ is also clear by comparing the left $(\rho = 0)$ and right $(\rho = 0.6)$ panels of Figure 1: greater efficiency gains are realized with larger correlations between Y_1 and Y_2.

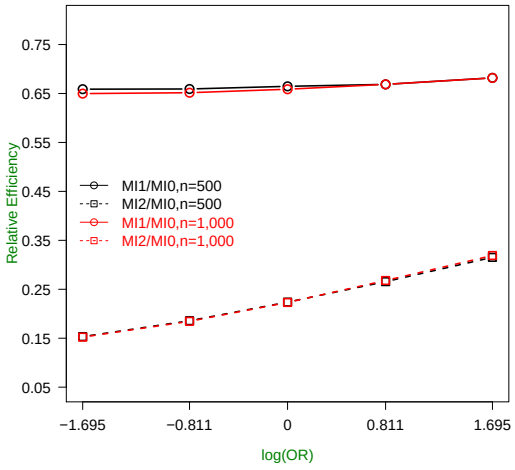
A second simulation study was carried out to evaluate the performance of the shrinkage estimator $\widehat{\boldsymbol{\theta}}_{\rm mis}$ when the (Y_1, Y_2) imputation model is misspecified. We generate bivariate (Y_1, Y_2) according to (10) but with a joint c.d.f. of (W_1, W_2) given by $F(w_1, w_2) = \exp\{-(e^{-w_1/\sigma} + e^{-w_2/\sigma})\}$, $\sigma \in \{1, 1.5\}$; if $Y = Y_1 - Y_2$, then $Y = \alpha_0 + \alpha_1 X + \sigma W$ where W follows a standard logistic distribution; in this scenario, the correct link function for Z is the logit link. We evaluate the performance of the following seven estimators: $\widehat{\boldsymbol{\theta}}_{\rm full}$, based on full data; $\widehat{\boldsymbol{\theta}}_{\rm cc}$, based on completely observed data; $\widehat{\boldsymbol{\theta}}_{\rm mi0}$; $\widehat{\boldsymbol{\theta}}_{\rm mi1}$; $\widehat{\boldsymbol{\theta}}_{\rm mi2}$, based on the convenient but incorrect bivariate normal model for (Y_1, Y_2); $\widehat{\boldsymbol{\theta}}^*_{\rm mi2}$, based on



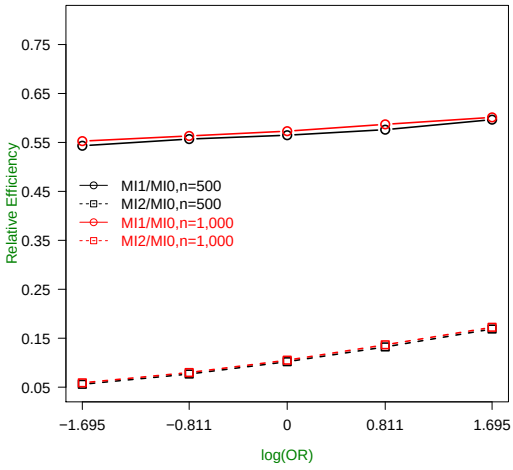
(a) Intercept for $Z|X$



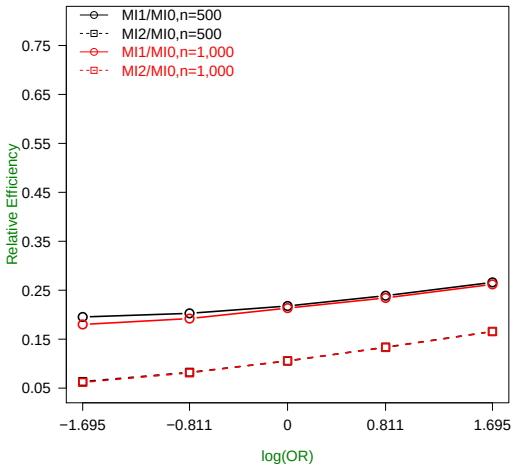
(b) Intercept for $Z|X$



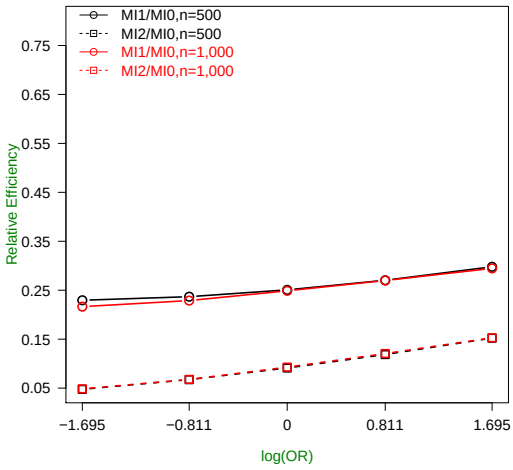
(c) X coefficient for $Z|X$



(d) X coefficient for $Z|X$

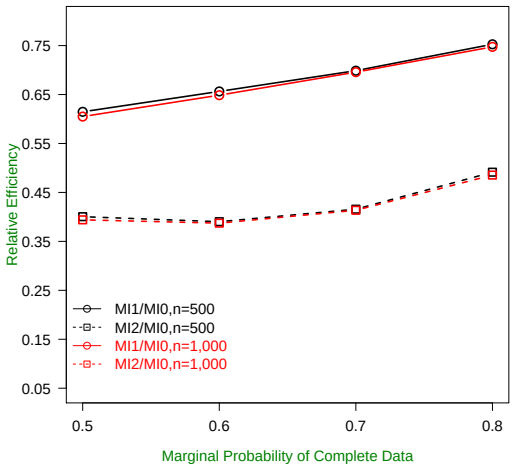


(e) Cut-off coefficient for $Z|X$

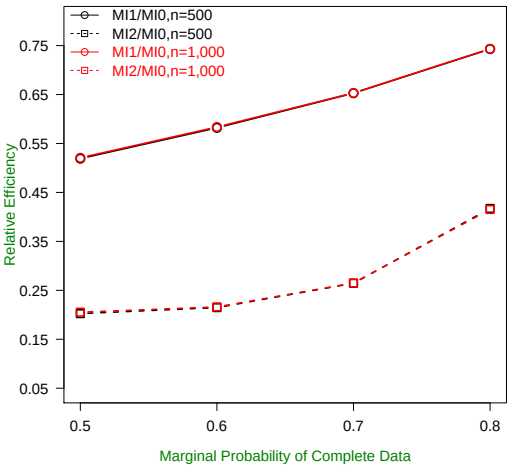


(f) Cut-off coefficient for $Z|X$

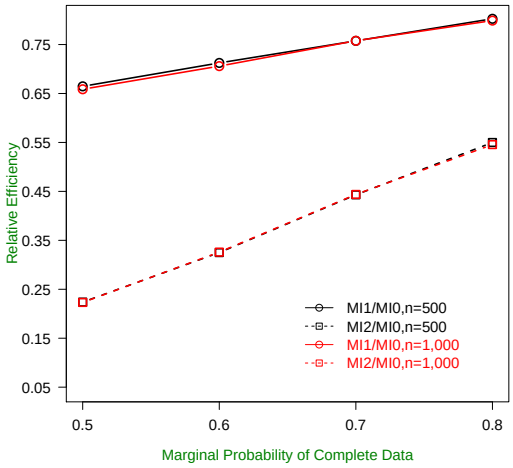
Figure 1: Plots of relative efficiency (RE) versus $\log \Psi$, comparing MI1 and MI2 versus MI0 when the marginal probability of complete data is fixed; $P(R_k = 1) = 0.5, k = 1, 2$. For the left panel, $\rho = 0$ and the right panel $\rho = 0.6$.



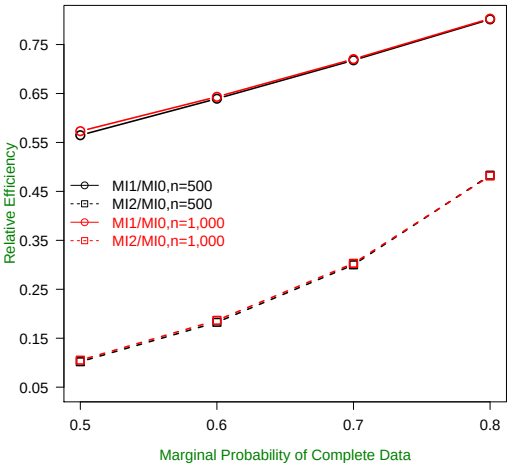
(a) Intercept for $Z|X$



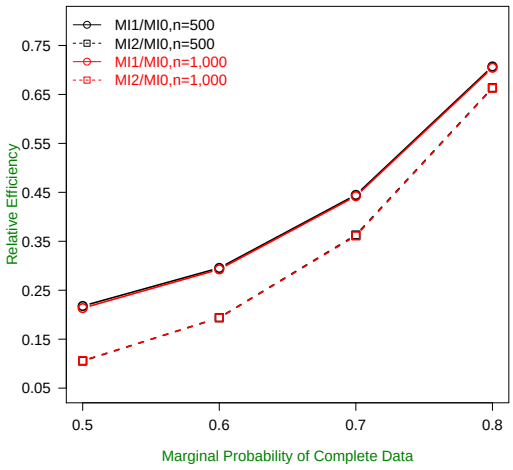
(b) Intercept for $Z|X$



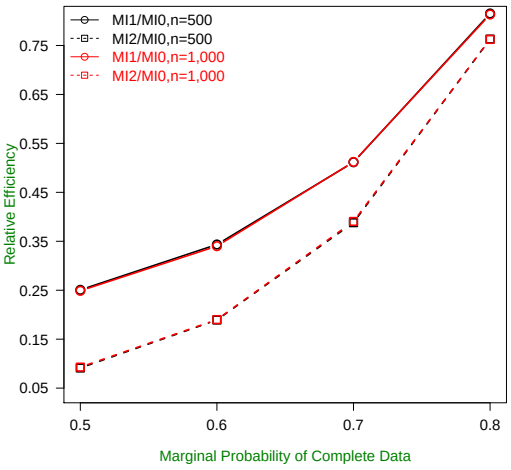
(c) X coefficient for $Z|X$



(d) X coefficient for $Z|X$



(e) Cut-off coefficient for $Z|X$



(f) Cut-off coefficient for $Z|X$

Figure 2: Plots of relative efficiency (RE) versus $P(R_k = 1), k = 1, 2$, comparing MI1 and MI2 vs MI0 when the odds ratio (OR) of R_1 and R_2 is fixed; $\Psi = 1$. For the left panel, $\rho = 0$ and the right panel $\rho = 0.6$.

the correct bivariate Gumbel distribution; and $\widehat{\boldsymbol{\theta}}_{\text{mis}}$, the shrinkage estimator based on $\widehat{\boldsymbol{\theta}}_{\text{mi1}}$ and $\widehat{\boldsymbol{\theta}}_{mi2}$. Our first four estimators $\widehat{\boldsymbol{\theta}}_{\text{full}}$, $\widehat{\boldsymbol{\theta}}_{\text{cc}}$, $\widehat{\boldsymbol{\theta}}_{\text{mi0}}$ and $\widehat{\boldsymbol{\theta}}_{\text{mi1}}$ are all based on the correct logistic distribution for W. The results are summarized in Tables 1 to 4. It can be seen that, $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$ has substantial bias and an empirical coverage probability which is incompatible with the nominal level, since it assumes an incorrect (Y_1, Y_2) model. The empirical bias of $\widehat{\boldsymbol{\theta}}_{\text{mis}}$ is much smaller and the 95% coverage probability of $\widehat{\boldsymbol{\theta}}_{\text{mis}}$ always performs well. For some situations, the performance of $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$ is numerically reasonable, especially when the marginal observed probability is large. The literature suggests that the numerical difference between the probit link and the logit link is small (Chambers & Cox, 1967). The results also show some robustness to misspecification of the bivariate normal imputation model. This point is further investigated in the example that follows.

6 Analysis of Body Mass Index in the Canadian Youth Smoking Survey

The prevalence of overweight and obese aboriginal youth in Canada is high (Hanley et al., 2000; Katzmarzyk, 2008). To examine this phenomenon and investigate the probability of being overweight/obese after compensating for missing data among off-reserve aboriginal (ORA) youth, we analyze a data set from students in Grades 9-12 participating in the 2010 Youth Smoking Survey, a representative sample of ORA in 9 Canadian provinces.

Our analysis includes 1,731 individuals, out of which 1,230 have complete observations ($R_1 = 1, R_2 = 1$); 118 have missing weights ($R_1 = 1, R_2 = 0$); 195 have missing heights ($R_1 = 0, R_2 = 1$); and 188 are missing both ($R_1 = 0, R_2 = 0$). According to WHO growth charts (Onis et al., 2007), the cut points for the overweight and obese designations are gender and age specific, i.e, $c_k = c_k(\texttt{sex, age})$, $i = 1, 2$; see the WHO website:

`http://www.who.int/growthref/who2007_bmi_for_age/en/.`

We consider the following five covariates: `sex` (X_1), `age` (X_2), a variable `sedent` (X_3) which characterizes the sedentary lifestyle of children based on the average time per day engaging in screen time activities (i.e. playing video games, surfing the internet, watching TV), `smoking` (X_4) indicating current smoker or former smoker versus non-smoker, defined consistently with Health Canada definitions for smoking status; and `selfesteem` (X_5), which is a score based on three items from the questionnaires with a higher score indicating higher self-esteem. Since the demographic factors `sex` and `age` are of particular interest to subject area researchers, we fit two models: Model 1 with $\boldsymbol{X} = (1, X_1, X_2, X_3, X_4, X_5)^T$; Model 2 with $\boldsymbol{X} = (1, X_1, X_2)^T$.

We analyze this data set with $p(Z|\boldsymbol{X}; \boldsymbol{\theta})$ modeled by a logit or probit link. First, we examine the normality of the residuals of BMI, weight and height before and after log-transformations based on Model 1 and Model 2 with completely observed data through the Q-Q plots; the data appear to be reasonably compatible with a normal assumption after a log-transformation; see Figure 3.

Five estimators are considered in the analysis: $\widehat{\boldsymbol{\theta}}_{\text{cc}}$, $\widehat{\boldsymbol{\theta}}_{\text{mi0}}$, $\widehat{\boldsymbol{\theta}}_{\text{mi1}}$, $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$, and $\widehat{\boldsymbol{\theta}}_{\text{mis}}$. The estimator $\widehat{\boldsymbol{\theta}}_{\text{mis}}$ shrinks $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$ towards $\widehat{\boldsymbol{\theta}}_{\text{mi1}}$, which does not rely on the specification of the bivariate distribution for (Y_1, Y_2). For each method we report the estimate and standard error (SE), calculated from the asymptotic variance. For the number of imputations, we consider $M = 10$ and $M = 20$. The efficiency gains of three proposed estimators $\widehat{\boldsymbol{\theta}}_{\text{mi1}}$, $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$, and $\widehat{\boldsymbol{\theta}}_{\text{mis}}$ over $\widehat{\boldsymbol{\theta}}_{\text{cc}}$ and $\widehat{\boldsymbol{\theta}}_{\text{mi0}}$ are apparent from Tables 5 and 6. The estimator of the coefficient of `sex` under MI2 with Model 1 and the logit link is different from others, which illustrates the possible effect of imputing with the BVN model. In some situations on the other hand, the performances of $\widehat{\boldsymbol{\theta}}_{\text{mi2}}$ estimators are numerically reasonable, which attests to the practical effectiveness of using the bivariate normal model for imputation.

Finally, we evaluate the estimates of the marginal probabilities of being overweight or obese by sex and age using the five methods based on Model 2. The estimates are plotted in Figures 4 and 5

Table 1: Empirical bias (Bias), standard error (SE) and coverage probability (CP) of seven estimators when $\sigma_0 = 1$, $\boldsymbol{\theta}_0 = (1, -1, 1)^T$; MI2 uses the convenient but incorrect BVN distribution and MI2* represents the method using the correct bivariate Gumbel distribution; all the entries are enlarged 100 times. Sample size $n = 500$.

									Multiple Imputation													
			FULL			CC			MI0			MI1			MI2			MI2*			MIS		
Ψ	$P(R_k=1)$		Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP
0.44	50%	θ_0	-0.1	14.8	95.1	-2.3	34.3	95.4	-2.8	35.4	95.7	-1.8	26.9	95.5	4.9	18.1	87.3	-0.9	16.4	95.0	-1.2	24.3	94.7
		θ_1	0.4	11.3	94.9	3.5	25.8	95.5	4.1	25.9	94.3	2.6	21.1	94.6	-4.2	12.7	89.1	1.1	11.3	95.1	1.3	18.0	95.0
		θ_2	0.2	10.0	95.2	2.3	20.9	95.1	2.7	22.2	95.6	1.9	10.4	94.7	-4.4	8.4	77.9	0.8	5.9	94.8	-1.1	9.2	95.2
0.44	80%	θ_0	-0.9	15.2	95.0	-1.7	23.0	95.0	-2.2	23.4	94.4	-1.7	20.7	94.8	3.9	17.3	89.0	-1.1	15.0	94.8	1.1	18.6	94.7
		θ_1	0.8	10.9	94.8	1.4	15.1	95.8	1.7	15.2	95.8	1.4	14.4	95.4	-2.0	12.3	91.9	1.0	10.8	95.2	1.2	13.3	94.8
		θ_2	0.5	9.0	95.2	0.8	10.3	95.1	0.9	11.1	94.8	0.8	9.1	95.1	0.2	9.0	94.5	0.7	8.7	95.1	0.6	8.9	95.3
1	50%	θ_0	-0.1	14.8	95.0	-3.5	31.1	95.2	-4.2	31.3	95.4	-2.6	24.9	94.9	4.1	17.6	88.7	-1.5	15.8	95.1	-1.7	21.9	95.2
		θ_1	0.5	10.6	94.4	3.5	22.8	94.7	4.1	23.2	95.7	2.5	18.7	94.7	-3.7	14.4	90.0	1.6	12.4	94.9	1.5	16.9	94.7
		θ_2	0.4	8.8	94.7	2.4	18.8	94.5	2.7	19.1	94.3	1.9	9.0	95.2	-3.9	8.4	78.5	1.0	5.9	94.5	1.2	9.2	94.7
1	80%	θ_0	-0.7	15.1	94.9	-1.0	22.9	95.1	-1.4	23.4	95.4	-1.0	20.6	95.3	4.1	18.3	89.1	-0.5	15.5	94.8	0.7	18.5	95.6
		θ_1	0.9	11.0	94.5	1.1	15.0	95.7	1.4	15.1	95.6	1.1	14.4	95.0	-2.1	13.3	93.4	0.8	10.8	95.2	1.0	13.1	95.3
		θ_2	0.9	10.3	95.3	0.9	11.0	94.9	1.0	11.3	95.3	0.9	9.0	95.2	0.3	9.1	95.5	0.8	8.9	94.8	0.7	9.4	94.8
2.25	50%	θ_0	-0.4	15.3	95.2	-2.5	27.9	94.8	-3.0	28.2	94.9	-2.3	21.8	95.1	4.1	16.5	91.3	-1.2	14.8	95.2	-1.4	20.3	95.3
		θ_1	0.2	11.3	95.0	2.2	20.0	95.1	2.7	19.7	95.6	1.9	16.6	95.4	-4.0	13.3	86.2	1.1	11.0	94.9	-1.3	14.8	95.2
		θ_2	0.4	9.2	94.5	1.6	17.4	95.3	1.8	16.8	95.5	1.4	8.8	94.8	-3.9	8.4	75.5	0.9	7.3	95.0	-1.2	8.4	95.0
2.25	80%	θ_0	-0.1	15.0	95.0	-0.5	21.5	95.6	-0.8	23.1	95.3	-0.8	21.2	94.7	4.2	18.2	88.4	-0.5	16.2	94.8	0.9	19.4	94.6
		θ_1	0.4	11.4	94.8	0.6	14.0	94.9	0.9	13.7	94.7	0.8	12.9	94.9	-2.3	12.4	94.3	0.5	11.4	95.1	-1.3	13.3	94.7
		θ_2	0.9	9.4	95.6	1.0	10.0	95.0	1.1	9.9	95.3	1.0	9.1	95.1	0.4	8.8	95.1	0.9	9.0	95.0	0.8	8.9	95.0

Table 2: Empirical bias (Bias), standard error (SE) and coverage probability (CP) of seven estimators when $\sigma_0 = 1$, $\boldsymbol{\theta}_0 = (1, -1, 1)^T$; MI2 uses the convenient but incorrect BVN distribution and MI2* represents the method using the correct bivariate Gumbel distribution; all the entries are enlarged 100 times. Sample size $n = 1,000$.

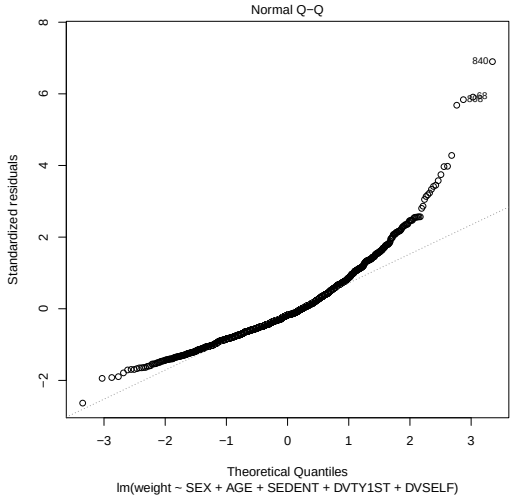
									Multiple Imputation													
			FULL			CC			MI0			MI1			MI2			MI2*			MIS		
Ψ	$P(R_k=1)$		Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP
0.44	50%	θ_0	0.1	10.6	95.0	-1.6	23.5	94.6	-1.9	24.0	94.6	-1.0	18.7	95.2	5.5	12.2	88.5	-0.0	11.4	95.0	0.8	15.8	94.2
		θ_1	-0.1	8.3	94.8	1.2	17.2	95.7	1.5	18.0	95.1	0.6	13.8	94.9	-5.3	8.9	81.0	0.0	8.3	95.1	-1.3	12.1	94.4
		θ_2	0.1	7.3	94.8	1.3	15.0	95.5	1.5	15.1	95.5	1.2	6.7	94.6	-4.7	5.2	72.3	0.4	4.3	95.2	-0.8	6.5	95.6
0.44	80%	θ_0	-0.2	11.4	95.1	-0.8	17.0	94.6	-1.0	17.2	94.7	-0.8	16.0	94.7	4.5	13.1	86.1	-0.5	10.6	95.3	0.9	15.2	95.3
		θ_1	0.2	7.0	95.3	0.6	10.7	95.4	0.7	11.1	95.2	0.6	10.0	95.2	-2.6	9.4	90.4	0.4	8.2	94.8	-0.4	10.0	94.9
		θ_2	0.0	7.1	95.0	0.1	8.0	94.6	0.2	8.3	94.8	0.2	6.4	94.7	-0.4	5.6	94.5	0.1	6.0	94.9	-0.5	5.9	94.3
1	50%	θ_0	-0.3	11.0	94.8	-1.1	21.8	95.0	-1.5	22.2	95.0	-1.0	17.3	95.0	4.3	13.3	90.4	-0.7	10.9	95.2	0.8	16.1	95.4
		θ_1	0.5	7.3	94.9	1.8	14.5	95.2	2.1	15.0	95.2	1.4	12.8	94.8	-4.3	9.2	87.4	0.7	8.1	94.9	-0.8	11.0	94.6
		θ_2	0.4	7.3	94.9	1.1	13.3	95.1	1.3	14.4	95.2	1.0	7.3	95.0	-4.6	5.7	75.6	0.4	5.4	94.8	-0.4	6.2	94.5
1	80%	θ_0	-0.2	11.2	95.1	0.5	15.6	95.0	0.3	16.0	95.6	0.4	15.4	94.8	5.3	12.9	84.7	0.3	11.4	94.8	2.0	14.2	94.6
		θ_1	0.1	7.8	95.2	-0.3	10.0	95.6	-0.2	10.2	95.3	-0.2	10.4	94.7	-3.2	8.8	88.1	-0.2	8.4	95.0	-1.2	9.2	94.7
		θ_2	0.1	7.2	95.0	0.2	7.0	95.4	0.2	8.3	95.1	0.2	6.3	94.9	-0.4	6.0	95.6	0.1	6.0	95.2	-0.3	6.0	95.4
2.25	50%	θ_0	0.0	11.2	95.0	-0.8	19.3	95.2	-1.0	19.8	95.1	-0.5	16.3	95.1	4.9	12.2	89.9	-0.2	11.1	94.8	1.1	14.0	95.2
		θ_1	0.2	8.3	94.7	1.3	14.0	94.8	1.5	14.0	94.7	0.8	11.6	95.3	-4.6	8.7	82.6	0.4	8.2	94.7	-1.0	10.5	94.7
		θ_2	0.1	6.8	94.8	1.0	11.8	95.7	1.2	12.2	95.4	0.9	6.4	94.6	-4.5	5.0	76.4	0.4	5.0	95.1	-0.4	6.1	95.0
2.25	80%	θ_0	-0.0	9.9	95.0	0.1	15.5	95.5	-0.1	16.4	95.3	-0.1	14.2	95.2	4.7	13.3	91.2	-0.1	11.0	95.0	1.5	13.4	94.8
		θ_1	0.3	8.3	95.0	0.2	10.1	94.7	0.4	10.4	94.8	0.3	9.3	94.6	-2.6	9.0	93.4	0.4	8.9	94.8	-0.7	9.0	94.6
		θ_2	0.4	7.3	94.7	0.4	8.4	94.9	0.5	8.4	94.8	0.4	6.2	95.1	-0.2	5.6	95.2	0.4	6.2	95.4	0.7	6.2	94.5

Table 3: Empirical bias (Bias), standard error (SE) and coverage probability (CP) of seven estimators when $\sigma_0 = 1.5$, $\boldsymbol{\theta}_0 = (0.67, -0.67, 0.67)^T$; MI2 uses the convenient but incorrect BVN distribution and MI2* represents the method using the correct bivariate Gumbel distribution; all the entries are enlarged 100 times. Sample size $n = 500$.

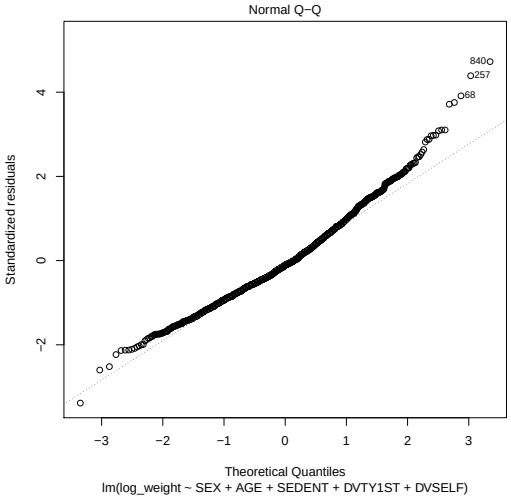
									Multiple Imputation													
			FULL			CC			MI0			MI1			MI2			MI2*			MIS		
Ψ	$P(R_k=1)$		Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP
0.44	50%	θ_0	0.1	13.9	95.1	-1.5	31.8	95.0	-1.9	32.4	95.7	-1.2	26.2	95.6	4.0	15.8	90.4	-0.5	15.4	95.1	-0.6	22.3	95.3
		θ_1	0.3	10.4	94.7	2.6	23.2	95.8	2.9	24.2	95.8	2.0	19.6	95.3	-3.4	12.1	88.7	0.8	10.4	94.8	1.2	16.6	94.7
		θ_2	0.2	8.1	94.8	1.3	16.6	95.6	1.5	17.3	95.3	1.2	7.4	94.8	-3.7	6.4	82.5	0.5	5.5	94.8	-0.6	7.0	95.1
0.44	80%	θ_0	-0.6	12.8	95.2	-1.1	22.5	95.2	-1.5	23.3	94.9	-1.3	21.1	95.0	3.8	16.4	92.3	-0.7	15.3	94.7	0.7	19.2	95.4
		θ_1	0.5	10.3	95.1	0.9	14.0	95.2	1.1	15.1	94.8	1.0	13.0	94.7	-2.2	11.4	93.4	0.6	10.3	95.2	-0.6	12.3	95.1
		θ_2	0.1	6.6	94.7	0.5	7.7	95.2	0.5	8.4	95.0	0.5	7.4	94.7	-0.6	6.9	94.5	0.4	7.0	94.7	0.6	7.0	95.1
1	50%	θ_0	0.3	14.1	95.3	-1.8	28.7	95.0	-2.3	29.4	95.0	-1.6	24.4	95.1	3.5	17.0	92.7	-0.9	15.3	94.8	-1.1	21.1	95.1
		θ_1	0.3	10.3	94.8	2.3	20.5	94.9	2.8	20.9	94.8	1.8	18.3	95.4	-3.1	11.8	93.5	1.1	11.4	95.1	1.3	16.0	95.3
		θ_2	0.6	7.2	94.6	1.8	14.6	94.8	2.0	15.2	94.8	1.5	7.3	94.7	-3.2	6.4	89.5	0.7	5.4	94.5	0.8	6.3	94.4
1	80%	θ_0	-0.4	14.0	95.1	-0.5	22.2	95.0	-0.7	23.3	94.8	-0.7	20.4	95.1	4.1	16.6	93.7	-0.1	15.3	94.6	0.8	19.1	94.9
		θ_1	0.6	9.9	95.0	0.6	13.5	95.2	0.8	14.3	95.1	0.8	13.3	94.8	-2.3	11.7	93.4	0.4	11.4	95.1	-0.8	12.3	94.7
		θ_2	0.6	7.2	94.5	0.8	7.8	94.8	0.9	8.2	94.9	0.7	7.3	95.2	-0.3	7.2	94.8	0.7	7.2	94.7	0.9	7.3	94.6
2.25	50%	θ_0	-0.6	14.2	94.9	-2.1	25.0	95.3	-2.3	26.3	95.4	-1.8	22.2	95.2	3.1	16.4	93.2	-1.0	14.3	94.7	-1.4	19.2	95.1
		θ_1	0.2	10.3	94.8	1.3	18.4	94.7	1.5	19.1	94.7	1.3	16.0	94.9	-3.2	12.2	92.1	0.8	10.4	95.1	-0.8	14.1	95.3
		θ_2	0.1	7.2	94.6	0.7	12.8	94.9	0.7	14.2	95.0	0.9	6.4	94.5	-3.4	6.0	84.9	0.5	5.4	94.8	-0.8	6.1	94.7
2.25	80%	θ_0	-0.1	14.4	95.0	-0.4	22.0	95.3	-0.8	22.4	95.2	-0.6	20.4	94.8	3.9	17.3	90.1	-0.4	16.3	94.6	0.8	19.1	94.9
		θ_1	0.1	9.8	94.7	0.3	13.7	95.2	0.5	14.4	95.2	0.4	13.3	94.7	-2.5	12.0	94.2	0.3	11.4	95.1	-0.6	12.3	94.8
		θ_2	0.1	7.2	94.6	0.2	8.0	94.8	0.3	8.3	94.7	0.4	7.0	94.9	-0.7	7.0	94.5	0.2	6.4	95.2	0.5	7.1	94.8

Table 4: Empirical bias (Bias), standard error (SE) and coverage probability (CP) of seven estimators when $\sigma_0 = 1.5$, $\boldsymbol{\theta}_0 = (0.67, -0.67, 0.67)^T$; MI2 uses the convenient but incorrect BVN distribution and MI2* represents the method using the correct bivariate Gumbel distribution; all the entries are enlarged 100 times. Sample size $n = 1,000$.

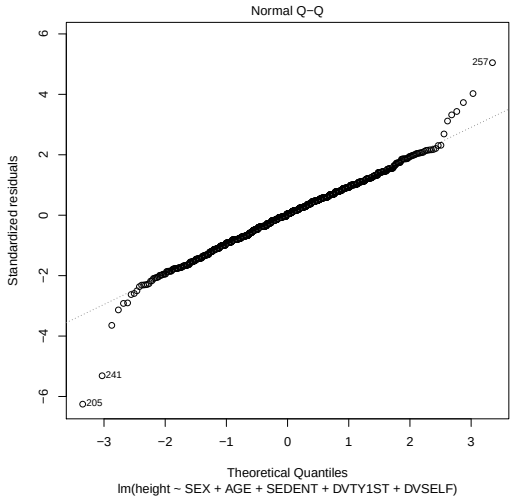
									Multiple Imputation													
			FULL			CC			MI0			MI1			MI2			MI2*			MIS		
Ψ	$P(R_k=1)$		Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP
0.44	50%	θ_0	0.2	10.0	95.0	-1.1	21.8	95.6	-1.3	22.4	95.2	-0.5	18.2	95.1	4.4	11.3	92.5	0.1	10.1	95.1	0.9	16.0	94.7
		θ_1	-0.1	6.9	94.8	0.7	14.8	95.0	0.9	16.2	95.1	0.2	13.3	94.8	-4.4	8.4	84.2	-0.1	7.2	94.9	-1.3	11.1	95.1
		θ_2	0.1	5.4	94.8	1.0	12.0	95.1	1.1	12.3	95.0	0.9	5.3	93.3	-3.9	4.4	76.5	0.3	3.4	94.6	-0.7	5.0	93.9
0.44	80%	θ_0	-0.2	9.9	95.1	-0.5	16.7	94.5	-0.6	17.2	95.1	-0.5	15.2	95.2	4.3	12.4	90.1	-0.2	11.1	95.3	1.0	14.3	94.8
		θ_1	0.3	7.2	94.9	0.5	10.1	95.0	0.6	10.2	95.2	0.5	10.1	95.3	-2.6	8.4	92.3	0.3	7.4	94.7	-0.5	9.2	95.1
		θ_2	0.2	5.3	94.7	0.4	6.1	94.6	0.4	6.1	94.7	0.4	5.2	94.5	-0.7	5.3	93.7	0.4	5.2	94.7	0.5	5.2	94.6
1	50%	θ_0	-0.2	10.0	95.1	-0.7	19.7	95.2	-0.9	20.1	95.1	-0.7	17.2	94.9	3.3	12.4	93.1	-0.5	10.7	94.8	0.9	15.1	94.6
		θ_1	0.4	7.3	94.9	1.3	14.1	95.1	1.5	14.3	95.0	1.0	12.3	94.8	-3.3	8.4	92.6	0.6	6.8	94.7	-0.8	10.8	94.9
		θ_2	0.2	5.4	94.8	0.4	11.1	95.0	0.6	11.2	95.1	0.6	5.0	94.6	-3.8	4.8	82.3	0.2	3.7	95.0	-0.3	5.0	94.8
1	80%	θ_0	-0.2	10.0	95.2	0.6	16.2	95.0	0.4	16.4	94.9	0.5	14.7	94.8	4.9	12.1	92.5	0.3	11.0	95.1	2.0	13.3	94.8
		θ_1	0.1	6.9	94.8	-0.4	9.9	95.1	-0.3	10.2	94.7	-0.3	9.3	94.6	-3.2	8.3	87.4	-0.2	8.4	94.7	-1.3	9.0	95.0
		θ_2	-0.0	5.1	94.9	-0.0	6.0	95.0	-0.0	6.0	95.1	0.0	5.1	95.0	-1.0	5.1	94.0	-0.0	5.1	95.1	-0.3	5.1	95.3
2.25	50%	θ_0	0.1	10.0	95.1	-0.2	18.2	95.2	-0.4	18.3	95.3	-0.1	15.3	95.0	3.9	11.0	87.9	0.1	10.2	95.0	1.1	14.4	95.1
		θ_1	0.1	7.2	94.7	0.8	12.7	94.9	1.0	13.1	94.8	0.5	10.8	94.7	-3.7	9.2	90.4	0.2	8.8	94.9	-0.9	10.0	94.6
		θ_2	0.0	5.0	94.8	0.9	9.1	94.6	1.0	9.4	94.5	0.7	4.3	93.5	-3.6	4.3	82.7	0.4	4.2	94.8	-0.7	4.3	94.5
2.25	80%	θ_0	-0.0	10.0	95.2	0.1	15.1	95.0	-0.0	15.5	95.1	0.0	14.3	94.9	4.3	12.1	91.5	-0.1	11.1	95.1	1.4	13.8	94.8
		θ_1	0.2	6.9	94.8	0.2	10.0	94.9	0.2	10.4	94.8	0.2	9.3	95.4	-2.6	7.7	92.4	0.3	7.8	94.8	-0.7	8.2	95.6
		θ_2	0.2	4.9	94.7	0.3	5.8	95.3	0.3	6.3	95.2	0.3	5.1	95.1	-0.7	5.1	88.6	0.2	5.1	94.8	-0.5	5.2	95.3



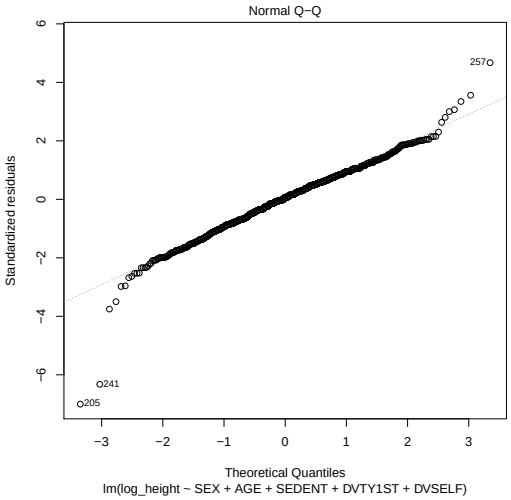
(a) Q-Q plot of "weight"



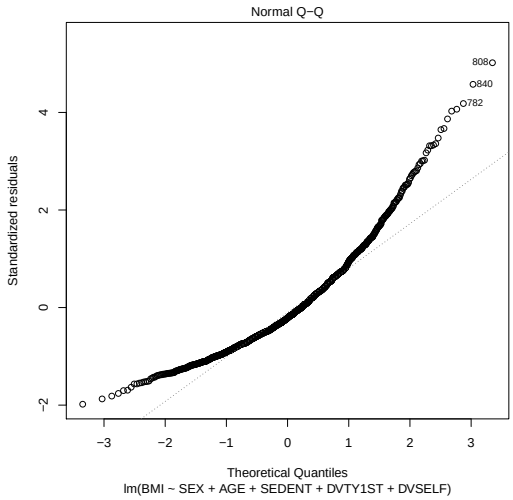
(b) Q-Q plot of "log weight"



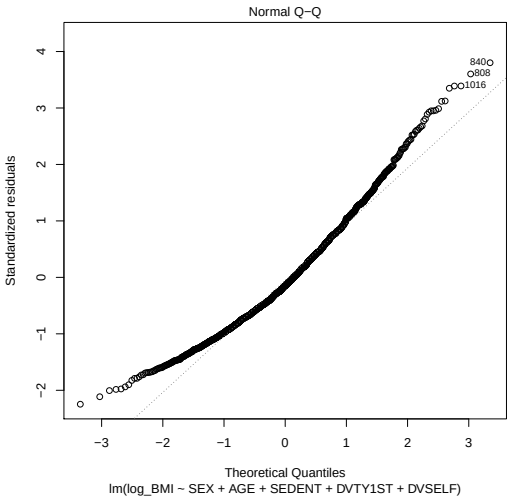
(c) Q-Q plot of "height"



(d) Q-Q plot of "log height"



(e) Q-Q plot of "BMI"



(f) Q-Q plot of "log BMI"

Figure 3: The Q-Q plots of weight, height and BMI before and after log-transformation based on Model 1, fitted on completely observed real data.

Table 5: Estimates and standard errors (SE) based on different models for five methods of analysis of the Canadian Youth Smoking Survey.

				Multiple Imputation							
		CC		MI0		MI1		MI2		MIS	
Model	Covariate	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Model 1	Intercept	-4.261	0.886	-4.252	0.895	-4.526	0.793	-4.463	0.790	-4.467	0.788
Logit link	X_1	-0.414	0.124	-0.413	0.125	-0.409	0.116	-0.360	0.116	-0.390	0.115
	X_2	-0.186	0.048	-0.185	0.049	-0.193	0.045	-0.203	0.045	-0.200	0.045
	X_3	-0.023	0.015	-0.027	0.015	-0.023	0.014	-0.022	0.014	-0.022	0.014
	X_4	0.037	0.076	0.021	0.076	0.023	0.070	0.028	0.070	0.028	0.070
	X_5	0.078	0.026	0.073	0.026	0.078	0.023	0.077	0.023	0.077	0.023
	Cut-offs	0.325	0.018	0.328	0.018	0.342	0.016	0.341	0.015	0.341	0.015
Model 2	Intercept	-3.553	0.799	-3.589	0.807	-3.871	0.713	-3.732	0.712	-3.771	0.705
Logit link	X_1	-0.326	0.119	-0.313	0.120	-0.318	0.112	-0.310	0.113	-0.310	0.113
	X_2	-0.189	0.047	-0.195	0.048	-0.195	0.045	-0.208	0.045	-0.202	0.044
	Cut-offs	0.322	0.018	0.327	0.018	0.338	0.015	0.339	0.015	0.339	0.015
Model 1	Intercept	-2.339	0.518	-2.358	0.523	-2.578	0.469	-2.415	0.472	-2.506	0.465
Probit link	X_1	-0.249	0.073	-0.226	0.074	-0.244	0.069	-0.243	0.070	-0.243	0.070
	X_2	-0.114	0.029	-0.115	0.029	-0.120	0.027	-0.123	0.027	-0.123	0.027
	X_3	-0.013	0.009	-0.012	0.009	-0.013	0.009	-0.009	0.009	-0.011	0.009
	X_4	0.024	0.045	0.009	0.045	0.023	0.041	0.013	0.042	0.017	0.041
	X_5	0.049	0.015	0.052	0.015	0.051	0.013	0.050	0.014	0.050	0.014
	Cut-offs	0.187	0.010	0.187	0.010	0.199	0.008	0.196	0.008	0.197	0.008
Model 2	Intercept	-1.950	0.466	-1.901	0.471	-2.193	0.415	-2.041	0.424	-2.125	0.413
Probit link	X_1	-0.194	0.070	-0.186	0.071	-0.218	0.066	-0.192	0.067	-0.207	0.066
	X_2	-0.113	0.028	-0.115	0.028	-0.115	0.026	-0.122	0.026	-0.119	0.026
	Cut-offs	0.185	0.010	0.185	0.010	0.196	0.008	0.194	0.008	0.195	0.008

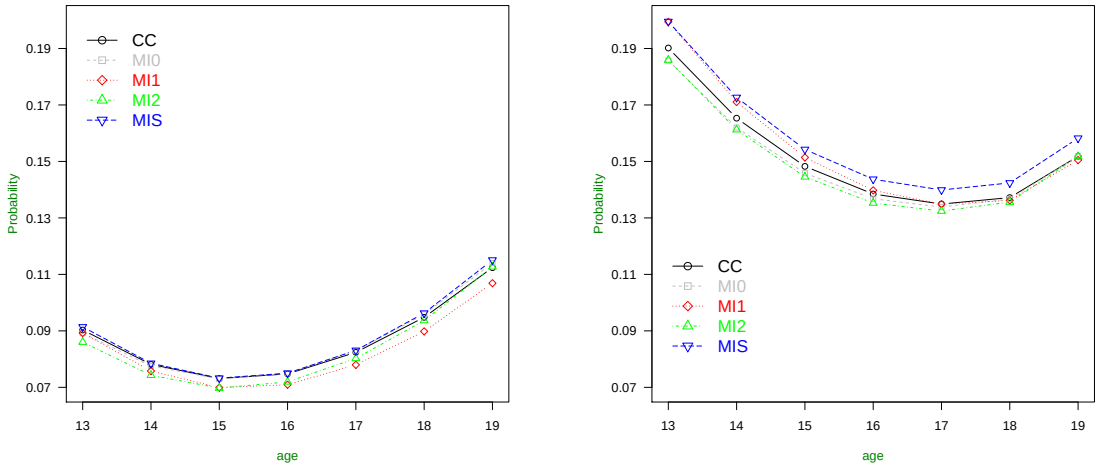
MI2 stands for MI2 method using the bivariate normal model for imputation. $M = 10$.

Table 6: Estimates and standard errors (SE) based on different models for five methods of analysis of the Canadian Youth Smoking Survey.

				Multiple Imputation							
		CC		MI0		MI1		MI2		MIS	
Model	Covariate	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Model 1	Intercept	-4.261	0.886	-4.219	0.891	-4.702	0.786	-4.157	0.789	-4.616	0.782
Logit link	X_1	-0.414	0.124	-0.404	0.124	-0.397	0.115	-0.397	0.116	-0.397	0.116
	X_2	-0.186	0.048	-0.191	0.049	-0.193	0.045	-0.209	0.046	-0.200	0.045
	X_3	-0.023	0.015	-0.024	0.015	-0.022	0.014	-0.022	0.014	-0.022	0.014
	X_4	0.037	0.075	0.033	0.076	0.047	0.069	0.005	0.070	0.040	0.069
	X_5	0.078	0.026	0.078	0.026	0.080	0.023	0.079	0.023	0.079	0.023
	Cut-offs	0.325	0.018	0.327	0.018	0.345	0.015	0.337	0.015	0.343	0.015
Model 2	Intercept	-3.553	0.799	-3.519	0.803	-3.922	0.704	-3.661	0.706	-3.825	0.698
Logit link	X_1	-0.326	0.119	-0.334	0.119	-0.321	0.111	-0.336	0.112	-0.333	0.111
	X_2	-0.189	0.047	-0.188	0.048	-0.194	0.044	-0.202	0.044	-0.200	0.044
	Cut-offs	0.322	0.018	0.320	0.018	0.340	0.015	0.335	0.015	0.338	0.015
Model 1	Intercept	-2.339	0.518	-2.329	0.521	-2.590	0.465	-2.438	0.472	-2.521	0.463
Probit link	X_1	-0.249	0.073	-0.235	0.073	-0.236	0.068	-0.241	0.069	-0.240	0.069
	X_2	-0.114	0.029	-0.116	0.029	-0.119	0.026	-0.123	0.027	-0.123	0.027
	X_3	-0.013	0.009	-0.014	0.009	-0.013	0.009	-0.012	0.009	-0.012	0.009
	X_4	0.024	0.045	0.023	0.045	0.024	0.041	0.017	0.042	0.019	0.041
	X_5	0.049	0.015	0.049	0.015	0.051	0.013	0.052	0.014	0.052	0.014
	Cut-offs	0.187	0.010	0.187	0.010	0.198	0.008	0.196	0.008	0.197	0.008
Model 2	Intercept	-1.950	0.466	-1.958	0.469	-2.160	0.417	-1.918	0.422	-2.114	0.416
Probit link	X_1	-0.194	0.070	-0.190	0.071	-0.203	0.066	-0.191	0.067	-0.194	0.066
	X_2	-0.113	0.028	-0.117	0.028	-0.117	0.026	-0.126	0.026	-0.120	0.026
	Cut-offs	0.185	0.010	0.188	0.010	0.196	0.008	0.192	0.008	0.195	0.008

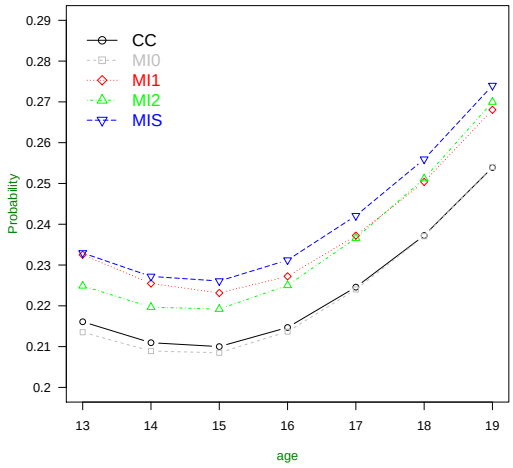
MI2 stands for MI2 method using the bivariate normal model for imputation. $M = 20$.

for the probit link. Conclusions based on the logit link or probit link are broadly similar, with the probability of being overweight or obese higher for males than females. For most situations, after compensating for missing data, the probability of being overweight increases slightly, while that of being obese decreases.

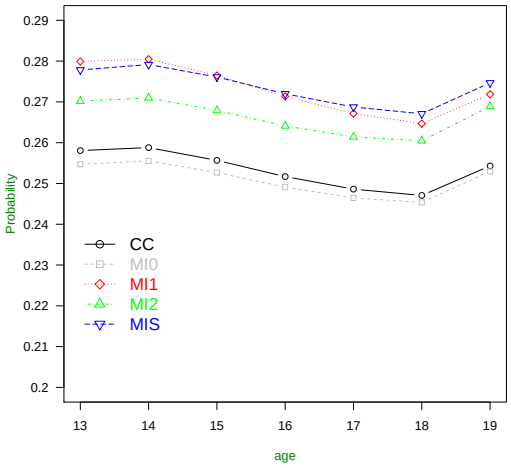


(a) Prevalence of Obesity (Female)

(b) Prevalence of Obesity (Male)



(c) Prevalence of Overweight (Female)

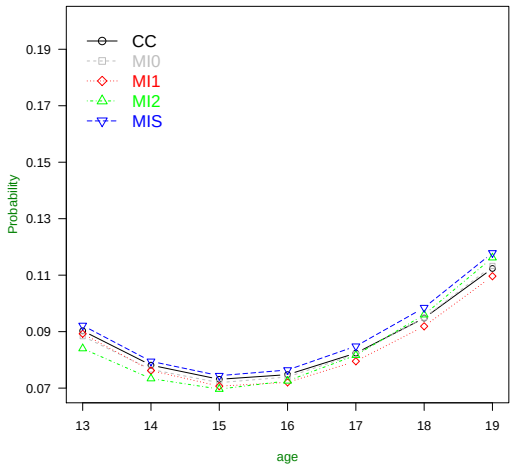


(d) Prevalence of Overweight (Male)

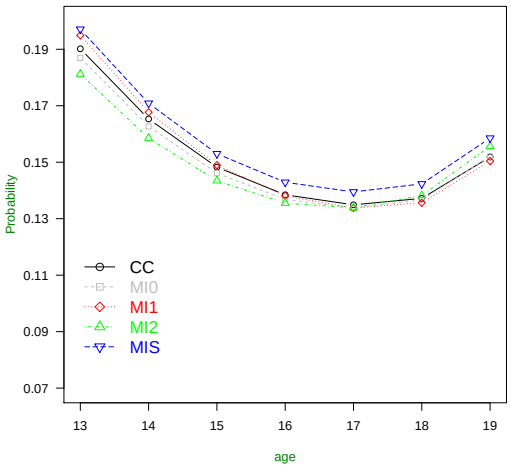
Figure 4: Plots of prevalence of overweight/obesity for female/male, aging 13-19 years old, comparing five methods based on Model 2 and Probit link. $M = 10$.

7 Discussion

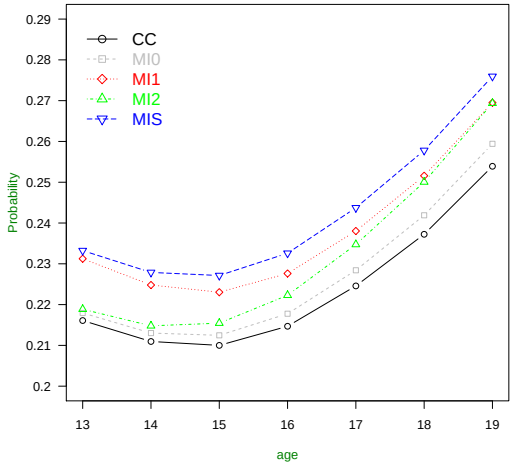
In this article we have discussed imputation strategies that exploit the multivariate nature of incomplete components of compound random variables. The efficiency gain realized comes from both the stronger model assumptions and the consequent better use of the available information regarding missing components. The study was restricted to a two-dimensional setting, but compound variables involving higher dimensional component variables routinely arise. When individual components are incomplete the net effect on the completeness of the compound variable can be substantial and so efficient imputation strategies are warranted.



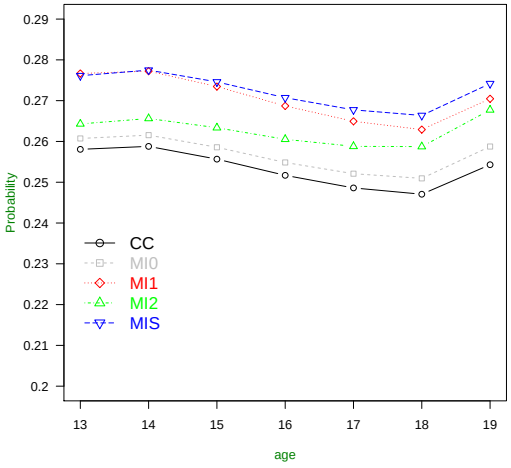
(a) Prevalence of Obesity (Female)



(b) Prevalence of Obesity (Male)



(c) Prevalence of Overweight (Female)



(d) Prevalence of Overweight (Male)

Figure 5: Plots of prevalence of overweight/obesity for female/male, aging 13-19 years old, comparing five methods based on Model 2 and Probit link. $M = 20$.

We adopt a multiple imputation technique to link models with responses Z, Y and (Y_1, Y_2) through their functional relations $Z = T_1(Y, \boldsymbol{X})$ and $Z = T_2(Y_1, Y_2, \boldsymbol{X})$. An EM algorithm could also be used in this setting but may not be as convenient as a MI procedure. Note also that at the first step of the MI2 procedure introduced in Section 4, an EM algorithm was used to find the preliminary MLE $\widehat{\boldsymbol{\eta}}$ in the model $p(Y_1, Y_2|\boldsymbol{X}; \boldsymbol{\eta})$ with partially observed (Y_1, Y_2) data.

While descriptive analyses of the distribution of body mass index were of interest here, this variable is often a covariate in regression models aiming to examine the association between obesity and other risk factors. In clinical studies it may be of interest to study the effect of obesity on risk of disease progression in diabetes, heart disease, and cancer. The multiple imputation strategies we develop can also be examined in these settings.

Finally we assume MAR throughout the whole paper. The procedure we proposed can be readily extended to deal with the nonignorable missing data scenario but the imputation models cannot be estimated based on available data. Analysts must therefore either incorporate external information to fit imputation models or specify them and view the resulting analyses as sensitivity analyses. Statistical analysis with nonignorable missing data had been investigated in the literature, see, for example, Glynn, Laird & Rubin (1993), Siddique & Belin (2008) and Zhao & Shao (2014). These ideas warrant further study for the case of incomplete compound variables.

APPENDIX

We first give the regularity conditions and then provide a sketch of the proof of Theorem 1.
Condition A for the Imputation Model: $\partial U(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}^T$ exists and is bounded in L^2;
Condition B for the Analysis Model: $\partial S(\boldsymbol{\theta}, \boldsymbol{\gamma})/\partial \boldsymbol{\theta}^T$ exists and is bounded in L^2; denote $e(\boldsymbol{\theta}, \boldsymbol{\gamma}) = E[S(\boldsymbol{\theta}, \boldsymbol{\gamma})]$, where $\partial e(\boldsymbol{\theta}, \boldsymbol{\gamma})/\partial \boldsymbol{\gamma}^T$ is continuous in $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ and $\partial e(\boldsymbol{\theta}, \boldsymbol{\gamma})/\partial \boldsymbol{\theta}$ is continuous in $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ and is nonsingular; there exists a $d > 0$, such that $E\{S(\boldsymbol{\theta}, \boldsymbol{\gamma})^{2+d}\}$ is finite;
Condition C for Stochastic Equicontinuity: Define

$$\mathcal{L}_n(\boldsymbol{\gamma}) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \left[\bar{S}^i(\boldsymbol{\theta}_0, \boldsymbol{\gamma}) - e(\boldsymbol{\theta}_0, \boldsymbol{\gamma})\right],$$

for every ϵ_1, $\epsilon_2 > 0$, there exists a $\epsilon > 0$ and an n_0 such that

$$P\left\{\sup_{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 : \|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\| < \epsilon} \|\mathcal{L}_n(\boldsymbol{\gamma}_1) - \mathcal{L}_n(\boldsymbol{\gamma}_2)\| \geq \epsilon_1\right\} \leq \epsilon_2$$

for all $n > n_0$, where $\|\cdot\|$ denotes the Euclidean norm.

Proof of Theorem 1. For the m-th complete data set, if $S_m^i(\boldsymbol{\theta}, \widehat{\boldsymbol{\gamma}}) = R_i \boldsymbol{S}(Z_i|\boldsymbol{X}_i; \boldsymbol{\theta}) + (1 - R_i)\boldsymbol{S}(\tilde{Z}_i^m(\widehat{\boldsymbol{\gamma}})|\boldsymbol{X}_i; \boldsymbol{\theta})$, then the estimator $\widehat{\boldsymbol{\theta}}^{*m}$ solves the equation $\frac{1}{n}\sum_{i=1}^{n} S_m^i(\boldsymbol{\theta}, \widehat{\boldsymbol{\gamma}}) = 0$. By condition B on the analysis model and the mean value theorem, we have

$$0 = \frac{1}{n}\sum_{i=1}^{n} S_m^i(\widehat{\boldsymbol{\theta}}^{*m}, \widehat{\boldsymbol{\gamma}}) = \frac{1}{n}\sum_{i=1}^{n} S_m^i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\gamma}}) + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial S_m^i(\tilde{\boldsymbol{\theta}}^{*m}, \widehat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}^{*m} - \boldsymbol{\theta}_0),$$

where $\tilde{\boldsymbol{\theta}}^{*m}$ is between $\widehat{\boldsymbol{\theta}}^{*m}$ and $\boldsymbol{\theta}_0$. Therefore, we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}^{*m} - \boldsymbol{\theta}_0) = n^{-1/2} I_c^{-1} \sum_{i=1}^{n} S_m^i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\gamma}}) + o_p(1).$$

By combining the M estimators, we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\text{mi1}} - \boldsymbol{\theta}_0) = n^{-1/2} I_c^{-1} \sum_{i=1}^{n} \left\{ M^{-1} \sum_{m=1}^{M} S_m^i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\gamma}}) \right\} + o_p(1).$$

Note that

$$\begin{aligned} & n^{-1/2} \sum_{i=1}^{n} \left\{ M^{-1} \sum_{m=1}^{M} S_m^i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\gamma}}) \right\} \\ = \; & n^{-1/2} \sum_{i=1}^{n} \left\{ M^{-1} \sum_{m=1}^{M} S_m^i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\gamma}}) \right\} - n^{-1/2} \sum_{i=1}^{n} \left\{ M^{-1} \sum_{m=1}^{M} S_m^i(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0) \right\} \\ & + n^{-1/2} \sum_{i=1}^{n} \left\{ M^{-1} \sum_{m=1}^{M} S_m^i(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0) \right\}. \end{aligned} \tag{11}$$

If $\lambda(\boldsymbol{\gamma}, \boldsymbol{\theta}_0) = \mathrm{E}[S_m^i(\boldsymbol{\theta}_0, \boldsymbol{\gamma})]$, then by stochastic equicontinuity (condition C), (11) is

$$\sqrt{n}(\lambda(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\theta}_0) - \lambda(\boldsymbol{\gamma}_0, \boldsymbol{\theta}_0)) + o_p(1) = \frac{\partial \lambda(\boldsymbol{\gamma}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\gamma}^T}\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0} \sqrt{n}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) + o_p(1).$$

It can be shown that $\partial \lambda(\boldsymbol{\gamma}, \boldsymbol{\theta}_0)/\partial \boldsymbol{\gamma}\big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0} = J_c - J_o$, and combined with condition A for the analysis model and (6), we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\text{mi1}} - \boldsymbol{\theta}_0) = n^{-1/2} I_c^{-1} \sum_{i=1}^{n} \left\{ M^{-1} \sum_{m=1}^{M} S_m^i(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0) + (J_c - J_o) Q^i \right\} + o_p(1).$$

Therefore,

$$\begin{aligned} \Omega &= \mathrm{Var}\left[M^{-1} \sum_{m=1}^{M} S_m^i(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0) + (J_c - J_o) Q^i \right] \\ &= M^{-1}(I_c - I_o) + I_o + (J_c - J_o)\mathrm{Var}(Q)(J_c - J_o)^T + \mathrm{E}(S_o Q^T)(J_c - J_o)^T + (J_c - J_o)\mathrm{E}(Q S_o^T) \\ &= M^{-1}(I_c - I_o) + I_o + J_c \mathrm{Var}(Q) J_c^T - J_o \mathrm{Var}(Q) J_o^T. \end{aligned}$$

■

Proof of Proposition 1. To prove

$$\begin{aligned} & \mathrm{E}(S^{\otimes 2})(\mathrm{E}(RS^{\otimes 2}))^{-1}\mathrm{E}(S^{\otimes 2}) - \mathrm{E}(RS^{\otimes 2})(\mathrm{E}(RS^{\otimes 2}))^{-1}\mathrm{E}(RS^{\otimes 2}) \\ \geq \; & \mathrm{E}(SU^T)(\mathrm{E}(RU^{\otimes 2}))^{-1}\mathrm{E}(US^T) - \mathrm{E}(RSU^T)(\mathrm{E}(RU^{\otimes 2}))^{-1}\mathrm{E}(RSU^T)^T \ , \end{aligned}$$

first note that

$$\begin{aligned} & \mathrm{E}(S^{\otimes 2})(\mathrm{E}(RS^{\otimes 2}))^{-1}\mathrm{E}(S^{\otimes 2}) - \mathrm{E}(RS^{\otimes 2})(\mathrm{E}(RS^{\otimes 2}))^{-1}\mathrm{E}(RS^{\otimes 2}) \\ = \; & \mathrm{E}(RS^{\otimes 2})(\mathrm{E}(RS^{\otimes 2}))^{-1}\mathrm{E}((1-R)S^{\otimes 2}) + \mathrm{E}((1-R)S^{\otimes 2})(\mathrm{E}(RS^{\otimes 2}))^{-1}\mathrm{E}(RS^{\otimes 2}) \\ & + \mathrm{E}((1-R)S^{\otimes 2})(\mathrm{E}(RS^{\otimes 2}))^{-1}\mathrm{E}((1-R)S^{\otimes 2}) \ , \end{aligned}$$

hence it is positive definite since both $\mathrm{E}(RS^{\otimes 2})$ and $\mathrm{E}((1-R)S^{\otimes 2})$ are positive definite. If there exists a matrix V, such that $(\mathrm{E}(RS^{\otimes 2}))^{-1} \geq V$, then

$$\begin{aligned} & \mathrm{E}(S^{\otimes 2})(\mathrm{E}(RS^{\otimes 2}))^{-1}\mathrm{E}(S^{\otimes 2}) - \mathrm{E}(RS^{\otimes 2})(\mathrm{E}(RS^{\otimes 2}))^{-1}\mathrm{E}(RS^{\otimes 2}) \\ \geq \; & \mathrm{E}(S^{\otimes 2})V\mathrm{E}(S^{\otimes 2}) - \mathrm{E}(RS^{\otimes 2})V\mathrm{E}(RS^{\otimes 2}), \end{aligned}$$

since both the left-hand side and right-hand side can be written as the summation of three terms as above, and each term on the left-hand side is $\geq$ its right-hand-side counterpart because both $\mathrm{E}(RS^{\otimes 2})$ and $\mathrm{E}((1-R)S^{\otimes 2})$ are positive definite.

Denote the $p \times p$ matrix $K = \partial\boldsymbol{\theta}/\partial\boldsymbol{\gamma}$ evaluated at the true values, giving $K(\mathrm{E}(RU^{\otimes 2}))^{-1}K^T \leq (\mathrm{E}(RS^{\otimes 2}))^{-1}$, since the left hand side is the variance of $\boldsymbol{\theta}$ derived from the Y model, while the right hand side is the variance of $\boldsymbol{\theta}$ derived from the Z model, Z is a function of Y. Therefore

$$\begin{aligned} & \mathrm{E}(S^{\otimes 2})(\mathrm{E}(RS^{\otimes 2}))^{-1}\mathrm{E}(S^{\otimes 2}) - \mathrm{E}(RS^{\otimes 2})(\mathrm{E}(RS^{\otimes 2}))^{-1}\mathrm{E}(RS^{\otimes 2}) \\ \geq\ & \mathrm{E}(S^{\otimes 2})K(\mathrm{E}(RU^{\otimes 2}))^{-1}K^T\mathrm{E}(S^{\otimes 2}) - \mathrm{E}(RS^{\otimes 2})K(\mathrm{E}(RU^{\otimes 2}))^{-1}K^T\mathrm{E}(RS^{\otimes 2}). \end{aligned}$$

To complete the proof, we only need to show $\mathrm{E}(SU^T) = \mathrm{E}(S^{\otimes 2})K$ and $\mathrm{E}(RS^{\otimes 2})K = \mathrm{E}(RSU^T)$. If $a(Z;\boldsymbol{\theta}) = (\mathrm{E}S^{\otimes 2})^{-1}S$, since $\mathrm{E}[a(Z;\boldsymbol{\theta})] = 0$, we have

$$\int a(Z;\boldsymbol{\theta})p(Z;\boldsymbol{\gamma})dZ = 0,$$

which implies

$$0 = \frac{\partial}{\partial\boldsymbol{\gamma}}\int a(Z;\boldsymbol{\theta})p(Z;\boldsymbol{\gamma})dZ = \int \frac{\partial a(Z;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}Kp(Z;\boldsymbol{\gamma})dZ + \int a(Z;\boldsymbol{\theta})Up(Z;\boldsymbol{\gamma})dZ,$$

so, $0 = -K + (\mathrm{E}S^{\otimes 2})^{-1}\mathrm{E}(SU^T)$. The same technique can be used to show $\mathrm{E}(RS^{\otimes 2})K = \mathrm{E}(RSU^T)$.

■

Influence function of $\widehat{\boldsymbol{\theta}}_{\mathrm{mi2}}$.
Introducing an index i to distinguish between different individuals, we note $\widehat{\boldsymbol{\eta}}$ is obtained from solving $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}_0^i = 0$. Therefore,

$$\sqrt{n}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) = n^{-\frac{1}{2}}\sum_{i=1}^{n}T^i + o_p(1),$$

where $T = \{E(\boldsymbol{V}_0^{\otimes 2})\}^{-1}\boldsymbol{V}_0$. Also let $G_c = E(\boldsymbol{S}\boldsymbol{V}^T)$ and $G_o = E(\boldsymbol{S}\boldsymbol{V}_0^T)$. As in the proof of Theorem 1, the asymptotically linear representation of $\widehat{\boldsymbol{\theta}}_{\mathrm{mi2}}$ can be written as

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathrm{mi2}} - \boldsymbol{\theta}_0) = n^{-1/2}I_c^{-1}\sum_{i=1}^{n}\left\{M^{-1}\sum_{m=1}^{M}S_m^i(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) + (G_c - G_o)T^i\right\} + o_p(1),$$

where $S_m^i(\boldsymbol{\theta},\widehat{\boldsymbol{\eta}}) = R_i\cdot\boldsymbol{S}(Z_i|\boldsymbol{X}_i;\boldsymbol{\theta}) + (1-R_i)\cdot\boldsymbol{S}(\tilde{Z}_i^m(\widehat{\boldsymbol{\eta}})|\boldsymbol{X}_i;\boldsymbol{\theta})$.

Proof of Proposition 2. As in Proposition 1, since

$$\begin{aligned} & \mathrm{E}(SV^T)(\mathrm{E}V_o^{\otimes 2})^{-1}\mathrm{E}(SV^T)^T - \mathrm{E}(SV_o^T)(\mathrm{E}V_o^{\otimes 2})^{-1}(\mathrm{E}(SV_o^T))^T \\ \leq\ & \mathrm{E}(SV^T)(\mathrm{E}V_o^{\otimes 2})^{-1}\mathrm{E}(SV^T)^T - \mathrm{E}(RSV^T)(\mathrm{E}V_o^{\otimes 2})^{-1}(\mathrm{E}(RSV^T))^T \end{aligned}$$

is clear, we only need to prove

$$\begin{aligned} & \mathrm{E}(SV^T)(\mathrm{E}V_o^{\otimes 2})^{-1}\mathrm{E}(SV^T)^T - \mathrm{E}(RSV^T)(\mathrm{E}V_o^{\otimes 2})^{-1}(\mathrm{E}(RSV^T))^T \\ \leq\ & \mathrm{E}(SV^T)(\mathrm{E}(RV^{\otimes 2}))^{-1}(\mathrm{E}(SV^T))^T - \mathrm{E}(RSV^T)(\mathrm{E}(RV^{\otimes 2}))^{-1}(\mathrm{E}(RSV^T))^T \\ \leq\ & \mathrm{E}(SU^T)(\mathrm{E}(RU^{\otimes 2}))^{-1}(\mathrm{E}(SU^T))^T - (\mathrm{E}(RSU^T))(\mathrm{E}(RU^{\otimes 2}))^{-1}(\mathrm{E}(RSU^T))^T. \end{aligned}$$

The first inequality is clear since $(\mathrm{E}V_o^{\otimes 2})^{-1} \leq (\mathrm{E}(RV^{\otimes 2}))^{-1}$. The proof of the second inequality is similar to that of Proposition 1. Let $L = \partial\boldsymbol{\gamma}/\partial\boldsymbol{\eta}$ be evaluated at the true values, so $L(\mathrm{E}(RV^{\otimes 2}))^{-1}L^T \leq (\mathrm{E}(RU^{\otimes 2}))^{-1}$. Then,

$$\begin{aligned} & \mathrm{E}(SU^T)(\mathrm{E}(RU^{\otimes 2}))^{-1}(\mathrm{E}(SU^T))^T - (\mathrm{E}(RSU^T))(\mathrm{E}(RU^{\otimes 2}))^{-1}(\mathrm{E}(RSU^T))^T \\ \geq\ & \mathrm{E}(SU^T)L(\mathrm{E}(RV^{\otimes 2})^{-1}L^T\mathrm{E}(SU^T)^T - (\mathrm{E}(RSU^T))L(\mathrm{E}(RV^{\otimes 2}))^{-1}L^T(\mathrm{E}(RSU^T))^T. \end{aligned}$$

Finally, the proof that $\mathrm{E}(SU^T)L = \mathrm{E}(SV^T)$ and $\mathrm{E}(RSU^T)L = (\mathrm{E}(RSV^T))$ is straightforward from Proposition 1.

■

Feasibility of the (Y_1, Y_2) Model with the Logit Link.
Here the problem is to find two random variables, such that their difference follows a logistic distribution. The difference between two independent and identically distributed Gumbel distributed variables is logistic, i.e., if the c.d.f. of (ϵ_1, ϵ_2) is $F(s,t) = \exp\{-(e^{-s/\sigma} + e^{-t/\sigma})\}$, $0 < \sigma < \infty$, then the c.d.f. for $\epsilon = \epsilon_1 - \epsilon_2$ is $\exp(x/\sigma)/\{1 + \exp(x/\sigma)\}$, a logistic distribution.

Result 1. Define a bivariate random variable (ϵ_1, ϵ_2) whose cumulative distribution function is of the form $F(s,t) = \exp\{-(e^{-s/\sigma} + e^{-t/\sigma})^\sigma\}, 0 < \sigma \leq 1$. The c.d.f. of $\epsilon_1 - \epsilon_2$ follows a logistic distribution of the form $\exp(x/\sigma)/\{1 + \exp(x/\sigma)\}$.

Note that the marginal distributions of ϵ_1 and ϵ_2 are Gumbel (or type-I extreme value) with a mean equal to the Euler–Mascheroni constant $\gamma \approx 0.5772$. The positive parameter σ approximately characterizes the correlation between ϵ_1 and ϵ_2, while they become independent when $\sigma = 1$. It is also possible to create (ϵ_1, ϵ_2) whose marginal distributions are different.

Result 2. Suppose (ξ_1, ξ_2) are distributed as above, and define $\epsilon_1 = \xi_1 + t\xi_2$, $\epsilon_2 = (1+t)\xi_2$, then, when $t \neq 0$, $\mathrm{Var}(\epsilon_1) \neq \mathrm{Var}(\epsilon_2)$, and $\epsilon_1 - \epsilon_2$ follows a logistic distribution of the form $\exp(x/\sigma)/\{1 + \exp(x/\sigma)\}$.

ACKNOWLEDGEMENTS

This research is supported by a Fields Institute Postdoctoral Fellowship to Jiwei Zhao, Discovery Grants from the Natural Sciences and Engineering Research Council of Canada to Richard Cook and to Changbao Wu. Richard Cook is a Canada Research Chair in Statistical Methods for Health Research. The authors thank Professors Scott Leatherdale and Tara Elton-Marshall for collaboration and permission to use the data from the Canadian Youth Smoking Survey. The authors thank the Editor, the Associate Editor and two anonymous referees for their insightful comments and useful suggestions, which have significantly improved the quality of the paper.

REFERENCES

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*, 2nd ed., John Wiley & Sons, New York.

Chambers, E. A. & Cox, D. R. (1967). Discrimination between alternative binary response models. *Biometrika*, 54, 573-578.

Chen, Y. H., Chatterjee, N., & Carroll, R. J. (2009). Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of the American Statistical Association*, 104, 220-233.

Cole, T. J. & Green, P. J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine*, 11, 1305-1319.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39, 1-38.

Elton-Marshall, T., Leatherdale, S. T., Currie, C., & Brown, K. S. (2014). An examination of the factors associated with overweight and obesity among off-reserve aboriginal youth in Canada.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd ed., Springer, New York.

Fries, J. F., Spitz, P. W., & Young, D. Y. (1981). The dimensions of health outcomes: the health assessment questionnaire, disability and pain scales. *The Journal of Rheumatology*, 9, 789-793.

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1993). Multiple imputation in mixture models for non-ignorable nonresponse with follow-ups. *Journal of the American Statistical Association*, 88(423), 984–993.

Hanley, A. J., Harris, S. B., Gittelsohn, J., Wolever, T. M., Saksvig, B., & Zinman, B. (2000). Overweight among children and adolescents in a native Canadian community: prevalence and associated factors. *The American Journal of Clinical Nutrition*, 71, 693-700.

Hedley, A. A., Ogden, C. L., Johnson, C. L., Carroll, M. D., Curtin, L. R., & Flegal, K. M. (2004). Prevalence of overweight and obesity among US children, adolescents, and adults, 1999-2002. *Journal of the American Medical Association*, 291, 2847-2850.

Katzmarzyk, P. T. (2008). Obesity and physical activity among aboriginal Canadians. *Obesity*, 16, 184-190.

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

Krebs, N. F., Himes, J. H., Jacobson, D., Nicklas, T. A., Guilday, P., & Styne, D. (2007). Assessment of child and adolescent overweight and obesity. *Pediatrics*, 120, 193-228.

Lamon-Fava, S., Wilson, P. W., & Schaefer, E. J. (1996). Impact of body mass index on coronary heart disease risk factors in men and women the Framingham offspring study. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 16, 1509-1515.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed., John Wiley & Sons, New York.

Lu, K., Jiang, L., & Tsiatis, A. A. (2010). Multiple imputation approaches for the analysis of dichotomized responses in longitudinal studies with missing data. *Biometrics*, 66, 1202-1208.

McHorney, C. A., Johne, W., & Anastasiae, R. (1993). The MOS 36-item short-form health survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care*, 31, 247-263.

Nielsen, S. F. (2003). Proper and improper multiple imputation. *International Statistical Review*, 71, 593-607.

Onis, M. D., Onyango, A. W., Borghi, E., Siyam, A., Nishida, C., & Siekmann, J. (2007). Development of a WHO growth reference for school-aged children and adolescents. *Bulletin of the World Health Organization*, 85, 660-667.

Robins, J. M. & Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In Jewell N. P., Dietz, K., Farewell, V. T. (Ed.) *AIDS Epidemiology: Methodological Issues*, Birkhauser, Boston, 297-331.

Robins, J. M. & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90, 122-129.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866.

Robins, J. M. & Wang, N. (2000). Inference for imputation estimators. *Biometrika*, 87, 113-124.

Rubin, D. B. (1978). Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse (with discussion). *American Statistical Association Proceedings of the Section on Survey Research Methods*, 20-34.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Shen, C. W. & Chen, Y. H. (2013). Model selection of generalized estimating equations with multiply imputed longitudinal data. *Biometrical Journal*, 55, 899-911.

Siddique, J., & Belin, T. R. (2008). Using an approximate Bayesian bootstrap to multiply impute nonignorable missing data. *Computational Statistics & Data Analysis*, 53(2), 405–415.

Stroncek, D. F. & Rebulla, P. (2007). Platelet transfusions. *The Lancet*, 370, 427-438.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*, Springer Science + Business Media, New York.

Wang, N. & Robins, J. M. (1998). Large sample theory for parametric multiple imputation procedures. *Biometrika*, 85, 935-948.

Ware, J. E. & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Medical Care*, 30, 473-483.

Wei, G. C. & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor mans data augmentation algorithms. *Journal of the American Statistical Association*, 85, 699-704.

Zhao, J. & Shao, J. (2015). Semiparametric pseudo likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2014.983234