Published in final edited form as:
Scand Stat Theory Appl. 2009 December 1; 36(4): 686–712. doi:10.1111/j.1467-9469.2009.00644.x.

Diagnostic Measures for Generalized Linear Models with Missing Covariates

HONGTU ZHU, **JOSEPH G. IBRAHIM**, and **XIAOYAN SHI**
Department of Biostatistics, University of North Carolina at Chapel Hill

Abstract

In this paper, we carry out an in-depth investigation of diagnostic measures for assessing the influence of observations and model misspecification in the presence of missing covariate data for generalized linear models. Our diagnostic measures include case-deletion measures and conditional residuals. We use the conditional residuals to construct goodness-of-fit statistics for testing possible misspecifications in model assumptions, including the sampling distribution. We develop specific strategies for incorporating missing data into goodness-of-fit statistics in order to increase the power of detecting model misspecification. A resampling method is proposed to approximate the *p*-value of the goodness-of-fit statistics. Simulation studies are conducted to evaluate our methods and a real data set is analysed to illustrate the use of our various diagnostic measures.

Keywords

Cook's distance; goodness-of-fit; missing covariates; residuals

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 distributional and modelling assumptions at various stages. Therefore, the resulting estimates and tests may be sensitive to these assumptions. For this reason, sensitivity analyses are commonly performed to check the sensitivity of the parameters of interest with respect to the model assumptions.

Diagnostic measures such as residuals and Cook's distance have been widely used to identify influential observations in various regression models, such as generalized linear models (GLMs) (Cox & Snell, 1968; Cook & Weisberg, 1982; Davison & Tsai, 1992; Zhu *et al.*, 2001). In addition, diagnostic measures, such as residuals, can be used to construct goodness-of-fit statistics to detect any systematic discrepancies between the data and the fitted values obtained from the model (Stute, 1997; Lin *et al.*, 2002). However, to the best of our knowledge, virtually no literature exists for developing diagnostic measures such as residuals, Cook's distance and goodness-of-fit statistics in GLMs with missing covariate data.

The aim of this paper is to systematically investigate various diagnostic measures for GLMs with missing at random (MAR) covariates as well as not missing at random (NMAR) covariates, often referred to as non-ignorably missing covariates. MAR also includes missing completely at random (MCAR) covariates as a special case. Data are said to be MCAR if the

Hongtu Zhu, Department of Biostatistics, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420, USA. hzhu@bios.unc.edu.

failure to observe a value does not depend on any data, either observed or missing, whereas data are said to be MAR if, conditional on the observed data, the failure to observe a value does not depend on the data that are unobserved. The missing data mechanism is NMAR if the failure to observe a value depends on the value that would have been observed (Ibrahim *et al.*, 2005). We propose two case-deletion measures, namely Cook's distance and the Q-displacement, based on the conditional expectation of the complete-data log-likelihood function in the expectation–maximization (EM) algorithm (Zhu *et al.*, 2001). We formally define conditional residuals and examine their properties under different missing data mechanisms, such as MAR and NMAR, and then we develop conditional residual processes to construct goodness-of-fit statistics. Moreover, we develop specific strategies for incorporating missing covariate data into the goodness-of-fit statistics in order to increase the power of detecting model misspecification.

The model assessment methodology we develop here is crucial for missing data problems and the first of its kind. It is important as:

i. it often turns out that covariates with missing values may in fact lead to cases with influential observations and one cannot just delete the cases with missing values and carry out a complete case analysis to examine which cases are influential;

ii. developing methods for assessing MAR and NMAR models as part of a sensitivity analysis is one of the most important problems in missing data, and the diagnostic and goodness-of-fit methodology we develop here is perfectly suited for this problem;

iii. model assessment and goodness-of-fit in the presence of missing data is a very important problem whose development is quite different from methods based on complete data, as one needs to appropriately define residuals and other quantities in the context of missing data, and these statistics have very different small and large sample properties and operating characteristics than statistics based on complete data methods.

To motivate the proposed methodology, we consider data on 191 patients from two Eastern Cooperative Oncology Group clinical trials (Ibrahim *et al.*, 1999), which is discussed in more detail in section 5. The primary interest here was to find how the number of cancerous liver nodes (response) when entering the trials is predicted by six other baseline characteristics: time since diagnosis of the disease (in weeks); two biochemical markers (each classified as normal or abnormal), alpha-fetoprotein and anti-hepatitis B antigen; associated jaundice (yes, no); body mass index (weight in kilograms divided by the square of height in metres); and age (in years). From these six covariates, three had missing data and the remaining covariates were completely observed. The three with missing data were time since diagnosis of the disease, alpha-fetoprotein and anti-hepatitis B antigen, with 8.9%, 5.8% and 18.3% missingness percentages, yielding a total missingness percentage of 29%. Table 1 shows all the potentially influential cases, where cases 10, 15, 65 and 160 have abnormally large response values and case 131 has an extreme covariate value in time since diagnosis compared with the rest of the cases. In this paper, we will develop a formal methodology to assess such cases. In section 5, we revisit this data set and use our proposed methodology to determine whether these cases are influential or not.

The rest of this paper is organized as follows. In section 2, we review the model assumptions for GLMs with missing covariates and the EM algorithm for calculating maximum likelihood estimates (MLEs). In section 3, we develop new diagnostic measures, including case-deletion diagnostics and conditional residuals and examine their properties. We construct goodness-of-fit statistics based on conditional residuals. We present several simulation studies in section 4, and analyse the liver cancer data set in section 5. We conclude the paper with some final remarks in section 6. Proofs are collected in the Appendix.

2. Preliminaries

Consider n independent observations $(\mathbf{x}_1, \mathbf{z}_1, y_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n, y_n)$, where y_i is the response variable, $\mathbf{x}_i$ is a p_1-dimensional vector of completely observed covariates, and $\mathbf{z}_i$ is a p_2-dimensional vector of partially observed covariates. Moreover, let $\mathbf{z}_{m,i}$ and $\mathbf{z}_{o,i}$ denote the missing and observed components of $\mathbf{z}_i$ respectively. Let $\mathbf{r}_i$ be a p_2-dimensional random vector, whose kth component, r_{ik} equals 1, if z_{ik} is observed for subject i, and 0, if z_{ik} is missing, where z_{ik} is the kth component of $\mathbf{z}_i$. Under the NMAR setting, we need to specify the joint distribution of $(\mathbf{x}_i, \mathbf{z}_i, \mathbf{r}_i, y_i)$ for each i. It is common to decompose $p(\mathbf{x}_i, \mathbf{z}_i, \mathbf{r}_i, y_i|\eta)$ into a product of three conditional distributions as follows:

$$p(\mathbf{x}_i, \mathbf{z}_i, \mathbf{r}_i, y_i|\eta)=p(y_i|\mathbf{x}_i, \mathbf{z}_i, \eta)p(\mathbf{x}_i, \mathbf{z}_i|\eta)p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \eta), \qquad (1)$$

where η denotes the vector of all unknown parameters as defined below.

Modelling $(\mathbf{x}_i, \mathbf{z}_i, \mathbf{r}_i, y_i)$ usually involves three levels of assumptions. We assume a GLM for the conditional distribution of y_i given $(\mathbf{x}_i, \mathbf{z}_i)$ (Ibrahim, 1990; Ibrahim & Lipsitz, 1996; Lipsitz & Ibrahim, 1996; Little & Rubin, 2002). Specifically, y_i given $(\mathbf{x}_i, \mathbf{z}_i)$ has a density in the exponential family

$$p(y_i|\mathbf{x}_i, \mathbf{z}_i, \beta, \tau)=\exp\{a_i^{-1}(\tau)[\, y_i\theta_i(\beta) - b(\theta_i(\beta))]+c(y_i, \tau)\}, \qquad (2)$$

$i = 1, \ldots, n$, indexed by the canonical parameter θ_i and the scale parameter τ, where the functions $b(\cdot)$ and $c(\cdot, \cdot)$ determine a particular distributional family in the class, such as the binomial, normal or Poisson distribution. The functions $a_i(\tau)$ are commonly of the form $a_i(\tau)=\tau^{-1}k_i^{-1}$, where the k_is are known weights. Further, the θ_is satisfy the equations $\theta_i = \theta(\mu_i)$, $i = 1, \ldots, n$, and $\mu_i=g((\mathbf{x}_i', \mathbf{z}_i')\beta)$ are the components of $\mu= E\,(\mathbf{y}|\mathbf{x}, \mathbf{z}, \beta, \tau)$, where $g(\cdot)$ is a known link function and $\beta = (\beta_1, \ldots, \beta_p)'$ is a p-dimensional vector of regression coefficients ($p = p_1 + p_2$). The GLMs include many well-known regression models, such as normal linear regression, logistic and probit regression, Poisson regression, gamma regression and some proportional hazards models (McCullagh & Nelder, 1989).

We also need to specify a distribution for the missing covariates $\mathbf{z}_i$. For large p, modelling the covariates usually involves several assumptions. To reduce the number of parameters, we follow Lipsitz & Ibrahim (1996) and Ibrahim *et al.* (1999) and write $p(\mathbf{z}_i|\mathbf{x}_i, \alpha)$ as a sequence of one-dimensional conditional distributions:

$$p(\mathbf{z}_i|\mathbf{x}_i, \alpha)=p(z_{ip_2}|z_{i(p_2-1)}, \ldots, z_{i1}, \mathbf{x}_i, \alpha)\cdots p(z_{i1}|\mathbf{x}_i, \alpha), \qquad (3)$$

where α is a subvector of η. Furthermore, we typically assume specific parametric forms for these one-dimensional conditional distributions. As the $\mathbf{x}_i$s are fully observed, it is not necessary to specify a distribution for $\mathbf{x}_i$.

One way of modelling the missing data mechanism $p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \xi)$ is to use a joint log-linear model (Lipsitz & Ibrahim, 1996). Following Ibrahim *et al.* (1999), another way of modelling $p(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i, \xi)$ is to assume that

$$p(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i, \xi)=p(r_{ip_2}|r_{i(p_2-1)}, \ldots, r_{i1}, \mathbf{x}_i, \mathbf{z}_i, y_i, \xi)\cdots p(r_{i2}|r_{i1}, \mathbf{x}_i, \mathbf{z}_i, y_i, \xi)p(r_{i1}|\mathbf{x}_i, \mathbf{z}_i, y_i, \xi). \qquad (4)$$

It is common to use logistic regression models for the binary variables r_{ij}.

The EM algorithm has been a popular technique for obtaining the MLE of $\eta = (\beta, \tau, \alpha, \xi)'$ in GLMs with missing covariate data (Little & Schluchter, 1985; Schluchter & Jackson, 1989; Ibrahim, 1990; Ibrahim & Lipsitz, 1996; Lipsitz & Ibrahim, 1996, 1998; Little & Rubin, 2002). Let

$$D_c = \{(y_j, \mathbf{x}_j, \mathbf{z}_j, \mathbf{r}_j): j = 1, \ldots, n\}$$

be the complete data,

$$D_o = \{\mathbf{d}_{o,i} = (y_i, \mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i): i = 1, \ldots, n\}$$

be the observed data, and $D_m = (\mathbf{z}_{m,1}, \ldots, \mathbf{z}_{m,n})$ be the missing data. At the sth step of the EM algorithm, given $\eta^{(s)}$, the E-step involves evaluating the Q-function, given by

$$\begin{aligned} Q(\eta|\eta^{(s)}) &= E[L_c(\eta|D_c)|D_o, \eta^{(s)}] \\ &= \sum_{i=1}^{n} \int \log[p(y_i|\mathbf{x}_i, \mathbf{z}_i, \beta, \tau)] p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta^{(s)}) d\mathbf{z}_{m,i} \\ &+ \sum_{i=1}^{n} \int \log[p(\mathbf{x}_i, \mathbf{z}_i|\alpha)] p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta^{(s)}) d\mathbf{z}_{m,i} \\ &+ \sum_{i=1}^{n} \int \log[p(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i, \xi)] p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta^{(s)}) d\mathbf{z}_{m,i} \\ &= Q_1(\beta, \tau|\eta^{(s)}) + Q_2(\alpha|\eta^{(s)}) + Q_3(\xi|\eta^{(s)}), \end{aligned} \tag{5}$$

where $L_c(\eta|D_c) = \log p(D_c|\eta)$ is the complete-data log-likelihood function. The M-step consists of maximizing $Q_1(\beta, \tau|\eta^{(s)})$, $Q_2(\alpha|\eta^{(s)})$ and $Q_3(\xi|\eta^{(s)})$ separately (Ibrahim *et al.*, 1999).

Our main interest is to make valid inferences about β, and this requires the correct specification of all three levels of assumptions in (1). Misspecifying some of those modelling assumptions may introduce serious bias in β. Thus, it is crucial to assess the potential degree of misspecification at each of the three levels of assumptions in (1).

3. Diagnostic measures

We define the following two types of diagnostic measures: case-deletion measures and conditional residuals for formal and informal examination of the adequacy of a GLM with missing covariates. The two case-deletion measures, Cook's distance and the Q-displacement, can be used to examine the effects of deleting individual observations on the estimate of η. The conditional residuals carry important information about the influence of observations. We use the conditional residuals to construct goodness-of-fit statistics for testing the validity of particular model assumptions.

3.1. Case-deletion influence measures

To quantify the effects of deleting the ith observation on the MLE, $\hat{\eta}$ of η, we define the MLE of η for a subsample $D_{c[i]}$, in which the ith observation $\mathbf{d}_i = (y_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{r}_i)$ is deleted from D_c. For the subsample $D_{c[i]}$, we define $\mathrm{Q}_{[i]}(\eta|\hat{\eta})$ as

$$Q_{[i]}(\eta|\widehat{\eta})=E[L_c(\eta|D_{c[i]})|D_o,\widehat{\eta}],$$

where the expectation is taken with respect to $p(D_m|D_o, \hat{\eta})$. Then we define $\hat{\eta}_{[i]}$ as the maximizer of $Q_{[i]}(\eta|\hat{\eta})$. Following Zhu *et al.* (2001), we calculate a one-step approximation $\widehat{\eta}^1_{[i]}$ of $\hat{\eta}_{[i]}$ as follows:

$$\widehat{\eta}^1_{[i]}=\widehat{\eta}+\{-\partial^2_\eta Q(\eta|\widehat{\eta})\}^{-1}\partial_\eta Q_{[i]}(\eta|\widehat{\eta})|_{\eta=\widehat{\eta}}, \tag{6}$$

where ∂_η and ∂^2_η represent the first- and second-order derivatives with respect to η. In (6), several degrees of approximation are used, but this is usually adequate for diagnostic purposes (Cook & Weisberg, 1982; Zhu *et al.*, 2001). As $\partial_\eta Q(\eta|\hat{\eta})|_{\eta=\hat{\eta}} = \mathbf{0}$,

$$\begin{aligned}\partial_\eta Q_{[i]}(\eta|\widehat{\eta})|_{\eta=\widehat{\eta}}&=-\int\partial_\eta\log\{p(\mathbf{d}_i|\widehat{\eta})\}p(\mathbf{z}_{m,i}|\mathbf{x}_i,\mathbf{z}_{o,i},y_i,\mathbf{r}_i,\widehat{\eta})d\mathbf{z}_{m,i}\\&=-\partial_\eta\log p(\mathbf{x}_i,\mathbf{z}_{o,i},y_i,\mathbf{r}_i|\widehat{\eta}).\end{aligned}$$

We introduce two case-deletion measures to quantify the distance between the MLE of η with and without the ith observation deleted from the full sample (Cook & Weisberg, 1982; Zhu *et al.*, 2001). Cook's distance, denoted by $\mathrm{CD}_i(M)$, in this setting is defined as

$$\mathrm{CD}_i(M)=(\widehat{\eta}^1_{[i]}-\widehat{\eta})'M(\widehat{\eta}^1_{[i]}-\widehat{\eta}), \tag{7}$$

where M is chosen to be a positive definite matrix. For simplicity, we use CD_i to denote $\mathrm{CD}_i(M)$, when $M=-\partial^2_\eta Q(\eta|\widehat{\eta})|_{\eta=\widehat{\eta}}$. Similar to the likelihood displacement (Cook, 1986), the Q-displacement (Zhu *et al.*, 2001) is defined by

$$\mathrm{QD}_i=2\{Q(\widehat{\eta}|\widehat{\eta})-Q(\widehat{\eta}^1_{[i]}|\widehat{\eta})\}. \tag{8}$$

If the value of CD_i or QD_i is large, then the ith observation is influential. Similarly, we can also quantify the effects of deleting two or more observations on $\hat{\eta}$ (Cook & Weisberg, 1982, chapter 3). For simplicity, we omit those details here.

The diagnostic measures CD_i and QD_i can be decomposed as sums of three diagnostic measures for assumptions (2)–(4) due to the decomposition in (5). The matrix $\partial^2_\eta Q(\eta|\widehat{\eta})$ can be written as

$$\mathrm{diag}(\partial^2_{(\beta,\tau)}Q_1(\beta,\tau|\widehat{\eta}),\partial^2_\alpha Q_2(\alpha|\widehat{\eta}),\partial^2_\xi Q_3(\xi|\widehat{\eta})).$$

Thus, (6) can be written as

$$\begin{aligned}&(\widehat{\beta}^1_{[i]}, \widehat{\tau}^1_{[i]})=(\widehat{\beta}, \widehat{\tau})+\{-\partial^2_{(\beta,\tau)} Q_1(\widehat{\beta}, \widehat{\tau}|\widehat{\eta})\}^{-1} \partial_{(\beta,\tau)} Q_{1[i]}(\widehat{\beta}, \widehat{\tau}|\widehat{\eta}),\\ &\widehat{\alpha}^1_{[i]}=\widehat{\alpha}+\{-\partial^2_{\alpha} Q_2(\widehat{\alpha}|\widehat{\eta})\}^{-1} \partial_{\alpha} Q_{2[i]}(\widehat{\alpha}|\widehat{\eta}),\\ &\widehat{\xi}^1_{[i]}=\widehat{\xi}+\{-\partial^2_{\xi} Q_3(\widehat{\xi}|\widehat{\eta})\}^{-1} \partial_{\xi} Q_{3[i]}(\widehat{\xi}|\widehat{\eta}).\end{aligned} \tag{9}$$

Finally, $\mathrm{CD}_i = \mathrm{CD}_{i,1} + \mathrm{CD}_{i,2} + \mathrm{CD}_{i,3}$, where

$$\begin{aligned}&\mathrm{CD}_{i,1}=\{\partial_{(\beta,\tau)} Q_{1[i]}(\widehat{\beta}, \widehat{\tau}|\widehat{\eta})\}'\{-\partial^2_{(\beta,\tau)} Q_1(\widehat{\beta}, \widehat{\tau}|\widehat{\eta})\}^{-1}\{\partial_{(\beta,\tau)} Q_{1[i]}(\widehat{\beta}, \widehat{\tau}|\widehat{\eta})\},\\ &\mathrm{CD}_{i,2}=\{\partial_{\alpha} Q_{2[i]}(\widehat{\alpha}|\widehat{\eta})\}'\{-\partial^2_{\alpha} Q_2(\widehat{\alpha}|\widehat{\eta})\}^{-1}\{\partial_{\alpha} Q_{2[i]}(\widehat{\alpha}|\widehat{\eta})\},\\ &\mathrm{CD}_{i,3}=\{\partial_{\xi} Q_{3[i]}(\widehat{\xi}|\widehat{\eta})\}'\{-\partial^2_{\xi} Q_3(\widehat{\xi}|\widehat{\eta})\}^{-1}\{\partial_{\xi} Q_{3[i]}(\widehat{\xi}|\widehat{\eta})\}.\end{aligned} \tag{10}$$

Intuitively, $\mathrm{CD}_{i,1}$ is mainly associated with the effects of removing the ith observation on assumption (2), $\mathrm{CD}_{i,2}$ is for assumption (3) and $\mathrm{CD}_{i,3}$ is for assumption (4). Similarly, it follows from (5) that

$$\mathrm{QD}_i=\mathrm{QD}_{i,1}+\mathrm{QD}_{i,2}+\mathrm{QD}_{i,3}, \tag{11}$$

where

$$\begin{aligned}&\mathrm{QD}_{i,1}=2[\,Q_1(\widehat{\beta}, \widehat{\tau}|\widehat{\eta}) - Q_1(\widehat{\beta}^1_{[i]}, \widehat{\tau}^1_{[i]}|\widehat{\eta})],\\ &\mathrm{QD}_{i,2}=2[\,Q_2(\widehat{\alpha}|\widehat{\eta}) - Q_2(\widehat{\alpha}^1_{[i]}|\widehat{\eta})]\ \text{ and } \mathrm{QD}_{i,3}=2[\,Q_3(\widehat{\xi}|\widehat{\eta}) - Q_3(\widehat{\xi}^1_{[i]}|\widehat{\eta})].\end{aligned}$$

Thus, $\mathrm{QD}_{i,1}$ is mainly associated with the effects of removing the ith observation on assumption (2), $\mathrm{QD}_{i,2}$ is for assumption (3) and $\mathrm{QD}_{i,3}$ is for assumption (4). Moreover, using a Taylor's series expansion, it can be shown that $\mathrm{QD}_{i,k}$ is asymptotically equivalent to $\mathrm{CD}_{i,k}$ for each of $k = 1, 2, 3$.

3.2. Conditional residuals

Residuals are key tools for revealing departures from assumptions (2)–(4). As our primary interest is to make valid inferences on assumption (2), we define the residual for the ith observation as

$$R_i(\widehat{\eta})=y_i - g((\mathbf{x}'_i, \mathbf{z}'_i)\widehat{\beta}).$$

However, as $\mathbf{z}_{m,i}$ is missing, $R_i(\hat{\eta})$ cannot be directly calculated for those cases with missing covariates. Generally, there are many ways of ‘eliminating’ $\mathbf{z}_{m,i}$. Here, we focus on two kinds of conditional residuals as follows:

$$\mathrm{CR}_i^{(1)}(\eta)=y_i - E[\,g((\mathbf{x}'_i, \mathbf{z}'_i)\beta)|\mathbf{x}_i, \mathbf{z}_{o,i}], \tag{12}$$

$$\mathrm{CR}_i^{(2)}(\eta)=y_i-E[g((\mathbf{x}_i',\mathbf{z}_i')\beta)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i,y_i], \tag{13}$$

for $i = 1, \ldots, n$, where the expectations in (12) and (13) are taken with respect to $p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \eta)$ and $p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta)$, respectively. If there are no missing covariates in $\mathbf{z}_i$, then $\mathrm{CR}_i^{(1)}(\widehat{\eta})$ and $\mathrm{CR}_i^{(2)}(\widehat{\eta})$ reduce to $R_i(\hat{\eta})$. Thus, the conditional residuals $\mathrm{CR}_i^{(k)}(\widehat{\eta})$ for $k = 1, 2$ can be regarded as generalizations of residuals in GLMs (Cook & Weisberg, 1982). The conditional residuals in (12) and (13) are computationally attractive because the conditional expectations involved can be easily evaluated using Markov chain Monte Carlo (MCMC) methods (Chen *et al.*, 2000;Liu, 2003). We note that $\mathrm{CR}_i^{(1)}(\widehat{\eta})$ does not account for the missing data mechanism.

We examine several properties of the proposed conditional residuals. Through a better understanding of the properties of conditional residuals, we may develop both formal and informal diagnostic tools for the examination of the adequacy of assumption (2). We derive the expectations and variances of the proposed conditional residuals in the following theorems, whose assumptions and detailed proofs can be found in the Appendix.

Proposition 1—Suppose that assumptions C3 and C2 in the Appendix are true. We then have the following results:

i. $E[\mathrm{CR}_i^{(k)}(\eta_*)|\mathbf{x}_i]=E[\mathrm{CR}_i^{(k)}(\eta_*)]=0$ for k = 1, 2, where η_* is the true value of η. However, $E[\mathrm{CR}_i^{(k)}(\eta_*)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i,\eta_*]$ may not equal zero for k = 1, 2.

ii. If the missing data are MAR, then $\mathrm{CR}_i^{(2)}(\eta)=y_i-E[g((\mathbf{x}_i',\mathbf{z}_i')\beta)|\mathbf{x}_i,\mathbf{z}_{o,i},y_i]$ and

$$E\left[\frac{\mathrm{CR}_i^{(1)}(\eta_*)}{p(\mathbf{r}_i|\mathbf{x}_i,\mathbf{z}_{o,i},y_i,\xi_*)}|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i\right]=0. \tag{14}$$

iii. If $p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \xi) = p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, \xi)$, then

$$E[\mathrm{CR}_i^{(2)}(\eta_*)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i]=0,\ \ but\ \ E[\mathrm{CR}_i^{(1)}(\eta_*)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i]\neq 0.$$

iv.

$$\mathrm{CR}_i^{(1)}(\widehat{\eta})=CR_i^{(1)}(\eta_*)-[\Delta_{i1}^{(1)\prime}(\widehat{\beta}-\beta_*)+\Delta_{i2}^{(1)\prime}(\widehat{\alpha}-\alpha_*)][1+o_p(1)],$$

where

$$\Delta_{i1}^{(1)}=E[\partial_\beta g((\mathbf{x}_i',\mathbf{z}_i')\beta_*)|\mathbf{x}_i,\mathbf{z}_{o,i},\alpha_*]$$

and

$$\Delta_{i2}^{(1)}=E[g((\mathbf{x}_i',\mathbf{z}_i')\beta_*)\{\partial_\eta\log p(\mathbf{z}_{m,i}|\mathbf{x}_i,\mathbf{z}_{o,i},\alpha_*)\}'|\mathbf{x}_i,\mathbf{z}_{o,i},\alpha_*].$$

v. $\mathrm{CR}_i^{(2)}(\widehat{\eta})=\mathrm{CR}_i^{(2)}(\eta_*)-[\Delta_{i1}^{(2)\prime}(\widehat{\beta}-\beta_*)+\Delta_{i2}^{(2)\prime}(\widehat{\eta}-\eta_*)][1+o_p(1)],$

where

$$\Delta_{i1}^{(2)}=E[\,\partial_\beta g((\mathbf{x}_i',\mathbf{z}_i')\beta_*)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i,y_i,\eta_*]$$

and

$$\Delta_{i2}^{(2)}=E[\,g((\mathbf{x}_i',\mathbf{z}_i')\beta_*)\{\partial_\eta \log p(\mathbf{z}_{m,i}|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i,y_i,\eta_*)\}'|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i,y_i,\eta_*].$$

Proposition 1(i) shows that $E[\,\mathrm{CR}_i^{(k)}(\eta_*)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i]$ for $k = 1, 2$ are biased, whereas $E[\,\mathrm{CR}_i^{(k)}(\eta_*)|\mathbf{x}_i]$ and $E[\,\mathrm{CR}_i^{(k)}(\eta_*)]$ are unbiased. Proposition 1(ii) shows that the missing data indicators can be dropped from $\mathrm{CR}_i^{(2)}(\eta)$ under MAR covariates. The inverse weighted residuals are unbiased only for $\mathrm{CR}_i^{(1)}(\eta)$. Proposition 1(iii) shows that $E[\,\mathrm{CR}_i^{(2)}(\eta_*)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i]$ is unbiased when the missing mechanism is independent of y_i. Proposition 1(iv) and (v) give the first-order expansions of $\mathrm{CR}_i^{(1)}(\eta)$ and $\mathrm{CR}_i^{(2)}(\eta)$ respectively. In particular, the terms involving $\Delta_{i2}^{(1)}$ and $\Delta_{i2}^{(2)}$ are due to the presence of the missing data. The matrices $\Delta_{i1}^{(k)}$ and $\Delta_{i2}^{(k)}$ for $k = 1, 2$, can be calculated using MCMC methods (Chen *et al.*, 2000). For instance,

$$\begin{aligned}&\partial_\eta\{\log p(\mathbf{z}_{m,i}|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i,y_i,\eta_*)\}=\partial_\eta\{\log p(\mathbf{d}_i|\eta_*)\}\\&-E[\,\partial_\eta\{\log p(\mathbf{d}_i|\eta_*)\}|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i,y_i,\eta_*].\end{aligned} \tag{15}$$

Thus,

$$\begin{aligned}\Delta_{i2}^{(2)}=&E[\,g((\mathbf{x}_i',\mathbf{z}_i')\beta_*)\partial_\eta\{\log p(\mathbf{d}_i|\eta_*)\}|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i,y_i,\eta_*]\\&-E[\,g((\mathbf{x}_i',\mathbf{z}_i')\beta_*)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i,y_i,\eta_*]E[\,\partial_\eta\{\log p(\mathbf{d}_i|\eta_*)\}|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i,y_i,\eta_*].\end{aligned}$$

We can use MCMC methods to generate random samples from $p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \hat{\eta})$ and construct a consistent estimate for $\Delta_{i2}^{(2)}$.

The values of the standardized $\mathrm{CR}_i^{(k)}(\widehat{\eta})$ may be used to detect anomalous or influential observations (Cook & Weisberg, 1982). We define a standardized conditional residual as follows:

$$\mathrm{SCR}_i^{(k)}(\widehat{\eta})=\mathrm{CR}_i^{(k)}(\widehat{\eta})/\sigma_{i,k}(\widehat{\eta}), \tag{16}$$

where

$$\sigma_{i;1}(\eta)^2=\mathrm{var}[\,\mathrm{CR}_i^{(1)}(\eta)|\mathbf{x}_i,\mathbf{z}_{o,i}]\ \text{ and } \sigma_{i;2}(\eta)^2=\mathrm{var}[\,\mathrm{CR}_i^{(2)}(\eta)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i].$$

When model (1) is correctly specified, $\mathrm{SCR}_i^{(k)}(\widehat{\eta})$ and $\mathrm{CR}_i^{(k)}(\widehat{\eta})$ should oscillate around 0. We consider the *i*th observation as an ‘outlier’ if $|\mathrm{SCR}_i^{(k)}(\widehat{\eta})|$ is significantly greater than some threshold, such as 3. Moreover, if many $|\mathrm{SCR}_i^{(k)}(\widehat{\eta})|$ values are significantly greater than zero, then one should question whether assumption (2) is correct. It is also worthwhile to inspect $\mathrm{SCR}_i^{(k)}(\widehat{\eta})$ against some function of the data, such as the observed responses and a specific covariate, which may provide an assessment of the adequacy of assumption (2).

3.3. Goodness-of-fit test without incorporating missing data

There is an extensive literature on developing test statistics to check the correct specification of the conditional mean (17) for GLMs with no missing data (Su & Wei, 1991;Stute, 1997;Lin *et al.*, 2002;Stute & Zhu, 2002). However, to the best of our knowledge, no goodness-of-fit test statistics have ever been developed for GLMs with missing covariate data.

We may use the two types of conditional residuals proposed in the previous subsection to develop test statistics to formally check model assumptions in a GLM with missing covariates. However, for simplicity, we temporarily drop the superscript (k) in $\mathrm{CR}_i^{(k)}(\widehat{\eta})$, because the results below hold for both types of conditional residuals. These test statistics are originally designed to test the following null and alternative hypotheses:

$$\begin{aligned} H_0^{(0)}&:E[y|\mathbf{x},\mathbf{z}]=g((\mathbf{x}',\mathbf{z}')\beta) \quad \text{for some } \eta,\\ H_1^{(0)}&:E[y|\mathbf{x},\mathbf{z}]-g((\mathbf{x}',\mathbf{z}')\beta)\neq 0 \ \text{ for all } \eta. \end{aligned} \tag{17}$$

However, because some components of **z** are missing, we may wish to test the equality

$$h(\eta|\mathbf{x})=E[y|\mathbf{x}]-E[g((\mathbf{x}',\mathbf{z}')\beta)|\mathbf{x}]=E\{\mathrm{CR}(\eta)|\mathbf{x}\}=0.$$

Thus, instead of testing $H_0^{(0)}$ against $H_1^{(0)}$, we test the following null and alternative hypotheses:

$$\begin{aligned} H_0^{(1)}&:h(\eta|\mathbf{x})=0 \ \text{ for some } \eta\\ H_1^{(1)}&:h(\eta|\mathbf{x})\neq 0 \ \text{ for all } \eta. \end{aligned} \tag{18}$$

Note that $h(\eta|\mathbf{x}) = 0$ is only a necessary condition of $E[y|\mathbf{x}, \mathbf{z}] = g((\mathbf{x}', \mathbf{z}')\beta)$. Thus, accepting $h(\eta|\mathbf{x}) = 0$ does not imply the acceptance of $H_0^{(0)}$.

We can construct statistics for testing $H_0^{(1)}$ as follows. Following theorem 1 in Bierens (1992), $E\{\mathrm{CR}(\eta)|\mathbf{x}\} = 0$ is equivalent to $E\{\mathrm{CR}(\eta)|\mathbf{x}'\boldsymbol{\phi}\} = 0$ for any $\boldsymbol{\phi}\in R^{p1}$. Thus, as shown in lemma 1 of Escanciano (2006), $H_0^{(1)}$ is equivalent to

$$E\{\mathrm{CR}(\eta)\mathbf{1}(\mathbf{x}'\phi\leq t)\}=0 \tag{19}$$

for almost every ($\boldsymbol{\phi}$, t). To test $H_0^{(1)}$, we may define a stochastic process as follows:

$$I_1((\phi,t);\eta)=n^{-1/2}\sum_{i=1}^{n}\mathbf{1}(\mathbf{x}_i'\phi\leq t)\mathrm{CR}_i(\eta),$$

where

$$(\phi, t) \in \Pi = \{\phi \in R^{p_1}: \|\phi\| = 1\} \times [-\infty, \infty],$$

in which ‖·‖ is the common L_2-norm in Euclidean space. Graphically, for a specific direction $\boldsymbol{\phi}$, we can plot $I_1((\boldsymbol{\phi}, t); \eta)$ as a function of t and use it as an exploratory tool for the detection of model misspecification along the direction $\boldsymbol{\phi}$ (Lin *et al.*, 2002). For instance, we may set $\boldsymbol{\phi} = (\hat{\beta}_1, \ldots, \hat{\beta}_{p1})'$.

Theoretically, we regard $I_1((\boldsymbol{\phi}, t); \eta)$ as a stochastic process indexed by $(\boldsymbol{\phi}, t)$ and then we use $I_1((\boldsymbol{\phi}, t); \eta)$ to construct two test statistics. We first define a conditional Kolmogorov test (CK) as

$$\mathrm{CK}_1 = \max_{(\phi, t)} |I_1((\phi, t); \eta)|. \tag{20}$$

We also define a Cramer–von Mises test as follows:

$$\mathrm{CM}_1 = \int_\Pi |I_1((\phi, t); \widehat{\eta})|^2 F_{n,\phi}(\mathrm{d}t)\mathrm{d}\phi, \tag{21}$$

where $F_{n,\boldsymbol{\phi}}(u)$ is the empirical distribution function of $\{\mathbf{x}_i'\phi : i = 1, \ldots, n\}$ (Stute, 1997). Large values of CM_1 and CK_1 lead to rejection of $H_0^{(1)}$.

We note that CM_1 has several distinctive features (Escanciano, 2006). The statistic CM_1 has a closed form (Escanciano, 2006, appendix B), whereas computing the Kolmogorov-type supremum statistic of residual process involves high-dimensional maximizations. Particularly, when the dimension of the covariate vector is high or even moderate, it can be computationally demanding to compute the Kolmogorov supremum statistic. Thus, CM_1 avoids the problem of the curse of dimensionality. We are now led to theorem 1.

Theorem 1—Suppose that assumptions C1–C7 in the Appendix are true. Under the null hypothesis $H_0^{(1)}$, we then have the following results:

i.

$$\sqrt{n}(\widehat{\eta} - \eta_*) = n^{-1/2} \sum_{i=1}^{n} \psi_{n,i} + o_p(1),$$

with $\psi_{n,i} = M_n(\eta_*)^{-1} \dot{\ell}_i(\eta_*)$, where

$$\dot{\ell}_i(\eta_*) = \partial_\eta \log p(\mathbf{d}_{o,i}|\eta_*) \ \ and \ \ M_n(\eta_*) = n^{-1} \sum_{i=1}^{n} E[\partial_\eta^2 \log p(\mathbf{d}_{o,i}|\eta_*)].$$

ii. $(I_1(\cdot;\eta_*), \sqrt{n}(\widehat{\eta} - \eta_*)')'$ converges in distribution to $(G_1(\cdot), v_1')'$, where $(G_1(\cdot), v_1')$ is a mean zero Gaussian process with covariance function

$$C_1((\phi_1,t_1),(\phi_2,t_2))=\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\begin{pmatrix}\mathrm{CR}_i(\eta_*)\mathbf{1}(\mathbf{x}_i'\phi_1\le t_1)\\ \psi_{n,i}(\eta_*)\end{pmatrix}\times\begin{pmatrix}\mathrm{CR}_i(\eta_*)\mathbf{1}(\mathbf{x}_i'\phi_2\le t_2)\\ \psi_{n,i}(\eta_*)\end{pmatrix}'.$$

iii. CK_1 and CM_1 converge in distribution to

$$\sup_{(\phi,t)}|G_1(\phi,t)+\Delta_1(\phi,t)'v_1| \ \mathit{and}\ \int_{\Pi}|G_1(\phi,t)+\Delta_1(\phi,t)'v_1|^2F_\phi(\mathrm{d}t)\mathrm{d}\phi,$$

respectively, where $F_{\boldsymbol{\phi}}(t)$ is the limiting cumulative distribution function of $F_{n,\boldsymbol{\phi}}(t)$ and $\Delta_1(\boldsymbol{\phi}, t)$ is defined by

$$\Delta_1(\phi,t)=\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}E\{\mathbf{1}(\mathbf{x}_i'\phi\le t)\partial_\eta[\mathrm{CR}_i(\eta_*)]\}.$$

Theorem 1 formally characterizes the asymptotic null distributions of CK_1 and CM_1. Therefore, we may directly approximate those distributions in order to calculate the *p*-values of the test statistics CK_1 and CM_1.

The next result establishes the asymptotic distributions of CK_1 and CM_1 under a sequence of local alternatives converging to the null at a parametric rate $n^{-1/2}$. We consider the local alternatives such that $p(y_i|\mathbf{x}_i, \mathbf{z}_i)$ belongs to the exponential family (2) and

$$E[y_i|\mathbf{x}_i,\mathbf{z}_i]=g((\mathbf{x}_i',\mathbf{z}_i')\beta_*)+n^{-1/2}g_0(\mathbf{x}_i,\mathbf{z}_i) \tag{22}$$

for $i = 1, \ldots, n$, where $g_0(\mathbf{x}_i, \mathbf{z}_i)$ is a function of $(\mathbf{x}_i, \mathbf{z}_i)$. Let

$$\theta_i(t)=\dot{b}^{-1}(g((\mathbf{x}_i',\mathbf{z}_i')\beta_*)+tg_0(\mathbf{x}_i,\mathbf{z}_i)),$$

where $\dot{b}$ denotes $\partial_t b(t)$ and $\dot{b}^{-1}(\cdot)$ is the inverse function of $\dot{b}(\cdot)$. Then, we have

$$\theta_i=\theta_i(n^{-1/2})=\dot{b}^{-1}(g((\mathbf{x}_i',\mathbf{z}_i')\beta_*)+n^{-1/2}g_0(\mathbf{x}_i,\mathbf{z}_i)).$$

Thus, the true distribution of y_i given $(\mathbf{x}_i, \mathbf{z}_i)$, denoted by $p(y_i|\mathbf{x}_i, \mathbf{z}_i, n^{-1/2})$, is

$$\exp\{a_i^{-1}(\tau_*)[y_i\theta_i(n^{-1/2})-b(\theta_i(n^{-1/2}))]+c(y_i,\tau_*)\}. \tag{23}$$

Moreover, $p(\mathbf{x}_i, \mathbf{z}_i|\alpha_*)$ and $p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \xi_*)$ are unchanged. We are now led to the following results.

Theorem 2—Suppose that assumptions C1–C7 in the Appendix and the sequence of models in (23) are true. We then have the following results:

i. $\sqrt{n}(\widehat{\eta} - \eta_*)$ converges in distribution to $\nu_1 + A_1$, where ν_1 is the same normal distribution as in theorem 1 and

$$A_1 = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\psi_{n,i}E[a_i(\tau_*)^{-1}\partial_i\theta_i(0)(y_i - g((\mathbf{x}_i', \mathbf{z}_i')\beta_*))|\mathbf{d}_{o,i}].$$

ii. $I_1(\cdot; \eta_*)$ converges in distribution to $G_1(\cdot) + A_2(\cdot)$, where $G_1(\cdot)$ is the same process as in theorem 1. In addition,

$$A_2(\phi, t) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\mathbf{1}(\phi'\mathbf{x}_i \le t)E[g_0(\mathbf{x}_i, \mathbf{z}_i)|\mathbf{x}_i, \mathbf{z}_{o,i}]$$

for $\mathrm{CR}_i^{(1)}$, whereas

$$A_2(\phi, t) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\mathbf{1}(\phi'\mathbf{x}_i \le t)E[g_0(\mathbf{x}_i, \mathbf{z}_i)|\mathbf{d}_{o,i}]$$

for $\mathrm{CR}_i^{(2)}$.

iii. CK_1 and CM_1 converge in distribution to

$$\sup_{(\phi,t)}|G_1(\phi, t) + A_2(\phi, t) + \Delta_1(\phi, t)'(\nu_1 + A_1)|$$

and

$$\int_{\Pi}|G_1(\phi, t) + A_2(\phi, t) + \Delta_1(\phi, t)'(\nu_1 + A_1)|^2 F_\phi(\mathrm{d}t)\mathrm{d}\phi$$

respectively.

3.4. Goodness-of-fit test incorporating missing data

We propose to use the missing covariates $\mathbf{z}_i$ to improve the power of $I_1((\boldsymbol{\phi}, t); \eta)$ in detecting the misspecification of $g((\mathbf{x}', \mathbf{z}')\beta)$. Recall that $h(\eta|\mathbf{x}) = 0$ is only a necessary condition of $E[y|\mathbf{x}, \mathbf{z}] = g((\mathbf{x}', \mathbf{z}')\beta)$. Because $\mathbf{1}(\mathbf{x}'\boldsymbol{\phi} \le t)$ in $I_1((\boldsymbol{\phi}, t); \eta)$ does not involve the missing covariates $\mathbf{z}$, we may lose power in detecting the misspecification of $H_0^{(0)}$ in the missing covariate space. In particular, if the fraction of missing covariates is small, then it is very inefficient to drop all the information in $\mathbf{z}$.

We may test whether $H_0^{(0)}$ is true using the additional information contained in the missing covariates. Letting $\mathbf{z}_{m,i}(\alpha) = E[\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \alpha]$, we suggest replacing $\mathbf{z}_{m,i}$ by $\mathbf{z}_{m,i}(\hat{\alpha})$, which is an imputed missing covariate vector. However, developing test statistics based on the imputed missing covariates depends on the specific missing data mechanism.

We first consider the case that $p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i)$ is independent of y_i. Using proposition 1(iii), we can show that

$$E[\mathrm{CR}_i^{(2)}(\eta_*)\mathbf{1}(\mathbf{c}'_{i*}\tilde{\phi}\le t)|\mathbf{x}_i,\mathbf{z}_{o,i}]=E\{E[\mathrm{CR}_i^{(2)}(\eta_*)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i]\mathbf{1}(\mathbf{c}'_{i*}\tilde{\phi}\le t)|\mathbf{x}_i,\mathbf{z}_{o,i}\}=0, \tag{24}$$

for all $i = 1, \ldots, n$, where

$$(\tilde{\phi},t)\in\Pi=\{\tilde{\phi}\in R^{p_1+p_2}:\tilde{\phi}'\tilde{\phi}=1\}\times[-\infty,\infty],$$

and $\mathbf{c}_{i*} = \mathbf{c}_i(\alpha_*) = (\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{z}_{m,i}(\alpha_*))$. In addition, $\mathbf{c}_i(\alpha)$ is defined as

$$\mathbf{c}_i(\alpha)=(\mathbf{x}_i,r_{i1}z_{i1}+(1-r_{i1})E[z_{i1}|\mathbf{x}_i,\mathbf{z}_{o,i},\alpha],\ldots,r_{ip_2}z_{ip_2}+(1-r_{ip_2})E[z_{ip_2}|\mathbf{x}_i,\mathbf{z}_{o,i},\alpha]).$$

Let $\hat{\mathbf{c}}_i = \mathbf{c}_i(\hat{\alpha})$. We are thus able to incorporate the additional information from $\mathbf{z}_{o,i}$ into the indicator function $\mathbf{1}(\hat{\mathbf{c}}'_i\tilde{\phi}\le t)$. Following the reasoning in (24), we now propose the stochastic process:

$$I_2((\tilde{\phi},t);\eta)=n^{-1/2}\sum_{i=1}^{n}\mathbf{1}(\hat{\mathbf{c}}'_i\tilde{\phi}\le t)\mathrm{CR}_i^{(2)}(\eta). \tag{25}$$

We first suggest plotting $I_2((\tilde{\boldsymbol{\phi}}, t)$ against t for a specific $\tilde{\boldsymbol{\phi}}$ as an exploratory tool for detecting the form of misspecification of assumption (2). For instance, we may set $\tilde{\boldsymbol{\phi}}=\hat{\beta}$. Then, we develop the corresponding CK and CM statistics based on $I_2((\tilde{\boldsymbol{\phi}}, t); \hat{\eta})$, denoted by CK_2 and CM_2. Large values of CK_2 and CM_2 lead to rejection of the hypothesis that $E[\mathrm{CR}_i^{(2)}(\eta)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i]=0.$

Secondly, suppose that the missing data are MAR. Using proposition 1(ii), we can show that for $i = 1, \ldots, n$,

$$\begin{aligned}&E\left[\frac{\mathrm{CR}_i^{(1)}(\eta_*)}{p(\mathbf{r}_i|\mathbf{x}_i,\mathbf{z}_{o,i},y_i,\xi_*)}\mathbf{1}(\mathbf{c}'_{i*}\tilde{\phi}\le t)|\mathbf{x}_i,\mathbf{z}_{o,i}\right]\\&=E\left\{E\left[\frac{\mathrm{CR}_i^{(1)}(\eta_*)}{p(\mathbf{r}_i|\mathbf{x}_i,\mathbf{z}_{o,i},y_i,\xi_*)}|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i\right]\mathbf{1}(\mathbf{c}'_{i*}\tilde{\phi}\le t)|\mathbf{x}_i,\mathbf{z}_{o,i}\right\}=0.\end{aligned}$$

Then, we propose an inverse weighted process as follows:

$$I_3((\tilde{\phi},t);\eta)=n^{-1/2}\sum_{i=1}^{n}\mathbf{1}(\hat{\mathbf{c}}'_i\tilde{\phi}\le t)\frac{\mathrm{CR}_i^{(1)}(\eta)}{p(\mathbf{r}_i|\mathbf{x}_i,\mathbf{z}_{o,i},y_i,\xi)}. \tag{26}$$

We may plot $I_3((\tilde{\boldsymbol{\phi}}, t)$ against t for a specific $\tilde{\boldsymbol{\phi}}$ as an exploratory tool for detecting the assumption of MAR. Similar to (21) and (20), we can develop the corresponding CK and CM statistics based on $I_3((\tilde{\boldsymbol{\phi}}, t); \hat{\eta})$ and denote them by CK_3 and CM_3. Large values of CK_3 and CM_3 lead to rejection of the hypothesis that

$$E\left[\frac{\mathrm{CR}_i^{(1)}(\eta_*)}{p(\mathbf{r}_i|\mathbf{x}_i,\mathbf{z}_{o,i},y_i,\xi_*)}|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i\right]=0.$$

Similar to theorems 1 and 2, we can establish the asymptotic distributions of CK_k and CM_k and their power behaviour under local alternatives for $k = 2, 3$. For simplicity, we only include the asymptotic null distributions of $I_2((\tilde{\boldsymbol{\phi}}, t); \eta_*)$ below.

Theorem 3—Suppose that assumptions C1–C8 in the Appendix are true. Under the null hypothesis $H_0^{(0)}$, $\mathrm{I}_2(\cdot; \eta_*)$ converges in distribution to $\mathrm{G}_2(\cdot)$, where $\mathrm{G}_2(\cdot)$ is a mean zero Gaussian process with covariance function

$$C_2((\tilde{\phi}_1,t_1),(\tilde{\phi}_2,t_2))=\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}[\mathrm{CR}_i^{(2)}(\eta_*)]^2\mathbf{1}(\mathbf{c}_{i*}'\tilde{\phi}_1\le t_1)\mathbf{1}(\mathbf{c}_{i*}'\tilde{\phi}_2\le t_2).$$

The CK_k and CM_k for $k = 2, 3$ differ from CK_1 and CM_1 in several aspects. The CK_1 and CM_1 focus on testing $H_0^{(1)}$ regardless of the missing data mechanism and the type of conditional residual, whereas large values of CK_k and CM_k for $k = 2, 3$ can be caused by the misspecification of the missing data mechanism. CK_1 and CM_1 can be used in either MAR or NMAR settings and regardless of the types of conditional residual, and are mainly used to examine the validity regarding the assumptions of the sampling distribution of $(y|\mathbf{x})$. CK_k and CM_k for $k = 2, 3$ are most useful in examining issues related to the sampling distribution of $(y|\mathbf{x}, \mathbf{z})$ and the missing data mechanism. Specifically, CM_2 (or CK_2) addresses the form of the missing data mechanism, and in particular, whether the missing data mechanism depends on the response variable. CM_3 (or CK_3) addresses the more general issue of whether the missing data mechanism is NMAR. For instance, CK_2 and CM_2 test whether $E[\mathrm{CR}_i^{(2)}(\eta_*)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i]$ equals zero or not, whereas CK_3 and CM_3 test whether

$$E\left[\frac{\mathrm{CR}_i^{(1)}(\eta_*)}{p(\mathbf{r}_i|\mathbf{x}_i,\mathbf{z}_i,y_i,\xi_*)}|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i\right]$$

equals zero or not. The rejection of $E[\mathrm{CR}_i^{(2)}(\eta_*)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i]=0$ may be caused by the dependence of $p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \xi)$ on the response y_i, while the rejection of

$$E\left[\frac{\mathrm{CR}_i^{(1)}(\eta_*)}{p(\mathbf{r}_i|\mathbf{x}_i,\mathbf{z}_i,y_i,\xi_*)}|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i\right]=0$$

can be caused by NMAR covariate data. Thus, CK_k and CM_k, $k = 2, 3$ are useful goodness-of-fit statistics for testing the missing data mechanism.

3.5. Resampling method

In the following, we devise a resampling method to approximate the *p*-value of CK_1. We can develop similar methods for CK_k, CM_j, $k = 2, 3$, $j = 1, 2, 3$. There are four steps in generating the stochastic processes that have the same asymptotic distributions as $I_1((\boldsymbol{\phi}, t); \hat{\eta})$.

Step 1. Generate independent and identically distributed (i.i.d.) random samples, $\{v_i^{(q)}:i=1,\ldots,n\}$, from an $N(0, 1)$ distribution for $q = 1, \ldots, Q$, where Q is the number of replications, say $Q = 1000$.

Step 2. Calculate

$$\mathrm{I}_1((\phi,t);\widehat{\eta})^{(q)}=n^{1/2}\sum_{i=1}^{n}v_i^{(q)}\{\mathrm{CR}_i(\widehat{\eta})\mathbf{1}(\mathbf{x}_i'\phi\le t)-\widehat{\Delta}_1(\phi,t)\psi_{ni}\}$$

where

$$\widehat{\Delta}_1(\phi,t)=n^{-1}\sum_{i=1}^{n}\partial_\eta\mathrm{CR}_i(\widehat{\eta})\mathbf{1}(\mathbf{x}_i'\phi\le t).$$

Note that conditional on the observed data, as $I_1((\boldsymbol{\phi}, t); \hat{\eta})^{(q)}$ is the sum of independent but not identically distributed stochastic process, it follows from some mild conditions that $I_1((\boldsymbol{\phi}, t); \hat{\eta})^{(q)}$ converges weakly to the desired Gaussian process in theorem 1 as $n \to \infty$ (Kosorok, 2003; van der Vaart & Wellner, 1996; Stute *et al*., 1998).

Step 3. Calculate the test statistics

$$\mathrm{CK}_1^{(q)}=\sup_{(\phi,t)}|I_1((\phi,t);\widehat{\eta})^{(q)}|$$

and obtain $\{\mathrm{CK}_1^{(q)}:q=1,\ldots,Q\}$.

Step 4. Calculate the *p*-value of CK_1 using $\{\mathrm{CK}_1^{(q)}:q=1,\ldots,Q\}$.

4. Simulation studies

We conducted Monte Carlo simulations to examine the finite-sample performance of the various diagnostic measures proposed here. First, we applied case-deletion measures and standardized conditional residuals to a simulated data set based on a linear model, in which an outlier was added. We expected that the diagnostic measures would detect the outlier. Secondly, we evaluated the rejection rates of the type I and type II errors for CM_1 based on the conditional residuals $\mathrm{CR}_i^{(1)}$ and $\mathrm{CR}_i^{(2)}$, for CM_2 and for CM_3 respectively. For the sake of simplicity, we omitted the results based on CK_k for $k = 1, 2, 3$ to save space, as they have similar type I and type II errors as CM_k. Furthermore, we evaluated the rejection rates for CM_1 based on the conditional residuals $\mathrm{CR}_i^{(1)}$ and $\mathrm{CR}_i^{(2)}$, and did the same thing for CM_2, for a logistic regression simulation.

4.1. Case-deletion measures and conditional residuals for the linear model

We considered the linear model

$$y_i=\beta_0+\beta_1 x_i+\beta_2 z_i+\varepsilon_i, \tag{27}$$

where the ε_is are i.i.d. and $\varepsilon_i \sim N(0, \tau)$, $i = 1, \ldots, n$. We assume that y_i and x_i are completely observed for $i = 1, \ldots, n$, but the covariate z_i may be missing for some cases. We set $n = 100$, $\beta_0 = \beta_1 = \beta_2 = 1$ and $\tau = 1$. Moreover, we independently generated 100 random vectors (x_i, z_i) from an $N_2(\mathbf{0}, \mathbf{I}_2)$ distribution. We also assumed the covariates are MAR,

$$p(r_i=1|x_i, z_i, y_i)=\frac{\exp(\xi_0+\xi_1 x_i)}{1+\exp(\xi_0+\xi_1 x_i)}, \tag{28}$$

with $\xi_0 = -1.5$ and $\xi_1 = 1.0$ to obtain an average missingness fraction of 20%.

We first changed the last response y_{100} to $y_{100} + 5.0$ in order to add an outlier to the data set. We fit the linear model assuming an MAR z_i for the simulated data and a normal distribution for z_i. Then we calculated case-deletion measures and conditional residuals for each observation. The last observation was classified as the most influential observation by CD_i and $\mathrm{CD}_{i,1}$, but not $\mathrm{CD}_{i,2}$, because we only changed y_{100} in the response space (Fig. 1A–C). Specifically, $\mathrm{CD}_{100} = 4.378$ is much larger than the second largest $\mathrm{CD}_i = 0.33$ (Fig. 1A). Moreover, we obtained similar findings based on QD_i, $\mathrm{QD}_{i,1}$ and $\mathrm{QD}_{i,2}$ (not presented here). The standardized conditional residuals with $\mathrm{SCR}_{100}^{(1)}=3.256$ also identified y_{100} as an influential observation (Fig. 1D).

Now, instead of changing the last response y_{100}, we changed z_{100} to $z_{100} + 5.0$ to add an outlier in the covariate space, and fit the same linear model assuming an MAR z_i. The last observation was classified as the most influential observation by CD_i, $\mathrm{CD}_{i,1}$ and $\mathrm{CD}_{i,2}$ (Fig. 1E–G). In contrast to the previous case in which y_{100} was changed, both $\mathrm{CD}_{i,1}$ and $\mathrm{CD}_{i,2}$ detected the influential observation z_{100} (Fig. 1F and G), because changing z_{100} affected the first two components of (1). The standardized conditional residuals $\mathrm{SCR}_i^{(k)}$ for $k = 1, 2$ identified the last observation as influential (Fig. 1H).

4.2. Goodness-of-fit statistics for the linear model

We systematically assessed the goodness-of-fit statistics based on the conditional residuals developed in section 3 under various scenarios. We used 500 replications to calculate the p-values of all test statistics. The significance level was always fixed at 0.05.

We considered three groups of simulation studies. The first group of simulation studies was to compare the finite-sample performance of CM_1 using either $\mathrm{CR}_i^{(1)}$ or $\mathrm{CR}_i^{(2)}$ under two scenarios. In the first scenario, we simulated 500 data sets from

$$y_i=\beta_0+\beta_1 x_i+\beta_2 z_i+c x_i^2+\varepsilon_i \quad \text{for } i=1, \ldots, 100,$$

where (x_i, z_i) were generated from an $N_2(\mathbf{0}, \mathbf{I}_2)$ distribution, the ε_is are i.i.d. and $\varepsilon_i \sim N(0, \tau)$, $i = 1, \ldots, n$, and c is in the range [0, 1]. We set $\beta_0 = \beta_1 = \beta_2 = 1$. We assumed that the covariate z_i has a normal distribution. We considered two missing data mechanisms: MCAR and MAR. Under MAR, the missing data mechanism was given by (28), in which we set $\xi_1 = 1.0$ and ξ_0 with values -1.5, -0.5 and 0.5 to obtain average missing data fractions of 20%, 40% and 60%

respectively. Then we fit $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$ under MAR, and thus the fitted model would be misspecified if $c \neq 0$ and the misspecification is due to the fully observed covariate x_i.

The top half of Table 2 shows the rejection rates of CM_1 based on both $\mathrm{CR}_i^{(1)}$ and $\mathrm{CR}_i^{(2)}$ for this scenario. The type I error rates are accurate across all missingness fractions. The CM_1 based on $\mathrm{CR}_i^{(2)}$ is uniformly more powerful than that based on $\mathrm{CR}_i^{(1)}$. Consistent with our expectations, the power for detecting misspecification of the model increases with $|c|$ for CM_1. The missing data fraction slightly influences the power of detecting model misspecification for CM_1.

In the second scenario, we generated 500 data sets from $y_i=\beta_0+\beta_1 x_i+\beta_2 z_i+cz_i^2+\varepsilon_i$, whereas the rest of the set-up remained the same as in the first scenario described earlier. We fit $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$ assuming MAR, and thus the model would be misspecified if $c \neq 0$ and the misspecification is due to the missing covariate z_i. The rejection rates are shown in the second half of Table 2. We found that CM_1 based on both $\mathrm{CR}_i^{(1)}$ and $\mathrm{CR}_i^{(2)}$ cannot detect the misspecification of cz_i^2, because CM_1 did not incorporate the missing covariate z_i. Comparing the top half with the bottom half in Table 2 reveals the importance of incorporating the misspecified covariate in the indicator function $\mathbf{1}(\mathbf{x}_i'\phi \leq t)$.

The second group of simulation studies was to assess the finite-sample performance of CM_2. First, we evaluated the power of CM_2 in detecting the misspecification of $E[y_i \mid x_i, z_i]$. We used the same two scenarios as in the first group of simulations, and in each case, we fit the linear model $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$ assuming z_i is MAR.

The first half of Table 3 shows the results for the first scenario where the misspecification is due to x_i, and the second half of the table is for the second scenario where the misspecification is due to z_i. The type I errors rates of CM_2 are accurate across all missingness fractions. For both scenarios, the power for detecting misspecification of the model increased with $|c|$ for CM_2 and the missing data fraction influences the power in detecting model misspecification (i.e. $|c| \neq 0$). Compared with Table 2, when $y_i=\beta_0+\beta_1 x_i+\beta_2 z_i+cx_i^2+\varepsilon_i$ is the true model (i.e. the first scenario), CM_1 based on $\mathrm{CR}_i^{(2)}$ is slightly more powerful than CM_2 in detecting the presence of cx_i^2. However, if $y_i=\beta_0+\beta_1 x_i+\beta_2 z_i+cz_i^2+\varepsilon_i$ is the true model (i.e. the second scenario), then CM_2 is much more powerful than CM_1 based on $\mathrm{CR}_i^{(2)}$. This indicates that incorporating the missing data can increase the power of detecting model misspecification due to cz_i^2.

We checked the influence of the misspecified parametric assumptions for the covariate distribution on the finite-sample performance of CM_2. Again, we used the same two settings as before except for one change: z_i was generated from $U[-3, 3]$, a uniform distribution, instead of an $N(0, 1)$ distribution. We fit the linear model $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$ and assumed that z_i is MAR and has a normal distribution. The first half of Table 4 shows the results for the first scenario where the misspecification is due to x_i, and the second half of the table is for the second scenario where the misspecification is due to z_i. Compared with Table 3, when $y_i=\beta_0+\beta_1 x_i+\beta_2 z_i+cx_i^2+\varepsilon_i$ is the true model (i.e. the first scenario), the misspecified covariate distribution for z_i has little effect on the statistical power of detecting the presence of cx_i^2. However, if $y_i=\beta_0+\beta_1 x_i+\beta_2 z_i+cz_i^2+\varepsilon_i$ is the true model (i.e. the second scenario), the misspecified covariate distribution for z_i has a clear effect on the statistical power of detecting the presence of cz_i^2, especially when the missing data fraction is large. This indicates that the covariate distribution may have a profound effect on the finite-sample performance of our goodness-of-fit tests.

Moreover, we assessed the power of CM$_2$ in detecting whether the missing data mechanism depends on the response variable. Specifically, 500 data sets were generated from $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$ assuming z_i is MAR,

$$p(r_i=1|x_i, z_i, y_i)=\frac{\exp(\xi_0+\xi_1 x_i+ay_i)}{1+\exp(\xi_0+\xi_1 x_i+ay_i)},$$

for $i = 1, \ldots, 100$, where$\beta_0 = \beta_1 = \beta_2 = 1$ and $\varepsilon_i \sim N(0, 1)$. We fit the linear model $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$ under (28). The rejection rates were 0.045, 0.198, 0.328 and 0.358 for $a = 0.0$, 1.0, 1.5 and 2.0 respectively. Thus, CM$_2$ can detect the dependence of the missing data mechanism on the response for large values of $|a|$.

The third group of simulation studies was to assess the finite-sample performance of CM$_3$. First, we evaluated the power of CM$_3$ in detecting the misspecification of $E[y_i | x_i, z_i]$ when the missing data mechanism is dependent on the response variable. We simulated 500 data sets using the second scenario in the first group of simulation studies, and then we fit the linear model assuming an MAR mechanism

$$p(r_i=1|x_i, z_i, y_i)=\frac{\exp(\xi_0+\xi_1 y_i)}{1+\exp(\xi_0+\xi_1 y_i)}, \tag{29}$$

with various values of ξ_0 and ξ_1 to obtain the desired average missing data fractions. The rejection rates of CM$_3$ were 0.051, 0.380, 0.514 and 0.594 for $c = 0.0$, 0.5, 1.0 and 1.5, respectively, assuming a 60% missingness fraction for z_i.

Furthermore, we assessed the power of CM$_3$ in detecting whether the missing data mechanism is non-ignorable. The 500 data sets were generated from $y_i = 1 + x_i + z_i + \varepsilon_i$ for $i = 1, \ldots, 100$, where $\varepsilon_i \sim N(0, 1)$, and the missing data mechanism is

$$p(r_i=1|x_i, z_i, y_i)=\frac{\exp(\xi_0+\xi_1 y_i+az_i)}{1+\exp(\xi_0+\xi_1 y_i+az_i)}.$$

Three average missingness fractions of 20%, 40% and 60% were used. We fit the linear model $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$ assuming (29). The rejection rates of CM$_3$ were 0.046, 0.16, 0.262 and 0.422 for $a = 0.0$, 1.0, 1.5 and 2.0, respectively, for the 60% missingness fraction.

4.3. Goodness-of-fit statistics for the logistic regression model

We then considered the logistic regression model. The first group of simulation studies was to compare the finite-sample performance of CM$_1$ using either $\mathrm{CR}_i^{(1)}$ or $\mathrm{CR}_i^{(2)}$ under the two similar scenarios as in the previous section. In the first scenario, we simulated 500 data sets from

$$p(y_i=1|x_i, z_i)=\frac{\exp(\beta_0+\beta_1 x_i+\beta_2 z_i+c x_i^2)}{1+\exp(\beta_0+\beta_1 x_i+\beta_2 z_i+c x_i^2)},$$

for $i = 1, \ldots, 200$, where the c was in the range [0, 1]. We set $\beta_0 = \beta_1 = \beta_2 = 1$. We considered two missing data mechanisms: MCAR and MAR. For MAR, the mechanism was given by (28), in which we set $\xi_1 = 1.0$ and ξ_0 with values $-1.5, -0.5$ and 0.5 to obtain average missing data fractions of 20%, 40% and 60% respectively. Then we fit

$$p(y_i=1|x_i, z_i)=\frac{\exp(\beta_0+\beta_1 x_i+\beta_2 z_i)}{1+\exp(\beta_0+\beta_1 x_i+\beta_2 z_i)},$$

assuming an MAR mechanism, and thus the fitted model would be misspecified if $c \neq 0$ and the misspecification is due to x_i.

The results shown in the first half of Table 5 are similar to those from the linear model. The type I error rates of CM_1 based on both $\mathrm{CR}_i^{(1)}$ and $\mathrm{CR}_i^{(2)}$ are accurate across all missingness fractions. The CM_1 based on $\mathrm{CR}_i^{(2)}$ is uniformly more powerful than that based on $\mathrm{CR}_i^{(1)}$. The power for detecting misspecification of the model increased with $|c|$ for CM_1. The missing data fraction slightly influences the power of detecting model misspecification for CM_1.

In the second scenario, we generated 500 data sets from

$$p(y_i=1|x_i, z_i)=\frac{\exp(\beta_0+\beta_1 x_i+\beta_2 z_i+c z_i^2)}{1+\exp(\beta_0+\beta_1 x_i+\beta_2 z_i+c z_i^2)},$$

whereas the rest of the set-up remained the same as in the first scenario. We fit the model ignoring the term $c z_i^2$, and thus the model would be misspecified if $c \neq 0$ and the misspecification is due to z_i. The results are shown in the second half of Table 5. Similar to the linear model, CM_1 based on both $\mathrm{CR}_i^{(1)}$ and $\mathrm{CR}_i^{(2)}$ cannot detect the misspecification of $c z_i^2$, because CM_1 did not incorporate the missing covariate z_i.

Similarly to the linear model, we assessed the finite-sample performance of CM_2 using the same two scenarios. The first half of Table 6 shows the results for the first scenario where the misspecification is due to x_i, and the second half of the table is for the second scenario where the misspecification is due to z_i. The type I errors rates of CM_2 are accurate across all missingness fractions. And for both scenarios, the power for detecting misspecification of the model increased with $|c|$ for CM_2 and the missing data fraction influences the power in detecting model misspecification (i.e. $|c| \neq 0$). Compared with Table 5, for the first scenario, CM_1 based on $\mathrm{CR}_i^{(2)}$ is slightly more powerful than CM_2 in detecting the presence of $c x_i^2$. However, for the second scenario, CM_2 is much more powerful than CM_1 based on $\mathrm{CR}_i^{(2)}$. This indicates that incorporating the missing data can increase the power of detecting model misspecification due to $c z_i^2$.

5. Liver cancer data

To illustrate our proposed methods, we considered data on 191 patients from two Eastern Cooperative Oncology Group clinical trials as mentioned in section 1 (Ibrahim *et al.*, 1999). We are interested in how the number of cancerous liver nodes (y) when entering the trials is predicted by six other baseline characteristics: time since diagnosis of the disease (in weeks) (z_1); two biochemical markers (each classified as normal or abnormal), alpha-fetoprotein (z_2) and anti-hepatitis antigen (z_3); associated jaundice (yes, no) (x_1); body mass index (weight in kilograms divided by the square of height in metres) (x_2); and age (in years) (x_3).

We used a Poisson regression model, given by

$$p(y_i|\mathbf{x}_i, \mathbf{z}_i, \beta) \propto \exp[y_i(\mathbf{v}_i'\beta) - \exp(\mathbf{v}_i'\beta)]$$

where $\mathbf{v}_i'=(1, x_{i1}, x_{i2}, x_{i3}, z_{i1}, z_{i2}, z_{i3})$ is the 1 × 7 vector of covariates including an intercept, and $\beta = (\beta_0, \beta_1, \dots, \beta_6)'$ are the corresponding regression coefficients. The logarithm of the time since diagnosis was used to achieve approximate normality. As only $\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3})$ have missing values, we need to consider a joint distribution only for these covariates given $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$. As z_{i2} and z_{i3} were both dichotomous, it was reasonable to model their conditional univariate distributions using logistic regressions. Thus

$$p(z_{i1}, z_{i2}, z_{i3}|\mathbf{x}_i, \alpha)=p(z_{i3}|z_{i1}, z_{i2}, \mathbf{x}_i, \alpha_3)p(z_{i2}|z_{i1}, \mathbf{x}_i, \alpha_2)p(z_{i1}|\mathbf{x}_i, \alpha_1),$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ and $(z_{i3} \mid z_{i1}, z_{i2}, \mathbf{x}_i)$ is a logistic regression with probability of success

$$p(z_{i3}=1|z_{i1}, z_{i2}, \mathbf{x}_i, \alpha_3)=\frac{\exp(\alpha_{30}+\alpha_{31}z_{i1}+\alpha_{32}z_{i2}+\alpha_{3x}'\mathbf{x}_i)}{1+\exp(\alpha_{30}+\alpha_{31}z_{i1}+\alpha_{32}z_{i2}+\alpha_{3x}'\mathbf{x}_i)},$$

and $\alpha_{3x}'=(\alpha_{33}, \alpha_{34}, \alpha_{35})$. Similarly,

$$p(z_{i2}=1|z_{i1}, \mathbf{x}_i, \alpha_2)=\frac{\exp(\alpha_{20}+\alpha_{21}z_{i1}+\alpha_{2x}'\mathbf{x}_i)}{1+\exp(\alpha_{20}+\alpha_{21}z_{i1}+\alpha_{2x}'\mathbf{x}_i)},$$

and $\alpha_{2x}'=(\alpha_{22}, \alpha_{23}, \alpha_{24})$. In addition, we took a normal distribution for the missing covariate z_1, specifically, $z_{i1} \sim N(\alpha_{11}, \alpha_{12})$, $i = 1, \dots, n$, and $\alpha_1'=(\alpha_{11}, \alpha_{12})$.

We assumed that the missing covariates are MAR and calculated the MLE of (β, α) using the EM algorithm. The case-deletion diagnostic measures CD_i identified cases 10, 15, 65, 131 and 160 as influential, among which $\mathrm{CD}_{i,1}$ identified cases 10, 15, 65 and 160, whereas $\mathrm{CD}_{i,2}$ identified case 131 (Fig. 2A–C). These findings confirmed the suspected cases reported in Table 1. The QD_i, $\mathrm{QD}_{i,1}$ and $\mathrm{QD}_{i,2}$ gave similar results (not presented here). The standardized conditional residuals, $\mathrm{SCR}^{(1)}$, detected cases 10, 15, 65 and 160 as influential observations (Fig. 2D) and $\mathrm{SCR}^{(2)}$ gave similar results (not presented).

The *p*-values of the goodness-of-fit test using CM_1 based on $\text{CR}_i^{(1)}$ and $\text{CR}_i^{(2)}$ were 0.56 and 0.48, respectively, whereas the *p*-value of the goodness-of-fit test using CM_2 was 0.06. We drew the residual plot of $I_1((\hat{\beta}, t); \hat{\eta})$ against *t* using $\text{CR}_i^{(2)}$ (Fig. 2E) and the residual plot of $I_2((\hat{\beta}, t); \hat{\eta})$ against *t* (Fig. 2F). In Fig. 2E and F, the observed pattern is shown by the thick drawn curve, and 15 simulated resampling realizations are shown by the thin dashed curves. The *p*-values for the supremum test with 500 realizations are 0.78 in Fig. 2E and 0.12 in Fig. 2F. All these indicated that either $E(y_i|\mathbf{v}_i) \neq \exp(\mathbf{v}_i'\beta)$ or the missing data mechanism depended on the response variable. So, we considered the following MAR mechanism,

$$p(\mathbf{r}_i|\mathbf{x}_i, y_i) = p(r_{i3}|r_{i1}, r_{i2}, \mathbf{x}_i, y_i, \xi_2) p(r_{i2}|r_{i1}, \mathbf{x}_i, y_i, \xi_2) p(r_{i1}|\mathbf{x}_i, y_i, \xi_1),$$

where $p(r_{i2} \mid r_{i1}, \mathbf{x}_i, y_i, \xi_2)$ and $p(r_{i1} \mid \mathbf{x}_i, y_i, \xi_1)$ are

$$\begin{aligned} p(r_{i1}|y_i, \mathbf{x}_i, \xi_1) &= \frac{\exp(f_{i1})}{1+\exp(f_{i1})}, \\ p(r_{i2}|r_{i1}, y_i, \mathbf{x}_i, \xi_2) &= \frac{\exp(f_{i2})}{1+\exp(f_{i2})}, \\ p(r_{i3}|r_{i1}, r_{i2}, y_i, \mathbf{x}_i, \xi_3) &= \frac{\exp(f_{i3})}{1+\exp(f_{i3})}, \end{aligned}$$

in which

$$\begin{aligned} f_{i1} &= \xi_{10} + \xi_{11} x_{i1} + \xi_{12} x_{i2} + \xi_{13} x_{i3} + \xi_{14} y_i, \\ f_{i2} &= \xi_{20} + \xi_{21} x_{i1} + \xi_{22} x_{i2} + \xi_{23} x_{i3} + \xi_{24} y_i + \xi_{25} r_{i1}, \end{aligned}$$

and

$$f_{i3} = \xi_{30} + \xi_{31} x_{i1} + \xi_{32} x_{i2} + \xi_{33} x_{i3} + \xi_{34} y_i + \xi_{35} r_{i1} + \xi_{36} r_{i2}.$$

We found that the missing data mechanism of z_{i1} depended on the response variable, so we should use CM_3 for the goodness-of-fit test. The goodness-of-fit test using CM_3 was not significant ($p = 0.56$), indicating that the model fit well.

6. Discussion

We have derived goodness-of-fit statistics in the presence of missing data based on novel definitions of case-deletion and residual diagnostics. The asymptotic properties of the goodness-of-fit measures based on conditional residuals were also derived, as well as MCMC algorithms for carrying out the EM algorithm. The simulation studies and liver cancer data set showed very promising results for the proposed methods. Future work in this area includes extending the methodologies to the Cox proportional hazards model with right censored survival data and missing covariates, as well as to parametric and semiparametric models for longitudinal data with MAR or NMAR response and/or covariate data.

We also note several limitations of our proposed tests. The first limitation is that we assume parametric distributions throughout the paper, whereas the goodness-of-fit tests focus on testing the regression function. It is very interesting to extend the definitions of conditional residuals and associated test statistics to semiparametric models. In addition, our preliminary results have shown that the misspecified distributions can have profound effects on the finite-sample performance of our proposed test statistics. The second limitation is that it is difficult to pinpoint the cause of rejection of the null hypothesis and subsequently to suggest an alternative model. This limitation is inherent in all omnibus tests based on integrated regressions (Stute, 1997). All of these issues merit further research, and we will study them in our future work.

Acknowledgments

The authors thank two anonymous referees, the Associate Editor and the Editor for very useful comments that improved the presentation of the paper. This work was supported in part by NSF grants SES-06-43663 and BCS-08-26844 and NIH grants UL1-RR025747-01 and R21AG033387 to Dr Zhu and NIH grants GM 70335 and CA 74015 to Dr Ibrahim.

References

Andrews, DWK. Empirical process methods in econometrics. In: Engle, RF.; McFadden, DL., editors. Handbook of econometrics. Vol. IV. North-Holland; Amsterdam: 1994. p. 2248-2292.

Andrews DWK. Estimation when a parameter is on a boundary. Econometrica 1999;67:1341–1383.

Bierens H. Consistent model specification tests. J Econometrics 1982;20:105–134.

Chen, MH.; Shao, QM.; Ibrahim, JG. Monte Carlo methods in Bayesian computation. Springer-Verlag; New York: 2000.

Cook RD. Assessment of local influence (with discussion contributions). J Roy Statist Soc Ser B 1986;48:133–169.

Cook, RD.; Weisberg, S. Residuals and influence in regression. Chapman & Hall; London: 1982.

Cox DR, Snell EJ. A general definition of residuals (with discussion contributions). J Roy Statist Soc Ser B 1968;30:248–275.

Davison AC, Tsai CL. Regression model diagnostics. Int Statist Rev 1992;60:337–355.

Escanciano JC. A consistent diagnostic test for regression models using projection. Econometric Theory 2006;22:1030–1051.

Ibrahim JG. Incomplete data in generalized linear models. J Amer Statist Assoc 1990;85:765–769.

Ibrahim JG, Lipsitz SR. Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. Biometrics 1996;52:1071–1078. [PubMed: 8805768]

Ibrahim JG, Lipsitz SR, Chen MH. Missing covariates in generalized linear models when the missing data mechanism is nonignorable. J Roy Statist Soc Ser B 1999;61:173–190.

Ibrahim JG, Chen MH, Lipsitz SR. Monte Carlo EM for missing covariates in parametric regression models. Biometrics 1999;55:591–596. [PubMed: 11318219]

Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. Missing-data methods for generalized linear models: a comparative review. J Amer Statist Assoc 2005;100:332–346.

Kosorok MR. Bootstraps of sums of independent but not identically distributed stochastic processes. J Multivariate Anal 2003;84:299–318.

Kosorok, MR. Introduction to empirical processes and semiparametric inference. Springer-Verlag; New York: 2007.

Lehmann, EL.; Romano, JP. Testing statistical hypotheses. Vol. 3. Springer-Verlag; New York: 2006.

Lin DY, Wei LJ, Ying ZL. Model-checking techniques based on cumulative residuals. Biometrics 2002;58:1–12. [PubMed: 11890304]

Lipsitz SR, Ibrahim JG. A conditional model for incomplete covariates in parametric regression models. Biometrika 1996;83:916–922.

Lipsitz SR, Ibrahim JG. Estimating equations with incomplete categorical covariates in the Cox model. Biometrics 1998;54:1002–1013. [PubMed: 9750248]

Little, RJA.; Rubin, DB. Statistical analysis with missing data. Vol. 2. John Wiley; New York: 2002.

Little RJA, Schluchter M. Maximum likelihood estimation for mixed continuous and categorical data with missing values. Biometrika 1985;72:497–512.

Liu, J. Monte Carlo strategies in scientific computing. Springer-Verlag; New York: 2003.

McCullagh, P.; Nelder, JA. Generalized linear models. Vol. 2. Chapman & Hall; London: 1989.

Ossiander M. A central limit theorem under metric entropy with bracketing. Ann Probab 1987;15:897–919.

Schluchter M, Jackson K. Log-linear analysis of censored survival data with partially observed covariates. J Amer Statist Assoc 1989;84:42–52.

Stute W. Nonparametric model checks for regression. Ann Statist 1997;25:613–641.

Stute W, Gonzlez-Manteiga W, Presedo-Quindimil M. Bootstrap approximations in model checks for regression. J Amer Statist Assoc 1998;93:141–149.

Stute W, Zhu LX. Model checks for generalized linear models. Scand J Statist 2002;29:535–545.

Su JQ, Wei LJ. A lack-of-fit test for the mean function in a generalized linear model. J Amer Statist Assoc 1991;86:420–426.

Van der Vaart, AW.; Wellner, JA. Weak convergence and empirical processes. Springer-Verlag; New York: 1996.

Zhu HT, Lee SY, Wei BC, Zhou J. Case-deletion measures for models with incomplete data. Biometrika 2001;88:727–737.

Appendix: Assumptions and Proofs

The following assumptions are needed to facilitate development of our methods, although they may not be the weakest possible conditions.

(C1) η_* is unique and an interior point of Υ, where Υ is a compact set in $R^{\dim(\eta)}$.

(C2) $\hat{\eta} \to \eta_*$ in probability as $n \to \infty$.

(C3) For each i, $\ell(\mathbf{d}_i; \eta) = \log p(\mathbf{d}_i; \eta)$ is three times continuously differentiable on Υ and $|\partial_j \ell(\mathbf{d}_i; \eta)|^2$ and $|\partial_j \partial_k \ell(\mathbf{d}_i; \eta)|$ are dominated by an integrable function B_i $(\mathbf{d}_i)$ for all j, k = 1, …, d, where $\partial_j = \partial/\partial\eta_j$.

(C4) For each $\varepsilon > 0$, there exists a finite K such that

$$\sup_{n\geq 1} n^{-1}\sum_{i=1}^{n} E[\,B_i(\mathbf{d}_i)^2\mathbf{1}\{B_i(\mathbf{d}_i)>K\}]<\varepsilon$$

for all n, where $\mathbf{1}\{B_i\,(\mathbf{d}_i) > K\}$ is the indicator function of $B_i\,(\mathbf{d}_i) > K$.

(C5)

$$\lim_{n\to\infty} n^{-1}\left\{-\sum_{i=1}^{n}\partial_\eta^2\ell(\mathbf{d}_{o,i};\eta_*)\right\}=A(\eta_*)\ \text{and}\ \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\{\partial_\eta\ell(\mathbf{d}_{o,i};\eta_*)\partial_\eta\ell(\mathbf{d}_{o,i};\eta_*)'\}=B(\eta_*),$$

where $A(\eta_*)$ is non-singular and $B(\eta_*)$ is positive definite.

(C6) Let $\rho((\boldsymbol{\phi}, t), (\boldsymbol{\phi}_*, t_*))$ be the limit of $\rho_n((\boldsymbol{\phi}_n, t_n), (\rho_{*n}, t_{*n}))$, where

$$\rho_n((\phi_n,t_n),(\phi_{*n},t_{*n}))=\left(n^{-1}\sum_{i=1}^{n}E[\,|\mathrm{CR}_i(\eta_*)|^2|\mathbf{1}(\phi_n'\mathbf{x}\leq t_n)-\mathbf{1}(\phi_{*n}'\mathbf{x}\leq t_{*n})|^2\right)^{1/2}.$$

For any sequences $\{(\boldsymbol{\phi}_n, t_n)\}$ and $\{(\boldsymbol{\phi}_{*n}, t_{*n})\}$, $\rho_n((\boldsymbol{\phi}_n, t_n), (\boldsymbol{\phi}_{*n}, t_{*n}))$ converges to zero when $\rho((\boldsymbol{\phi}_n, t_n), (\boldsymbol{\phi}_{*n}, t_{*n})) \to 0$ as $n \to \infty$. A similar condition also holds for $I_3((\tilde{\boldsymbol{\phi}}, t); \eta_*)$.

(C7) $\Delta_1(\boldsymbol{\phi}, t)$ and $F_{\boldsymbol{\phi}}(\mathrm{d}t)\,\mathrm{d}\boldsymbol{\phi}$ are absolutely continuous with respect to Lebesgue measure on Π.

(C8) For any small $a_0 > 0$, we assume that

$$\sup_{(\alpha,\phi,t)\in\mathcal{A}\times\Pi} P\{-\delta<[\mathbf{c}_i(\alpha)'\phi-t]/V_i<\delta\} \le C_0\delta^{c_1},$$

where C_0 and c_1 are two positive scalars,

$$\mathcal{A}=\{\alpha:\|\alpha-\alpha_*\|_2 \le a_0\} \text{ and } \sup_{\alpha\in\mathcal{A}}\|\partial_\alpha[\mathbf{c}_i(\alpha)]\|_2^2+\sup_{\alpha\in\mathcal{A}}\|\mathbf{c}_i(\alpha)\|_2^2+1=V_i(\mathbf{x}_i,\mathbf{z}_{o,i})^2.$$

Comments

Condition C1 is a standard identifiability condition. Some sufficient conditions for condition C2 have been widely presented in the literature; see Van der Vaart & Wellner (1996) and Andrews (1999). Conditions C3–C5 are required to ensure the asymptotic normality of $\hat{\eta}$. Condition C6 is required to invoke the central limit theory for the sums of independent but not identically distributed stochastic processes (Van der Vaart & Wellner, 1996; Kosorok, 2007). Condition C7 is required to ensure the asymptotic distributions of the Cramer–von Mises test statistics. C8 is required to invoke Ossiander's entropy conditions (Ossiander, 1987; Andrews, 1994).

Proof of proposition 1

For ease of exposition, we omit η_* in some notation, such as $p(\mathbf{r}_i \mid y_i, \mathbf{x}_i, \mathbf{z}_i)$.

i. For brevity, we only consider $\mathrm{CR}_i^{(1)}(\eta_*)$. It can be shown that

$$\begin{aligned}&E[E[g((\mathbf{x}_i',\mathbf{z}_i')\beta)|\mathbf{x}_i,\mathbf{z}_{o,i}]|\mathbf{x}_i]=E[g((\mathbf{x}_i',\mathbf{z}_i')\beta)|\mathbf{x}_i],\\&E[y_i|\mathbf{x}_i]=\int y_i p(y_i|\mathbf{x}_i,\mathbf{z}_i)p(\mathbf{z}_i|\mathbf{x}_i)p(\mathbf{r}_i|y_i,\mathbf{x}_i,\mathbf{z}_i)\mathrm{d}y_i\mathrm{d}\mathbf{z}_i\mathrm{d}\mathbf{r}_i\\&=\int g((\mathbf{x}_i',\mathbf{z}_i')\beta_*)p(\mathbf{z}_i|\mathbf{x}_i)\mathrm{d}\mathbf{z}_i,\end{aligned}$$

which yields $E[\mathrm{CR}_i^{(1)}(\eta_*)|\mathbf{x}_i]=0$. Furthermore, $E[\mathrm{CR}_i^{(1)}(\eta_*)]=E\{E[\mathrm{CR}_i^{(1)}(\eta_*)|\mathbf{x}_i]\}=0$. However, it can be shown that

$$\begin{aligned}&E[y_i|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i]=\frac{\int y_i p(y_i|\mathbf{x}_i,\mathbf{z}_i)p(\mathbf{x}_i,\mathbf{z}_i)p(\mathbf{r}_i|y_i,\mathbf{x}_i,\mathbf{z}_i)\mathrm{d}\mathbf{z}_{m,i}\mathrm{d}y_i}{\int p(y_i|\mathbf{x}_i,\mathbf{z}_i)p(\mathbf{x}_i,\mathbf{z}_i)p(\mathbf{r}_i|y_i,\mathbf{x}_i,\mathbf{z}_i)\mathrm{d}\mathbf{z}_{m,i}\mathrm{d}y_i}\\&\ne E[g((\mathbf{x}_i',\mathbf{z}_i')\beta_*)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i].\end{aligned}$$

ii. For MAR covariates, we have $p(\mathbf{r}_i \mid \mathbf{x}_i, \mathbf{z}_i, y_i) = p(\mathbf{r}_i \mid \mathbf{x}_i, \mathbf{z}_{o,i}, y_i)$. It can be shown that

$$\begin{aligned}&E[g((\mathbf{x}_i',\mathbf{z}_i')\beta)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i,y_i]=\frac{\int g((\mathbf{x}_i',\mathbf{z}_i')\beta)p(y_i|\mathbf{x}_i,\mathbf{z}_i)p(\mathbf{x}_i,\mathbf{z}_i)p(\mathbf{r}_i|y_i,\mathbf{x}_i,\mathbf{z}_{o,i})\mathrm{d}\mathbf{z}_{m,i}}{\int p(y_i|\mathbf{x}_i,\mathbf{z}_i)p(\mathbf{x}_i,\mathbf{z}_i)p(\mathbf{r}_i|y_i,\mathbf{x}_i,\mathbf{z}_{o,i})\mathrm{d}\mathbf{z}_{m,i}}\\&=\frac{\int g((\mathbf{x}_i',\mathbf{z}_i')\beta)p(y_i|\mathbf{x}_i,\mathbf{z}_i)p(\mathbf{x}_i,\mathbf{z}_i)\mathrm{d}\mathbf{z}_{m,i}}{\int p(y_i|\mathbf{x}_i,\mathbf{z}_i)p(\mathbf{x}_i,\mathbf{z}_i)\mathrm{d}\mathbf{z}_{m,i}}\\&=E[g((\mathbf{x}_i',\mathbf{z}_i')\beta)|\mathbf{x}_i,\mathbf{z}_{o,i},y_i].\end{aligned}$$

Thus, we have

$$\mathrm{CR}_i^{(2)}(\eta)=y_i - E[g((\mathbf{x}_i', \mathbf{z}_i')\beta)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i]=y_i - E[g((\mathbf{x}_i', \mathbf{z}_i')\beta)|\mathbf{x}_i, \mathbf{z}_{o,i}, y_i].$$

Furthermore, it can be shown that

$$E\left[\frac{\mathrm{CR}_i^{(1)}}{p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_{o,i}, y_i)}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i\right]=\frac{\int \mathrm{CR}_i^{(1)} p(y_i|\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{x}_i, \mathbf{z}_i)\mathrm{d}\mathbf{z}_{m,i}\mathrm{d}y_i}{\int p(y_i|\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_{o,i})\mathrm{d}\mathbf{z}_{m,i}\mathrm{d}y_i}=0.$$

iii. Using $p(\mathbf{r}_i \mid \mathbf{x}_i, \mathbf{z}_i, y_i) = p(\mathbf{r}_i \mid \mathbf{x}_i, \mathbf{z}_i)$, we obtain

$$\begin{aligned} E[y_i|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i]&=\frac{\int y_i p(y_i|\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i)\mathrm{d}\mathbf{z}_{m,i}\mathrm{d}y_i}{\int p(y_i|\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i)\mathrm{d}\mathbf{z}_{m,i}\mathrm{d}y_i} \\ &=\frac{\int g((\mathbf{x}_i', \mathbf{z}_i')\beta_*)p(\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i)\mathrm{d}\mathbf{z}_{m,i}}{\int p(\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i)\mathrm{d}\mathbf{z}_{m,i}}=E[g((\mathbf{x}_i', \mathbf{z}_i')\beta_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i]. \end{aligned}$$

Thus, $E[\mathrm{CR}_i^{(2)}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i]=0$ and $E[\mathrm{CR}_i^{(1)}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i] \neq 0$.

iv. Using first-order Taylor's series expansions yields the desired results.

Proof of theorem 1

(i) Conditions C1–C5 are sufficient for establishing (i) (Andrews, 1994; Van der Vaart & Wellner, 1996).

(ii) First, we can prove weak convergence of $I_1(\cdot; \eta_*)$ using a standard argument of empirical process theory. The finite-dimensional marginals of $I_1(\cdot; \eta_*)$ converge weakly to the corresponding marginals of the zero-mean Gaussian process $G_1(\cdot)$. This can be proved by using assumptions C3 and C4. Because

$$\mathcal{F}=\{f(\phi, t)=\mathrm{CR}(\eta_*)\mathbf{1}(\phi'\mathbf{x} \le t):(\phi, t) \in \Pi\}$$

is a VC (Vapnik and Cervonenkis) class, which satisfies the universal entropy condition (Van der Vaart & Wellner, 1996, sections 2.5 and 2.6), the tightness of $I_1(\cdot; \eta_*)$ follows from the Donsker Theorem (Van der Vaart & Wellner, 1996, section 2.11). Second, the convergence of $\sqrt{n}(\widehat{\eta} - \eta_*)$ follows from the standard Lindeberg–Feller theorem. Third, we can prove the joint convergence of $I_1(\cdot; \eta_*)$ and $\sqrt{n}(\widehat{\eta} - \eta_*)$ using the Cramer–Wold device and empirical process theory.

(iii) It can be shown from a Taylor's series expansion that

$$\begin{aligned} &I_1((\phi, t);\widehat{\eta})=I_1((\phi, t);\eta_*)+n^{1/2}(\widehat{\eta} - \eta_*)n^{-1}\sum_{i=1}^{n}\partial_\eta[\mathrm{CR}_i(\eta_*)]\mathbf{1}(\phi'\mathbf{x}_i \le t) \\ &+n^{1/2}(\widehat{\eta} - \eta_*)n^{-1}\sum_{i=1}^{n}\{\partial_\eta[\mathrm{CR}_i(\widetilde{\eta})] - \partial_\eta[\mathrm{CR}_i(\eta_*)]\}\mathbf{1}(\phi'\mathbf{x}_i \le t), \end{aligned} \tag{30}$$

where $\|\tilde{\eta} - \eta_*\| \le \|\hat{\eta} - \eta_*\| \to 0$. It follows from the law of large numbers and assumptions C3 and C4 that

$$n^{-1}\sum_{i=1}^{n}\{\partial_\eta[\mathrm{CR}_i(\tilde{\eta})]-\partial_\eta[\mathrm{CR}_i(\eta_*)]\}\mathbf{1}(\phi'\mathbf{x}_i\le t)$$

converges to zero uniformly in $(\boldsymbol{\phi}, t)$ in probability (Van der Vaart & Wellner, 1996). Similarly,

$$n^{-1}\sum_{i=1}^{n}\partial_\eta[\mathrm{CR}_i(\eta_*)]\mathbf{1}(\phi'\mathbf{x}_i\le t)$$

converges to $\Delta_1(\boldsymbol{\phi}, t)$ uniformly in $(\boldsymbol{\phi}, t)$ in probability. Because $n^{1/2}(\hat{\eta}-\eta_*)$ is asymptotically normal and $\Delta_1(\boldsymbol{\phi}, t)$ is uniformly continuous, the second term of (30) on the right-hand side is asymptotically tight. As we have already established weak convergence of $I_1((\boldsymbol{\phi}, t); \eta_*)$, we can use a standard argument of empirical process theory to establish that $I_1(\cdot; \hat{\eta})$ converges weakly to $G_1(\cdot) + \Delta_1(\cdot)'v_1$ as $n \to \infty$. Applying the continuous mapping theorem ensures that CK_1 converges in distribution to

$$\sup_{(\phi,t)}|G_1(\phi,t)+\Delta_1(\phi,t)'v_1|.$$

To prove weak convergence of CM_1, we use proposition 7.27 of Kosorok (2007) to prove that $\mathrm{CM}_1 = \int_\Pi|I_1((\boldsymbol{\phi}, t); \hat{\eta})|^2F_{n,\boldsymbol{\phi}}(\mathrm{d}t)\,\mathrm{d}\boldsymbol{\phi}$ converges weakly to

$$\int_\Pi|G_1(\phi,t)+\Delta_1(\phi,t)'v_1|^2F_\phi(\mathrm{d}t)\mathrm{d}\phi\ \ \text{as}\, n\to\infty,$$

as $I_1((\boldsymbol{\phi}, t); \hat{\eta})$ converges weakly to $G_1(\boldsymbol{\phi}, t) + \Delta_1(\boldsymbol{\phi}, t)'v_1 \in \ell^\infty(\Pi)$ and

$$\sup_{t\in[-\infty,\infty]}|F_{n,\phi}(t)-F_\phi(t)|\to 0\ \ \text{as}\, n\to\infty.$$

Proof of theorem 2

(i) We define $\ell_n(t) = \log p(D_o; t)$, where

$$p(D_o;t)=\prod_{i=1}^{n}\int p(y_i|\mathbf{x}_i,\mathbf{z}_i,t)p(\mathbf{x}_i,\mathbf{z}_i)p(\mathbf{r}_i|\mathbf{x}_i,\mathbf{z}_i,y_i)\mathrm{d}\mathbf{z}_{m,i}.$$

The true density function of $\mathbf{d}_{o,i}$ under local alternatives equals $p(D_o; n^{-1/2})$. Using a Taylor's series expansion, we get

$$\ell_n(n^{-1/2})=\ell_n(0)+n^{-1/2}\partial_t\ell_n(0)+0.5n^{-1}\partial_t^2\ell_n(0)+o_p(1),$$

where ∂_t = d/dt and ∂_t^2=d^2/dt^2. In particular, we have

$$\partial_t \ell_n(0)=\sum_{i=1}^{n} E[a_i(\tau_*)^{-1}\partial_t\theta_i(0)(y_i-\mu_i)|\mathbf{d}_{o,i}, t=0],$$

where the conditional expectation is taken with respect to $\mathbf{z}_{m,i}$ given $\mathbf{d}_{o,i}$ under t = 0. Under p (D_o; t = 0), ($\sqrt{n}(\widehat{\eta}-\eta_*)$, $\ell_n(n^{-1/2}) - \ell_n(0)$) can be approximated by

$$\left(n^{-1/2}\sum_{i=1}^{n}\psi_{n,i}, \partial_t\ell_n(0)n^{-1/2}\right)+(0, -0.5n^{-1}[-\partial_t^2\ell_n(0)])+o_p(1).$$

Following the arguments in example 12.3.8 of Lehmann & Romano (2006), we can show that under local alternative hypotheses, $\sqrt{n}(\widehat{\eta}-\eta_*)$ converges in distribution to $A_1 + v_1$.

(ii) For simplicity, we only consider $\mathrm{CR}_i^{(1)}$. The process $I_1((\boldsymbol{\phi}, t); \eta_*)$ can be represented as

$$I_1((\phi,t);\eta_*)=n^{-1/2}\sum_{i=1}^{n}\mathbf{1}(\phi'\mathbf{x}_i\le t)(y_i-E[g((\mathbf{x}_i',\mathbf{z}_i')\beta_*)+n^{-1/2}g_0(\mathbf{x}_i,\mathbf{z}_i)|\mathbf{x}_i,\mathbf{z}_{o,i}])$$
$$+n^{-1}\sum_{i=1}^{n}\mathbf{1}(\phi'\mathbf{x}_i\le t)E[g_0(\mathbf{x}_i,\mathbf{z}_i)|\mathbf{x}_i,\mathbf{z}_{o,i}],$$

in which the first term on the right-hand side converges weakly to $G_1(\cdot)$ by using similar arguments as in theorem 1(ii). In addition, it follows from the law of large numbers that

$$n^{-1}\sum_{i=1}^{n}\mathbf{1}(\phi'\mathbf{x}_i\le t)E[g_0(\mathbf{x}_i,\mathbf{z}_i)|\mathbf{x}_i,\mathbf{z}_{o,i}]$$

converges to $A_2(\boldsymbol{\phi}, t)$ uniformly in probability.

(iii) Following similar arguments as in theorem 1(iii), we use a Taylor's series expansion to show that

$$I_1((\phi,t);\widehat{\eta})=I_1((\phi,t);\eta_*)+n^{1/2}(\widehat{\eta}-\eta_*)n^{-1}\sum_{i=1}^{n}\partial_\eta[\mathrm{CR}_i(\eta_*)]\mathbf{1}(\phi'\mathbf{x}_i\le t)$$
$$+n^{1/2}(\widehat{\eta}-\eta_*)n^{-1}\sum_{i=1}^{n}\{\partial_\eta[\mathrm{CR}_i(\widetilde{\eta})]-\partial_\eta[\mathrm{CR}_i(\eta_*)]\}\mathbf{1}(\phi'\mathbf{x}_i\le t),$$

where $\|\tilde{\eta}-\eta_*\|_2 \le \|\hat{\eta}-\eta_*\|_2$. Similar to the arguments in theorem 1(iii), we can use standard arguments of empirical processes and the results in theorem 2(i) and (ii) to complete the proof of (iii).

Proof of theorem 3

The proof of theorem 3 consists of two steps as follows. In step 1, we need to prove that $I_2((\tilde{\boldsymbol{\phi}}, t); \eta_*)$ can be represented as

$$n^{-1/2}\sum_{i=1}^{n}\mathbf{1}(\mathbf{c}'_{i,*}\ \tilde{\phi}\leq t)\mathrm{CR}_i^{(2)}(\eta_*)+n^{-1/2}\sum_{i=1}^{n}[\,\mathbf{1}(\widetilde{\mathbf{c}}'_i\ \tilde{\phi}\leq t)-\mathbf{1}(\mathbf{c}'_{i,*}\ \tilde{\phi}\leq t)\mathrm{CR}_i^{(2)}(\eta_*), \tag{31}$$

where $\mathbf{c}_{i,*} = \mathbf{c}_i\ (\alpha_*)$. We first show that the second term of (31) converges to zero uniformly in probability and a sufficient condition is that

$$\left\{n^{-1/2}\sum_{i=1}^{m}\mathbf{1}(\mathbf{c}_i(\alpha)'\ \tilde{\phi}\leq t)\mathrm{CR}_i^{(2)}(\eta_*):\kappa=(\alpha,\tilde{\phi},t)\in\mathcal{A}\times\Pi\right\}$$

is stochastically equicontinuous, where $\mathcal{A} = \{\alpha: \|\alpha-\alpha_*\|_2 \leq a_0\}$ for a sufficiently small $a_0 > 0$. We invoke Ossiander's entropy condition to show that

$$\mathcal{M}=\{\mathbf{1}(\mathbf{c}(\alpha)'\ \tilde{\phi}\leq t)\mathrm{CR}^{(2)}(\eta_*):\kappa=(\alpha,\tilde{\phi},t)\in\mathcal{A}\times\Pi\}$$

is a type IV class (Ossiander, 1987; Andrews, 1994). We need to check the following condition:

$$\sup_i E\{[\mathrm{CR}_i^{(2)}(\eta_*)]^2\sup_{\kappa_1:\left\|\kappa_1-\kappa\right\|_2<\delta}|\mathbf{1}(\mathbf{c}_i(\alpha_1)'\tilde{\phi}_1\leq t_1)-\mathbf{1}(\mathbf{c}_i(\alpha)'\ \tilde{\phi}\leq t)|^2\}\leq C\delta^{c_1}, \tag{32}$$

where $\kappa_1 = (\alpha_1, \tilde{\boldsymbol{\phi}}_1, t_1)$ and C and c_1 are some finite positive constants. The left-hand side of (32) can be bounded above by

$$\begin{aligned}&\sup_i E\{E[\,[\mathrm{CR}_i^{(2)}(\eta_*)]^2|\mathbf{d}_{o,i}]\sup_{\kappa_1:\left\|\kappa_1-\kappa\right\|_2<\delta}|\mathbf{1}(\mathbf{c}_i(\alpha_1)'\tilde{\phi}_1\leq t_1)-\mathbf{1}(\mathbf{c}_i(\alpha)'\ \tilde{\phi}\leq t)|\}\\&=\sup_i E\{[\mathrm{CR}_i^{(2)}(\eta_*)]^2\sup_{\kappa_1:\left\|\kappa_1-\kappa\right\|_2<\delta}|\mathbf{1}(\mathbf{c}_i(\alpha_1)'\tilde{\phi}_1\leq t_1)-\mathbf{1}(\mathbf{c}_i(\alpha)'\ \tilde{\phi}\leq t)|\}\\&\leq\sup_i E\{[\mathrm{CR}_i^{(2)}(\eta_*)]^4\}^{1/2}\sup_i E\left[\sup_{\kappa_1:\left\|\kappa_1-\kappa\right\|_2<\delta}|\mathbf{1}(\mathbf{c}_i(\alpha_1)'\tilde{\phi}_1\leq t_1)-\mathbf{1}(\mathbf{c}_i(\alpha)'\ \tilde{\phi}\leq t)|\right]^{1/2},\end{aligned}$$

in which we have used the Cauchy–Schwartz inequality twice and

$$|\mathbf{1}(S_1)-\mathbf{1}(S_2)|^2=|\mathbf{1}(S_1)-\mathbf{1}(S_2)|$$

for any two sets S_1 and S_2. As

$$E[\mathrm{CR}_i^{(2)}(\eta_*)]^4 = E[\mathrm{CR}_1^{(2)}(\eta_*)]^4 \text{ for all } i,$$

it follows from condition C2 that

$$\sup_i E[\mathrm{CR}_i^{(2)}(\eta_*)]^4 = E[\mathrm{CR}_1^{(2)}(\eta_*)]^4 < \infty.$$

Let $h_i(\kappa) = \mathbf{c}_i(\alpha)' \boldsymbol{\phi} - t$. It follows from a Taylor's series expansion that

$$h_i(\kappa) = h_i(\kappa_1) + \partial_\kappa h_i(\tilde{\kappa})'(\kappa - \kappa_1), \text{ where } \left\|\tilde{\kappa} - \kappa_1\right\|_2 \le \left\|\kappa - \kappa_1\right\|_2.$$

Thus, we have

$$|h_i(\kappa) - h_i(\kappa_1)| \le \left\|\partial_\kappa h_i(\tilde{\kappa})\right\|_2 \left\|\kappa - \kappa_1\right\|_2, \text{ where } \partial_\kappa h_i(\tilde{\kappa}) = (\partial_\alpha[\mathbf{c}_i(\alpha)'\phi]', \mathbf{c}_i(\alpha)', 1)'.$$

Then, we have

$$\left\|\partial_\kappa h_i(\tilde{\kappa})\right\|_2^2 \le \left\|\partial_\alpha[\mathbf{c}_i(\alpha)]\right\|_2^2 + \left\|[\mathbf{c}_i(\alpha)]\right\|_2^2 + 1 \le V_i.$$

Using $|\mathbf{1}(S_1) - \mathbf{1}(S_2)| \le \mathbf{1}(S_1 \cap S_2^c) + \mathbf{1}(S_2 \cap S_1^c)$ and condition C7, we can further show that

$$\begin{aligned} & E\left[\sup_{\kappa_1 : \left\|\kappa_1 - \kappa\right\|_2 < \delta} |\mathbf{1}(h_i(\kappa) \le h_i(\kappa) - h_i(\kappa_1)) - \mathbf{1}(h_i(\kappa) \le 0)|\right] \\ & \le E[\mathbf{1}(-\sqrt{V_i}\delta \le h_i(\kappa) \le \sqrt{V_i}\delta)] \le C_0 \delta^{c_1}. \end{aligned}$$

In step 2, we follow the arguments of theorem 1(ii) to prove that

$$n^{-1/2} \sum_{i=1}^n \mathbf{1}(\mathbf{c}_{i,*}' \tilde{\phi} \le t) \mathrm{CR}_i^{(2)}(\eta_*)$$

converges to $G_2(\cdot)$ in distribution.

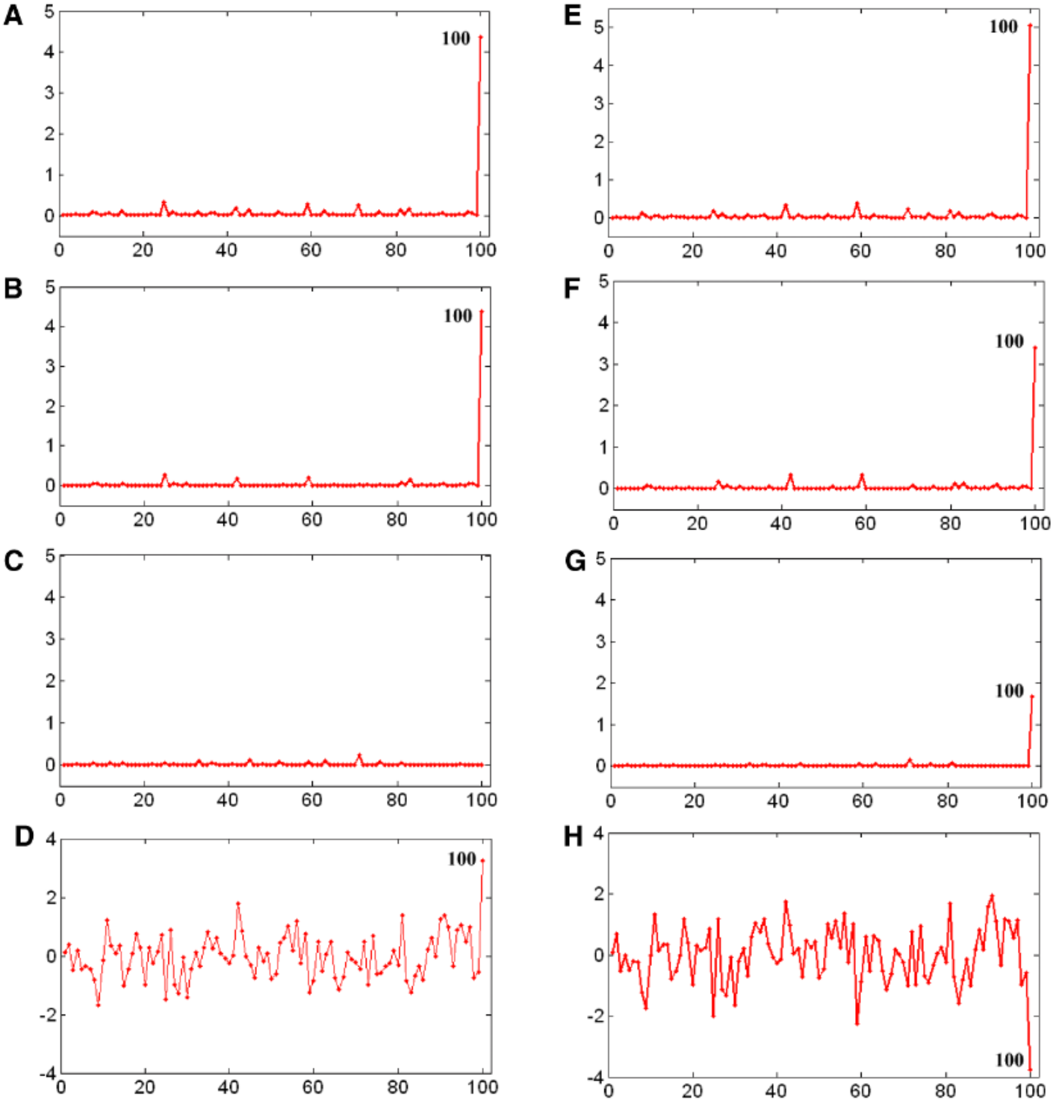

Fig. 1.

Index plots of diagnostic measures from two simulated data sets: (A) $\mathrm{CD}_{i;}$ (B) $\mathrm{CD}_{i,1}$; (C) $\mathrm{CD}_{i,2}$; (D) $\mathrm{SCR}_i^{(1)}$; (E) $\mathrm{CD}_{i;}$ (F) $\mathrm{CD}_{i,1}$; (G) $\mathrm{CD}_{i,2}$; (H) $\mathrm{SCR}_i^{(1)}$. Column one shows the results from the simulated data with y_{100} as an influential point, whereas column two shows the results from the simulated data with z_{100} as an influential point.

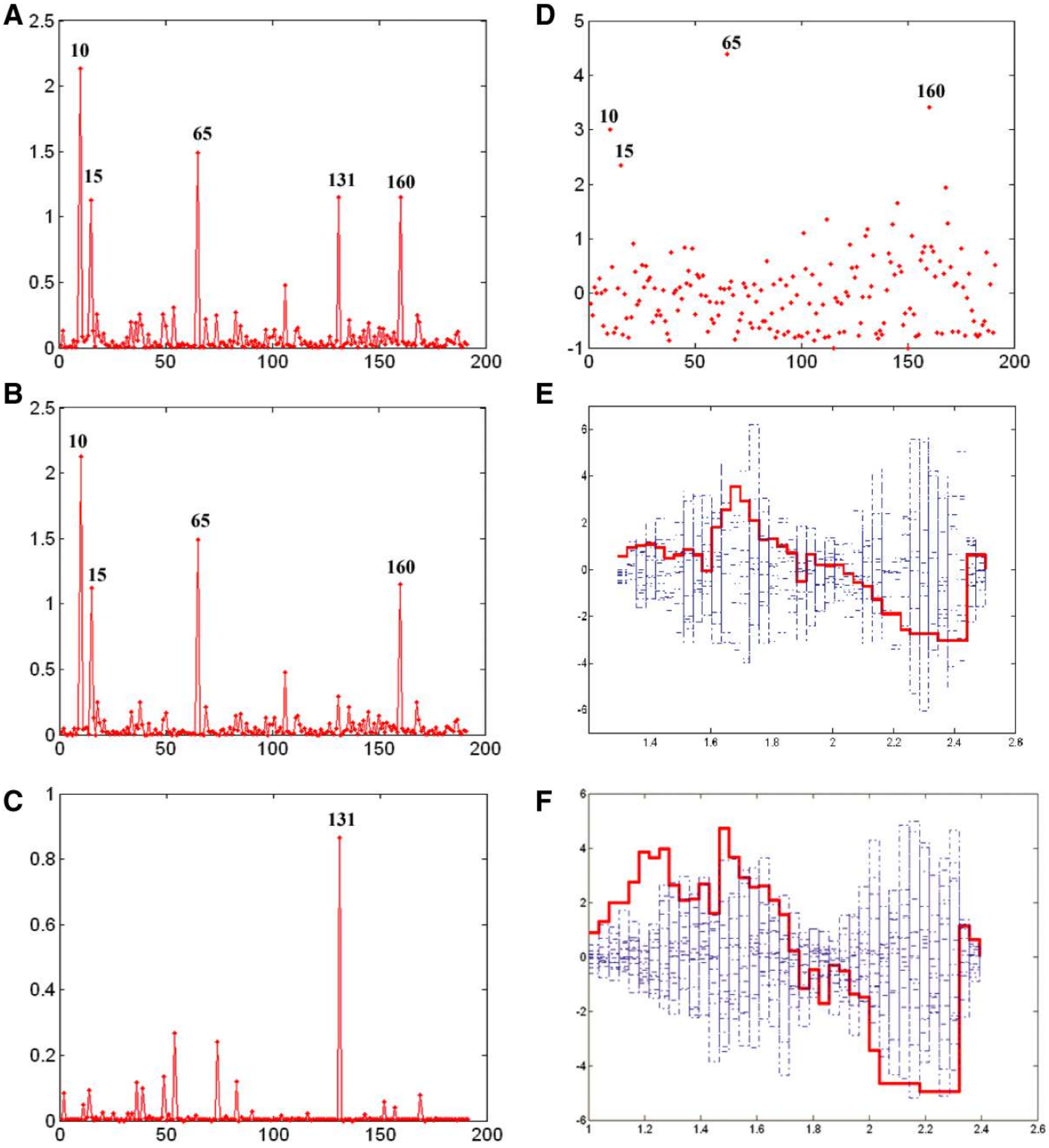

Fig. 2.

Liver cancer data: index plots of diagnostic measures: (A) CD$_i$, (B) CD$_{i,1}$, (C) CD$_{i,2}$, (D) $\mathrm{SCR}_i^{(1)}$; (E) residual plot of $I_1((\hat{\beta}, t); \hat{\eta})$ against t using $\mathrm{CR}_i^{(2)}$, (F) residual plot of $I_2((\hat{\beta}, t); \hat{\eta})$ against t. In (E) and (F), the observed pattern is shown by the thick drawn curve and 15 simulated resampling realizations are shown by the thin dashed curves.

Table 1

The five influential cases and the corresponding responses and covariates in the liver cancer data. The 65th and 131st observations have missing ‘Anti-hepatitis antigen’

Observation number	Number of cancerous nodes	Time	Alpha-fetoprotein	Anti-hepatitis	Jaundice	BMI	Age
10	61	2.29	0	1	1	18.81	31.06
15	21	0.57	1	0	1	23.73	42.29
65	23	2.57	1	.	1	23.72	70.52
131	6	320.86	0	.	0	20.31	66.19
160	21	1.14	1	0	1	22.94	65.40

Table 2

Rejection rates for CM$_1$ using either $\mathrm{CR}_i^{(1)}$ or $\mathrm{CR}_i^{(2)}$ at the 5% significance level for the linear model. The first half shows the results for the first scenario where the misspecification is due to x$_i$ and the second half of the table for the second scenario where the misspecification is due to z$_i$

	MCAR						MAR					
	20%		40%		60%		20%		40%		60%	
c	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$
$y_i=\beta_0+\beta_1 x_i+\beta_2 z_i+c x_i^2+\varepsilon_i$												
0	0.054	0.046	0.036	0.046	0.046	0.044	0.062	0.040	0.044	0.046	0.038	0.042
0.2	0.312	0.424	0.192	0.316	0.188	0.298	0.248	0.344	0.234	0.324	0.212	0.282
0.4	0.786	0.874	0.652	0.818	0.658	0.802	0.798	0.864	0.700	0.826	0.706	0.772
0.6	0.966	0.974	0.928	0.972	0.922	0.938	0.968	0.980	0.952	0.978	0.934	0.936
0.8	0.986	0.992	0.978	0.990	0.976	0.986	0.992	0.996	0.990	0.994	0.958	0.964
1.0	1.000	1.000	0.984	0.994	0.986	0.992	1.000	1.000	1.000	1.000	0.996	1.000
$y_i=\beta_0+\beta_1 x_i+\beta_2 z_i+c z_i^2+\varepsilon_i$												
0	0.054	0.046	0.036	0.046	0.046	0.044	0.062	0.040	0.044	0.046	0.038	0.042
0.2	0.062	0.062	0.034	0.056	0.032	0.034	0.058	0.048	0.032	0.028	0.038	0.042
0.4	0.044	0.048	0.058	0.058	0.018	0.028	0.040	0.036	0.030	0.044	0.036	0.040
0.6	0.046	0.052	0.046	0.050	0.034	0.048	0.032	0.036	0.028	0.034	0.030	0.036
0.8	0.050	0.062	0.040	0.054	0.048	0.056	0.040	0.042	0.038	0.040	0.032	0.042
1.0	0.052	0.068	0.048	0.058	0.044	0.052	0.046	0.050	0.034	0.052	0.038	0.044

Table 3

Rejection rates for CM_2 at the 5% significance level for the linear model. The first half shows the results for the first scenario where the misspecification is due to x_i and the second half of the table for the second scenario where the misspecification is due to z_i

	MCAR			**MAR**		
c	**20%**	**40%**	**60%**	**20%**	**40%**	**60%**
$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c x_i^2 + \varepsilon_i$						
0	0.044	0.046	0.044	0.046	0.044	0.042
0.2	0.264	0.164	0.170	0.222	0.210	0.204
0.4	0.756	0.648	0.622	0.716	0.678	0.630
0.6	0.938	0.906	0.854	0.946	0.906	0.890
0.8	0.972	0.970	0.966	0.990	0.970	0.938
1.0	1.000	0.984	0.980	1.000	0.986	0.980
$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c z_i^2 + \varepsilon_i$						
0	0.044	0.046	0.044	0.046	0.044	0.042
0.2	0.176	0.086	0.058	0.182	0.080	0.058
0.4	0.504	0.254	0.092	0.464	0.262	0.112
0.6	0.730	0.336	0.130	0.756	0.408	0.146
0.8	0.790	0.540	0.192	0.820	0.552	0.188
1.0	0.882	0.558	0.208	0.850	0.600	0.280

Table 4

Rejection rates for CM_2 at the 5% significance level for the linear model with a misspecified covariate. The missing covariates z_i are generated from a uniform distribution, whereas a normal distribution is assumed for z_i when we fit the data. The first half shows the results for the first scenario where the misspecification is due to x_i and the second half of the table is for the second scenario where the misspecification is due to z_i

	MCAR			MAR		
c	**20%**	**40%**	**60%**	**20%**	**40%**	**60%**
$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c x_i^2 + \varepsilon_i$						
0	0.040	0.053	0.053	0.040	0.042	0.038
0.2	0.262	0.246	0.244	0.298	0.176	0.210
0.4	0.766	0.760	0.664	0.782	0.728	0.698
0.6	0.962	0.936	0.946	0.954	0.942	0.880
0.8	0.980	0.964	0.964	0.994	0.980	0.974
1.0	0.998	0.982	0.980	0.994	0.984	0.982
$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c z_i^2 + \varepsilon_i$						
0	0.040	0.054	0.044	0.040	0.038	0.042
0.2	0.090	0.064	0.056	0.078	0.076	0.042
0.4	0.178	0.120	0.058	0.164	0.078	0.050
0.6	0.420	0.164	0.088	0.446	0.178	0.088
0.8	0.680	0.358	0.102	0.664	0.332	0.130
1.0	0.840	0.510	0.152	0.852	0.508	0.150

Table 5

Rejection rates for CM_1 using either $CR_i^{(1)}$ or $CR_i^{(2)}$ at the 5% significance level for the logistic regression model. The first half shows the results for the first scenario where the misspecification is due to x_i and the second half of the table is for the second scenario where the misspecification is due to z_i

	MCAR						MAR					
	20%		40%		60%		20%		40%		60%	
c	$\mathbf{CR}_i^{(1)}$	$\mathbf{CR}_i^{(2)}$	$\mathbf{CR}_i^{(1)}$	$\mathbf{CR}_i^{(2)}$	$\mathbf{CR}_i^{(1)}$	$\mathbf{CR}_i^{(2)}$	$\mathbf{CR}_i^{(1)}$	$\mathbf{CR}_i^{(2)}$	$\mathbf{CR}_i^{(1)}$	$\mathbf{CR}_i^{(2)}$	$\mathbf{CR}_i^{(1)}$	$\mathbf{CR}_i^{(2)}$
$\text{logit}(p(y_i = 1 \square x_i, z_i)) = \beta_0 + \beta_1 x_i + \beta_2 z_i + c x_i^2$												
0	0.062	0.056	0.060	0.054	0.060	0.052	0.058	0.056	0.064	0.058	0.062	0.058
0.2	0.282	0.346	0.162	0.242	0.144	0.208	0.278	0.324	0.212	0.316	0.200	0.282
0.4	0.592	0.662	0.486	0.520	0.448	0.480	0.584	0.632	0.560	0.618	0.546	0.602
0.6	0.804	0.856	0.896	0.892	0.876	0.884	0.818	0.826	0.822	0.834	0.804	0.826
0.8	0.952	0.980	0.922	0.932	0.896	0.906	0.960	0.986	0.920	0.944	0.906	0.914
1.0	1.000	1.000	0.940	0.964	0.906	0.942	1.000	1.000	0.938	0.958	0.900	0.936
$\text{logit}(p(y_i = 1 \square x_i, z_i)) = \beta_0 + \beta_1 x_i + \beta_2 z_i + c z_i^2$												
0	0.034	0.052	0.036	0.050	0.046	0.054	0.056	0.054	0.054	0.056	0.038	0.062
0.2	0.042	0.054	0.054	0.050	0.058	0.054	0.050	0.058	0.062	0.048	0.044	0.062
0.4	0.050	0.060	0.052	0.054	0.052	0.058	0.054	0.054	0.046	0.058	0.052	0.054
0.6	0.048	0.054	0.048	0.056	0.044	0.068	0.032	0.058	0.048	0.060	0.058	0.066
0.8	0.056	0.058	0.062	0.054	0.058	0.054	0.058	0.054	0.044	0.064	0.062	0.058
1.0	0.062	0.054	0.068	0.068	0.062	0.066	0.058	0.060	0.064	0.062	0.068	0.068

Table 6

Rejection rates for CM_2 at the 5% significance level for a logistic regression model. The first half shows the results for the first scenario where the misspecification is due to x_i and the second half of the table is for the second scenario where the misspecification is due to z_i

	MCAR			MAR		
c	**20%**	**40%**	**60%**	**20%**	**40%**	**60%**
$\mathrm{logit}(p(y_i = 1 \square x_i, z_i)) = \beta_0 + \beta_1 x_i + \beta_2 z_i + c x_i^2$						
0	0.052	0.044	0.042	0.056	0.056	0.058
0.2	0.200	0.164	0.166	0.224	0.216	0.204
0.4	0.480	0.322	0.306	0.446	0.336	0.322
0.6	0.736	0.604	0.584	0.740	0.628	0.592
0.8	0.800	0.774	0.764	0.856	0.776	0.778
1.0	0.912	0.884	0.876	0.910	0.876	0.870
$\mathrm{logit}(p(y_i = 1 \square x_i, z_i)) = \beta_0 + \beta_1 x_i + \beta_2 z_i + c z_i^2$						
0	0.054	0.056	0.058	0.058	0.054	0.058
0.2	0.172	0.066	0.056	0.180	0.066	0.068
0.4	0.324	0.222	0.090	0.326	0.210	0.092
0.6	0.562	0.306	0.118	0.540	0.288	0.144
0.8	0.642	0.442	0.176	0.666	0.454	0.174
1.0	0.884	0.508	0.200	0.856	0.544	0.242