Hindawi
Journal of Mathematics
Volume 2021, Article ID 7570222, 8 pages
https://doi.org/10.1155/2021/7570222

Research Article

Identification and Estimation of Graphical Models with Nonignorable Nonresponse

Lingju Chen,[1] **Shaoxin Hong**,[2] **and Bo Tang**[3]

[1]*College of Mathematics and Data Science (Software College), Minjiang University, Fuzhou 350108, China*
[2]*The Center for Economic Research, Shandong University, Jinan 250100, China*
[3]*Department of Mathematics and Statistics, University of North Carolina, Charlotte, NC 28223, USA*

Correspondence should be addressed to Shaoxin Hong; henryhong@sdu.edu.cn

Received 24 September 2021; Accepted 30 October 2021; Published 8 December 2021

Academic Editor: Jiancheng Jiang

We study the identification and estimation of graphical models with nonignorable nonresponse. An observable variable correlated to nonresponse is added to identify the mean of response for the unidentifiable model. An approach to estimating the marginal mean of response is proposed, based on simulation imputation methods which are introduced for a variety of models including linear, generalized linear, and monotone nonlinear models. The proposed mean estimators are $\sqrt{N}$-consistent, where N is the sample size. Finite sample simulations confirm the effectiveness of the proposed method. Sensitivity analysis for the untestable assumption on our augmented model is also conducted. A real data example is employed to illustrate the use of the proposed methodology.

1. Introduction

The problem of missing data in practice has attracted much attention for decades. Rubin [1] gave the weakest general conditions under which the nonresponse is ignorable; that is, ignoring the process that causes missing can still result in correct inference when both missing and observing are at random. Alternatively, the nonresponse is said to be nonignorable (Little and Rubin [2]) if it depends on the value of the possibly unobserved outcome. Groves, Presser, and Dipko [3] illustrated that if the missing mechanism is not ignorable, then a complete-case analysis which excludes missing data could result in highly biased estimates. For more discussions about nonresponse bias, one can refer to the work of Little and Rubin [2], Ibrahim and Lipsitz [4], and Goves [5], among others.

To address the problems of nonignorable nonresponse, weighting adjustments, which adjust the estimates by rescaling each unit's sample weight proportionally to the inverse of its response probability, are frequently used. The methods used for adjusting include poststratification by Holt and Smith [6], Calibration by Kott [7], and raking-ratio estimation by Deville, Särndal, and Sautory [8]. Weighting adjustments are model-based approaches, that is, the population values are treated as realizations of random variables that are distributed according to a superpopulation [9], and auxiliary information is incorporated into various models to describe nonrespondent behavior with respect to the variables of target. For instance, Greenlees, Reece, and Zieschang [10] conducted linear regression to analyze the unobserved income data assuming that nonresponse income depends on the unobserved value. Fay [11] and Baker and Laird [12] provided a family of estimable hierarchical log-linear models for the joint distribution of the data and the response indicator.

As pointed out by Little [13], the fully parametric approach is sensitive to failure of the assumed parametric model. Based on the exponential tilting model, Kim and Yu [14] proposed a semiparametric estimation method of mean functionals with nonignorable missing data. Riddles, Kim and Im [15] presented an approach of maximum likelihood estimation that uses parametric model assumptions about the variable of interest among the respondents only. Zhao, Tang, Qu, and Jiang [16] considered the parametric propensity model and studied semiparametric estimating equations inference by the nonparametric imputation method. Guo, Ma, and Wang [17] generalized the propensity model to semiparametric form and investigated the

estimation of the parametric copula model. There are also some works using graphic models to depict nonresponse mechanisms in the literature on the modeling approach. For examples, Ma, Geng, and Hu [18] used graphic models with temporal structure to describe nonresponse mechanisms for binary income in a longitudinal study. Wang, Chen, Geng, and Zhou [19] extended the work of Ma, Geng, and Hu [18] to derive the maximum likelihood estimator for the parameter of the binomial proportion and its associated variance. More existing works along this line can be found in the work of Little [13], Fay [11], and Forster and Smith [20].

In this paper, we present a new method to tackle the problem of nonignore nonresponse. A completely observable variable correlated to Y is added to identify the mean of the response Y. Adjustments are applied by a fixing response approach which divides the population into two strata: one consists of respondents and the other of nonrespondents [21]. Then, an approach to estimating the marginal mean of response Y is proposed based on simulation imputation methods under linear, generalized linear, and monotone nonlinear models, respectively. It is shown that the proposed estimators are $\sqrt{N}$-consistent, where N is the sample size. The effectiveness of the method is demonstrated via simulations. Application to real data of the method is also considered. Simulations show that our new method is successful in modeling the mean of response Y. Since the assumptions on the models are untestable, we propose to assess sensitivity to the assumption of our methods.

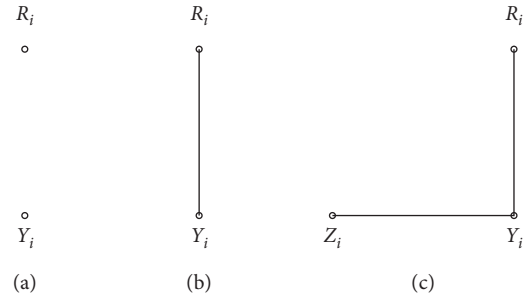
The outline of this paper is as follows. In Section 2, we introduce the graphical models. Section 3 develops the simulation imputation methods for linear, generalized linear, and monotone nonlinear models. In Section 4, the generalized linear model is extended to cope with other covariates. Section 5 introduces empirical standard errors to monitor the accuracy of bootstrap imputations. A simulation study is conducted in Section 6. Sensitivity analysis is given in Section 7. Section 8 illustrates an application to the mental health dataset. Proofs are given in Appendix.

2. Graphical Models

The response variable is denoted by Y. We assume that $\{Y_i\}_{i=1}^N$ are independent and identically distributed response variables for N subjects in the study and $\{R_i\}_{i=1}^N$ are the respective status indexes for the responses, where $R_i = 1$ or 0, which depends on Y_i response or nonresponse, respectively. Identifying the marginal mean $E(Y)$ of Y is an important problem in itself, and the treatment effect can be obtained once the marginal mean is available. However, if one uses only observed Y_i's with $R_i = 1$ to estimate $E(Y)$, it may result in a large biased estimator.

Example 1. We consider the mixture model with $Y_i = Y_{i1}R_i + Y_{i2}(1 - R_i)$, where $Y_{i1} \sim N(0, 1)$, $Y_{i2} \sim N(-4, 1)$, and R_i is a binary variable independent of Y_{i1} and Y_{i2} and satisfies that $P(R_i = 1) = 1 - P(R_i = 0) = 0.5$, for $i = 1, \ldots, n$. Then, the marginal mean of Y is $E(Y) = -2$, and the mean of observed Y is 0. Thus, if one ignores those nonresponses with $R_i = 0$, then the bias is 2.

To deal with this problem, we introduce the following three graphical models:



The graphical models have the following statistical meanings:

(i) Model (a): $E(Y_i|R_i) = E(Y_i)$, which means that Y_i is uncorrelated to R_i.

(ii) Model (b): $E(Y_i|R_i) = E(Y_i)$; that is, Y_i is correlated to R_i.

(iii) Model (c) (model (c) is untestable; i.e., one cannot test if $E(Z_i\, Y_i,\, R_i) = E(Z_i\, Y_i)$ holds. The assumption is proposed based on the information from specific experts. We will perform sensitivity analysis on the assumption for our results): $E(Z_i|Y_i, R_i) = E(Z_i|Y_i)$, which shows that Z_i is uncorrelated to R_i conditional on Y_i.

The marginal mean of the response Y is identifiable for model (a), but not for model (b), since one observes only those values of response variable with $R_i = 1$ (the response group). To solve this problem, we now introduce a surrogate Z of Y, which is completely observable in model (c), to identify the marginal mean of Y. We call model (c) the augmented model of model (b), which is identifiable for the marginal mean under certain conditions. The motivation for adding such a completely observed variable Z is from a study conducted by us for the income of the inhabitants in an area of Beijing. Since the income variable Y is relevant to private matters, it is subject to informative missing. Neglecting of the missing values may result in very biased estimation for the income levels in the area. This motivates us to add a completely observed variable Z, the size of houses that every inhabitant possesses. Note that, conditional on the income variable Y, the missing mechanism can be regarded as uncorrelated to the housing variable Z so that the model (c) holds in this example.

In the following, we consider the identification of model (c) and propose a method to estimate the marginal mean of Y_i.

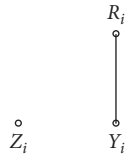
3. Simulation Imputation

Since the nonresponse subjects are nonignorable, biased estimation may be resulted if only the subjects in the response group with $R_i = 1$ are considered. It is necessary to use the information from the nonresponse so that a consistent estimator of the marginal mean of Y can be obtained. To this end, we introduce an approach to achieving this objective. Our idea is to impute the nonresponse by simulation.

3.1. Linear Models. Suppose that Z_i and Y_i satisfy the following linear model:

$$E(Z_i|Y_i) = \alpha + \beta Y_i, \quad i = 1, \ldots, N. \tag{1}$$

If $\beta = 0$, then model (c) is equivalent to the following model (d).



Model (d)

Since model (c) includes model (d), using a standard argument in hypothesis testing for general linear models, we can test if H_0: $\beta = 0$ holds via constructing testing statistic, $T(Z, Y)$, say, based on the estimator of β from model (c). If (d) holds, then $E(Y_i)$ is unidentifiable, and one must find a variable Z correlated with Y to solve this problem.

If $\beta = 0$, then it follows from (1) that

$$E(Y_i) = \frac{\{E(Z_i) - \alpha\}}{\beta}. \tag{2}$$

Thus, identifying $E(Y_i)$ is equivalent to estimating α and β because $E(Z_i)$ is estimatable. From model (c), we know that $E(Z_i|Y_i) = E(Z_i|Y_i, R_i = 1)$; then, regression of Z_i on Y_i for those observed Y_i with $R_i = 1$, we get the $\sqrt{N}$-consistent estimators of α and β (for example, the simple least squares estimators). We denote them by $\widehat{\alpha}$ and $\widehat{\beta}$, respectively. Let the estimator of $E(Z_i)$ be $\widehat{E}(Z)$. Then, the marginal mean of Y can be estimated as

$$\widehat{E}(Y_i) = \frac{\{\widehat{E}(Z_i) - \widehat{\alpha}\}}{\widehat{\beta}}. \tag{3}$$

The abovementioned estimation method is useful for the linear model, but it is inflexible for other models, for example, the generalized linear model and nonlinear model considered later. We here introduce another flexible estimating method, which is based on bootstrap. It uses the fact

$$E(Y) = E(Y|R = 1)P(R = 1) + E(Y|R = 0)P(R = 0). \tag{4}$$

To estimate $E(Y)$, we need to estimate $E(Y|R = 1)$ and $E(Y|R = 0)$, where the former can be estimated by the average of Y_i in the response group and the latter can be estimated by the average of imputations of Y_i for the nonresponse group. This method is referred to as "*simulation imputation*," which is detailed in the following:

(i) Fitting model (1) with a data subset: the subjects are partitioned into two groups, one is of $R_i = 1$ (the response group), the other $R_i = 0$ (the nonresponse group). Without loss of generality, assume $R_i = 1$ for $i = 1, \ldots, N_1$, and 0 others. Based on those subjects with responses observed (i.e., $i = 1, \ldots, N_1$), we regress$\{Z_i\}$ on $\{Y_i\}$ using model (1) and obtain the estimators of α and β, denoted by $\widehat{\alpha}$ and $\widehat{\beta}$, respectively.

(ii) Bootstrap residuals (The bootstrap based on residuals was proposed by Efron [22]. Because ε_i's are of mean zero, the empirical distribution function of the residuals is centered at 0): the residuals are denoted by $\widehat{\varepsilon}_i = Z_i - \widehat{\alpha} - \widehat{\beta}Y_i$ for $i = 1, \ldots, N_1$. Let $\overline{\varepsilon}_i = N_1^{-1} \sum_{i=1}^{N_1} \widehat{\varepsilon}_i$. We resample M samples with replacement, each with size $N_2 = N - N_1$, from the empirical distribution function of the centered residuals$\{\widehat{\varepsilon}_i - \overline{\varepsilon}, i = 1, N_1\}$, and denote the M samples by

$$\varepsilon^{*(m)}_{N_1+1}, \ldots, \varepsilon^{*(m)}_{N_1+N_2}, \quad \text{for } m = 1, \ldots, M. \tag{5}$$

(iii) Prediction of Y_j in the nonresponse group: let $Y_j^{(m)} = (Z_j - \widehat{\alpha} - \varepsilon_j^{*(m)})/\widehat{\beta}$, for $j = N_1 + 1, \ldots, N_1 + N_2$. We predict Y_j as the average of $Y_j^{(m)}$ over $m = 1, \ldots, M$. The average is denoted by $Y_j^{(*)}$.

(iv) Estimator of the marginal mean of Y:

$$\widehat{E}(Y) = \omega_1 N_1^{-1} \sum_{i=1}^{N_1} Y_i + (1 - \omega_1) N_2^{-1} \sum_{j=N_1+1}^{N_1+N_2} Y_j^*. \tag{6}$$

Theoretical justification of the abovementioned "*simulation imputation*" method can be built by using standard bootstrap theory. Especially, we have the following consistency result for the estimator $\widehat{E}(Y)$, which is proved in Appendix.

Theorem 1. *If model (1) is correct and $\beta = 0$, then the estimator in (4) for the marginal mean of is $\sqrt{N}$-consistent.*

3.2. Logistic Regression Models. As an alternative to model (1), the following generalized linear model is used to model the relation between binary Z_i and Y_i:

$$\begin{aligned} E(Z_i|Y_i) &= \mu(\eta_i), \\ g(E(Z_i|Y_i)) &= \alpha + \beta Y_i, \\ \mathrm{Var}(Z_i|Y_i) &= \phi\mu'(\eta_i), \end{aligned} \tag{7}$$

where $\phi > 0$ is an unknown dispersion parameter, $\eta_i = \eta(Y_i) = \alpha + \beta Y_i$ is a linear predictor, $\mu(\eta) = e^{\eta}/(1 + e^{\eta})$ is a known differentiable function with derivative $\mu\prime(\eta) = e^{\eta}/(1 + e^{\eta})^2 \mu(1 - \mu) > 0$, and g is the logit link defined as (see the work of Nelder and Wedderburn [23] and McCullagh and Nelder [24]):

$$g(\mu) = \mathrm{log}it(\mu) = \log\left\{\frac{\mu}{(1 - \mu)}\right\}. \tag{8}$$

In particular, when Z_i is binary with values 1 and 0, $E(Z_i|Y_i) = P(Z_i = 1|Y_i)$, and one reasonable choice for $g(\cdot)$ is the logit link with $g(p) = \log\{p/(1 - p)\}$. From model (7), we obtain that

$$Y_i = \frac{\{g(\mu(\eta_i)) - \alpha\}}{\beta}, \tag{9}$$

if β is not zero. Note that by the conditional uncorrelation in model (c), we have $E(Z_i|Y_i) = E(Z_i|Y_i, R_i = 1)$ and $Var(Z_i|Y_i) = Var(Z_i|Y_i, R_i = 1)$. Therefore, model (7) can be estimated by those data in the response group. Let $(\widehat{\alpha}, \widehat{\beta})$ be such estimators of (α, β) for model (7). Then, $\mu(\eta(y)) \equiv E(Z|Y = y)$ is nonparametrically estimable via regressing Z_i on Y_i for those observed Ys with $R = 1$. We denote the resulting nonparametric estimator by $\widehat{\mu}(\eta(y)) = \widehat{E}(Z_i|Y_i = y)$, for example, using the local linear smoothing in the work of Fan and Gijbels [25] and Jiang [26]. Then, by (6), we estimate $E(Y)$ by

$$\widehat{E}(Y) = \frac{\left\{N_1^{-1}\sum_{i=1}^{N_1} g\left(\widehat{\mu}\left(\eta\left(Y_i\right)\right)\right) - \widehat{\alpha}\right\}}{\widehat{\beta}}. \tag{10}$$

However, the estimator $\widehat{E}(Y)$ may suffer from substantial loss of efficiency since it uses only a portion of sample. Furthermore, in finite sample settings, the bias of nonparametric estimator $\widehat{E}(Z_i|Y_i)$ may yield a highly biased estimate of the marginal mean of Y, since it can be magnified by the nonlinear link function. However, this approach can be used to compare with the following estimation method and justify the appropriateness of the parametric form of $\mu(\cdot)$ in (7) for modeling real data.

Note that $E(R|Y, Z) = E(R|Y)$; it follows that we can model the missing mechanism through

$$P(R = 1 \mid Y) = \frac{\exp\left(\phi_0 + \phi_1 Y\right)}{1 + \exp\left(\phi_0 + \phi_1 Y\right)}. \tag{11}$$

We here extend the previous "*simulation imputation*" approach to model (7), from which one can develop imputation of nonresponse Y_i from equation (9). Let $\varepsilon_i = \{Z_i - \mu(\eta_i)\}/\sqrt{\mu'(\eta_i)}$. Then, ε_i's are white noises of mean zero and variance ϕ and $Z_i = \mu(\eta_i) + \sqrt{\mu'(\eta_i)}\varepsilon_i$.

Specifically, it proceeds as follows:

(i) Fitting model (7) with a data subset: based on those observations from the response group, we regress $\{Z_i\}$ on $\{Y_i\}$ using model (7) and obtain the estimators $(\widehat{\alpha}, \widehat{\beta})$ of (α, β), the fitted values $\mu(\widehat{\eta}_i) = g^{-1}(\widehat{\alpha} + \widehat{\beta}Y_i)$ with $\widehat{\eta}_i = \widehat{\alpha} + \widehat{\beta}Y_i$, and Pearson's residuals.

$$\widehat{\varepsilon}_i = \frac{\{Z_i - \mu(\widehat{\eta}_i)\}}{\sqrt{\mu'(\widehat{\eta}_i)}}, \quad i = 1, \ldots, N_1. \tag{12}$$

(ii) Bootstrap residuals: we resample M samples, each with size $N2 = N - N1$, from the empirical distribution function of centered Pearson's residuals (where $\bar{\varepsilon}$ is the average of $\widehat{\varepsilon}_i$'s), and denote the M samples by $\varepsilon_{N_1+1}^{*(m)}, \ldots, \varepsilon_{N_1+N_2}^{*(m)}$, for $m = 1, \ldots, M$.

(iii) Prediction of Y_j in the nonresponse group: $Y_j^{(m)} = (Z_j - \widehat{\alpha} - \varepsilon_j^{*(m)})/\beta$, for $j = N_1 + 1, \ldots, N_1 + N_2$, and we find η_j such that $\varepsilon_j^{*(m)} = \left\{Z_j - \mu(\eta_j)\right\}/\sqrt{\mu'(\widehat{\eta}_j)}$ or equivalently

$$Z_j - \mu_j = \sqrt{\mu_j\left(1 - \mu_j\right)}\varepsilon_j^{*(m)}. \tag{13}$$

Then, $\mu_j = \varepsilon_j^{*(m)2}/\left\{1 + \varepsilon_j^{*(m)2}\right\}$ if $Z_j = 0$ and $\mu_j = 1/\left\{1 + \varepsilon_j^{*(m)^2}\right\}$ if $Z_j = 1$. The solution is denoted by $\widehat{\eta}_j^{(m)}$:

$$\widehat{\eta}_j^{(m)} = g\left(\mu_j\right) = \text{logit}\left(\mu_j\right). \tag{14}$$

The compute $Y_j^{(m)} = \left\{g(\mu(\widehat{\eta}_j^{(m)})) - \widehat{\alpha}\right\}/\widehat{\beta}$, and we use the average $Y_j^* = M^{-1}\sum_{m=1}^{M} Y_j^{(m)}$ as prediction of Y_j.

(iv) Estimator of the marginal mean of Y:

$$\widehat{E}(Y) = \omega_1 N_1^{-1}\sum_{i=1}^{N_1} Y_i + (1 - \omega_1)N_2^{-1}\sum_{j=N_1+1}^{N_1+N_2} Y_j^*. \tag{15}$$

The abovementioned bootstrap Pearson's residuals procedure in (i) and (ii) is standard in the generalized linear models (see, for example, page 341 of the work of Shao and Tu [27]). The "*simulation imputation*" method uses bootstrap data from Pearson's residuals to yield predicted values for the nonresponses.

Theorem 2. *If model (7) is correct and $\beta \neq 0$, then the result in Theorem 1 continues to hold.*

3.3. Monotone Nonlinear Models. We consider the following monotone nonlinear relation between Z_i and Y_i:

$$Z_i = f\left(Y_i\right) + \varepsilon_i, \tag{16}$$

where $f(\cdot)$ is a known monotone nonlinear function and $E(\varepsilon_i|Y_i) = 0$. By reversing the relation in (16), we get $Y_i = f^{-1}(Z_i - \varepsilon_i)$. Using a simulation imputation method similar to that in Section 3.1, we can obtain the estimator of $E(Y_i)$. This model requires one to find a surrogate Z for Y such that $E(Z|Y)$ is monotone.

Up to now, one may wonder if the models in (1), (7), and (16) are appropriate in practice. This involves in model selection and diagnostic analysis. Traditional model diagnostic tools are useful for this problem. The problem can also be addressed by nonparametrically modeling those responses with $R_i = 1$ in the exploration analysis stage if a moderate sample is available, so that one can test if some parametric model holds, based on a nonparametric testing statistics such as the generalized likelihood ratio test in the work of Fan, Zhang, and Zhang [28], Fan and Jiang [29], and Fan and Jiang [30].

4. Extension

Previous results cannot cope well with the cases in the presence of other covariates. We here extend them in the framework of the generalized linear models. The extension to other models can be similarly made. In parallel with model (7), we consider the following model with completely observed covariates X of dimension d:

$$\begin{cases} E(Z_i|Y_i, X_i) = \mu(\eta_i), \\ g(E(Z_i|Y_i, X_i)) = \alpha + \beta Y_i + \gamma^T X_i, \\ Var(Z_i|Y_i, X_i) = \phi\mu'(\eta_i), \end{cases} \tag{17}$$

where $\eta_i = \alpha + \beta Y_i + \gamma^T X_i$ is a linear predictor, g is a canonical link function, and $\mu(\cdot)$ is a known differentiable function with derivative $\mu'(\cdot) > 0$.

As in Section 3.2, the same "simulation imputation" approach can be used to estimate the mean of Y, but with $\widehat{\alpha}$ in Section 3.2 replaced by $\widehat{\alpha} + \widehat{\gamma}^T X_i$ in the present setting. This approach will be readdressed in our real example.

5. Empirical Standard Errors

The empirical standard error (*ESE*) is used to monitor the accuracy of convergence. For the *simulation imputation* methods mentioned above, the sampling number M is generally required to be large enough to ensure the prediction for the nonresponses with accuracy in a reasonable range. For a specific application, what should M be? Naturally, a good choice of M should yield an accurate predicted value of the nonresponse. This motivates us to use the *ESE*s of the simulation imputations to assess the accuracy of the predicted value of the nonresponse. In our real data analysis, M is taken as 1000, and the *ESE*s of the imputations Y_j^* are reported.

6. Simulations

To investigate the performance of our procedure, we compare our estimators with the observed average of the responses in the following three models. The number of simulations is 600 and the number of bootstrapping samples for each simulation is taken as $M = 1000$. The deviation, $\{\widehat{E}(Y) - E(Y)\}$, of each estimator from the true mean in the 600 simulations is computed and displayed via box plots, which depict the distributions of the deviations. The more the deviation concentrates at 0, the better the corresponding estimator is.

Example 2. We consider a linear model with $P(R_i = 1) = 0.5$, $Y_{i1} \sim N(0, 2^2)$, $Y_{i2} \sim N(1, 2^2)$, $Y_i = Y_{i1}R_i + Y_{i2}(1 - R_i)$, $Z_i = \alpha + \beta Y_i + \varepsilon_i$, and $\varepsilon_i \sim N(0, 0.5^2)$. Sample size is $N = 100$, where $\alpha = 1$ and $\beta = 0.5$ Thus, the true mean $E(Y)$ equals 0.5.

Example 3. Let $P(R_i = 1) = 0.5$, $Y_{i1} \sim N(0, 2^2)$, $Y_{t2} \sim N(-4, 2^2)$, and $Y_i = Y_{i1}R_i + Y_{i2}(1 - R_i), i = 1, \cdots, 100$, which is mixed normal. Then, $E(Y) = -2$. We generate binary Z_i with outcomes 0 and 1 from the logistic model.

$$P(Z_i = 1|Y_i) = \frac{\exp(\alpha + \beta Y_i)}{\{1 + \exp(\alpha + \beta Y_i)\}} \tag{18}$$

where $\alpha = 3$ and $\beta = 1$.

Example 4. Consider the monotone nonlinear model with $f(y) = a + y/(b + y)$, where $a = 0.5$ and $b = 1$. For $i = 1, 100$, let $P(R_i = 1) = 0.5$, $Y_{i1} \sim N(0, 1)$, $Y_{i2} \sim N(2, 1)$, and $Y_i = Y_{i1}R_i + Y_{i2}(1 - R_i)$, which is mixed normal. The marginal mean of Y is $E(Y) = 1$.

The boxplots of the deviations among 600 simulations for Examples 1–3 are reported in Figures 1(a)–1(c), respectively. The abovementioned three examples show that the proposed estimator is consistent, but the observed average of the response (with $R_i = 1$) is very biased because it ignores the information from the nonresponse subjects.

7. Sensitivity Analysis

For identification of model (c), we have assumed R_i is uncorrelated to Z_i conditional on Y_i. This assumption is untestable and made with knowledge of the specific experts. Naturally, one may ask if our method exhibits robustness to some extent against the assumption. This motivates us to assess the sensitivity of our estimators to the assumption.

Example 5. To assess the sensitivity of our estimators to the assumption on uncorrelation of Z_i with R_i conditional on Y_i, we set samples size $N = 100$, $P(R_i = 1) = 0.5$, $Y_{i1} \sim N(0, 1)$, $Y_{i2} \sim N(3, 1)$, $Y_i = Y_{i1}R_i + Y_{i2}(1 - R_i)$, $\varepsilon_i = 0.5\,N(0, 1)$, and $Z_i = 1 + 0.5Y_i - S * R_i + \varepsilon_i$. We consider different values of S at grid points on the interval $[-0.2, 0.2]$. For each given S, the deviations of the estimators among 600 simulations were computed. The median of the deviations for each disturbing magnitude S is reported in Figure 1(d). When S increases, the conditional correlation gets stronger. Our estimator seems to perform reasonably well against the appropriate departure from the conditional uncorrelation, but the sample average of observed response does not.

8. Real Data Analysis

We now use the mental health dataset to demonstrate how the proposed procedure works in a typical application.

This dataset is from a study of mental health of children in Connecticut. It was previously analyzed by Ibrahim, Lipsitz, and Horton (2001). There are totally 2486 subjects in study and six related variables:

Father: parental status of the household (father figure present = 0; no father figure present = 1).

health: physical health of the child (no health problem = 0; fair or poor health = 1). t_{rept}: teacher's report of the psychopathology of the child (normal = 0; abnormal = 1; and missing = .); p_{rept}: parents' report of the psychopathology of the child (normal = 0; abnormal = 1; and missing = .); *counts*: # of observations. *pctage*: percentage(=count/total sample size; total sample size = 2486).

Of the six variables, we choose the first five variables for analysis, since the 6th variable *pctage* is uniquely determinated by the 5th variable *counts*. The outcome of main interest, "t_{rept}," is missing for 1061(42.7%) subjects, but the variable "p_{rept}," which relates to "t_{rept}" and can serve as a surrogate for it, is observed for all subjects. Note that, as discussed in the abovementioned paper, once conditional on the surrogate "p_{rept}," the missing mechanism can be regarded as independent of the outcome "t_{rept}," although unconditionally the missing mechanism "R" (equals 0 for the missing and 1 for the observed) seems to depend on the

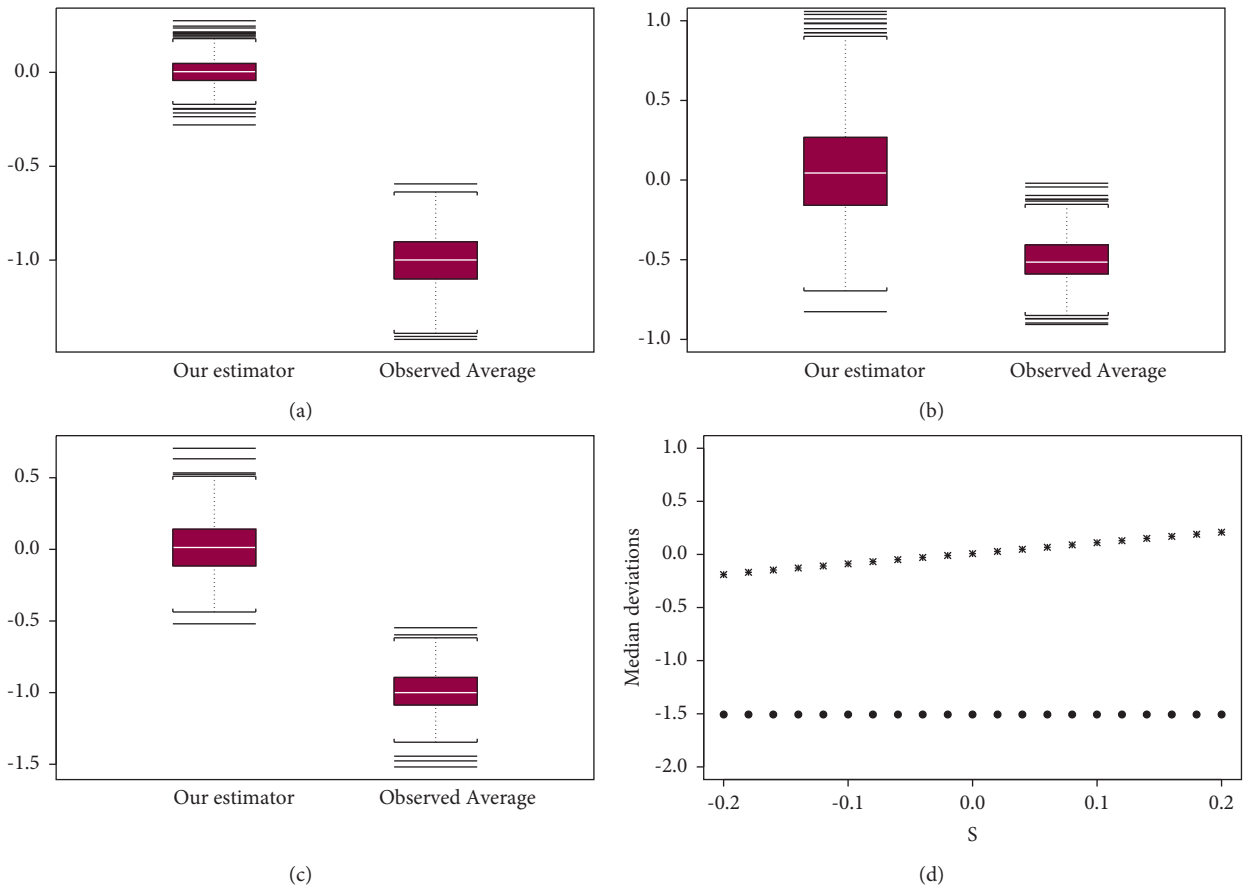


FIGURE 1: (a–c) The box plots of the distributions for the deviations of the estimators from the true mean among 600 simulations in Examples 1–3, respectively; (d) sensitivity analysis for Example 5. Dotted: the average of observed responses; –: our estimator.

outcome. Then, the model (c) seems reasonable to depict the relation among t_{rept}, p_{rept}, and R.

Our interest is to estimate the marginal mean of the response variable "t_{rept}." We use model (5) and the estimation approach in Section 3.2 and set $Z = p_{rept}$, $Y = t_{rept}$, and the number of bootstrap sampling $M = 1000$. The estimated mean, variance of the response t_{rept}, and the average of the *ESE*s of the imputations Y_j^* for the nonresponses are reported in Table 1. By incorporating the information on the first two covariates, *father* and *health*, we get the estimated mean and variance of the response from model (10), which can be found in third column of Table 1.

However, if only using the observations of t_{rept}, its marginal mean is computed to be 0.1832 with variance 0.1497. Obviously, our estimator not only rectifies the value of the marginal mean, but is more efficient than the simple average of the observed values of Y. The small values of the *ESE*s reflect high accuracy of the simulation imputations.

Appendix

Proofs of Theorems

Proof. of Theorem 1. The empirical distribution function for $i = 1, \cdots, N_1$ is denoted by $F_{N_1}(\varepsilon)$. Since $\varepsilon_j^{*(m)}$'s are drawn from $F_{N_1}(\varepsilon)$, conditional on the original sample points with $R_i = 1$, $\varepsilon_j^{*(m)} \sim F_{N1}(\varepsilon)$ for $j = N_1 + 1, \ldots, N_1 + N_2$. Hence, $E\{\varepsilon_j^{*(m)} | F_{N_1}\} = \int x\, dF_{N_1}(x) = \int x\, dF(x) + \int x\, d[F_{N_1}(x) - F(x)] = \int x\, dF(x) + O_p(1/\sqrt{N}) = O_p(1/\sqrt{N})$, since $\int x\, dF(x) = 0$. Let $\overline{\varepsilon_j^*} = M^{-1} \sum_{m=1}^{M} \varepsilon_j^{*(m)}$. Then,

$$\begin{aligned}
\sum_{j=N_1+1}^{N} Y_j^* &= \sum_{j=N_1+1}^{N} \sum_{m=1}^{M} \frac{M^{-1}\left[Z_j - \widehat{\alpha} - \varepsilon_j^{*(m)}\right]}{\widehat{\beta}}, \\
&= \sum_{j=N_1+1}^{N} \frac{\left(Z_j - \widehat{\alpha} - \overline{\varepsilon_j^{*(m)}}\right)}{\widehat{\beta}}, \\
&= \sum_{j=N_1+1}^{N} \frac{\left\{Z_j - \alpha - \varepsilon_j - (\widehat{\alpha} - \alpha) - \left(\overline{\varepsilon_j^{*(m)}} - \varepsilon_j\right)\right\}}{\widehat{\beta}}, \\
&= \sum_{j=N_1+1}^{N} \frac{\left\{\beta Y_j - (\widehat{\alpha} - \alpha) - \left(\overline{\varepsilon_j^{*(m)}} - \varepsilon_j\right)\right\}}{\widehat{\beta}}.
\end{aligned} \tag{19}$$

Note that

Table 1: Estimated mean and variance of trept.

	Marginal	Model (5)	Model (10)
Mean	0.1832	0.0956	0.1203
Variance	0.1497	0.000103	0.000104
ESEs	N/A	1.4969	1.5719

$$\widehat{\alpha} - \alpha = O_p\left(\frac{1}{\sqrt{N}}\right), \quad \widehat{\beta} - \beta = O_p\left(\frac{1}{\sqrt{N}}\right). \tag{20}$$

Using the standard argument for bootstrap consistency, it can be shown that

$$N^{-1}\sum_{j=N_1}^{N}\left(\overline{\varepsilon_j^*} - \varepsilon_j\right) = N^{-1}\sum_{j=N_1}^{N}\overline{\varepsilon_j^*} + N^{-1}\sum_{j=N_1}^{N}\varepsilon_j \equiv L_{n1} + L_{n2}. \tag{21}$$

By calculating the mean and variance, it is easy to see that

$$L_{n1} = N^{-1}\sum_{j=N_1+1}^{N} M^{-1}\sum_{m=1}^{M}\varepsilon_j^{*(m)} = N^{-1}\sum_{j=N_1+1}^{N}\widehat{E}_{F_{N_1}}\left(\varepsilon_j^{*(m)}\right) + O_p\left(\frac{1}{\sqrt{N}}\right). \tag{22}$$

It follows that

$$N^{-1}\sum_{j=N_1+1}^{N} Y_j^* = \left[N^{-1}\sum_{j=N_1+1}^{N} Y_j\right]\left(1 + O_p\left(1/\sqrt{N}\right)\right) + O_p\left(1/\sqrt{N}\right). \tag{23}$$

This, combined with (4), yields $\sqrt{N}-$ consistency of $\widehat{E}(Y)$.

Proof. of Theorem 2. The result follows from a similar argument as Theorem 1.

Data Availability

The data are public and available in the reference. They are also available upon request via emailing to the first author.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[2] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley, New York, NY, USA, 1987.

[3] R. M. Groves, S. Presser, and S. Dipko, "The role of topic interest in survey participation decisions," *Public Opinion Quarterly*, vol. 68, no. 1, pp. 2–31, 2004.

[4] J. G. Ibrahim and S. R. Lipsitz, "Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable," *Biometrics*, vol. 52, no. 3, pp. 1071–1078, 1996.

[5] R. M. Groves, "Nonresponse rates and nonresponse bias in household surveys," *Public Opinion Quarterly*, vol. 70, no. 5, pp. 646–675, 2006.

[6] D. Holt and T. M. F. Smith, "Post stratification," *Journal of the Royal Statistical Society: Series A*, vol. 142, no. 1, pp. 33–46, 1979.

[7] P. S. Kott, "Using calibration weighting to adjust for nonresponse and coverage errors," *Survey Methodology*, vol. 32, pp. 133–142, 2006.

[8] J.-C. Deville, C.-E. Särndal, and O. Sautory, "Generalized raking procedures in survey sampling," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 1013–1020, 1993.

[9] R. J. A. Little, "Models for nonresponse in sample surveys," *Journal of the American Statistical Association*, vol. 77, no. 378, pp. 237–250, 1982.

[10] J. S. Greenlees, W. S. Reece, and K. D. Zieschang, "Imputation of missing values when the probability of response depends on the variable being imputed," *Journal of the American Statistical Association*, vol. 77, no. 378, pp. 251–261, 1982.

[11] R. E. Fay, "Causal models for patterns of nonresponse," *Journal of the American Statistical Association*, vol. 81, no. 394, pp. 354–365, 1986.

[12] S. G. Baker and N. M. Laird, "Regression analysis for categorical variables with outcome subject to nonignorable nonresponse," *Journal of the American Statistical Association*, vol. 83, no. 401, pp. 62–69, 1988.

[13] R. J. A. Little, "Nonresponse adjustments in longitudinal surveys: models for categorical data," *Bulletin of the International Statistical Institute*, vol. 15, pp. 1–15, 1985.

[14] J. K. Kim and C. L. Yu, "A semiparametric estimation of mean functionals with nonignorable missing data," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 157–165, 2011.

[15] M. K. Riddles, J. K. Kim, and J. Im, "A propensity-score-adjustment method for nonignorable nonresponse," *Journal of Survey statistics and Methodology*, vol. 4, no. 2, pp. 215–245, 2016.

[16] P. Zhao, N. Tang, A. Qu, and D. Jiang, "Semiparametric estimating equations inference with nonig- norable missing data," *Statistica Sinica*, vol. 27, pp. 89–113, 2017.

[17] F. Guo, W. Ma, and L. Wang, "Semiparametric estimation of copula models with nonignorable missing data," *Journal of Nonparametric Statistics*, vol. 32, no. 1, pp. 109–130, 2020.

[18] W.-Q. Ma, Z. Geng, Z. Geng, and Y.-H. Hu, "Identification of graphical models for nonignorable nonresponse of binary outcomes in longitudinal studies," *Journal of Multivariate Analysis*, vol. 87, no. 1, pp. 24–45, 2003.

[19] X. Wang, H. Chen, Z. Geng, and X. Zhou, "Using auxiliary data for binomial parameter estimation with nonignorable nonresponse," *Communications in Statistics - Theory and Methods*, vol. 41, no. 19, pp. 3468–3478, 2012.

[20] J. J. Forster and P. W. F. Smith, "Model-based inference for categorical survey data subject to non-ignorable non-response," *Journal of the Royal Statistical Society: Series B*, vol. 60, no. 1, pp. 57–70, 1998.

[21] W. D. Kalsbeek, "A conceptual review of survey error due to nonresponse," *American Statistical Associ- ation Proceedings of the Section on Survey Research Method*, vol. 43, pp. 131–136, 1980.

[22] B. Efron, "Bootstrap methods: another look at the jackknife," *The Annals of Statistics*, vol. 7, pp. 1–26.

[23] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society: Series A*, vol. 135, no. 3, pp. 370–384, 1972.
[24] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Chapman & Hall, London, UK, 1989.
[25] J. Fan and I. Gijbels, "Variable bandwidth and local linear regression smoothers," *Annals of Statistics*, vol. 20, 1992.
[26] J. Jiang, "Multivariate functional-coefficient regression models for nonlinear vector time series data," *Biometrika*, vol. 101, pp. 689–702, 2014.
[27] J. Shao and D. Tu, *The Jackknife and Bootstrap*, Springer-Verlag, Berlin, Germany, 1995.
[28] J. Fan, C. M. Zhang, and J. Zhang, "Generalized likelihood ratio statistics and wilks phe- nomenon," *The Annals of Statistics*, vol. 29, pp. 153–193, 2001.
[29] J. Fan and J. Jiang, "Nonparametric inferences for additive models," *Journal of the American Statistical Association*, vol. 100, no. 471, pp. 890–907, 2005.
[30] J. Fan and J. Jiang, "Nonparametric inference with generalized likelihood ratio tests," *Test*, vol. 16, no. 3, pp. 409–444, 2007.