Hindawi
Journal of Probability and Statistics
Volume 2019, Article ID 1019303, 10 pages
https://doi.org/10.1155/2019/1019303

Research Article

An Alternative Sensitivity Approach for Longitudinal Analysis with Dropout

Amal Almohisen,[1] **Robin Henderson**,[2] **and Arwa M. Alshingiti**[1]

[1]*Department of Statistics and Operations Research, College of Science, King Saud University, Riyadh, Saudi Arabia*
[2]*School of Mathematics and Statistics, Newcastle University, Newcastle Upon Tyne, UK*

Correspondence should be addressed to Amal Almohisen; amalmoh@ksu.edu.sa

Received 22 March 2019; Accepted 4 June 2019; Published 1 July 2019

Academic Editor: Hyungjun Cho

In any longitudinal study, a dropout before the final timepoint can rarely be avoided. The chosen dropout model is commonly one of these types: Missing Completely at Random (MCAR), Missing at Random (MAR), Missing Not at Random (MNAR), and Shared Parameter (SP). In this paper we estimate the parameters of the longitudinal model for simulated data and real data using the Linear Mixed Effect (LME) method. We investigate the consequences of misspecifying the missingness mechanism by deriving the so-called least false values. These are the values the parameter estimates converge to, when the assumptions may be wrong. The knowledge of the least false values allows us to conduct a sensitivity analysis, which is illustrated. This method provides an alternative to a local misspecification sensitivity procedure, which has been developed for likelihood-based analysis. We compare the results obtained by the method proposed with the results found by using the local misspecification method. We apply the local misspecification and least false methods to estimate the bias and sensitivity of parameter estimates for a clinical trial example.

1. Introduction

Missing data are common in various settings, including surveys, clinical trials, and longitudinal studies. Methods for handling missing data strongly depend on the mechanism that generated the missing values as well as the distributional and modeling assumptions at various stages. This study focuses only on Missing at Random and Missing Not at Random dropout models, under a Linear Mixed Effect (LME) model.

Much of the literature on missing data problems assumes the dropout model is only MAR and not MNAR, but this assumption is clearly limited [1]. The consequences of misspecifying the missingness mechanism are investigated by deriving the so-called least false values, which are the values the parameter estimates converge to when the assumptions may be wrong. Derivation and illustration of theoretical least false values for the LME method are made under Missing at Random (MAR) and Missing Not at Random (MNAR) dropout. The misspecified dropout model MAR is assumed in this study.

Copas and Eguchi [2] gave a formula to estimate the bias under such misspecification using a likelihood approach. As the LME is a likelihood-based method, the estimates obtained through the Copas and Eguchi method can be compared with the LME least false estimates. The procedure will be applied by adding a tilt to the MAR dropout model to provide what Copas and Eguchi call local misspecification.

The local model uncertainty is elaborated as proposed by Copas and Eguchi [2] and illustrated both when model misspecification is present and when the data is incomplete. Furthermore, we find that the Copas and Eguchi method gives very similar results to the least false method. Misspecification will be dealt with assuming MAR where actually the truth is MNAR. Beside Copas and Eguchi [2], many other authors have developed methods to assess the sensitivity of inference under the MAR assumption [3, 4]. Moreover, Lin et al. [5] extended the Copas and Eguchi method and assumed a doubly misspecified model while having only single misspecification. Also, there has been interest in the Copas and Eguchi method from a Bayes perspective [6–10]. Recently, [11] performed simulation based sensitivity analysis.

In Section 2, the LME method is presented and we show how to calculate the least false values. A description of the Copas and Eguchi method is provided in Section 3.1, followed by an example in Section 3.2. A simulation study is described in Section 4. The Copas and Eguchi bias estimate results are studied and examined with the least false values derived from the LME method, and we then show the coverage of nominal confidence intervals. A sensitivity analysis is conducted to assess how inference can depend on missing data. In Section 5, the methods are applied to data from a clinical trial with two treatments and two measurement times as introduced and analysed by Matthews et al. [12]. We compared the results obtained by the proposed method with the results found by using the Copas and Eguchi method.

2. Linear Mixed Effect (LME) Method

A statistical model containing fixed effects and random effects is called a mixed effect model. These models have been shown to be effective in many disciplines in the biological, physical, and social sciences. Usually a linear form is assumed.

Reference [13] gave a definition of the response Y in the LME model which is of the form:

$$\begin{pmatrix} measured \\ response \end{pmatrix} = \begin{pmatrix} covariate \\ effects \end{pmatrix} + \begin{pmatrix} random \\ effects \end{pmatrix} + \begin{pmatrix} measurement \\ error \end{pmatrix}. \tag{1}$$

For example, a simplified version of the Liard and Ware [14] mixed model approach for longitudinal data would include a random effect in the intercept term in a model for responses. If Y_{ij} is the response at time j on subject i, the model is

$$Y_{ij} = \mu_{ij} + U_i + \epsilon_{ij} \tag{2}$$

where μ_{ij} is the marginal mean, which will usually be a linear function of covariates, ϵ_{ij} is independent Gaussian noise, and U_i is a realisation of a zero mean scalar Gaussian random variable. Since U_i has zero mean, the marginal mean of Y_{ij} remains μ_{ij} after integrating out U_{ij}. However, since U_i is common to all j, we get dependence between observations on the same subject. For example, if U_i is positive, then all values would tend to be above the marginal mean and so on. In the context of longitudinal data, some reviews of linear mixed models can be found in [15, 16].

2.1. Assumptions. Suppose there are n individuals in a study and each provides longitudinal responses Y and dropout information R. Generally, we will assume a linear model for Y (in the absence of dropout) and logistic models for the probability of *continuing* to the next timepoint $t + 1$ given that a subject is still under observation at time t. At times, we refer to a *true* or *generating* model as the way in which data are obtained and to an *assumed* or *fitting* model as that chosen by the analyst for estimation.

For simplicity in this work, the study assumes that there are just two observations or treatment periods. The methods are of course more general.

At time 1, there is a measurement provided for all subjects, denoted by Y_{i1} for subject i. Then at time 2, some subjects are dropped out before measurement. Let R_i=1 indicate that there is a measurement at time 2 and R_i=0 otherwise. Let Y_i=$(Y_{i1}, Y_{i2})^T$ and assume $E[Y_i]$=$x_i\beta$ where β is a parameter vector of dimension p and x_i is the design matrix associated with subject i, which is of dimension $2 \times p$. The standard model assumes just one covariate and is

$$\begin{aligned} Y_{1i} &= \beta_1^G + \beta_2^G x_i + U_i + \epsilon_{1i} \\ Y_{2i} &= \beta_3^G + \beta_4^G x_i + U_i + \epsilon_{2i} \\ Y_i &= X_i\beta^G + U_i\mathbf{1} + \epsilon_i, \end{aligned} \tag{3}$$

where $Y_i = \begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix}$, $X_i = \begin{pmatrix} 1 & x_i & 0 & 0 \\ 0 & 0 & 1 & x_i \end{pmatrix}$, $\beta^G = (\beta_1^G, \beta_2^G, \beta_3^G, \beta_4^G)^T$, $\mathbf{1} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\epsilon_i = \begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{pmatrix}$, and $x_i \sim N(0, \sigma_x^2)$, $U_i \sim N(0, \sigma_U^2)$, $\epsilon_{1i} \sim N(0, \sigma_{\epsilon_1}^2)$ and $\epsilon_{2i} \sim N(0, \sigma_{\epsilon_2}^2)$. Let $\sigma_1^2 = \sigma_U^2 + \sigma_{\epsilon_1}^2$, $\sigma_2^2 = \sigma_U^2 + \sigma_{\epsilon_2}^2$ and $\rho = \sigma_U^2/\sigma_1\sigma_2$.

Returning to the general case, the influence of missing data depends on the missingness mechanism, that is, the probability model for missingness. Knowing the reason for the missingness is obviously helpful to handle missing data. There are four general *missingness mechanisms* as introduced by Little and Rubin [17] and Wu and Carroll [18]. They are Missing Completely at Random (MCAR), Missing at Random (MAR), Missing Not at Random (MNAR), and Shared Parameter (SP).

For simplicity in this investigation, the parameters are assumed to be common between timepoints. Let the dropout parameters be $\theta = (\theta_0, \theta_1)$. The MAR dropout logistic model is then

$$\begin{aligned} \pi_i(\theta) &= P(R_i = 1 \mid Y_{1i}, Y_{2i}) = \frac{e^{\theta_0 + \theta_1 Y_{1i}}}{1 + e^{\theta_0 + \theta_1 Y_{1i}}} \\ &= \operatorname{expit}(\theta_0 + \theta_1 Y_{1i}). \end{aligned} \tag{4}$$

The missingness is called Missing Not at Random, if it depends on unrecorded information, which predicts the missing values. An example is that a patient was unsatisfied with a particular treatment, and thus this patient is more likely to quit the study. If missingness is not at random, then some bias is expected in inferences.

Let the dropout parameters now be (θ, θ_2). The MNAR version for the two-timepoint example is the logistic model:

$$\begin{aligned} \pi_i(\theta, \theta_2) &= P(R_i = 1 \mid Y_{1i}, Y_{2i}) = \frac{e^{\theta_0 + \theta_1 Y_{1i} + \theta_2 Y_{2i}}}{1 + e^{\theta_0 + \theta_1 Y_{1i} + \theta_2 Y_{2i}}} \\ &= \operatorname{expit}(\theta_0 + \theta_1 Y_{1i} + \theta_2 Y_{2i}). \end{aligned} \tag{5}$$

2.2. LME Least False. In this section, the Linear Mixed Effect (LME) method is investigated, which is based on a maximum likelihood estimating approach. The performance of the LME method under MAR and MNAR dropout is examined. Derivation and illustration of theoretical least false values are made. Assuming a Gaussian random intercept model, the score equation of current interest is [19]

$$\sum_{i=1}^{n}\left[R_i\left\{X_i^TV^{-1}\left(Y_i-X_i\widehat{\beta}\right)\right\}+\frac{(1-R_i)}{\sigma_1^2}\left\{x_{i1}\left(Y_{i1}-x_{i1}^T\widehat{\beta}\right)\right\}\right]=0 \tag{6}$$

where $Y_i=(Y_{i1},Y_{i2})$, X_i is a 2×4 design matrix associated with subject i which is $X_i=\begin{pmatrix}1 & x_i & 0 & 0\\ 0 & 0 & 1 & x_i\end{pmatrix}$, and we will use x_{i1}^T as notation for the first row of X_i; thus $x_{i1}^T=(1,x_i,0,0)$, $\widehat{\beta}^T=(\widehat{\beta}_1,\widehat{\beta}_2,\widehat{\beta}_3,\widehat{\beta}_4)$, and $V=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}$. We can rearrange the terms in (6) to be

$$\sum_{i=1}^{n}\left[R_i\left\{X_i^TV^{-1}X_i\right\}+\frac{(1-R_i)}{\sigma_1^2}\left\{x_{i1}x_{i1}^T\right\}\right]\widehat{\beta} \tag{7}$$

$$=\sum_{i=1}^{n}\left[R_i\left\{X_i^TV^{-1}Y_i\right\}+\frac{(1-R_i)}{\sigma_1^2}\left\{x_{i1}Y_{i1}\right\}\right]. \tag{8}$$

These components are in detail

$$V^{-1}=K\begin{pmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2\\ -\rho\sigma_1\sigma_2 & \sigma_1^2\end{pmatrix} \tag{9}$$

where $K=1/\sigma_1^2\sigma_2^2(1-\rho^2)$, and

$$X_i^TV^{-1}X_i=K\begin{pmatrix}\sigma_2^2 & \sigma_2^2x_i & -\rho\sigma_1\sigma_2 & -\rho\sigma_1\sigma_2x_i\\ \sigma_2^2x_i & \sigma_2^2x_i^2 & -\rho\sigma_1\sigma_2x_i & -\rho\sigma_1\sigma_2x_i^2\\ -\rho\sigma_1\sigma_2 & -\rho\sigma_1\sigma_2x_i & \sigma_1^2 & \sigma_1^2x_i\\ -\rho\sigma_1\sigma_2x_i & -\rho\sigma_1\sigma_2x_i^2 & \sigma_1^2x_i & \sigma_1^2x_i^2\end{pmatrix}. \tag{10}$$

Also

$$x_{i1}x_{i1}^{\mathrm{T}}=\begin{pmatrix}1 & x_i & 0 & 0\\ x_i & x_i^2 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{pmatrix}. \tag{11}$$

Similarly for the right hand side of (8)

$$X_i^TV^{-1}Y_i=K\begin{pmatrix}\sigma_2^2Y_{i1}-\rho\sigma_1\sigma_2Y_{i2}\\ \sigma_2^2Y_{i1}x_i-\rho\sigma_1\sigma_2Y_{i2}x_i\\ \sigma_1^2Y_{i2}-\rho\sigma_1\sigma_2Y_{i1}\\ \sigma_1^2Y_{i2}x_i-\rho\sigma_1\sigma_2Y_{i1}x_i\end{pmatrix}. \tag{12}$$

Finally

$$x_{i1}Y_{i1}=\begin{pmatrix}Y_{i1}\\ x_iY_{i1}\\ 0\\ 0\end{pmatrix}. \tag{13}$$

We assume independent and identically distributed responses, with finite variance for the covariate and error distributions, and dropout probabilities bounded away from both zero and one. On dividing all sums by n, the weak law of large numbers applies and we can replace the sums with expectations as follows:

$$E\left[R\left\{X^TV^{-1}X\right\}+\frac{(1-R)}{\sigma_1^2}\left\{x_1x_1^T\right\}\right]\beta^*=E\left[R\left\{X^TV^{-1}Y\right\}+\frac{(1-R)}{\sigma_1^2}\left\{x_1Y_1\right\}\right]. \tag{14}$$

In the left hand side of (14), there will be two parts. First

$$E\left[R\left\{X^TV^{-1}X\right\}\right] \tag{15}$$

$$=K\begin{pmatrix}\sigma_2^2E[R] & \sigma_2^2E[Rx] & -\rho\sigma_1\sigma_2E[R] & -\rho\sigma_1\sigma_2E[Rx]\\ \sigma_2^2E[Rx] & \sigma_2^2E\left[Rx^2\right] & -\rho\sigma_1\sigma_2E[Rx] & -\rho\sigma_1\sigma_2E\left[Rx^2\right]\\ -\rho\sigma_1\sigma_2E[R] & -\rho\sigma_1\sigma_2E[Rx] & \sigma_1^2E[R] & \sigma_1^2E[Rx]\\ -\rho\sigma_1\sigma_2E[Rx] & -\rho\sigma_1\sigma_2E\left[Rx^2\right] & \sigma_1^2E[Rx] & \sigma_1^2E\left[Rx^2\right]\end{pmatrix}, \tag{16}$$

and second

$$\frac{(1-R)}{\sigma_1^2}x_1x_1^T=\frac{1}{\sigma_1^2}\begin{pmatrix}1-E[R] & E[x]-E[Rx] & 0 & 0\\ E[x]-E[Rx] & E\left[x^2\right]-E\left[Rx^2\right] & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{pmatrix}. \tag{17}$$

Similarly, the right hand side is

$$K\begin{pmatrix}\sigma_2^2E\left[RY_1\right]-\rho\sigma_1\sigma_2E\left[RY_2\right]\\ \sigma_2^2E\left[RY_1x\right]-\rho\sigma_1\sigma_2E\left[RY_2x\right]\\ \sigma_1^2E\left[RY_2\right]-\rho\sigma_1\sigma_2E\left[RY_1\right]\\ \sigma_1^2E\left[RY_2x\right]-\rho\sigma_1\sigma_2E\left[RY_1x\right]\end{pmatrix}+\frac{1}{\sigma_1^2}\begin{pmatrix}E\left[Y_1\right]-E\left[RY_1\right]\\ E\left[Y_1x\right]-E\left[RY_1x\right]\\ 0\\ 0\end{pmatrix}. \tag{18}$$

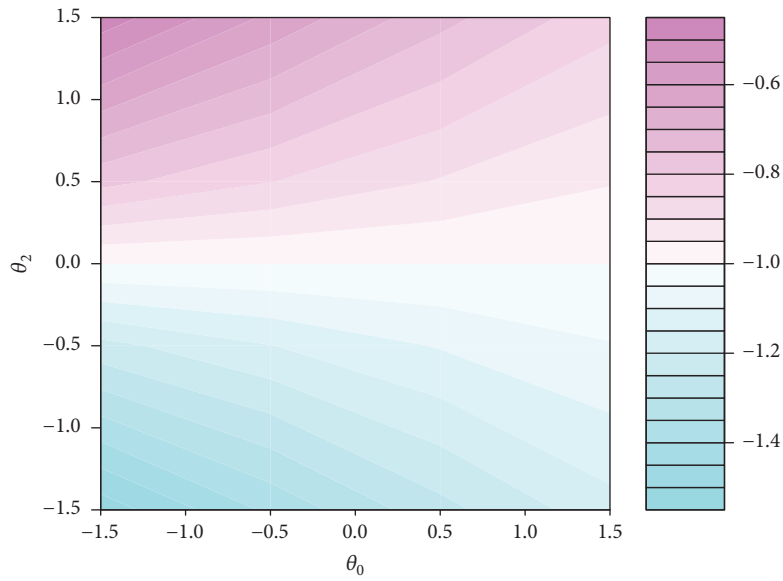

Figure 1: Contour plot of β_3^* under MNAR.

Expressions for $E[R]$, $E[Rx]$, $E[Rx^2]$, $E[RY_1]$, $E[RY_2]$, $E[RY_1x]$, and $E[RY_2x]$ have been obtained under different dropout models. For illustration, we show calculation of $E[R]$ under MAR in the Supplementary Materials available at the journal website (available here).

Finally to find the least false value β^*, the inverse of the matrix has been considered in the left hand side of (14) and we multiply this inverse by the matrix in the right hand side, which will yield the array of the least false values $\beta^{*T} = (\beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*)$. In the following section, we present simulations regarding how the LME method performs under MAR and MNAR dropout model.

2.3. Numerical Investigation. A scalar $N(0,1)$ variable x is generated, and then the longitudinal means are generated $\mu_1=\beta_1 + \beta_2 x$, $\mu_2=\beta_3 + \beta_4 x$. This was followed by (Y_1, Y_2) from a bivariate normal distribution with mean (μ_1, μ_2). Missingness was generated from (4) and (5) for the MAR and MNAR models, respectively. In all of the following simulations, unless it is stated otherwise, the parameters $\beta=(-2,-2,-1,-1)$, $\sigma_x = \sigma_1 = \sigma_2 = 1$, $\rho = 0.5$ were followed. In the following, we show the effect of dropout on the limiting values β_3^* and β_4^*.

As LME provides consistent estimates under MAR, the least false values β_3^* and β_4^* are not affected by changing the dropout probabilities under MAR. Therefore, only MNAR concentrations were considered. From a contour plot of β_3^* under MNAR (Figure 1), in order to minimise the bias in β_3^*, θ_1 should be chosen to be around zero. For negative θ_1, the dropout is associated with large U, so Y_1 and Y_2 both tend to be low if dropout does not occur. Hence β_3^* is lower than it should be. The opposite happens for a positive θ_1.

Figure 2 shows a contour plot of β_4^* under MNAR. Here, negative bias is obtained as θ_1 moves away from zero in either direction. Such an attenuation of regression effect is common when there are errors in variables [20]. It seems that a similar effect is obtained here.

Having obtained least false values, we propose their use in sensitivity analyses. Before doing so, a sensitivity procedure is investigated for local misspecification as proposed by Copas and Eguchi [2].

3. The Effect of Local Misspecification of the Dropout Model When Using Likelihood-Based Methods under the MAR Assumption

In the previous section, we investigated the consequences of misspecifying the missingness mechanism by deriving the so-called least false values, which are the values the parameter estimates converge to when the assumptions may be wrong.

As an alternative, Copas and Eguchi [2] give a formula to estimate the bias under such misspecification using a likelihood approach. As the LME is a likelihood-based method, we can compare the Copas and Eguchi method with the LME least false estimates. The procedure will be applied by adding a tilt to the MAR dropout model to provide what Copas and Eguchi [2] call local misspecification.

3.1. Description of Copas and Eguchi Method. We use the notation of Copas and Eguchi [2], denoting by Z complete data and by Y incomplete data. There are two types of model: the true model and the assumed model. The true model is also called the generating model and it means how the data are actually generated or simulated. On the other hand, the assumed model or what is also known as the fitting model is what we fit to data. The true model for complete data is denoted by $g_Z = g_Z(z;\psi)$ and the corresponding true model for incomplete data is $g_Y = g_Y(y:\psi)$ which can be derived from g_Z. Here ψ is a generic (vector) parameter. The assumed or working model is a parametric model $f_Z = f_Z(z;\psi)$ which gives the distribution of Z, and its marginal density is $f_Y = f_Y(y:\psi)$.

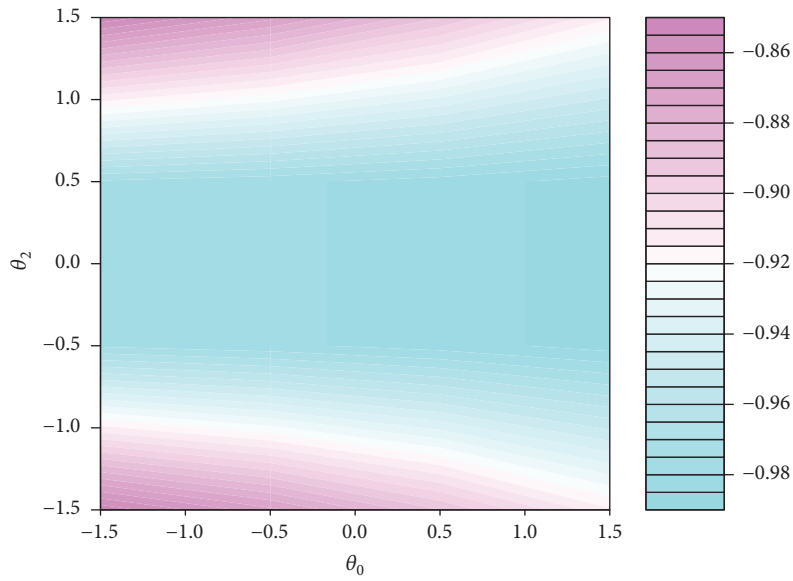

Figure 2: Contour plot of β_4^* under MNAR.

Thus

$$f_Y = \int_{(y)} f_Z \mathrm{d}z \tag{19}$$

where the notation (y) means integration over all missing values in Z that are consistent with the observed Y.

A method is provided to approximate the bias in the estimation of the parameters of the misspecified model following Copas and Eguchi [2]. We consider MAR as the working model and MNAR as the true model. Thus, the misspecification is caused by assuming MAR but the truth is MNAR.

Suppose there is a random sample of n observations, and the true model is given by g_Z, which is defined by (16) in Copas and Eguchi [2] as a tilt model:

$$g_Z = g_Z(z;\psi,\varepsilon,u_Z) = f_Z(z;\psi)\exp\{\varepsilon u_Z(z;\psi)\}. \tag{20}$$

Thus, the misspecification is determined by the quantity $\varepsilon u_Z(z;\psi)$. In this, ε, which is assumed to be small, measures the size of misspecification while $u_Z(z;\psi)$ determines its direction. We assume $u_Z(z;\psi)$ has zero mean and unit variance under the working model f_Z. The misspecification is local because ε is small. Hence, g_Z is close to f_Z and can be written as

$$\frac{g_Z}{f_Z} = \exp\{\varepsilon u_Z(z;\psi)\}. \tag{21}$$

Now if the model actually used to fit the data is $f_Z(z;\psi)$, then the limiting value of the MLE $\widehat{\psi}$ as $n \longrightarrow \infty$ is given by equation (18) in Copas and Eguchi [2]

$$\begin{aligned}\psi_{g_Z} &= \arg_\psi\left[E_g\{s_Z(z;\psi)\} = 0\right]\\ &= \psi + \varepsilon I_Z^{-1} E_{f_Z}\{u_Z(z;\psi)\, s_Z(z;\psi)\},\end{aligned} \tag{22}$$

where $s_Z(\cdot;\psi) = \partial\{\log(f_Z)\}/\partial\psi$ and $I_Z = E[-\partial^2\{\log(f_Z)\}/\partial\psi\partial\theta^T]$ are the score and information matrix for the model f_Z, respectively.

However, f_Y will be considered as the working model for the marginal data. Copas and Eguchi [2] show that if (20) is true and ε is small, then a similar approximation holds for the marginal data Y, i.e.,

$$g_Y = g_Y(y;\psi,\varepsilon,u_Y) = f_Y(y;\psi)\exp\{\varepsilon u_Y(y;\psi)\} \tag{23}$$

where again $u_Y(y;\psi)$ has zero mean and unit variance. In this case according to (19) in Copas and Eguchi [2] the limiting value is

$$\psi_{g_Y} \approx \psi + \varepsilon I_Y^{-1} E_f[u_Y s_Y] = \psi + I_Y^{-1} E_f[\varepsilon u_Y s_Y] \tag{24}$$

where $s_Y(\cdot;\psi) = \partial\{\log(f_Y)\}/\partial\psi$ and $I_Y = E[-\partial^2\{\log(f_Y)\}/\partial\psi\partial\psi^T]$ are the score and information matrix for the model f_Y, respectively. To calculate the bias, $I_Y^{-1}E_f[\varepsilon u_Y s_Y]$, tilt εu_Y. In the next section, how to calculate this amount under MAR and MNAR in our setting of two timepoints will be determined.

3.2. Copas and Eguchi Method for Two-Timepoint Example. The bias consists of, as shown in (24), the score, information matrix, and the tilt. In order to calculate these components, the likelihood model in use is defined. Under MAR, either of the following equivalent formulations can be selected:

$$\begin{aligned}L &= \left(f(Y_1,Y_2)\,P(R=1\mid Y_1,Y_2)\right)^R\\ &\quad\cdot\left(f(Y_1)\,P(R=0\mid Y_1)\right)^{1-R}\end{aligned} \tag{25}$$

$$\begin{aligned}&= \left(f(Y_2\mid Y_1)\,f(Y_1)\,P(R=1\mid Y_1,Y_2)\right)^R\\ &\quad\cdot\left(f(Y_1)\,P(R=0\mid Y_1)\right)^{1-R}.\end{aligned} \tag{26}$$

The conditional distribution of Y_2 given Y_1 is needed quite a lot in this section. Hence, for simplicity, we use Y_{21} to denote this quantity. Since $f(Y_1,Y_2)$ is bivariate normal in this assumed model, $Y_{21} \sim N(\mu_{21},\sigma_{21})$ where $\mu_{21} = \mu_2 +$

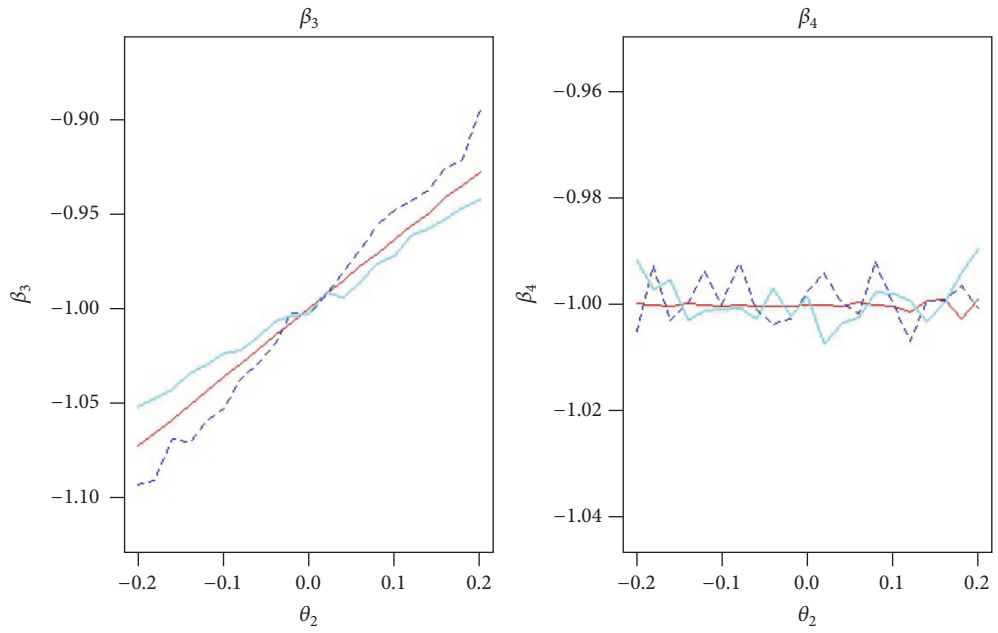

FIGURE 3: Comparison 1: $\beta = (-2, -2, -1, -1)$, $\theta = (-0.5, 0)$. The blue lines (dotted lines) are simulation estimates using maximum likelihood, the red lines (solid lines) are Copas and Eguchi estimates, and the light blue lines are the LME least false estimates.

$(\sigma_2/\sigma_1)\rho(Y_1 - \mu_1)$ and $\sigma_{21} = \sigma_2\sqrt{1-\rho^2}$. Also, the complete data is $Z = (Y_1, Y_2, R)$ and incomplete data is $Y = (Y_1, Y_2^{(R)}, R)$ where

$$Y_2^{(R)} = \begin{cases} Y_2, & R = 1 \\ \text{undefined}, & R = 0. \end{cases} \tag{27}$$

Therefore, at $R = 1$, Y=Z, but Y will differ from Z at $R = 0$.

In addition, the models are defined as f_Z, f_Y, g_Z, and g_Y. MAR is assumed as the working model or misspecified model. Under MAR, there is $P(R = 1 \mid Y_1, Y_2) = P(R = 1 \mid Y_1)$; then from (25) the working model for complete data by assuming $R = 1$ is

$$f_Z = f(Y_1, Y_2)\, P(R = 1 \mid Y_1). \tag{28}$$

Similarly, from (26) the working model for incomplete data by assuming $R = 0$ is

$$f_Y = f(Y_1)\, P(R = 0 \mid Y_1). \tag{29}$$

Under MNAR, if there is complete data, then we will always set $R = 1$. Thus, from (25), the true model for complete data is

$$g_Z = f(Y_1, Y_2)\, P(R = 1 \mid Y_1, Y_2). \tag{30}$$

The true model for incomplete data on the other hand is the marginal density:

$$g_Y = \int_y g_Z dz = \int_{Y_2^R} f(Y_1, Y_2)\, P(R = 1 \mid Y_1, Y_2)\, dY_2^R. \tag{31}$$

Note that the integral is over the missing values Y_2^R. Referring to (27), the missing values Y_2 are undefined in case that $R = 0$.

This means that in order to use Copas and Eguchi's ideas, we should convert the specific g_Y in (31) into the general form of (23). To do this, we will redefine the MNAR model in tilt form:

$$P(R = 1 \mid Y_1, Y_2) = \text{expit}\{\theta_0 + \theta_1 Y_1\} \exp\{\varepsilon u_Y\}. \tag{32}$$

Here $\varepsilon = \theta_2\sigma_{21}$ and $u_Y = u_Y(y;\theta) = (Y_2 - \mu_{21})/\sigma_{21}$. For small θ_2 this is a good approximation to the logistic MNAR model.

Calculation of the terms needed for the bias expression (24) is now possible and follows directly. Details are in the Supplementary Materials available at the journal website.

4. Simulation Study

We use the same simulation setup as before. The limiting values β_3^* and β_4^* are compared using different methods under MAR and MNAR dropout models. Next, the local model uncertainty will be elaborated as proposed by Copas, and we illustrate how to apply it both when model misspecification is present and when the data is incomplete. We find that the Copas and Eguchi [2] method gives very similar results to the least false. Misspecification will be dealt with assuming MAR where actually the truth is MNAR.

4.1. Comparing the Copas and Eguchi Method with LME Least False Results. In this section, the parameter estimates are affected when a MAR model is fitted to data that are MNAR, and compared with the values that the Copas and Eguchi method predicts.

The sample size is 10000, and 10 simulations are used. We used large samples here, as our first task is to check the accuracy of the large-sample approximations underpinning the least false values. The aim is to show the variation in treatment effect estimates as θ_2 varies. A grid of θ_2 from -0.2 to 0.2 is selected. Figure 3 is produced when β=(-2,-2,-1,-1)

Table 1: CI coverage in percent for the estimated β_3 and β_4 at assumed θ_2=0. We use θ_2^T for the true θ_2, θ_2^A for the assumed value in adjusting the estimates, $(\beta_3^{**}, \beta_4^{**})$ for the Copas and Eguchi adjustment method, and (β_3^*, β_4^*) for the least false adjustment method. Results based on 1000 samples of size 1000.

θ_2^T	θ_2^A	β_3^{**}	β_4^{**}	β_3^*	β_4^*
-0.10	0.00	84.80	95.30	84.90	95.30
-0.09	0.00	85.90	96.70	85.80	96.70
-0.06	0.00	92.00	94.70	91.90	94.70
-0.03	0.00	95.00	95.00	95.10	94.90
0.00	0.00	94.70	95.10	94.70	95.10
0.03	0.00	95.20	94.20	95.00	94.20
0.06	0.00	91.70	94.70	91.70	94.70
0.09	0.00	88.00	95.00	87.80	95.00
0.10	0.00	83.40	95.10	83.60	95.00

Table 2: CI coverage for the estimated β_3 and β_4 in percent at assumed θ_2=-0.10. We use θ_2^T for the true θ_2, θ_2^A for the assumed value in adjusting the estimates, $(\beta_3^{**}, \beta_4^{**})$ for the Copas and Eguchi adjustment method, and (β_3^*, β_4^*) for the least false adjustment method. Results based on 1000 samples of size 1000.

θ_2^T	θ_2^A	β_3^{**}	β_4^{**}	β_3^*	β_4^*
-0.10	-0.10	95.30	95.40	95.70	95.10
-0.09	-0.10	94.80	95.10	95.50	94.80
-0.06	-0.10	95.80	95.20	94.90	95.60
-0.03	-0.10	92.50	94.90	92.00	95.10
0.00	-0.10	89.30	95.30	87.40	95.40
0.03	-0.10	83.70	93.80	81.20	93.80
0.06	-0.10	74.40	95.10	70.50	95.00
0.09	-0.10	62.20	96.00	58.30	95.80
0.10	-0.10	62.70	93.90	57.80	94.20

and θ=(-0.5,0), which gives dropout rate around 40%. Here the blue lines (dotted lines) are simulation estimates using maximum likelihood, the red lines (solid lines) are Copas and Eguchi estimates, and the light blue lines are the LME least false estimates. These show that the least false, simulations, and Copas and Eguchi [2] results all match well. Therefore, we can use the least false results for bias correction as an alternative to Copas and Eguchi.

4.2. CI Coverage for the Estimated β_3 and β_4. The Copas and Eguchi and LME least false values show how estimates are biased by assuming MAR when the data are MNAR. The misspecification parameter is θ_2, with θ_2=0 meaning no misspecification. If the value of θ_2 was known, then the parameter estimates will be adjusted to take into account the misspecification. This idea will be illustrated in this section.

For a range of true (generating) θ_2, 1000 samples are simulated, each of size 1000. This is a realistic number for applications. In each case, β_3 and β_4 are estimated using maximum likelihood under a MAR assumption. Afterwards, the estimates are adjusted using either the estimated Copas and Eguchi bias or the bias arising through least false calculations, in both cases taking an *assumed* θ_2. Coverage of the resulting nominal 95% confidence intervals is then recorded. The estimated confidence interval width is not adjusted, just its location.

Tables 1 and 2 give the results. Here we use θ_2^T for the true θ_2, and θ_2^A denotes the assumed value used in adjusting the estimates. Also, $(\beta_3^{**}, \beta_4^{**})$ are used for the Copas and Eguchi adjustment method and (β_3^*, β_4^*) for the least false adjustment method.

In Table 1, the assumed θ_2 is zero, meaning no correction. Results at the correct value of θ_2^A =0 are good. Otherwise, the CI for β_3 goes badly wrong. Note that there is no correction here, so the Copas and Eguchi and least false results should be the same. Small differences are just because of the different calculations that are involved. For example, the least false calculation needs an estimate of σ_x but the Copas and Eguchi one does not. The CI coverage is noted for β_4 which is not too much affected at any true θ_2 in the range (-0.1,+0.1). For example, at θ_2^A=-0.1, the CI coverage for β_4 is about 95%, whereas there is undercoverage for β_3 when θ_2^T deviates from zero. For example at θ_2^T=-0.1, the CI coverage for β_3 is about 85%. This indicates that β_4 is less sensitive to the misspecification than β_3 in this scenario.

In Table 2, the assumed value is taken of θ_2=-0.1, which means that dropout is associated with high Y_2. Note that, in contrast to the Table 2, there is correction here, so the Copas and Eguchi and least false results will not be the same; for example, at θ_2^T=+0.1, the CI coverage for β_3^{**} is about 62.7%, but the CI coverage for β_3^* is about 57.8%. However, both

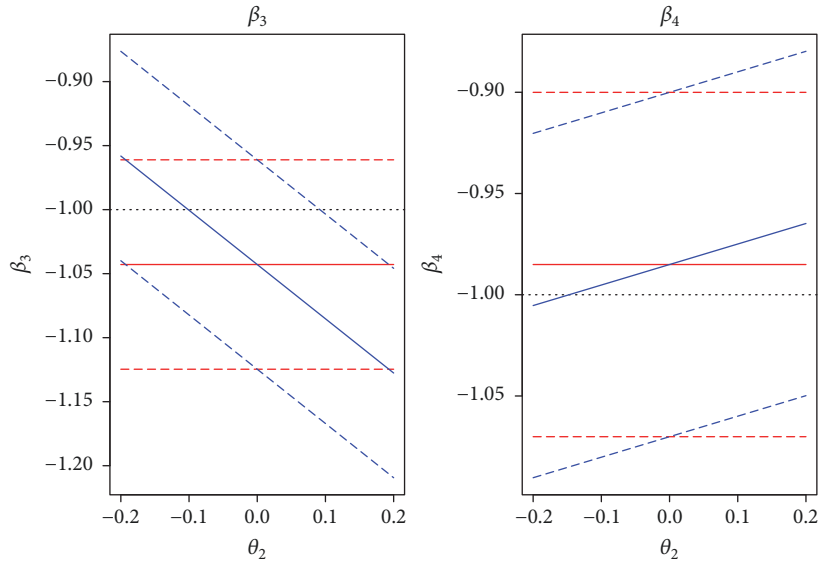

FIGURE 4: CI under MAR: $\beta = (-2, -2, -1, -1)$, $\theta = (-0.5, -0.5)$, θ_2^T=0. The blue lines are the adjusted estimates, and red lines are the unadjusted estimates. The horizontal dotted lines are at the true values.

estimates β_3^{**} and β_3^* have undercoverage as θ_2^T goes further from the assumed value -0.1.

4.3. Sensitivity Analysis. Of course, in practice θ_2 is not known. For any given data set, a sensible sensitivity procedure would mean plotting bias-corrected estimates and confidence intervals for a range of assumed θ_2 values. Here, a grid of assumed θ_2 is used from -0.2 to 0.2. We will show that, for each limiting value calculated by the Copas and Eguchi method, the simulated values are within noise of the theoretical values for large sample sizes (n=10000). The noise is estimated from the simulations; that is, a confidence interval is achieved from the simulations with reassurance that the population values are present. A correct MAR model is obtained and after that, under true MNAR, MAR is assumed.

Figure 4 illustrates the case when MAR is the correct model ($\theta_2 = 0$) and the unadjusted confidence intervals (red lines) include the true parameter values (β_3=-1 and β_4=-1), as in this case so do the adjusted ones (blue lines). The horizontal lines are at the true values. We note that β_3^{**} decreases as θ_2 increases whereas β_4^{**} increases as θ_2 increases. Note that β_4 has a wider CI than β_3.

Figure 5 has the true θ_2=0.1, so the study has fitted MAR to data that are really MNAR. The lines cross at θ_2=0 because the same MAR model is fitted. The important point is that better estimates of the true β's are obtained at the correct θ_2. Also, as mentioned in Figure 4, β_4 has wider CI than β_3.

Note that, both under MAR and MNAR, β_3 and β_4 have opposite trends; β_3 decreases as θ_2 increases whereas β_4 increases as θ_2 increases.

5. Application: Sensitivity Analysis for Clinical Trial

In this section, the method is illustrated using a real data example. The data is considered from a clinical trial with two treatments and two measurement times as introduced and analysed by Mathews et al. [12]. The covariates are only treatment type and time. The parameter vector is $(\beta_1, \beta_2, \beta_3, \beta_4)$, ignoring any time interaction. There are 422 subjects, assigned to either treatment A or B. Treatment A is associated with treatment effect x=1 and treatment B is when x=0. Then, at time 2, the mean of the group receiving treatment B is β_3 and the mean of the group receiving treatment A is $\beta_3+\beta_4$. At time 1, all subjects provided a response, but 24.4% dropped out by time 2. There are 212 subjects receiving treatment A, but only 126 provided a response at time 2 and the other 86 dropped out. Hence the missingness percentage is about 40%. The dropout reason is not known. For treatment B, there are 210 subjects, of which 193 subjects continued to time 2 and hence there are 17 that did not, and this gave around 8% missingness.

A sensitivity analysis approach (over a grid of θ_2) using the Copas and Eguchi and LME methods is shown in Figure 6. The blue lines use the Copas and Eguchi method and the red lines use the least false method. The idea is to adjust the estimate to compensate for bias from a misspecified MAR fit. Consequently, for example, if the least false value is known under MAR to underestimate a parameter, the difference for the estimate is added to back-calculate. Dashes are the CIs, based on the MAR standard errors. The first plot shows confidence intervals for the treatment B mean as the assumed value of θ_2 changes. The horizontal line is the estimate under MAR. The second plot shows the confidence intervals for the mean of treatment A. The third plot is the difference in means between treatment A and B, which yields the treatment effect means, i.e., β_4 means. In the first plot, the horizontal line is at -0.74 which is the same value for the LME estimate for β_3. Again, the LME estimate for β_4 is about -0.40 in Figure 2. Also, note that $\beta_3+\beta_4$ equals -1.15. This supports the finding here and shows better results.

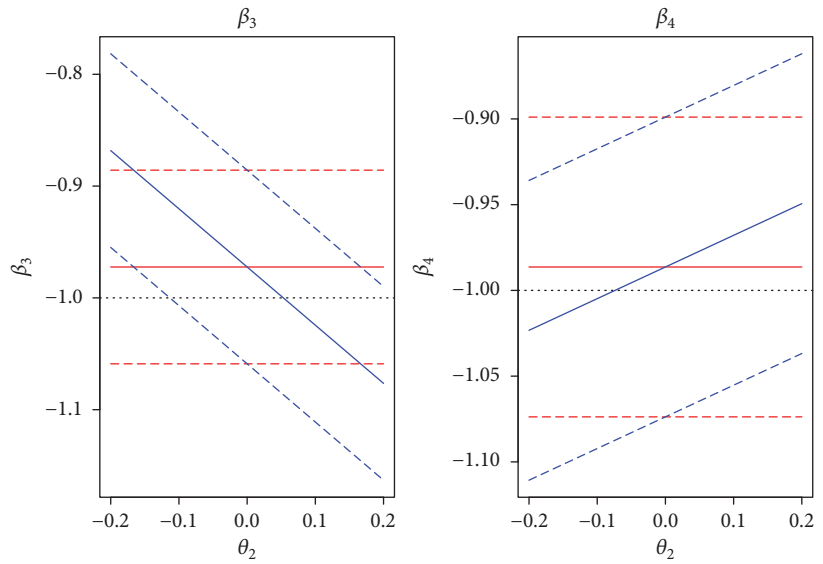

FIGURE 5: CI under MNAR: $\beta = (-2, -2, -1, -1)$, $\theta = (-0.5, -0.5)$, θ_2^T=0.1. The blue lines are the adjusted estimates, and red lines are the unadjusted estimates. The horizontal lines are at the true values.

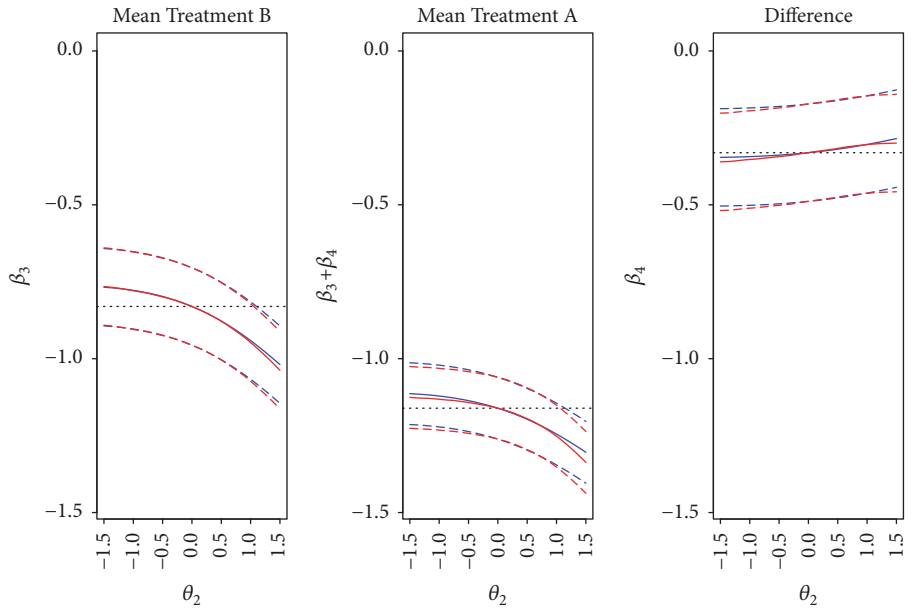

FIGURE 6: Clinical trial example: 95% CI for β_3, $\beta_3 + \beta_4$, and β_4. The blue lines use the Copas and Eguchi method, the red lines use the least false method, and the horizontal line is at the MAR estimate.

The first thing to note is how close the least false and Copas and Eguchi estimates are. There is almost no difference over this range of θ_2. We take θ_2 from -1.5 to +1.5. The value of $\widehat{\theta_1}$ under MAR is -1.66, meaning the range of θ_2 allows Y_2 to have the same order of effect as Y_1. Clearly at large values of θ_2, there is concern that the misspecification is not local, which is the assumption of Copas and Eguchi. However, the least false results apply to any misspecification, not necessarily local, and the fact that Copas and Eguchi estimate is so close to the least false one suggests that it can work well even under quite large misspecification.

When θ_2 is negative, the estimates get adjusted upwards, and the opposite is true for positive θ_2. This makes sense: at negative θ_2, large Y_2 values have low probability of staying in the trial. Hence the observed means are lower than they would be in the hypothetical no-dropout situation, so we adjust upwards.

The estimates seem to be affected more at the positive θ_2 than at the negative one. At the very largest θ_2 shown, there would be a significant change in the value of the estimated true mean. However, there is very little effect of misspecification on the difference between means (third subplot), as the adjustments essentially cancel.

6. Conclusion

We considered the Linear Mixed Effect models (maximum likelihood method) for handling missing data. Then, by deriving the so-called least false values, we investigated the consequences of misspecifying the missingness mechanism. The closed form expressions were given to calculate the least false values β_3^* and β_4^*. The knowledge of these least false values allowed us to conduct sensitivity analysis, which was illustrated for the LME method.

Copas and Eguchi [2] gave a formula to estimate the bias under the misspecification. We derived and explored the Copas and Eguchi approximation for the bias raised by the misspecification of the working model. The results found by using Copas and Eguchi method are compared with the results obtained by the method proposed. Also, we applied the Copas and Eguchi method to estimate the bias for the real data example.

Moreover, we explained how to use a sensitivity analysis to see how the methods work under a range of θ_2. We found that the Copas and Eguchi method and LME least false match very well. Both gave very close results over the grid of θ_2 considered. This suggests that the least false method can provide a credible alternative to Copas and Eguchi in sensitivity analysis. In fact, it might be preferred since there is no assumption of local misspecification. Finally, we illustrated the results using example data from a clinical trial with two measurement times.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This project was supported by King Saud University, Deanship of Scientific Research, College of Science Research Center.

Supplementary Materials

Supplementary material is available online at the journal website. There are two sections. One is for the least false estimate, which shows the calculations for E[R] in (18). The other section illustrates the Copas and Eguchi bias. It shows the calculation of the terms needed for the bias expression (24). *(Supplementary Materials)*

References

[1] K. Mohan and J. Pearl, "Graphical models for processing missing data," Tech. Rep. R-473, Department of Computer Science, University of California, Los Angeles, Calif, USA, 2017.

[2] J. Copas and S. Eguchi, "Local model uncertainty and incomplete-data bias," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 67, no. 4, pp. 459–513, 2005.

[3] X. de Luna and M. Lundin, "Sensitivity analysis of the unconfoundedness assumption with an application to an evaluation of college choice effects on earnings," *Journal of Applied Statistics*, vol. 41, no. 8, pp. 1767–1784, 2014.

[4] P. R. Rosenbaum, *Design of Observational Studies*, Springer, 2010.

[5] N. X. Lin, J. Q. Shi, and R. Henderson, "Doubly misspecified models," *Biometrika*, vol. 99, no. 2, pp. 285–298, 2012.

[6] N. Tang, S. Chow, J. G. Ibrahim, and H. Zhu, "Bayesian sensitivity analysis of a nonlinear dynamic factor analysis model with nonparametric prior and possible nonignorable missingness," *Psychometrika*, vol. 82, no. 4, pp. 875–903, 2017.

[7] H. Zhu, J. G Ibrahim, and N. Tang, "Bayesian sensitivity analysis of statistical models with missing data," *Statistica Sinica*, vol. 24, pp. 871–896, 2014.

[8] M. J. Daniels and J. W. Hogan, *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*, CRC Press, New York, NY, USA, 2008.

[9] H. Zhu, J. G. Ibrahim, and N. Tang, "Bayesian influence analysis: a geometric approach," *Biometrika*, vol. 98, no. 2, pp. 307–323, 2011.

[10] K. Imai, Y. Lu, and A. Strauss, "Bayesian and likelihood inference for 2 × 2 ecological tables: An incomplete-data approach," *Political Analysis*, vol. 16, no. 1, pp. 41–69, 2008.

[11] P. Yin and J. Q. Shi, "Simulation-based sensitivity analysis for non-ignorably missing data," *Statistical Methods in Medical Research*, vol. 28, no. 1, pp. 289–308, 2017.

[12] J. N. Matthews, R. Henderson, D. M. Farewell, W. Ho, and L. R. Rodgers, "Dropout in crossover and longitudinal studies: Is complete case so bad?" *Statistical Methods in Medical Research*, vol. 23, no. 1, pp. 60–73, 2014.

[13] P. Diggle, D. Farewell, and R. Henderson, "Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal," *Applied Statistics*, vol. 56, no. 5, pp. 499–550, 2007.

[14] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982.

[15] A. Cnaan, N. M. Laird, and P. Slasor, "Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data," *Statistics in Medicine*, vol. 16, no. 20, pp. 2349–2380, 1997.

[16] G. Molenberghs and G. Verbeke, "A review on linear mixed models for longitudinal data, possibly subject to dropout," *Statistical Modelling*, vol. 1, pp. 235–269, 2001.

[17] R. Little and D. Rubin, *Statistical Analysis with Missing Data*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2002.

[18] M. C. Wu and R. J. Carroll, "Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process," *Biometrics*, vol. 44, no. 1, pp. 175–188, 1988.

[19] A. Almohisen, *Least-false and local misspecification methods for longitudinal data with dropout [PhD thesis]*, Newcastle University, Newcastle Upon Tyne, UK, 2017.

[20] R. Carroll, D. Ruppert, L. Stefanski, and C. Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective*, Chapman and Hall, 2nd edition, 2006.