Iterated imputation estimation for generalized linear models with missing response and covariate values

Fang Fang^a, Jun Shao^{a,b,*}

^a*School of Statistics, East China Normal University, 500 Dongchuan Road, Shanghai, 200241, China.*

^b*Department of Statistics, University of Wisconsin, 1300 University Ave., Madison, Wisconsin 53706, U.S.A.*

Abstract

A new approach named as the iterated imputation estimation is proposed for parameter estimation in generalized linear models with missing values in both response and covariates and data are missing at random. The proposed approach is much faster and easier to implement than the method of maximum likelihood or weighted estimating equation. It can be applied by directly using any existing software package for generalized linear models and treating the imputed values as observed in each iteration, which brings great convenience in programming. Theoretical results for the algorithm convergence of the iterated imputation estimation and the asymptotic distribution of the proposed estimator are obtained. Simulation studies and an illustrative example show that the iterated imputation estimation works quite well considering the trade-off between computational burden and estimation efficiency compared with the maximum likelihood estimation.

Keywords: arbitrary missing pattern, iteration convergence, imputation, maximum likelihood, missing at random, missing covariate

1. Introduction

Missing data is a common phenomenon in many applications in areas such as clinical trials, economics, sample surveys, and social sciences. Con-

*Corresponding author.

Email addresses: ffang@sfs.ecnu.edu.cn (Fang Fang), shao@stat.wisc.edu (Jun Shao)

ventional statistical methods cannot be directly applied to incomplete data in most situations. We consider regression problems with both response and covariate variables having missing values. Most existing methods dealt with either missing response or missing covariate data (see, e.g., Little, 1992; Robins, Rotnitzky and Zhao, 1994; Lipsitz, Ibrahim and Zhao, 1999; Fitzmaurice et al., 2001; Tang, Little and Raghunathan, 2003; Ibrahim et al., 2005). For linear regression models with missing response and covariate data, Chen et al. (2008) discussed theoretical properties for inference using maximum likelihood estimation via EM algorithm when the data is missing at random; Shao (2013) proposed some asymptotically unbiased and consistent estimators via direct estimation or imputation. For linear mixed models and generalized linear mixed models, Stubbendick and Ibrahim (2003, 2006) assumed full parametric models when the missing data mechanism is nonignorable and estimated the parameters by the maximum likelihood method. For longitudinal data when both responses and covariates are missing at random, Shardell and Miller (2008) and Chen, Yi and Cook (2010) proposed several estimation methods based on inverse probability weighted estimating equations, in which a parametric model for the missing probability needs to be correctly specified.

In this paper, we consider a generalized linear model (GLM) involving n sampled subjects with independent and identically distributed data (y_i, x_i, z_i) , $i = 1, \dots, n$, where y_i is the outcome or response from subject i , x_i and z_i are p - and q -dimensional random vectors of covariates, y_i and some components of x_i may be missing, and z_i has no missing data. We assume missing at random (MAR), i.e., the probabilities of missing y_i and components of x_i only depend on z_i and x_i^{obs} , the observed part of x_i (Little and Rubin, 2002), but no parametric forms of the probabilities are assumed and the missing data mechanisms of y_i and x_i can be correlated each other. There are many methods for handling missing data under MAR (e.g., Little and Rubin, 2002; Kim and Shao, 2013). The simplest method is the complete case analysis that ignores the sampled subjects with incomplete data and applies the conventional statistical method. This approach is valid under the GLM but is quite inefficient if the size of incomplete data is large, especially when x_i is multivariate.

Under MAR, the maximum likelihood estimation (MLE) can be adopted in a similar way to Lipsitz, Ibrahim and Zhao (1999) and Chen et al. (2008). The MLE specifies conditional distributions of y_i given x_i and z_i , $p(y_i|x_i, z_i, \beta)$, and x_i given z_i , $p(x_i|z_i, \alpha)$, where β is the unknown regression parameter vec-

tor of interest and α is an unknown vector of nuisance parameters. The parameters can be consistently estimated by maximizing the observed likelihood with a Monte Carlo EM algorithm (Lipsitz, Ibrahim and Zhao, 1999; Ibrahim et al., 2005). In iteration $t + 1$ of the algorithm, a Monte Carlo sample of size L needs to be generated from the distribution $p(x_i^{mis}|y_i, x_i^{obs}, z_i, \hat{\beta}^{(t)}, \hat{\alpha}^{(t)})$ or $p(y_i, x_i^{mis}|x_i^{obs}, z_i, \hat{\beta}^{(t)}, \hat{\alpha}^{(t)})$ for every subject i with missing covariate values (response can be observed or missed), where x_i^{mis} is the missing components of x_i and $(\hat{\beta}^{(t)}, \hat{\alpha}^{(t)})$ is the parameter vector estimated in iteration t . The distribution $p(x_i^{mis}|y_i, x_i^{obs}, z_i, \hat{\beta}^{(t)}, \hat{\alpha}^{(t)})$ does not have an explicit form (due to the nonlinearity in GLMs) and sampling techniques such as the Gibbs sampler and the adaptive rejection algorithm of Gilks and Wild (1992) need to be used, which could be quite time-consuming especially when $p > 1$.

Alternatively, one may use the weighted estimating equation (WEE) that was mainly developed by Robins, Rotnizky and Zhao (1994), Zhao, Lipsitz and Lew (1996), and Lipsitz, Ibrahim and Zhao (1999). The WEE with an augmented term has a double robustness property in the sense that the estimation is consistent when either $p(x_i|z_i, \alpha)$ or $\pi_i = p(\delta_i^y = 1, \delta_i^x = 1|x_i^{obs}, z_i)$ is correctly specified, where $\delta_i^y = 1$ if y_i is observed, $\delta_i^y = 0$ if y_i is missing, $\delta_i^x = 1$ if x_i is fully observed and $\delta_i^x = 0$ if some components of x_i are missing. However, the efficiency of WEE relies on the correct specifications of both $p(x_i|z_i, \alpha)$ and π_i ; and modeling π_i is not easy especially when $p > 1$. Furthermore, the WEE has even heavier computational burden than the MLE since a EM-Type algorithm is also needed and the Monte Carlo sampling procedure needs to be done for all the subjects (not just the subjects with $\delta_i^x = 0$ as in the MLE).

Kim (2011) proposed a parametric fraction imputation (PFI) method using the idea of importance sampling and calibration weighting to reduce the computational burden of the MLE and WEE. However, the PFI method needs to carefully arrange the imputed data and apply a weighted GLM, where the weights have to be updated in each iteration. Moreover, the PFI needs a large number of multiple imputations to achieve the consistency and efficiency of the resulting estimators. Although a calibrated PFI was proposed for a moderate number of multiple imputations, it brings more complexity in computation.

In this paper, we propose an iterated imputation estimation (IIE) approach. It has much less computational burden than the MLE and WEE and hence is more applicable for large samples with multivariate x_i having missing values. The IIE can directly use any existing software package for

GLMs by treating the imputed values as observed in each iteration, which brings great convenience in programming. The methodology is described in Section 2. Some theoretical results concerning the convergence of the iterative algorithm and asymptotic distribution of the proposed estimator are presented in Section 3. Some simulations in Section 4 show that the IIE is much faster than the MLE and WEE but its efficiency loss is minor. The capability of the IIE to handle multivariate x_i and a relatively large sample size is also checked. An illustrative example is given in Section 5. Some concluding remarks are given in Section 6. All the proofs are in the Appendix.

2. Iterated Imputation Estimation

2.1. Notation and model

Under a generalized linear model, the conditional distribution of y_i given x_i and z_i has a density $p(y_i|x_i, z_i, \beta, \tau) = \exp\{\tau^{-1}(y_i\eta_i - b(\eta_i)) + c(y_i, \tau)\}$, where b and c are known functions, $\tau > 0$ is an unknown dispersion parameter, $\eta_i = \eta(\beta_x^T x_i + \beta_z^T z_i)$, β_x and β_z are p - and q -dimensional subvectors of β , a^T denotes the transpose of a , and η is a known one-to-one, continuously differentiable function. This includes many useful regression models as special cases, such as normal linear regression, logistic regression, probit regression, Poisson regression, gamma regression, etc. The covariate z_i may contain a constant component so that the corresponding component of β_z is the intercept effect. Since our main interest is to estimate β with missing values, without loss of generality, we assume throughout that $\tau = 1$ and write $p(y_i|x_i, z_i, \beta) = p(y_i|x_i, z_i, \beta, \tau = 1)$. If there is no missing data, β is estimated by maximizing the full data likelihood function, or equivalently, by resolving the full data score equation

$$S(\beta) = \sum_{i=1}^n \frac{\partial \log\{p(y_i|x_i, z_i, \beta)\}}{\partial \beta} = \sum_{i=1}^n g(x_i, z_i, \beta) \{y_i - h(\beta_x^T x_i + \beta_z^T z_i)\} = 0 \quad (1)$$

with $h(\beta_x^T x_i + \beta_z^T z_i) = b'(\eta_i) = E(y_i|x_i, z_i)$, $g(x_i, z_i, \beta) = \{\partial h(\beta_x^T x_i + \beta_z^T z_i) / \partial \beta\} / v_i$, and $v_i = b''(\eta_i) = \text{Var}(y_i|x_i, z_i)$.

When y_i and some components of x_i are missing for some subjects, the equation in (1) cannot be solved. The complete case analysis replaces (1) with

$$S_{cc}(\beta) = \sum_{i=1}^n \delta_i g(x_i, z_i, \beta) \{y_i - h(\beta_x^T x_i + \beta_z^T z_i)\} = 0, \quad (2)$$

where $\delta_i = \delta_i^y \delta_i^x$. The solution is asymptotically valid as $n \rightarrow \infty$ under MAR, but it is inefficient especially when x_i is multivariate since the number of subjects with incomplete data could be large.

Under MAR, the parameter α in $p(x_i|z_i, \alpha)$ can be consistently estimated using methods in the literature (e.g., Little and Rubin, 2002; Kim and Shao, 2013) without any model on the missing x_i -value mechanism. Hence, in what follows we assume that a consistent estimator $\hat{\alpha}$ of α has been obtained using data (x_i^{obs}, z_i) , $i = 1, \dots, n$.

2.2. IIE with univariate missing covariate

To illustrate the idea, we first consider a univariate x_i . When some x_i 's are missing, it is tempting to impute missing x_i 's based on z_i 's and estimate β by resolving the estimating equation

$$S_{cc}(\beta) + \sum_{i=1}^n \delta_i^y (1 - \delta_i^x) g(x_i^{imp}, z_i, \beta) \{y_i - h(\beta_x x_i^{imp} + \beta_z^T z_i)\} = 0, \quad (3)$$

where $S_{cc}(\beta)$ is the score function based on complete cases as defined in (2) and $x_i^{imp} = E(x_i|z_i, \hat{\alpha})$ is an imputed value of x_i . This method is quite easy to implement since we can directly use any existing software package for GLMs by treating x_i^{imp} as observed. Although this method works for linear h (Little, 1992), it produces an inconsistent estimator of β when h is nonlinear and $\beta_x \neq 0$, because

$$E\{y_i - h(\beta_x x_i^{imp} + \beta_z^T z_i) | z_i\} = E\{h(\beta_x x_i + \beta_z^T z_i) | z_i\} - h\{E(\beta_x x_i + \beta_z^T z_i | z_i)\} \neq 0.$$

The problem can be fixed if we could find an imputation procedure such that

$$h(\beta_x x_i^{imp} + \beta_z^T z_i) = \mu_i^{imp}, \quad (4)$$

where $\mu_i^{imp} = E\{h(\beta_x x_i + \beta_z^T z_i) | z_i\}$ is an imputed value of $\mu_i = h(\beta_x x_i + \beta_z^T z_i)$ with a missing x_i . When h is nonlinear, however, equation (4) cannot be achieved unless β is known. Our idea is to replace (4) by an iterative equation

$$h(\hat{\beta}_x^{(t)} x_i^{imp(t)} + \hat{\beta}_z^{(t)T} z_i) = \mu_i^{imp(t)}, \quad t = 1, 2, \dots, \quad (5)$$

i.e., if β_x and β_z are estimated by $\hat{\beta}_x^{(t)}$ and $\hat{\beta}_z^{(t)}$ at the t th step, then we can treat $\hat{\beta}_x^{(t)}$ and $\hat{\beta}_z^{(t)}$ as known values and choose our imputed value $x_i^{imp(t)}$ at the t th iteration to satisfy equation (5) in solving (3) to obtain $\hat{\beta}_x^{(t+1)}$ and $\hat{\beta}_z^{(t+1)}$. This leads to the following *iterated imputation estimation* (IIE) algorithm.

1. For each subject i with $\delta_i^y = 1$ and $\delta_i^x = 0$, sample $\{x_{i,l}, l = 1, \dots, L\}$ from the distribution $p(x_i|z_i, \hat{\alpha})$.
2. Having $\hat{\beta}^{(t)} = (\hat{\beta}_x^{(t)}, \hat{\beta}_z^{(t)})$ at the t th iteration step, compute

$$\mu_i^{imp(t)} = \frac{1}{L} \sum_{l=1}^L h(\hat{\beta}_x^{(t)} x_{i,l} + \hat{\beta}_z^{(t)\top} z_i) \quad (6)$$

and then impute missing x_i by

$$x_i^{imp(t)} = \{h^{-1}(\mu_i^{imp(t)}) - \hat{\beta}_z^{(t)\top} z_i\} / \hat{\beta}_x^{(t)} \quad (7)$$

for all i with $\delta_i^y = 1$ and $\delta_i^x = 0$.

3. Having imputed $x_i^{imp(t)}$ for all i with $\delta_i^y = 1$ and $\delta_i^x = 0$, obtain $\hat{\beta}^{(t+1)}$ by solving (3) with x_i^{imp} replaced by $x_i^{imp(t)}$.
4. Carry out steps 2-3 for $t = 1, \dots, T$ until $\|\hat{\beta}^{(T)} - \hat{\beta}^{(T-1)}\|$ is smaller than a preset small positive constant C .

We should emphasize that the same Monte Carlo sample in step 1 can be used for all t to save computational time. Note that $\mu_i^{imp(t)}$ in (6) is a Monte Carlo approximation to $E\{h(\beta_x x_i + \beta_z^\top z_i) | z_i, \hat{\beta}^{(t)}\}$ that typically does not have an explicit form when h is nonlinear. Equation (7) is simply an inverse version of (5) and imputation has to be carried out at each iteration. When $\hat{\beta}_x^{(t)} \approx 0$, the numerator of the right side of (7) ≈ 0 as well and $x_i^{imp(t)} \approx E(x_i | z_i, \hat{\alpha})$. Although β is estimated iteratively, at each iteration equation (3) with $x_i^{imp} = x_i^{imp(t)}$ can be easily solved using any available software package for GLMs. Thus, compared with the naive imputation method of solving (3) without iteration, the increased computational effort in IIE is the iterations in steps 2-3, which does not involve Monte Carlo sampling.

2.3. IIE with multivariate missing covariate

We now consider multivariate x_i . Let x_i^{imp} be an imputed value of x_i^{mis} , the missing components of x_i . According to our previous discussion, we can obtain a consistent estimator if

$$h\left(\beta_x^\top (x_i^{obs\top}, x_i^{imp\top})^\top + \beta_z^\top z_i\right) = \mu_i^{imp}, \quad (8)$$

where $\mu_i^{imp} = E\{h(\beta_x^\top x_i + \beta_z^\top z_i) | x_i^{obs}, z_i\}$. Unlike the univariate case, the solution of x_i^{imp} to (8) for multivariate x_i is not unique and hence more

constraints are needed. If there is no requirement (8), a natural imputed value of x_i^{mis} is $x_i^{imp(0)} = E(x_i^{mis}|x_i^{obs}, z_i, \hat{\alpha})$. Thus, we can choose x_i^{imp} so that (8) holds and meanwhile it is as close as possible to $x_i^{imp(0)}$. This leads to the following IIE algorithm for multivariate x_i .

1. For subject i with $\delta_i^y = 1$ and $\delta_i^x = 0$, sample $\{x_{i,l}^{mis}, l = 1, \dots, L\}$ from the distribution $p(x_i^{mis}|x_i^{obs}, z_i, \hat{\alpha})$.
2. Having $\hat{\beta}^{(t)} = (\hat{\beta}_x^{(t)}, \hat{\beta}_z^{(t)})$ at the t th iteration step, compute

$$\mu_i^{imp(t)} = \frac{1}{L} \sum_{l=1}^L h \left(\hat{\beta}_x^{(t)\top} (x_i^{obs\top}, x_{i,l}^{mis\top})^\top + \hat{\beta}_z^{(t)\top} z_i \right)$$

and then impute missing x_i^{mis} by minimizing $\|x - x_i^{imp(0)}\|^2$ over x satisfying (8), i.e.,

$$x_i^{imp(t)} = x_i^{imp(0)} + \frac{\hat{\beta}_{x,i}^{mis(t)}}{\|\hat{\beta}_{x,i}^{mis(t)}\|^2} \left\{ h^{-1}(\mu_i^{imp(t)}) - \hat{\beta}_x^{(t)\top} (x_i^{obs\top}, x_i^{imp(0)\top})^\top - \hat{\beta}_z^{(t)\top} z_i \right\}$$

for all i with $\delta_i^y = 1$ and $\delta_i^x = 0$, where $\hat{\beta}_{x,i}^{mis(t)}$ contains the components of $\hat{\beta}_x^{(t)}$ corresponding to components of x_i^{mis} .

3. Having imputed $x_i^{imp(t)}$, obtain $\hat{\beta}^{(t+1)}$ by solving (3) with x_i^{imp} replaced by $(x_i^{obs}, x_i^{imp(t)})$.
4. Carry out steps 2-3 for $t = 1, \dots, T$ until $\|\hat{\beta}^{(T)} - \hat{\beta}^{(T-1)}\|$ is smaller than a preset small positive constant C .

2.4. Discussion

For multivariate x_i , we need to generate random samples from the distribution $p(x_i^{mis}|x_i^{obs}, z_i, \hat{\alpha})$. This task is not difficult in many practical situations: (i) $p(x_i|z_i, \hat{\alpha})$ is multivariate normal, in which each $p(x_i^{mis}|x_i^{obs}, z_i, \hat{\alpha})$ is also multivariate normal with an explicit form and hence can be easily sampled. (ii) Given z_i , components of x_i (could be discrete or continuous) are conditionally independent, in which $p(x_i^{mis}|x_i^{obs}, z_i, \hat{\alpha}) = p(x_i^{mis}|z_i, \hat{\alpha})$ is a product of several marginal conditional distributions that can be easily sampled. (iii) A mixture of (i) and (ii).

The IIE approach has much less computational burden than the MLE and WEE. First, the IIE approach requires generating random samples from the

distribution $p(x_i^{mis}|x_i^{obs}, z_i, \hat{\alpha})$, whereas the MLE and WEE require generating samples from $p(x_i^{mis}|y_i, x_i^{obs}, z_i, \hat{\beta}^{(t)}, \hat{\alpha}^{(t)})$ and $p(y_i, x_i^{mis}|x_i^{obs}, z_i, \hat{\beta}^{(t)}, \hat{\alpha}^{(t)})$, which are more complicated under GLMs. Even for normal (x_i, z_i) , in which case IIE can be easily implemented, the density $p(x_i^{mis}|y_i, x_i^{obs}, z_i, \hat{\beta}^{(t)}, \hat{\alpha}^{(t)})$ does not have an explicit form unless the density $p(y_i|x_i, z_i, \beta)$ is also normal and needs to be handled by complicated sampling techniques. Second, the IIE requires sampling only once for all iterations, whereas sampling has to be repeated in the iterations for the MLE and WEE. In addition, any existing software package for GLMs can be directly applied in the IIE, whereas in the MLE and WEE, the sampled data and weights need to be arranged carefully before applying the weighted GLM programming.

The IIE may be less efficient than the MLE because (A) it simplifies computation by not directly using the observed y_i 's in the imputation of missing covariates and (B) it does not use the observed covariate values corresponding to missing y_i 's in estimation equation (3). For (A), the IIE has a comparable efficiency with the MLE when y_i is a weaker predictor of the missing covariates than the observed covariates, which occurs in many applications. Furthermore, although our imputation does not directly use the observed y_i 's, the IIE leverages the information carried in the responses during the iteration procedure when solving (3). For (B), although observed covariate values corresponding to missing y_i 's are not used in (3), they are utilized in the estimation of $\hat{\alpha}$ so that the efficiency loss may be negligible.

3. Theoretical Properties

3.1. Algorithm convergence

We first study the convergence of the sequence $\{\hat{\beta}^{(t)}, t = 1, 2, \dots, \}$ from the proposed iterated imputation estimation. Each $\hat{\beta}^{(t)}$ actually depends on n , but a subscript is omitted for notation simplicity. In this subsection we temporarily assume that α is known and omit notation α in the following equations, which does not affect our discussion on convergence since $\hat{\alpha}$ is consistent. Define

$$l(\beta|\phi) = \sum_{i=1}^n l_i(\beta|\phi) = \sum_{i=1}^n \delta_i^y \left\{ \delta_i^x l(\beta; y_i, x_i, z_i) + (1 - \delta_i^x) l(\beta; y_i, x_i^{obs}, x_i(\phi), z_i) \right\} \quad (9)$$

$$S(\beta|\phi) = \sum_{i=1}^n S_i(\beta|\phi) = \sum_{i=1}^n \delta_i^y \left\{ \delta_i^x S(\beta; y_i, x_i, z_i) + (1 - \delta_i^x) S(\beta; y_i, x_i^{obs}, x_i(\phi), z_i) \right\}, \quad (10)$$

where $l(\beta; y_i, x_i, z_i) = \log p(y_i|x_i, z_i, \beta)$, $S(\beta; y_i, x_i, z_i) = \partial l(\beta; y_i, x_i, z_i)/\partial \beta$, $\phi = (\phi_x, \phi_z)$, and $x_i(\phi)$ is a function of (x_i^{obs}, z_i, ϕ) defined as the minimizer of $\|x - E(x_i^{mis}|x_i^{obs}, z_i)\|^2$ over x in the space of x_i^{mis} subject to $h(\phi_x^T(x_i^{obs^T}, x^T)^T + \phi_z^T z_i) = E\{h(\phi_x^T x_i + \phi_z^T z_i)|x_i^{obs}, z_i\}$. Let β_0 be the true value of β . Under MAR, i.e., (δ_i^y, δ_i^x) is conditional independent of (y_i, x_i^{mis}) given (x_i^{obs}, z_i) , we have $E\{S(\beta_0|\beta_0)\} = 0$. A detailed proof is given in the Appendix.

In the proposed iterated imputation estimation algorithm, $\hat{\beta}^{(t+1)}$ is a solution to the estimating equation $S(\beta|\hat{\beta}^{(t)}) = 0$. The following theorem shows the convergence properties of the iterated imputation estimation and the proof is given in the Appendix.

Theorem 1: Assume the following conditions:

- (C1) The parameter space of β (and also for ϕ), denoted by B , is a compact and convex subset of R^{p+q} . The true value of β , denoted by β_0 , is an interior point of B .
- (C2) $l(\beta|\beta_0)$ as defined by (9) with $\phi = \beta_0$ is a function of β strictly concave in a neighborhood of β_0 and has a unique maximizer in B . Also, β_0 is the only solution to $E\{S(\beta|\beta_0)\} = 0$, where $S(\beta|\beta_0)$ is given by (10) with $\phi = \beta_0$.
- (C3) $E\{\sup_{\beta, \phi \in B} \|\partial l_1(\beta|\phi)/\partial \phi\|\} < \infty$.
- (C4) $|E\{\partial S_1(\beta_0|\phi)/\partial \phi^T\}_{\phi=\beta_0}| < E\{-\partial S_1(\beta|\beta_0)/\partial \beta^T\}_{\beta=\beta_0}$.

We have the following conclusions.

- (i) For any fixed t , if $\hat{\beta}^{(t)}$ converges to β_0 in probability as $n \rightarrow \infty$, then its one-step update in the iterated imputation estimation algorithm $\hat{\beta}^{(t+1)}$ also converges to β_0 in probability.
- (ii) If $\hat{\beta}^{(1)}$ converges to β_0 in probability, then as $n \rightarrow \infty$,

$$P\left(\|\hat{\beta}^{(t)} - \hat{\beta}\| \geq \|\hat{\beta}^{(t+1)} - \hat{\beta}\| \text{ for all } t \text{ and } S(\hat{\beta}|\hat{\beta}) = 0\right) \rightarrow 1,$$

where $\hat{\beta} = \lim_{t \rightarrow \infty} \hat{\beta}^{(t)}$ for each fixed n .

Theorem 1 tells us that if we choose a consistent estimator $\hat{\beta}^{(1)}$ as the starting point for the iterated imputation estimation algorithm, then, with probability tending to 1, the sequence $\{\hat{\beta}^{(t)}, t \geq 1\}$ converges monotonically to $\hat{\beta}$ for every fixed n and, as $n \rightarrow \infty$, $\hat{\beta}$ is a consistent solution to $S(\beta|\beta) = 0$. Also, each iteration makes the estimated β closer to the limit $\hat{\beta}$. Under our assumed missing data mechanism, we can use the estimator from the complete case analysis as $\hat{\beta}^{(1)}$, which is consistent but may be inefficient.

3.2. Asymptotic distribution and variance estimation

Since $\hat{\beta}$ is a consistent solution to $S(\beta|\beta) = 0$, we can show that $\hat{\beta} - \beta_0$ is asymptotically normal with mean 0 and some covariance matrix, using a standard asymptotic analysis. For statistical inference, we need to estimate the asymptotic covariance matrix of $\hat{\beta}$. Note that $S_i(\beta|\beta)$ also depends on α since it involves the conditional expectation $E(x_i^{mis}|x_i^{obs}, z_i, \alpha)$. So the variation of $\hat{\alpha}$ has to be considered. Rewrite $S_i(\beta|\beta)$ as $S_i(\beta, \alpha)$ and assume that an asymptotically efficient estimator $\hat{\alpha}$ is obtained by solving an estimating equation $U(\alpha) = \sum_{i=1}^n U_i(\alpha) = 0$. A possible choice for $U_i(\alpha)$ is the score function with complete cases, i.e., $U_i(\alpha) = \delta_i^x \frac{\partial \log p(x_i|z_i, \alpha)}{\partial \alpha}$. Then the regular sandwich form of the asymptotic covariance matrix (White, 1982) can be derived. Let $\gamma = (\beta, \alpha)$ and

$$Q(\gamma) = \sum_{i=1}^n Q_i(\gamma) = \sum_{i=1}^n \begin{pmatrix} S_i(\beta, \alpha) \\ U_i(\alpha) \end{pmatrix}.$$

The asymptotic covariance matrix of $\hat{\beta}$ is given by V/n , where V is the upper $(p+q) \times (p+q)$ block of $\left(E \frac{\partial Q_i}{\partial \gamma}\right)^{-1} E(Q_i Q_i^T) \left(E \frac{\partial Q_i^T}{\partial \gamma}\right)^{-1}$. Note that $ES_i U_i^T = 0$. Then, some calculation shows that

$$V = \left(E \frac{\partial S_i}{\partial \beta}\right)^{-1} \left\{ ES_i S_i^T + E \frac{\partial S_i}{\partial \alpha} \left(E \frac{\partial U_i}{\partial \alpha}\right)^{-1} E(U_i U_i^T) \left(E \frac{\partial U_i^T}{\partial \alpha}\right)^{-1} E \frac{\partial S_i^T}{\partial \alpha} \right\} \left(E \frac{\partial S_i^T}{\partial \beta}\right)^{-1}.$$

To estimate V , the relevant expectations are replaced by sample averages respectively and $\gamma = (\beta, \alpha)$ is replaced by $(\hat{\beta}, \hat{\alpha})$.

4. Simulation Study

4.1. Univariate missing covariate

We performed some simulation studies to compare the performance of several methods. In particular, we would like to compare the computational

time and efficiency of the proposed IIE with the MLE and WEE. A univariate x_i was first considered, since the MLE and WEE could be implemented much easier for univariate x_i than multivariate x_i .

The simulation was carried out with 1000 replications and sample size $n = 250$. The dichotomous outcomes y_i were independently generated with $\text{logit}(P(y_i = 1|x_i, z_i)) = 1 - x_i + z_i$, where $z_i \sim N(0, 1)$, and $x_i|z_i \sim N(1 + z_i^2, 2^2)$. The response y_i was observed with probability $\Phi(1 - 0.5z_i)$ and the covariate x_i was observed with probability $\pi_i = \text{expit}(-1 + |z_i| + z_i)$. The missing data proportions for y_i and x_i were about 19% and 56%. The averages of $\delta_i^y \delta_i^x$, $\delta_i^y(1 - \delta_i^x)$, $(1 - \delta_i^y)\delta_i^x$, and $(1 - \delta_i^y)(1 - \delta_i^x)$ were about 33%, 48%, 11% and 8%, respectively. We applied the following 7 methods for comparison: 1. FULL: the standard logistic regression with full data; 2. CC: the complete case analysis, i.e., solving equation (2); 3. NI: the naive imputation, i.e., solving equation (3) with $x_i^{imp} = E(x_i|z_i, \hat{\alpha})$; 4. IIE: the proposed iterated imputation estimation approach; 5. MLE: the maximum likelihood estimation; 6. WEE_c: the weighted estimation equation with the correct $\pi_i = \text{expit}(\gamma_1 + \gamma_2|z_i| + \gamma_3 z_i)$; 7. WEE_m: the weighted estimation equation with a misspecified $\pi_i = \text{expit}(\gamma_1 + \gamma_2 z_i + \gamma_3 z_i^2)$. In methods 4-7, the Monte Carlo sample size L was 1,000, and the convergence criterion was that the absolute difference between the t th and $(t + 1)$ th iterations was less than $C = 10^{-2}$. We used the estimator from the complete case analysis as $\hat{\beta}^{(1)}$ in the iterations.

On the average, the numbers of iterations required by IIE, MLE, WEE_c, and WEE_m were 4.2, 5.3, 7.1 and 7.8. Each replication of the IIE, MLE, WEE_c and WEE_m took about 0.5 second, 40 seconds, 1.5 minutes and 1.6 minutes, respectively, on a Lenovo IdeaCentre B545 personal computer with R programs. Table 1 reports estimates, standard deviations, relative efficiencies (mean square error ratio to the FULL method) and the coverage probabilities of approximate 95% confidence intervals for β based on the 7 methods under consideration. The NI produces biased estimates and the coverage probabilities are much smaller than 95%, while the other methods have small estimation biases and reasonable coverage probabilities. The IIE is more efficient than CC and less efficient than MLE. However, the efficiency loss compared with the MLE is minor. The WEE is almost unbiased even if π_i is misspecified, but it does not have better efficiency than IIE and MLE.

Table 1: Simulation results for univariate x_i

Method	Quantity	$\beta_{int} = 1$	$\beta_x = -1$	$\beta_z = 1$
FULL	EST	1.026	-1.036	1.027
	SD	0.250	0.137	0.241
	REL	1.00	1.00	1.00
	CP	94.8	95.0	94.9
CC	EST	1.090	-1.099	1.089
	SD	0.486	0.274	0.482
	REL	3.85	4.23	4.08
	CP	95.4	96.4	96.3
NI	EST	0.980	-0.903	0.719
	SD	0.356	0.185	0.252
	REL	2.00	2.18	2.42
	CP	90.9	86.0	75.3
IIE	EST	1.044	-1.060	1.050
	SD	0.441	0.268	0.371
	REL	3.10	3.76	2.39
	CP	94.0	96.1	96.3
MLE	EST	1.062	-1.073	1.056
	SD	0.420	0.262	0.362
	REL	2.84	3.71	2.28
	CP	95.2	94.8	96.1
WEE_c	EST	1.087	-1.094	1.075
	SD	0.473	0.308	0.410
	REL	3.65	5.18	2.95
	CP	96.4	95.5	95.1
WEE_m	EST	1.092	-1.098	1.080
	SD	0.477	0.313	0.419
	REL	3.73	5.36	3.09
	CP	94.9	96.1	95.3

FULL: full data

CC: complete case analysis

NI: naive imputation

IIE: iterated imputation estimation

MLE: maximum likelihood estimation

WEE_c: WEE with the correct π_i WEE_m: WEE with a misspecified π_i

EST: parameter estimation

SD: standard deviation

REL: relative efficiency (MSE ratio to FULL)

CP: coverage probabilities in %

4.2. Multivariate missing covariate

To check the performance of the IIE with multivariate x_i and the capability to handle a relatively large sample size, we performed a simulation study with a 4-dimensional x_i and a 6-dimensional z_i (including intercept). The simulation was carried out with 1,000 replications and sample size $n = 2,000$. The dichotomous outcomes y_i were independently generated with $\text{logit}(P(y_i = 1|x_i, z_i)) = -1 + x_{i1} + x_{i2} + x_{i3} + x_{i4} + z_{i1} + z_{i2} + z_{i3} + z_{i4} + z_{i5}$, where $z_{i1}, z_{i2}, z_{i3}, z_{i4}, z_{i5} \sim N(1, 1)$,

$$x_{i1}, x_{i2}, x_{i3}|z_i, \alpha \sim N \left(\begin{pmatrix} \alpha_1 z_{i1} + \alpha_2 z_{i2}^2 \\ \alpha_3 z_{i2} + \alpha_4 z_{i3}^2 \\ \alpha_5 z_{i3} + \alpha_6 z_{i4}^2 \end{pmatrix}, \begin{pmatrix} 1 & \alpha_7 & 0 \\ \alpha_7 & 1 & \alpha_7 \\ 0 & \alpha_7 & 1 \end{pmatrix} \right),$$

x_{i4} is binary with $P(x_{i4} = 1|z_i, \alpha) = \text{expit}(\alpha_8 z_{i4} + \alpha_9 z_{i5}^2)$,

and $\alpha = (0.5, -0.5, 0.5, -0.5, 0.5, -0.5, 0.2, 0.5, -0.5)$. The response y_i was observed with probability $\Phi(z_{i1} + z_{i2}z_{i3} + z_{i4}z_{i5})$. When missing x_i components were considered, x_{i1} was observed with probability $\text{expit}(1.5z_{i1})$, x_{i2} was observed with probability $\text{expit}(1.5z_{i1} + 1.5z_{i2})$, x_{i3} was observed with probability $\Phi(1.5z_{i3})$, and x_{i4} was observed with probability $\Phi(1.5z_{i4} + 1.5z_{i5})$. The average missing data proportions for y_i , x_{i1} , x_{i2} , x_{i3} , and x_{i4} were 13%, 25%, 14%, 20% and 11%, respectively. The averages of $\delta_i^y \delta_i^x$, $\delta_i^y(1 - \delta_i^x)$, $(1 - \delta_i^y)\delta_i^x$, and $(1 - \delta_i^y)(1 - \delta_i^x)$ were about 45%, 42%, 3% and 10%.

Note that this is the scenario (iii) in the first paragraph of Section 2.4 with a multivariate x_i having both discrete and continuous components. In this setting, 1000 replications of the MLE and WEE are pretty time-consuming. Hence, we just considered four methods: FULL, CC, NI and IIE. We use the same Monte Carlo sample size $L = 1,000$ and convergence criteria $C = 0.01$ as in the previous subsection although a larger L and a smaller C can be used since IIE is much faster.

The average number of iterations required for the IIE was about 4 with a maximum of 8. Each IIE replication took less than 1 second. Table 2 reports estimates, standard deviations, standard errors (estimated standard deviations), and coverage probabilities of approximate 95% confidence intervals for β based on the four methods and normal approximation. The standard errors based on full data, the complete case analysis and the naive imputation were obtained by the standard formulas in generalized linear models. The standard errors for the iterated imputation estimation method were estimated using the formula in Section 3.2.

Table 2: Simulation results for multivariate x_i

Parameter	Quantity	Method			
		Full Data	CC	NI	IIE
intercept $\beta_{int} = -1$	$\hat{\beta}_{int}$	-1.015	-1.053	-0.973	-1.027
	SD	0.204	0.522	0.245	0.280
	SE	0.202	0.510	0.239	0.275
	CP	94.9	94.6	94.4	94.2
$\beta_{x1} = 1$	$\hat{\beta}_{x1}$	1.014	1.043	0.924	1.026
	SD	0.080	0.151	0.089	0.108
	SE	0.079	0.146	0.089	0.107
	CP	95.0	94.6	85.0	95.6
$\beta_{x2} = 1$	$\hat{\beta}_{x2}$	1.012	1.037	0.955	1.053
	SD	0.079	0.158	0.086	0.103
	SE	0.080	0.155	0.087	0.103
	CP	95.4	95.2	91.1	95.0
$\beta_{x3} = 1$	$\hat{\beta}_{x3}$	1.013	1.037	0.884	0.984
	SD	0.081	0.154	0.088	0.106
	SE	0.079	0.149	0.086	0.102
	CP	94.4	94.8	69.9	94.4
$\beta_{x4} = 1$	$\hat{\beta}_{x4}$	1.013	1.043	0.912	1.016
	SD	0.183	0.347	0.207	0.235
	SE	0.182	0.339	0.205	0.233
	CP	95.0	94.4	92.4	94.6
$\beta_{z1} = 1$	$\hat{\beta}_{z1}$	1.012	1.040	0.965	1.019
	SD	0.101	0.217	0.110	0.128
	SE	0.099	0.210	0.107	0.124
	CP	95.1	94.1	92.9	93.6
$\beta_{z2} = 1$	$\hat{\beta}_{z2}$	1.014	1.048	0.917	1.012
	SD	0.114	0.239	0.131	0.152
	SE	0.114	0.234	0.130	0.151
	CP	95.2	94.8	89.7	95.4
$\beta_{z3} = 1$	$\hat{\beta}_{z3}$	1.012	1.038	0.992	1.069
	SD	0.113	0.283	0.130	0.151
	SE	0.114	0.277	0.128	0.149
	CP	95.5	94.8	94.5	94.2
$\beta_{z4} = 1$	$\hat{\beta}_{z4}$	1.013	1.037	0.887	0.981
	SD	0.110	0.238	0.126	0.145
	SE	0.109	0.236	0.125	0.144
	CP	95.0	95.5	84.4	96.4
$\beta_{z5} = 1$	$\hat{\beta}_{z5}$	1.011	1.036	0.918	1.019
	SD	0.096	0.199	0.106	0.123
	SE	0.095	0.195	0.105	0.122
	CP	95.3	94.8	86.4	94.0

CC: complete case analysis

NI: naive imputation

IIE: iterated imputation estimation

SD: standard deviation

SE: standard error (estimated standard deviation)

CP: coverage probability in %

The complete case analysis and iterated imputation estimation are approximately unbiased and the naive imputation is biased. The iterated imputation estimation performs much better than the complete case analysis in terms of the standard deviation, since the complete case analysis ignores 55% subjects on average. Our proposed variance estimation for the iterated imputation estimation performs quite well and so is the coverage probability of approximate 95% confidence intervals.

5. An Illustrative Example

To illustrate the usage of our proposed method, we consider a real data example from the United Kingdom General Practice Research Database (Tannen, Weiner and Xie, 2008). The study was to mimic known clinical trial results from the European trial on reduction of cardiac events with Perindopril in patients with stable coronary artery disease (EUROPA, 2003). The study investigated whether angiotensin-converting enzyme inhibitors improve cardiovascular outcomes in patients without congestive heart failure at high risk for cardiovascular disease. In our analysis, we investigate mortality rate for the treatment group with $n = 2,347$ patients. The primary outcome is 5-year death rate and there are seven baseline characteristics. The binary outcome y equals 1 if the patient died within 5 years from the baseline and 0 if the patient survived. The seven baseline covariates are body mass index (BMI), systolic blood pressure (SBP), AGE, SEX (1=male, 0=female), alcohol drinking history (ETOH, 1=yes, 0=no), alcohol addiction (ETOH_M, 1=yes, 0=no), and number of prior events (EVENT) such as acute myocardial infarction, coronary artery bypass surgery, etc. The covariates AGE, SEX, ETOH, ETOH_M, and EVENT are fully observed, and hence these 5 covariates and the intercept form the 6-dimensional vector z . The covariates BMI (x_1) and SBP (x_2) are missing for 540 (23.0%) and 614 (26.2%) of the 2,347 patients, respectively. To apply our method, we impose the MAR assumption, which was assumed in previous publications using this data set (e.g., Tannen, Weiner and Xie, 2009; Yu et al., 2012; Tannen et al., 2013). The outcome y has no missing value in the original data. However, it is quite common that the patients dropout from the study for some reasons in many clinical trials. We manually generate some missing values of y for illustration. The outcome y is assumed to be observed with probability $\text{expit}(-1 + 0.05 * \text{AGE} + 0.02 * \text{EVENT})$. As a result, the outcome y is missing for 217 (9.2%) of the patients. There are 1351 (57.6%) patients having

no missing data, 779 (33.2%) patients having observed response and some missing covariates, 138 (5.9%) patients having missing response and fully observed covariates, and 79 (3.4%) patients having missing response and some missing covariates. We assume that the missing x -value mechanism is conditional independent of (y, x) given z and $\text{logit}(P(y = 1|x, z)) = \beta_x^T x + \beta_z^T z$. We consider four methods: CC, NI, MLE and IIE.

To apply the MLE or IIE, we need to specify the conditional distribution $x|z$. We assume that x_1 and x_2 are independent conditional on z . For x_1 , we assume that $x_1|z \sim N(\alpha_1 + \alpha_2^T g(z), \alpha_3)$, where $g(z)$ is a vector function of the elements of z including pairwise interactions. Using all patients with observed x_1 , we fit a linear model including all the main effects and pairwise interactions of z and keep all the parameters significant at the .20 level of significance. The effects remained in the model are AGE, SEX, EVENT and interaction between AGE and SEX. For x_2 , we assume a gamma model, $x_2|z \sim \text{Gamma}(\alpha_4, \exp(\alpha_5 + \alpha_6^T h(z))/\alpha_4)$, where $h(z)$ is similar to $g(z)$. Based on a complete case gamma regression analysis with all the parameters at .20 significance level, the effects remained in the model are SEX, interactions between AGE and SEX, AGE and ETOH, SEX and EVENT, and ETOH and EVENT. The results are shown in Table 3.

Table 3: Estimates of α in the real data example

Distribution	Parameter	Estimate	Standard Error	P Value
$p(x_1 z, \alpha)$	Intercept	38.501	1.1715	0.000
	AGE	-0.152	0.0164	0.000
	SEX	-6.080	1.4391	0.000
	EVENT	0.103	0.0737	0.163
	AGE*SEX	0.074	0.0211	0.000
$p(x_2 z, \alpha)$	Intercept	4.962	0.0053	0.000
	SEX	-0.105	0.0231	0.000
	AGE*SEX	0.001	0.0003	0.000
	AGE*ETOH	-0.000	0.0002	0.028
	SEX*EVENT	-0.005	0.0032	0.142
	ETOH*EVENT	0.009	0.0030	0.002

The Monte Carlo sample size L for the maximum likelihood and iterated imputation estimation was 1,000. The convergence criteria was that the absolute difference between the t th and $(t + 1)$ th iterations was less

than 10^{-2} . When y , x_1 and x_2 have missing values, in each iteration, the MLE needs to sample from the following 6 different conditional distributions: $p(x_1|y, x_2, z)$, $p(x_2|y, x_1, z)$, $p(x_1, x_2|y, z)$, $p(y, x_1|x_2, z)$, $p(y, x_2|x_1, z)$, and $p(y, x_1, x_2|z)$. The first three distributions do not have explicit forms and need to be handled by adaptive rejective sampling and the third distribution further requires Gibbs sampler. Meanwhile, the IIE only needs to sample once from $p(x_1|z)$ and $p(x_2|z)$ that both have explicit forms. Hence, the implementation of the IIE is much easier than the MLE and the computational time is much shorter. Actually, using the complete case analysis estimates as starting values, the MLE spent 5 iterations and 40 minutes for convergence, while the IIE only spent 4 iterations and 2 seconds.

Table 4 shows estimates, standard errors, and p-values for the regression coefficients based on the complete case analysis, naive imputation, maximum likelihood estimation and iterated imputation estimation. For all four methods, the effects of AGE, SEX, and intercept are highly significant and the coefficient estimates are also close. All four methods also agree that the effects of ETOH, ETOH_M, and BMI are not significant. For the effects of EVENT and SBP, however, there are some disagreements among the four methods. The effect of EVENT is not significant for the complete case analysis, whereas it is significant for the other three methods with similar estimates. This may be due to the inefficiency of the complete case analysis, which has the largest standard error in all cases. The effect of SBP is significant or marginally significant for the maximum likelihood and complete case analysis methods, whereas it is not significant for the iterated imputation estimation and the naive imputation. However, the significant SBP estimates are negative, indicating that patients with lower SBP have higher death rate, which is somewhat contradictory to common knowledge. A closer look at the results in Table 4 reveals that the significant SBP effect obtained by the maximum likelihood estimation is a result of a particularly small standard error, indicating that the standard error for the maximum likelihood estimation may be too optimistic. The same phenomenon exists in the estimation for BMI, where the maximum likelihood estimation has smaller standard error but almost the same estimated effect compared with the iterated imputation estimation.

While the complete case analysis may be misleading or failing in detecting significant effects and the maximum likelihood estimation takes 40 minutes to compute and produces unreasonable result, the naive imputation method produces almost the same results as the iterated imputation estimation in

Table 4: Estimates of β in the real data example

Parameter	Method	Estimate	Standard Error	P Value
Intercept	CC	-6.827	1.3510	0.000
	NI	-7.194	1.1111	0.000
	MLE	-7.222	0.8780	0.000
	IIE	-7.268	1.1250	0.000
AGE	CC	0.069	0.0107	0.000
	NI	0.074	0.0079	0.000
	MLE	0.074	0.0074	0.000
	IIE	0.074	0.0079	0.000
SEX	CC	0.674	0.2026	0.001
	NI	0.672	0.1485	0.000
	MLE	0.673	0.1405	0.000
	IIE	0.673	0.1442	0.000
ETOH	CC	0.116	0.2273	0.609
	NI	-0.031	0.1659	0.850
	MLE	-0.031	0.1615	0.850
	IIE	-0.031	0.1653	0.850
ETOH_M	CC	-0.078	0.2198	0.722
	NI	-0.033	0.1658	0.844
	MLE	-0.032	0.1569	0.837
	IIE	-0.032	0.1609	0.841
EVENT	CC	0.100	0.0604	0.097
	NI	0.129	0.0452	0.004
	MLE	0.129	0.0455	0.005
	IIE	0.129	0.0469	0.006
BMI	CC	0.021	0.0218	0.342
	NI	0.012	0.0183	0.512
	MLE	0.014	0.0125	0.278
	IIE	0.014	0.0185	0.451
SBP	CC	-0.010	0.0059	0.053
	NI	-0.007	0.0050	0.151
	MLE	-0.007	0.0036	0.037
	IIE	-0.007	0.0050	0.150

CC: complete case analysis

NI: naive imputation

MLE: maximum likelihood estimation

IIE: iterated imputation estimation

this example, because of the fact that the covariates with missing values, BMI and SBP, actually do not have any significant effect.

6. Concluding Remarks

We propose an iterated imputation estimation (IIE) approach to deal with the missing response and covariates problem in GLMs. The proposed IIE has much less computational burden than the MLE and WEE and hence is more applicable for multivariate covariates with a large number of incomplete data. In particular, the IIE can directly use any existing software package for GLMs by treating the imputed values as observed in each iteration, which brings a great convenience in programming. Although it is less efficient than the MLE, the IIE is a reasonable alternative to the MLE by considering the tradeoff between efficiency and computational time and simplicity especially when we need to deal with large data sets in the “big data” era.

The consistency of IIE requires a correctly specified $p(x_i|z_i, \alpha)$. However, instead of sampling from the conditional distribution $p(x_i^{mis}|x_i^{obs}, z_i, \hat{\alpha})$ with a parametric form, we can also use the nonparametric distribution estimation for sampling. Since $p(x_i^{mis}|x_i^{obs}, z_i) = p(x_i^{mis}|x_i^{obs}, z_i, \delta_i^x = 1)$ under our missing x_i -value mechanism, we can apply the standard nonparametric product kernel estimator (Li and Racine, 2007) on the complete covariate cases to get a consistent estimator of $p(x_i^{mis}|x_i^{obs}, z_i)$.

Finally, all results in this paper rely on the MAR assumption, which is hard to be verified using data due to missing data. A careful study of the missing data mechanism is needed prior to applying methods relying on MAR. Evidence from previous studies may help.

Appendix A. Proofs

Proof of $E\{S(\beta_0|\beta_0)\} = 0$:

Under the missing at random assumption, (δ_i^y, δ_i^x) is conditional independent of (y_i, x_i^{mis}) given (x_i^{obs}, z_i) . Therefore,

$$\begin{aligned} E[\delta_i^y \delta_i^x S(\beta_0; y_i, x_i, z_i)] &= E[E(\delta_i^y \delta_i^x | x_i^{obs}, z_i) E(S(\beta_0; y_i, x_i, z_i) | x_i^{obs}, z_i)] \\ &= E[E(\delta_i^y \delta_i^x | x_i^{obs}, z_i) E(E(S(\beta_0; y_i, x_i, z_i) | x_i, z_i) | x_i^{obs}, z_i)] \\ &= 0, \end{aligned}$$

and

$$\begin{aligned} & E[\delta_i^y(1 - \delta_i^x)S(\beta_0; y_i, x_i^{obs}, x_i(\beta_0), z_i)] \\ &= E[E(\delta_i^y(1 - \delta_i^x)|x_i^{obs}, z_i)E(S(\beta_0; y_i, x_i^{obs}, x_i(\beta_0), z_i)|x_i^{obs}, z_i)], \end{aligned}$$

where $x_i(\beta_0)$ is a function of $(x_i^{obs}, z_i, \beta_0)$ and satisfies

$$h(\beta_0^T(x_i^{obs}, x_i(\beta_0), z_i)) = E\{h(\beta_{0x}^T x_i + \beta_{0z}^T z_i)|x_i^{obs}, z_i\}.$$

So further we have

$$\begin{aligned} & E(S(\beta_0; y_i, x_i^{obs}, x_i(\beta_0), z_i)|x_i^{obs}, z_i) \\ &= E(g(x_i^{obs}, x_i(\beta_0), z_i)[y_i - h(\beta_0^T(x_i^{obs}, x_i(\beta_0), z_i))]|x_i^{obs}, z_i) \\ &= E(g(x_i^{obs}, x_i(\beta_0), z_i)[y_i - E\{h(\beta_{0x}^T x_i + \beta_{0z}^T z_i)|x_i^{obs}, z_i\}]|x_i^{obs}, z_i) \\ &= g(x_i^{obs}, x_i(\beta_0), z_i)[E(y_i|x_i^{obs}, z_i) - E\{E(y_i|x_i, z_i)|x_i^{obs}, z_i\}] \\ &= 0. \end{aligned}$$

This shows that $E\{S(\beta_0|\beta_0)\} = 0$.

Proof of Theorem 1:

We first establish the following three lemmas.

Lemma 1: Denote $K = 2 \times E \sup_{\beta, \phi \in B} \left\| \frac{\partial l_1(\beta|\phi)}{\partial \phi} \right\|$, which is a positive finite constant. As $n \rightarrow \infty$,

$$P(n^{-1}|l(\beta|\phi_1) - l(\beta|\phi_2)| \leq K\|\phi_1 - \phi_2\|, \forall \beta, \phi_1, \phi_2 \in B) \rightarrow 1.$$

Proof: Between ϕ_1 and ϕ_2 , there exists $\phi^* \in B$ such that

$$\begin{aligned} & P(n^{-1}|l(\beta|\phi_1) - l(\beta|\phi_2)| \leq K\|\phi_1 - \phi_2\|) \\ &= P\left(\left|\frac{1}{n} \frac{\partial l(\beta|\phi^*)}{\partial \phi^T}(\phi_1 - \phi_2)\right| \leq K\|\phi_1 - \phi_2\|\right) \\ &\geq P\left(\left\|\frac{1}{n} \frac{\partial l(\beta|\phi^*)}{\partial \phi}\right\| \leq K\right) \\ &\geq P\left(\frac{1}{n} \sum_{i=1}^n \sup_{\beta, \phi \in B} \left\| \frac{\partial l_i(\beta|\phi)}{\partial \phi} \right\| \leq K\right) \\ &\geq P\left(\left|\frac{1}{n} \sum_{i=1}^n \sup_{\beta, \phi \in B} \left\| \frac{\partial l_i(\beta|\phi)}{\partial \phi} \right\| - E \sup_{\beta, \phi \in B} \left\| \frac{\partial l_1(\beta|\phi)}{\partial \phi} \right\| \right| \leq \frac{K}{2}\right) \\ &\rightarrow 1. \end{aligned}$$

Lemma 2: For any $\epsilon > 0$, define $\tau_n = n^{-1} \left\{ l(\tilde{\beta}|\beta_0) - \sup_{\beta \in B, \|\beta - \tilde{\beta}\| > \epsilon/2} l(\beta|\beta_0) \right\}$, where $\tilde{\beta}$ is the unique maximizer of $l(\beta|\beta_0)$, and

$$\tau = E(l_1(\beta_0|\beta_0)) - \sup_{\beta \in B, \|\beta - \beta_0\| > \epsilon/2} E(l_1(\beta|\beta_0)).$$

Then, τ is a positive constant (depends on ϵ) and $P(\tau_n \geq \tau/2) \rightarrow 1$ as $n \rightarrow \infty$.

Proof: Since $\tilde{\beta}$ converges to β_0 in probability as $n \rightarrow \infty$, τ_n converges to τ in probability as B is compact. Thus, we just need to show that τ is positive. This is guaranteed by condition (C2) along with the facts that the function $E(l_1(\beta|\beta_0))$ is strictly concave in a neighborhood of β_0 and $\partial E(l_1(\beta|\beta_0))/\partial \beta|_{\beta=\beta_0} = E(S_1(\beta_0|\beta_0)) = 0$.

Lemma 3: There exists a sequence of estimator $\{\tilde{\beta}_n\}$ such that $P(S(\tilde{\beta}_n|\tilde{\beta}_n) = 0) \rightarrow 1$ and $\tilde{\beta}_n$ converges to β_0 in probability.

Proof: The proof is trivial and omitted since $E(S(\beta_0|\beta_0)) = 0$.

Proof of (i) in Theorem 1: For any $\epsilon > 0$, define

$A_n = \{\omega \in \Omega : n^{-1} |l(\beta|\phi_1) - l(\beta|\phi_2)| \leq K \|\phi_1 - \phi_2\|, \text{ for any } \beta, \phi_1, \phi_2 \in B\}$ and $B_n = \{\omega \in \Omega : \tau_n(\omega) \geq \frac{\tau}{2}\}$, where Ω is the sample space in which all the random variables are defined. By Lemma 1 and Lemma 2, $P(A_n) \rightarrow 1$ and $P(B_n) \rightarrow 1$ as n goes to infinity. For any $\omega \in A_n \cap B_n$, when $\|\hat{\beta}^{(t)} - \beta_0\| \leq \tau/8K$,

$$n^{-1} |l(\beta|\hat{\beta}^{(t)}) - l(\beta|\beta_0)| \leq K \|\hat{\beta}^{(t)} - \beta_0\| \leq \frac{\tau}{8} \text{ for any } \beta \in B. \quad (\text{A.1})$$

Hence, when $\|\beta - \tilde{\beta}\| > \frac{\epsilon}{2}$, where $\tilde{\beta}$ is given in Lemma 2, we have

$$n^{-1} l(\beta|\hat{\beta}^{(t)}) \leq n^{-1} l(\beta|\beta_0) + \frac{\tau}{8} \quad (\text{A.2})$$

$$\begin{aligned} &\leq \sup_{\beta: \|\beta - \tilde{\beta}\| > \frac{\epsilon}{2}} n^{-1} l(\beta|\beta_0) + \frac{\tau}{8} \\ &= n^{-1} l(\tilde{\beta}|\beta_0) - \tau_n(\omega) + \frac{\tau}{8} \\ &\leq n^{-1} l(\tilde{\beta}|\beta_0) - \frac{\tau}{2} + \frac{\tau}{8} \end{aligned} \quad (\text{A.3})$$

$$\begin{aligned} &\leq n^{-1} l(\tilde{\beta}|\hat{\beta}^{(t)}) + \frac{\tau}{8} - \frac{\tau}{2} + \frac{\tau}{8} \\ &= n^{-1} l(\tilde{\beta}|\hat{\beta}^{(t)}) - \frac{\tau}{4}, \end{aligned} \quad (\text{A.4})$$

where (A.2) and (A.4) follow (A.1), and (A.3) holds since $\omega \in B_n$. Therefore, the maximizer of $l(\beta|\hat{\beta}^{(t)})$, i.e., $\hat{\beta}^{(t+1)}$ must satisfy $\|\hat{\beta}^{(t+1)} - \tilde{\beta}\| \leq \epsilon/2$ when $\omega \in A_n \cap B_n$ and $\|\hat{\beta}^{(t)} - \beta_0\| \leq \tau/8K$. Hence,

$$\begin{aligned} P\left(\|\hat{\beta}^{(t+1)} - \beta_0\| \leq \epsilon\right) &\geq P\left(\|\hat{\beta}^{(t+1)} - \tilde{\beta}\| \leq \frac{\epsilon}{2}, \|\tilde{\beta} - \beta_0\| \leq \frac{\epsilon}{2}\right) \\ &\geq P\left(\omega \in A_n \cap B_n, \|\hat{\beta}^{(t)} - \beta_0\| \leq \frac{\tau}{8K}, \|\tilde{\beta} - \beta_0\| \leq \frac{\epsilon}{2}\right) \\ &\rightarrow 1, \end{aligned} \tag{A.5}$$

where (A.5) holds since $P(A_n) \rightarrow 1$, $P(B_n) \rightarrow 1$, and both $\hat{\beta}^{(t)}$ and $\tilde{\beta}$ converge to β_0 in probability. This finishes the proof of (i) in Theorem 1.

Proof of (ii) in Theorem 1: Since $\hat{\beta}^{(1)}$ converges to β_0 in probability, by (i) in Theorem 1 and Lemma 3, $\hat{\beta}^{(t)}$ converges to β_0 in probability, and $\hat{\beta}^{(t)} - \tilde{\beta}_n$ converges to 0 in probability for any fixed t . Thus, we can use the Taylor expansion of $S(\hat{\beta}^{(t+1)}|\hat{\beta}^{(t)})$ on the point $(\tilde{\beta}_n, \tilde{\beta}_n)$ to get

$$0 = S(\hat{\beta}^{(t+1)}|\hat{\beta}^{(t)}) = S(\tilde{\beta}_n|\tilde{\beta}_n) + \frac{\partial S(\beta^*|\phi^*)}{\partial \beta^T}(\hat{\beta}^{(t+1)} - \tilde{\beta}_n) + \frac{\partial S(\beta^*|\phi^*)}{\partial \phi^T}(\hat{\beta}^{(t)} - \tilde{\beta}_n),$$

where β^* is between $\hat{\beta}^{(t+1)}$ and $\tilde{\beta}_n$, and ϕ^* is between $\hat{\beta}^{(t)}$ and $\tilde{\beta}_n$. Both β^* and ϕ^* converge to β_0 in probability. Therefore

$$\begin{aligned} &\hat{\beta}^{(t+1)} - \tilde{\beta}_n \\ &= \left(-\frac{1}{n} \frac{\partial S(\beta^*|\phi^*)}{\partial \beta^T}\right)^{-1} \left(\frac{1}{n} \frac{\partial S(\beta^*|\phi^*)}{\partial \phi^T}\right) (\hat{\beta}^{(t)} - \tilde{\beta}_n) \\ &\cong \left(-\frac{1}{n} \frac{\partial S(\beta|\beta_0)}{\partial \beta^T}\Big|_{\beta=\beta_0}\right)^{-1} \left(\frac{1}{n} \frac{\partial S(\beta_0|\phi)}{\partial \phi^T}\Big|_{\phi=\beta_0}\right) (\hat{\beta}^{(t)} - \tilde{\beta}_n) \\ &= \left\{I_{(p+q) \times (p+q)} - \left(-\frac{1}{n} \frac{\partial S(\beta|\beta_0)}{\partial \beta^T}\Big|_{\beta=\beta_0}\right)^{-1} \left(-\frac{1}{n} \frac{\partial S(\beta|\beta)}{\partial \beta^T}\Big|_{\beta=\beta_0}\right)\right\} (\hat{\beta}^{(t)} - \tilde{\beta}_n) \\ &\cong \{I_{(p+q) \times (p+q)} - I_2(\beta_0)^{-1} I_1(\beta_0)\} (\hat{\beta}^{(t)} - \tilde{\beta}_n), \end{aligned}$$

where the notation $A_n \cong B_n$ means that $B_n^{-1}A_n = 1 + R_n$ for some R_n which converges to 0 in probability,

$$I_1(\beta_0) = \mathbb{E} \left(-\frac{\partial S_1(\beta|\beta)}{\partial \beta^T} \Big|_{\beta=\beta_0} \right), \text{ and } I_2(\beta_0) = \mathbb{E} \left(-\frac{\partial S_1(\beta|\beta_0)}{\partial \beta^T} \Big|_{\beta=\beta_0} \right).$$

Therefore,

$$\hat{\beta}^{(t)} - \tilde{\beta}_n \cong \{I_{(p+q) \times (p+q)} - I_2(\beta_0)^{-1} I_1(\beta_0)\}^{t-1} (\hat{\beta}^{(1)} - \tilde{\beta}_n),$$

which converges to 0 as t goes to infinity with condition (C4). Then (ii) of Theorem 1 holds with $\hat{\beta} = \tilde{\beta}_n$.

Acknowledgement

The authors would like to thank the reviewers for their valuable comments and suggestions which improved the presentation of this paper. Fang Fang's research was partially supported by Shanghai Nature Science Foundation 15ZR1410300, Program of Shanghai Subject Chief Scientist (14XD1401600), and the 111 Project (B14019). Jun Shao's research was partially supported by the 111 Project (B14019) and US National Science Foundation grant DMS-1305474.

References

- Chen, B., Yi, G. Y. and Cook, R. J. (2010). Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Association* **105**, 336-353.
- Chen, Q., Ibrahim, J.G., Chen, M.-H. and Senchaudhuri, P. (2008). Theory and inference for regression models with missing responses and covariates. *Journal of Multivariate Analysis* **99**, 1302-1331.
- EUROPA (2003). Efficacy of perindopril in reduction of cardiovascular events among patients with stable coronary artery disease: randomized double-blind, placebo-controlled, multicentre trial. *Lancet* **362**, 782-788.
- Fitzmaurice, G. M., Lipsitz, S. R., Molenberghs, G. and Ibrahim, J. G. (2001). Bias in Estimating Association Parameters for Longitudinal Binary Responses with Dropouts. *Biometrics* **57**, 15-21.
- Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337-348.
- Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R. and Herring, A. H. (2005). Missing-Data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association* **100**, 332-346.

- Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika* **98**, 119-132.
- Kim, J. K. and Shao, J. (2013). *Statistical Methods for Incomplete Data Analysis*. Chapman & Hall/CRC, New York.
- Li, Q. and Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton University Press.
- Lipsitz, S. R., Ibrahim, J. G. and Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association* **94**, 1147-1160.
- Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association* **87**, 1227-1237.
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*, second edition. Wiley, New York.
- Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimating of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846-866.
- Shao, J. (2013). Estimation and imputation in linear regression with missing values in both response and covariate. *Statistics and Its Interface* **6**, 361-368.
- Shardell, M. and Miller, R. (2008). Weighted estimating equations for longitudinal studies with death and non-monotone missing time-dependent covariates and outcomes. *Statistics in Medicine* **27**, 1008-1025.
- Stubbendick, A. L. and Ibrahim, J. G. (2003). Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics* **59**, 1140-1150.
- Stubbendick, A. L. and Ibrahim, J. G. (2006). Likelihood based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statistica Sinica* **16**, 1143-1167.
- Tang, G., Little, R. J. A. and Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* **90**, 747-764.

- Tannen, R. L., Weiner, M. G. and Xie, D. (2008). Replicated studies of two randomized trials of angiotensin-converting enzyme inhibitors: further empiric validation of the 'prior event rate ratio' to adjust for unmeasured confounding by indication. *Pharmacoepidemiology and Drug Safety* **17**, 671-685.
- Tannen, R. L., Weiner, M. G. and Xie, D. (2009). Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomized controlled trial findings. *British Medical Journal* **338**, 395-399.
- Tannen, R.L., Xie, D., Wang, X., Yu, M. and Weiner, M. G. (2013). A new "Comparative Effectiveness" assessment strategy using the THIN database: comparison of the cardiac complications of pioglitazone and rosiglitazone. *Pharmacoepidemiology and Drug Safety* **22**, 86-97.
- White, H. (1982). Maximum likelihood estimation under mis-specified models. *Econometrica* **50**, 1-26.
- Yu, M., Xie, D., Wang, X., Weiner, M. G. and Tannen, R.L. (2012). Prior event rate ratio adjustment: numerical studies of a statistical method to address unrecognized confounding in observational studies. *Pharmacoepidemiology and Drug Safety* **21**, 60-68.
- Zhao, L. P., Lipsitz, S. R. and Lew, D. (1996). Regression analysis with missing covariate data using estimating equations. *Biometrics* **52**, 1165-1182.