Yanmei Xie* and Biao Zhang

Empirical Likelihood in Nonignorable Covariate-Missing Data Problems

DOI 10.1515/ijb-2016-0053

Abstract: Missing covariate data occurs often in regression analysis, which frequently arises in the health and social sciences as well as in survey sampling. We study methods for the analysis of a nonignorable covariate-missing data problem in an assumed conditional mean function when some covariates are completely observed but other covariates are missing for some subjects. We adopt the semiparametric perspective of Bartlett et al. (Improving upon the efficiency of complete case analysis when covariates are MNAR. Biostatistics 2014;15:719–30) on regression analyses with nonignorable missing covariates, in which they have introduced the use of two working models, the working probability model of missingness and the working conditional score model. In this paper, we study an empirical likelihood approach to nonignorable covariate-missing data problems with the objective of effectively utilizing the two working models in the analysis of covariate-missing data. We propose a unified approach to constructing a system of unbiased estimating equations, where there are more equations than unknown parameters of interest. One useful feature of these unbiased estimating equations is that they naturally incorporate the incomplete data into the data analysis, making it possible to seek efficient estimation of the parameter of interest even when the working regression function is not specified to be the optimal regression function. We apply the general methodology of empirical likelihood to optimally combine these unbiased estimating equations. We propose three maximum empirical likelihood estimators of the underlying regression parameters and compare their efficiencies with other existing competitors. We present a simulation study to compare the finite-sample performance of various methods with respect to bias, efficiency, and robustness to model misspecification. The proposed empirical likelihood method is also illustrated by an analysis of a data set from the US National Health and Nutrition Examination Survey (NHANES).

Keywords: complete case analysis, efficiency, empirical likelihood, influence function, linear space, missing covariates, missing not at random, projection, regression, residual, unbiased estimating function

1 Introduction

Missing data is an important practical problem commonly found in medical sciences, social sciences, and many other related disciplines. For example, in sample surveys, some sampled individuals do not complete the questionnaire because of non-contact, refusal to respond, or some other reason. In panel surveys and clinical trials, attrition arises when members of the panel or subjects of the longitudinal study group drop out prior to the end of the study and do not return. In the past decade, there has been a great deal of interest in statistical methods for studying missing data problems, including missing response, missing covariate, or both. There are three major missing-data mechanisms, as discussed in Little and Rubin [1]. The first and simplest case is called missing completely at random (MCAR), i.e., missingness does not depend on any observed or missing quantities. The second case is missing at random (MAR), i.e., missingness depends only on the observed quantities. The third case is called missing not at random (MNAR) or nonignorable missing, i.e., missingness depends on the unobserved quantities; this is the most difficult case to handle. Most of

***Corresponding author: Yanmei Xie,** Department of Mathematics and Statistics, The University of Toledo, Toledo, OH 43606, USA, E-mail: Yanmei.Xie@rockets.utoledo.edu
Biao Zhang, Department of Mathematics and Statistics, The University of Toledo, Toledo, OH 43606, USA, E-mail: bzhang@utnet.utoledo.edu

the past works have mainly focused on ignorable missing data problems involving MCAR and MAR missing data mechanisms. For a review on the analyses of ignorable missing data problems, we refer the readers to Little and Rubin [1]. In this paper, we focus on the study of a nonignorable covariate-missing data problem motivated by the alcohol and blood pressure data from the US National Health and Nutrition Examination Survey (NHANES) [2] as described below.

The NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States. This survey produces vital and health statistics for the Nation and provides an important basis for public health interventions. For the Demographics Data, the Examination Data, and the Questionnaire Data from the 2003–2004 NHANES, age and gender were fully observed, but the other variables such as education, systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI), income, and alcohol consumption had missing values. As argued by Little and Zhang [3], it is plausible and reasonable to assume that the missingness in education, BMI and the two blood pressure measures was missing at random (MAR) due to missed visits, but missingness of household income was thought more likely to be missing not at random (MNAR), since the propensity to respond to the income question in sample surveys is likely to depend on and possibly determined by the underlying value of the income variable, with those with low or high income generally considered less likely to respond than others. In a recent and related work, Bartlett et al. [4] have considered data on alcohol consumption and blood pressure from the 2003–2004 NHANES and have argued that missingness in alcohol consumption is likely to depend largely on alcohol consumption itself and is thus missing not at random (MNAR). To study the effect of socio-economic status (income and education) on blood pressure measures, Little and Zhang [3] have performed the regression analysis of blood pressure on income and education, adjusting for age, gender and body mass index after conditioning on the subsample of cases with household income fully observed. In addition, Bartlett et al. [4] have considered the regression analysis on the dependence of systolic blood pressure on the average number of alcoholic drinks consumed per day, with adjustment for age and body mass index. Both regression analyses can be regarded as nonignorable covariate-missing data problems.

Under the ignorable covariate-missing data mechanism, there are a variety of statistical methods available in the literature on analysis of missing data, including (a) complete case analysis (CCA), (b) inverse probability weighting (IPW) in the spirit of Horvitz and Thompson [5], which weights each of the complete data by the inverse of its inclusion probability, and (c) augmented inverse probability weighted complete-case (AIPWCC) proposed by Robins et al. [6], which incorporates the incomplete data to increase efficiency while reducing the bias. By contrast, under the nonignorable covariate-missing data mechanism, since the probability of nonignorable missingness in a covariate variable is dependent on the value of that variable, it is generally difficult to specify a model for the nonignorable covariate-missing data mechanism. To overcome this difficulty to handle nonignorable covariate-missing data problems, Bartlett et al. [4] have proposed an augmented complete case analysis by modeling the probability of missingness on the fully observed variables under the assumption that missingness in a covariate depends on the value of that covariate, but is conditionally independent of the outcome variable. Under this missing not at random (MNAR) mechanism, the commonly used CCA approach is valid and gives rise to consistent estimates, but is inefficient in that it does not make use of all observed information by disregarding all incomplete cases. Like the AIPWCC approach, the augmented CCA approach utilizes all the observed data by drawing on the information available from both complete and incomplete cases and thus improves upon the efficiency of CCA through specification of an additional model for the probability of missingness given the fully observed variables. The estimating function used in Bartlett et al. [4] can be viewed as the difference of two separate estimating functions, but other linear combinations of these two estimating functions are also possible. Thus, the question arises as how to combine the two sets of estimating functions adopted in the augmented complete case analysis of Bartlett et al. [4].

Our objective in this paper is to explore the use of empirical likelihood methods in the nonignorable covariate-missing data problem described in Little and Zhang [3] and Bartlett et al. [4] to effectively incorporate incomplete cases into the analysis of covariate-missing data and thus to improve estimation efficiency

when the data are not missing at random. We propose to construct a system of unbiased estimating equations, where there are more equations than unknown parameters of interest. These estimating equations are always unbiased for any given working regression function as long as the working probability model of missingness is correctly specified. One useful feature of these unbiased estimating equations is that they naturally incorporate the incomplete data into the data analysis, making it possible to seek efficient estimation of the parameter of interest even when the working regression function is possibly not specified to be the optimal regression function. We apply the general methodology of empirical likelihood to effectively and optimally combine these unbiased estimating equations when the number of estimating equations is greater than the number of parameters of interest. Furthermore, we propose to study maximum empirical likelihood estimation of the underlying regression parameters based on the aforementioned system of unbiased estimating equations. Moreover, we apply the proposed empirical likelihood estimators to the analysis of the alcohol and blood pressure data from the 2003–2004 NHANES.

As a nonparametric method, empirical likelihood was introduced by Owen [7, 8] for constructing confidence intervals or regions for the mean and other parameters. One advantage of using empirical likelihood is that the shape of the confidence region is determined automatically by the data. Qin and Lawless [9] have demonstrated that the empirical likelihood method can be used to solve estimating equations when the number of estimating equations exceeds the number of parameters. For theoretical developments of the empirical likelihood method, we refer readers to Hall and La Scala [10] and Owen [11]. For the analysis of missing data using the empirical or nonparametric likelihood method, see, for example, Wang and Rao [12, 13], Chen et al. [14], Liang et al. [15], Stute et al. [16], Qin and Zhang [17], and Wang and Dai [18], among others.

This paper is organized as follows. In Section 2, we propose three unbiased estimating functions and study maximum empirical likelihood estimation of the underlying regression parameters. In Section 3, we study efficiency comparison between the proposed and existing estimators. In Section 4, the proposed methodology is illustrated using a data set from the US National Health and Nutrition Examination Survey (NHANES). Simulation results are presented in Section 5 and concluding remarks are given in Section 6. Proofs of the main theoretical results are delegated to an Appendix in Section 7.

2 Methodology

In clinical trials and sample surveys, completely observed covariate information is often not available for every subject. We consider a regression analysis of an outcome variable Y on a vector Z of covariates which are always observed and a vector X of covariates which can be missing for some subjects. Our interest lies in the estimation of the unknown $p \times 1$ vector of regression parameters, β, characterized by the conditional mean model

$$Y = \mu(X, Z, \beta) + \varepsilon \tag{1}$$

for a specified, possibly nonlinear, link function $\mu(X, Z, \beta)$, where the random error ε satisfies $E(\varepsilon|X, Z) = 0$ so that $E(Y|X, Z) = \mu(X, Z, \beta)$, and the joint distribution of the regressors (X, Z) is left completely unspecified. Conditional mean models include familiar linear and logistic regression models. The regression analysis with missing covariate data has received much attention recently in the statistical literature; see, for example, Little and Rubin [1], Little and Schluchter [19], Ibrahim [20], Little [21], Ibrahim and Lipsitz [22], Lipsitz and Ibrahim [23], and Ibrahim et al. [24]. Let D denote a binary non-missing indicator such that $D = 1$ if the covariate X is observed and $D = 0$ if X is missing. When the covariate vector X is missing at random, the probability of X being observed is commonly modeled by a parametric model for the propensity score $P(D = 1|Y = y, Z = z) = P(D = 1|Y = y, X = x, Z = z)$; the parameters in this parametric propensity score model can be consistently estimated by the fully observed data on (Y, Z, D). The validity of the commonly used IPW and AIPWCC methods is contingent upon the correct specification of a propensity score model for the missingness mechanism $P(D = 1|Y = y, Z = z)$ under the MAR assumption.

Let $(Y_1, X_1, Z_1, D_1), \ldots, (Y_n, X_n, Z_n, D_n)$ denote a random sample of (Y, X, Z, D). We are interested in estimating the regression parameter β based on the observed data $\{(Y_i, D_i X_i, Z_i, D_i),\ i = 1, \ldots, n\}$ available for analysis. Let $U(Y, X, Z, \beta)$ denote a $p \times 1$ specific full-data unbiased estimating function for β; a common choice for $U(Y, X, Z, \beta)$ is $A(X, Z, \beta)\{Y - \mu(X, Z, \beta)\}$ under the conditional mean model (1), where $A(X, Z, \beta)$ is a vector of p known functions of (X, Z) and β. In the absence of missing data, since $E\{U(Y, X, Z, \beta)\} = 0$, a full-data estimator $\hat{\beta}_f$ of β solves the full-data estimating equation: $\sum_{i=1}^{n} U(Y_i, X_i, Z_i, \beta) = 0$. With missing covariate data, a complete-case estimator $\hat{\beta}_c$ of β solves the complete-case estimating equation

$$\sum_{i=1}^{n} D_i U(Y_i, X_i, Z_i, \beta) = 0; \tag{2}$$

it is well known that $\hat{\beta}_c$ can be biased when X is not missing completely at random (MCAR). Under the nonignorable missing-data mechanism in which the missingness of X is dependent on the value of X itself, we cannot directly estimate the underlying missingness mechanism $P(D = 1|Y = y, X = x, Z = z)$ on the basis of the observed data $\{(Y_i, D_i X_i, Z_i, D_i),\ i = 1, \ldots, n\}$ due to missingness of X_i whenever $D_i = 0$. As a result, the propensity score approach based on modeling $P(D = 1|Y = y, X = x, Z = z)$ is not applicable to the estimation of β under model (1) with nonignorable missingness of X; in particular, the MAR-based IPW and AIPWCC methods cannot be applied to estimate β in this context.

Under the assumption that the missingness of X is independent of the outcome variable Y conditional on covariates X and Z, namely, D and Y are conditionally independent given X and Z or $D \perp Y|X, Z$, the complete-case estimator $\hat{\beta}_c$ is a consistent estimator of the regression parameter β. Although the complete-case analysis using the estimating eq. (2) provides valid inferences for β under the conditional independence assumption of $D \perp Y|X, Z$, the resulting complete-case estimator $\hat{\beta}_c$ is inefficient in that it fails to draw on the observed information contained in the incomplete cases. To attempt to improve upon the efficiency of the complete-case analysis, Bartlett et al. [4] have proposed an additional model for the probability of missingness given the fully observed outcome variable Y and the fully observed covariate Z by making the same conditional independence assumption for missingness as in the complete-case analysis. Specifically, under the conditional independence assumption of $D \perp Y|X, Z$, the probability of X being observed is described by the probability model $\pi_0(y, z) = P(D = 1|Y = y, Z = z)$. Let $\pi(y, z; \gamma)$ represent a working probability model for $\pi_0(y, z)$, where $\pi(y, z; \gamma)$ is a strictly positive function of (y, z, γ) and γ is a $q \times 1$ vector parameter. Let γ_0 denote the truth of γ; if the working probability model is correctly specified, then $\pi(y, z; \gamma_0) = \pi_0(y, z) = P(D = 1|Y = y, Z = z)$. Since the model $\pi(y, z; \gamma)$ for $P(D = 1|Y = y, Z = z)$ only involves fully observed variables Y and Z, we can estimate γ by the method of maximum likelihood based on the fully observed data $(Y_1, Z_1, D_1), \ldots, (Y_n, Z_n, D_n)$. Indeed, we can estimate γ by the maximum likelihood estimator $\hat{\gamma}_n$, which maximizes the binomial likelihood:

$$L_B(\gamma) = \prod_{i=1}^{n} \{\pi(Y_i, Z_i; \gamma)\}^{D_i} \{1 - \pi(Y_i, Z_i; \gamma)\}^{1-D_i}$$

and is a solution to the following system of score equations:

$$U_n(\gamma) \equiv \frac{1}{n} \frac{\partial \log L_B(\gamma)}{\partial \gamma} = \frac{1}{n} \sum_{i=1}^{n} \frac{\{D_i - \pi(Y_i, Z_i; \gamma)\} \pi_\gamma(Y_i, Z_i; \gamma)}{\pi(Y_i, Z_i; \gamma)\{1 - \pi(Y_i, Z_i; \gamma)\}} = 0, \tag{3}$$

where $\pi_\gamma(y, z; \gamma) = \partial \pi(y, z; \gamma)/\partial \gamma$. Although the additional model $\pi(y, z; \gamma)$ for the probability of missingness is not a model for the underlying missingness mechanism $P(D = 1|Y = y, X = x, Z = z)$ under the conditional independence assumption, it allows us to develop an alternative approach to estimation of β. Under the conditional independence assumption of $D \perp Y|X, Z$ and based on the working probability model $\pi(y, z; \gamma)$ for the missingness of X, Bartlett et al. [4] have proposed an augmented CCA estimation method for estimating β. To describe their method, we posit a working conditional score model for

$m_0(Y, Z; \beta, \xi_0) = E\{U(Y, X, Z, \beta)|Y, Z, D = 1\}$ in terms of a parametric model $m(Y, Z; \beta, \xi)$ with the same dimension as β, where $m(Y, Z; \beta, \xi)$ is an arbitrary, user specified, and possibly data dependent working regression function of $(Y, Z; \beta, \xi)$ and ξ is an additional finite-dimensional parameter with its true value denoted as ξ_0. For each fixed (γ, ξ), let $\hat{\beta}_{\text{acc}}(\gamma, \xi)$ be defined as a solution to the system of augmented complete-case estimating equations:

$$\sum_{i=1}^{n}\left[D_i U(Y_i, X_i, Z_i, \beta) - \{D_i - \pi(Y_i, Z_i; \gamma)\}m(Y_i, Z_i; \beta, \xi)\right] = 0. \tag{4}$$

Then the augmented complete case (ACC) estimators of β are given, respectively, by $\hat{\beta}_{\text{acc0}} = \hat{\beta}_{\text{acc}}(\gamma_0, \hat{\xi}_n)$ when γ_0 is known and $\hat{\beta}_{\text{acc}} = \hat{\beta}_{\text{acc}}(\hat{\gamma}_n, \hat{\xi}_n)$ when γ_0 is unknown, where $\hat{\xi}_n$ is an estimator of ξ based on the complete data $\{(Y_i, X_i, Z_i) : D_i = 1, i = 1, \ldots, n\}$; for example, $\hat{\xi}_n$ may be chosen to be the possibly nonlinear least squares estimator of ξ from the regression of $U(Y_i, X_i, Z_i; \hat{\beta}_c)$ on (Y_i, Z_i) among the complete cases $\{i : D_i = 1\}$. We assume throughout that $\hat{\xi}_n$ is $\sqrt{n}$-consistent for some constant ξ^* or $\hat{\xi}_n - \xi^* = O_p(n^{-1/2})$ under suitable regularity conditions. When the working probability model $\pi(y, z; \gamma)$ is correctly specified for $P(D = 1|Y = y, Z = z)$, Bartlett et al. [4] have shown under suitable regularity conditions that the augmented complete case (ACC) estimator $\hat{\beta}_{\text{acc}}$ is consistent and asymptotically normal for any working regression function $m(Y, Z; \beta, \xi)$. They have also shown that for a given choice of $U(Y, X, Z, \beta)$, the optimal choice for the working regression function $m(Y, Z; \beta, \xi)$ is given by

$$m_0(Y, Z; \beta, \xi_0) = E\{U(Y, X, Z, \beta)|Y, Z, D = 1\}. \tag{5}$$

When the working regression function $m(y, x; \beta, \xi)$ is correctly specified to be the optimal regression function in eq. (5), we have $\xi^* = \xi_0$. In this case, the augmented complete case estimator $\hat{\beta}_{\text{acc}}$ improves upon the efficiency of the complete case estimator $\hat{\beta}_c$ by minimizing the variance of $\hat{\beta}_{\text{acc}}$ among all the choices of the working regression function $m(Y, Z; \beta, \xi)$ for any given choice of $U(Y, X, Z, \beta)$. In addition, Bartlett et al. [4] have proposed a modified estimator of β, which is guaranteed to be at least as efficient as the complete case estimator $\hat{\beta}_c$ for any choice of the working regression function $m(Y, Z; \beta, \xi)$.

To explore the empirical likelihood approach to the estimation of the regression parameter β under model (1) with covariate X subject to nonignorable missing, we are motivated by eqs (3) and (4) to define the following estimating functions on the basis of the working probability model $\pi(y, z; \gamma)$ and working regression model $m(y, z; \beta, \xi)$:

$$g_1(Y, Z, D, X; \beta) = DU(Y, Z, X; \beta), g_2(Y, Z, D; \beta, \gamma, \xi) = \{D - \pi(Y, Z; \gamma)\} m(Y, Z; \beta, \xi),$$
$$g_3(Y, Z, D; \gamma) = \frac{\{D - \pi(Y, Z; \gamma)\}\pi_\gamma(Y, Z; \gamma)}{\pi(Y, Z; \gamma)\{1 - \pi(Y, Z; \gamma)\}}, g(Y, Z, D, X; \beta, \gamma, \xi) = \begin{pmatrix} g_1(Y, Z, D, X; \beta) \\ g_2(Y, Z, D; \beta, \gamma, \xi) \end{pmatrix},$$
$$G(Y, Z, D, X; \beta, \gamma, \xi) = \begin{pmatrix} g_1(Y, Z, D, X; \beta) \\ g_2(Y, Z, D; \beta, \gamma, \xi) \\ g_3(Y, Z, D; \gamma) \end{pmatrix}.$$

The first estimating function $g_1(Y, Z, D, X; \beta)$ is identical to the complete case estimating function in eq. (2) and has mean zero or $E_F\{g_1(Y, Z, D, X; \beta_0)\} = 0$ when the conditional independence assumption of $D \perp Y|X, Z$ holds, where $(Y, Z, D, X) \sim F$. The second estimating function $g_2(Y, Z, D; \beta, \gamma, \xi)$ involves both subjects with X observed and those with X missing and has mean zero or $E_F\{g_2(Y, Z, D; \beta, \gamma_0, \xi)\} = 0$ for any working regression model $m(y, z; \beta, \xi)$ provided that the probability model $\pi(y, z; \gamma)$ for $P(D = 1|Y = y, Z = z)$ is correctly specified, since then $E\{D - \pi(Y, Z; \gamma_0)|Y, Z\} = 0$. The third estimating function $g_3(Y, Z, D; \gamma)$ is the score function of γ and is optional when the true probability model $\pi(y, z; \gamma_0)$ for $P(D = 1|Y = y, Z = z)$ is completely known. Furthermore, when the working propensity score $\pi(y, z; \gamma)$ is correctly specified for $P(D = 1|Y = y, Z = z)$, we have $E_F\{g_3(Y, Z, D; \gamma_0)\} = 0$.

2.1 Maximum empirical likelihood estimation of β when γ_0 is known

In this subsection, we assume that the true probability model $\pi(y, z; \gamma_0)$ that we have adopted for modeling $P(D = 1|Y = y, Z = z)$ is completely known. When γ_0 is known, we impose constraints on $g_1(Y, Z, D, X; \beta)$ and $g_2(Y, Z, D; \beta, \gamma_0, \hat{\xi}_n)$ and maximize the nonparametric likelihood $L_F = \prod_{i=1}^n p_i$ subject to the constraints

$$\sum_{i=1}^n p_i = 1, \qquad p_i \geq 0, \qquad \sum_{i=1}^n p_i g(Y_i, Z_i, D_i, X_i; \beta, \gamma_0, \hat{\xi}_n) = 0,$$

where $p_i = dF(Y_i, Z_i, D_i, X_i)$ for $i = 1, \ldots, n$ and $\hat{\xi}_n$ is some $\sqrt{n}$-consistent estimator of ξ based on the complete cases $\{i : D_i = 1\}$. For fixed β, an application of the Lagrange multipliers method shows that the maximum value of L_F is attained at

$$\tilde{p}_{1i}(\beta) = \frac{1}{n} \frac{1}{1 + \hat{\lambda}_{1n}^\tau g(Y_i, Z_i, D_i, X_i; \beta, \gamma_0, \hat{\xi}_n)},$$

for $i = 1, \ldots, n$, where the Lagrange multiplier $\hat{\lambda}_{1n} = \hat{\lambda}_{1n}(\beta)$ is determined by

$$U_{1n}(\lambda) \equiv U_{1n}(\lambda, \beta, \hat{\xi}_n) = \frac{1}{n} \sum_{i=1}^n \frac{g(Y_i, Z_i, D_i, X_i; \beta, \gamma_0, \hat{\xi}_n)}{1 + \lambda^\tau g(Y_i, Z_i, D_i, X_i; \beta, \gamma_0, \hat{\xi}_n)} = 0. \tag{6}$$

After profiling the p_{1i}'s, the profile log likelihood function of β is given by

$$\ell_1(\beta) = \sum_{i=1}^n \log p_i = -\sum_{i=1}^n \log[1 + \hat{\lambda}_{1n}^\tau(\beta) g(Y_i, Z_i, D_i, X_i; \beta, \gamma_0, \hat{\xi}_n)] - n \log n. \tag{7}$$

Let $\hat{\beta}_{\text{el1}}$ denote the maximum empirical likelihood estimator of β, which maximizes $\ell_1(\beta)$. The asymptotic distribution of $\hat{\beta}_{\text{el1}}$ is established in the following theorem.

Theorem 1 : Under the regularity conditions (A1)–(A6) stated in the Appendix, if the true probability model $\pi(y, z; \gamma_0) = \pi_0(y, z) = P(D = 1|Y = y, Z = z)$ is known, we can write

$$\hat{\beta}_{\text{el1}} - \beta_0 = -\frac{1}{n} \sum_{i=1}^n C_1^{-1}\{g_1(Y_i, Z_i, D_i, X_i; \beta_0) - C_4 C_5^{-1} g_2(Y_i, Z_i, D_i; \beta_0, \gamma_0, \xi^*)\} + o_p(n^{-1/2}),$$

where C_1, C_4, C_5 are defined in eqs (14) and (19), respectively. As a result, $\sqrt{n}(\hat{\beta}_{\text{el1}} - \beta_0) \xrightarrow{d} N(0, \Sigma_{\text{el1}})$, where

$$\Sigma_{\text{el1}} = C_1^{-1} \text{Var}\{g_1(Y, Z, D, X; \beta_0) - C_4 C_5^{-1} g_2(Y, Z, D; \beta_0, \gamma_0, \xi^*)\}(C_1^{-1})^\tau.$$

2.2 Maximum empirical likelihood estimation of β when γ_0 is unknown

In practice, the true probability model $\pi(y, z; \gamma_0)$ is often unknown because it is difficult to specify a completely known probability model for $\pi_0(y, z) = P(D = 1|Y = y, Z = z)$. Thus, we assume that $\pi(y, z; \gamma)$ is the posited working probability model for $\pi_0(y, z)$. We consider two different approaches to estimation of β without and with the constraints on $g_3(Y, Z, D; \gamma_0)$. In the first approach, we impose no constraints on $g_3(Y, Z, D; \gamma)$, while the second approach imposes constraints on $g_3(Y, Z, D; \gamma)$.

In the first approach, we impose constraints on two estimating functions $g_1(Y, Z, D, X; \beta)$ and $g_2(Y, Z, D; \beta, \gamma, \xi)$, and employ the maximum likelihood estimator $\hat{\gamma}_n$ of γ. Specifically, we maximize the nonparametric likelihood $L_F = \prod_{i=1}^n p_i$ subject to the constraints $\sum_{i=1}^n p_i = 1$, $p_i \geq 0$, $\sum_{i=1}^n p_i g(Y_i, Z_i, D_i, X_i; \beta, \hat{\gamma}_n, \hat{\xi}_n) = 0$, where $p_i = dF(Y_i, Z_i, D_i, X_i)$ for $i = 1, \ldots, n$. Similar to eqs (6) and (7), we

propose to estimate β by the maximum empirical likelihood estimator $\hat{\beta}_{\text{el2}}$, which maximizes the profile log empirical likelihood function of β:

$$\ell_2(\beta) = -\sum_{i=1}^{n} \log\{1 + \hat{\lambda}_{2n}^{\tau}(\beta) g(Y_i, Z_i, D_i, X_i; \beta, \hat{\gamma}_n, \hat{\xi}_n)\} - n \log n,$$

where the Lagrange multiplier $\hat{\lambda}_{2n} = \hat{\lambda}_{2n}(\beta)$ solves

$$U_{2n}(\lambda) \equiv U_{2n}(\lambda, \beta, \hat{\gamma}_n, \hat{\xi}_n) = \frac{1}{n} \sum_{i=1}^{n} \frac{g(Y_i, Z_i, D_i, X_i; \beta, \hat{\gamma}_n, \hat{\xi}_n)}{1 + \lambda^{\tau} g(Y_i, Z_i, D_i, X_i; \beta, \hat{\gamma}_n, \hat{\xi}_n)} = 0.$$

The motivation of the second approach is to improve the efficiency of $\hat{\beta}_{\text{el2}}$ by adding the score estimating function $g_3(Y, Z, D; \gamma)$ to the constraints associated with $\ell_2(\beta)$. Thus, in the second approach, we impose constraints on $g_1(Y, Z, D, X; \beta)$, $g_2(Y, Z, D; \beta, \hat{\gamma}_n, \hat{\xi}_n)$, and $g_3(Y, Z, D; \hat{\gamma}_n)$ and maximize $L_F = \prod_{i=1}^{n} p_i$ subject to the constraints $\sum_{i=1}^{n} p_i = 1$, $p_i \geq 0$, $\sum_{i=1}^{n} p_i G(Y_i, Z_i, D_i, X_i; \beta, \hat{\gamma}_n, \hat{\xi}_n) = 0$. This leads to estimation of β by maximizing the profile log empirical likelihood function $\ell_3(\beta) = \ell_3\big(\beta, \hat{\lambda}_{3n}(\beta), \hat{\gamma}_n\big)$ of β, where

$$\ell_3(\beta, \lambda, \gamma) = -\sum_{i=1}^{n} \log\{1 + \lambda^{\tau} G(Y_i, Z_i, D_i, X_i; \beta, \gamma, \hat{\xi}_n)\} - n \log n, \tag{8}$$

and, for fixed β, the Lagrange multiplier $\hat{\lambda}_{3n} = \hat{\lambda}_{3n}(\beta)$ is determined by

$$U_{3n}(\lambda) \equiv U_{3n}(\lambda, \beta, \hat{\gamma}_n, \hat{\xi}_n) = \frac{1}{n} \sum_{i=1}^{n} \frac{G(Y_i, Z_i, D_i, X_i; \beta, \hat{\gamma}_n, \hat{\xi}_n)}{1 + \lambda^{\tau} G(Y_i, Z_i, D_i, X_i; \beta, \hat{\gamma}_n, \hat{\xi}_n)} = 0. \tag{9}$$

Let $\hat{\beta}_{\text{el3}}$ denote the maximum empirical likelihood estimator of β that maximizes $\ell_3(\beta)$. The following theorem summarizes the large-sample results of $\hat{\beta}_{\text{el2}}$ and $\hat{\beta}_{\text{el3}}$.

Theorem 2 : Under the regularity conditions (A1)-(A6) stated in the Appendix, if the working probability model $\pi(y, z; \gamma)$ is correctly specified, we can write

$$\begin{aligned}
\hat{\beta}_{\text{el2}} - \beta_0 = -\frac{1}{n} \sum_{i=1}^{n} C_1^{-1} \Big[& g_1(Y_i, Z_i, D_i, X_i; \beta_0) \\
& - C_4 C_5^{-1} \{ g_2(Y_i, Z_i, D_i; \beta_0, \gamma_0, \xi^*) - C_2 S_{\gamma}^{-1} g_3(Y_i, Z_i, D_i; \gamma_0) \} \Big] + o_p(n^{-1/2}), \\
\hat{\beta}_{\text{el3}} - \beta_0 = -\frac{1}{n} \sum_{i=1}^{n} C_1^{-1} \Big[& g_1(Y_i, Z_i, D_i, X_i; \beta_0) \\
& - C_7 \{ g_2(Y_i, Z_i, D_i; \beta_0, \gamma_0, \xi^*) - C_2 S_{\gamma}^{-1} g_3(Y_i, Z_i, D_i; \gamma_0) \} \Big] + o_p(n^{-1/2}),
\end{aligned}$$

where S_{γ}, C_2, C_7 are, respectively, defined in eqs (11), (14), and (19). As a result, $\sqrt{n}(\hat{\beta}_{\text{el2}} - \beta_0) \xrightarrow{d} N(0, \Sigma_{\text{el2}})$ and $\sqrt{n}(\hat{\beta}_{\text{el3}} - \beta_0) \xrightarrow{d} N(0, \Sigma_{\text{el3}})$, where

$$\begin{aligned}
\Sigma_{\text{el2}} &= C_1^{-1} \text{Var}[g_1(Y, Z, D, X; \beta_0) - C_4 C_5^{-1} \{ g_2(Y, Z, D; \beta_0, \gamma_0, \xi^*) - C_2 S_{\gamma}^{-1} g_3(Y, Z, D; \gamma_0) \}] (C_1^{-1})^{\tau}, \\
\Sigma_{\text{el3}} &= C_1^{-1} \text{Var}[g_1(Y, Z, D, X; \beta_0) - C_7 \{ g_2(Y, Z, D; \beta_0, \gamma_0, \xi^*) - C_2 S_{\gamma}^{-1} g_3(Y, Z, D; \gamma_0) \}] (C_1^{-1})^{\tau}.
\end{aligned}$$

The asymptotic expansions of the maximum empirical likelihood estimators $\hat{\beta}_{\text{el1}}$, $\hat{\beta}_{\text{el2}}$, and $\hat{\beta}_{\text{el3}}$ in Theorems 1 and 2 provide a revealing insight into the potential efficiency gain by employing the empirical

likelihood method to incorporate the working probability model of missingness and the working regression model into the analysis of nonignorable covariate-missing data problems; we will discuss the efficiency comparison of $\hat{\beta}_{\mathrm{el}1}$, $\hat{\beta}_{\mathrm{el}2}$, and $\hat{\beta}_{\mathrm{el}3}$ with other competitive estimators in the next section. To make inference for β, we need to consistently estimate the asymptotic covariance matrices $\Sigma_{\mathrm{el}1}$, $\Sigma_{\mathrm{el}2}$, and $\Sigma_{\mathrm{el}3}$ given in Theorems 1 and 2. Here, we present an empirical-likelihood-based estimator of $\Sigma_{\mathrm{el}3}$; the empirical-likelihood-based estimators of $\Sigma_{\mathrm{el}1}$ and $\Sigma_{\mathrm{el}2}$ can be similarly derived. Throughout this paper, let $E_{3n}(\cdot)$ denote the empirical likelihood mean operator associated with $\hat{\beta}_{\mathrm{el}3}$. For example, $E_{3n}(Y) = \sum_{i=1}^{n} \tilde{p}_{3i}(\hat{\beta}_{\mathrm{el}3})Y_i$, where

$$\tilde{p}_{3i}(\beta) = \frac{1}{n}\frac{1}{1+\hat{\lambda}_{3n}^{\tau}G(Y_i, Z_i, D_i, X_i; \beta, \hat{\gamma}_n, \hat{\xi}_n)}, \qquad i = 1, \ldots, n.$$

Then the asymptotic covariance matrix $\Sigma_{\mathrm{el}3}$ of $\hat{\beta}_{\mathrm{el}3}$ can be consistently estimated by

$$\tilde{\Sigma}_{\mathrm{el}3} = \tilde{C}_1^{-1}E_{3n}\left(\left[g_1(Y, Z, D, X; \hat{\beta}_{\mathrm{el}3}) - \tilde{C}_7\{g_2(Y, Z, D; \hat{\beta}_{\mathrm{el}3}, \hat{\gamma}_n, \hat{\xi}_n) - \tilde{C}_2\tilde{S}_{\gamma}^{-1}g_3(Y, Z, D; \hat{\gamma}_n)\}\right]^{\otimes 2}\right)(\tilde{C}_1^{-1})^{\tau},$$

where

$$\begin{aligned}
&\tilde{C}_1 = E_{3n}\left\{D\frac{\partial U(Y, Z, X; \hat{\beta}_{\mathrm{el}3})}{\partial \beta^{\tau}}\right\}, \qquad \tilde{C}_2 = E_{3n}\{m(Y, Z; \hat{\beta}_{\mathrm{el}3}, \hat{\xi}_n)\pi_{\gamma}^{\tau}(Y, Z; \hat{\gamma}_n)\},\\
&\tilde{C}_4 = E_{3n}[DU(Y, Z, X; \hat{\beta}_{\mathrm{el}3})\{1 - \pi(Y, Z; \hat{\gamma}_n)\}m^{\tau}(Y, Z; \hat{\beta}_{\mathrm{el}3}, \hat{\xi}_n)]\\
&\tilde{C}_5 = E_{3n}[\pi(Y, Z; \hat{\gamma}_n)\{1 - \pi(Y, Z; \hat{\gamma}_n)\}m(Y, Z; \hat{\beta}_{\mathrm{el}3}, \hat{\xi}_n)m^{\tau}(Y, Z; \hat{\beta}_{\mathrm{el}3}, \hat{\xi}_n)],\\
&\tilde{C}_6 = E_{3n}\left[DU(Y, Z, X; \hat{\beta}_{\mathrm{el}3})\frac{\pi_{\gamma}^{\tau}(Y, Z; \hat{\gamma}_n)}{\pi(Y, Z; \hat{\gamma}_n)}\right], \qquad \tilde{S}_{\gamma} = E_{3n}\left[\frac{\pi_{\gamma}(Y, Z; \hat{\gamma}_n)\pi_{\gamma}^{\tau}(Y, Z; \hat{\gamma}_n)}{\pi(Y, Z; \hat{\gamma}_n)\{1 - \pi(Y, Z; \hat{\gamma}_n)\}}\right],\\
&\tilde{C}_7 = (\tilde{C}_4 - \tilde{C}_6\tilde{S}_{\gamma}^{-1}\tilde{C}_2^{\tau})(\tilde{C}_5 - \tilde{C}_2\tilde{S}_{\gamma}^{-1}\tilde{C}_2^{\tau})^{-1}.
\end{aligned}$$

Here we define $A^{\otimes 2} = AA^{\tau}$ for matrix A. The asymptotic normality of $\hat{\beta}_{\mathrm{el}3}$ in Theorem 2 implies that an approximate level $1 - \alpha$ Wald confidence region for the regression parameter β is the ellipsoid determined by all β such that

$$n(\hat{\beta}_{\mathrm{el}3} - \beta)^{\tau}\tilde{\Sigma}_{\mathrm{el}3}^{-1}(\hat{\beta}_{\mathrm{el}3} - \beta) \le \chi_p^2(1 - \alpha),$$

where $\chi_p^2(\alpha)$ is the lower (100α)th percentile of the χ_p^2-distribution, satisfying $P\{\chi_p^2 \le \chi_p^2(\alpha)\} = \alpha$. The hypothesis $H_0 : \beta = \beta_0$ is rejected in favor of $H_1 : \beta \ne \beta_0$, at a level of significance approximately α, if the observed Wald test statistic $n(\hat{\beta}_{\mathrm{el}3} - \beta_0)^{\tau}\tilde{\Sigma}_{\mathrm{el}3}^{-1}(\hat{\beta}_{\mathrm{el}3} - \beta_0) > \chi_p^2(1 - \alpha)$. In a similar manner, we can construct confidence intervals and test statistics for components of β.

3 Efficiency comparison

To compare the proposed maximum empirical likelihood estimators $\hat{\beta}_{\mathrm{el}1}$, $\hat{\beta}_{\mathrm{el}2}$, and $\hat{\beta}_{\mathrm{el}3}$ with other existing estimators, we employ the notion of influence functions. For an exposition on influence functions, see, for example, Tsiatis [25, Chapter 3]. For two matrices M_1 and M_2, the notations $M_1 \le M_2$ and $M_2 \ge M_1$ mean that $M_2 - M_1$ is nonnegative definite. Recall that $\hat{\beta}_c$ is the complete-case estimator defined in eq. (2), and $\hat{\beta}_{\mathrm{acc}0}$ and $\hat{\beta}_{\mathrm{acc}}$ are the augmented complete-case estimators defined in eq. (4). In addition to $\hat{\beta}_{\mathrm{acc}}$, Bartlett et al. [4] have also proposed a modified estimator $\hat{\beta}_{\mathrm{acc}2}$ which, for a given choice of working regression model $m(y, z; \beta, \xi)$, solves

$$\sum_{i=1}^{n}\left[D_iU(Y_i, X_i, Z_i, \beta) - \hat{C}_7\{D_i - \pi(Y_i, Z_i; \hat{\gamma}_n)\}m(Y_i, Z_i; \beta, \hat{\xi}_n)\right] = 0,$$

where $\hat{C}_7 = (\hat{C}_4 - \hat{C}_6\hat{S}_\gamma^{-1}\hat{C}_2^\tau)(\hat{C}_5 - \hat{C}_2\hat{S}_\gamma^{-1}\hat{C}_2^\tau)^{-1}$ is a consistent estimator of C_7 defined in eq. (19) and

$$\hat{C}_2 = E_n\{m(Y,Z;\hat{\beta}_c,\hat{\xi}_n)\pi_\gamma^\tau(Y,Z;\hat{\gamma}_n)\}, \quad \hat{C}_4 = E_n[DU(Y,Z,X;\hat{\beta}_c)\{1-\pi(Y,Z;\hat{\gamma}_n)\}m^\tau(Y,Z;\hat{\beta}_c,\hat{\xi}_n)]$$
$$\hat{C}_5 = E_n[\pi(Y,Z;\hat{\gamma}_n)\{1-\pi(Y,Z;\hat{\gamma}_n)\}m(Y,Z;\hat{\beta}_c,\hat{\xi}_n)m^\tau(Y,Z;\hat{\beta}_c,\hat{\xi}_n)],$$
$$\hat{C}_6 = E_n\left[DU(Y,Z,X;\hat{\beta}_c)\frac{\pi_\gamma^\tau(Y,Z;\hat{\gamma}_n)}{\pi(Y,Z;\hat{\gamma}_n)}\right], \qquad \hat{S}_\gamma = E_n\left[\frac{\pi_\gamma(Y,Z;\hat{\gamma}_n)\pi_\gamma^\tau(Y,Z;\hat{\gamma}_n)}{\pi(Y,Z;\hat{\gamma}_n)\{1-\pi(Y,Z;\hat{\gamma}_n)\}}\right].$$

Throughout $E_n(\cdot)$ represents the empirical mean operator. Bartlett et al. [4] have shown that $\hat{\beta}_{\text{acc2}}$ is at least as efficient as $\hat{\beta}_c$.

We first compare $\hat{\beta}_{\text{el1}}$ with the complete-case estimator $\hat{\beta}_c$ and the augmented complete-case estimator $\hat{\beta}_{\text{acc0}}$. According to Bartlett et al. [4], the influence function of $\hat{\beta}_c$ is given by

$$\psi_c(Y,Z,D,X;\beta_0) = -C_1^{-1}g_1(Y,Z,D,X;\beta_0).$$

Furthermore, it can be shown that the influence function of $\hat{\beta}_{\text{acc0}}$ is equal to

$$\psi_{\text{acc0}}(Y,Z,D,X;\beta_0,\gamma_0) = -C_1^{-1}\{g_1(Y,Z,D,X;\beta_0) - g_2(Y,Z,D;\beta_0,\gamma_0,\xi^*)\}.$$

Moreover, the asymptotic expansion of $\hat{\beta}_{\text{el1}}$ in Theorem 2.1 implies that the influence function of $\hat{\beta}_{\text{el1}}$ is given by

$$\psi_{\text{el1}}(Y,Z,D,X;\beta_0,\gamma_0,\xi^*) = -C_1^{-1}\{g_1(Y,Z,D,X;\beta_0) - C_4C_5^{-1}g_2(Y,Z,D;\beta_0,\gamma_0,\xi^*)\}.$$

To compare the asymptotic performances of $\hat{\beta}_{\text{el1}}$, $\hat{\beta}_c$, and $\hat{\beta}_{\text{acc0}}$, let

$$\Lambda_2 = \left\{Ag_2(Y,Z,D;\beta_0,\gamma_0,\xi^*) \text{ for all } 1\times p \text{ real vectors } A\right\}$$

denote the linear space spanned by $g_2(Y,Z,D;\beta_0,\gamma_0,\xi^*)$ and let $\Pi\{g_1(Y,Z,D,X;\beta_0)|\Lambda_2\}$ stand for the projection of $g_1(Y,Z,D,X;\beta_0)$ onto the space Λ_2. It can be shown after some algebra that

$$\Pi\{g_1(Y,Z,D,X;\beta_0)|\Lambda_2\} = C_4C_5^{-1}g_2(Y,Z,D;\beta_0,\gamma_0,\xi^*).$$

Thus, the influence function of $\hat{\beta}_{\text{el1}}$ can be alternatively written as

$$\psi_{\text{el1}}(Y,Z,D,X;\beta_0,\gamma_0,\xi^*) = \psi_c(Y,Z,D,X;\beta_0) - \Pi\{\psi_c(Y,Z,D,X;\beta_0)|\Lambda_2\}.$$

This implies that $\psi_{\text{el1}}(Y,Z,D,X;\beta_0,\gamma_0,\xi^*)$ is the residual from the projection of $\psi_c(Y,Z,D,X;\beta_0)$ onto the space Λ_2. It now follows from the theory of influence functions that

$$\Sigma_{\text{el1}} = \text{Var}\{\psi_{\text{el1}}(Y,Z,D,X;\beta_0,\gamma_0,\xi^*)\} \le \text{Var}\{\psi_c(Y,Z,D,X;\beta_0)\} \equiv \Sigma_c.$$

In addition, since the second term in $\psi_{\text{acc0}}(Y,Z,D,X;\beta_0,\gamma_0)$ belongs to the space Λ_2, we have

$$\Sigma_{\text{el1}} = \text{Var}\{\psi_{\text{el1}}(Y,Z,D,X;\beta_0,\gamma_0,\xi^*)\} \le \text{Var}\{\psi_{\text{acc0}}(Y,Z,D,X;\beta_0,\gamma_0)\} \equiv \Sigma_{\text{acc0}}.$$

These two facts demonstrate that $\hat{\beta}_{\text{el1}}$ is asymptotically at least as efficient as both $\hat{\beta}_c$ and $\hat{\beta}_{\text{acc0}}$ for any working regression function $m(y,z;\beta,\xi)$, whether or not it correctly identifies the optimal regression function $E\{U(Y,X,Z,\beta)|Y,Z,D = 1\}$. Nevertheless, $\hat{\beta}_{\text{acc0}}$ is not guaranteed to be asymptotically at least as efficient as $\hat{\beta}_c$.

Next, we compare the three proposed estimators $\hat{\beta}_{\text{el1}}$, $\hat{\beta}_{\text{el2}}$, and $\hat{\beta}_{\text{el3}}$. The asymptotic expansions of $\hat{\beta}_{\text{el2}}$ and $\hat{\beta}_{\text{el3}}$ in Theorem 2.2 imply that their influence functions are given by

$$\begin{aligned}
&\psi_{\text{el2}}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*) \\
&\quad = C_1^{-1}\big[g_1(Y, Z, D, X; \beta_0) - C_4 C_5^{-1}\{g_2(Y, Z, D; \beta_0, \gamma_0, \xi^*) - C_2 S_\gamma^{-1} g_3(Y, Z, D; \gamma_0)\}\big], \\
&\psi_{\text{el3}}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*) \\
&\quad = -C_1^{-1}\big[g_1(Y, Z, D, X; \beta_0) - C_7\big\{g_2(Y, Z, D; \beta_0, \gamma_0, \xi^*) - C_2 S_\gamma^{-1} g_3(Y, Z, D; \gamma_0)\big\}\big].
\end{aligned}$$

To compare the asymptotic performances among $\hat{\beta}_{\text{el1}}$, $\hat{\beta}_{\text{el2}}$, and $\hat{\beta}_{\text{el3}}$, let

$$\Lambda_3 = \left\{ A g_3(Y, Z, D; \gamma_0) \text{ for all } 1 \times q \text{ real vectors } A \right\}$$

denote the linear space spanned by $g_3(Y, Z, D; \gamma_0)$. It can be shown after some algebra that

$$\Pi\{g_2(Y, Z, D; \beta_0, \gamma_0, \xi^*) | \Lambda_3\} = C_2 S_\gamma^{-1} g_3(Y, Z, D; \gamma_0). \tag{10}$$

The residual from the projection of $g_2(Y, Z, D; \beta_0, \gamma_0, \xi^*)$ onto the space Λ_3 is orthogonal to components of the estimating function $g_3(Y, Z, D; \gamma_0)$ and is given by

$$h_{23}(Y, Z, D; \beta_0, \gamma_0, \xi^*) \equiv g_2(Y, Z, D; \beta_0, \gamma_0, \xi^*) - C_2 S_\gamma^{-1} g_3(Y, Z, D; \gamma_0) = \Pi\{g_2(Y, Z, D; \beta_0, \gamma_0, \xi^*) | \Lambda_3^\perp\},$$

where $\Lambda_3^\perp$ is the orthogonal complement of Λ_3. Moreover, let

$$\Lambda_{23} = \left\{ A h_{23}(Y, Z, D; \beta_0, \gamma_0, \xi^*) \text{ for all } 1 \times q \text{ real vectors } A \right\}$$

represent the linear space spanned by $h_{23}(Y, Z, D; \beta_0, \gamma_0, \xi^*)$. Then it is seen that the projection of $g_1(Y, Z, D, X; \beta_0)$ onto the space Λ_{23} is given by

$$\Pi\{g_1(Y, Z, D, X; \beta_0) | \Lambda_{23}\} = C_7 h_{23}(Y, Z, D; \beta_0, \gamma_0, \xi^*).$$

This, together with eq. (10), implies that $\psi_{\text{el3}}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*)$ is the residual from the projection of $\psi_c(Y, Z, D, X; \beta_0)$ onto the space Λ_{23}, namely,

$$\psi_{\text{el3}}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*) = \psi_c(Y, Z, D, X; \beta_0) - \Pi\{\psi_c(Y, Z, D, X; \beta_0) | \Lambda_{23}\}.$$

In other words, $\psi_{\text{el3}}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*)$ is the residual from the projection of $\psi_c(Y, Z, D, X; \beta_0)$ onto the linear space spanned by the residual from the projection of $g_2(Y, Z, D; \beta_0, \gamma_0, \xi^*)$ onto the linear space spanned by $g_3(Y, Z, D; \gamma_0)$. Again, the theory of influence functions implies that

$$\Sigma_{\text{el3}} = \text{Var}\{\psi_{\text{el3}}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*)\} \le \text{Var}\{\psi_c(Y, Z, D, X; \beta_0)\} = \Sigma_c.$$

Furthermore, since the second term in $\psi_{\text{el2}}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*)$ belongs to the space Λ_{23}, we have

$$\Sigma_{\text{el3}} = \text{Var}\{\psi_{\text{el3}}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*)\} \le \text{Var}\{\psi_{\text{el2}}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*)\} = \Sigma_{\text{el2}}.$$

Consequently, $\hat{\beta}_{\text{el3}}$ is asymptotically at least as efficient as both $\hat{\beta}_c$ and $\hat{\beta}_{\text{el2}}$ for any working regression function $m(y, z; \beta, \xi)$, whether or not it correctly identifies the optimal regression function $E\{U(Y, X, Z, \beta) | Y, Z, D = 1\}$. Since the empirical likelihood method leading to $\hat{\beta}_{\text{el2}}$ imposes a smaller number of constrains and hence

involves a smaller number of Lagrange multipliers, $\hat{\beta}_{el2}$ is computationally less intensive than $\hat{\beta}_{el3}$. On the other hand, use of $\hat{\beta}_{el2}$ does not ensure that its asymptotic efficiency is at least as good as $\hat{\beta}_c$, through specification of a working probability model for the probability of missingness. Moreover, there is no guarantee that either $\hat{\beta}_{el2}$ or $\hat{\beta}_{el3}$ improves upon the asymptotic efficiency of $\hat{\beta}_{el1}$, indicating that we are not guaranteed to improve on the asymptotic efficiency of estimation for β by estimating the probability of missingness even when it is known. This phenomenon contrasts with the well-known counter-intuitive fact under MAR scenarios where we can improve on the efficiency of estimation for a parameter by estimating the propensity score even when it is known.

Finally, we compare the proposed estimators $\hat{\beta}_{el2}$ and $\hat{\beta}_{el3}$ with the augmented complete-case estimators $\hat{\beta}_{acc}$ and $\hat{\beta}_{acc2}$. According to Bartlett et al. [4], the influence function of $\hat{\beta}_{acc}$ is

$$\psi_{acc}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*) = -C_1^{-1}\{g_1(Y, Z, D, X; \beta_0) - h_{23}(Y, Z, D; \beta_0, \gamma_0, \xi^*)\}.$$

Moreover, it is seen that the influence function of $\hat{\beta}_{acc2}$ is identical to that of $\hat{\beta}_{el3}$, i.e.,

$$\psi_{acc2}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*) = \psi_{el3}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*).$$

Thus, $\hat{\beta}_{el3}$ and $\hat{\beta}_{acc2}$ are asymptotically equivalent. Now since the second term in the influence function $\psi_{acc}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*)$ belongs to the space Λ_{23}, we have

$$\begin{aligned}\Sigma_{el3} &= \mathrm{Var}\{\psi_{el3}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*)\} = \mathrm{Var}\{\psi_{acc2}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*)\} \\ &\le \mathrm{Var}\{\psi_{acc}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*)\} \equiv \Sigma_{acc}.\end{aligned}$$

As a result, $\hat{\beta}_{el3}$ and $\hat{\beta}_{acc2}$ are asymptotically at least as efficient as both $\hat{\beta}_{el2}$ and $\hat{\beta}_{acc}$ for any given working regression model $m(y, z; \beta, \xi)$ posited for $E\{U(Y, X, Z, \beta)|Y, Z, D = 1\}$. However, there does not appear to have a direct asymptotic efficiency comparison between $\hat{\beta}_{el2}$ and $\hat{\beta}_{acc}$, though the simulation study in Section 5 indicates that $\hat{\beta}_{el2}$ has a better small-sample performance than $\hat{\beta}_{acc}$ in terms of root mean square error.

We close this section by pointing out that when the working regression function $m(Y, Z; \beta, \xi)$ correctly specifies the true regression function $E\{U(Y, X, Z, \beta)|Y, Z, D = 1\}$, so that $m(Y, Z; \beta_0, \xi_0) = E\{U(Y, X, Z, \beta_0)|Y, Z, D = 1\}$, it can be shown after some algebra that $C_2 = C_6$, $C_4 = C_5$, and $C_7 = I_p$, and hence

$$\begin{aligned}\psi_{el1}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*) &= \psi_{acc0}(Y, Z, D, X; \beta_0, \gamma_0), \\ \psi_{el3}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*) &= \psi_{el2}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*) = \psi_{acc2}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*) \\ &= \psi_{acc}(Y, Z, D, X; \beta_0, \gamma_0, \xi^*).\end{aligned}$$

This implies that $\hat{\beta}_{el1}$ and $\hat{\beta}_{acc0}$ are asymptotically equivalent when γ_0 is known and that $\hat{\beta}_{el2}$, $\hat{\beta}_{el3}$, $\hat{\beta}_{acc}$, and $\hat{\beta}_{acc2}$ are asymptotically equivalent when γ_0 is unknown. As discussed earlier, the asymptotic efficiency of $\hat{\beta}_{el1}$ and $\hat{\beta}_{acc0}$ with γ_0 known differs from that of $\hat{\beta}_{el2}$, $\hat{\beta}_{el3}$, $\hat{\beta}_{acc}$, and $\hat{\beta}_{acc2}$ with γ_0 unknown, but all these estimators, $\hat{\beta}_{el1}$, $\hat{\beta}_{acc0}$, $\hat{\beta}_{el2}$, $\hat{\beta}_{el3}$, $\hat{\beta}_{acc}$, and $\hat{\beta}_{acc2}$ are asymptotically at least as efficient as $\hat{\beta}_c$.

4 Application to the NHANES data

To demonstrate our proposed methods, we revisit the Demographics Data, the Examination Data, and the Questionnaire Data from the 2003–2004 NHANES. We focus on the dependence of systolic blood pressure (SBP) on the average number of alcoholic drinks consumed per day, with adjustment for age and body mass index (BMI). As in Bartlett et al. [4], we consider the dataset drawn from 2,418 male participants, in which 278 men have missing values in SBP and 181 men have missing values in BMI. However, age was fully observed for all participants. According to Little and Zhang [3] and Bartlett et al. [4], it is reasonable to assume that

Table 1: Parameter estimates and standard errors (in parentheses) in a conditional mean model of SBP (mmHg) (centered at 125 mmHg) on log (average number of drinks consumed per day +1), BMI, age (decades above 50) and age^2 ((decades above 50)^2).

Estimator	Variable				
	Constant	log (no.drinks+1)	BMI (kg/m^2)	Age	Age^2
$\hat{\beta}_c$	−1.9286 (0.7980)	1.2665 (0.5831)	0.4143 (0.0798)	3.9431 (0.2606)	0.2646 (0.1431)
$\hat{\beta}_{acc}$	−2.0663 (0.6530)	1.2494 (0.3879)	0.3893 (0.0656)	3.8732 (0.2296)	0.3147 (0.1078)
$\hat{\beta}_{acc2}$	−2.0131 (0.6137)	1.2180 (0.2764)	0.3920 (0.0650)	3.8699 (0.2273)	0.3023 (0.1073)
$\hat{\beta}_{el2}$	−2.1825 (0.6270)	1.4116 (0.3117)	0.3860 (0.0650)	3.9268 (0.2282)	0.3056 (0.1078)
$\hat{\beta}_{el3}$	−2.0063 (0.6137)	1.2372 (0.2763)	0.3956 (0.0650)	3.8710 (0.2273)	0.2990 (0.1073)

the missingness in SBP and BMI is missing completely at random (MCAR) due to missed visits, and hence an analysis with these participants omitted shall not produce biased estimation and is thus valid. Consequently, we apply the proposed methods to the remaining 2,111 participants, of which 720 (34%) men have missing values in alcohol consumption. In addition to the MCAR assumption, Bartlett et al. [4] have argued that missingness in alcohol consumption is missing not at random (MNAR) because it is likely dependent largely on alcohol consumption itself, and that missingness in alcohol consumption is independent of SBP conditional on alcohol consumption, age, and BMI.

In our application, we fit a linear regression model by regressing SBP on the covariates: log (no.drinks+1), BMI, and age (with both linear and quadratic effects). As suggested by Bartlett et al. [4], a log transformation for the alcohol variable is performed to reduce the influence on parameter estimates caused by the few observations with extremely large values. The aforementioned assumption on the conditional independence of the alcohol consumption missingness and SBP given alcohol consumption, age, and BMI entails that the CCA estimates $\hat{\beta}_c$ are unbiased. For the calculation of the estimates $\hat{\beta}_{acc}$, $\hat{\beta}_{acc2}$, $\hat{\beta}_{el2}$, and $\hat{\beta}_{el3}$, we postulate a logistic regression model for the working probability model with covariates age, BMI, SBP, and SBP^2 and a negative binomial regression model for the working regression model with covariates age, age^2, BMI, BMI^2, SBP, and SBP^2.

The results of the analysis are given in Table 1. The results show that the proposed empirical likelihood estimates $\hat{\beta}_{el2}$ and $\hat{\beta}_{el3}$ are quite similar to the estimates $\hat{\beta}_{acc}$ and $\hat{\beta}_{acc2}$ of Bartlett et al. [4] and that all four of these estimates have smaller standard errors than the CCA estimate $\hat{\beta}_c$. Moreover, the methods that yield $\hat{\beta}_{el3}$ and $\hat{\beta}_{acc2}$ produce almost identical results in terms of estimated coefficients and standard errors and, in particular, render the lowest standard errors among all of the methods. Overall, our analysis indicate that an increased alcohol consumption per day is associated with an increased SBP, as expected.

5 Simulations

In this section, we conduct a simulation study to compare the proposed empirical likelihood based estimators with CCA, multiple imputation, IPW, and ACC estimators. We generate the data using the procedure of Bartlett et al. [4], in which the non-missing indicator variable D is simulated with $P(D = 1) = 0.5$ and the variables (X, Z, Y) are generated from a trivariate normal distribution conditional on D :

$$\begin{pmatrix} X \\ Z \\ Y \end{pmatrix} | D \sim N \left\{ \begin{pmatrix} \alpha_X D \\ \alpha_Z D \\ \alpha_Y D \end{pmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XZ} & \sigma_{XY} \\ \sigma_{XZ} & \sigma_Z^2 & \sigma_{ZY} \\ \sigma_{XY} & \sigma_{ZY} & \sigma_Y^2 \end{bmatrix} \right\}$$

with

$$\alpha_Y = \frac{\alpha_X(\sigma_{XY}\sigma_Z^2 - \sigma_{XZ}\sigma_{ZY}) + \alpha_Z(\sigma_{ZY}\sigma_X^2 - \sigma_{XY}\sigma_{XZ})}{\sigma_X^2\sigma_Z^2 - \sigma_{XZ}^2}$$

such that $D \perp Y|(X, Z)$ is satisfied.

The working probability model used in our simulation study is the logistic regression model:

$$\pi(y, z; \gamma_0) = P(D = 1|Y = y, Z = z) = \frac{1}{1 + \exp\{-(\gamma_C + \gamma_Y y + \gamma_Z z)\}}.$$

To calculate a working regression model for the optimal regression function $m_0(Y, Z; \beta, \xi_0)$, we posit a parametric working model $f(X|D = 1, Y, Z; \xi)$ for the true conditional density function $f(X|D = 1, Y, Z)$ of X given $(D = 1, Y, Z)$. Then the conditional expectation $m(Y, Z; \beta, \xi)$ of X given $(D = 1, Y, Z)$ under $f(X|D = 1, Y, Z; \xi)$ is the corresponding working regression model for $m_0(Y, Z; \beta, \xi_0)$ and is calculated by Monte-Carlo integration. The conditional mean model of interest is specified as

$$Y|X, Z \sim N\left(\beta_0 + \beta_X X + \beta_Z Z, \sigma^2\right).$$

We consider two scenarios in our simulation study:

(a) the working regression function $m(Y, Z; \beta, \xi)$ is correctly specified;
(b) the working regression function $m(Y, Z; \beta, \xi)$ is misspecified.

For each scenario, we generate 1,000 Monte Carlo data sets, each composed of either $n = 500$ or $n = 1500$ subjects according to two different setups:

(1) SETUP A: $\alpha_X = 1$, $\alpha_Z = 0$, $\sigma_X^2 = \sigma_Z^2 = \sigma_Y^2 = 0.9$, $\sigma_{XZ} = \sigma_{ZY} = \sigma_{XY} = 0.25$;
(2) SETUP B: $\alpha_X = 1$, $\alpha_Z = 0$, $\sigma_X^2 = 0.9$, $\sigma_Z^2 = 1$, $\sigma_Y^2 = 0.8$, $\sigma_{XZ} = 0.1$, $\sigma_{ZY} = 0.25$, $\sigma_{XY} = 0.2$.

Under scenario (a), the working regression models for SETUP A and SETUP B are both correctly specified as $m(Y, Z; \beta, \xi_0) = E(X|D = 1, Y, Z) = \xi_C + \xi_Z Z + \xi_Y Y$. Under scenario (b), we postulate $m(Y, Z; \beta, \xi) = 1 + 0.5Z^2 + 0.5Y^2$ for $E(X|D = 1, Y, Z)$ under SETUP A and $m(Y, Z; \beta, \xi) = 1.2 + 0.5Z^2 + 0.4Y^2$ for $E(X|D = 1, Y, Z)$ under SETUP B. For each scenario and setup, we examine the bias, standard deviation (SD), and root mean squared error (RMSE) of the following estimators:

(i) CCA estimator $\hat{\beta}_c$.
(ii) Multiple imputation estimator $\hat{\beta}_{MI}$ by imputing X 10 times from a normal linear regression model for $X|Y, Z$.
(iii) IPW estimator $\hat{\beta}_{IPW}$ based on weights from a logistic regression model with Y and Z as the covariates.
(iv) $\hat{\beta}_{acc0}$ and $\hat{\beta}_{el1}$ with γ_0 known. In this case, the logistic working probability function $\pi(y, z; \gamma_0)$ satisfies

$$\gamma_Z = \frac{\alpha_Z \sigma_Y^2 - \alpha_Y \sigma_{ZY}}{\sigma_Y^2 \sigma_Z^2 - \sigma_{ZY}^2}, \quad \gamma_Y = \frac{\alpha_Y \sigma_Z^2 - \alpha_Z \sigma_{ZY}}{\sigma_Y^2 \sigma_Z^2 - \sigma_{ZY}^2}, \quad \gamma_C = \frac{-\frac{\alpha_Z^2}{2}\sigma_Y^2 + \alpha_Y \alpha_Z \sigma_{ZY} - \frac{\alpha_Y^2}{2}\sigma_Z^2}{\sigma_Y^2 \sigma_Z^2 - \sigma_{ZY}^2}.$$

(v) $\hat{\beta}_{acc}$, $\hat{\beta}_{acc2}$, $\hat{\beta}_{el2}$, and $\hat{\beta}_{el3}$ with γ_0 unknown using an estimated logistic working probability function.

The simulation results are summarized in Tables 2–5.

Tables 2 and 3, with sample sizes $n = 500$ and $n = 1500$, respectively, show the performances of different estimators under SETUP A and SETUP B when the working regression model is correctly specified. The results in Tables 2 and 3 can be summarized as follows.

(1) As expected, the biases of the CCA estimator $\hat{\beta}_c$, the ACC estimators $\hat{\beta}_{acc0}$, $\hat{\beta}_{acc}$, $\hat{\beta}_{acc2}$, and the empirical likelihood estimators $\hat{\beta}_{el1}$, $\hat{\beta}_{el2}$ and $\hat{\beta}_{el3}$ are all negligible under both setups and for both sample sizes.
(2) The biases of the MI estimator $\hat{\beta}_{MI}$ and IPW estimator $\hat{\beta}_{IPW}$ are significantly larger than those of the other estimators in most cases for both sample sizes. This is not surprising because both the MI and IPW estimators require the MAR assumption and are biased when missing is not at random (MNAR).
(3) When γ_0 is known, $\hat{\beta}_c$, $\hat{\beta}_{acc0}$ and $\hat{\beta}_{el1}$ have similar biases, but $\hat{\beta}_{acc0}$ and $\hat{\beta}_{el1}$ have much smaller standard deviations and hence smaller RMSEs for β_0 and β_Z. For β_X, these three estimators have similar RMSEs. In

Table 2: Bias, SD, and RMSE based on 1,000 simulations with sample size $n = 500$. The working regression model is correctly specified.

Estimator	β_0		β_X		β_Z	
	Bias (SD)	RMSE	Bias (SD)	RMSE	Bias (SD)	RMSE
(1) SETUP A: Missing % = 51.0%						
$\hat{\beta}_{c}$	0.0012 (0.0834)	0.0834	−0.0010 (0.0627)	0.0628	0.0018 (0.0636)	0.0637
$\hat{\beta}_{MI}$	−0.2783 (0.0715)	0.2874	0.1760 (0.0609)	0.1862	−0.0494 (0.0445)	0.0665
$\hat{\beta}_{IPW}$	−0.1121 (0.0771)	0.1361	0.0007 (0.0543)	0.0543	0.0314 (0.0429)	0.0532
$\hat{\beta}_{acc0}$	−0.0032 (0.0756)	0.0756	−0.0001 (0.0647)	0.0647	−0.0001 (0.0471)	0.0471
$\hat{\beta}_{el1}$	−0.0043 (0.0745)	0.0746	0.0012 (0.0638)	0.0638	−0.0003 (0.0473)	0.0473
$\hat{\beta}_{acc}$	−0.0005 (0.0850)	0.0850	0.0004 (0.0649)	0.0649	−0.0015 (0.0481)	0.0481
$\hat{\beta}_{acc2}$	0.0037 (0.0843)	0.0844	−0.0017 (0.0631)	0.0631	−0.0013 (0.0483)	0.0484
$\hat{\beta}_{el2}$	−0.0005 (0.0840)	0.0840	0.0005 (0.0633)	0.0633	−0.0011 (0.0482)	0.0482
$\hat{\beta}_{el3}$	−0.0016 (0.0835)	0.0835	0.0036 (0.0619)	0.0620	−0.0023 (0.0489)	0.0490
(2) SETUP B: Missing % = 49.4%						
$\hat{\beta}_{c}$	−0.0010 (0.0781)	0.0781	0.0009 (0.0575)	0.0575	−0.0011 (0.0524)	0.0525
$\hat{\beta}_{MI}$	−0.1129 (0.0646)	0.1301	0.0128 (0.0546)	0.0561	−0.0179 (0.0388)	0.0427
$\hat{\beta}_{IPW}$	−0.1019 (0.0681)	0.1226	0.0040 (0.0486)	0.0488	0.0091 (0.0389)	0.0400
$\hat{\beta}_{acc0}$	−0.0042 (0.0669)	0.0671	0.0012 (0.0576)	0.0576	−0.0004 (0.0385)	0.0385
$\hat{\beta}_{el1}$	−0.0047 (0.0666)	0.0667	0.0019 (0.0576)	0.0576	−0.0004 (0.0384)	0.0384
$\hat{\beta}_{acc}$	−0.0014 (0.0785)	0.0785	0.0013 (0.0578)	0.0578	−0.0013 (0.0387)	0.0387
$\hat{\beta}_{acc2}$	0.0011 (0.0781)	0.0781	0.0005 (0.0573)	0.0573	−0.0015 (0.0387)	0.0387
$\hat{\beta}_{el2}$	−0.0020 (0.0779)	0.0779	0.0019 (0.0575)	0.0576	−0.0013 (0.0386)	0.0386
$\hat{\beta}_{el3}$	−0.0026 (0.0779)	0.0779	0.0029 (0.0576)	0.0577	−0.0013 (0.0386)	0.0386

addition, $\hat{\beta}_{el1}$ has slightly smaller RMSEs than $\hat{\beta}_{acc0}$ for β_0, β_X and β_Z. These observations made for both sample sizes are in agreement with the theoretical results.

(4) When γ_0 is unknown, $\hat{\beta}_{acc}$, $\hat{\beta}_{acc2}$, $\hat{\beta}_{el2}$ and $\hat{\beta}_{el3}$ have similar RMSEs. Compared to $\hat{\beta}_{c}$, these estimators have similar biases but largely reduced standard deviations for β_Z and thus gain efficiency for β_Z. As for the estimation of β_0 and β_X, these estimators have similar RMSEs to $\hat{\beta}_{c}$. These results obtained from both sample sizes confirm that if the working regression model is correctly specified, $\hat{\beta}_{acc}$, $\hat{\beta}_{acc2}$, $\hat{\beta}_{el2}$ and $\hat{\beta}_{el3}$ are asymptotically equivalent and at least as efficient as $\hat{\beta}_{c}$ when γ_0 is unknown.

(5) When $n = 500$, $\hat{\beta}_{acc}$ has the largest RMSEs for β_0 and β_X under both setups. Furthermore, $\hat{\beta}_{el3}$ has the lowest RMSEs for β_0 and β_X under SETUP A and the lowest RMSEs for β_0 and β_Z under SETUP B. By contrast, when $n = 1500$, the proposed empirical likelihood estimators $\hat{\beta}_{el2}$ and $\hat{\beta}_{el3}$ have relatively smaller RMSEs than the ACC estimators $\hat{\beta}_{acc}$ and $\hat{\beta}_{acc2}$. In addition, $\hat{\beta}_{el3}$ performs best in terms of RMSEs under both SETUP A and SETUP B.

(6) When the sample size is increased from $n = 500$ to $n = 1500$, there is a large reduction in standard deviations and RMSEs for each estimator.

Tables 4 and 5, with sample sizes $n = 500$ and $n = 1500$, respectively, pertain to the situation where the working regression model is misspecified under both setups. The results in Tables 4 and 5 can be summarized as follows.

(1) Similar to the first scenario, $\hat{\beta}_{c}$ has negligible biases, whereas $\hat{\beta}_{MI}$ and $\hat{\beta}_{IPW}$ have very large biases and hence are biased under both setups and for both sample sizes.

(2) Under SETUP A with γ_0 known and for both sample sizes, $\hat{\beta}_{acc0}$ has larger RMSEs than $\hat{\beta}_{c}$ for both β_0 and β_X, although they are still unbiased. By contrast, $\hat{\beta}_{el1}$ has much lower RMSEs for β_0 and β_Z and a similar RMSE for β_X compared to $\hat{\beta}_{c}$.

(3) Under SETUP B with γ_0 known and for both sample sizes, $\hat{\beta}_{acc0}$ has relatively higher biases and larger RMSEs than $\hat{\beta}_{c}$ for both β_0 and β_X, although these estimators are unbiased. Again, $\hat{\beta}_{el1}$ has much smaller RMSEs for β_0 and β_Z and a similar RMSE for β_X compared to $\hat{\beta}_{c}$. Moreover, $\hat{\beta}_{el1}$ always performs better than

Table 3: Bias, SD, and RMSE based on 1,000 simulations with sample size $n = 1500$. The working regression model is correctly specified.

Estimator	β_0		β_X		β_Z	
	Bias (SD)	**RMSE**	**Bias** (SD)	**RMSE**	**Bias** (SD)	**RMSE**
(1) SETUP A: Missing % = 48.6%						
$\hat{\beta}_{\mathrm{c}}$	−0.0016 (0.0509)	0.0510	0.0017 (0.0366)	0.0367	−0.0005 (0.0359)	0.0359
$\hat{\beta}_{\mathrm{MI}}$	−0.2728 (0.0432)	0.2762	0.1727 (0.0354)	0.1763	−0.0468 (0.0266)	0.0538
$\hat{\beta}_{\mathrm{IPW}}$	−0.1086 (0.0435)	0.1170	0.0005 (0.0305)	0.0305	0.0321 (0.0261)	0.0414
$\hat{\beta}_{\mathrm{acc0}}$	−0.0019 (0.0449)	0.0450	0.0012 (0.0373)	0.0373	0.0012 (0.0282)	0.0282
$\hat{\beta}_{\mathrm{el1}}$	−0.0026 (0.0445)	0.0446	0.0020 (0.0367)	0.0368	0.0009 (0.0280)	0.0281
$\hat{\beta}_{\mathrm{acc}}$	−0.0015 (0.0513)	0.0513	0.0015 (0.0374)	0.0374	0.0005 (0.0286)	0.0286
$\hat{\beta}_{\mathrm{acc2}}$	−0.0004 (0.0512)	0.0512	0.0012 (0.0367)	0.0367	0.0003 (0.0284)	0.0284
$\hat{\beta}_{\mathrm{el2}}$	−0.0018 (0.0510)	0.0511	0.0019 (0.0368)	0.0368	0.0004 (0.0283)	0.0283
$\hat{\beta}_{\mathrm{el3}}$	−0.0015 (0.0510)	0.0510	0.0016 (0.0367)	0.0367	0.0005 (0.0283)	0.0283
(2) SETUP B: Missing % = 50.3%						
$\hat{\beta}_{\mathrm{c}}$	0.0010 (0.0468)	0.0468	−0.0009 (0.0339)	0.0339	0.0001 (0.0315)	0.0315
$\hat{\beta}_{\mathrm{MI}}$	−0.1086 (0.0387)	0.1153	0.0108 (0.0322)	0.0339	−0.0172 (0.0233)	0.0289
$\hat{\beta}_{\mathrm{IPW}}$	−0.0996 (0.0404)	0.1075	0.0031 (0.0277)	0.0279	0.0090 (0.0231)	0.0248
$\hat{\beta}_{\mathrm{acc0}}$	−0.0001 (0.0398)	0.0398	−0.0005 (0.0340)	0.0340	−0.0005 (0.0230)	0.0230
$\hat{\beta}_{\mathrm{el1}}$	0.0000 (0.0397)	0.0397	−0.0005 (0.0339)	0.0339	−0.0004 (0.0230)	0.0230
$\hat{\beta}_{\mathrm{acc}}$	0.0007 (0.0469)	0.0469	−0.0006 (0.0340)	0.0340	−0.0005 (0.0232)	0.0232
$\hat{\beta}_{\mathrm{acc2}}$	0.0019 (0.0469)	0.0469	−0.0011 (0.0339)	0.0339	−0.0005 (0.0232)	0.0233
$\hat{\beta}_{\mathrm{el2}}$	0.0008 (0.0467)	0.0467	−0.0007 (0.0339)	0.0339	−0.0004 (0.0232)	0.0232
$\hat{\beta}_{\mathrm{el3}}$	0.0009 (0.0467)	0.0467	−0.0008 (0.0339)	0.0339	−0.0004 (0.0232)	0.0232

$\hat{\beta}_{\mathrm{acc0}}$ in terms of RMSE. These results confirm our asymptotic theory that when γ_0 is known, the empirical likelihood estimator $\hat{\beta}_{\mathrm{el1}}$ is at least as efficient as the CCA estimator $\hat{\beta}_{\mathrm{c}}$ whether or not the working regression function is correctly specified. In contrast, the ACC estimator $\hat{\beta}_{\mathrm{acc0}}$ is not guaranteed to be at least as efficient as $\hat{\beta}_{\mathrm{c}}$.

(4) Under SETUP A with γ_0 unknown and for both sample sizes, $\hat{\beta}_{\mathrm{acc}}$ has larger RMSEs than $\hat{\beta}_{\mathrm{c}}$ for β_0 and β_X; both estimators are unbiased. It is seen that $\hat{\beta}_{\mathrm{acc2}}$, $\hat{\beta}_{\mathrm{el2}}$ and $\hat{\beta}_{\mathrm{el3}}$ have similar RMSEs; they all gain efficiency for β_Z compared to $\hat{\beta}_{\mathrm{c}}$, while having similar efficiency to $\hat{\beta}_{\mathrm{c}}$ for β_0 and β_X. In particular, when $n = 500$, $\hat{\beta}_{\mathrm{el3}}$ has the lowest RMSEs for both β_0 and β_X and has a similar RMSE to $\hat{\beta}_{\mathrm{acc2}}$ for β_Z. By contrast, when $n = 1500$, $\hat{\beta}_{\mathrm{el3}}$ has slightly smaller RMSEs than both $\hat{\beta}_{\mathrm{el2}}$ and $\hat{\beta}_{\mathrm{acc2}}$, and thus has the lowest RMSEs under SETUP A.

(5) Under SETUP B with γ_0 unknown, $\hat{\beta}_{\mathrm{acc}}$ has larger RMSEs than $\hat{\beta}_{\mathrm{c}}$ for β_X for both sample sizes, although they are unbiased. When $n = 500$, $\hat{\beta}_{\mathrm{el3}}$ has the lowest RMSEs for β_0 and β_X and a similar RMSE for β_Z compared to $\hat{\beta}_{\mathrm{acc2}}$ and $\hat{\beta}_{\mathrm{el2}}$. For $n = 1500$, we observe that $\hat{\beta}_{\mathrm{acc2}}$, $\hat{\beta}_{\mathrm{el2}}$, and $\hat{\beta}_{\mathrm{el3}}$ have almost same RMSEs; they all gain efficiency for β_Z compared to $\hat{\beta}_{\mathrm{c}}$, while having similar efficiency to $\hat{\beta}_{\mathrm{c}}$ for β_0 and β_X. These results support our asymptotic theory that when γ_0 is unknown, $\hat{\beta}_{\mathrm{el3}}$ and $\hat{\beta}_{\mathrm{acc2}}$ are asymptotically equivalent and at least as efficient as $\hat{\beta}_{\mathrm{c}}$ whether or not the working regression function is correctly specified. On the other hand, $\hat{\beta}_{\mathrm{acc}}$ is not guaranteed to be at least as efficient as $\hat{\beta}_{\mathrm{c}}$.

(6) The biases of $\hat{\beta}_{\mathrm{acc0}}$ in absolute value are at least 20% larger for $n = 500$ than those for $n = 1500$ under both setups, except for the case with β_Z under SETUP B. By contrast, the biases of $\hat{\beta}_{\mathrm{acc}}$ in absolute value are at least 43% larger for $n = 500$ than those for $n = 1500$ for β_0 and β_X under both setups; the biases of $\hat{\beta}_{\mathrm{acc}}$ are, nevertheless, smaller for $n = 500$ than for $n = 1500$ for β_Z under both setups. Again, we observe that when the sample size is increased from $n = 500$ to $n = 1500$, there is a large reduction in standard deviations and RMSEs for each estimator.

In summary, our simulation results with sample sizes $n = 500$ and $n = 1500$ indicate that $\hat{\beta}_{\mathrm{el2}}$ always performs better than $\hat{\beta}_{\mathrm{acc}}$ in terms of RMSE whether or not the working regression model is correctly specified, although there is no direct asymptotic efficiency comparison between them in our asymptotic theory. Also in

Table 4: Bias, SD, and RMSE based on 1,000 simulations with sample size $n = 500$. The working regression model is misspecified.

Estimator	β_0		β_x		β_z	
	Bias (SD)	RMSE	Bias (SD)	RMSE	Bias (SD)	RMSE
(1) SETUP A: Missing % = 49.6%						
$\hat{\beta}_c$	−0.0017 (0.0854)	0.0854	0.0008 (0.0629)	0.0629	−0.0016 (0.0611)	0.0612
$\hat{\beta}_{MI}$	−0.2668 (0.0711)	0.2761	0.1678 (0.0606)	0.1784	−0.0475 (0.0440)	0.0647
$\hat{\beta}_{IPW}$	−0.1062 (0.0713)	0.1279	−0.0007 (0.0511)	0.0511	0.0300 (0.0437)	0.0530
$\hat{\beta}_{acc0}$	−0.0063 (0.0961)	0.0963	0.0097 (0.0892)	0.0897	−0.0050 (0.0516)	0.0518
$\hat{\beta}_{el1}$	−0.0001 (0.0750)	0.0750	0.0020 (0.0644)	0.0644	−0.0019 (0.0469)	0.0470
$\hat{\beta}_{acc}$	−0.0026 (0.0909)	0.0910	0.0018 (0.0705)	0.0705	−0.0008 (0.0479)	0.0479
$\hat{\beta}_{acc2}$	0.0007 (0.0857)	0.0857	0.0002 (0.0629)	0.0629	−0.0011 (0.0467)	0.0467
$\hat{\beta}_{el2}$	−0.0032 (0.0862)	0.0863	0.0028 (0.0640)	0.0640	−0.0012 (0.0469)	0.0469
$\hat{\beta}_{el3}$	−0.0066 (0.0850)	0.0852	0.0022 (0.0629)	0.0629	−0.0021 (0.0469)	0.0469
(2) SETUP B: Missing % = 49.6%						
$\hat{\beta}_c$	−0.0012 (0.0779)	0.0780	0.0026 (0.0568)	0.0569	0.0004 (0.0541)	0.0541
$\hat{\beta}_{MI}$	−0.1089 (0.0651)	0.1269	0.0132 (0.0548)	0.0563	−0.0168 (0.0402)	0.0436
$\hat{\beta}_{IPW}$	−0.0996 (0.0696)	0.1215	0.0065 (0.0480)	0.0484	0.0095 (0.0400)	0.0411
$\hat{\beta}_{acc0}$	−0.0053 (0.0852)	0.0854	0.0093 (0.0799)	0.0804	−0.0012 (0.0421)	0.0421
$\hat{\beta}_{el1}$	−0.0019 (0.0666)	0.0667	0.0047 (0.0572)	0.0574	−0.0005 (0.0412)	0.0412
$\hat{\beta}_{acc}$	−0.0030 (0.0801)	0.0802	0.0048 (0.0605)	0.0607	−0.0005 (0.0406)	0.0406
$\hat{\beta}_{acc2}$	0.0014 (0.0788)	0.0788	0.0019 (0.0569)	0.0569	−0.0006 (0.0407)	0.0407
$\hat{\beta}_{el2}$	−0.0029 (0.0781)	0.0782	0.0048 (0.0570)	0.0572	−0.0004 (0.0408)	0.0408
$\hat{\beta}_{el3}$	−0.0052 (0.0775)	0.0777	0.0014 (0.0568)	0.0568	−0.0016 (0.0407)	0.0408

Table 5: Bias, SD, and RMSE based on 1,000 simulations with sample size $n = 1500$. The working regression model is misspecified.

Estimator	β_0		β_x		β_z	
	Bias (SD)	RMSE	Bias (SD)	RMSE	Bias (SD)	RMSE
(1) SETUP A: Missing % = 50.0%						
$\hat{\beta}_c$	0.0003 (0.0477)	0.0477	−0.0013 (0.0356)	0.0356	−0.0012 (0.0354)	0.0354
$\hat{\beta}_{MI}$	−0.2748 (0.0416)	0.2780	0.1732 (0.0347)	0.1767	−0.0495 (0.0254)	0.0557
$\hat{\beta}_{IPW}$	−0.1085 (0.0422)	0.1164	−0.0019 (0.0298)	0.0298	0.0300 (0.0247)	0.0389
$\hat{\beta}_{acc0}$	0.0052 (0.0499)	0.0502	−0.0011 (0.0444)	0.0444	0.0026 (0.0284)	0.0285
$\hat{\beta}_{el1}$	−0.0042 (0.0421)	0.0423	−0.0006 (0.0359)	0.0359	−0.0007 (0.0272)	0.0272
$\hat{\beta}_{acc}$	−0.0001 (0.0494)	0.0494	−0.0007 (0.0386)	0.0386	−0.0013 (0.0275)	0.0275
$\hat{\beta}_{acc2}$	0.0011 (0.0476)	0.0477	−0.0014 (0.0355)	0.0355	−0.0015 (0.0270)	0.0270
$\hat{\beta}_{el2}$	−0.0005 (0.0477)	0.0477	−0.0003 (0.0356)	0.0356	−0.0014 (0.0270)	0.0270
$\hat{\beta}_{el3}$	0.0001 (0.0475)	0.0475	−0.0009 (0.0355)	0.0355	−0.0013 (0.0269)	0.0270
(2) SETUP B: Missing % = 50.7%						
$\hat{\beta}_c$	−0.0016 (0.0453)	0.0453	0.0017 (0.0330)	0.0330	−0.0014 (0.0306)	0.0306
$\hat{\beta}_{MI}$	−0.1117 (0.0382)	0.1181	0.0137 (0.0319)	0.0347	−0.0172 (0.0220)	0.0279
$\hat{\beta}_{IPW}$	−0.1018 (0.0388)	0.1089	0.0054 (0.0278)	0.0283	0.0094 (0.0220)	0.0239
$\hat{\beta}_{acc0}$	−0.0031 (0.0484)	0.0485	0.0037 (0.0446)	0.0448	−0.0012 (0.0230)	0.0230
$\hat{\beta}_{el1}$	−0.0023 (0.0385)	0.0386	0.0025 (0.0330)	0.0331	−0.0010 (0.0224)	0.0224
$\hat{\beta}_{acc}$	−0.0021 (0.0459)	0.0459	0.0022 (0.0345)	0.0346	−0.0013 (0.0220)	0.0220
$\hat{\beta}_{acc2}$	−0.0008 (0.0453)	0.0453	0.0015 (0.0330)	0.0330	−0.0014 (0.0220)	0.0220
$\hat{\beta}_{el2}$	−0.0024 (0.0452)	0.0453	0.0025 (0.0329)	0.0330	−0.0014 (0.0221)	0.0222
$\hat{\beta}_{el3}$	−0.0017 (0.0453)	0.0453	0.0018 (0.0330)	0.0330	−0.0013 (0.0220)	0.0220

our simulation scenarios, $\hat{\beta}_{el3}$ performs at least as well as $\hat{\beta}_{acc2}$, even though the reduction in RMSE is sometimes very slight. Finally, the nine estimators we have considered in Tables 2–5 are not directly comparable between the two cases, γ_0 is known versus γ_0 is unknown.

6 Concluding remarks

In this paper, we have studied empirical likelihood methods for estimating the regression coefficients in a nonignorable covariate-missing data problem under the conditional independence assumption advocated by Bartlett et al. [4]. We have proposed three unbiased estimating functions and three empirical likelihood-based estimators of the regression coefficients through specification of a working probability model of missingness and a working regression model. The proposed empirical likelihood methods can be viewed as pseudo-empirical likelihood methods for estimation of β because we have employed the maximum likelihood estimator $\hat{\gamma}_n$ of γ rather than the maximum empirical likelihood estimator of γ. We have also made efficiency comparisons of the three proposed estimators with other existing estimators. The simulation results indicate that the proposed empirical likelihood estimators have competitive finite sample properties in terms of bias and root mean square error. The proposed empirical likelihood approach is illustrated using an analysis of the alcohol and blood pressure data from the US National Health and Nutrition Examination Survey (NHANES).

An alternative empirical likelihood method to the pseudo empirical likelihood method discussed in this paper is to simultaneously estimate β and γ by maximizing the profile empirical likelihood function jointly with respect to (β, γ); this gives rise to studying the maximum empirical likelihood estimator of (β, γ). Another issue is to study the empirical likelihood ratio test for testing the linear null hypothesis $H_0 : C\beta = C\beta_0$ with C being a $r \times p$ matrix; this leads to studying the constraint maximum empirical likelihood estimation of β under linear constraints on the regression parameter in nonignorable covariate-missing data problems. These considerations pave an avenue for further exploration.

Acknowledgment: We are grateful to the Editor, Professor Moulinath Banerjee, two reviewers for a number of helpful comments and suggestions, and to Dr. Jonathan Bartlett for providing the NHANES dataset.

Appendix: Proofs

Here we provide a proof for Theorem 2. The proof of Theorem 1 is similar to that of Theorem 2 and is therefore omitted here.

Regularity Conditions

(A1) For all (y, z), $\pi(y, z; \gamma)$ admits all third partial derivatives $\dfrac{\partial^3 \pi(y, z; \gamma)}{\partial \gamma_i \partial \gamma_j \partial \gamma_k}$ for all γ in a neighborhood of the true value γ_0 and $\left|\frac{\partial^3 \pi(y,z;\gamma)}{\partial \gamma_i \partial \gamma_j \partial \gamma_k}\right|$ is bounded by an integrable function for all γ in this neighborhood.
(A2) For all (y, z), $\pi(y, z; \gamma) \in (0, 1)$ for all γ in a neighborhood of γ_0.
(A3) The matrix S_γ is positive definite and $E\{||\pi_\gamma(Y, Z, \gamma_0)||^2\} < \infty$, where S_γ is defined in eq. (11) below.
(A4) For all (y, z, x), $U(y, z, x; \beta)$ admits all second partial derivatives $\frac{\partial^2 U(y,z,x;\beta)}{\partial \beta_i \partial \beta_j}$ for all β in a neighborhood of β_0 and $\left|\frac{\partial^2 U(y,z,x;\beta)}{\partial \beta_i \partial \beta_j}\right|$ is bounded by an integrable function in this neighborhood. Also, $\partial U(y, z, x; \beta)/\partial \beta$ has finite second-order moments and $E\{U^\tau(Y, Z, X, \beta_0)U(Y, Z, X, \beta_0)\} < \infty$.
(A5) For all (y, z), $m(y, z; \beta, \xi)$ admits all second partial derivatives $\frac{\partial^2 m(y,z;\beta,\xi)}{\partial \beta_i \partial \beta_j}$ and $\frac{\partial^2 m(y,z;\beta,\xi)}{\partial \xi_k \partial \xi_l}$ for all (β, ξ) in a neighborhood of (β_0, ξ^*). Moreover, $\left|\frac{\partial^2 m(y,z;\beta,\xi))}{\partial \beta_i \partial \beta_j}\right|$ and $\left|\frac{\partial^2 m(y,z;\beta,\xi))}{\partial \xi_k \partial \xi_l}\right|$ are bounded by an integrable function in this neighborhood.
(A6) The matrices C_1, C_5, and $C_5 - C_2 S_\gamma^{-1} C_2^\tau$ are invertible, where C_1, C_2, and C_5 are, respectively, defined in eqs (14) and (19) below.

Proof of Theorem 2: For parts (a) and (b), under regularity conditions (A1)-(A3), it can be shown that the maximum likelihood estimator $\hat{\gamma}_n$ solves the score equation $U_n(\gamma) = 0$ in eq. (3) and has influence function $\psi_{\hat{\gamma}_n}(Y, X, D; \gamma_0) = S_\gamma^{-1} g_3(Y, Z, D; \gamma_0)$, where

$$S_\gamma = E\left[\frac{\pi_\gamma(Y,Z;\gamma_0)\pi_\gamma^\tau(Y,Z;\gamma_0)}{\pi(Y,Z;\gamma_0)\{1-\pi(Y,Z;\gamma_0)\}}\right]. \tag{11}$$

Since $\hat\beta_{\text{el3}}$ maximizes $\ell_3(\beta)$ in eq. (8) for each fixed λ, it is seen from eqs (8), (9), and (3) that $(\hat\lambda_{3n}, \hat\beta_{\text{el3}}, \hat\gamma_n)$ satisfies the equations $U_{3n}(\hat\lambda_{3n}, \hat\beta_{\text{el3}}, \hat\gamma_n, \hat\xi_n) = 0$, $Q_{3n}(\hat\lambda_{3n}, \hat\beta_{\text{el3}}, \hat\gamma_n, \hat\xi_n) = 0$, and $U_n(\hat\gamma_n) = 0$, where

$$Q_{3n}(\lambda,\beta,\gamma,\xi) = -\frac{1}{n}\frac{\partial\ell_3(\beta,\lambda,\gamma)}{\partial\beta} = \frac{1}{n}\sum_{i=1}^n \frac{\frac{\partial\lambda^\tau G(Y_i,Z_i,D_i,X_i;\beta,\gamma,\xi)}{\partial\beta}}{1+\lambda^\tau G(Y_i,Z_i,D_i,X_i;\beta,\gamma,\xi)}.$$

Under the regularity conditions in (A1)–(A6), it can be shown that $(\hat\beta_{\text{el3}}, \hat\lambda_{3n}, \hat\gamma_n)$ is an $\sqrt{n}$-consistent estimator of $(\beta_0, 0^{(2p+q)\times 1}, \gamma_0)$. An application of a first-order Taylor expansion gives

$$\begin{aligned}
0 = U_{3n}(\hat\lambda_{3n},\hat\beta_{\text{el3}},\hat\gamma_n,\hat\xi_n) &= U_{3n}(0,\beta_0,\gamma_0,\hat\xi_n) + \frac{\partial U_{3n}(\lambda_{3n}^*,\beta_n^*,\gamma_n^*,\xi^*)}{\partial\lambda^\tau}\hat\lambda_{3n} \\
&\quad + \frac{\partial U_{3n}(\lambda_{3n}^*,\beta_n^*,\gamma_n^*,\xi^*)}{\partial\beta^\tau}(\hat\beta_{\text{el3}}-\beta_0) + \frac{\partial U_{3n}(\lambda_{3n}^*,\beta_n^*,\gamma_n^*,\xi^*)}{\partial\gamma^\tau}(\hat\gamma_n-\gamma_0), \\
0 = Q_{3n}(\hat\lambda_{3n},\hat\beta_{\text{el3}},\hat\gamma_n,\hat\xi_n) &= Q_{3n}(0,\beta_0,\gamma_0,\hat\xi_n) + \frac{\partial Q_{3n}(\lambda_{3n}^*,\beta_n^*,\gamma_n^*,\xi^*)}{\partial\lambda^\tau}\hat\lambda_{3n} \\
&\quad + \frac{\partial Q_{3n}(\lambda_{3n}^*,\beta_n^*,\gamma_n^*,\xi^*)}{\partial\beta^\tau}(\hat\beta_{\text{el3}}-\beta_0) + \frac{\partial Q_{3n}(\lambda_{3n}^*,\beta_n^*,\gamma_n^*,\xi^*)}{\partial\gamma^\tau}(\hat\gamma_n-\gamma_0), \\
0 = U_n(\hat\gamma_n) &= U_n(\gamma_0) + \frac{\partial U_n(\gamma_n^*)}{\partial\gamma^\tau}(\hat\gamma_n-\gamma_0),
\end{aligned} \tag{12}$$

where $(\lambda_{3n}^*, \beta_n^*, \gamma_n^*)$ is some intermediate value between $(\hat\lambda_{3n}, \hat\beta_{\text{el3}}, \hat\gamma_n)$ and $(0, \beta_0, \gamma_0)$. As $n \to \infty$, it follows from the law of large numbers that

$$\begin{aligned}
&\frac{\partial U_{3n}(\lambda_{3n}^*,\beta_n^*,\gamma_n^*,\xi^*)}{\partial\lambda^\tau} \xrightarrow{p} -E\{G(Y,Z,D,X;\beta_0,\gamma_0,\xi^*)G^\tau(Y,Z,D,X;\beta_0,\gamma_0,\xi^*)\} = H_{11}, \\
&\frac{\partial U_{3n}(\lambda_{3n}^*,\beta_n^*,\gamma_n^*,\xi^*)}{\partial\beta^\tau} \xrightarrow{p} E\left\{\frac{\partial G(Y,Z,D,X;\beta_0,\gamma_0,\xi^*)}{\partial\beta^\tau}\right\} = H_{12} = \begin{pmatrix} C_1 \\ 0 \end{pmatrix}, \\
&\frac{\partial U_{3n}(\lambda_{3n}^*,\beta_n^*,\gamma_n^*,\xi^*)}{\partial\gamma^\tau} \xrightarrow{p} E\left\{\frac{\partial G(Y,Z,D,X;\beta_0,\gamma_0,\xi^*)}{\partial\gamma^\tau}\right\} = \begin{pmatrix} 0 \\ C_3 \end{pmatrix}, \\
&\frac{\partial Q_{3n}(\lambda_{3n}^*,\beta_n^*,\gamma_n^*,\xi^*)}{\partial\lambda^\tau} \xrightarrow{p} E\left\{\frac{\partial G^\tau(Y,Z,D,X;\beta_0,\gamma_0,\xi^*)}{\partial\beta}\right\} = H_{21} = H_{12}^\tau = (C_1^\tau,\ 0^\tau), \\
&\frac{\partial Q_{3n}(\lambda_{3n}^*,\beta_n^*,\gamma_n^*,\xi^*)}{\partial\beta^\tau} \xrightarrow{p} 0, \qquad \frac{\partial Q_{3n}(\lambda_{3n}^*,\beta_n^*,\gamma_n^*,\xi^*)}{\partial\gamma^\tau} \xrightarrow{p} 0, \qquad \frac{\partial U_n(\gamma_n^*)}{\partial\gamma^\tau} \xrightarrow{p} -S_\gamma,
\end{aligned} \tag{13}$$

where

$$C_1 = E\left\{D\frac{\partial U(Y,Z,X;\beta_0)}{\partial\beta^\tau}\right\}, \quad C_2 = E\{m(Y,Z;\beta_0,\xi^*)\pi_\gamma^\tau(Y,Z;\gamma_0)\}, \quad C_3 = -\begin{pmatrix} C_2 \\ S_\gamma \end{pmatrix}. \tag{14}$$

Furthermore, it can be shown after some algebra that

$$\begin{aligned}
U_{3n}(0,\beta_0,\gamma_0,\hat\xi_n) &= U_{3n}(0,\beta_0,\gamma_0,\xi^*) + o_p(n^{-1/2}), \\
Q_{3n}(0,\beta_0,\gamma_0,\hat\xi_n) &= Q_{3n}(0,\beta_0,\gamma_0,\xi^*) + o_p(n^{-1/2}).
\end{aligned} \tag{15}$$

Combining eqs (12), (13), and (15) yields

$$\begin{pmatrix} \hat{\lambda}_{3n} \\ \hat{\beta}_{\mathrm{el3}} - \beta_0 \end{pmatrix} = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & 0 \end{pmatrix}^{-1} \begin{pmatrix} -U_n(\beta_0, \gamma_0, \xi^*) \\ 0 \end{pmatrix} + o_p(n^{-1/2}) \tag{16}$$

$$= \begin{pmatrix} -(H_{11}^{-1} + H_{11}^{-1}H_{12}H_{22\cdot1}^{-1}H_{21}H_{11}^{-1}) \\ H_{22\cdot1}^{-1}H_{21}H_{11}^{-1} \end{pmatrix} U_{4n}(\beta_0, \gamma_0, \xi^*) + o_p(n^{-1/2}), \tag{17}$$

where $H_{22\cdot1} = -H_{21}H_{11}^{-1}H_{12}$ and

$$U_{4n}(\beta_0, \gamma_0, \xi^*) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} g_1(Y_i, Z_i, D_i, X_i; \beta_0) \\ g_{23}(Y_i, Z_i, D_i; \beta_0, \gamma_0, \xi^*) + C_3 S_\gamma^{-1} g_3(Y_i, Z_i, D_i; \gamma_0) \end{pmatrix},$$

$$g_{23}(Y, Z, D; \beta, \gamma, \xi) = \begin{pmatrix} g_2(Y, Z, D; \beta, \gamma, \xi) \\ g_3(Y, Z, D; \gamma) \end{pmatrix}. \tag{18}$$

As a result, it follows from eqs (17) and (18) that we can write

$$\begin{aligned} \hat{\beta}_{\mathrm{el3}} - \beta_0 &= H_{22\cdot1}^{-1}H_{21}H_{11}^{-1}U_{4n}(\beta_0, \gamma_0, \xi^*) + o_p(n^{-1/2}) \\ &= -\frac{1}{n}\sum_{i=1}^{n} C_1^{-1}\Big[g_1(Y_i, Z_i, D_i, X_i; \beta_0) - C_7\big\{g_2(Y_i, Z_i, D_i; \beta_0, \gamma_0, \xi^*) - C_2 S_\gamma^{-1} g_3(Y_i, Z_i, D_i; \gamma_0)\big\}\Big] \\ &\quad + o_p(n^{-1/2}), \end{aligned}$$

where

$$\begin{aligned} C_4 &= E[DU(Y, Z, X; \beta_0)\{1 - \pi(Y, Z; \gamma_0)\}m^\tau(Y, Z; \beta_0, \xi^*)], \\ C_5 &= E[\pi(Y, Z; \gamma_0)\{1 - \pi(Y, Z; \gamma_0)\}m(Y, Z; \beta_0, \xi^*)m^\tau(Y, Z; \beta_0, \xi^*)], \\ C_6 &= E\left[DU(Y, Z, X; \beta_0)\frac{\pi_\gamma^\tau(Y, Z; \gamma_0)}{\pi(Y, Z; \gamma_0)}\right], \qquad C_7 = (C_4 - C_6 S_\gamma^{-1} C_2^\tau)(C_5 - C_2 S_\gamma^{-1} C_2^\tau)^{-1}. \end{aligned} \tag{19}$$

This, together with Slutsky's theorem, implies that $\sqrt{n}(\hat{\beta}_{\mathrm{el3}} - \beta_0) \xrightarrow{d} N_p(0, \Sigma_{\mathrm{el3}})$. The proof is complete.

References

1. Little RJ, Rubin DB. Statistical analysis with missing data, 2nd ed. Hoboken, NJ: Wiley, 2002.
2. Centers for Disease Control and Prevention. National health and nutrition examination survey data. Hyattsville: Centers for Disease Control and Prevention, 2004.
3. Little RJ, Zhang N. Subsample ignorable likelihood for regression analysis with missing data. J R Stat Soc Ser C 2011;60: 591–605.
4. Bartlett JW, Carpenter JR, Tilling K, Vansteelandt, S. Improving upon the efficiency of complete case analysis when covariates are MNAR. Biostatistics 2014;15:719–30.
5. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 1952;47:663–85.
6. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 1994;89:846–66.
7. Owen AB. Empirical likelihood ratio confidence intervals for a single functional. Biometrika 1988;75:237–49.
8. Owen AB. Empirical likelihood confidence regions. Ann Stat 1990;18:90–120.
9. Qin J, Lawless JF. Empirical likelihood and general estimating equations. Ann Stat 1994;22:300–25.
10. Hall P, La Scala B. Methodology and algorithms of empirical likelihood. Int Stat Rev 1990;58:109–27.
11. Owen AB. Empirical likelihood. New York: Chapman & Hall/CRC, 2001.

12. Wang Q, Rao JN. Empirical likelihood-based inference under imputation for missing response data. Ann Stat 2002;30: 896–924.
13. Wang Q, Rao JN. Empirical likelihood-based inference in linear errors-in-covariables models with validation data. Biometrika 2002;89:345–58.
14. Chen SX, Leung DH, Qin J. Information recovery in a study with surrogate endpoints. J Am Stat Assoc 2003;98:1052–62.
15. Liang H, Wang SJ, Carroll RJ. Partially linear models with missing response variables and error-prone covariates. Biometrika 2007;94:185–98.
16. Stute W, Xue LG, Zhu LX. Empirical likelihood inference in nonlinear errors-in-covariables models with validation data. J Am Stat Assoc 2007;102:332–46.
17. Qin J, Zhang B. Empirical-likelihood-based inference in missing response problem and its application in observational studies. J R Stat Soc Ser B 2007;69:101–22.
18. Wang Q, Dai P. Semiparametric model-based inference in the presence of missing responses. Biometrika 2008;89:721–34.
19. Little RJ, Schluchter MD. Maximum likelihood estimation for mixed continuous and categorical data with missing values. Biometrika 1985;72:497–512.
20. Ibrahim JG. Incomplete data in generalized linear models. J Am Stat Assoc 1990;85:765–9.
21. Little RJ. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227–37.
22. Ibrahim JG, Lipsitz SR. Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. Biometrics 1996;52:1071–8.
23. Lipsitz SR, Ibrahim JG. A conditional model for incomplete covariates in parametric regression models. Biometrika 1996;83:916–22.
24. Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. Missing-data methods for generalized linear models: A comparative review. J Am Stat Assoc 2005;100:332–46.
25. Tsiatis AA. Semiparametric theory and missing data. New York: Springer, 2006.