View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Institutional Knowledge at Singapore Management University

Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Economics

School of Economics

1-2021

Nonignorable missing data, single index propensity score and profile synthetic distribution function

Xuerong CHEN

Denis H. Y. LEUNG

Jing QIN

Follow this and additional works at: https://ink.library.smu.edu.sg/soe_research

 Part of the Econometrics Commons

This Journal Article is brought to you for free and open access by the School of Economics at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Economics by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Nonignorable Missing Data, Single Index Propensity Score and Profile Synthetic Distribution Function

Xuerong Chen[a], Denis Heng-Yan Leung[b], and Jing Qin[c]

[a] Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu, China; [b] School of Economics, Singapore Management University, Singapore, Singapore; [c] Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, MD. Corresponding author: denisleung@smu.edu.sg

Published in Journal of Business & Economic Statistics, 2021 Jan, Advance online
DOI: 10.1080/07350015.2020.1860065

Abstract: In missing data problems, missing not at random is difficult to handle since the response probability or propensity score is confounded with the outcome data model in the likelihood. Existing works often assume the propensity score is known up to a finite dimensional parameter. We relax this assumption and consider an unspecified single index model for the propensity score. A pseudo-likelihood based on the complete data is constructed by profiling out a synthetic distribution function that involves the unknown propensity score. The pseudo-likelihood gives asymptotically normal estimates. Simulations show the method compares favorably with existing methods.

Keywords: Missing not at random, Nonignorable missing, Pseudo-conditional likelihood, Single index model, Synthetic distribution

1 Introduction

The problem of missing data has always been a concern for those who work with empirical research. One reason why missing data has captured so much attention from researchers is because the nature of the problem is complex. It is widely acknowledged that missing data does not arise from a unitary source under a well-defined situation. Rather, the causes and processes leading to missing data are varied and often a function of multiple factors, such as the population under study, the nature of the outcome, or the way the study is designed and conducted. A most challenging issue is that information about the missing data is usually scant, making it very difficult for researchers to determine the nature of missingness. In carrying out research with missing data, a common approach is to discard observations with missing data and to carry out a "complete case" analysis. Such an approach is rarely appropriate unless the missingness probability, or propensity score, does not vary with any observable or unobservable data. More often, the complete cases form a subsample with characteristics different from a random sample from the target population, so analysis based on this subsample will result in misleading inference.

In economics, missing data or nonresponse is endemic in many important surveys that are used routinely to investigate economic and social issues. The seriousness of the missing data issues in these dataset was brought to the fore by Lillard, Smith, and Welch (1986). The article reported that, for the Current Population Survey (CPS), the nonresponse rates among persons 14 years and older experienced a 10-fold increase between 1940 and 1982. This trend has continued in the years following the publication of the article, so that the nonresponse rate in parts of the CPS currently stands at 35% or higher (Kline and Santos 2013; Bollinger et al. 2019). It has been found that missing data and nonresponse in the CPS affects measures of important economic metrics, such as earning gaps between gender, race, and education groups (e.g., Lemieux 2006; Mulligan and Rubinstein 2008; Bollinger and Hirsch 2013; Maasoumi and Wang 2017, 2019; Bollinger et al. 2019). Unfortunately, the CPS is not alone in its nonresponse issues. Similar problems have been cited from the National Longitudinal Survey of Youths (NLSY), the Panel Study of Income Dynamics (PSID), and many other major surveys (e.g., Davey, Shanahan, and Schafer 2001; Schräpler 2004; Arpino, De Cao, and Peracchi 2014; Breunig 2017; Golsteyn and Hirsch 2019; Qin et al. 2019).

There is a rich body of statistical works for handling missing or nonresponse data. Most of these works are based on the frameworks defined in Rubin (1976, 1987). A common approach to analysis with missing data or nonresponse is to assume that data is either "missing completely at random" (MCAR) or "missing at random" (MAR), in the sense that missingness is either constant across all units or constant across all units when conditioned on observed values. These conditions allow valid inference to be drawn from observations with complete data, or "complete cases" but they are made based on convenience than plausibility. In practice, the probability or "propensity" (Rosenbaum and Rubin 1983) for missingness and nonresponse, more likely varies across the unobservable data, so that the data is "missing not at random" (MNAR) or "nonignorable." Previous works on nonignorable missingness

include full maximum likelihood models (Greenlees, Reece, and Zieschang 1982; Stasny 1985) and/or exclusion restrictions to obtain point identification (Lillard, Smith, and Welch 1986; Chen, Geng, and Zhou 2009; Kott and Chang 2010; Yang, Lorch, and Small 2014), or the "worst case" bounds on population moments that result when no assumptions regarding the missingness process are used (Manski 1989, 1990; Manski and Pepper 2000). Fully parametric approaches are known to be very sensitive to model misspecification. For most applied problems, the worst case bounds are overly conservative in the sense that they consider missingness processes unlikely to be found in practice.

Robins and Ritov (1997) showed that a fully nonparametric approach is not possible for point identification. Hence, to balance the extremes of a fully parametric model and a worst case bound approach, a semiparametric approach is taken in this article. Specifically, we develop a pseudo-likelihood (Besag 1975) approach. We assume a set of data with outcome Y and covariates $\mathbf{X}$ such that the covariates are always observed but the outcome is observed when $R = 1$ and is missing when $R = 0$ and missingness may be nonignorable. We use the notations $[\cdot]$, $[\cdot,\cdot]$, and $[\cdot|\cdot]$ to denote marginal, joint, and conditional distributions, respectively. Previous works on pseudo-likelihood analysis of MNAR data include Liang and Qin (2000), Tang, Little, and Raghunathan (2003), and Zhao and Shao (2015). When missingness is nonignorable, the $(Y, \mathbf{X}, R)$ is a nonrandom sample of the target population [Y,**X**] because $[Y, \mathbf{X}|R] \neq [Y, \mathbf{X}]$. The basic idea of a pseudo-likelihood approach is that when conditioned on the outcome and R, $(Y, \mathbf{X}, R = 1)$ is a random samples of $[\mathbf{X}|Y, R = 1]$. Liang and Qin (2000) assumed missingness can only depend on the outcome but not covariates (Liang and Qin (2000) also allowed covariates to be missing but the missingness probability for covariates is a function of solely the covariates) and formed a pseudo-likelihood using pairs of observations. Using the assumption that the nonresponse mechanism only depends on the outcome, Tang, Little, and Raghunathan (2003) factorized the joint distribution $[Y, \mathbf{X}, R = 1]$ as $[R = 1|Y][Y|\mathbf{X}][\mathbf{X}]$. They then used a parametric model for $[Y|\mathbf{X}]$ and a nonparametric model for $[\mathbf{X}]$. The key to the pseudo-likelihoods of Liang and Qin (2000) and Tang, Little, and Raghunathan (2003) is the missingness mechanism is only a function of the outcome. This condition allows the unknown missingness function to be factored out of the pseudo-likelihood. This condition is valid in certain situations. For example, in a survey of willingness to receive HIV testing, individuals might refuse to participate due to social stigma (Rueda et al. 2016), in which case, it can be argued that missingness is only dependent on a person latent HIV status, but independent of demographics or personal characteristics. The condition may also be plausible in threshold crossing models (Matzkin 1992) in which case a person's willingness to participate depends only on the latent outcome. This condition requires the missing mechanism to be identical across all covariates. Therefore, if the number of covariates is large or some of them are continuous, the condition is a severe restriction. This point was raised by Zhao and Shao (2015). In Zhao and Shao (2015), it is assumed the covariates $\mathbf{X}$ can be written as $(\mathbf{U}, \mathbf{Z})$ such that $\mathbf{Z}$ is not of interest. Furthermore, $\mathbf{Z}$ can be used as an instrument for missingness. They factorized the joint distribution $[Y, \mathbf{X}, R = 1]$ as $[R = 1|Y, \mathbf{U}][Y|\mathbf{X}][\mathbf{U}|\mathbf{Z}][\mathbf{Z}]$ and developed estimation method based on conditional distribution $[\mathbf{Z}|Y, \mathbf{U}, R = 1]$. They used a parametric model for $[Y|\mathbf{X}]$ and proposed nonparametric estimates for $[\mathbf{U}|\mathbf{Z}]$ and $[\mathbf{Z}]$, such that the propensity score can be factored out. In this article, we also allow the missingness process to depend on the outcome and possibly some covariates. In contrast to Zhao and Shao (2015), who used instrumental variables for identifiability, we use a semiparametric single index propensity score model based on some covariates that are independent of the missing indicator given the outcome and other covariates. The only restriction is there exists at least one continuous component covariate in the single index model. Unlike Zhao and Shao (2015), our method is founded on $[Y, \mathbf{X}|R = 1]$, which extracts more information in $[Y, \mathbf{U}|R = 1]$ than Zhao and Shao (2015). Moreover, we do not require nonparametric estimate of $[\mathbf{U}|\mathbf{Z}][\mathbf{Z}]$, which can be problematic if $[\mathbf{U}, \mathbf{Z}]$ is high dimensional. We show the likelihood using the complete data is a biased sampling version of the target likelihood function. We construct a synthetic distribution using a normalized version of the missingness function. We then profile this pseudo-likelihood to find estimates of the synthetic distribution function and related parameters. We note in passing that the pseudo-likelihoods mentioned here are associated with works in the areas of choice-based or response-based sampling (see, e.g., Cosslett 1981; Chen 2001; Ramalho and Smith 2013). Indeed, Ramalho and Smith (2013) showed that a missing data problem can be recast as a choice-based sampling problem. However, in choice-based or response-based sampling, the outcome Y (choice) is normally assumed to be discrete. We consider the more general case here where Y can be discrete or continuous. When Y is continuous, the distribution $[Y, \mathbf{X}]$ must be parametrized or estimated using nonparametric smoothing.

There are also other semiparametric works on MNAR problems that appeared in the literature. Kott and Chang (2010) assumed a parametric model for missingness and nonparametric model for $[Y, \mathbf{X}]$ and used calibration weights. Wang, Shao, and Kim (2014) also allowed a nonparametric model for $[Y, \mathbf{X}]$ but used nonresponse instruments as in Zhao and Shao (2015) and then used GMM for estimation. Kim and Yu (2011) assumed the missingness mechanism is known or can be estimated using external data and they modeled the missing data distribution as an exponential tilt of the data distribution of the observed. Also using the exponential tilt framework, Zhao, Zhao, and Tang (2013) used empirical likelihood estimators for drawing inferences and Shao and Wang (2016) but an instrumental variable. MNAR problems have also been studied in panel data situations. For longitudinal data with only two patterns of monotone missingness that depends only on the outcome, Little (1993, 1994) and Little and Wang (1996) showed that it is possible to identify the parameters. However, their methods do not apply to more general missing patterns.

The rest of this article is organized as follows. In Section 2, we describe the basic set-up of the nonignorable missing data problem and we introduce the proposed method. Large sample properties of the proposed method are examined in Section 3. Section 4 reports the results of a modest simulation study. In

Section 5, empirical application of the method is illustrated using a set of data. Concluding remarks are given in Section 6. Proofs are relegated to the online supplementary materials.

2. Pseudo-Conditional Likelihood Estimation

We begin with the parametric likelihood approach discussed in Greenlees, Reece, and Zieschang (1982). Assume an outcome Y and a p_1-dimensional covariate vector $\mathbf{X}$ be related by

$$Y = \mathbf{X}^T\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\boldsymbol{\beta}$ is a p_1-dimensional parameter, ϵ is a random error with density $f(\epsilon, \boldsymbol{\gamma})$, and $\boldsymbol{\gamma}$ is a p_2-dimensional parameter. Hence, the conditional distribution of the outcome given the covariates is given by $f(Y - \mathbf{X}^T\boldsymbol{\beta}, \boldsymbol{\gamma}) \equiv f(Y|\mathbf{X}, \boldsymbol{\Theta})$, where $\boldsymbol{\Theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$. We denote the marginal density of $\mathbf{X}$ by $g(\mathbf{X})$. Greenlees, Reece, and Zieschang (1982) assumed the outcome may be missing in some observations and the probability of observing a response, or propensity score, can be written as a logistic model linear in Y and $\mathbf{U}$, where $\mathbf{U}$ is a q-dimensional subvector of $\mathbf{X}$ or covariates distinct from $\mathbf{X}$. Under the assumption that ϵ is normally distributed, Greenlees, Reece, and Zieschang (1982) developed maximum likelihood estimators of the underlying parameters.

As noted by Little (1985), a parametric propensity function is at risk of misspecification. An alternative is a nonparametric representation for the propensity score. However, it suffers from the well-known curse of dimensionality problem. A natural balance between these extremes is a single index model (Powell, Stock, and Stoker 1989), which we adopt here. Let R be an indicator for missingness such that $R = 1$ if Y is observed and $R = 0$ otherwise. Then, we model the propensity score as follows

$$P(R = 1|Y, \mathbf{X}) = P(R = 1|Y, \mathbf{U}) = \pi(Y + \mathbf{U}^T\boldsymbol{\eta}), \tag{2}$$

where $\pi(\cdot)$ is an unknown function and $\boldsymbol{\eta}$ is an unknown parameter vector. It is easy to see that (2) includes the probit and logit models as special cases. Propensity score model (2) under the special case when $\boldsymbol{\eta}$ is known was considered in Little (1994) and Tang, Little, and Raghunathan (2003), even though neither carried out analysis of its properties. Here, we assume the propensity score depends on covariates $\mathbf{X}$ only through its subset $\mathbf{U}$ for the reason of identifiability (Wang, Shao, and Kim 2014). The specification of a unit coefficient for Y is not critical by noting that, for $a \neq 0$,

$$\begin{aligned} P(R = 1|Y, X) &= \pi(aY + \mathbf{U}^T\boldsymbol{\eta}) \\ &= \pi\{a(Y + \mathbf{U}^T\boldsymbol{\eta}/a)\} \\ &= \tilde{\pi}(Y + \mathbf{U}^T\tilde{\boldsymbol{\eta}}), \end{aligned}$$

where $\tilde{\pi}(t) = \pi(at)$ and $\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta}/a$. Since the form of π is not specified, we may replace $\pi(\cdot)$ by $\tilde{\pi}(\cdot)$ and $\boldsymbol{\eta}$ by $\tilde{\boldsymbol{\eta}}$ in (2).

Before giving details of the proposed estimators, we first consider identifiability issues arising from a semiparametric representation for $\pi(\cdot)$. From MNAR data, the conditional distribution of (Y, R) given $\mathbf{X}$ is $P(R = 1|Y, \mathbf{X})f(Y|\mathbf{X}, \boldsymbol{\Theta})$, which is unidentifiable if both $P(R = 1|Y, \mathbf{X})$ and $f(Y|\mathbf{X}, \boldsymbol{\Theta})$ are left unspecified (Robins and Ritov 1997). In our set-up, we parametrize $f(Y|\mathbf{X}, \boldsymbol{\Theta})$. The following lemma gives identifiability conditions under the single index model (2) for $P(R = 1|Y, \mathbf{X}) = \pi(Y + \mathbf{U}^T\boldsymbol{\eta})$.

Lemma 1. Without loss of generality, let the first component of $\mathbf{Z}$ be 1, and $\mathbf{Z}_{-1}$ be the remaining components of $\mathbf{Z}$. Write $f(Y|\mathbf{X}, \boldsymbol{\Theta}) = f(Y - (1, \mathbf{Z}_{-1}, \mathbf{U})^T\boldsymbol{\beta}, \boldsymbol{\gamma})$, where $\boldsymbol{\beta}^T = (\beta_0, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$. Assume $0 < P(R = 1) < 1$ and $\pi(\cdot) > 0$. Furthermore, suppose the following identifiability conditions hold:

- (S_1) There exist constants c, a_1, a_2 such that
$$f(s, \boldsymbol{\gamma}) = cf(a_1 + a_2 s, \tilde{\boldsymbol{\gamma}})$$
is true for all s, then $\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}, a_1 = 0, a_2 = 1$, and $c = 1$
- (S_2) At least one of the components of $\mathbf{Z}_{-1}$ is continuous with infinite support and its coefficient is not 0
- (S_3) The range of the support of at least one of the continuous components for $\mathbf{U}$ is infinite
- (S_4) The function $f(\cdot, \cdot)$ is continuous and differentiable with respect to $\boldsymbol{\Theta}$

If (S_1)–(S_4) hold, then $\pi(\cdot)$,$\boldsymbol{\eta}$ and $\boldsymbol{\Theta}$ are identifiable.

Clearly, the identifiability condition (S_1) is satisfied for commonly used distributions such as the normal, exponential, and those in the generalized exponential family. Conditions (S_2)–(S_4) similar to identify conditions in Ichimura (1988); see also Ichimura (1993) when the single index function is nonlinear.

For the special case that the propensity score is independent of $\mathbf{U}$, that is, $\pi(Y + \mathbf{U}^T\boldsymbol{\eta}) = \pi(Y)$, Tang, Little, and Raghunathan (2003) proposed a pseudo-likelihood based on the conditional distribution

$$\begin{aligned} [\mathbf{X}|Y, R = 1] &\sim \frac{\pi(Y)f(Y|\mathbf{X}, \boldsymbol{\Theta})g(\mathbf{X})}{\pi(Y)\int f(Y|\mathbf{X}, \boldsymbol{\Theta})g(\mathbf{X})d\mathbf{X}} \\ &= \frac{f(Y|\mathbf{X}, \boldsymbol{\Theta})g(\mathbf{X})}{\int f(Y|\mathbf{X}, \boldsymbol{\Theta})g(\mathbf{X})d\mathbf{X}}. \end{aligned} \tag{3}$$

Notice that the unspecified propensity score $\pi(Y)$ has been factored out in (3). Replacing $g(\mathbf{X})d\mathbf{X}$ by $dG_n(\mathbf{X})$ from the empirical marginal distribution of $\mathbf{X}$, under suitable conditions, the parameters in the model $f(Y|\mathbf{X}, \boldsymbol{\Theta})$ can be identified if it is normal density or more generally it is from generalized exponential family.

In practice, the propensity score may depend on the outcome as well as some covariates. This situation motivates a propensity score of the form $\pi(Y + \mathbf{U}^T\boldsymbol{\eta})$. The conditional distribution considered in Tang, Little, and Raghunathan (2003) becomes

$$[\mathbf{X}|Y, R = 1] \sim \frac{\pi(Y + \mathbf{U}^T\boldsymbol{\eta})f(Y|\mathbf{X}, \boldsymbol{\Theta})g(\mathbf{X})}{\int \pi(Y + \mathbf{U}^T\boldsymbol{\eta})f(Y|\mathbf{X}, \boldsymbol{\Theta})g(\mathbf{X})d\mathbf{X}}. \tag{4}$$

If $\boldsymbol{\eta}$ is known, then the problem can be reparametrized in terms of $Y^* = Y + \mathbf{U}^T\boldsymbol{\eta}$ and the method of Tang, Little, and Raghunathan (2003) can still be used. However, in the more general case that $\boldsymbol{\eta}$ is unknown, $\pi(Y + \mathbf{U}^T\boldsymbol{\eta})$ cannot be eliminated from (4). Therefore, Tang, Little, and Raghunathan's (2003) method cannot be used in general.

Write $\mathbf{X} = (\mathbf{Z}^T, \mathbf{U}^T)^T$. To eliminate $\pi(Y + \mathbf{U}^T\boldsymbol{\eta})$ in the likelihood, Zhao and Shao (2015) considered an alternative pseudo-likelihood based on the conditional distribution

$$\begin{aligned} [\mathbf{Z}|Y, \mathbf{U}, R = 1] &\sim \frac{\pi(Y + \mathbf{U}^T\boldsymbol{\eta})f(Y|\mathbf{X}, \boldsymbol{\Theta})h(\mathbf{U}|\mathbf{Z})}{\int \pi(Y + \mathbf{U}^T\boldsymbol{\eta})f(Y|\mathbf{X}, \boldsymbol{\Theta})h(\mathbf{U}|\mathbf{Z})d\mathbf{Z}} \\ &= \frac{f(Y|\mathbf{X}, \boldsymbol{\Theta})h(\mathbf{U}|\mathbf{Z})}{\int f(Y|\mathbf{X}, \boldsymbol{\Theta})h(\mathbf{U}|\mathbf{Z})d\mathbf{Z}}. \end{aligned}$$

Under suitable regularity conditions, Zhao and Shao (2015) showed that all parameters in f are identifiable in this pseudo-likelihood approach. Even though this approach can eliminate $\pi(Y + \mathbf{U}^T\boldsymbol{\eta})$, the conditional density $h(\mathbf{U}|\mathbf{Z})$ appears in the pseudo-likelihood. Therefore, either parametric assumptions are required or a nonparametric method, such as kernel smoothing is needed. When the dimension of $\mathbf{Z}$ is high, nonparametric methods are not feasible due to the notorious curse of dimensionality problem.

In this article, we also consider a pseudo-likelihood approach but our likelihood is based on the joint distribution of the complete observations, that is, $[Y, \mathbf{X}|R = 1]$. In the likelihood contribution, the complete data can be decomposed as

$$[Y, \mathbf{X}|R = 1] = [Y, \mathbf{U}|R = 1][\mathbf{Z}|Y, \mathbf{U}, R = 1]. \tag{5}$$

This approach is more efficient than the approach of Zhao and Shao (2015), since the first factor in (5) is ignored in their work. This conjecture is validated in the simulation study, which shows that in the situations we studied, the proposed method is uniformly more efficient than the Zhao and Shao (2015) method.

We observe that

$$[Y, \mathbf{X}|R = 1] \sim \frac{\pi(Y + \mathbf{U}^T\boldsymbol{\eta})f(Y|\mathbf{X}, \boldsymbol{\Theta})g(\mathbf{X})}{\int\int \pi(Y + \mathbf{U}^T\boldsymbol{\eta})f(Y|\mathbf{X}, \boldsymbol{\Theta})dG(\mathbf{X})dY},$$

where $G(\mathbf{X})$ is the corresponding distribution function of $g(\mathbf{X})$.

We study the semiparametric maximum likelihood estimation of $\boldsymbol{\Theta}$ and $\Pi = \int \pi(Y^*)dY^*$. Based on the best of our knowledge, in statistical literature all inferences on the propensity score $\pi(\cdot)$ are based on some parametric assumptions. It is our goal in this article to explore statistical inference for $\boldsymbol{\Theta}$ by treating $\pi(\cdot)$ nonparametrically.

2.1. Inference When η Is Known

Since the covariate $\mathbf{X}_i, i = 1, 2, \ldots, n$ is available for each observation, we can replace $G(\mathbf{X})$ by the empirical distribution $G_n(\mathbf{X}) = n^{-1}\sum_{i=1}^n I(\mathbf{X}_i \le \mathbf{X})$. Using the transformation $Y^* = Y + \mathbf{U}^T\boldsymbol{\eta}$,

$$\begin{aligned}
&\int\int \pi(Y + \mathbf{U}^T\boldsymbol{\eta})f(Y|\mathbf{X}, \boldsymbol{\Theta})dG_n(\mathbf{X})dY \\
&= \int\int \pi(Y^*)f(Y^* - \mathbf{U}^T\boldsymbol{\eta} - \mathbf{X}^T\boldsymbol{\beta}, \boldsymbol{\gamma})dG_n(\mathbf{X})dY^* \\
&= \int \pi(Y^*)\psi_n(Y^*, \boldsymbol{\Theta}, \boldsymbol{\eta})dY^*,
\end{aligned}$$

where $\psi_n(Y^*, \boldsymbol{\Theta}, \boldsymbol{\eta}) = n^{-1}\sum_{i=1}^n f(Y^* - \mathbf{X}_i^T\boldsymbol{\beta} - \mathbf{U}_i^T\boldsymbol{\eta}, \boldsymbol{\gamma})$.

Without loss of generality, we assume the outcome in the first n_1 observations are observed, that is, $R_i = 1, i = 1, \ldots, n_1, R_i = 0, i = n + 1, \ldots, n$. The pseudo-likelihood is

$$\begin{aligned}
L &= \prod_{i=1}^{n_1} \frac{\pi(Y_i + \mathbf{U}_i^T\boldsymbol{\eta})f(Y_i|\mathbf{X}_i, \boldsymbol{\Theta})}{\int\int \pi(Y + Z^T\boldsymbol{\eta})f(Y|\mathbf{X}, \boldsymbol{\Theta})dG_n(\mathbf{X})dY^*} \\
&= \prod_{i=1}^{n_1} \frac{\pi(Y_i^*)f(Y_i|\mathbf{X}_i, \boldsymbol{\Theta})}{\int \pi(Y^*)\psi_n(Y^*, \boldsymbol{\Theta}, \boldsymbol{\eta})dY^*}.
\end{aligned}$$

Note that $\int \pi(Y^*)dY^*$ is not necessarily 1. Let $\Pi = \int \pi(Y^*)dY^*$, then

$$r(Y^*) = \frac{\pi(Y^*)}{\Pi}$$

is a legitimate density. The pseudo-likelihood is proportional to

$$\prod_{i=1}^{n_1} \frac{r(Y_i^*)f(Y_i|\mathbf{X}_i, \boldsymbol{\Theta})}{\int r(Y^*)\psi_n(Y^*, \boldsymbol{\Theta}, \boldsymbol{\eta})dY^*}. \tag{6}$$

We can treat this problem as a biased sampling problem with underlying density $r(Y^*)$ and weight function $f(Y_i|\mathbf{X}_i, \boldsymbol{\Theta}), i = 1, 2, \ldots, n_1$. However, this problem differs from the conventional biased sampling problem discussed in Qin (2017, chap. 1), since in the current context, the roles of the underlying density and the weight function are interchanged.

Let $p_i = r(Y_i^*)dY_i^*, i = 1, \ldots, n_1$. If $\boldsymbol{\eta}$ is known, we can obtain the following semiparametric log pseudo-likelihood

$$\ell(\boldsymbol{\Theta}|\boldsymbol{\eta}) = \sum_{i=1}^{n_1}\Big[\log p_i + \log f(Y_i|\mathbf{X}_i, \boldsymbol{\Theta})\Big] - n_1 \log\Big[\sum_{i=1}^{n_1} p_i\psi_n(Y_i^*, \boldsymbol{\Theta}, \boldsymbol{\eta})\Big],$$

subject to the constraint $\sum_{i=1}^{n_1} p_i = 1, p_i \ge 0$. It is easy to show that

$$\begin{aligned}
&\hat{p}_i = \hat{r}(Y_i^*)dY_i^* = \frac{1/\psi_n(Y_i^*, \boldsymbol{\Theta}, \boldsymbol{\eta})}{C}, \quad \text{where} \\
&C = \sum_{j=1}^{n_1} \frac{1}{\psi_n(Y_j^*, \boldsymbol{\Theta}, \boldsymbol{\eta})}, \qquad i = 1, 2, \ldots, n_1
\end{aligned} \tag{7}$$

is the solution. After some algebra, we obtain the following profile log pseudo-likelihood

$$\ell(\boldsymbol{\Theta}|\boldsymbol{\eta}) = \sum_{i=1}^{n_1}\Big[\log f(Y_i|\mathbf{X}_i, \boldsymbol{\Theta}) - \log\psi_n(Y_i^*, \boldsymbol{\Theta}, \boldsymbol{\eta})\Big]. \tag{8}$$

Hence, an estimate of $\boldsymbol{\Theta}$ can be obtained by maximizing the semiparametric log pseudo-likelihood function $\ell(\boldsymbol{\Theta}|\boldsymbol{\eta})$ based on a given $\boldsymbol{\eta}$.

We can also understand the profile log pseudo-likelihood by an alternative conditional likelihood argument. Note that

$$[Y, \mathbf{X}|R = 1] \sim \frac{\pi(Y + \mathbf{U}^T\boldsymbol{\eta})f(Y|\mathbf{X}, \boldsymbol{\Theta})g(\mathbf{X})}{P(R = 1)}.$$

Let $Y^* = Y + \mathbf{U}^T\boldsymbol{\eta}$, then

$$[Y^*, \mathbf{X}|R = 1] \sim \frac{\pi(Y^*)f(Y^* - \mathbf{X}^T\boldsymbol{\beta} - \mathbf{U}^T\boldsymbol{\eta}, \boldsymbol{\gamma})g(\mathbf{X})}{P(R = 1)},$$

which gives

$$[Y^*|R = 1] \sim \frac{\pi(Y^*)\int f(Y^* - \mathbf{X}^T\boldsymbol{\beta} - \mathbf{U}^T\boldsymbol{\eta}, \boldsymbol{\gamma})g(\mathbf{X})d\mathbf{X}}{P(R = 1)}.$$

Consequently,

$$\begin{aligned}
[\mathbf{X}|R = 1, Y^*] &\sim \frac{f(Y^* - \mathbf{X}^T\boldsymbol{\beta} - \mathbf{U}^T\boldsymbol{\eta}, \boldsymbol{\gamma})g(\mathbf{X})}{\int f(Y^* - \mathbf{X}^T\boldsymbol{\beta} - \mathbf{U}^T\boldsymbol{\eta}, \boldsymbol{\gamma})g(\mathbf{X})d\mathbf{X}} \\
&= \frac{f(Y|\mathbf{X}, \boldsymbol{\Theta})g(\mathbf{X})}{\int f(Y^* - \mathbf{X}^T\boldsymbol{\beta} - \mathbf{U}^T\boldsymbol{\eta}, \boldsymbol{\gamma})g(\mathbf{X})d\mathbf{X}},
\end{aligned}$$

where the unknown $\pi(Y^*)$ has been factored out. Replacing $g(\mathbf{X})d\mathbf{X} = dG(\mathbf{X})$ by the marginal empirical distribution function $dG_n(\mathbf{X})$, we arrive at the profile log pseudo-likelihood (8). This pseudo-likelihood approach was also pointed out in Tang et al. (2003) when $\boldsymbol{\eta}$ is known.

The conditional likelihood argument eliminates the unknown propensity score $\pi(\cdot)$ for the case that $\boldsymbol{\eta}$ is known. In the next section, we consider the more commonly encountered situation when $\boldsymbol{\eta}$ is not known. In that situation, we find a profile "synthetic" distribution function technique for estimating $\boldsymbol{\eta}$ and Π.

2.2. Inference When η Is Unknown

In practice $\boldsymbol{\eta}$ is often unknown, the log pseudo-likelihood $\ell(\boldsymbol{\Theta}|\boldsymbol{\eta})$ discussed in (8) in the last section does not have information on $\boldsymbol{\eta}$. Instead we will use the incomplete observations $(R_i = 0, \mathbf{X}_i), i = n_1 + 1, \ldots, n$ to recover information about $\boldsymbol{\eta}$ through the construction of a binomial likelihood, discussed below. Since $(R_i, \mathbf{X}_i), i = 1, 2, \ldots, n$ are available for all observations, we use this fact to extract information from the propensity score function. For a given covariate $\mathbf{X}$, the propensity score can be written as

$$P(R = 1|\mathbf{X}) = \int \pi(Y + \mathbf{U}^T\boldsymbol{\eta}) f(Y|\mathbf{X}, \boldsymbol{\Theta}) dY \equiv \pi^*(\mathbf{X}, \boldsymbol{\Theta}, \boldsymbol{\eta}). \tag{9}$$

Notice as opposed to (2), in propensity score representation (9), the Y_i's that are unobservable when $R_i = 0$ have been eliminated. This fact allows us to extract information about the unknown parameter $\boldsymbol{\eta}$ in the propensity score function. Using (9), the binomial likelihood of $\boldsymbol{\eta}$ given $\boldsymbol{\Theta}$ is

$$L(\boldsymbol{\eta}|\boldsymbol{\Theta}) = \prod_{i=1}^{n} \left[\pi^*(\mathbf{X}_i, \boldsymbol{\Theta}, \boldsymbol{\eta})\right]^{R_i} \left[1 - \pi^*(\mathbf{X}_i, \boldsymbol{\Theta}, \boldsymbol{\eta})\right]^{1-R_i}. \tag{10}$$

We can make inference on $\boldsymbol{\eta}$ if $\pi^*(\mathbf{X}, \boldsymbol{\Theta}, \boldsymbol{\eta})$ can be estimated. To do this, we first note that,

$$\begin{aligned} P(R = 1) &= \int\int \pi(Y + \mathbf{U}^T\boldsymbol{\eta}) f(Y|\mathbf{X}, \boldsymbol{\Theta}) dG(\mathbf{X}) dY \\ &= \int\int \pi(Y^*) f(Y^* - \mathbf{X}^T\boldsymbol{\beta} - \mathbf{U}^T\boldsymbol{\eta}, \boldsymbol{\gamma}) dY^* dG(\mathbf{X}) \\ &= \Pi \int\int r(Y^*) f(Y^* - \mathbf{X}^T\boldsymbol{\beta} - \mathbf{U}^T\boldsymbol{\eta}, \boldsymbol{\gamma}) dY^* dG(\mathbf{X}). \end{aligned}$$

Replacing $dG(\mathbf{X})$ by $dG_n(\mathbf{X})$ and $P(R = 1)$ by n_1/n, we obtain

$$\frac{n_1}{n} = \Pi \int r(Y^*) \psi_n(Y^*, \boldsymbol{\Theta}, \boldsymbol{\eta}) dY^*.$$

Moreover replacing $r(Y^*)dY^*$ by $\hat{p}_i = \hat{r}(Y_i^*)dY_i^*$, $i = 1, 2, \ldots, n_1$, gives

$$\frac{n_1}{n} = \Pi \frac{n_1}{C}.$$

Therefore, we can estimate Π by

$$\hat{\Pi} = \frac{C}{n}. \tag{11}$$

Similarly, we can write

$$\begin{aligned} \pi^*(\mathbf{X}_i, \boldsymbol{\Theta}, \boldsymbol{\eta}) &= \int \pi(Y + \mathbf{U}_i^T\boldsymbol{\eta}) f(Y|\mathbf{X}_i, \boldsymbol{\Theta}) dY \\ &= \int \pi(Y^*) f(Y^* - \mathbf{X}_i^T\boldsymbol{\beta} - \mathbf{U}_i^T\boldsymbol{\eta}, \boldsymbol{\gamma}) dY^* \\ &= \Pi \int r(Y^*) f(Y^* - \mathbf{X}_i^T\boldsymbol{\beta} - \mathbf{U}_i^T\boldsymbol{\eta}, \boldsymbol{\gamma}) dY^* \end{aligned} \tag{12}$$

From (12), we estimate $\pi^*(\mathbf{X}_i, \boldsymbol{\Theta}, \boldsymbol{\eta})$ using (7), (11) and a sample equivalence of (12), as follows

$$\begin{aligned} \hat{\pi}^*(\mathbf{X}_i, \boldsymbol{\Theta}, \boldsymbol{\eta}) &= \hat{\Pi} \sum_{j=1}^{n_1} \hat{r}_j f(Y_j^* - \mathbf{X}_i^T\boldsymbol{\beta} - \mathbf{U}_i^T\boldsymbol{\eta}, \boldsymbol{\gamma}) \\ &= \frac{C}{n} \sum_{j=1}^{n_1} \frac{1}{\psi_n(Y_j^*, \boldsymbol{\Theta}, \boldsymbol{\eta})/C} f(Y_j^* - \mathbf{X}_i^T\boldsymbol{\beta} - \mathbf{U}_i^T\boldsymbol{\eta}, \boldsymbol{\gamma}) \\ &= \frac{1}{n} \sum_{j=1}^{n_1} \frac{f(Y_j^* - \mathbf{X}_i^T\boldsymbol{\beta} - \mathbf{U}_i^T\boldsymbol{\eta}, \boldsymbol{\gamma})}{\psi_n(Y_j^*, \boldsymbol{\Theta}, \boldsymbol{\eta})}. \end{aligned} \tag{13}$$

For a fixed $\boldsymbol{\Theta}$, the following log pseudo-likelihood function can be maximized with respect to $\boldsymbol{\eta}$,

$$\begin{aligned} \ell^*(\boldsymbol{\eta}|\boldsymbol{\Theta}) = \sum_{i=1}^{n} R_i \log[\hat{\pi}^*(\mathbf{X}_i, \boldsymbol{\Theta}, \boldsymbol{\eta})] + (1 - R_i) \\ \log[1 - \hat{\pi}^*(\mathbf{X}_i, \boldsymbol{\Theta}, \boldsymbol{\eta})] \end{aligned} \tag{14}$$

Combining (8) and (14), maximum pseudo-likelihood estimates of $\boldsymbol{\Theta}$ and $\boldsymbol{\eta}$ can be obtained using the following algorithm. Let $(\boldsymbol{\Theta}^{(0)}, \boldsymbol{\eta}^{(0)})$ be initial values of $(\boldsymbol{\Theta}, \boldsymbol{\eta})$. For $k = 1, 2, \ldots$, iterate between

1. for fixed $\boldsymbol{\eta}^{(k-1)}$, obtain $\boldsymbol{\Theta}^{(k)}$ by maximizing (8), $\ell(\boldsymbol{\Theta}|\boldsymbol{\eta}^{(k-1)})$, with respect to $\boldsymbol{\Theta}$;
2. let $Y_j^{*(k)} = Y_j + \mathbf{U}_j^T\boldsymbol{\eta}^{(k-1)}$ and, given $\boldsymbol{\Theta}^{(k)}$,

$$\hat{\pi}^{*(k)}(\mathbf{X}_i, \boldsymbol{\Theta}^{(k)}, \boldsymbol{\eta}) = \frac{1}{n} \sum_{j=1}^{n_1} \frac{f(Y_j^{*(k)} - \mathbf{X}_i^T\boldsymbol{\beta}^{(k)} - \mathbf{U}_i^T\boldsymbol{\eta}, \boldsymbol{\gamma}^{(k)})}{\psi_n(Y_j^{*(k)}, \boldsymbol{\Theta}^{(k)}, \boldsymbol{\eta})}.$$

maximize (14)

$$\begin{aligned} \ell^*(\boldsymbol{\eta}|\boldsymbol{\Theta}^{(k)}) = \sum_{i=1}^{n} R_i \log \hat{\pi}^{*(k)}(\mathbf{X}_i, \boldsymbol{\Theta}^{(k)}, \boldsymbol{\eta}) + (1 - R_i) \\ \log\{1 - \hat{\pi}^{*(k)}(\mathbf{X}_i, \boldsymbol{\Theta}^{(k)}, \boldsymbol{\eta})\} \end{aligned}$$

with respect to $\boldsymbol{\eta}$ to obtain $\boldsymbol{\eta}^{(k)}$.

At convergence, define the final values of $\boldsymbol{\Theta}^{(k)}, \boldsymbol{\eta}^{(k)}$ as $\hat{\boldsymbol{\Theta}}$ and $\hat{\boldsymbol{\eta}}$.

The method suggested in (13) gives an estimate of the integral of the propensity score, which offers indirect information on the form of the propensity score. Existing works either completely specify the form of the propensity score (e.g., Greenlees, Reece, and Zieschang 1982) or eliminate it from consideration (e.g., Tang, Little, and Raghunathan 2003).

3. Asymptotic Properties

In this section, we give regularity conditions and large sample results of the proposed maximum pseudo-likelihood estimate.

Let $\boldsymbol{\eta}_0$ be the true value of $\boldsymbol{\eta}$, $\boldsymbol{\Theta}(\boldsymbol{\eta}_0) = \boldsymbol{\Theta}_0$ and $\hat{\boldsymbol{\Theta}}(\hat{\boldsymbol{\eta}}) = \hat{\boldsymbol{\Theta}}$. Define $\psi(Y^*, \boldsymbol{\Theta}, \boldsymbol{\eta}) = \int f(Y^* - \mathbf{X}^T\boldsymbol{\beta} - \mathbf{U}^T\boldsymbol{\eta}, \boldsymbol{\gamma}) dG(\mathbf{X})$. Let $p = p_1 + p_2$, $\mathcal{A}$ be the parameter space of $\boldsymbol{\Theta}$, $\mathcal{D}$ be the parameter space of $\boldsymbol{\eta}$. The following regularity conditions are required:

- (C_1) $\mathcal{A}$ and $\mathcal{D}$ are compact subsets of $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively.
- (C_2) $\partial f(Y|\mathbf{X}, \boldsymbol{\Theta})/\partial\boldsymbol{\Theta}, \partial^2 f(Y|\mathbf{X}, \boldsymbol{\Theta})/\partial\boldsymbol{\Theta}^2, \partial^3 f(Y|\mathbf{X}, \boldsymbol{\Theta})/\partial\boldsymbol{\Theta}^3$ are bounded for $\boldsymbol{\Theta} \in \mathcal{A}$ and $\partial\pi(Y + \mathbf{U}^T\boldsymbol{\eta})/\partial\boldsymbol{\eta}, \partial^2\pi(Y + \mathbf{U}^T\boldsymbol{\eta})/\partial\boldsymbol{\eta}^2, \partial^3\pi(Y + \mathbf{U}^T\boldsymbol{\eta})/\partial\boldsymbol{\eta}^3$ are bounded for $\boldsymbol{\eta} \in \mathcal{D}$; $0 < f(Y|\mathbf{X}, \boldsymbol{\Theta}) < \infty$ is bounded for $\boldsymbol{\Theta} \in \mathcal{A}$, $0 < \pi(Y + \mathbf{U}^T\boldsymbol{\eta}) < \infty$ is bounded for $\boldsymbol{\eta} \in \mathcal{D}$.
- (C_3) $H_1(\boldsymbol{\Theta}|\boldsymbol{\eta})$ and $H^*(\boldsymbol{\eta}|\boldsymbol{\Theta})$, defined in the online Supplementary Materials, are nonsingular for $\boldsymbol{\Theta} \in \mathcal{A}$ and $\boldsymbol{\eta} \in \mathcal{D}$.
- (C_4) Σ_1, Σ_2, defined in the online Supplementary Materials, satisfy $\Sigma_1 < \infty, \|H_1(\boldsymbol{\Theta}_0|\boldsymbol{\eta}_0)\| < \infty$ and $\Sigma_2 < \infty, \|H^*(\boldsymbol{\Theta}_0|\boldsymbol{\eta}_0)\| < \infty$.

Among these conditions, (C_1) assumes compact parameter spaces, which is commonly used in the literature; (C_2) specifies the smoothness of $f(Y|\mathbf{X}, \boldsymbol{\Theta})$ and $\pi(Y + \mathbf{U}^T\boldsymbol{\eta})$; ($C_3$) ensures the second-order derivatives of the profile log pseudo-likelihood $\ell(\boldsymbol{\Theta}|\boldsymbol{\eta})$ and log pseudo-likelihood $\ell_n^*(\boldsymbol{\eta}|\boldsymbol{\Theta})$ are nonsingular; and (C_4) guarantees finiteness of the covariance matrix for the proposed estimators.

Consistency of the proposed estimators is established in the following theorem:

Theorem 1. Assume conditions of Lemma 1 and (C_1)–(C_3) hold, then $\hat{\boldsymbol{\Theta}} \xrightarrow{P} \boldsymbol{\Theta}_0$, and $\hat{\boldsymbol{\eta}} \xrightarrow{P} \boldsymbol{\eta}_0$.

Asymptotic normality of the proposed estimators is given in following theorem:

Theorem 2. Assume conditions of Lemma 1 and (C_1)–(C_4) hold, then

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{D} N(0, H^{*-1}\Sigma_1 H^{*-1}), \quad \text{and}$$
$$\sqrt{n}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0) \xrightarrow{D} N(0, H_1^{-1}\Sigma_2 H_1^{-1}),$$

where $H_1 = H_1(\boldsymbol{\Theta}_0|\boldsymbol{\eta}_0)$, $H^* = H^*(\boldsymbol{\eta}_0|\boldsymbol{\Theta}_0)$.

Write $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$, for any arbitrary vector $\mathbf{a}$. The variances of the proposed estimator of $\hat{\boldsymbol{\eta}}$ and $\hat{\Theta}$ can be estimated by $H_n^{*-1}(\hat{\boldsymbol{\eta}}|\hat{\boldsymbol{\Theta}})\hat{\Sigma}_1 H_n^{*-1}(\hat{\boldsymbol{\eta}}|\hat{\boldsymbol{\Theta}})$ and $H_{1n}^{-1}(\hat{\boldsymbol{\Theta}}|\hat{\boldsymbol{\eta}})\hat{\Sigma}_2 H_{1n}^{-1}(\hat{\boldsymbol{\Theta}}|\hat{\boldsymbol{\eta}})$, respectively, where

$$\hat{\Sigma}_1 = \frac{1}{n}\sum_{i=1}^{n}\phi_i^{\otimes 2}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\eta}}), \quad \hat{\Sigma}_2 = \frac{1}{n}\sum_{i=1}^{n}\phi_i^{*\otimes 2}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\eta}}),$$

$$H_n^*(\hat{\boldsymbol{\eta}}|\hat{\boldsymbol{\Theta}}) = H_{1n}^*(\hat{\boldsymbol{\eta}}|\hat{\boldsymbol{\Theta}}) + H_{2n}^*(\hat{\boldsymbol{\eta}}|\hat{\boldsymbol{\Theta}})H_{1n}(\hat{\boldsymbol{\Theta}}|\hat{\boldsymbol{\eta}})^{-1}H_{2n}(\hat{\boldsymbol{\Theta}}|\hat{\boldsymbol{\eta}}),$$

$$\phi_i(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\eta}}) = g_i^*(\hat{\boldsymbol{\eta}}|\hat{\boldsymbol{\Theta}}) - H_{2n}^*(\hat{\boldsymbol{\eta}}|\hat{\boldsymbol{\Theta}})H_{1n}^{-1}(\hat{\boldsymbol{\Theta}}|\hat{\boldsymbol{\eta}})g_i(\hat{\boldsymbol{\Theta}}|\hat{\boldsymbol{\eta}}),$$

$$\phi_i^*(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\eta}}) = g_i(\hat{\boldsymbol{\Theta}}|\hat{\boldsymbol{\eta}}) - H_{2n}(\hat{\boldsymbol{\Theta}}|\hat{\boldsymbol{\eta}})\big\{H_n^{*-1}(\hat{\boldsymbol{\eta}}|\hat{\boldsymbol{\Theta}})\big[g_i^*(\hat{\boldsymbol{\eta}}|\hat{\boldsymbol{\Theta}}) - H_{2n}^*(\hat{\boldsymbol{\eta}}|\hat{\boldsymbol{\Theta}})H_{1n}^{-1}(\hat{\boldsymbol{\Theta}}|\hat{\boldsymbol{\eta}})g_i(\hat{\boldsymbol{\Theta}}|\hat{\boldsymbol{\eta}})\big]\big\},$$

$$g_i(\hat{\boldsymbol{\Theta}}|\hat{\boldsymbol{\eta}}) = R_i\left[\frac{\partial f(Y_i|\mathbf{X}_i, \hat{\boldsymbol{\Theta}})/\partial\boldsymbol{\Theta}}{f(Y_i|\mathbf{X}_i, \hat{\boldsymbol{\Theta}})} - \frac{\partial\psi_n(Y_i^*, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\eta}})/\partial\boldsymbol{\Theta}}{\psi_n(Y_i^*, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\eta}})}\right],$$

$$g_i^*(\hat{\boldsymbol{\eta}}|\hat{\boldsymbol{\Theta}}) = R_i\frac{\partial\hat{\pi}^*(\mathbf{X}_i, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\eta}})/\partial\boldsymbol{\eta}}{\hat{\pi}^*(\mathbf{X}_i, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\eta}})} - (1 - R_i)\frac{\partial\hat{\pi}^*(\mathbf{X}_i, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\eta}})/\partial\boldsymbol{\eta}}{1 - \hat{\pi}^*(\mathbf{X}_i, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\eta}})},$$

where $H_{1n}(\boldsymbol{\Theta}|\boldsymbol{\eta}), H_{2n}(\boldsymbol{\Theta}|\boldsymbol{\eta}), H_{1n}^*(\boldsymbol{\eta}|\boldsymbol{\Theta})$, and $H_{2n}^*(\boldsymbol{\eta}|\boldsymbol{\Theta})$ are given in the online supplementary materials. The formula of the estimator for asymptotic variance is very complicated, hence bootstrap method is used to estimate the variance.

Remark 1. Our method is based on $[Y, \mathbf{X}|R = 1]$ while the method in Zhao and Shao (2015) is based on $[\mathbf{Z}|Y, \mathbf{U}, R = 1]$, therefore, we argue that the proposed method should be more efficient because it includes information of the conditional likelihood $[Y, \mathbf{U}|R = 1]$. However, it is difficult to make direct theoretical comparison between them since the variance of the two estimators depend on different quantities that cannot be compared directly. Specifically, the variance of $\hat{\boldsymbol{\Theta}}$ is affected by the $\hat{\boldsymbol{\eta}}$ and the empirical distribution function $G_n(\mathbf{X})$, while the variance of the Zhao and Shao (2015) estimator depends on the empirical distribution of $\mathbf{Z}$ and the estimated parameter in the parametric model of $\mathbf{U}$ given $\mathbf{Z}$.

Remark 2. Under our semiparametric setting, the unknown propensity score function $\pi(\cdot)$ and the marginal distribution function of $\mathbf{X}$ are infinite dimensional nuisance parameters, whereas $\boldsymbol{\eta}$ and $\boldsymbol{\Theta}$ are finite dimensional parameters of interest. Naturally, it is of interest to determine whether the proposed estimators can attain the semiparametric efficiency bounds. However, a concise analytic expression of the efficient score function is not obtainable in general. It worth noting that in existing literature, discussions of semiparametric efficiency under nonignorable missingness are based on a parametric model for propensity score and moment restriction model for the response and covariates. This is because the nuisance tangent space and hence the efficient score has simple and concise forms in that case, more detail see, for example, Rotnitzky and Robin (1997), Miao et al. (2019), and Morikawa and Kim (2019). The question of how to develop a method for studying semiparametric efficiency under more general settings is a topic worth further study.

4. Simulation Study

In this section, we evaluate the finite sample performance of the proposed maximum pseudo-likelihood estimator (MPCL). We compare MPCL to the complete case only method (CC), the maximum parametric likelihood estimation method based on correctly specified models (ML), and misspecified models (MML), the method proposed in Tang, Little, and Raghunathan (2003) (TLR) and the method proposed in Zhao and Shao (2015) (ZS). Let $\mathbf{X} = (X_1, \mathbf{U}^T = (X_2, X_3)^T)^T$, where $X_1 \sim N(0, 1)$, $X_2 \sim U(0, 1)$, and $X_3 \sim N(0, 1)$. The outcome Y is generated from the following model:

$$Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + \epsilon,$$

where $\epsilon \sim N(0, \gamma)$, $\beta_1 = 1.5, \beta_2 = -1, \beta_3 = 2, \beta_4 = 1$, and $\gamma = 1$. We consider four missing data scenarios, one using a logistic model and another using a probit model

$S_1 : P(R = 1|Y, \mathbf{U}) = \exp(Y + \boldsymbol{\eta}^T\mathbf{U})/[1 + \exp(Y + \boldsymbol{\eta}^T\mathbf{U})]$

$S_2 : P(R = 1|Y, \mathbf{U}) = \Phi(Y + \boldsymbol{\eta}^T\mathbf{U})$, where Φ is the standard normal distribution function

$S_3 : P(R = 1|Y, \mathbf{U}) = \exp(Y + \boldsymbol{\eta}^T\tilde{\mathbf{U}})/[1 + \exp(Y + \boldsymbol{\eta}^T\tilde{\mathbf{U}})]$, where $\tilde{\mathbf{U}} = (\tilde{U}_1, \tilde{U}_2)$ and $\tilde{U}_i = 1 - (0.5U_i)^2, i = 1, 2$

$S_4 : P(R = 1|Y, \mathbf{U}) = \exp(Y + \boldsymbol{\eta}^T\tilde{\mathbf{U}})/[1 + \exp(Y + \boldsymbol{\eta}^T\tilde{\mathbf{U}})]$, where $\tilde{\mathbf{U}} = (\tilde{U}_1, \tilde{U}_2)$ and $\tilde{U}_i = 1 - (1 - 0.5U_i)^2, i = 1, 2$

Notice that S_1 and S_2 are linear and monotonic in U_i, S_3 is nonmonotonic and symmetric in U_i (assuming the chance of a normal variate exceeding 6 standard deviations from the mean is practically zero), S_4 is nonmonotonic and asymmetric in U_i. For S_1 and S_2, we consider three different levels of missingness ratio: M_1: $\boldsymbol{\eta} = (0, 0)$; M_2: $\boldsymbol{\eta} = (-2, -1)$; M_3: $\boldsymbol{\eta} = (-4, -5)$. For S_3 and S_4, we use levels of missingness ratio: M_4: $\boldsymbol{\eta} = (-0.5, -0.5)$; M_5: $\boldsymbol{\eta} = (-2, -2)$, respectively.

The simulation missingness ratios in S_1 and S_2, under M_1–M_3, are approximately: 16%, 30%, 50% and 12%, 28%, 50%, respectively. Note that the propensity score depends only on the outcome Y under M_1, while it depends on both the outcome and the covariates under M_2 and M_3, that is, the requirement of Tang, Little, and Raghunathan (2003) is satisfied in M_1 but not under M_2 and M_3. For S_3 and S_4, the missingness ratios are approximately 30%.

The simulation results are given in Tables 1–8. We evaluate the performance of the estimators based on bias (BIAS), standard error (SE), estimated standard error (SEE, obtained by bootstrap), proximity of empirical confidence interval coverage to the nominal target coverage of 0.95 (CP), and mean square error (MSE). All simulations use a sample size of $n = 300$ with 500 replications and 300 bootstrap resampling for SEE. We used a probit model for S_1 and logistic model for S_2–S_4. For MML, the outcome model is always correctly specified while the propensity score is misspecified such that it uses a logit model when the actual is probit and vice versa. For ZS, the distribution of all covariates are assumed to be correctly specified.

To assess bias, we adopt a rule suggested by Olsen and Schafer (2001) that bias does not have undue influence on inference unless the standardized bias (bias over SE) exceeds 0.4. Applying that rule, we conclude that the estimates based on the proposed MPCL are unbiased. The sample SEs are similar to the corresponding SEEs. Moreover, the coverage probabilities are close to the target nominal level of 0.95. On the other hand, CC exhibits serious biases that lead to low coverage probabilities. As expected, the performance of ML based on a correctly specified data model and propensity score is the best among all methods. However, MML is seriously biased. For all methods, performance is inversely related to missingness ratio, in all scenarios.

Under M_1, the performance of TRL is slightly better than MPCL. This result is not unexpected since M_1 satisfies the requirements of TRL. Furthermore, TRL does not need to estimate the parameters in the propensity score; hence, the dimension of the unknown parameters for TRL is lower than that for the proposed method. Under M_2 and M_3, the propensity score depends on the covariates, in those cases, using TRL results in seriously biased estimates.

Table 1. Estimation results of M_1 under scenario S_1.

	MPCL						
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	−0.001	0.010	0.013	−0.007	−0.008	−0.046	0.033
SE	0.146	0.068	0.230	0.067	0.051	0.373	0.737
SEE	0.142	0.069	0.227	0.070	0.051	0.371	0.764
CP	0.942	0.944	0.942	0.964	0.938	0.947	0.952
MSE	0.021	0.005	0.053	0.005	0.003	0.141	0.542
	ML						
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	−0.005	0.010	0.018	−0.005	−0.007	−0.006	0.035
SE	0.132	0.065	0.218	0.066	0.048	0.221	0.409
SEE	0.129	0.065	0.216	0.064	0.047	0.213	0.407
CP	0.948	0.942	0.952	0.960	0.920	0.932	0.944
MSE	0.017	0.004	0.048	0.004	0.002	0.049	0.168
	MML						
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	−0.126	−0.022	0.109	0.011	0.046	0.257	−0.833
SE	0.137	0.067	0.224	0.064	0.053	0.176	0.306
SEE	0.137	0.067	0.226	0.066	0.053	0.165	0.311
CP	0.846	0.924	0.926	0.962	0.882	0.614	0.236
MSE	0.034	0.005	0.062	0.004	0.005	0.097	0.787
	TRL						
	β_1	β_2	β_3	β_4	γ		
BIAS	−0.007	0.007	0.017	−0.005	−0.007		
SE	0.142	0.067	0.230	0.067	0.050		
SEE	0.137	0.068	0.226	0.068	0.049		
CP	0.938	0.948	0.946	0.962	0.938		
MSE	0.020	0.005	0.053	0.005	0.003		
	ZS						
	β_1	β_2	β_3	β_4	γ		
BIAS	−0.003	−0.002	0.013	0.007	−0.008		
SE	0.186	0.085	0.312	0.151	0.088		
SEE	0.201	0.094	0.317	0.153	0.091		
CP	0.966	0.962	0.940	0.966	0.944		
MSE	0.034	0.007	0.097	0.023	0.008		
	CC						
	β_1	β_2	β_3	β_4	γ		
BIAS	0.197	0.077	−0.125	−0.075	−0.043		
SE	0.130	0.065	0.215	0.063	0.044		
SEE	0.126	0.063	0.213	0.063	0.043		
CP	0.636	0.736	0.914	0.762	0.790		
MSE	0.056	0.010	0.062	0.010	0.004		

It is also easy to observe that the proposed method is considerably more efficient than ZS, which confirms our conjecture in Section 2. In all simulations we carried out, the MPCL is consistently at least 2 times and can be up to more than 4 times (for β_4 and γ) more efficient than ZS, based on MSE. As suggested in Section 2, MPCL is based on the conditional distribution $Y, \mathbf{X}|R = 1$ while ZS is based on the conditional distribution, $\mathbf{U}|Y, Y^*, R = 1$, hence MPCL uses more information from the data than ZS. In the situations we studied, MPCL is more robust than ML and TRL, and considerably more efficient than ZS.

As pointed out by one of the referees, the proposed MPCL method involves iterated optimizations, and hence, computational effort of the method is a concern. For the simulation study, the average time per simulation (with no bootstrapping to illustrate the actual run time), all based on $n = 300$ observations, are (in seconds): 0.195, 1.528, 1.724, 2.124, 5.610, and 7.037, respectively, for CC, ML, MML, TLR, ZS, and MPCL.

Table 2. Estimation results of M_2 under scenario S_1.

	β_1	β_2	β_3	β_4	γ	η_1	η_2
	MPCL						
BIAS	0.021	0.012	−0.013	0.002	−0.014	−0.056	−0.041
SE	0.154	0.095	0.238	0.074	0.054	0.642	0.745
SEE	0.157	0.097	0.239	0.073	0.052	0.632	0.740
CP	0.946	0.946	0.954	0.938	0.928	0.937	0.955
MSE	0.024	0.009	0.057	0.006	0.003	0.415	0.557
	ML						
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	−0.033	0.045	0.011	0.007	−0.007	0.004	−0.012
SE	0.137	0.083	0.227	0.072	0.052	0.335	0.419
SEE	0.143	0.084	0.232	0.070	0.050	0.332	0.414
CP	0.958	0.938	0.954	0.946	0.928	0.954	0.942
MSE	0.019	0.007	0.052	0.005	0.003	0.107	0.177
	MML						
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	−0.142	−0.073	0.079	−0.004	0.037	0.767	−0.578
SE	0.154	0.088	0.247	0.072	0.057	0.272	0.322
SEE	0.150	0.087	0.241	0.071	0.056	0.259	0.325
CP	0.832	0.888	0.938	0.942	0.916	0.198	0.540
MSE	0.044	0.013	0.067	0.005	0.005	0.662	0.437
	TRL						
	β_1	β_2	β_3	β_4	γ		
BIAS	−0.237	−0.082	0.165	0.066	0.052		
SE	0.166	0.111	0.262	0.078	0.069		
SEE	0.162	0.106	0.255	0.075	0.065		
CP	0.696	0.868	0.896	0.850	0.906		
MSE	0.084	0.019	0.096	0.010	0.008		
	ZS						
	β_1	β_2	β_3	β_4	γ		
BIAS	−0.017	0.002	0.030	0.014	−0.007		
SE	0.223	0.122	0.363	0.169	0.099		
SEE	0.222	0.125	0.346	0.163	0.099		
CP	0.952	0.956	0.927	0.949	0.960		
MSE	0.050	0.015	0.132	0.029	0.010		
	CC						
	β_1	β_2	β_3	β_4	γ		
BIAS	0.240	0.171	−0.067	−0.053	−0.040		
SE	0.139	0.083	0.232	0.073	0.048		
SEE	0.141	0.083	0.232	0.070	0.047		
CP	0.622	0.464	0.924	0.856	0.842		
MSE	0.077	0.036	0.058	0.008	0.004		

Table 3. Estimation results of M_3 under scenario S_1.

	β_1	β_2	β_3	β_4	γ	η_1	η_2
	MPCL						
BIAS	0.033	0.029	0.018	0.004	−0.018	−0.048	0.034
SE	0.191	0.144	0.301	0.083	0.058	0.845	0.852
SEE	0.184	0.142	0.292	0.086	0.061	0.815	0.816
CP	0.930	0.934	0.928	0.952	0.940	0.955	0.932
MSE	0.038	0.022	0.091	0.007	0.004	0.714	0.725
	ML						
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	0.008	0.011	0.017	0.007	−0.014	−0.074	−0.042
SE	0.167	0.118	0.284	0.079	0.057	0.508	0.396
SEE	0.166	0.120	0.280	0.083	0.059	0.512	0.412
CP	0.942	0.954	0.966	0.938	0.964	0.944	0.958
MSE	0.029	0.014	0.081	0.006	0.003	0.263	0.158
	MML						
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	−0.152	−0.103	−0.003	−0.016	0.029	1.279	0.479
SE	0.169	0.123	0.294	0.084	0.065	0.414	0.345
SEE	0.173	0.124	0.290	0.083	0.065	0.418	0.345
CP	0.881	0.887	0.946	0.958	0.944	0.159	0.718
MSE	0.052	0.026	0.086	0.007	0.005	1.806	0.348
	TRL						
	β_1	β_2	β_3	β_4	γ		
BIAS	−0.547	−0.492	−0.276	0.092	0.180		
SE	0.245	0.223	0.383	0.098	0.133		
SEE	0.243	0.219	0.385	0.095	0.131		
CP	0.428	0.488	0.908	0.860	0.884		
MSE	0.359	0.292	0.222	0.018	0.050		
	ZS						
	β_1	β_2	β_3	β_4	γ		
BIAS	−0.013	0.014	0.036	0.019	−0.005		
SE	0.254	0.185	0.419	0.184	0.111		
SEE	0.261	0.181	0.426	0.186	0.114		
CP	0.966	0.950	0.953	0.958	0.963		
MSE	0.065	0.034	0.176	0.034	0.012		
	CC						
	β_1	β_2	β_3	β_4	γ		
BIAS	0.258	0.234	0.141	−0.039	−0.037		
SE	0.174	0.123	0.291	0.081	0.053		
SEE	0.167	0.121	0.283	0.083	0.055		
CP	0.646	0.490	0.902	0.926	0.874		
MSE	0.097	0.070	0.104	0.008	0.004		

As expected, the computing time is longest for MPCL, but manageable for practical purposes.

5. Empirical Illustration

For over half a century, the Peabody Picture Vocabulary Test (PPVT, Dunn and Dunn 2007) has been an important tool for measuring the receptive vocabulary in Standard American English. The test is standardized and age adjusted to be used for all age groups, native or nonnative English speakers, and independent of English proficiency level. In this section, we give results of a study of PPVT data collected as part of the National Longitudinal Survey of Child and Young Adult (NLSY79 Child). The NLSY79 Child survey is a longitudinal study that follows the biological children of women in NLSY79. As of 2016, more than 10,000 children have been interviewed in at least one survey round. The children in the survey are assessed and interviewed every two years. These assessments measure cognitive skills, temperament, motor and social development, self competence, behavioral issues, and their home environment. One of the assessments is the PPVT. There is a large literature on using PPVT as an index for family production models in economics (see, e.g., Becker 1981; Blau and Grossberg 1992, and references therein).

The PPVT is made up of a number of items. Each item consists of 4 pictures. The interviewer says a word loud and the child selects one of four pictures that best describes the word's meaning. Our sample comes from test results between 1986 and 1992. Our main interest is the trajectory of the test results in the intervening years. We limit our focus to children aged between 3 and 4 years at the 1986 assessment. These children were offered PPVT in 1986 and then again in. One of the research questions of interest is whether there is a change in scholastic aptitude and verbal skills over time and whether there is a gender bias

Table 4. Estimation results of M_1 under scenario S_2.

	β_1	β_2	β_3	β_4	γ	η_1	η_2
	MPCL						
BIAS	−0.001	0.011	0.013	−0.009	−0.008	−0.062	0.078
SE	0.141	0.068	0.226	0.065	0.051	0.294	0.719
SEE	0.135	0.067	0.219	0.067	0.050	0.326	0.704
CP	0.946	0.940	0.948	0.958	0.942	0.966	0.924
MSE	0.020	0.005	0.051	0.004	0.003	0.090	0.522
	ML						
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	−0.005	0.010	0.017	−0.006	−0.007	−0.009	0.033
SE	0.124	0.064	0.208	0.061	0.049	0.175	0.356
SEE	0.123	0.063	0.208	0.062	0.045	0.174	0.366
CP	0.958	0.934	0.944	0.954	0.940	0.952	0.958
MSE	0.015	0.004	0.044	0.004	0.002	0.031	0.128
	MML						
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	0.065	0.034	−0.041	−0.027	−0.030	−0.283	1.146
SE	0.128	0.065	0.215	0.061	0.047	0.224	0.439
SEE	0.125	0.063	0.213	0.063	0.044	0.217	0.423
CP	0.920	0.909	0.951	0.931	0.858	0.756	0.271
MSE	0.021	0.005	0.048	0.004	0.003	0.130	1.505
	TRL						
	β_1	β_2	β_3	β_4	γ		
BIAS	−0.007	0.008	0.019	−0.006	−0.007		
SE	0.136	0.067	0.224	0.064	0.050		
SEE	0.132	0.066	0.220	0.066	0.048		
CP	0.950	0.942	0.948	0.952	0.940		
MSE	0.018	0.005	0.051	0.004	0.003		
	ZS						
	β_1	β_2	β_3	β_4	γ		
BIAS	−0.015	−0.006	0.026	0.004	−0.004		
SE	0.190	0.083	0.313	0.150	0.093		
SEE	0.196	0.093	0.309	0.153	0.091		
CP	0.972	0.965	0.952	0.958	0.930		
MSE	0.036	0.007	0.098	0.022	0.009		
	CC						
	β_1	β_2	β_3	β_4	γ		
BIAS	0.208	0.094	−0.156	−0.093	−0.051		
SE	0.123	0.063	0.208	0.060	0.043		
SEE	0.120	0.062	0.205	0.061	0.041		
CP	0.574	0.634	0.886	0.650	0.724		
MSE	0.058	0.013	0.068	0.012	0.004		

Table 5. Estimation results of M_2 under scenario S_2.

	β_1	β_2	β_3	β_4	γ	η_1	η_2
	MPCL						
BIAS	0.007	0.021	0.017	−0.012	−0.012	−0.152	0.079
SE	0.157	0.102	0.245	0.070	0.056	0.541	0.647
SEE	0.153	0.098	0.236	0.071	0.052	0.551	0.658
CP	0.938	0.932	0.944	0.948	0.920	0.942	0.938
MSE	0.025	0.011	0.060	0.005	0.003	0.315	0.424
	ML						
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	−0.007	0.013	0.025	−0.008	−0.010	−0.021	−0.012
SE	0.143	0.089	0.235	0.067	0.052	0.268	0.333
SEE	0.140	0.085	0.228	0.068	0.050	0.291	0.357
CP	0.964	0.968	0.942	0.912	0.954	0.954	0.940
MSE	0.021	0.008	0.056	0.005	0.003	0.072	0.111
	MML						
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	0.083	0.074	−0.016	−0.018	−0.031	−0.900	0.772
SE	0.145	0.089	0.236	0.068	0.051	0.354	0.432
SEE	0.141	0.086	0.230	0.069	0.048	0.349	0.426
CP	0.892	0.858	0.951	0.951	0.868	0.293	0.598
MSE	0.028	0.013	0.056	0.005	0.004	0.934	0.782
	TRL						
	β_1	β_2	β_3	β_4	γ		
BIAS	−0.275	−0.107	0.195	0.082	0.068		
SE	0.170	0.116	0.263	0.074	0.072		
SEE	0.164	0.114	0.254	0.074	0.069		
CP	0.646	0.842	0.876	0.796	0.880		
MSE	0.104	0.025	0.107	0.012	0.010		
	ZS						
	β_1	β_2	β_3	β_4	γ		
BIAS	−0.015	0.004	0.027	0.011	−0.006		
SE	0.222	0.125	0.357	0.169	0.099		
SEE	0.223	0.129	0.342	0.162	0.098		
CP	0.956	0.956	0.950	0.957	0.960		
MSE	0.049	0.016	0.128	0.029	0.010		
	CC						
	β_1	β_2	β_3	β_4	γ		
BIAS	0.259	0.219	−0.049	−0.079	−0.045		
SE	0.142	0.089	0.234	0.068	0.049		
SEE	0.140	0.086	0.230	0.069	0.046		
CP	0.536	0.278	0.946	0.792	0.778		
MSE	0.087	0.056	0.057	0.011	0.004		

in the change. To obtain a proper assessment, we need to take into account maternal supply. Here, we use two measures of maternal supply, the average income of the mother between 1986 and 1992 and education attainment of the mother as of 1986. Hence, we further restrict the sample to only those whose mothers reported nonzero income in at least one year between 1986 and 1992. There are a total of $n = 557$ children who satisfy these criteria and who have valid assessments in 1986. By 1992 assessment, their age ranges between 9 and 10 years with a mean of 9.8 years. There are 282 males and 275 females.

A key characteristics of this sample is the significant amount of missing data. In 1992, there are only 387 valid assessments, giving a missing data rate of over 30%. There are a variety of reasons why a child might skip the assessment, for example, motivation, family influence, perceived poor performance, etc.. As a result, nonignorable missingness cannot be ruled out. Our goal here is to analyze the data using the proposed method, and to compare it with several other methods that make different assumptions on the missingness mechanisms.

Under the notations we defined in Section 1, we let the outcome Y be the difference in PPVT score between 1986 and 1992. The PPVT scores have been standardized to be normally distributed across age. Our main covariate of interest is Gender (1 = "Male," 0 = "Female"). Other covariates are Race (1 = "White," 0 = "Others"), Mother's income, Mother's education (1 = ">12 years," 0 = "≤ 12 years"). We obtain mother's income by taking the total income over the years a mother reported income (in the labor force) and dividing by the total hours of work reported during those years. We also created three binary dummy variables that classify the data by the four quartiles of the 1986 PPVT score, these are named Dummy_1, Dummy_2, Dummy_3. So a 1986 PPVT score in the 1st, 2nd, 3rd, and 4th quartiles would receive (0,0,0), (1,0,0), (0,1,0), (0,0,1), respectively, for

Table 6. Estimation results of M_3 under scenario S_2.

	β_1	β_2	β_3	β_4	γ	η_1	η_2
				MPCL			
BIAS	0.009	0.031	0.042	−0.012	−0.020	−0.057	0.017
SE	0.185	0.148	0.296	0.083	0.064	0.775	0.807
SEE	0.181	0.144	0.290	0.084	0.061	0.742	0.787
CP	0.942	0.932	0.928	0.948	0.920	0.929	0.915
MSE	0.034	0.023	0.089	0.007	0.005	0.603	0.650
				ML			
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	−0.006	0.020	0.044	−0.010	−0.016	−0.063	−0.021
SE	0.168	0.125	0.281	0.080	0.061	0.501	0.320
SEE	0.164	0.123	0.276	0.081	0.059	0.469	0.323
CP	0.944	0.934	0.942	0.962	0.932	0.922	0.952
MSE	0.028	0.016	0.081	0.006	0.004	0.254	0.103
				MML			
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	0.077	0.084	0.055	−0.011	−0.030	−1.192	−0.345
SE	0.160	0.124	0.278	0.084	0.062	0.462	0.376
SEE	0.169	0.125	0.285	0.082	0.058	0.471	0.401
CP	0.944	0.885	0.946	0.954	0.862	0.283	0.875
MSE	0.031	0.022	0.080	0.007	0.005	1.634	0.260
				TRL			
	β_1	β_2	β_3	β_4	γ		
BIAS	−0.693	−0.661	−0.375	0.122	0.251		
SE	0.311	0.314	0.430	0.101	0.202		
SEE	0.306	0.312	0.424	0.099	0.204		
CP	0.404	0.462	0.914	0.822	0.886		
MSE	0.577	0.535	0.325	0.025	0.104		
				ZS			
	β_1	β_2	β_3	β_4	γ		
BIAS	−0.009	0.013	0.033	0.015	−0.009		
SE	0.255	0.189	0.425	0.181	0.109		
SEE	0.263	0.190	0.426	0.186	0.114		
CP	0.956	0.950	0.946	0.955	0.955		
MSE	0.065	0.036	0.181	0.033	0.012		
				CC			
	β_1	β_2	β_3	β_4	γ		
BIAS	0.275	0.284	0.191	−0.064	−0.044		
SE	0.168	0.129	0.285	0.083	0.058		
SEE	0.168	0.127	0.283	0.082	0.055		
CP	0.614	0.384	0.88	0.882	0.828		
MSE	0.104	0.097	0.118	0.011	0.005		

Table 7. Estimation results under scenario S_3.

	β_1	β_2	β_3	β_4	γ	η_1	η_2
				MPCL			
BIAS	0.017	0.004	−0.010	−0.006	−0.010	−0.0835	−0.0995
SE	0.151	0.090	0.234	0.076	0.052	0.648	0.857
SEE	0.152	0.091	0.239	0.072	0.053	0.641	0.815
CP	0.948	0.946	0.948	0.924	0.924	0.951	0.921
MSE	0.023	0.008	0.055	0.006	0.003	0.426	0.743
				ML			
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	0.010	−0.001	−0.009	−0.005	−0.009	−0.040	0.068
SE	0.141	0.081	0.221	0.072	0.051	0.378	0.467
SEE	0.140	0.080	0.233	0.069	0.051	0.380	0.461
CP	0.948	0.946	0.952	0.934	0.926	0.932	0.938
MSE	0.020	0.007	0.049	0.005	0.003	0.144	0.223
				MML			
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	−0.121	−0.086	0.026	−0.001	0.041	0.833	−0.390
SE	0.149	0.084	0.234	0.074	0.056	0.305	0.372
SEE	0.146	0.083	0.240	0.071	0.056	0.297	0.367
CP	0.876	0.834	0.962	0.932	0.908	0.240	0.786
MSE	0.037	0.014	0.055	0.005	0.005	0.786	0.290
				TRL			
	β_1	β_2	β_3	β_4	γ		
BIAS	−0.160	−0.037	0.109	0.0551	0.036		
SE	0.158	0.099	0.249	0.079	0.061		
SEE	0.154	0.098	0.251	0.075	0.063		
CP	0.830	0.918	0.926	0.878	0.930		
MSE	0.051	0.011	0.074	0.009	0.005		
				ZS			
	β_1	β_2	β_3	β_4	γ		
BIAS	−0.011	−0.007	0.009	0.004	−0.008		
SE	0.232	0.112	0.357	0.162	0.091		
SEE	0.218	0.119	0.346	0.163	0.098		
CP	0.932	0.956	0.944	0.964	0.962		
MSE	0.054	0.013	0.127	0.026	0.008		
				CC			
	β_1	β_2	β_3	β_4	γ		
BIAS	0.238	0.173	−0.035	−0.066	−0.040		
SE	0.145	0.080	0.230	0.073	0.047		
SEE	0.139	0.079	0.232	0.069	0.046		
CP	0.608	0.392	0.942	0.820	0.832		
MSE	0.078	0.037	0.054	0.010	0.004		

$(\text{Dummy}_1, \text{Dummy}_2, \text{Dummy}_3)$. The regression equation is

$$Y = \beta_1 + \beta_2\text{Gender} + \beta_4\text{Race} + \beta_4\text{Mother's Income} + \beta_5\text{Mother's education} + \beta_6\text{Dummy}_1 + \beta_7\text{Dummy}_2 + \beta_8\text{Dummy}_3 + \epsilon, \quad \epsilon \sim N(0, \gamma^2).$$

The complete case method, CC, discards subjects with missing outcomes. In this case, analysis is based only on the 387 observations with both valid 1986 and 1992 PPVT scores. The method of Tang, Little, and Raghunathan (2003, TLR) assumes nonignorable missing but the propensity for missingness to depend only on the outcome but none of the covariates. The method of Zhao and Shao (2015, ZS) allows nonignorable missingness to depend on the outcome as well as the covariates but requires an instrument. The method proposed in this article makes the same assumption as ZS but does not require instrumental variables. We use (13) and (14) to estimate η, and we referred the solution as MCPL. For ZS, as suggested by Zhao and Shao (2015), binary instrumental variables are more robust. Furthermore, using binary instruments eases the computational burden. Based on the argument above that children who did poorly previously might be more likely to miss an assessment, we let the instrumental variable to be the 1986 PPVT score. Hence, we define $\mathbf{Z} = (\text{Dummy}_1, \text{Dummy}_2, \text{Dummy}_3)^T$. It is reasonable to assume that, $P(D = 1|Y, \mathbf{Z}) = P(D = 1|Y)$. In addition, scores in a previous assessment is likely to be correlated with changes in test score between 1986 and 1992. An additional hurdle for ZS is to obtain the nonparametric distribution of the covariates, conditional on the instruments. Among the covariates, 1986 PPVT score is continuous, whereas gender, race and mother's education are binary. We followed Zhao and Shao (2015, eqs. (6) and (8) therein) by using a combination of kernel density and discrete approximation to handle the discrete and continuous components of the conditional distribution. For all methods, we used 300 nonparametric bootstrap samples to estimate the SE of the estimates.

Table 8. Estimation results under scenario S_4.

	MPCL						
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	0.012	0.000	−0.016	−0.004	−0.010	0.003	0.120
SE	0.154	0.073	0.236	0.076	0.053	0.607	1.110
SEE	0.154	0.074	0.241	0.075	0.054	0.599	1.076
CP	0.946	0.948	0.950	0.934	0.944	0.951	0.989
MSE	0.024	0.005	0.056	0.006	0.003	0.367	1.243
	ML						
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	0.009	−0.002	−0.012	−0.002	−0.009	0.001	0.024
SE	0.141	0.067	0.226	0.068	0.050	0.523	0.475
SEE	0.141	0.067	0.229	0.068	0.051	0.507	0.461
CP	0.944	0.942	0.954	0.946	0.932	0.934	0.936
MSE	0.020	0.005	0.051	0.005	0.003	0.273	0.226
	MML						
	β_1	β_2	β_3	β_4	γ	η_1	η_2
BIAS	−0.109	−0.024	0.023	0.015	0.061	−0.230	0.092
SE	0.149	0.071	0.235	0.070	0.059	0.487	0.453
SEE	0.148	0.071	0.237	0.071	0.059	0.454	0.419
CP	0.886	0.916	0.952	0.944	0.852	0.922	0.910
MSE	0.034	0.006	0.056	0.005	0.007	0.289	0.213
	TRL						
	β_1	β_2	β_3	β_4	γ		
BIAS	0.021	0.005	−0.026	−0.007	−0.012		
SE	0.151	0.072	0.239	0.076	0.052		
SEE	0.151	0.071	0.241	0.074	0.053		
CP	0.942	0.936	0.938	0.934	0.946		
MSE	0.023	0.005	0.058	0.006	0.003		
	ZS						
	β_1	β_2	β_3	β_4	γ		
BIAS	−0.008	−0.006	0.002	0.009	−0.006		
SE	0.231	0.093	0.356	0.160	0.090		
SEE	0.219	0.100	0.340	0.162	0.097		
CP	0.930	0.964	0.934	0.950	0.970		
MSE	0.053	0.009	0.127	0.026	0.008		
	CC						
	β_1	β_2	β_3	β_4	γ		
BIAS	0.312	0.087	−0.207	−0.095	−0.056		
SE	0.139	0.066	0.224	0.068	0.044		
SEE	0.136	0.066	0.225	0.067	0.044		
CP	0.374	0.758	0.854	0.708	0.736		
MSE	0.117	0.012	0.093	0.014	0.005		

We use the CC estimates as initial values for the other methods. The results of the analysis are given in Table 9. The following trends emerge from the analysis. The estimates of the regression coefficients using CC, TLR, MCPL are all in the same direction. The results for ZS are drastically different from the other methods. Its estimates are very different in magnitudes. Furthermore, the SEs are very large. This is an indication of the curse of dimensionality problem discussed earlier.

We now discuss the results for CC, TLR, and MCPL. The estimates for race, mother's income, and mother's education point to a higher improvement in PPVT scores between 1986 and 1992, for children who are white, whose mothers has a higher income and whose mothers have >12 years of education. The results for the last two factors seem to explain themselves quite easily. Those whose mothers are more highly educated, with higher income, provide better support and possibly motivation for their children to improve. Blau and Grossberg (1992) also suggested that mothers with higher income could allocate a greater proportion of their time with their children on developmental activities. Many studies, for example, Champion et al. (2003) and references therein, suggested that children from minority background tend to do poorly in tests such as PPVT. These authors argued that children from minority or linguistically diverse groups may not have experience with or exposure to words that educators in schools expect them to know. If this is true, then there would be less opportunity for improvement over time as compared to whites. The results show that girls improved more over time than boys. This observation is consistent with many studies that showed a gender bias in favor of females in language tests (e.g., Chiu and McBride-Chang 2006, and references therein). The estimates for Dummy_1, Dummy_2, and Dummy_3 are all negative for all four methods, and importantly, in increasing (negative) magnitude. These dummy variables show comparison between the lowest quartile ($\text{Dummy}_1 = \text{Dummy}_2 = \text{Dummy}_3 = 0$) and the remaining quartiles, suggesting the higher the PPVT score in 1986, the lower the improvement, other factors being considered. This may be explained by that each child has a ceiling on his/her ability and there is more room for improvement for those who initially fared poorly.

As pointed out earlier, the results for ZS are quite different from the rest. In addition, its SEs are quite a bit higher than the rest. The poor results of ZS (as measured by the size of its SEs) can be attributed to two possible factors. First, ZS requires more models than other methods, such as MCPL, which only needs specification of the conditional distribution of the outcome. Furthermore, ZS also needs to specify the conditional distribution of the covariates that appear in the propensity score.

Table 9. Regression analysis of PPVT from the NLSY79 Children Survey.

	CC		ZS		TLR		MPCL	
Parameter	Estimate	SE[a]	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	78.8	1.83	76.4	15.2	79	2.03	79.8	2.19
Gender	1.23	1.18	3.75	4.43	0.35	1.45	0.797	1.55
Race	2.4	1.13	1.77	12.4	2.32	1.21	3.16	1.36
Mother's Income	2.45	1.24	−12.6	17	2.6	1.36	2.82	1.45
Mother's education	1.71	1.54	6.5	19.5	1.96	1.63	2.33	1.78
Dummy_1	−4.59	1.73	4.72	14.6	−3.58	1.76	−4.19	1.86
Dummy_2	−11.2	1.79	−15.6	12.1	−10.6	1.87	−11.6	2.09
Dummy_3	−18	1.74	−20	16.6	−18.5	1.95	−20.4	2.61
γ	11.5	0.57	16.1	8.3	11.7	0.80	12.5	0.98

[a]Based on 300 bootstrap samples.

That distribution is complex when the covariates are of mixed categorical and continuous types, such as the data considered here.

Comparing CC to TLR and MCPL, the coefficients for CC are all slightly attenuated, which supports the hypothesis that there may be some nonrandom missingness. In particular, if children who did poorly have a higher propensity for missing a future assessment, then the remaining subsample of complete cases would be more similar than reality. Once adjustment for nonrandom missingness has been made, the difference between the 1986 low and high scorers is restored, leading to higher values in the parameter estimates.

6. Concluding Remarks

This article proposes a new method that overcomes the well-known limitation of Tang, Little, and Raghunathan's (2003) method by allowing the propensity score to depend on the outcome as well as the covariates. Compared to the method of Zhao and Shao (2015), our method is more efficient since our likelihood includes information on the relationship between the outcome and the covariates affecting missingness. In the literature, profile likelihood is only known to work in a limited number of semiparametric models, such as Vardi's (1985) biased sampling models and Cox's (1972) regression models. In fact, even if a full likelihood is available, viz.,

$$\prod_{i=1}^{n}\left[\pi(Y_i+\mathbf{U}_i^T\boldsymbol{\eta})f(Y_i|\mathbf{X}_i,\boldsymbol{\Theta})\right]^{R_i}\left[\int\{1-\pi(Y+\mathbf{U}_i^T\boldsymbol{\eta})\}f(Y_i|\mathbf{X}_i,\boldsymbol{\Theta})dY\right]^{1-R_i}$$

profiling in the manner of the current article is not possible, since the support points $\pi(\cdot)$ are unknown. By working with a pseudo-likelihood based only on the complete data, we successfully converted this problem to a biased sampling problem.

The proposed method uses an iterative optimization algorithm. It is difficult to guarantee the objective functions are convex/concave. For nonconvex optimization problems, the choice of initial values is an important factor for the speed of convergence. Common choices of initial values include: (1) random starting point (generated from some random distributions which are defined on the parameter space); (2) the "best" among different random starting points (for each set of initial values, the objective functions are calculated and the set with the optimal objective function value is considered as the "best" and used as the starting point); (3) estimates from other estimators. It is also possible to develop other methods according to the characteristics of the data and the objective function. For the simulations and empirical application, we tried to maximize the pseudo-likelihood using different initial values. The answers were similar in all cases. So we used the complete case estimates as initial values. A similar strategy was employed by Zhao and Shao (2015) and Tang, Little, and Raghunathan (2003). For small sample size, we do not exclude the possibility that some initial values may lead to local solutions. If this happens, we suggest using multiple initial values and then choosing the one with the largest pseudo-likelihood as the solution. In our experience, this method works well.

Supplementary Materials

The proofs of the asymptotic properties are given in the Supplementary Materials.

Acknowledgments

The authors would like to thank an associate editor and two anonymous referees for comments and suggestions, that have led to a much improved article.

Funding

Chen's work was supported by the National Natural Science Foundation of China (NSFC) (11871402, 11931014) and the Fundamental Research Funds for the Central Universities (JBK1806002).

References

Arpino, B., De Cao, E., and Peracchi, F. (2014), "Using Panel Data for Partial Identification of Human Immunodeficiency Virus Prevalence When Infection Status Is Missing Not at Random," *Journal of the Royal Statistical Society*, Series A, 177, 587–606. [1]

Becker, G. (1981), *A Treatise on the Family*, Cambridge, MA: Harvard University Press. [8]

Besag, J. (1975), "Statistical Analysis of Non-Lattice Data," *Journal of the Royal Statistical Society*, Series D, 24, 179–195. [2]

Blau, F., and Grossberg, A. (1992), "Maternal Labor Supply and Children's Cognitive Development," *The Review of Economics and Statistics*, 74, 474–481. [8,11]

Bollinger, C. R., and Hirsch, B. T. (2013), "Is Earnings Nonresponse Ignorable?," *The Review of Economics and Statistics*, 95, 407–416. [1]

Bollinger, C. R., Hirsch, B. T., Hokayem, C. M., and Ziliak, J. P. (2019), "Trouble in the Tails? What We Know About Earnings Nonresponse 30 Years After Lillard, Smith, and Welch," *Journal of Political Economy*, 127, 2143–2185. [1]

Breunig, C. (2017), "Testing Missing at Random Using Instrumental Variables," SFB 649 Discussion Papers SFB649DP2017-007, Humboldt University, Berlin, Germany. [1]

Champion, T., Hyter, Y., Mccabe, A., and Bland-Stewart, L. (2003), "'A Matter of Vocabulary': Performances of Low-Income African American Head Start Children on the Peabody Picture Vocabulary Test-III," *Communication Disorders Quarterly*, 24, 121–127. [11]

Chen, H., Geng, Z., and Zhou, X.-H. (2009), "Identifiability and Estimation of Causal Effects in Randomized Trials With Noncompliance and Completely Nonignorable Missing Data," *Biometrics*, 65, 675–682. [2]

Chen, K. (2001), "Parametric Models for Response-Biased Sampling," *Journal of the Royal Statistical Society*, Series B, 63, 775–789. [2]

Chiu, M. M., and McBride-Chang, C. (2006), "Gender, Context, and Reading: A Comparison of Students in 43 Countries," *Scientific Studies of Reading*, 10, 331–362. [11]

Cosslett, S. R. (1981), "Maximum Likelihood Estimator for Choice-Based Samples," *Econometrica*, 49, 1289–1316. [2]

Cox, D. R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society (Series B)*, 34, 187–220. [12]

Davey, A., Shanahan, M. J., and Schafer, J. L. (2001), "Correcting for Selective Nonresponse in the National Longitudinal Survey of Youth Using Multiple Imputation," *The Journal of Human Resources*, 36, 500–519. [1]

Dunn, L., and Dunn, D. (2007), *The Peabody Picture Vocabulary Test, PPVT-4* (4th ed.), Minneapolis, MN: Pearson. [8]

Golsteyn, B. H. H., and Hirsch, S. (2019), "Are Estimates of Intergenerational Mobility Biased by Non-Response? Evidence From the Netherlands," *Social Choice and Welfare*, 52, 29–63. [1]

Greenlees, J. S., Reece, W. S., and Zieschang, K. Y. (1982), "Imputation of Missing Values When the Probability of Response Depends on the Variable Being Imputed," *Journal of the American Statistical Association*, 77, 251–261. [2,3,5]

Ichimura, H. (1988), "Estimation of Single Index Models," PhD thesis, Department of Economics, Massachusetts Institute of Technology. [3]

——— (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58, 71–120. [3]

Kim, J. K., and Yu, C. Y. (2011), "A Semi-Parametric Estimation of Mean Functionals With Non-Ignorable Missing Data," *Journal of the American Statistical Association*, 106, 157–165. [2]

Kline, P., and Santos, A. (2013), "Sensitivity to Missing Data Assumptions: Theory and an Evaluation of the U.S. Wage Structure," *Quantitative Economics*, 4, 231–267. [1]

Kott, P. S., and Chang, T. (2010), "Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse," *Journal of the American Statistical Association*, 105, 1265–1275. [2]

Lemieux, T. (2006), "Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?," *American Economic Review*, 96, 461–498. [1]

Liang, K.-Y., and Qin, J. (2000), "Regression Analysis Under Non-Standard Situations: A Pairwise Pseudolikelihood Approach," *Journal of the Royal Statistical Society*, Series B, 62, 773–786. [2]

Lillard, L., Smith, J. P., and Welch, F. (1986), "What Do We Really Know About Wages? The Importance of Nonreporting and Census Imputation," *Journal of Political Economy*, 94, 489–506. [1,2]

Little, R. J. A. (1985), "Nonresponse Adjustments in Longitudinal Surveys: Models for Categorical Data," *Bulletin of the International Statistical Institute*, 15, 1–15. [3]

——— (1993), "Pattern-Mixture Models for Multivariate Incomplete Data," *Journal of the American Statistical Association*, 88, 125–134. [2]

——— (1994), "A Class of Pattern-Mixture Models for Normal Incomplete Data," *Biometrika*, 81, 471–483. [2,3]

Little, R. J. A., and Wang, Y. (1996), "Pattern-Mixture Models for Multivariate Incomplete Data With Covariates," *Biometrics*, 52, 98–111. [2]

Maasoumi, E., and Wang, L. (2017), "What Can We Learn About the Racial Gap in the Presence of Sample Selection?," *Journal of Econometrics*, 199, 117–130. [1]

——— (2019), "The Gender Gap Between Earnings Distributions," *Journal of Political Economy*, 127, 2438–2504. [1]

Manski, C. F. (1989), "Anatomy of the Selection Problem," *Journal of Human Resources*, 24, 343–360. [2]

——— (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 80, 319–323. [2]

Manski, C. F., and Pepper, J. V. (2000), "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010. [2]

Matzkin, R. L. (1992), "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models," *Econometrica*, 60, 239–270. [2]

Miao, W., Liu, L., Tchetgen, E., and Geng, Z. (2019), "Identification, Doubly Robust Estimation, and Semiparametric Efficiency Theory of Nonignorable Missing Data With a Shadow Variable," arXiv no. 1509.02556v3. [6]

Morikawa, K., and Kim, J. (2019), "Semiparametric Optimal Estimation With Nonignorable Nonresponse Data," arXiv no. 1612.09207v2. [6]

Mulligan, C., and Rubinstein, Y. (2008), "Selection, Investment, and Women's Relative Wages Over Time," *The Quarterly Journal of Economics*, 123, 1061–1110. [1]

Olsen, M. K., and Schafer, J. L. (2001), "A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data," *Journal of the American Statistical Association*, 96, 730–745. [7]

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430. [3]

Qin, D., van Huellen, S., Elshafie, R., Liu, Y., and Moraitis, T. (2019), "A Principled Approach to Assessing Missing-Wage Induced Selection Bias," Working Papers Series No. 216, SOAS Department of Economics. [1]

Qin, J. (2017), *Biased Sampling, Over-Identified Parameter Problems and Beyond*, Singapore: Springer-Verlag. [4]

Ramalho, E. A., and Smith, R. J. (2013), "Discrete Choice Non-Response," *The Review of Economic Studies*, 80, 343–364. [2]

Robins, J. M., and Ritov, Y. (1997), "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-Parametric Models," *Statistics in Medicine*, 16, 285–319. [2,3]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [1]

Rotnitzky, A., and Robin, J. M. (1997), "Analysis of Semi-Parametric Regression Models With Non-Ignorable Non-Response," *Statistics in Medicine*, 16, 81–102. [6]

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592. [1]

——— (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley. [1]

Rueda, S., Mitra, S., Chen, S., Gogolishvili, D., Globerman, J., Chambers, L., Wilson, M., Logie, C. H., Shi, Q., Morassaei, S., and Rourke, S. B. (2016), "Examining the Associations Between HIV-Related Stigma and Health Outcomes in People Living With HIV/AIDS: A Series of Meta-Analyses," *BMJ Open*, 6, e011453. [2]

Schräpler, J.-P. (2004), "Respondent Behavior in Panel Studies: A Case Study for Income Nonresponse by Means of the German Socio-Economic Panel (SOEP), *Sociological Methods & Research*, 33, 118–156. [1]

Shao, J., and Wang, L. (2016), "Semiparametric Inverse Propensity Weighting for Nonignorable Missing Data," *Biometrika*, 103, 175–187. [2]

Stasny, E. A. (1985), "Modeling Nonignorable Non-Response in Panel Data," in *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 349–354. [2]

Tang, G., Little, R. J. A., and Raghunathan, T. E. (2003), "Analysis of Multivariate Missing Data With Nonignorable Nonresponse," *Biometrika*, 90, 747–764. [2,3,5,6,7,10,12]

Vardi, Y. (1985), "Empirical Distribution in Selection Bias Models," *The Annals of Statistics*, 13, 178–203. [12]

Wang, S., Shao, J., and Kim, J. K. (2014), "An Instrumental Variable Approach for Identification and Estimation With Nonignorable Nonresponse," *Statistica Sinica*, 24, 1097–1116. [2,3]

Yang, F., Lorch, S. A., and Small, D. S. (2014), "Estimation of Causal Effects Using Instrumental Variables With Nonignorable Missing Covariates: Application to Effect of Type of Delivery NICU on Premature Infants," *The Annals of Applied Statistics*, 8, 48–73. [2]

Zhao, H., Zhao, P., and Tang, N. (2013), "Empirical Likelihood Inference for Mean Functionals With Nonignorably Missing Response Data," *Computational Statistiscs and Data Analysis*, 66, 101–116. [2]

Zhao, J., and Shao, J. (2015), "Semiparametric Pseudo-Likelihoods in Generalized Linear Models With Nonignorable Missing Data," *Journal of the American Statistical Association*, 110, 1577–1590. [2,3,4,6,10,12]