Identification and Inference With Nonignorable Missing Covariate Data

Wang Miao and Eric Tchetgen Tchetgen

Peking University and Harvard University

July 27, 2017

Abstract

We study identification of parametric and semiparametric models with missing covariate data. When covariate data are missing not at random, identification is not guaranteed even under fairly restrictive parametric assumptions, a fact that is illustrated with several examples. We propose a general approach to establish identification of parametric and semiparametric models when a covariate is missing not at random. Without auxiliary information about the missingness process, identification of parametric models is strongly dependent on model specification. However, in the presence of a fully observed shadow variable that is correlated with the missing covariate but otherwise independent of the missingness conditional on the covariate, identification is more broadly achievable, including in fairly large semiparametric models. Special consideration is given to the generalized linear models with the missingness process unrestricted. Under such a setting, the outcome model is identified for a number of familiar generalized linear models, and we provide counterexamples when identification fails. For estimation, we describe an inverse probability weighted estimator that incorporates the shadow variable to estimate the propensity score model, and we evaluate its performance via simulations. We further illustrate the shadow variable approach with a real data example about home price in China.

Keywords: Identification; Missing covariate data; Missing not at random; Shadow variable.

1. INTRODUCTION

Missing data are commonly encountered in many socioeconomic and biomedical studies. Methods to account for missing outcome data in regression analysis figure prominently in the literature. However, missing covariate data is also a longstanding problem in applied research. In the early history of missing data analysis, Glasser (1964), Afifi & Elashoff (1966) and Haitovsky (1968) studied the missing covariate problem in regression analysis; Edgett (1956), Anderson (1957) and Buck (1960) studied the problem in the context of multivariate analysis. Rubin (1976) formalized the concept of missing data mechanism as a separate process from the full data law of primary scientific interest. The missing data mechanism is said missing at random, if it is independent of missing values after conditioning on the observed data, and it is said missing not at random otherwise. For analysis of data missing at random, there currently exist a variety of methods such as likelihood-based approaches (Dempster et al., 1977; Horton & Laird, 2001; Ibrahim, 1990), imputation and multiple imputation (Rubin & Schenker, 1986; Vach & Schumacher, 1993; Rubin, 1987), and semiparametric methods (Zhao et al., 1996; Robins et al., 1994).

In many empirical studies, however, covariate data will often be missing not at random, i.e., the missingness is related to missing covariate values even after conditioning on the observed data. Most of the aforementioned methods have previously been adapted to deal with covariate data missing not at random. Comprehensive reviews of statistical research on this topic include Ibrahim et al. (1999), Little & Zhang (2011) and Ibrahim et al. (2005). Validity of existing estimation methods relies on first establishing identification. Identification means that the parameter of interest, for example, the regression coefficient of the outcome on the missing covariate, is uniquely determined by the observed data. Without identification, statistical inference is generally of limited interest and may often be misleading. Under missingness at random, the joint distribution of all variables of interest is identified without parametric assumptions (Little & Rubin, 2002); however, as pointed out by Baker & Laird (1988), under missingness not at random, identification is not always guaranteed. Fay (1986) and Ma et al. (2003) used graphical models to represent missingness mechanisms, and they studied identification for longitudinal categorical variables that are missing not at random. In the context of missing outcome data, Tang et al. (2003), Wang et al. (2014), Zhao & Shao (2015) and Miao et al. (2017) studied identification of several parametric and semiparametric models, and they presented counterexamples when identification fails; Kott (2014), Wang et al. (2014) and D'Haultfoeuille (2010) noted that a fully observed shadow variable can sometimes be used to improve identification under missingness not at random, and we have previously demonstrated identification of a class of location-scale models with a shadow variable (Miao et al., 2015). Such a variable is associated with the potentially unobserved variable conditional on the observed data but independent of the missingness process conditional both on the observed data and missing variable (Kott, 2014).

Identification is also crucial and challenging for covariate data missing not at random; yet the literature on this topic is somewhat sparse. In this paper, we illustrate difficulty for identification of nonginorable missing covariate data in section 2. We establish a general framework for studying identification with missing covariate data in section 3 and we illustrate with several parametric models. In section 4, we use a shadow variable for the missing covariate to improve identification in semiparametric models where the missingness process is unspecified, and we establish identification conditions of a large family of the generalized linear models. In section 5, we describe an inverse probability weighted estimator, which incorporates the shadow variable to estimate the nonignorable missingness process. We evaluate its performance via simulations in section 6, and further illustrate with a real data example about home price in China. In section 7 we include some discussions, such as the difference between the missing outcome problem and the missing covariate problem.

2. POTENTIAL DIFFICULTY FOR IDENTIFICATION

Throughout, we let Y denote the fully observed outcome variable and (X, Z) the vector of covariates with Z fully observed and X subject to missingness. We let R denote the missing indicator of X with $R = 1$ if X is observed and $R = 0$ otherwise. For notational convenience, we suppress Z in this section. The observed data include (Y, R) for all samples, and X only for those with $R = 1$. The goal of missing data analysis is to make inference about the full data distribution $\text{pr}(x, y)$ and the missingness process (or propensity score) $\text{pr}(r = 1 \mid x, y)$, based on the observed data distribution that is captured by $\mathrm{pr}(y, r = 0)$ and $\mathrm{pr}(x, y, r = 1)$. Recovery of the full data law and the missingness process from the observed data distribution is the fundamental identification challenge in missing data problems. It can be formally expressed as below.

Definition 1. For a model $\mathrm{pr}(x, y, r; \theta)$ indexed by θ that may have a finite dimensional component as well as nonparametric components, the parameter θ is said to be identified from the observed data, if there exists a one-to-one mapping between the parameter space $\Theta = \{\theta\}$ and the space of observed data distribution $\{\mathrm{pr}(y, r = 0; \theta), \mathrm{pr}(x, y, r = 1; \theta); \theta \in \Theta\}$.

When data are missing at random, i.e., $R \perp\!\!\!\perp X \mid Y$, it is well known that the joint distribution $\mathrm{pr}(x, y, r)$ is nonparametrically identified, because $\mathrm{pr}(x, y, r = 0) = \mathrm{pr}(x \mid y, r = 0)\mathrm{pr}(y, r = 0)$ and $\mathrm{pr}(x \mid y, r = 0) = \mathrm{pr}(x \mid y, r = 1)$. When data are missing not at random, however, $\mathrm{pr}(x \mid y, r = 0) \neq \mathrm{pr}(x \mid y, r = 1)$, and thus one cannot ignore the missing data mechanism to make inference (Little & Rubin, 2002; Ibrahim et al., 1999). As shown in the following example, even when fairly restrictive parametric models are correctly specified for $\mathrm{pr}(x, y, r)$, identification is not guaranteed, and selection bias due to missing data cannot necessarily be eliminated.

Example 1. We consider a joint normal model encoded in $\mathrm{pr}(x) \sim N(\gamma, \lambda)$ and $\mathrm{pr}(y \mid x) \sim N(\beta_0 + \beta_1 x, \phi)$, and a logistic propensity score model: $\mathrm{logit}\ \mathrm{pr}(r = 1 \mid x, y) = \alpha_0 + \alpha_1 x + \alpha_2 y$. Letting $(\lambda, \phi, \beta_1) = (1.25, 0.8, 0.4)$, one can verify that the two sets of parameters $(\gamma, \beta_0, \alpha_0, \alpha_1, \alpha_2) = (0, 0, 2, -2, 1)$ and $(2, -0.8, -2, 2, -1)$ result in identical observed data distribution $\mathrm{pr}(y, r = 0)$ and $\mathrm{pr}(x, y, r = 1)$. Therefore, $(\gamma, \beta_0, \alpha_0, \alpha_1, \alpha_2)$ are not identified from the observed data.

The normal-logistic model is commonly used in missing data analysis. However, Example 1 shows potential lack of identification of such models, when the missingness process depends on the potentially unobserved covariate. Without identification, there may exist different laws of the full data that have identical observed data distribution, and thus even at asymptopia one cannot determine which is the truth based on the observed data. In this case, estimation for the parameters is of limited interest in practice. Despite its importance, identification for missing covariate data is not extensively studied in the literature. In subsequent sections, we propose a general framework to establish identification of regression analysis with missing covariate data.

3. A GENERAL FRAMEWORK FOR IDENTIFICATION

We consider a model $\mathrm{pr}(x, y, z, r; \theta)$ indexed by θ. Without loss of generality, we assume

Assumption 1. There exists a one-to-one mapping between the parameter space $\Theta = \{\theta\}$ and the joint distribution space $\{\mathrm{pr}(x, y, z, r; \theta); \theta \in \Theta\}$.

Our goal is to identify θ or $\mathrm{pr}(x, y, z, r; \theta)$ from the observed data distribution that is captured by $\{\mathrm{pr}(y, z, r = 0; \theta), \mathrm{pr}(x, y, z, r = 1; \theta)\}$, which is equivalent to $\{\mathrm{pr}(z; \theta), \mathrm{pr}(y \mid z; \theta), \mathrm{pr}(x, y, r = 1 \mid z; \theta)\}$. Suppose θ_1 and θ_2 are two candidate values falling in the parameter space of θ and result in identical distribution of the observed data:

$$\begin{aligned} \mathrm{pr}(z; \theta_1) &= \mathrm{pr}(z; \theta_2) \\ \mathrm{pr}(y \mid z; \theta_1) &= \mathrm{pr}(y \mid z; \theta_2), \\ \mathrm{pr}(x, y, r = 1 \mid z; \theta_1) &= \mathrm{pr}(x, y, r = 1 \mid z; \theta_2). \end{aligned}$$

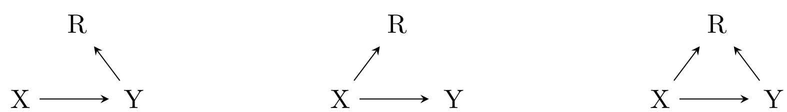
The three equations characterize all values at which identification fails and which must be ruled out to guarantee identification. We have the following condition for identification.

Condition 1. The parameter θ is identified if for any two candidate values θ_1 and θ_2 such that $\mathrm{pr}(z; \theta_1) = \mathrm{pr}(z; \theta_2)$ and $\mathrm{pr}(y \mid z; \theta_1) = \mathrm{pr}(y \mid z; \theta_2)$ almost surely, the following inequality holds with a positive probability

$$\frac{\mathrm{pr}(x, y \mid z; \theta_1)}{\mathrm{pr}(x, y \mid z; \theta_2)} \neq \frac{\mathrm{pr}(r = 1 \mid x, y, z; \theta_2)}{\mathrm{pr}(r = 1 \mid x, y, z; \theta_1)}. \tag{1}$$

Condition 1 is a sufficient condition for identification. Inequality (1) involves the missingness process, which provides a convenient approach to check identification for selection models where separate parametric/semiparametric models are specified for the propensity score $\mathrm{pr}(r = 1 \mid x, y, z)$ and the full data distribution $\mathrm{pr}(x, y, z)$. In subsequent sections, we focus on identification under the selection model parametrization, and in the Supplementary Material we extend results to the pattern-mixture parametrization (Little, 1993). Here we provide several examples to illustrate how to apply Condition 1 in selection models. For notational convenience, we suppress Z in these examples.

Example 2. We verify identification of the missing at random mechanism: $R \perp\!\!\!\perp X \mid Y$ by checking Condition 1. Following the approach of Fay (1986), such a missingness mechanism can also be encoded in the directed acyclic graph model of Figure 1 (i), where the arrow between X and R is not present. It is plausible in a retrospective study such as a case control study in which X is ascertained only after Y is determined, so that Y may in fact directly influence whether or not X is missing. For any two candidate models $\text{pr}(x, y, r; \theta_1)$ and $\text{pr}(x, y, r; \theta_2)$ such that $\text{pr}(y; \theta_1) = \text{pr}(y; \theta_2)$, the ratio of the propensity score models $\text{pr}(r = 1 \mid y; \theta_1)/\text{pr}(r = 1 \mid y; \theta_2)$ is a function only of y. However, $\text{pr}(x, y; \theta_1)/\text{pr}(x, y; \theta_2)$ must vary with x and thus (1) holds. Therefore, θ is identified according to Condition 1.



(i) MAR (ii) MNAR (iii) MNAR

Figure 1: Directed acyclic graph models for different missingness mechanisms for x.

Example 3. Bartlett et al. (2014) considered estimation under the missingness mechanism encoded in the graph model of Figure 1 (ii), which is missingness not at random. The graph depicts a prospective study in which Y is ascertained only after X is observed, and therefore, it is reasonable to assume that Y cannot determine whether X is missing, provided that a participant is not able to anticipate her outcome at baseline. Considering an outcome model $\text{pr}(y \mid x, \theta)$, for any θ_1, θ_2, the ratio $\{\text{pr}(y \mid x; \theta_1)\text{pr}_1(x)\}/\{\text{pr}(y \mid x; \theta_2)\text{pr}_2(x)\}$ must vary with y, and thus cannot equal the ratio of two propensity score models, a function only of x. Therefore, θ indexing the outcome model $\text{pr}(y \mid x; \theta)$ is identified, although, the covariate distribution $\text{pr}(x)$ may not be.

When the missingness process depends on either the missing covariate X (Example 3) or the fully observed outcome Y (Example 2), identification is well established (Little & Rubin, 2002), thus, we have simply confirmed identification by verifying Condition 1. We further provide several examples to illustrate the case where missingness depends both on X and Y.

Example 4. In empirical studies, the covariate and outcome of interest are often binary. We note that identification is not guaranteed for binary variables when the missingness depends both on X and Y, i.e., the missingness mechanism encoded in Figure 1 (iii). Considering the logistic models for binary X and Y:

$$\begin{aligned}\text{logit pr}(y = 1 \mid x; \beta) &= \beta_0 + \beta_1 x,\\ \text{logit pr}(r = 1 \mid x, y; \alpha) &= \alpha_0 + \alpha_1 x + \alpha_2 y,\end{aligned}$$

one can verify that $\text{pr}(y = 1)$ and $\text{pr}(r = 1, x \mid y)$ are identical under these two settings: $\alpha = (-0.4, -0.4, 0.2)$, $\beta = (-0.359, 0.6)$, $\text{pr}(x = 1) = 0.597$, and $\alpha' = (0.468, -1.64, 0.338)$, $\beta' = (-0.361, 0.488)$, $\text{pr}'(x = 1) = 0.737$. As a result, the parameters are not identified.

In the binary example, one can also follow the "parameter counting" approach to check identification (Baker & Laird, 1988). In Example 4, the model contains six unknown parameters: (α, β) and $\text{pr}(x = 1)$, but the observed data distribution only has five degrees of freedom: $\text{pr}(x, y, r = 1)$ for $x, y = 0$ or 1 and $\text{pr}(y = 1, r = 0)$, which provides five estimating equations of the unknown parameters. Thus, the solution for $\{\alpha, \beta, \text{pr}(x = 1)\}$ is not unique with more parameters than estimating equations, i.e., the parameters are not identified. For a continuous covariate or a semiparametric model, however, the number of unknown parameters and degrees of freedom of the observed data are difficult to characterize. In this case, the "parameter counting" approach often does not apply. But, it is quite convenient to apply Condition 1 for parametric models.

Example 5. Continuation of Example 1. The missingness mechanism can be encoded in the graph of Figure 1 (iii). The model for the joint distribution is indexed by $\theta = (\gamma, \lambda, \alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2, \phi)$. Considering the respective models indexed by θ and θ', we have

$$\log \frac{\text{pr}(x, y; \theta)}{\text{pr}(x, y; \theta')} = -\frac{(y - \beta_0 - \beta_1 x)^2}{2\phi} + \frac{(y - \beta_0' - \beta_1' x)^2}{2\phi'} - \frac{(x - \gamma)^2}{2\lambda} + \frac{(x - \gamma')^2}{2\lambda'}, \quad (2)$$

which is a linear combination of y^2, y, x^2, xy and x; and we have

$$\log \frac{\text{pr}(r = 1 \mid y, x; \theta')}{\text{pr}(r = 1 \mid y, x; \theta)} = \alpha_0' + \alpha_1' x + \alpha_2' y + \log \frac{1 + \exp\{-\alpha_0 - \alpha_1 x - \alpha_2 y\}}{1 + \exp\{\alpha_0' + \alpha_1' x + \alpha_2' y\}}. \quad (3)$$

For $(\alpha_0, \alpha_1, \alpha_2) = -(\alpha_0', \alpha_1', \alpha_2')$, (3) is a linear combination of x and y. Thus, (2) and (3) can equal for certain values of the parameters such as those given in Example 1, i.e., $\text{pr}(r = 1, x, y; \theta) = \text{pr}(r = 1, x, y; \theta')$. We can further verify that $\text{pr}(y)$ is identical under those two settings. Therefore, θ cannot be identified. In particular, $(\gamma, \alpha_0, \alpha_1, \alpha_2, \beta_0)$ cannot be identified, but (λ, ϕ, β_1) can be identified by noting that when (2) equals (3), the coefficients of y^2, xy and x^2 must be zero in (2).

Examples 1 and 5 show potential lack of identification for the normal model when the covariate is missing not at random. In this case, the slope of the outcome model, i.e., β_1 is identified but the intercept β_0 is not. In contrast to the normal model, the following example establishes identification of a certain exponential regression model.

Example 6. Consider a normal model for the covariate: $X \sim N(\mu, \sigma_1^2)$, an exponential regression model for the outcome variable: $Y \sim \eta(x) \exp\{-y\eta(x)\}$ with $\eta(x) = \exp(\beta_0 + \beta_1 x)$ and $\beta_1 \neq 0$, and a logistic propensity score model: $\text{logit pr}(r = 1 \mid x, y) = \alpha_0 + \alpha_1 x + \alpha_2 y$, then all parameters are identified.

In a breast cancer study, Lipsitz et al. (1999) applied the Weibull regression $Y \sim \sigma_2 y^{\sigma_2 - 1} \exp\{-y^{\sigma_2}\eta(x) + \log(\eta(x))\}$ to model the time to treatment failure, without formally establishing identification of the model. The Weibull regression model is more general than the exponential regression model. We show in the appendix that identification does hold as well.

4. IDENTIFICATION WITH A SHADOW VARIABLE

In Examples 1–6, identification or lack thereof is completely determined by the specific parametric model being considered, and therefore, it is unclear whether a general identification framework is available without all of the restrictions on the models. However, when a shadow variable for the missing covariate is fully observed, identification is often possible even in fairly large semiparametric models. A shadow variable is associated with the potentially missing variable conditional on the observed data, but independent of the missingness process conditional both on the observe data and the potentially missing variable (Kott, 2014). The following definition formalizes these conditions.

Definition 2. A fully observed variable Z is a shadow variable for X, if $Z \not\perp\!\!\!\perp X \mid Y$ and $Z \perp\!\!\!\perp R \mid (Y, X)$.

Definition 2 formalizes the idea that the shadow variable affects the missingness only through its association with the missing covariate and the fully observed outcome. Figure 2 is a directed acyclic graph encoding the definition.

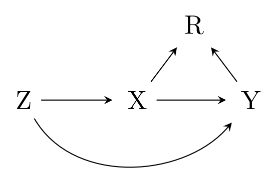


Figure 2: A directed acyclic graph model for the shadow variable.

The shadow variable for a missing covariate may be available in many empirical studies, where a fully observed proxy or a mismeasured version of the missing covariate is available. For example, in a study of mental health of children in Connecticut (Zahner et al., 1992; Horton & Laird, 2001), researchers were interested in the correlation between children's mental health status and utilization of mental health service. The measure of psychopathology used in the study was based on the teacher's assessment that had 43% missing values, however, a separate parental report was complete. The parental report is a proxy for the teacher's assessment, but it is unlikely to be related to the teacher's response rate conditional on other covariates and her assessment of the student; in this case the parental assessment constitutes a valid shadow variable. Such a variable introduces additional restrictions on the missingness process, and thus provides better opportunity for identification under missingness not at random. For example, non-identification of the binary case (Example 4) is completely resolved with a binary shadow variable.

Example 7. Continuation of Example 4. Suppose Z is a valid shadow variable for X. Because $Z \perp\!\!\!\perp R \mid (X, Y)$, we have $\mathrm{pr}(z \mid x, y) = \mathrm{pr}(z \mid x, y, r = 1)$ for all (y, x, z). For arbitrary y, one can solve the linear equation $\mathrm{pr}(z \mid y) = \sum_x \mathrm{pr}(z \mid x, y, r = 1)\mathrm{pr}(x \mid y)$ for $\mathrm{pr}(x \mid y)$. Note that $Z \not\perp\!\!\!\perp X \mid Y$, the solution is unique, i.e., $\mathrm{pr}(x \mid y)$ is identified. One can further solve $\mathrm{pr}(r = 1, x \mid y) = \mathrm{pr}(r = 1 \mid x, y)\mathrm{pr}(x \mid y)$ to identify the propensity score $\mathrm{pr}(r = 1 \mid x, y)$. Thus, one can identify the joint distribution $\mathrm{pr}(x, y, z, r) = \mathrm{pr}(y)\mathrm{pr}(x \mid y)\mathrm{pr}(z \mid x, y)\mathrm{pr}(r = 1 \mid x, y)$. See the appendix for additional details.

It has been previously noted by Ma et al. (2003) that for the binary case of Example 7, the joint distribution $\mathrm{pr}(x, y, z, r)$ can be identified explicitly as a function of the observed data distribution when a binary shadow variable is available. For more complicated models such as a semiparametric model with a continuous covariate, identification is not as straightforward as that of Example 7. For such cases, the following proposition is convenient to check identification of the outcome model, even the propensity score model is nonparametric.

Proposition 1. Consider models $\mathrm{pr}(y \mid x, z; \theta)$ and $\mathrm{pr}(x|z; \xi)$, if for any $\theta_1 \neq \theta_2$ and for all function $h(x, y)$, $\mathrm{pr}(x, y|z; \theta_1, \xi_1)/\mathrm{pr}(x, y|z; \theta_2, \xi_2) \neq h(x, y)$ with a positive probability, then the parameter θ indexing the outcome model $\mathrm{pr}(y \mid x, z; \theta)$ is identified.

The proposition follows from the fact that under the shadow variable assumption, the ratio of any two different propensity score models is not a function of z, and thus from Condition 1, θ must be identified if the ratio $\mathrm{pr}(x, y \mid z; \theta_1, \xi_1)/\mathrm{pr}(x, y \mid z; \theta_2, \xi_2)$ varies with z for distinct values θ_1 and θ_2. We further consider identification for the generalized linear models that are commonly used in practice. We suppose X and Z are continuous variables, and we assume the following models:

$$\mathrm{pr}(x \mid z; \gamma, \lambda) = \exp\left\{\frac{x \cdot \eta_1 - B_1(\eta_1)}{\lambda} + A_1(x, \lambda)\right\}, \tag{4}$$

$$\mathrm{pr}(y \mid x, z; \beta, \phi) = \exp\left\{\frac{y \cdot \eta_2 - B_2(\eta_2)}{\phi} + A_2(y, \phi)\right\}, \tag{5}$$

with dispersion parameters $\phi, \lambda > 0$, and known functions $A_1, A_2, B_1, B_2, \eta_1(z; \gamma) = \eta_1(\gamma_0 + \gamma_1 z)$ and $\eta_2(x, z; \beta) = \eta_2(\beta_0 + \beta_1 z + \beta_2 x)$. We assume that the functions are infinitely often differentiable and that for all (γ, λ) in the parameter space, the exponential family $\mathrm{pr}(x \mid z; \gamma, \lambda)$ is full rank (Shao, 2003, page 96), i.e., the range of $\eta_1(z; \gamma)/\lambda$ contains an open set. Note that the propensity score model is unspecified except for $Z \perp\!\!\!\perp R \mid (Y, X)$. We have the following identification results for such models.

Theorem 1. Suppose Z is a shadow variable and assume the generalized linear models (4)–(5), we have

(a) if η_2 is a linear function, then β_1/ϕ is identified;

(b) if η_2 a linear function, and $B_2^{(2)}$, the second-order derivative of B_2 is not a linear function, then (β_1, β_2, ϕ) are identified;

(c) if η_2 is a nonlinear function, then (β_1, β_2) are identified.

In the Supplementary Material, we prove Theorem 1 by verifying the condition of Proposition 1. The theorem establishes identification of the coefficients of Z and X in the outcome model $\text{pr}(y \mid x, z)$ except when η_2 is a linear function and B_2 is a cubic or quadratic function. From Theorem 1, (β_1, β_2) of the logistic model

$$\text{pr}(y \mid x, z; \beta) = \exp\{y(\beta_0 + \beta_1 z + \beta_2 x) - \log\{1 + \exp(\beta_0 + \beta_1 z + \beta_2 x)\}\},$$

must be identified. When η_2 is a linear function and B_2 is a quadratic function, i.e., $\text{pr}(y \mid x, z)$ follows a normal model, we observe that even though Z is correlated with X, Z may be independent of X after conditioning on Y, i.e., the shadow variable assumption is not met. We have the following counterexample for identification of the normal model.

Example 8. Consider the normal models $\text{pr}(y \mid x, z) = N(\beta_1 z + \beta_2 x, \phi)$ and $\text{pr}(x \mid z) = N(\gamma_1 z, \lambda)$ indexed by $\theta = (\beta_1, \beta_2, \phi, \gamma_1, \lambda)$. For the two sets of values $\theta_1 = (1, 1, 1, 1, 1)$ and $\theta_2 = (1.5, 0.5, 1.5, 1, 2)$, one can verify

$$\frac{\text{pr}(x, y \mid z; \theta_1)}{\text{pr}(x, y \mid z; \theta_2)} = \exp\left\{-\frac{1}{2}\log(3) - \frac{1}{6}(y - 2x)^2\right\},$$

which does not vary with z. Consider the following two models for the missingness process:

$$\text{logit pr}_2(r = 1 \mid x, y) = -\text{logit pr}_1(r = 1 \mid x, y) = -\frac{1}{2}\log(3) - \frac{1}{6}(y - 2x)^2,$$

then one can verify that the two data generating mechanisms, encoded in $\text{pr}(x, y \mid z, \theta_i)$ and $\text{pr}_i(r = 1 \mid x, y)$ for $i = 1, 2$, have identical observed data distribution. Thus, θ is not identified from the observed data. But $\beta_1/\phi = 1$ is identified, a fact that is consistent with Theorem 1 (a).

The example shows potential lack of identification of normal models. But it should be noted that non-identification only happens at certain values of the parameter space.

Theorem 2. For the normal models $\mathrm{pr}(y \mid x, z) = N(\beta_0 + \beta_1 z + \beta_2 x, \phi)$ and $\mathrm{pr}(x \mid z) = N(\gamma_0 + \gamma_1 z, \lambda)$, all parameters are identified if $\beta_1\beta_2/\phi - \gamma_1/\lambda \neq 0$.

From Theorem 2, normal models are generally identifiable except for a specific subset of the parameter space. The condition $\beta_1\beta_2/\phi - \gamma_1/\lambda \neq 0$ in fact characterizes the subset of data generating mechanisms that violate the shadow variable assumption. The following submodels offer better identification results as they involve fewer parameters than models (4)–(5).

$$\eta_2(x, z; \beta) = \eta_2(\beta_0 + \beta_2 x), \quad \beta_2 \neq 0; \tag{6}$$

$$\eta_2(x, z; \beta) = \eta_2(\beta_0 + \beta_1 z), \quad \beta_1 \neq 0. \tag{7}$$

Theorem 3. For model (6), (β_0, β_2, ϕ) are identified, and for model (7), (β_0, β_1, ϕ) are identified.

5. ESTIMATION

Inverse probability weighted estimation (Horvitz & Thompson, 1952; Robins et al., 1994; Scharfstein et al., 1999) is one of the most influential methods for missing data analysis. The approach employs a propensity score model $\pi(x, y; \alpha) = \mathrm{pr}(r = 1 \mid x, y; \alpha)$, for example, a logistic model $\mathrm{logit}\,\{\pi(x, y; \alpha)\} = \alpha_0 + \alpha_1 x + \alpha_2 y$. If $\alpha_1 \neq 0$, the model accommodates a nonignorable missingness process. With fully observed data, α can be consistently estimated by standard maximum likelihood. Alternatively, one may solve estimating functions of the following form: $\widehat{E}[\{r/\pi(x, y; \widehat{\alpha}) - 1\}G(x, y)] = 0$, with $\widehat{E}$ denoting the empirical expectation, $G(x, y)$ a user-specified vector function of dimension equal to that of α, and $E[\partial\{r/\pi(x, y; \alpha)\}/\partial\alpha \times G(x, y)]$ nonsingular for all α. For instance, one may naturally choose $G(x, y) = (1, x, y)$ for the logistic propensity score model. But, when X has missing values, neither approach is feasible. Nevertheless, when a shadow variable Z is fully observed, one can solve the following modified estimating equation with $G(x, y)$ replaced by $G(z, y)$,

$$\widehat{E}\left[\left\{\frac{r}{\pi(x, y; \widehat{\alpha})} - 1\right\} G(z, y)\right] = 0. \tag{8}$$

Incorporating $\pi(x, y; \hat{\alpha})$ obtained from (8), one can solve

$$\widehat{E}\left\{\frac{r}{\pi(x, y; \widehat{\alpha})} S(x, y, z; \widehat{\beta}, \widehat{\phi})\right\} = 0, \tag{9}$$

for $(\hat{\beta}, \hat{\phi})$, with the score function $S(x, y, z; \beta, \phi) = \partial \log\{\text{pr}(y \mid x, z; \beta; \phi)\}/\partial(\beta, \phi)$. With a valid shadow variable Z, we show that replacing $G(x, y)$ with $G(z, y)$ does not compromise unbiasedness of the estimating equations (8)–(9).

Theorem 4. If the propensity score model $\pi(x, y; \alpha)$ is correctly specified, then (8) is an unbiased estimating equation for α, i.e., at the true value of α

$$E\left[\left\{\frac{r}{\pi(x, y; \alpha)} - 1\right\} G(z, y)\right] = 0.$$

If further the outcome model $\text{pr}(y \mid x, z; \beta, \phi)$ is correctly specified, then (9) is an unbiased estimating equation for (β, ϕ), i.e., at the true value of (α, β, ϕ)

$$E\left\{\frac{r}{\pi(x, y; \alpha)} S(x, y, z; \beta, \phi)\right\} = 0.$$

Provided unbiasedness of the estimating equations, the consistency and asymptotic normality of $(\hat{\alpha}, \hat{\beta}, \hat{\phi})$ can be obtained under standard regularity conditions given by Newey & McFadden (1994, Theorem 6.1), and the asymptotic variance and 95% confidence interval can be obtained based on asymptotic normality. Such asymptotic properties follow from the general theory of estimating equations, which has been very well established in the literature. We refer readers to Newey & McFadden (1994); Robins et al. (1994); Shao (2003); Tsiatis (2006) for the technical details. Specific choices of G can generally affect efficiency but not consistency of the estimators. In the Supplementary Material, we characterize the optimal choice of G within our class of estimating equations, which typically follows from the general framework by Newey & McFadden (1994, Theorem 5.3).

Inverse probability weighted (IPW) estimation as applied in this paper is not new except that we use the shadow variable to assist with identification and estimation of the propensity score model in (8). There exists a large literature on properties and extensions of IPW estimation, to name a few, Horvitz & Thompson (1952); Robins et al. (1994); Wang et al. (2014); Shao & Wang (2016). However, we note that IPW estimation enjoys several attractive properties. For instance, the identification strategy for nonparametric propensity score models developed under the missing outcome setting by Sun et al. (2016) can be extended to assess identification for the missing covariate problem; the semiparametric IPW estimation developed for the missing outcome problem by Shao & Wang (2016) can be extended to the missing covariate problem to relax stringent parametric model assumptions. Moreover, such extensions are often achieved by simply switching X and Y in the propensity score model. It is noteworthy that alternative fully likelihood based or Bayesian based approaches also exist for estimation in the present context, e.g., imputation methods (Rubin & Schenker, 1986). However, to account for missing covariate data, these methods require additionally specifying a model for $\text{pr}(x \mid y, z)$ or $E(x \mid y, z)$, and therefore are more sensitive to model misspecification and possible lack of coherence between models for $\text{pr}(x \mid y, z)$ and $\text{pr}(y \mid x, z)$.

6. NUMERICAL EXAMPLES

6.1 Simulation Studies

We study the finite sample performance of the proposed inverse probability weighted estimator via simulations. We generate the shadow variable Z from $N(0, 1)$, $X \sim N(0.5 + 0.5z, 1)$, and $Y \sim N(\beta_0 + \beta_1 z + \beta_2 x, 1)$ with $(\beta_0, \beta_1, \beta_2) = (0.5, 1.5, -0.5)$. We generate R from $\text{logit}\,\{\text{pr}(r = 1 \mid x, y)\} = \alpha_0 + \alpha_1 y + \alpha_2 x$ with $(\alpha_0, \alpha_1, \alpha_2) = (0.5, -1, 1)$, and treat the samples of X with $R = 0$ as missing values. Under such a setting, the missing data proportion is about 39%. We simulate 1000 independent data sets under sample sizes 500 and 1500. We apply three methods to analyze the data sets, including inverse probability weighting, complete-case analysis, and full data maximum likelihood estimation in which we pretend to have the missing values. Results are summarized in the boxplots of Figure 3. As expected, the full data maximum likelihood estimation always performs best, with smallest bias and variance. But the full data maximum likelihood estimation is not feasible in practice as missing data arise. However, with the shadow variable incorporated, the inverse probability weighted estimator performs reasonably well. Both bias and variance are relatively small under moderate sample sizes; as sample size increases, the bias and variance decrease and the coverage probability of the 95% confidence interval approximates the nominal level as shown in Table 1. But the estimator obtained from complete-case analysis has large bias, and it cannot be alleviated as sample size increases.

Table 1: Coverage probability of the 95% confidence interval.

	α_1	α_2	β_1	β_2
N=500	0.959	0.954	0.911	0.931
1500	0.943	0.939	0.929	0.945

Note: Confidence intervals are constructed based on asymptotic normality of the estimators.

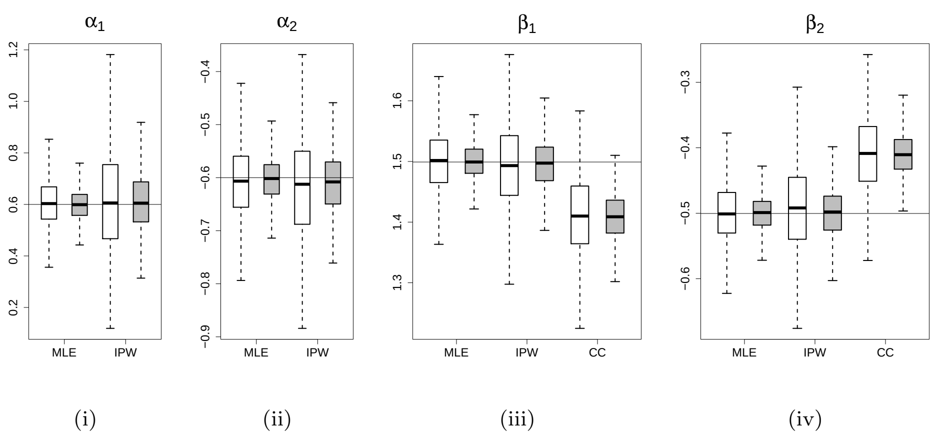


Figure 3: Boxplots for the estimators.

Note: Data are analyzed with inverse probability weighting (IPW), complete-case analysis (CC), and full data maximum likelihood estimation (MLE). In each boxplot, white boxes are for sample size 500 and gray ones for 1500. The horizontal line marks the true value of the parameter.

6.2 China Home Pricing example

For illustration, we apply the shadow variable approach to a data set extracted from China Family Panel Studies (CFPS). Details of the survey can be found at `http://www.isss.edu.cn/cfps/EN/`. The dataset we used consists of 5534 homeowners. We are interested in the effect of family income (`faminc`) on home price (`houspr`). Other covariates include gender, age, education status, and family size of the homeowners being investigated, and location, distance to the downtown (`dist`), year of construction, size, type, and tidiness of their homes. Family income has 1896 (34.2%) missing values, while the other variables are fully observed. Home price increases as the distance to the downtown decreases, and homeowners living closed to the downtown are more likely to be wealthy. Thus, `dist` is highly correlated both with home price and family income, but it is reasonable to assume that `dist` does not affect the response propensity of homeowners after conditioning on home price, family income, and other covariates. As such, we use `dist` as a shadow variable for family income. We analyze the dataset with a linear outcome model and a logistic propensity score model, and summarize the results in Table 2. The results for the propensity score model provides significant empirical evidence of nonignorable missingness of family income, with a coefficient of -0.440 and 95% confidence interval of $(-0.773, -0.107)$, which indicates that homeowners with high family income tend not to respond to the survey. In the outcome model, the coefficient of `faminc` is 0.148 with confidence interval $(0.104, 0.191)$, which shows a significant positive effect of family income on home price. The results also confirm that home price increases as the distance to downtown decreases and home size increases, and that a newer home has a higher price.

7. DISCUSSION

In this paper, we have given careful and formal consideration to the challenging problem of identification in regression analysis when covariate data are missing not at random. Through several examples, we have demonstrated that identification is often not possible in such settings even under fairly stringent parametric conditions. We have also discussed identification in settings where a shadow variable might be available. Apart from our framework, one may apply the identification and estimation results established for the missing outcome problem by simply switching X and Y. For instance, applying the identification results established for the missing outcome problem given by Wang et al. (2014), Miao et al. (2015) and D'Haultfoeuille (2010), one can check if $\text{pr}(x \mid y, z)$ follows a location-scale model to assess identification for the missing covariate problem. However, we do not recommend this approach because it is often infeasible in practice. First, in the

Table 2: Results for the China home pricing example.

Outcome model			Propensity score model		
gender	-0.242	(0.182, 0.302)	gender	0.358	(0.234, 0.483)
age	0.018	(0.015, 0.021)	age	0.015	(0.008, 0.021)
educ	-0.019	(-0.030, -0.009)	educ	0.081	(0.049, 0.113)
urban	0.497	(0.390, 0.604)	urban	0.737	(0.599, 0.875)
year	-0.021	(-0.025, -0.018)	famsz	0.059	(0.008, 0.110)
size	0.039	(0.035, 0.043)	faminc	-0.440	(-0.773, -0.107)
type	0.746	(0.656, 0.836)	houspr	0.013	(-0.072, 0.098)
famsz	-0.017	(-0.036, 0.001)			
tidy	0.088	(0.064, 0.111)			
faminc	0.148	(0.104, 0.191)			
dist	-1.972	(-2.107, -1.838)			

Note: Point estimates and confidence intervals (in brackets) for the models, with 7 variables (after stepwise selection) included in the propensity score model and 11 in the outcome model.

missing covariate problem $\mathrm{pr}(x \mid y, z)$ is treated as a nuisance parameter, which however may not be informative to identify $\mathrm{pr}(y \mid x, z)$, the parameter of primary interest; taking the missingness mechanism $\{R \perp\!\!\!\perp Y \mid (X, Z),\ R \not\perp\!\!\!\perp X \mid Z\}$ as an example, $\mathrm{pr}(x \mid y, z)$ may not be identified and thus cannot be used to identify $\mathrm{pr}(y \mid x, z)$; nevertheless, in Example 3 we have verified that $\mathrm{pr}(y \mid x, z)$ is in fact identified! Second, in order to apply the missing outcome framework, one has to work out $\mathrm{pr}(x \mid y, z)$ based on the model for $\mathrm{pr}(y \mid x, z)$, but the result can be very complicated so that one cannot easily verify whether $\mathrm{pr}(x \mid y, z)$ satisfies the identification conditions; for instance, under a logistic model for $\mathrm{pr}(y \mid x, z)$ and a normal model for $\mathrm{pr}(x \mid z)$, the model for $\mathrm{pr}(x \mid y, z)$ has no closed form and one cannot apply previous identification results. Last, although sometimes one may parametrize both $\mathrm{pr}(y \mid x, z)$ and $\mathrm{pr}(x \mid y, z)$, but one is more likely to encounter model misspecification and lack of coherence between models for $\mathrm{pr}(x \mid y, z)$

and $\text{pr}(y \mid x, z)$, which may lead to even larger bias; for instance, model incoherence and misspecification occurs if linear regression models are specified both for $E(y \mid x, z)$ and $E(x \mid y, z)$ and when the error terms have complicated distributions. Therefore, it is favorable to directly assess identification for the missing covariate problem, except for certain situations where model incoherence and misspecification is not a serious issue.

The shadow variable plays a central role in identification of the semiparametric models where the propensity score $\text{pr}(r = 1 \mid x, y)$ is left unspecified. The definition of shadow variable in this paper is closed to the "instrumental variable" described by D'Haultfoeuille (2010), Wang et al. (2014) and Shao & Wang (2016), but is different from the conventional instrumental variable in the econometrics literature, where an instrumental variable is independent of the potentially missing variable but associated with its missingness. In econometrics, the instrumental variable approach has a longstanding tradition initiated by Wright (1928) and Goldberger (1972) and further developed by Imbens & Angrist (1994), Angrist et al. (1996) and Heckman (1997). Recently, Sun et al. (2016) and Tchetgen Tchetgen & Wirth (2017) implemented such an instrumental variable to establish identification conditions for nonignorable missing data, which only involve the propensity score $\text{pr}(r = 1 \mid x, y)$. Their work can be generalized to the missing covariate problem when an instrumental variable for X is available and the propensity score model is specified, and thus is a useful complement to this paper.

ACKNOWLEDGMENT

We thank the editor, associate editor, and two referees for their valuable comments. The work is partially supported by the China Scholarship Council and the National Institute of Health.

SUPPLEMENTARY MATERIAL

The online supplementary material includes identification results for the pattern-mixture parametrization, efficiency issue for (8), useful lemmas, and proofs of the theorems.

APPENDIX

This appendix includes additional details for Examples 1, 2, 3, 6, and 7.

Details for Example 1

Note that $\text{pr}(y, r = 0) = \text{pr}(y) - \text{pr}(y, r = 1)$ and $\text{pr}(y, r = 1) = \int_x \text{pr}(x, r = 1 \mid y)\text{pr}(y)\text{dx}$, we only need to show that these two different settings lead to the identical distributions of $\text{pr}(y)$ and $\text{pr}(x, r = 1 \mid y)$. One can verify that $\text{pr}(y) = N(0, 1)$ and

$$\text{pr}(x, r = 1 \mid y) = (2\pi)^{-1/2} \exp\left\{-\frac{(y - 2x)^2}{8}\right\} \frac{\exp(2 - 2x + y)}{1 + \exp(2 - 2x + y)}.$$

Details for Example 2

Suppose $\text{pr}(x \mid y; \theta_1)/\text{pr}(x \mid y; \theta_2) = h(y)$ for some function $h(y)$, then for all y we have

$$\int_x \text{pr}(x \mid y; \theta_1)dx = \int_x \text{pr}(x \mid y; \theta_2)h(y)dx = h(y) = 1,$$

which contradicts $\text{pr}(x \mid y; \theta_1) \neq \text{pr}(x \mid y; \theta_2)$. Therefore, $\text{pr}(x \mid y; \theta_1)/\text{pr}(x \mid y; \theta_2)$ must vary with x.

Details for Example 3

We only need to prove that $\text{pr}(y \mid x; \theta_1)/\text{pr}(y \mid x; \theta_2)$ varies with y; otherwise, suppose $\text{pr}(y \mid x; \theta_1)/\text{pr}(y \mid x; \theta_2) = h(x)$ for some function $h(x)$, then for all x we have

$$\int_y \text{pr}(y \mid x; \theta_1)dy = \int_y \text{pr}(y \mid x; \theta_2)h(x)dy = h(x) = 1,$$

which contradicts $\text{pr}(y \mid x; \theta_1) \neq \text{pr}(y \mid x; \theta_2)$. Therefore, $\text{pr}(y \mid x; \theta_1)/\text{pr}(y \mid x; \theta_2)$ and thus $\{\text{pr}(y \mid x; \theta_1)\text{pr}_1(x)\}/\{\text{pr}(y \mid x; \theta_2)\text{pr}_2(x)\}$ must vary with y.

Details for Example 6

We use a proof by contradiction to show identification of the parameters. Suppose that there were two sets of parameters resulting in the identical distribution $\text{pr}(x, y, r = 1)$:

$$\begin{aligned} &\exp(\beta_0 + \beta_1 x) \exp\{-y \exp(\beta_0 + \beta_1 x)\} \tfrac{1}{\sigma_1} \Phi(\tfrac{x-\mu}{\sigma_1}) \tfrac{\exp(\alpha_0+\alpha_1 x+\alpha_2 y)}{1+\exp(\alpha_0+\alpha_1 x+\alpha_2 y)} \\ &= \exp(\beta_0' + \beta_1' x) \exp\{-y \exp(\beta_0' + \beta_1' x)\} \tfrac{1}{\sigma_1'} \Phi(\tfrac{x-\mu'}{\sigma_1'}) \tfrac{\exp(\alpha_0'+\alpha_1' x+\alpha_2' y)}{1+\exp(\alpha_0'+\alpha_1' x+\alpha_2' y)}, \end{aligned} \tag{10}$$

with Φ the probability density function of $N(0, 1)$. Taking logarithm on both sides and rearranging the terms, we have

$$\begin{aligned} &c - \{\tfrac{(x-\mu)^2}{2\sigma_1^2} - \tfrac{(x-\mu')^2}{2\sigma_1'^2}\} + (\beta_1 - \beta_1' + \alpha_1 - \alpha_1')x + (\alpha_2 - \alpha_2')y \\ &= y\{\exp(\beta_0 + \beta_1 x) - \exp(\beta_0' + \beta_1' x)\} + \log \tfrac{1+\exp(\alpha_0+\alpha_1 x+\alpha_2 y)}{1+\exp(\alpha_0'+\alpha_1' x+\alpha_2' y)}, \end{aligned} \tag{11}$$

with $c = \{\beta_0 - \beta_0' + \alpha_0 - \alpha_0' - \log(\sigma_1) + \log(\sigma_1')\}$. For arbitrary y, the left hand side of (11) is a linear combination of x and x^2. But for $\beta_0 \neq \beta_0'$ or $\beta_1 \neq \beta_1'$, note that

$\beta_1, \beta_1' \neq 0$, the right hand side of (11) must include an exponential term of x; and it cannot equal the left hand side of (11). Thus, we must have $\beta_0 = \beta_0'$ and $\beta_1 = \beta_1'$, and (10) reduces to

$$\frac{1}{\sigma_1}\Phi\left(\frac{x-\mu}{\sigma_1}\right)\frac{\exp(\alpha_0+\alpha_1 x+\alpha_2 y)}{1+\exp(\alpha_0+\alpha_1 x+\alpha_2 y)} = \frac{1}{\sigma_1'}\Phi\left(\frac{x-\mu'}{\sigma_1'}\right)\frac{\exp(\alpha_0'+\alpha_1' x+\alpha_2' y)}{1+\exp(\alpha_0'+\alpha_1' x+\alpha_2' y)}.$$

By the same argument of Miao et al. (2017) for identification of normal densities, the identity holds only for $\mu = \mu'$, $(\alpha_0, \alpha_1, \alpha_2) = (\alpha_0', \alpha_1', \alpha_2')$ and $\sigma_1 = \sigma_1'$. Therefore, all parameters are identified.

The Weibull regression $Y \sim \sigma_2 y^{\sigma_2 - 1}\exp\{-y^{\sigma_2}\eta(x) + \log(\eta(x))\}$ is a generalization of the exponential regression model. We first prove identification of σ_2, and then identification of other parameters follows from identification of the exponential regression model. For the Weibull regression, we follow the proof for the exponential regression and then obtain a parallel version of (11):

$$\begin{aligned} &c - \{\tfrac{(x-\mu)^2}{2\sigma_1^2} - \tfrac{(x-\mu')^2}{2\sigma_1'^2}\} + (\beta_1 - \beta_1' + \alpha_1 - \alpha_1')x + (\alpha_2 - \alpha_2')y + (\sigma_2 - \sigma_2')\log(y) \\ &\quad = \{y^{\sigma_2}\exp(\beta_0 + \beta_1 x) - y^{\sigma_2'}\exp(\beta_0' + \beta_1' x)\} + \log\tfrac{1+\exp(\alpha_0+\alpha_1 x+\alpha_2 y)}{1+\exp(\alpha_0'+\alpha_1' x+\alpha_2' y)}. \end{aligned} \tag{12}$$

For arbitrary x, the left hand side of (12) is a linear combination of y and $\log(y)$. But for $\sigma_2 \neq \sigma_2'$, the right hand side of (11) must include a power term of y, and it is not equal to the left hand side of (11). Thus, we must have $\sigma_2 = \sigma_2'$. Letting $\widetilde{Y} = Y^{\sigma_2}$, then $\widetilde{Y} \sim \exp\{-\widetilde{y}\eta(x) + \log(\eta(x))\}$, which is an exponential regression model. Applying the identification result of the exponential regression model, we obtain identification of the remaining parameters.

Details for Example 7

When X and Z are binary, for arbitrary y, we solve the equation $\text{pr}(z = 1 \mid y) = \sum_{x=0,1}\text{pr}(z = 1 \mid x, y, r = 1)\text{pr}(x \mid y)$ for $\text{pr}(x = 1 \mid y)$. Note that $\text{pr}(x = 1 \mid y) + \text{pr}(x = 0 \mid y) = 1$, we have

$$\text{pr}(x = 1 \mid y) = \frac{\text{pr}(z = 1 \mid y) - \text{pr}(z = 1 \mid x = 0, y, r = 1)}{\text{pr}(z = 1 \mid x = 1, y, r = 1) - \text{pr}(z = 1 \mid x = 0, y, r = 1)}.$$

Under the assumption $Z \not\perp\!\!\!\perp X \mid Y = y$ for any y, $\text{pr}(z = 1 \mid x = 1, y) \neq \text{pr}(z = 1 \mid x = 0, y)$, thus, $\text{pr}(z = 1 \mid x = 1, y, r = 1) \neq \text{pr}(z = 1 \mid x = 0, y, r = 1)$ by the shadow variable assumption $Z \perp\!\!\!\perp R \mid (X, Y)$. Therefore, the solution for $\text{pr}(x = 1 \mid y)$ is unique.

REFERENCES

AFIFI, A. & ELASHOFF, R. (1966). Missing observations in multivariate statistics I. Review of the literature. *Journal of the American Statistical Association* **61**, 595–604.

ANDERSON, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association* **52**, 200–203.

ANGRIST, J., IMBENS, G. & RUBIN, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**, 444–455.

BAKER, S. G. & LAIRD, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association* **83**, 62–69.

BARTLETT, J. W., CARPENTER, J. R., TILLING, K. & VANSTEELANDT, S. (2014). Improving upon the efficiency of complete case analysis when covariates are MNAR. *Biostatistics* **15**, 719–730.

BUCK, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society. Series B (Methodological)* **22**, 302–306.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1–38.

D'HAULTFOEUILLE, X. (2010). A new instrumental method for dealing with endogenous selection. *Journal of Econometrics* **154**, 1–15.

EDGETT, G. L. (1956). Multiple regression with missing observations among the independent variables. *Journal of the American Statistical Association* **51**, 122–131.

FAY, R. E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association* **81**, 354–365.

Glasser, M. (1964). Linear regression analysis with missing observations among the independent variables. *Journal of the American Statistical Association* **59**, 834–844.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica* **40**, 979–1001.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society. Series B (Methodological)* **30**, 67–82.

Heckman, J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *The Journal of Human Resources* **32**, 441–462.

Horton, N. J. & Laird, N. M. (2001). Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. *Biometrics* **57**, 34–42.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association* **85**, 765–769.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R. & Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association* **100**, 332–346.

Ibrahim, J. G., Lipsitz, S. R. & Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Methodological)* **61**, 173–190.

Imbens, G. & Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–475.

Kott, P. S. (2014). Calibration weighting when model and calibration variables can differ. In *Contributions to Sampling Statistics*, F. Mecatti, L. P. Conti & G. M. Ranalli, eds. Cham: Springer, pp. 1–18.

Lipsitz, S. R., Ibrahim, J. G., Chen, M.-H. & Peterson, H. (1999). Non-ignorable missing covariates in generalized linear models. *Statistics in Medicine* **18**, 2435–2448.

Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134.

Little, R. J. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* Wiley: New York.

Little, R. J. & Zhang, N. (2011). Subsample ignorable likelihood for regression analysis with missing data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **60**, 591–605.

Ma, W. Q., Geng, Z. & Hu, Y. H. (2003). Identification of graphical models for nonignorable nonresponse of binary outcomes in longitudinal studies. *Journal of Multivariate Analysis* **87**, 24–45.

Miao, W., Ding, P. & Geng, Z. (2017). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association* **111**, 1673–1683.

Miao, W., Tchetgen Tchetgen, E. & Geng, Z. (2015). Identification and doubly robust estimation of data missing not at random with a shadow variable. Technical report.

Newey, W. K. & McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, R. F. Engle & D. L. McFadden, eds., vol. 4. Amsterdam: Elsevier, pp. 2111–2245.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika* **63**, 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Rubin, D. B. & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* **81**, 366–374.

Scharfstein, D. O., Rotnitzky, A. & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94**, 1096–1120.

Shao, J. (2003). *Mathematical Statistics.* New York: Springer, 2nd ed.

Shao, J. & Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika* **103**, 175–187.

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J. & Tchetgen, E. T. (2016). Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica,* in press.

Tang, G., Little, R. J. & Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* **90**, 747–764.

Tchetgen Tchetgen, E. J. & Wirth, K. E. (2017). A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics,* in press.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data.* New York: Springer.

Vach, W. & Schumacher, M. (1993). Logistic regression with incompletely observed categorical covariates: a comparison of three approaches. *Biometrika* **80**, 353–362.

Wang, S., Shao, J. & Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica* **24**, 1097–1116.

Wright, P. G. (1928). *Tariff on Animal and Vegetable Oils.* New York: Macmillan.

Zahner, G. E., Pawelkiewicz, W., DeFrancesco, J. J. & Adnopoz, J. (1992). Children's mental health service needs and utilization patterns in an urban community: an epidemiological assessment. *Journal of the American Academy of Child & Adolescent Psychiatry* **31**, 951–960.

Zhao, J. & Shao, J. (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association* **110**, 1577–1590.

Zhao, L. P., Lipsitz, S. & Lew, D. (1996). Regression analysis with missing covariate data using estimating equations. *Biometrics* **52**, 1165–1182.

Wang Miao
Guanghua School of Management
Peking University
Haidian District, Beijing 100871
E-mail: mwfy@pku.edu.cn

Eric Tchetgen Tchetgen
Department of Biostatistics
Harvard University
Boston, Massachusetts 02115
E-mail: etchetge@hsph.harvard.edu