Statistics and Its Interface Volume 16 (2023) 617–627

Fine-tuned sensitivity analysis for non-ignorable missing data mechanism in linear regression models

Rong Zhu, Peng Yin, and Jian Qing Shi*

Missing data is a widespread problem in many fields, such as statistical analysis in medical research. The missing data mechanism (MDM) is overly complicated in many cases, and the most complex one is the non-ignorable missingness. In this paper, we analyse the incomplete data bias of maximum likelihood estimates on the inference of linear regression models with non-ignorable missing covariate specifically, where the working model always has a small departure from the true model. The incomplete data bias has been divided into two parts because of two types of uncertainties, one is the misspecified distribution between covariates, the other is the misspecified MDM. We identify the key sensitivity parameters in MDM and further propose generative MDM models, leading to a non-implausible set which quantify a smaller range of possible solutions comparing to the conventional sensitivity analysis and worst-case study. Our analysis focuses the sensitivity of MDM modelling in the missing covariate problems. Numerical examples are presented in both simulation studies and a real data example.

AMS 2000 subject classifications: Primary 62D10, 62J05; secondary 62F25.

Keywords and phrases: Bias analysis, Model uncertainty, Non-ignorable missing data, Incomplete data bias, Sensitivity analysis.

1. INTRODUCTION

We consider multi-variable regression analysis of a n-dimensional vector of incomplete data set $\mathcal{D} = (D_1, \dots, D_n)$ where each D_i are independent from the other. Some observations are missing, and a Bernoulli distributed indicator vector R is defined to record the missingness. A standard approach is to assume a parametric model of $[R, \mathcal{D}]$ with parameters estimated by maximum likelihood of the model. Selection model formulation (see Ch.12 in Little and Rubin, 2002) factorizes the distribution of $[R, \mathcal{D}]$ into a model for $[\mathcal{D}]$ and a model for $[R|\mathcal{D}]$, where the latter is well known as the MDM.

When the missing at random (MAR) assumption Little and Rubin (2002) is plausible, the selection model formulation seems compelling because it leads to likelihood ignorable for complete density family. However, as pointed out by Little (1993), valid inference is rather based on the knowledge of MDM; if assumptions about the MDM are misspecified, say from a non-ignorable missing (i.e. missing not at random – MNAR) data model, the parameter estimation will be biased (Diggle and Kenward, 1994). Beyond of the maximum likelihood method, the inverse probability weighted estimating equations approach (Robins, Rotnitzky and Zhao, 1994) and the multiple imputation method (Ibrahim et al., 2005) also require correct specification of the MDM to ensure unbiased analysis. Maity, Pradhan and Das (2019) proposed a likelihood-based method to reduce the bias of the estimation of logistic regression with non-ignorable missing responses, where the penalty term is by multiplying the likelihood by a non-informative Jeffreys prior.

A specific model for non-ignorable missing data is required for valid inference, which may be fitted by a parametric model (e.g. logistic model, probit model) or a semi-parametric model (see e.g. Kim and Yu, 2011). Troxel, Ma and Heitjan (2004) derived an index of sensitivity to non-ignorability to capture the extent of sensitivity under the assumption of ignorability based on a Taylor-series approximation to the non-ignorable likelihood. Wang, Shao and Kim (2014) applied a non-response instrument variable for the generalized method of moments to deal with the estimation and identification with non-ignorable non-response. Zhao and Shao (2015) proposed a semi-parametric pseudo-likelihood approach with an instrumental variable to identify and estimated parameters in the generalized linear model with non-ignorable missing data. Gao et al. (2016) developed the non-linear sensitivity indices to study with the non-ignorable missingness in both outcome and covariates. Guo, Ma and Wang (2020) constructed the bias-corrected semi-parametric estimating equations for the copula models with non-ignorable missing data to improve the efficiency of the estimation. Yuan et al. (2020) defined the non-linear indexes of local sensitivity to non-ignorability for sensitivity analysis to non-ignorable missingness in intensive longitudinal data, which is a computationally feasible method. Zhang

*Corresponding author.

and Wang (2020) proposed a smoothed weighted empirical likelihood to estimate the coefficient for quantile regression with non-ignorable missing. However, it is not always theoretically possible to *"characterize the set of all estimable parameters for this class of models given a certain choice of variables"* (see p.14 in Ibrahim and Molenberghs, 2009). The problem with these concerns as we may encounter in the missing data comprises the non-identifiability and sensitivity of MDM. A critical look at handling missing covariate data in epidemiological studies was presented in Vach and Blettner (1991) and then Greenland and Finkle (1995), where the sensitivity of MDM and misspecification of covariate models were investigated by simulation studies. To date, there are not enough literatures investigating the problem of MDM sensitivity in theory and exploring the potential key risk of misspecified non-ignorable missing data. Thus, we focus on local sensitivity evaluation about MDM misspecification and we would summarize the sensitivity problems of missing covariate in linear regression model into theoretical form. We utilize likelihood-based incomplete data bias measurement (Copas, 2013) to assess the effect of model sensitivity, and simplify the complex uncertainty quantities by some key sensitivity parameters. Then, local sensitivity analysis (with worst case) (Copas and Eguchi, 2005) and generative models (with small range) (Yin and Shi, 2019) are used to examine the severity of MDM sensitivity.

In this paper, we start our analysis from maximum likelihood estimation (MLE) under a working model, i.e., the model we usually used in practice; for example, a linear regression problem under MAR assumption for the data with partly missing covariates. The local sensitivity analysis approach is to investigate theoretically how sensitive the final conclusion or the bias is to the assumption by examining models in local area of the working model in a functional space (each model can be treated as a function in a functional space). And this sensitivity almost always exists under non-ignorable missingness. We further decompose the bias into two types: one is related to the distribution among covariates and the other is to the MDM assumption; both with non-identifiability issues. Next, we would estimate the plausible values of MLEs of parameters in worst case or in a small range, where worst case is derived by maximizing squared standardized bias. Yin and Shi (2019) proposed the generative modelling for MDM (GM-MDM), which attempted to investigate the possibility of each MDM model assumption and offered a plausible set of the sensitivity parameters. We applied their idea to the above problems and provide a non-implausible set which quantify a range of possible solutions even if the true MDM is MNAR. The similar idea has also been used in Andrianakis et al. (2017). We focus our discussion on the missing covariates problem in a linear regression model in this paper, but the method can be applied to other related problems.

The rest of the paper is organized as follows. Section 2 discusses the main problem of MDM misspecification with missing covariate problems and how to evaluate the incomplete data bias. The linear regression models with missing covariate under continuous and binary confounder are discussed separately. Both the local sensitivity analysis and GM-MDM are discussed in Section 3. Section 4 performs the simulation studies under different models of non-ignorable missingness and covariate distributions. In Section 5, we present the real data example analysis. Further discussion will be addressed in Section 6. Technical proofs, some detailed derivative processes, and some extra numerical results are presented in Supplementary Materials http://intlpress.com/site/pub/files/_supp/sii/2023/0016/0004/SII-2023-0016-0004-s002.pdf.

2. LOCAL SENSITIVITY ANALYSIS WITH MDM UNCERTAINTY

Given collected data, we specify a model $\{f(\cdot,\theta), \theta \in \Theta\}$ to inference parameter θ. However in practice, the true data generating distribution, denoted as $g(\cdot)$, is not always equal to the specified model or working model $f(\cdot)$. Copas and Eguchi (2005) discussed the model uncertainty issue and incomplete data bias analysis. For complete data Z and incomplete data Y, the parametric model $g_Z(z;\theta)$ and its marginal model $g_Y(y;\theta) = \int_{(y)} g_Z \mathrm{d}z$ specify the distribution of Z and Y respectively, where $\int_{(y)}$ means integration with respect to z over the missing set. In many cases, inference is based on a working model f_Z, while data Z is in fact generated by a nearby distribution g_Z. To formulate distribution in a local neighbourhood of f_Z, let $u_Z(z;\theta)$ be any scalar function of Z and parameter θ, standardized to have mean zero and variance one under the model f_Z. Then for small values of ϵ, the sampling model

$$g_Z = g_Z(z;\theta,\epsilon,u_Z) = f_Z(z;\theta)\exp\{\epsilon u_Z(z;\theta)\} \tag{2.1}$$

is non-negative and integrates to one, thus, identifies a distribution in the neighbourhood of f_Z. And the distribution of Y induced by g_Z is $g_Y = g_Y(y;\theta,\epsilon,u_Y) = \int_{(y)} f_Z(z;\theta)\exp\{\epsilon u_Z(z;\theta)\}\mathrm{d}z \approx f_Y(y;\theta)\exp\{\epsilon u_Y(y;\theta)\}$, where $u_Y(y;\theta) = \mathrm{E}_f\{u_Z(z;\theta)|Y = y\}$, which integrates the missing data out, and f_Y is the corresponding marginal model of f_Z for incomplete data: $f_Y = \int_{(y)} f_Z \mathrm{d}z$.

Under the identifiability condition (see Lin, Shi and Henderson, 2012), the incomplete data bias b_θ is the first-order approximation to the difference $\theta_{g_Y} - \theta_{g_Z}$, which is given by

$$\theta_{g_Y} - \theta_{g_Z} \approx b_\theta = \epsilon \mathrm{E}_f[u_Z(z;\theta)\{I_Y^{-1}s_Y(y;\theta) - I_Z^{-1}s_Z(z;\theta)\}], \tag{2.2}$$

where $\theta_{g_Y} = \arg_\theta[\mathrm{E}_{g_Y}\{s_Y(y;\theta)\} = 0]$, $\theta_{g_Z} = \arg_\theta[\mathrm{E}_{g_Z}\{s_Z(z;\theta)\} = 0]$, $s_Y(y;\theta) = \partial \log f_Y(y;\theta)/\partial\theta$ and $s_Z(z;\theta) = \partial \log f_Z(z;\theta)/\partial\theta$ are the score functions, $I_Y = \mathrm{E}(s_Y^2(y;\theta))$ and $I_Z = \mathrm{E}(s_Z^2(y;\theta))$ are the information matrices of f_Y and f_Z respectively, and $\arg_\theta[\cdot]$ returns the set of all possible values of θ from the brackets.

2.1 Missing covariate in regression models

With missing data problem, there are many uncertainty occurrences in model fitting when we consider the distributions of observed and unobserved covariates and MDM. Now, we specifically focus on the uncertainty analysis for linear regression models with missing covariate problem.

Assume there is an experiment design comprised of a response variable T, and the fully observed covariate X and the partially missing covariate C. For example, in genetic epidemiology field, the response variable T (also named as "phenotype") may be a continuous trait (e.g., blood pressure). Variable X may represent the environment variable or patient demographic such as patient's age, gender or drug therapy. Variable C may represent a confounding data, e.g. diet or smoking status, which may have a number of samples with data unobserved. With non-ignorable MDM (MNAR), the data generating model under complete data $Z = \{T, X, C, R\}$ is:

$$g_Z = f_{T|XC}(t|x, c; \theta) f_{XC}(x, c) f(r|t, x, c), \tag{2.3}$$

where R is a missing indicator, which sets to 0 if the corresponding C is missing and 1 otherwise, and $f(r|t, x, c)$ is the model for the true MDM. We usually do not know the true form of $f(r|t, x, c)$, so we may instead use:

$$f_Z = f_{T|XC}(t|x, c; \theta) f_{XC}(x, c) f_1(r|t, x, c), \tag{2.4}$$

where $f_1(r|t, x, c)$ is a working model, which is usually selected as a parametric model (e.g. logistic linear model). The difference between the two models indicates the problem of MDM sensitivity, i.e. $\exp\{\epsilon_R u_R\} = \frac{f(r|t,x,c)}{f_1(r|t,x,c)}$. If we use the notation in Equation (2.1), the misspecification part $\exp\{\epsilon u_Z\}$ is re-written as $\exp\{\epsilon_R u_R\}$ to present the uncertainty source caused by the different MDM specifically.

True MDM model $f(r|t, x, c)$ is usually unknown, thus this misspecification under MNAR is common. The fitted model $f_1(r|t, x, c)$ cannot be examined by goodness-of-fit test due to the partly missingness of c. Actually, it is often fitted by a MAR model in practice, i.e. $f_1(r|t, x)$ or $f_1(r|x)$. For simplification, we only consider $f(r|x, c)$ and $f_1(r|x)$ as the true and working MDM models in the following part. In this case, u_R directs non-ignorable missing data into ignorable missing data frame, and maps the complex uncertain non-ignorable missing data pattern into a much simpler and identifiable model, but with stronger assumptions and often the cost of incomplete data bias.

When we further consider the specification of the covariate density, the model fit of $f(c|x, r = 0)$ is very difficult, thus, we estimate the distribution of (X, C) based on the observed data or often simply assume X and C are independent:

$$f_Z^* = f_{T|XC}(t|x, c; \theta) f_X(x) f_C(c) f_1(r|x). \tag{2.5}$$

Similarly, we can use $\exp\{\epsilon_{XC} u_{XC}\} = \frac{f_{XC}(x,c)}{f_X(x) f_C(c)}$ as the misspecification of covariate distribution.

The corresponding working model f_Y^* for the incomplete data $Y = \{T, X, C^{(r)}, R\}$, where $C^{(r=1)}$ is observed and $C^{(r=0)}$ is missing, is given by

$$f_Y^* = \int_{(y)} f_Z^* \mathrm{d}z = f_{T|XC}^r(t|x, c^{(r)}; \theta) f_X(x) f_C^r(c) f_1(r|x), \tag{2.6}$$

with $f_{T|XC}^r(t|x, c^{(r)}; \theta) = \begin{cases} f_{T|XC}(t|x, c; \theta), & r = 1; \\ f_{T|X}(t|x; \theta), & r = 0 \end{cases}$ and $f_C^r(c) = \begin{cases} f_C(c), & r = 1; \\ 1, & r = 0. \end{cases}$ It is easy to notice that model f_Y^* is fully identifiable and parameters can be estimated by MLE method.

We are interested in estimating θ, and we would like to know the influence of the misspecified assumptions on the parameter estimation. The following theorem gives formula to calculate the double unavoidable misspecification quantities with the incomplete data.

Theorem 1. *The data generating distribution for complete data Z is noted as $g_Z = g_Z(z; \theta, \epsilon_R, \epsilon_{XC}, u_R, u_{XC}) = f_Z^*(z; \theta) \exp(\epsilon_{XC} u_{XC}) \exp(\epsilon_R u_R)$, where f_Z^* is the working model, and the limiting value of MLE is denoted θ_{g_Z}. Correspondingly, the sampling distribution under incomplete data Y is g_Y, which is the marginal model of g_Z: i.e. $g_Y = f_Y^*(y; \theta) \exp(\epsilon_R u_{R|Y}) \exp(\epsilon_{XC} u_{XC|Y})$, where $u_{XC|Y} = E_{f_Z^*}(u_{XC}(z; \theta)|Y)$ and $u_{R|Y} = E_{f_Z}(u_R(z; \theta)|Y)$. So, we use the model $f_Y^*(y; \theta)$ to fit the observations sampling from g_Y, and the limiting value of MLE under Y is denoted θ_{g_Y}. The incomplete data bias b_θ under some identifiability conditions is given by*

$$\begin{aligned} \theta_{g_Y} - \theta_{g_Z} \approx b_\theta \quad = \quad & \epsilon_R {I_Y^*}^{-1} E_{f_Z}(u_R s_Y^*) - \epsilon_R {I_Z^*}^{-1} E_{f_Z}(u_R s_Z^*) \\ & + \epsilon_{XC} {I_Y^*}^{-1} E_{f_Z^*}(u_{XC} s_Y^*) \\ & - \epsilon_{XC} {I_Z^*}^{-1} E_{f_Z^*}(u_{XC} s_Z^*), \end{aligned} \tag{2.7}$$

with s_Y^* and I_Y^* as score function and information matrix under model f_Y^*, while s_Z^* and I_Z^* as those under f_Z^*.

The proof of Theorem 1 is given in Section S1 in Supplementary Materials. In Theorem 1, g_Z is non-negative and integrates to one and including first-order terms in $(\epsilon_R u_R, \epsilon_{XC} u_{XC})$ and so identifies a distribution in the neighbourhood of f_Z^*. Two types of biases are decomposed and will be investigated separately. Specifically, the first two terms in the bias expression in (2.7) can be described as the MDM bias:

$$b_R = \epsilon_R {I_Y^*}^{-1} \mathrm{E}_{f_Z}(u_R s_Y^*) - \epsilon_R {I_Z^*}^{-1} \mathrm{E}_{f_Z}(u_R s_Z^*), \tag{2.8}$$

and the last two terms as the covariate bias:

$$b_{XC} = \epsilon_{XC} {I_Y^*}^{-1} \mathrm{E}_{f_Z^*}(u_{XC} s_Y^*) - \epsilon_{XC} {I_Z^*}^{-1} \mathrm{E}_{f_Z^*}(u_{XC} s_Z^*), \tag{2.9}$$

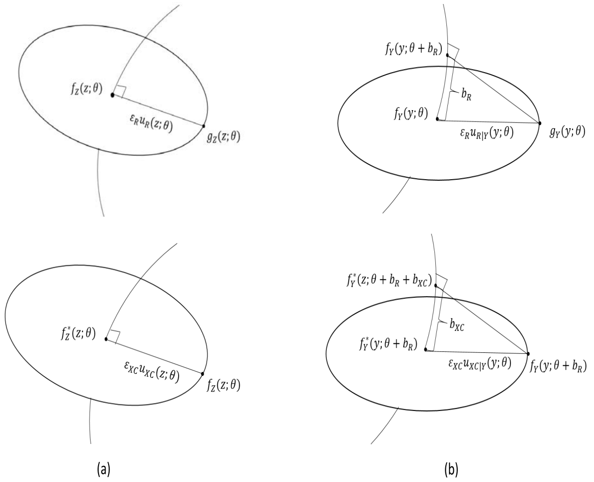

Figure 1. Distributions of Z and Y: (a) complete-data model – the true model g_Z in (2.3) is in a neighbourhood of the model f_Z in (2.4) with a misspecified MDM, and in addition f_Z is in a neighbourhood of the model f_Z^* in (2.5) with a further misspecified distribution between X and C; (b) incomplete observed data model – project the model for Z to the model for Y.

and $b_\theta = b_R + b_{XC}$. In a regression model with missing covariates, the MDM bias b_R is mainly caused by the nonidentifiability of MDM under MNAR and a misspecified MDM is used, and the covariate bias b_{XC} is mainly caused by the misspecified distribution between X and C. For the latter, a special case, as used in practice often, is to set a model under a (wrong) strong assumption for the correlation between the fully observed covariate X and the partially missing covariate C.

To geometrically visualize the relationship between the true models and the working models of the complete data Z and the incomplete data Y, we provide Figure 1, which also illustrates how the incomplete data bias arises. For any given θ, the misspecification quantities $\epsilon_{XC} u_{XC}(z;\theta)$ and $\epsilon_R u_R(z;\theta)$ are the vectors joining $f_Z^*(z;\theta)$ to $f_Z(z;\theta)$ and $f_Z(z;\theta)$ to $g_Z(z;\theta)$, respectively. These vectors have "length" ϵ_{XC} and ϵ_R, and "direction" given by the unit vectors $u_{XC}(z;\theta)$ and $u_R(z;\theta)$ respectively.

2.2 MDM uncertainty under a linear regression model

For complete data $Z = \{T, X, C, R\}$, the linear fixed effect model can be written as

$$t|(x,c) \sim \mathrm{N}(\theta_0 + \theta_X x + \theta_C c, \sigma^2), \tag{2.10}$$

where σ^2 is the variance of the error term. A normal distributed model is assumed here. And the model with the incomplete data $Y = \{T, X, C^{(r)}, R\}$ is $t|(x,c,r) \sim \mathrm{N}(\theta_0 + \theta_X x + r\theta_C c, \sigma^2 + (1-r)\theta_C^2\sigma_c^2)$. For complete case, the incomplete data model is the same with the complete data model; while for incomplete case, the model $t|(x, r = 0) \sim \mathrm{N}(\theta_0 + \theta_X x, \sigma^2 + \theta_C^2\sigma_c^2)$ is similar to the missing confounder problem (Copas and Eguchi, 2005). Here we use the maximum likelihood method to estimate the parameters $\theta = (\theta_0, \theta_X, \theta_C)^{\mathrm{T}}$ and the incomplete data bias b_θ.

For complete data Z and incomplete data Y, the log-likelihoods, the score functions, and the Fisher information matrices for the linear model are $l_Z^*(\theta; z), l_Y^*(\theta; y), s_Z^*(z;\theta), s_Y^*(y;\theta), I_Z^*$ and I_Y^* respectively, whose equations are listed in Section S2 in Supplementary Materials. To simplify notations, we define $v = (c, cx, \theta_C^2\sigma_c^2(c^2 - \sigma_c^2))^{\mathrm{T}}$. Let $f_{R|Z} = f(r|x,c)$ and $f_{R|Y} = f_1(r|x)$. Since $\mathrm{E}_{f_{T|XC}}(s_Z) = 0$ for all x and c, the MDM bias is

$$b_R \approx \frac{\theta_C {I_Y^*}^{-1}}{\sigma_Y^2} \mathrm{E}_{f_{XC}}\left\{(f_{R=1|Y} - f_{R=1|Z})v\right\}, \tag{2.11}$$

and the covariate bias is

$$b_{XC} \approx \frac{\theta_C {I_Y^*}^{-1}}{\sigma_Y^2} \{\mathrm{E}_{f_{XC}}(v f_{R=0|Y}) - \mathrm{E}_{f_X f_C}(v f_{R=0|Y})\}, \tag{2.12}$$

where $\mathrm{E}_{f_{XC}}(\cdot)$ indicates the expectation under the joint distribution $f_{XC}(x,c)$, while $\mathrm{E}_{f_X f_C}(\cdot)$ indicates the expectation under the independent distribution $f_X(x) f_C(c)$. Equation (2.11) contains the term $f_{R=1|Y} - f_{R=1|Z}$, which reflects the difference between the true MDM and the working MDM that we assume. The MDM bias b_R will disappear if the true MDM is known, i.e. $f_{R=1|Y} - f_{R=1|Z} = 0$. Equation (2.12) contains the term $\mathrm{E}_{f_{XC}}(v f_{R=0|Y}) - \mathrm{E}_{f_X f_C}(v f_{R=0|Y})$, which is influenced by the correlation of X and C. The covariate bias b_{XC} will be zero if the correlation between X and C does not exist, i.e. $\mathrm{E}_{f_{XC}}(v f_{R=0|Y}) = \mathrm{E}_{f_X f_C}(v f_{R=0|Y})$. The parameter θ_C is the effect of covariate C and σ_c^2 is the variance of that covariate, the term $\theta_C^2\sigma_c^2$ in $\sigma_Y^2 = \sigma^2 + \theta_C^2\sigma_c^2$ measures the error in the assumption that $\mathrm{corr}(x,c) = 0$ in the working model f_Z^*. Both θ_C and σ_Y^2 have influence in the MDM bias and the covariate bias.

Remark 1. In Section S2 in Supplementary Materials, we provided another form for b_R and b_{XC} which linked to some quantities have clear physical explanation. They may be used in practice to understand how the bias is related to those quantities if we have some prior knowledge. We also offer the detailed derivation procedure for the following biases under continuous and binary confounders.

For simplicity, we suppose X and C have the multivariate normal distribution $(X, C)^{\mathrm{T}} \sim \mathrm{N}(\mu_{XC}, \Sigma_{XC})$ with $\mu_{XC} = (\mu_x, \mu_c)^{\mathrm{T}}$ and correlation ρ, then the MDM bias in Equation (2.11) can be rewritten as

$$b_R \approx \frac{\theta_C {I_Y^*}^{-1}\{f_1(r=1) - f(r=1)\}}{\sigma_Y^2} \mathrm{E}(v|r=1), \tag{2.13}$$

where $f_1(r=1)$ and $f(r=1)$ are the observed probabilities for the working MDM and the true MDM respectively, and $\mathrm{E}(\cdot|r=1)$ is the expectation given $r=1$. This formula literally shows how the MDM bias correlated to the quantities for the working MDM and the true MDM. The key drivers of sensitivity relies on $f_1(r=1)-f(r=1)$, which is the difference between the working MDM and the true MDM. When the working MDM is equal to the true MDM, i.e. $f_1(r=1)=f(r=1)$, there is no uncertainty for the MDM, and no MDM bias. The covariate bias in Equation (2.12) is

(2.14)
$$b_{XC} \approx \frac{\theta_C {I_Y^*}^{-1}}{\sigma_Y^2} f_1(r=0)\rho\frac{\sigma_c}{\sigma_x}\begin{pmatrix} \mathrm{E}(x|r=0)-\mu_x \\ \mathrm{E}(x^2|r=0)-\mu_x\mathrm{E}(x|r=0) \\ \xi \end{pmatrix},$$

where $\xi = \theta_C^2\sigma_c^2[\rho\frac{\sigma_c}{\sigma_x}\{\mathrm{E}(x^2|r=0)-2\mu_x\mathrm{E}(x|r=0)+\mu_x^2-\sigma_x^2\}+2\mu_c\{\mathrm{E}(x|r=0)-\mu_x\}]$. The covariate bias correlates with the key sensitivity parameter from covariate distribution, which is the correlation coefficient $\mathrm{corr}(x,c)=\rho$ between X and C. Specially, if the covariate correlation $\rho=0$, the incomplete data bias will not exist, which measures the cost of misspecification of covariate correlation.

Suppose X is a binary variable, such as the treatment variable, i.e. $X \sim \mathrm{B}(1,p_x)$, and the conditional distribution of $C|X$ has mean $\mu_{c|x}$ and variance $\sigma_{c|x}^2$. For simplification, we suppose $\sigma_{c|x=0}^2=\sigma_{c|x=1}^2$ and denote them as σ_0^2. By calculation, we have that the mean and variance of C are $\mu_c = p_x\mu_{c|x=1}+(1-p_x)\mu_{c|x=0}$ and $\sigma_c^2=\sigma_0^2+p_x(1-p_x)(\mu_{c|x=1}-\mu_{c|x=0})^2$ respectively, and the covariance between X and C is $\mathrm{cov}(x,c)=p_x(1-p_x)(\mu_{c|x=1}-\mu_{c|x=0})$, which is determined by $\mu_{c|x=1}-\mu_{c|x=0}$, i.e. the difference of the mean values of C under the treatment group $(X=1)$ and the control group $(X=0)$. Under this case, the MDM bias in Equation (2.11) is

$$b_R \approx \frac{\theta_C {I_Y^*}^{-1}\{f_1(r=1)-f(r=1)\}}{\sigma_Y^2}\mathrm{E}(v|r=1), \tag{2.15}$$

and the covariate bias in Equation (2.12) is

$$\begin{aligned} b_{XC} &\approx \frac{\theta_C {I_Y^*}^{-1}\mathrm{cov}(x,c)}{\sigma_Y^2} \\ &\quad\times\begin{pmatrix} f_1(r=0|x=1)-f_1(r=0|x=0) \\ f_1(r=0|x=1) \\ \Xi_1 \end{pmatrix}, \end{aligned} \tag{2.16}$$

where $\Xi_1 = \theta_C^2\sigma_c^2\{f_1(r=0|x=1)-f_1(r=0|x=0)\}(\mu_{c|x=1}+\mu_{c|x=0})$. And the biases are influenced by $f_1(r=1)-f(r=1)$ and $\mathrm{cov}(x,c)$ respectively. Since the covariance is $\mathrm{cov}(x,c)=p_x(1-p_x)(\mu_{c|x=1}-\mu_{c|x=0})$, the term $\mu_{c|x=1}-\mu_{c|x=0}$ shows the correlation between X and C, and also implies the difference of the mean values of C under binary treatment variable X, thus, it influences the covariate bias b_{XC} significantly.

3. SENSITIVITY ANALYSIS

3.1 Local sensitivity analysis

The method of finding the most sensitive direction u_Z in Equation (2.1) by maximizing squared standardized bias (SSB) has been introduced in Copas and Eguchi (2005). Here, we also can derive the most sensitive directions u_R and u_{XC} based on Copas and Eguchi (2005)'s idea.

Denote $\theta=(\theta_0,\theta_X,\theta_C)^{\mathrm{T}}$, the scalar parameter $\phi=d^{\mathrm{T}}\theta$, the estimation of ϕ under the working model f_Y^* as $\hat{\phi}=d^{\mathrm{T}}\hat{\theta}_{g_Y}$, and the incomplete data bias $b_\phi=d^{\mathrm{T}}b_\theta$. The squared standardized bias can be written as

$$\begin{aligned} \mathrm{SSB} &= \frac{b_\phi^2}{n\mathrm{var}_{f_Y^*}(\hat{\phi})} = \frac{\{d^{\mathrm{T}}(b_R+b_{XC})\}^2}{d^{\mathrm{T}}{I_Y^*}^{-1}d} \\ &\leq 2\frac{(d^{\mathrm{T}}b_R)^2}{d^{\mathrm{T}}{I_Y^*}^{-1}d}+2\frac{(d^{\mathrm{T}}b_{XC})^2}{d^{\mathrm{T}}{I_Y^*}^{-1}d}. \end{aligned} \tag{3.1}$$

Next, we consider the two terms above respectively. By some technical details, we have $\frac{(d^{\mathrm{T}}b_R)^2}{d^{\mathrm{T}}{I_Y^*}^{-1}d} \leq \epsilon_R^2\{1-\lambda_{\min}(\Lambda^*)\}$, where $\lambda_{\min}(\Lambda^*)$ is the minimum eigenvalue of $\Lambda^* = {I_Y^*}^{1/2}{I_Z^*}^{-1}{I_Y^*}^{1/2}$. The equality holds when $u_R = \frac{d^{\mathrm{T}}({I_Y^*}^{-1}s_Y^*-{I_Z^*}^{-1}s_Z^*)}{\{d^{\mathrm{T}}({I_Y^*}^{-1}-{I_Z^*}^{-1})d\}^{1/2}}$ and $d={I_Y^*}^{1/2}\nu_{\min}(\Lambda^*)$, where $\nu_{\min}(\Lambda^*)$ is the eigenvector of Λ^* with the smallest eigenvalue. We put the expression of d into u_R and obtain that $u_R=\frac{\nu_{\min}^{\mathrm{T}}(\Lambda^*){I_Y^*}^{1/2}({I_Y^*}^{-1}s_Y^*-{I_Z^*}^{-1}s_Z^*)}{\{1-\lambda_{\min}(\Lambda^*)\}^{1/2}}$, which is the "worst case" direction for the MDM bias. Similarly, the worst or the most sensitive direction of u_{XC} is similar to u_R, because the term $({I_Y^*}^{-1}s_Y^*-{I_Z^*}^{-1}s_Z^*)$ is the same. Thus, the squared standardized bias is

$$\begin{aligned} \mathrm{SSB} &= \frac{b_\phi^2}{n\mathrm{var}_{f_Y^*}(\hat{\phi})} \\ &\leq 2\epsilon_R^2\{1-\lambda_{\min}(\Lambda^*)\}+2\epsilon_{XC}^2\{1-\lambda_{\min}(\Lambda^*)\}. \end{aligned} \tag{3.2}$$

Remark 2. See Section S3 in Supplementary Materials for some technical details. Although Equations (3.1) and (3.2) provide upper bound of the bias, it is difficult to use the general results in practice. Therefore, a very conservative method of double-the-variance was given in Copas and Eguchi (2005). We propose to use a generative modelling approach in the next section to quantify the range of the bias.

3.2 Generative modelling for MDM

The idea of generative modelling for MDM (GM-MDM) is first proposed by Yin and Shi (2019), which is used to do the sensitivity analysis with MNAR data. In contrast to the method discussed in Section 3.1 which focuses on the worst-case direction but is not often easy to find meaningful results, GM-MDM approach is to investigate the plausibility of the values for some inestimable parameters involved in the model for missing data. For example, if the data is

MNAR for the linear model discussed in Section 2.2, and the true MDM is $f(r=1|x,c) = \text{expit}(1 + x + \eta c)$, where $\text{expit}(a) = \exp(a)/\{1+\exp(a)\}$. The parameter η, namely sensitive parameter, is inestimable since part of c is missing, and the MDM depends on the missing part as well. The basic idea of GM-MDM is to investigate all potential candidate of η and remove those implausible values. This will lead to a so called non-implausible (NIP) set of η. As we will show later by using numerical examples, it can usually provide some meaningful and attractive results in practice. The GM-MDM is often implemented using Monte Carlo techniques.

A general procedure of GM-MDM is described as follows.

Step 1. The observed data set is $\mathcal{D}_{obs} = \{t_i, x_i, c_i^{(r_i=1)}, r_i\}_{i=1}^n$, where $\{c_i^{(r_i=0)}\}_i$ are missing. Under the assumption that the working MDM is MAR, i.e. $f_1(r=1|x) = \text{expit}(w_0 + w_1 x)$, we can obtain the coefficient estimator $\hat{\theta}_{gY} = \hat{\theta}_{MAR}$ using $\mathcal{D}_{obs}$.

Step 2. Impute the missing $\{c_i\}_{i,r_i=0}$ by

$$f(c|t_i,x_i,r_i=0) = f(c|t_i,x_i,r_i=1) \times \frac{\frac{f(r_i=0|t_i,x_i,c)}{f(r_i=1|t_i,x_i,c)}}{\mathrm{E}\{\frac{f(r_i=0|t_i,x_i,c)}{f(r_i=1|t_i,x_i,c)}|t_i,x_i,r_i=1\}}, \tag{3.3}$$

where the true MDM $f(r_i=1|t_i,x_i,c)$ depends on η, which is inestimable. So, we select a value of η from a proposal distribution or a candidate set, denoted by Γ. The imputed c_i is $\hat{c}_i = c_i I(r_i=1) + \tilde{c}_i I(r_i=0)$, where $\tilde{c}_i$ is the one generated from (3.3). Thus, we obtain the "complete" data $\{t_i, x_i, \hat{c}_i\}_{i=1}^n$, we therefore use it to calculate all the unknown parameters and denote it by $\hat{\theta}_\eta$.

Step 3. Re-generate t by $t_{i,\eta} = \hat{\theta}_{0,\eta} + \hat{\theta}_{X,\eta} x_i + \hat{\theta}_{C,\eta}\hat{c}_i + \epsilon_i$, where $\epsilon_i \sim \mathrm{N}(0, \hat{\sigma}^2)$, and then denote the dataset generated by the given η as $\mathcal{D}_\eta = \{t_{i,\eta}, x_i, \hat{c}_i, r_i\}_{i=1}^n$ and denote $\mathcal{D}_\eta^{obs}$ as the corresponding part of $\mathcal{D}_{obs}$.

Step 4. Compare the distance $s(\mathcal{D}_{obs}, \mathcal{D}_\eta^{obs})$ between the original observed data set $\mathcal{D}_{obs}$ and the corresponding generated data set $\mathcal{D}_\eta^{obs}$.

Step 5. Determine the non-implausible (NIP) set Γ_α for the sensitivity parameters η.

Remark 3. We know that $\mathcal{D}_{obs}$ is corresponding to the true model, while $\mathcal{D}_\eta^{obs}$ is corresponding to an assumed model controlled by η. If the distribution of $\mathcal{D}_\eta^{obs}$ is very different to the distribution of $\mathcal{D}_{obs}$, the underline assumed model is wrong. The closeness of the two distributions can be assessed by a distance measure, e.g., K-nearest-neighbour (KNN) distance or Kolmogorov-Smirnov (KS) distance defined by the two-sample Kolmogorov-Smirnov statistic. In the simulation study, we calculate the KS distance between $\{t_i\}_{i,r_i=1}$ and $\{t_{i,\eta}\}_{i,r_i=1}$ to express this closeness, i.e. $s_{KS} = s_{KS}(\{t_i\}_{i,r_i=1}, \{t_{i,\eta}\}_{i,r_i=1}) = \sqrt{\frac{n_1}{2}} \sup_t |\hat{F}(t) - \hat{F}_\eta(t)|$, where n_1 is the sample size for the observed data $\{t_i\}_{i,r_i=1}$, $\hat{F}(t)$ and $\hat{F}_\eta(t)$ are the empirical distribution functions of the observed data $\{t_i\}_{i,r_i=1}$ and the generated sample $\{t_{i,\eta}\}_{i,r_i=1}$ respectively, and sup is the supremum function. Define the KS set as $\Gamma_{KS,\alpha} = \{\eta : s_{KS} < c(\alpha)\}$, where the value of $c(\alpha)$ is generated by $c(\alpha) = \sqrt{-\log(\alpha/2)/2}$ (Knuth, 1998) based on the KS test. A non-parametric test based on any distance measures can be conducted to remove less plausible candidates. In *Step 5*, by using a permutation method to define an "achieved significance level" (ASL) with bootstrap in Yin and Shi (2019). We first sample $(\mathcal{D}_{boot,1}, \mathcal{D}_{boot,2})$ from the combined data set $\mathcal{D}^* = (\mathcal{D}_{obs}, \mathcal{D}_\eta^{obs})$ and calculate $s^* = s(\mathcal{D}_{boot,1}, \mathcal{D}_{boot,2})$, where $\mathcal{D}_\eta^{obs}$ is the generated dataset given η, $\mathcal{D}_{boot,1}$ is a sample randomly selected from $\mathcal{D}^*$, and so is $\mathcal{D}_{boot,2}$. We repeat this step with a sufficiently large number, and calculate the proportion of $\{s^* \geq s(\mathcal{D}_{obs}, \mathcal{D}_\eta^{obs})\}$. This is the empirical value of the following ASL value

$$\mathrm{ASL}_\eta = \mathrm{Pr}_{H_0}(s^* \geq s(\mathcal{D}_{obs}, \mathcal{D}_\eta^{obs})), \tag{3.4}$$

where H_0 is the null hypothesis stating that $\mathcal{D}_{obs}$ and $\mathcal{D}_\eta^{obs}$ come from the same distribution. For a given significance level, a non-implausible (NIP) set of sensitivity parameters is defined as $\Gamma_\alpha = \{\eta : \mathrm{ASL}_\eta > \alpha\}$. Thus Γ_α and the corresponding θ_η give a range of plausible solutions. All the models, which are not in the range, should be discarded. Note that the NIP set Γ_α is general and suit for any distance, while KS set $\Gamma_{KS,\alpha}$ is only suit for KS distance based on KS test. The detailed explanation of each step of the GM-MDM can be found in Section S4 in Supplementary Materials.

Theorem 2. *In the GM-MDM, if the true value of the sensitivity parameter η_{true} is contained in Γ, then it will be selected into the non-implausible (NIP) set Γ_α asymptotically given a significance level α.*

The proof of Theorem 2 is given in Section S5 in Supplementary Materials. Although we are almost sure the true model can be selected, we cannot guarantee the consistency. The size of NIP does not depend on the sample size only, it also depends on the degree of the ignorability of the missing data. The latter cannot be eliminated by big data, it is controlled by the randomness of the sample.

4. SIMULATION STUDIES

Now we perform simulation studies to further examine the sensitivity of non-ignorable missing data. In order to show the performance of the adjusted estimator with the MDM bias b_R and the covariate bias b_{XC}, we design two settings to report the results, one is to compare the coverage rate (CR) with the results under MAR; another is to show the local sensitivity by the "achieved significance level" (ASL) values and KS distances with a range of sensitivity parameters. Besides, more simulation studies to check whether the missing rate will impact the performance of our method have been considered, see the details in Section S6.3 in Supplementary Materials.

Table 1. The biases b_R, b_{XC}, estimation and 95% coverage rate (CR) for θ_X.

n	η	ρ	b_R	b_{XC}	$\theta_{X,MAR}$	$\theta_{X,gz}$	CR$_{MAR}$	CR
500	0.1	-0.3	0.001	-0.079	0.919	0.997	0.634	0.936
		0	0.001	-0.001	1.001	1.001	0.956	0.944
		0.3	0.001	0.080	1.086	1.005	0.604	0.936
	0.5	-0.3	0.005	-0.078	0.925	0.998	0.650	0.962
		0	0.006	0.000	1.007	1.000	0.954	0.958
		0.3	0.005	0.081	1.090	1.004	0.570	0.940
	1	-0.3	0.011	-0.078	0.922	0.990	0.658	0.904
		0	0.011	0.000	1.011	1.000	0.948	0.950
		0.3	0.009	0.082	1.096	1.005	0.524	0.928
2000	0.1	-0.3	0.001	-0.080	0.919	0.998	0.114	0.940
		0	0.001	0.001	1.001	0.999	0.962	0.956
		0.3	0.001	0.079	1.083	1.003	0.100	0.934
	0.5	-0.3	0.005	-0.079	0.922	0.996	0.142	0.928
		0	0.006	0.000	1.006	1.000	0.936	0.944
		0.3	0.005	0.080	1.088	1.003	0.070	0.936
	1	-0.3	0.011	-0.078	0.927	0.994	0.180	0.920
		0	0.012	0.001	1.013	1.000	0.926	0.950
		0.3	0.009	0.082	1.094	1.003	0.034	0.934

4.1 Simulation studies for coverage rate

We consider a linear regression model, which is the same as that in Equation (2.10), i.e. $t|(x,c) \sim \mathrm{N}(\theta_0 + \theta_X x + \theta_C c, \sigma^2)$, where the true value of parameters $\theta = (\theta_0, \theta_X, \theta_C)^{\mathrm{T}}$ is chosen as $(1,1,1)^{\mathrm{T}}$ and $\sigma^2 = 1$. The covariates $(X, C)^{\mathrm{T}}$ follow a multivariate normal distribution with mean vector $(0,0)^{\mathrm{T}}$ and covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, where the correlation coefficient $\rho = -0.3, 0$ and 0.3. We denote this case as the normal case, besides, we also consider the binary case when $X \sim \mathrm{B}(1, p_x)$, see Section S6 in Supplementary Materials for the results of additional simulations. We drop some of the observations of C with $f(r=1|x,c) = \mathrm{expit}\{(1+x+\eta c)/2\}$, where $\eta = 0.1$, 0.5 and 1. The working MDM is MAR with $f_1(r=1|x,c) = \mathrm{expit}(w_0 + w_1 x)$, where the coefficients θ and w_0, w_1 can be estimated under the working MAR with observed data. We simulate complete observations with sample size $n = \{500, 2000\}$ and replicate 500 times. The missing $\{c_i\}_{i,r_i=0}$ have been imputed by 20 times for the stability of results.

The MDM bias (b_R), the covariate bias (b_{XC}), the estimators of θ (θ_{MAR}) under the working MAR, the adjusted estimators ($\theta_{gz} = \theta_{MAR} - b_R - b_{XC}$ from Equation (2.7)), CR of θ_{MAR} under the working MAR (CR$_{MAR}$), and the adjusted CR of θ_{gz} (CR) with 95% confidence interval are listed in Table 1 for the parameter of interest θ_X.

In general, we can see that the adjusted $\theta_{X,gz}$ is much closer to 1 than $\theta_{X,MAR}$, where 1 is the true value of θ_X. The coverage rate CR is larger than CR$_{MAR}$ in most of the cases. The MDM bias b_R increases as η grows, which is caused by the departure between the working MDM and the true MDM. Besides, the CR$_{MAR}$ is so small when $|\rho|$ is large, e.g. $\rho = \pm 0.3$, while there is a significant growth for adjusted CR in this case. When $\rho = 0$, which means there is no correlation between X and C, the covariate bias b_{XC} is so much close to zero.

4.2 Simulation studies for ASL and KS

The settings are similar as before, except that we only consider the correlation coefficient $\rho = 0.3$ as examples. The sensitivity parameter $\eta = 1, 3$ and 5 in MDM, the sample size $n = 2000$, and the time of repetition is 100. We set the $\eta \in [-20, 20]$, which is spaced by 0.5 as the interval, thus, we have 81 candidate sensitive parameters, and the biases are calculated based on each candidate η. The missing $\{c_i\}_{i,r_i=0}$ have been imputed by 20 times for the stability of results. The bootstrap process is used to calculate the KS distances between the observed data $\{t_i\}_{i,r_i=1}$ and the generated data $\{t_{i,\eta}\}_{i,r_i=1}$ under each sensitive parameter η, thus the ASL values can be estimated and then the non-implausible (NIP) set of η can be obtained, where the bootstrap time is 100, and the given significance levels $\alpha = 5\%$ and 10%. We use $c(\alpha = 5\%) = \sqrt{-\log(\alpha/2)/2} \approx 1.358$ and $c(\alpha = 10\%) = \sqrt{-\log(\alpha/2)/2} \approx 1.224$ as the critical value for the KS distances to define the KS set, i.e. $\{\eta : s_{KS} < c(\alpha)\}$. The procedures of the GM-MDM method can refer to Section 3.2.

We calculate the estimation θ_{MAR} and 95% confidence interval (CI$_{MAR}$) under MAR, which are marked with dotted lines in Figure 2. The adjusted estimation θ_{gz} and the corresponding 95% confidence interval (CI$_{gz}$) are marked with solid lines. And the dashed vertical line is for the true value of η, while the dashed horizontal line is for the true value of θ_X.

Figure 2 shows the estimation of θ_X and the corresponding 95% CI under the MAR and the MNAR given different

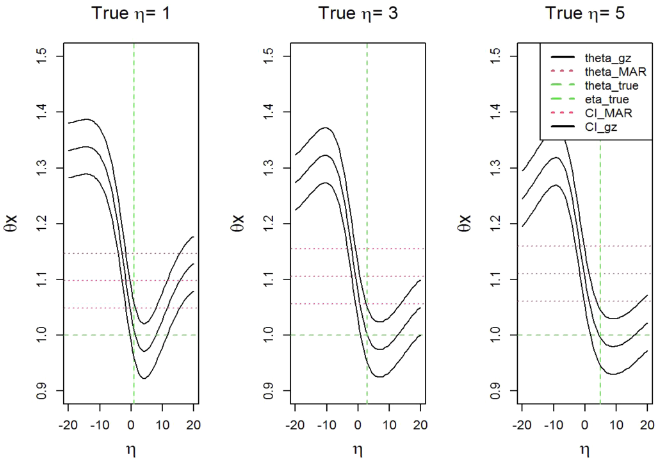

Figure 2. The 95% confidence interval (CI) of the estimator for θ_X under MAR and GM-MDM with true $\eta = 1, 3$ and 5, where the missing percentages are 0.392, 0.418 and 0.438 respectively.

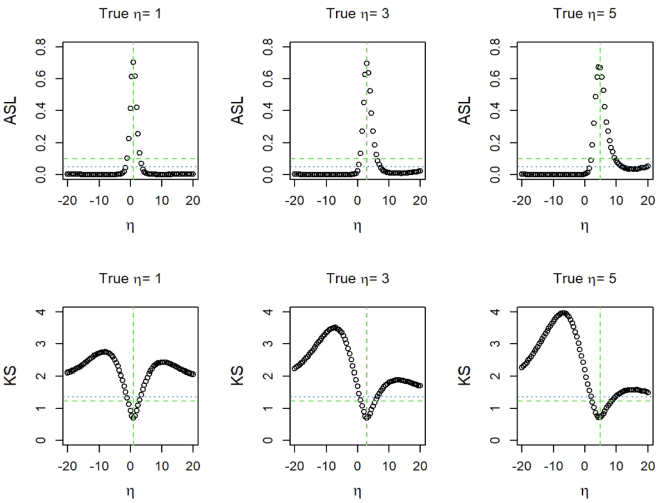

Figure 3. The ASL values and KS distances with true $\eta = 1, 3$ and 5.

η, where the true $\eta = 1, 3$ and 5 respectively. We can see that the CI_{gz} contains the true θ_X, which is equal to 1, under the true η in any case, however, the true θ_X is not included in CI_{MAR}, which is over estimated under MAR. The deficient performance of the estimation under MAR is caused by the fact that MAR is not the true MDM of the simulated data. As η is far away from its true value, the incomplete data bias is large, which results in that CI_{gz} cannot contain the true θ_X when η is far away from its true value.

Figures 3 reports the ASL and KS distances with $\eta = 1, 3$ and 5 respectively. The dots are the values of ASL and KS under each given η, the dashed vertical line is for the true value of η, the dotted horizontal line is $\alpha = 5\%$ for ASL and $c(5\%) = 1.358$ for KS value, while the dashed horizontal line is $\alpha = 10\%$ for ASL and $c(10\%) = 1.224$ for KS value.

The NIP sets $\Gamma_{\alpha=5\%}$ of sensitivity parameters η with significance level $\alpha = 5\%$ are $[-1.0, 3.5]$, $[0.5, 7.0]$ and $[2.0, 11.5]$, while $\Gamma_{\alpha=10\%}$ are $[-1.0, 3.0]$, $[1.0, 6.0]$ and $[2.5, 9.5]$ for $\eta = 1, 3$ and 5 respectively, which are selected when the ASL values are larger than the critical values 5% and 10%. The KS sets $\Gamma_{KS,\alpha=5\%}$ of sensitivity parameters η are $[-1.0, 3.0]$, $[1.0, 6.0]$ and $[2.5, 9.5]$, while $\Gamma_{KS,\alpha=10\%}$ are $[-0.5, 2.5]$, $[1.5, 5.5]$ and $[2.5, 8.5]$ for $\eta = 1, 3$ and 5 respectively, which are selected when the KS values are less than the critical value $c(\alpha = 5\%) = 1.358$ and $c(\alpha = 10\%) = 1.224$. We can see that the true value η is contained in each NIP set and KS set, and the sets with significance level $\alpha = 5\%$ are slightly wider than these with $\alpha = 10\%$. Although the GM-MDM method does not offer a point estimation of the sensitivity parameter, a convincing interval of the sensitivity parameter can be obtained. When the sensitivity parameter is away from its true value, the distance becomes larger, which is as we expect.

5. REAL DATA EXAMPLE

To validate the performance of the proposed method, we consider a real data example, which is to predict the acute toxicity of diverse chemicals based on two molecular descriptors, towards the fathead minnow (Pimephales promelas). The manufacturers can use the prediction results to prove that their products are safe for human health and the environment. The fish toxicity data set comes from the UC Irvine Machine Learning Repository (Dua and Graff, 2017), which contains the information about the toxicity towards fish and molecular descriptors of 908 chemicals. The toxicity of diverse chemicals towards the fathead minnow is defined as LC_{50} 96 hours, which is the concentrations causing death in 50% of test fathead minnows over a test duration of 96 hours. The two molecular descriptors are SM1_Dz and MLOGP, where SM1_Dz is the descriptor calculated from 2D matrices derived from the molecular graph, and MLOGP is the octanol-water partitioning coefficient, which is considered the driving force of narcosis. We are interested in the dependence of LC_{50} (T) on SM1_Dz (X) and MLOGP (C). Since the dataset is fully observed, we use the artificial missing method to miss some values of MLOGP, where the supposed true MDM is MNAR with $f(r = 1|x, c) = \mathrm{expit}\{(1+x+\eta c)/2\}$ and $\eta = 1, 3$ and 5. The missing percentages are 0.394, 0.421 and 0.439 respectively.

We fitted a linear regression model (using ordinary least squares) for LC_{50} (T) with SM1_Dz (X) and MLOGP (C) as covariates. First, we standardize the variables to reduce the errors caused by dimensional difference, self-variation, or large numerical difference, which can reveal statistically significant findings that we might otherwise miss. Then, we use the supposed true MDM to delete some values of MLOGP (C). Thus, we can obtain the missing dataset, and the following analysis is based on this dateset. The GM-MDM process in Section 3.2 is implemented, where the range of the

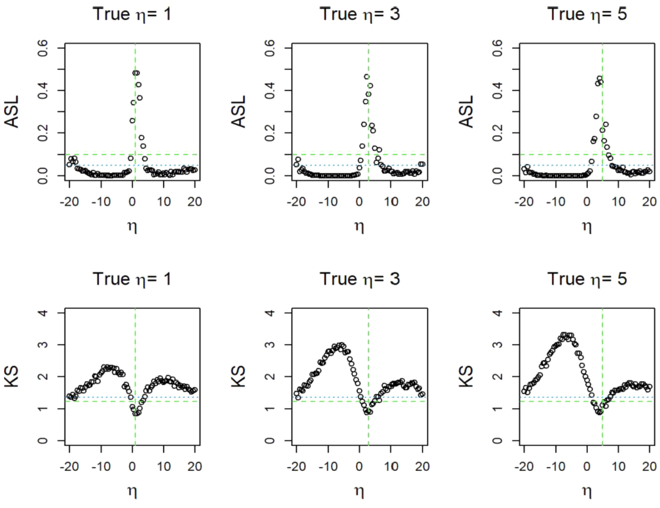

Figure 4. The ASL values and KS distances with true $\eta = 1, 3$ and 5 in fish toxicity dataset.

sensitivity parameter η is $[-20, 20]$ spaced by 0.5. The sample size of this dataset is 908, and the missing $\{c_i\}_{i,r_i=0}$ have been imputed by 20 times for the stability of results. The bootstrap process is used to calculate the KS distances between the observed data $\{t_i\}_{i,r_i=1}$ and the generated data $\{t_{i,\eta}\}_{i,r_i=1}$ under each sensitive parameter η, thus the ASL values can be estimated and then the non-implausible (NIP) set of η can be obtained, where the bootstrap time is 100, and the given significance level $\alpha = 5\%$ and 10%. The procedures of the GM-MDM can refer to Section 3.2.

To compare with our proposed GM-MDM method, we consider other three methods which are commonly used to deal with missing data, i.e. complete case analysis (CCA), multiple imputation (MI) (Rubin, 1987), and the estimation with MAR $f(r = 1|x) = \text{expit}(w_0 + w_1 x)$ (MAR). The CCA method assumes that the missingness in C is independent of T, X and C, thus, coefficient estimation is obtained based on the samples, where the missing data are deleted during the analysis process. The estimations based on MI method are calculated with the mice package in R (Buuren and Groothuis-Oudshoorn, 2011). We set meth="norm" in 'mice' function, which calculates imputations for univariate missing data by Bayesian linear regression, and calculate the estimations after the imputation by 'with' function. And estimations of the coefficient θ with MAR method are implemented by assuming the true MDM is MAR with $f(r = 1|x) = \text{expit}(w_0 + w_1 x)$.

The ASL values and KS distances under true MDM with $\eta = 1, 3$ and 5 are showed in Figure 4. The panels on the upside are the ASL values, where the NIP sets $\Gamma_{\alpha=5\%}$ of sensitivity parameters η with significance level $\alpha = 5\%$ are $[-0.5, 4.0]$, $[0.5, 7.0]$ and $[1.5, 7.5]$, while $\Gamma_{\alpha=10\%}$ are $[0, 3.5]$, $[1.0, 6.5]$ and $[2.0, 6.5]$ for $\eta = 1, 3$ and 5 respectively. The panel on the downside are the KS distances, where the KS sets $\Gamma_{KS,\alpha=5\%}$ of sensitivity parameters η are [0, 3.5], [1.0, 6.5] and [2.0, 7.0], while $\Gamma_{KS,\alpha=10\%}$ are $[0.0, 3.0]$, $[1.5, 4.5]$ and $[2.0, 6.0]$ for $\eta = 1, 3$ and 5 respectively. We can see that all the NIP sets and KS sets cover the true value η.

The coefficient estimations by the three compared methods and the NIP sets and KS sets of the coefficients are reported in Table 2, where the values in round brackets are the standard errors from Complete case analysis (CCA) and Multiple Imputation (MI), while the ranges in square brackets are the NIP sets and KS sets of the coefficients from GM-MDM method. The standard error of CCA is much larger than that of the MI in each setting, which means the CCA method is not as much stable as the MI method. The coefficients of the CCA, MI and MAR methods are not much similar, which means if we use different methods to analysis, the different results may obtain. The advantage of the GM-MDM method is that it offers a reasonable interval estimation instead of a point estimation of the coefficients, which can avoid the overestimation or underestimation with the inappropriate method.

6. DISCUSSION

The sensitivity of the MDM is common and difficult since lack of randomization or lack of identifiability (Copas and Eguchi, 2005). Without considering those effects could lead to biases (Vach and Blettner, 1991; Greenland and Finkle, 1995). In this paper, we mainly focus on investigating the sensitivity analysis of non-ignorable missing covariate in the linear regression model. We use the local bias analysis method to measure the uncertainty and extent the general model uncertainty problem to the non-ignorable missing covariate problem, where two kinds of uncertainties have been considered, one is the MDM uncertainty, the other is the correlation between observed covariate and missing covariate. The sensitivity measurement, or so-called "incomplete data bias" is calculated and interpreted by meaningful and interpretable quantities, such as the covariate variable correlation, the conditional mean difference of the missing covariate.

To identify the key sensitivity parameter in the incomplete data bias is difficult but crucial. This could illustrate the most crucial factors that drive the non-ignorable sensitivity. As for the continuous covariates case, we used the approximate likelihood-based bias evaluation, which identifies covariate correlation $\text{corr}(x, c)$ as the key sensitivity drive of covariate misspecification; the difference of the observation probability between the working MDM and the true MDM $f_1(r = 1) - f(r = 1)$ as the key drive of MDM uncertainty measure. As for the binary covariates case, we find that the conditional mean difference of $C|X$, i.e. $\text{E}(c|x = 1) - \text{E}(c|x = 0)$ influences the covariate misspecification bias, where $\text{E}(c|x = 1) - \text{E}(c|x = 0)$ is proportional to the covariance $\text{cov}(x, c)$; while the difference

Table 2. Estimations of model parameters in fish toxicity dataset. (The missing percentages are 0.394, 0.421 and 0.439 respectively.)

η	Estimator	Intercept	SM1_Dz (X)	MLOGP (C)
1	CCA	-0.011 (0.03)	0.277 (0.028)	0.597 (0.032)
	MI	-0.077 (0.013)	0.286 (0.008)	0.625 (0.015)
	MAR	-0.091	0.264	0.807
	GM-MDM ($\Gamma_{\alpha=5\%}$)	[-0.108, 0.127]	[0.252, 0.295]	[0.508, 0.619]
	GM-MDM ($\Gamma_{KS,\alpha=5\%}$)	[-0.076, 0.107]	[0.254, 0.285]	[0.527, 0.619]
	GM-MDM ($\Gamma_{\alpha=10\%}$)	[-0.076, 0.107]	[0.254, 0.285]	[0.527, 0.619]
	GM-MDM ($\Gamma_{KS,\alpha=10\%}$)	[-0.076, 0.087]	[0.254, 0.285]	[0.546, 0.619]
3	CCA	0 (0.032)	0.249 (0.026)	0.612(0.034)
	MI	-0.193 (0.016)	0.287 (0.011)	0.694(0.02)
	MAR	-0.184	0.248	0.888
	GM-MDM ($\Gamma_{\alpha=5\%}$)	[-0.165, 0.158]	[0.236, 0.282]	[0.490, 0.681]
	GM-MDM ($\Gamma_{KS,\alpha=5\%}$)	[-0.129, 0.140]	[0.236, 0.275]	[0.504, 0.677]
	GM-MDM ($\Gamma_{\alpha=10\%}$)	[-0.129, 0.140]	[0.236, 0.275]	[0.504, 0.677]
	GM-MDM ($\Gamma_{KS,\alpha=10\%}$)	[-0.106, 0.067]	[0.242, 0.263]	[0.581, 0.677]
5	CCA	-0.012 (0.035)	0.233 (0.026)	0.634 (0.037)
	MI	-0.274 (0.016)	0.285 (0.01)	0.749 (0.014)
	MAR	-0.241	0.24	0.938
	GM-MDM ($\Gamma_{\alpha=5\%}$)	[-0.174, 0.121]	[0.251, 0.269]	[0.528, 0.723]
	GM-MDM ($\Gamma_{KS,\alpha=5\%}$)	[-0.145, 0.099]	[0.251, 0.268]	[0.544, 0.703]
	GM-MDM ($\Gamma_{\alpha=10\%}$)	[-0.145, 0.086]	[0.251, 0.268]	[0.557, 0.703]
	GM-MDM ($\Gamma_{KS,\alpha=10\%}$)	[-0.145, 0.065]	[0.251, 0.268]	[0.575, 0.703]

$f_1(r=1) - f(r=1)$ has effect on the MDM bias. We further perform simulation studies and real data example to support the robustness of our theoretic derivation.

Several crucial points need to be addressed. One important point is that the MDM misspecification can be a serious problem, depending on how we "cut the cake". Several parameters drive those sensitivity, include the correlation between observed covariates and missing covariates, the distribution parameters of the missing covariates. Conventional sensitivity analysis can be adopted to investigate which of them are more sensitive than the other, as shown in our discussion. In practice, we may measure the worst case of incomplete data bias and adjust the parameter estimation. More recommended approach comes from GM-MDM, which examines the plausibility of each choice of sensitivity parameter by evidence. The other key point is that MDM bias cannot be avoided by simply increasing sample size. We should always give caution concerns about the MDM assumptions and specifications. Regarding the identifiability issue, the specification of non-ignorable MDM is impossible without further help, and the pattern of missing data is usually not reflected from observed data only. MNAR model is plausible in many places and needs to be examined.

There are some limitations for our method. First of all, we only derive the approach under the selection model framework, and do not consider other frameworks, such as shared parameter framework and pattern mixture framework, due to the different factorisation of these models. Besides, we consider the misspecification of the missing covariate distribution and the missing data mechanism, which can be extended to the misspecification of the response model. However, the problem of identifiability under the misspecification of the response model needs more attention. Furthermore, our method is based on the linear regression model. Study on our method under the more generalised models is interesting, although they are more complicated with high-dimensional covariates and non-linear case. Future work can be addressed for these concerns.

SUPPLEMENTARY MATERIALS

The proof of Theorem 1, the detailed proving processes of some equations in Section 2.2 and Section 3.1, the explanation of each step of the GM-MDM in Section 3.2, the proof of Theorem 2, additional simulation results for the binary case and the effect in the performance of our method with different missing percentage are available in Sections S1 – S6 in Supplementary Materials.

ACKNOWLEDGEMENTS

Rong Zhu and Jian Qing Shi (corresponding author) acknowledge financial support from the Medical Research Council [grant number: MR/M025152/2]; Peng Yin acknowledges financial support from the National Natural Science Foundation of China (NNSFC) [grant number: 11801542]. The authors thank the associate editor and referees for insightful comments that have substantially improved the quality of the paper. The usual disclaimer applies.

Received 24 August 2021

REFERENCES

Andrianakis, I., Vernon, I., McCreesh, N., McKinley, T., Oakley, J., Nsubuga, R., Goldstein, M. and White, R. (2017). History matching of a complex epidemiological model of human immunodeficiency virus transmission by using variance emulation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **66** 717-740. MR3670414

Buuren, S. V. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* **45** 1–67.

Copas, J. B. (2013). A likelihood-based sensitivity analysis for publication bias in meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62** 47–66. MR3042310

Copas, J. and Eguchi, S. (2005). Local model uncertainty and incomplete-data bias (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 459–513. MR2168201

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **43** 49–73.

Dua, D. and Graff, C. (2017). UCI Machine Learning Repository: QSAR fish toxicity Data Set. http://archive.ics.uci.edu/ml/datasets/QSAR+fish+toxicity#.

Gao, W., Hedeker, D., Mermelstein, R. and Xie, H. (2016). A scalable approach to measuring the impact of nonignorable nonresponse with an EMA application. *Statistics in medicine* **35** 5579–5602. MR3580927

Greenland, S. and Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology* **142** 1255–1264.

Guo, F., Ma, W. and Wang, L. (2020). Semiparametric estimation of copula models with nonignorable missing data. *Journal of Nonparametric Statistics* **32** 109-130. MR4077755

Ibrahim, J. G. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *Test* **18** 1–43. MR2495958

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R. and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association* **100** 332–346. MR2166072

Kim, J. K. and Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association* **106** 157–165. MR2816710

Knuth, D. E. (1998). *Art of Computer Programming, Volume 2: Seminumerical Algorithms*. Addison-Wesley Professional. MR3077153

Lin, N. X., Shi, J. Q. and Henderson, R. (2012). Doubly misspecified models. *Biometrika* **99** 285–298. MR2931254

Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88** 125–134.

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

Maity, A. K., Pradhan, V. and Das, U. (2019). Bias reduction in logistic regression with missing responses when the missing data mechanism is nonignorable. *The American Statistician* **73** 340–349. MR4027874

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89** 846–866. MR1294730

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

Troxel, A. B., Ma, G. and Heitjan, D. F. (2004). An index of local sensitivity to nonignorability. *Statistica Sinica* **14** 1221–1237. MR2126350

Vach, W. and Blettner, M. (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American Journal of Epidemiology* **134** 895–907.

Wang, S., Shao, J. and Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica* **24** 1097–1116. MR3241279

Yin, P. and Shi, J. Q. (2019). Simulation-based sensitivity analysis for non-ignorably missing data. *Statistical Methods in Medical Research* **28** 289–308. MR3894529

Yuan, C., Hedeker, D., Mermelstein, R. and Xie, H. (2020). A tractable method to account for high-dimensional nonignorable missing data in intensive longitudinal data. *Statistics in Medicine* **39** 2589–2605. MR4133127

Zhang, T. and Wang, L. (2020). Smoothed empirical likelihood inference and variable selection for quantile regression with nonignorable missing response. *Computational Statistics & Data Analysis* **144** https://doi.org/10.1016/j.csda.2019.106888. MR4038215

Zhao, J. and Shao, J. (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association* **110** 1577–1590. MR3449056

Rong Zhu
Medical Research Council Biostatistics Unit
School of Clinical Medicine
University of Cambridge
Cambridge
U.K.
E-mail address: zhu_rong@amss.ac.cn

Peng Yin
Shenzhen Institutes of Advanced Technology
Chinese Academy of Sciences, Shenzhen
China
E-mail address: peng.yin@siat.ac.cn

Jian Qing Shi
Department of Statistics and Data Science
Southern University of Science and Technology
Shenzhen
National Center for Applied Mathematics Shenzhen
China
E-mail address: shijq@sustech.edu.cn