Semi-Parametric Methods for Missing Data and Causal Inference

Semi-Parametric Methods for Missing Data and Causal Inference

A dissertation presented

by

BaoLuo Sun

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

May 2016

Dissertation Advisor: Professor Eric Joel Tchetgen Tchetgen BaoLuo Sun

Semi-Parametric Methods for Missing Data and Causal Inference

Abstract

In this dissertation, we propose methodology to account for missing data as well as a strategy to account for outcome heterogeneity.

Missing data occurs frequently in empirical studies in health and social sciences, often compromising our ability to make accurate inferences. An outcome is said to be missing not at random (MNAR) if, conditional on the observed variables, the missing data mechanism still depends on the unobserved outcome. In such settings, identification is generally not possible without imposing additional assumptions. Identification is sometimes possible, however, if an exogeneous instrumental variable (IV) is observed for all subjects such that it satisfies the exclusion restriction that the IV affects the missingness process without directly influencing the outcome. In chapter 1, we provide necessary and sufficient conditions for nonparametric identification of the full data distribution under MNAR with the aid of an IV. In addition, we give sufficient identification conditions that are more straightforward to verify in practice. For inference, we focus on estimation of a population outcome mean, for which we develop a suite of semiparametric estimators that extend methods previously developed for data missing at random. Specifically, we propose inverse probability weighted estimation, outcome regression based estimation and doubly robust estimation of the mean of an outcome subject to MNAR. For illustration, the methods are used to account for selection bias induced by HIV testing refusal in the evaluation of HIV seroprevalence in Mochudi, Botswana, using interviewer characteristics such as gender, age and years of experience as IVs.

The development of coherent missing data models to account for nonmonotone missing at random (MAR) data by inverse probability weighting (IPW) remains to date largely

unresolved. As a consequence, IPW has essentially been restricted for use only in monotone MAR settings. In chapter 2, we propose a class of models for nonmonotone missing data mechanisms that spans the MAR model, while allowing the underlying full data law to remain unrestricted. For parametric specifications within the proposed class, we introduce an unconstrained maximum likelihood estimator for estimating the missing data probabilities which is easily implemented using existing software. To circumvent potential convergence issues with this procedure, we also introduce a constrained Bayesian approach to estimate the missing data process which is guaranteed to yield inferences that respect all model restrictions. The efficiency of standard IPW estimation is improved by incorporating information from incomplete cases through an augmented estimating equation which is optimal within a large class of estimating equations. We investigate the finite-sample properties of the proposed estimators in extensive simulations and illustrate the new methodology in an application evaluating key correlates of preterm delivery for infants born to HIV infected mothers in Botswana, Africa.

When a risk factor affects certain categories of a multinomial outcome but not others, outcome heterogeneity is said to be present. A standard epidemiologic approach for modeling risk factors of a categorical outcome typically entails fitting a polytomous logistic regression via maximum likelihood estimation. In chapter 3, we show that standard polytomous regression is ill-equipped to detect outcome heterogeneity, and will generally understate the degree to which such heterogeneity may be present. Specifically, nonsaturated polytomous regression will often a priori rule out the possibility of outcome heterogeneity from its parameter space. As a remedy, we propose to model each category of the outcome as a separate binary regression. For full efficiency, we propose to estimate the collection of regression parameters jointly by a constrained Bayesian approach which ensures that one remains within the multinomial model. The approach is straightforward to implement in standard software for Bayesian estimation.

Contents

Acknowledgments

I am grateful to my thesis advisor, Eric Tchetgen Tchetgen, for his guidance, patience and encouragement. I would also like to thank my dissertation committee members, James Robins and Sebastien Haneuse for their invaluable insights and advice. Their brilliance and mentorships continue to inspire me as I embark on the long research journey ahead.

I cherish the friendships of many professors and schoolmates whom I have met and worked with over the years, especially in the departments of Biostatistics and Epidemiology, Harvard T. H. Chan School of Public Health. I thank Boston for her culture of learning, beautiful sceneries and the occasional harsh winters which allowed me to finish this thesis, as well as Singapore's Agency for Science, Technology and Research for supporting my studies.

This dissertation is dedicated to my family.

Instrumental Variable Identification and Doubly Robust Estimation for Outcome Missing Not at Random

BaoLuo Sun[1], Lan Liu[1], Wang Miao[1,4], Kathleen Wirth[2,3], James Robins[1,2] and Eric J. Tchetgen Tchetgen[1,2]

Departments of Biostatistics[1], Epidemiology[2] and Immunology and Infectious Diseases[3], Harvard T.H. Chan School of Public Health. Beijing International Center for Mathematical Research[4], Peking University.

1.1 Introduction

Selection bias is a major problem in health and social sciences, and is said to be present if, in an empirical study, features of the underlying population of primary interest are entangled with features of the selection process not of scientific interest. Selection bias can occur in practice due to incomplete data, if the observed sample is not representative of the true underlying population. While various ad hoc methods exist to adjust for missing data, such methods may be subject to bias unless under fairly strong assumptions. For example, complete-case analysis is easy to implement and is routinely used in practice. However, complete-case analysis is well-known to produce biased estimates when the outcome is not missing completely at random (MCAR) (Little and Rubin, 2002). Progress can still be made if data are missing at random (MAR), such that the missing data mechanism is independent of unobserved variables conditional on observed data. Principled methods for handling MAR data abound, including likelihood-based procedures (Little and Rubin, 2002; Horton and Laird, 1998), multiple imputation (Rubin, 1987; Kenward and Carpenter, 2007a; Horton and Lipsitz, 2001; Schafer, 1999), inverse probability weighting (Robins et al., 1994; Tsiatis, 2006; Li et al., 2013) and doubly robust estimation (Scharfstein et al., 1999; Lipsitz et al., 1999; Robins et al., 2000; Robins and Rotnitzky, 2001; Neugebauer and van der Laan, 2005; Tsiatis, 2006; Tchetgen Tchetgen, 2009).

The MAR assumption is strictly not testable in a nonparametric model without an additional assumption (Gill et al., 1997; Potthoff et al., 2006) and is often untenable. An outcome is said to be missing not at random (MNAR) if, conditional on the observed data, the missingness process remains dependent on the unobserved outcome (Little and Rubin, 2002). Identification is generally no longer available under MNAR without an additional assumption (Robins and Ritov, 1997). A possible approach is to make sufficient parametric assumptions (Little and Rubin, 2002; Roy, 2003; Wu and Carroll, 1988) about the full data distribution for identification. However, this approach can fail even with commonly used fully parametric models (Miao et al., 2014; Wang et al., 2014). Other existing strategies for MNAR include positing sufficiently stringent modeling restrictions on a model for the missing data process (Rotnitzky et al., 1998) or obtaining sensitivity anal-

ysis and bounds (Moreno-Betancur and Chavance, 2013; Kenward and Carpenter, 2007b; Robins et al., 2000; Vansteelandt et al., 2007). Another common identification approach under MNAR involves leveraging an instrumental variable (IV) (Manski, 1985; Winship and Mare, 1992). Heckman's framework (Heckman, 1979, 1997) is perhaps the most common IV approach used primarily in economics and other social sciences to account for outcome MNAR. A valid IV is known to satisfy the following conditions:

(i) the IV is not directly related to the outcome in the underlying population, conditional on a set of fully observed covariates, and

(ii) the IV is associated with the missingness mechanism conditional on the fully observed covariates.

Therefore a valid IV must predict a person's propensity to have an observed outcome, without directly influencing the outcome.

One can in principle use a valid IV to obtain a nonparametric test of the MAR assumption. However access to an IV does not point identify the joint distribution of the full data and therefore its functionals. Heckman's selection model (Heckman, 1979) is generally not known to be identifiable without an assumption of bivariate normal latent error in defining the model (Wooldridge, 2010). Estimation using Heckman-type selection models may be sensitive to these parametric assumptions (Winship and Mare, 1992; Puhani, 2000), although there has been significant work towards relaxing the assumptions (Manski, 1985; Newey et al., 1990; Das et al., 2003; Newey, 2009). An alternative sufficient identification condition was considered by Tchetgen Tchetgen and Wirth (2013) which involves restricting the functional form of the selection bias function due to non-response on a given scale for the outcome (mean additive, mean multiplicative or logistic) under the specified model. However, as shown in simulation studies below, their approach is likely sensitive to bias due to model misspecification, and a more robust approach is warranted.

In this paper, we develop a general framework for nonparametric identification of selection models based on an IV. We describe necessary and sufficient conditions for identifiability of the full data distribution with a valid IV. For inference we focus on estimation of an outcome mean, although the proposed methods are easy to adapt to other

functionals. We develop three semiparametric approaches that extend analogous methods previously developed under missing at random (MAR) settings: inverse probability weighting (IPW), outcome regression (OR) and doubly robust (DR) estimation. The consistency of each estimator relies on correctly specified models for parts of the joint full data law. Extensive simulation studies are used to investigate the finite sample properties of the proposed estimators. For illustration, the methods are used to account for selection bias induced by HIV testing refusal in the evaluation of HIV seroprevalence in Mochudi, Botswana, using interviewer characteristics including gender, age and years of experience as IVs. All proofs are delegated to an appendix.

1.2 Notation and Assumptions

Suppose that one has observed n independent and identically distributed observations (X, RY, R, Z) with fully observed covariates X and R is the indicator of whether the person's outcome Y is observed. The variable Z is a fully observed IV that satisfies assumptions (i) and (ii) formalized below. In the evaluation of HIV prevalence in Mochudi, X includes all demographic and behavioral variables collected for all persons in the sample, while HIV status Y may be missing for individuals who failed to be tested, i.e. with $R = 0$. Let $\tilde{\pi}(X, Z) = \Pr(R = 1|X, Z)$ denote the propensity score for the missingness mechanism given (X, Z). As a valid IV, we will assume that Z satisfies the following assumptions.

Assumption 1.

(IV.1) Exclusion restriction:

$$P_{Y|X,Z}(y|x, z) = P_{Y|X}(y|x) \quad \text{for all } x, z.$$

(IV.2) IV relevance:

$$\tilde{\pi}(x, z) - \tilde{\pi}(x, z') \neq 0 \quad \text{for all } x.$$

Exclusion restriction (IV.1) states that the IV and the outcome are conditionally independent given X in the underlying population, that is the IV does not have a direct effect

on the outcome, which places restrictions on the full data law for identification. IV relevance requires that the IV remains associated with the missingness mechanism even after conditioning on X, which allows for full rank conditions in estimation. In spite of (IV.2), (IV.1) implies that Z cannot reduce the dependence between R and Y, therefore under MNAR $\pi(x, y, z) = P(R = 1|x, y, z)$ remains a function of y even after conditioning on (x, z). In addition, (IV.1) and (IV.2) implies that under MNAR the IV remains relevant in $\pi(x, y, z)$ conditional on (x, y). Both of these facts will be used repeatedly throughout. $\tilde{\pi}(x, z)$ is typically referred to as the propensity score for the missingness process, and we shall likewise refer to $\pi(x, y, z)$ as the extended propensity score.

1.3 Identification

Although (IV.1) reduces the number of unknown parameters in the full data law, identification is still only available for a subset of all possible full data laws. As an illustration, consider the case of binary outcome and IV. For simplicity and without loss of generality, we omit covariates X. Assumption (IV.1) implies $P(z, y) = P(y)P(z)$. We are only able to identify the quantities $P(z, y|R = 1)$, $P(z|R = 0)$, $P(R = 1)$ from the observed data. These quantities are functions of the unknown parameters: $P(Z = 1)$, $P(Y = 1)$, and $P(R = 1|z, y)$. So we have six unknown parameters, but only five available independent equations, one for each empirically identified parameter given above. As a result, the full data law is not identifiable, and $P(Y = 1)$ is not identifiable.

The IV model becomes identifiable once one sufficiently restricts the class of models for the joint distribution of (Z, Y, R). Let $\mathcal{P}_\theta(R, Z, Y)$, $\mathcal{P}_\eta(Z)$ and $\mathcal{P}_\xi(Y)$ denote the collection of such candidates for $P(R = 1|z, y)$, $P(z)$ and $P(y)$, respectively. Members of the sets are indexed by parameters θ, η and ξ, which may be infinite dimensional. The identifiability of the model is determined by the relationship between its members. Proofs for all results and examples are relegated to an appendix.

Theorem 1. Suppose that Assumption 1 holds, then the joint distribution $P(z, y, r)$ is identifiable if and only if $\mathcal{P}_\theta(R, Z, Y)$ and $\mathcal{P}_\xi(Y)$ satisfy the following condition: for any pair of candidates in the model $\{P_{\theta_1}(R = 1|z, y), P_{\xi_1}(y)\}$ and $\{P_{\theta_2}(R = 1|z, y), P_{\xi_2}(y)\}$, their

ratios are not equal:

$$\frac{P_{\theta_1}(R=1|z,y)}{P_{\theta_2}(R=1|z,y)} \neq \frac{P_{\xi_2}(y)}{P_{\xi_1}(y)} \tag{1.1}$$

for at least one value of z and y.

Theorem 1 presents a necessary and sufficient condition for identifiability of the joint distribution of the full data, and thus a sufficient condition for identifiability of its functionals. Although condition (1.1) of Theorem 1 can be readily verified for parametric distributions, it may not be so for semiparametric and nonparametric distributions with less tractable forms for $\mathcal{P}_\theta(R,Z,Y)$ and $\mathcal{P}_\xi(Y)$. We have the following corollary which provides a more convenient condition to verify.

Corollary 1. Suppose that Assumption 1 holds, then the joint distribution $P(z,y,r)$ is identifiable if the ratio $P_{\theta_1}(R=1|z,y)/P_{\theta_2}(R=1|z,y)$ is either a constant or varies with z for any two elements $P_{\theta_1}(R,Z,Y)$ and $P_{\theta_2}(R,Z,Y)$ of the model.

Although Corollary 1 provides a sufficient condition for identification of the joint distribution of the full data, it is still possible to characterize the identifiability of a large class of parametric or semi-parametric models by verifying the condition in the corollary, which we illustrate with several examples.

Example 1. For binary outcome with binary instrument, consider the candidate set

$$\mathcal{P}_\theta(R,Z,Y) = \left\{P(R=1|Z,Y) = \text{expit}\left[\theta_0+\theta_1 Z+\theta_2 Y+\theta_3 ZY\right] : (\theta_0,\theta_1,\theta_2,\theta_3)\in\mathbb{R}^4\right\},$$

which are saturated in (Z,Y). It can be shown that candidates from this set do not satisfy inequality (1.1) in Theorem 1 and therefore the joint distribution of (Z,Y,R) cannot be identified without reducing the dimension of θ through modeling assumptions. By Corollary 1, the joint distribution can be identified for the candidate set

$$\mathcal{P}_\theta(R,Z,Y) = \left\{P(R=1|Z,Y) = \text{expit}\left[\theta_0+\theta_1 Z+\theta_2 Y\right] : (\theta_0,\theta_1,\theta_2)\in\mathbb{R}^3\right\},$$

i.e. with the additional assumption that the association of the outcome Y with the missingness mechanism is constant within levels of Z, on the logit scale. A more general result on the identifiability of separable logistic missing data mechanisms, which also holds for continuous Y and Z, is given in Example 2.

Example 2. The separable logistic missing data mechanism

$$\mathcal{P}_\theta(R, Z, Y) = \{P(R = 1|Z, Y) = \text{expit}[q(Z) + h(Y)]\}, \tag{1.2}$$

is identifiable, where $q(\cdot)$ and $h(\cdot)$ are unknown functions differentiable with respect to Z and Y respectively. Identification also holds if either or both of Z and Y are binary or discrete random variables.

Example 3. The separable probit missing data data mechanism

$$\mathcal{P}_\theta(R, Z, Y) = \{P(R = 1|Z, Y) = \Phi[q(Z) + h(Y)]\}, \tag{1.3}$$

is identifiable, where $q(\cdot)$ and $h(\cdot)$ are unknown functions assumed to be differentiable with respect to Z and Y respectively.

Example 4. Under MAR, the missing data mechanism

$$\mathcal{P}_\theta(R, Z, Y) = \{P(R = 1|Z, Y) = g(Z)\} \tag{1.4}$$

is identifiable. Its members satisfy the conditions of Corollary 1. It is clear that the ratio of any pair of members is either a constant or varies with Z.

1.4 Estimation and Inference

In this section, we consider estimation and inference under a variety of semiparametric IV models that are assumed to satisfy the identifiability conditions of Theorem 1. We denote the collection of such identifiable models as $\mathcal{M}_{\text{IV}}$. As a measure of departure from MAR, we introduce the selection bias function

$$\eta(x, y, z) = \log \left\{ \frac{P(R = 1|x, y, z)}{P(R = 0|x, y, z)} \Big/ \frac{P(R = 1|x, Y = 0, z)}{P(R = 0|x, Y = 0, z)} \right\}. \tag{1.5}$$

η quantifies the degree of association between Y and R given (X, Z) on the log odds ratio scale. Under MAR, $P(R = 1|x, y, z) = P(R = 1|x, z)$ and $\eta = 0$. Using the modeling framework proposed by Chen (2007), under mild conditions the conditional density $P(r, y|x, z)$ can be represented in terms of the selection bias function η and baseline densities as

$$P(r, y|x, z) = C(x, z)^{-1} \exp[(r - 1)\eta(x, y, z)] f(y|R = 1, x, z) P(r|Y = 0, x, z), \tag{1.6}$$

where $C(x,z) = P(R=1|Y=0,x,z) + P(R=0|Y=0,x,z)E\{\exp[-\eta(x,Y,z)]|R=1,x,z\} < +\infty$ for all (x,z) is a normalizing constant. We can therefore characterize the joint data law conditional on X as

$$P(r,y,z|x) = C(x,z)^{-1}\exp[(r-1)\eta(x,y,z)]f(y|R=1,x,z)P(r|Y=0,x,z)q(z|x). \quad (1.7)$$

By (1.6), the selection bias function η needs to be correctly specified for any of the three proposed estimators to be consistent. This is significant in that for any observed data law and each selection bias function η, one can identify a unique full data law that marginalizes to the observed data law (Scharfstein et al., 2003). Absent of restrictions such as Assumption 1, the selection bias function is not identifiable from the observed data law since different values of η can lead to the same observed data likelihood. In order to address this identification problem, an existing strategy is sensitivity analysis whereby one conducts inferences assuming $\eta(\zeta_0)$ is completely known and repeats the analysis upon varying the assumed value of ζ_0 (Robins et al., 2000; Rotnitzky et al., 1998, 2001; Scharfstein et al., 1999; Vansteelandt et al., 2007). A different approach is possible with an IV since η is in principle identified under Theorem 1 from the observed data and therefore needs not be assumed known. It is impossible to disentangle the full data law from the selection process without evaluating η. Therefore, we will proceed by assuming that although a priori unknown, one can correctly specify a model $\eta(\zeta)$ for the selection bias function which can be estimated from the observed data. To fix ideas, throughout we suppose that one aims to make inferences about the population mean $\phi = E(Y)$, although the proposed methods are easy to extend to other full data functionals.

Although identification results given in the previous section in principle allow for nonparametric inference, in practice estimation often involves specifying parametric models, at least for parts of the full data law. This will generally be the case when a large number of covariates X and Z are present and therefore the curse of dimensionality precludes the use of nonparametric regression to model the association between Y and R given (X,Y) (Robins and Ritov, 1997). IPW estimation typically requires a correctly specified model for the extended propensity score $\pi(x,y,z)$. The extended propensity score under logit

link function is

$$\pi(x, y, z) = 1/\{1 + \exp[-\eta(x, y, z) - \lambda(x, z)]\}, \tag{1.8}$$

where $\eta(x, y, z)$ is the selection bias function given in (1.5) and $\lambda(x, z) = \log\{P(R = 1|Y = 0, x, z)/P(R = 0|Y = 0, x, z)\}$ is the baseline missing data model, a function of $P(r|Y = 0, x, z)$. While IPW estimation can be applied for any proper link function for the extended propensity score, as we show below, DR estimation relies on using the logit link function to model the extended propensity score. We consider IPW estimation in the model

$$\mathcal{M}_{\text{IPW}} = \{P(r, y, z|x) : \eta(x, y, z; \zeta), P(r|Y = 0, x, z; \omega), q(z|x; \xi)\} \subset \mathcal{M}_{\text{IV}},$$

where the parametric models indexed by (ζ, ω, ξ) respectively are assumed to be correctly specified, and the baseline outcome model $f(y|R = 1, x, z)$ in (1.7) is unrestricted.

Outcome regression based estimation under MAR typically requires a model for $f(y|R = 1, x, z) = f(y|x, z)$, which can be estimated based on complete-cases. However, under MNAR $f(y|R = 1, X, Z) \neq f(y|R = 0, X, Z)$ and estimation of $f(y|R = 0, x, z)$ is difficult since outcome is not observed for this subpopulation. When a valid IV is available, the same conditional density $f(y|R = 0, x, z)$ has an equivalent exponential-tilt representation

$$f(y|r, x, z) = \frac{P(y, r|x, z)}{\int P(y, r|x, z)\mathrm{d}\mu(y)} = \frac{\exp[-(1 - r)\eta(x, y, z)]f(y|R = 1, x, z)}{E\{\exp[-(1 - r)\eta(x, Y, z)]|R = 1, x, z\}}, \tag{1.9}$$

and therefore relies on correctly specified models for the selection bias function η and baseline outcome model $f(y|R = 1, x, z)$ for complete-cases. We consider OR estimation in the model

$$\mathcal{M}_{\text{OR}} = \{P(r, y, z|x) : \eta(x, y, z; \zeta), P(y|R = 1, x, z; \theta), q(z|x; \xi)\} \subset \mathcal{M}_{\text{IV}},$$

which allows the baseline missing data model $P(r|Y = 0, x, z)$ to remain unrestricted.

We also propose a doubly robust estimator which is consistent in the union model $\mathcal{M}_{\text{IPW}} \cup \mathcal{M}_{\text{OR}}$. The DR estimator holds appeal in that it remains consistent if the conditional density $q(z|x)$ is correctly specified, and either one of the models $P(y|R = 0, X, Z; \theta)$ or $P(r|Y, X, Z; \omega)$, but not necessarily both, is correctly specified, thus rendering it more robust against model misspecifications.

Throughout the next section, we let $\hat{\theta}_{\text{MLE}}$ and $\hat{\xi}_{\text{MLE}}$ denote the maximum likelihood estimators of the parametric models $P(y|R=1,x,z;\theta)$ and $q(z|x;\xi)$ respectively, and let $\mathbb{P}_n$ denote the empirical measure $\mathbb{P}_n f(O) = n^{-1}\sum_i f(O_i)$.

1.4.1 Inverse probability weighted estimation under $\mathcal{M}_{\text{IPW}}$

IPW is a well known approach to acount for missing data under MAR. In this section we describe an analogous approach under MNAR. Standard approaches to estimating the propensity score under MAR such as maximum likelihood say of a logistic regression model of the propensity score cannot be used here since $\pi(x,y,z)$ depends on Y which is only observed when $R=1$. Therefore, we propose an alternative method of moments approach which resolves this difficulty. Within the model $\mathcal{M}_{\text{IPW}}$, $(\hat{\zeta},\hat{\omega})$ solves

$$\mathbb{P}_n\left\{\boldsymbol{U}^{\text{IPW}}\left(\hat{\xi}_{\text{MLE}},\zeta,\omega\right)\right\} = \mathbf{0} \tag{1.10}$$

where $\boldsymbol{U}^{\text{IPW}}(\cdot)$ consists of the estimating equations

$$\mathbb{P}_n\left\{\left[\frac{R}{\pi(\zeta,\omega)}-1\right]\boldsymbol{h}_1(X,Z)\right\} = \mathbf{0} \tag{1.11}$$

$$\mathbb{P}_n\left\{\frac{R}{\pi(\zeta,\omega)}\boldsymbol{g}(X,Y)\left\{\boldsymbol{h}_2(Z,X) - E\left[\boldsymbol{h}_2(Z,X)\middle|X;\hat{\xi}_{\text{MLE}}\right]\right\}\right\} = \mathbf{0}. \tag{1.12}$$

Equations (1.11) and (1.12) estimate unknown parameters in $P(r|Y=0,x,z;\omega)$ and $\eta(x,y,z;\zeta)$ respectively, where $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are arbitrary functions of (x,z) with same dimensions as ω and ζ respectively. Specific choices of $(\boldsymbol{h}_1,\boldsymbol{h}_2,\boldsymbol{g})$ can generally affect efficiency but not consistency. Optimal choices are described in the next section. To illustrate IPW estimation, suppose that Z is binary and consider the following logistic model for the extended propensity score

$$\text{logit}\,\pi(X,Y,Z) = \omega_0+\omega_1 X+\omega_2 XZ+\zeta Y, \quad \eta=(\omega_0,\omega_1,\omega_2,\zeta).$$

Thus, $\eta(x,y,z;\zeta)=\zeta y$ and $\text{logit}\,P(R=1|Y=0,x,z;\omega)=\omega_0+\omega_1 x+\omega_2 xz$. Suppose further that $q(Z=1|x;\xi)=B(x;\xi)=\left\{1+\exp\left[-(1,x)^T\xi\right]\right\}^{-1}$. We obtain $(\hat{\zeta},\hat{\omega})$ by solving

$$\mathbb{P}_n\left\{\left[\frac{R}{\pi(\zeta,\omega)}-1\right](1,X,XZ)^T\right\} = 0$$

$$\mathbb{P}_n\left\{\frac{R}{\pi(\zeta,\omega)}Y\left\{Z-B(X;\hat{\xi}_{\text{MLE}})\right\}\right\} = 0.$$

Proposition 1. Consider a model $\mathcal{M}_{\text{IPW}} \subset \mathcal{M}_{\text{IV}}$ which satisfies the identification condition in Theorem (1). Then the IPW estimator

$$\hat{\phi}^{\text{IPW}} = \mathbb{P}_n \left\{ \frac{RY}{\pi(\hat{\eta})} \right\} \tag{1.13}$$

is consistent and asymptotically normal as $n \to \infty$, that is

$$\sqrt{n} \left(\hat{\phi}^{\text{IPW}} - \phi_0 \right) \xrightarrow{d} N(0, V)$$

in model $\mathcal{M}_{\text{IPW}}$ under suitable regularity conditions, where V is given in (1.14) below.

Let $\boldsymbol{M}(\delta)$ represent the stacked vector of the following estimating functions: score functions for estimating ξ, $\boldsymbol{U}^{\text{IPW}}(\xi, \zeta, \omega)$ and $G(\phi, \zeta, \omega) = \left\{ \frac{RY}{\pi(\zeta,\omega)} - \phi \right\}$, where $\delta = (\zeta, \omega, \xi, \phi)$. Then under standard regularity conditions for M-estimation (Newey and McFadden, 1993), the asymptotic variance V is given by the diagonal entry corresponding to ϕ of the following variance-covariance matrix

$$\left[E \left\{ \frac{\partial \boldsymbol{M}(\delta)}{\partial \delta^T} \Big|_{\delta_0} \right\} \right]^{-1} E \left\{ \boldsymbol{M}(\delta_0) \boldsymbol{M}(\delta_0)^T \right\} \left[E \left\{ \frac{\partial \boldsymbol{M}(\delta)}{\partial \delta^T} \right\} \Big|_{\delta_0} \right]^{-1^T}, \tag{1.14}$$

where $\delta_0 = (\zeta_0, \omega_0, \xi_0, \phi_0)$ is the probability limit of $\hat{\delta} = (\hat{\zeta}, \hat{\omega}, \hat{\xi}, \hat{\phi})$. A consistent sandwich estimator for the above asymptotic variance can be constructed by evaluating unknown expectations as sample means at the estimated parameter value $\hat{\delta}$.

1.4.2 Outcome regression estimation under $\mathcal{M}_{\text{OR}}$

We consider inferences under a parametric model for the outcome, i.e. under model $\mathcal{M}_{\text{OR}}$. Using the parametrization given in (1.9), consider the parametric model

$$P(y|R=0, x, z; \zeta, \hat{\theta}_{\text{MLE}}) = \frac{\exp\left[-\eta(x, y, z; \zeta)\right] f\left(y|R=1, x, z; \hat{\theta}_{\text{MLE}}\right)}{E\left\{\exp[-\eta(x, Y, z; \zeta)]|R=1, x, z; \hat{\theta}_{\text{MLE}}\right\}}.$$

We arrive at the estimate $\hat{\zeta}$ of the selection bias function $\eta(\zeta)$ as the solution of the estimating equation

$$\begin{aligned}
&\mathbb{P}_n \left\{ \boldsymbol{U}^{\text{OR}} \left(\zeta, \hat{\xi}_{\text{MLE}}, \hat{\theta}_{\text{MLE}}, \boldsymbol{q}_1, \boldsymbol{q}_2 \right) \right\} = \\
&\mathbb{P}_n \left\{ \boldsymbol{q}_1(X, Z) - E \left[\boldsymbol{q}_1(X, Z) \Big| X; \hat{\xi}_{\text{MLE}} \right] \right\} \left\{ (1-R) E \left(\boldsymbol{q}_2(X, Y) \Big| R=0, X, Z; \zeta, \hat{\theta}_{\text{MLE}} \right) + R \boldsymbol{q}_2(X, Y) \right\} \\
&= \mathbf{0},
\end{aligned} \tag{1.15}$$

where $\boldsymbol{q}_1, \boldsymbol{q}_2$ are vectors of the same dimensions as ζ.

Proposition 2. Consider a model $\mathcal{M}_{\text{OR}} \subset \mathcal{M}_{\text{IV}}$ which satisfies the identification condition in Theorem (1). Then the outcome regression estimator

$$\hat{\phi}^{\text{OR}} = \mathbb{P}_n \left\{ RY + (1-R)E\left(Y \middle| R=0, X, Z; \hat{\zeta}, \hat{\theta}_{\text{MLE}}\right)\right\} \tag{1.16}$$

is consistent and asymptotically normal as $n \to \infty$, that is

$$\sqrt{n}\left(\hat{\phi}^{\text{OR}} - \phi_0\right) \xrightarrow{d} N\left(0, V\right)$$

in model $\mathcal{M}_{\text{OR}}$ under suitable regularity conditions.

A consistent sandwich estimator for the asymptotic variance V analogous to (1.14) is straightforward to obtain.

1.4.3 Doubly robust estimation under $\mathcal{M}_{\text{IPW}} \cup \mathcal{M}_{\text{OR}}$

Estimation approaches described thus far depend on correct specifications of missing data model and outcome model for the IPW and OR estimators respectively. Here we propose a doubly robust estimator that remains consistent if the selection bias function η and the conditional density $q(z|x;\xi)$ are correctly specified, and any one of two models $P(y|R,X,Z;\theta)$ or $P(r|Y,X,Z;\omega)$ is correctly specified, but not necessarily both. we first derive a DR estimator $\hat{\zeta}_{\text{DR}}$ of the selection bias function $\eta(\zeta)$ that remains consistent in $\mathcal{M}_{\text{IPW}} \cup \mathcal{M}_{\text{OR}}$. In this vein, let

$$\begin{aligned}
&\boldsymbol{G}^{\text{DR}}\left(R, X, Y, Z; \zeta, \omega, \hat{\theta}_{\text{MLE}}, \boldsymbol{u}\right) \\
&= \frac{R}{\pi(\zeta,\omega)}\boldsymbol{u}(X,Y) - \frac{R-\pi(\zeta,\omega)}{\pi(\zeta,\omega)} E\left(\boldsymbol{u}(X,Y)\middle| R=0, X, Z; \zeta, \hat{\theta}_{\text{MLE}}\right) \\
&= \frac{R}{\pi(\zeta,\omega)}\left\{\boldsymbol{u}(X,Y) - E\left(\boldsymbol{u}(X,Y)\middle| R=0, X, Z; \zeta, \hat{\theta}_{\text{MLE}}\right)\right\} \\
&+ E\left(\boldsymbol{u}(X,Y)\middle| R=0, X, Z; \zeta, \hat{\theta}_{\text{MLE}}\right),
\end{aligned} \tag{1.17}$$

where $\boldsymbol{u}(X,Y)$ is of the same dimensions as ζ. We obtain $(\hat{\zeta}_{\text{DR}}, \hat{\omega})$ as the solution to the estimating equation (1.11) combined with

$$\begin{aligned}
&\mathbb{P}_n\left\{\boldsymbol{U}^{\text{DR}}\left(\zeta, \omega, \hat{\theta}_{\text{MLE}}, \hat{\xi}_{\text{MLE}}, \boldsymbol{u}, \boldsymbol{v}\right)\right\} = \\
&\mathbb{P}_n\left\{ \left[\boldsymbol{v}(X,Z) - E\left(\boldsymbol{v}(X,Z)\middle| X; \hat{\xi}_{\text{MLE}}\right)\right]\left[\boldsymbol{G}^{\text{DR}}\left(R, X, Y, Z; \zeta, \omega, \hat{\theta}_{\text{MLE}}, \boldsymbol{u}\right)\right]\right\} = \boldsymbol{0}.
\end{aligned} \tag{1.18}$$

Proposition 3. Consider a union model $(\mathcal{M}_{\text{IPW}} \cup \mathcal{M}_{\text{OR}}) \subset \mathcal{M}_{\text{IV}}$which satisfies the identification condition in Theorem (1). Then the doubly robust estimator

$$\hat{\phi}^{\text{DR}} = \mathbb{P}_n \left\{ \boldsymbol{G}^{\text{DR}} \left(R, X, Y, Z, \hat{\zeta}_{\text{DR}}, \hat{\omega}, \hat{\theta}_{\text{MLE}}, \boldsymbol{u} \right) \right\} \tag{1.19}$$

where $\boldsymbol{u}(X, Y) = Y$ is consistent and asymptotically normal as $n \to \infty$, that is

$$\sqrt{n} \left(\hat{\phi}^{\text{DR}} - \phi_0 \right) \xrightarrow{d} N(0, V)$$

in the union model $\mathcal{M}_{\text{IPW}} \cup \mathcal{M}_{\text{OR}}$ under suitable regularity conditions.

The notion of doubly robust estimation was first introduced in the context of semiparametric non-response models under MAR (Scharfstein et al., 1999), and the approach was further studied by others (Lipsitz et al., 1999; Robins et al., 2000; Lunceford and Davidian, 2004; Neugebauer and van der Laan, 2005) with theoretical underpinnings given by Robins and Rotnitzky (2001) and van der Laan and Robins (2003). A doubly robust version of estimating equation (1.19) of mean outcome under MNAR was previously described by Vansteelandt et al. (2007) who, as described earlier, assume that the selection bias function η is known a priori within the context of a sensitivity analysis. An important contribution of the current paper is to derive a large class of DR estimators of the selection bias using an IV. To the best of our knowledge, this is the first time that a DR estimator for the mean outcome has been constructed in the context of an IV for data subject to MNAR.

1.5 Semiparametric Estimation Theory

In the semiparametric model given by Assumption 1, we consider estimation of ζ and the full data functional $\phi = E(Y)$ based on the observed data $O = (R, RY, X, Z)$ by deriving their respective ortho-complement nuisance tangent spaces $\mathcal{N}_{\zeta}^{O,\perp}$ and $\mathcal{N}_{\phi}^{O,\perp}$. Throughout this section, spaces refer to sub-spaces of the Hilbert space of mean zero measurable random functions with bounded second moments, equipped with covariance inner product evaluated at the truth. The spaces $\mathcal{N}_{\zeta}^{O,\perp}$ and $\mathcal{N}_{\phi}^{O,\perp}$ are of interest since for sufficiently smooth models, the influence functions of regular and asymptotically linear (RAL) estimators for ζ and ϕ are normalized versions of elements in these ortho-complement nui-

sance tangent spaces respectively (Tsiatis, 2006). For example, the set of influence functions of all RAL estimators of ϕ is the space $IF_\phi = \left\{E(AS_\phi)^{-1}A; A \in \mathcal{N}_\phi^{O,\perp}\right\}$, where an element of $\mathcal{N}_\phi^{O,\perp}$ is a one-dimensional function of the observed data and S_ϕ is the score for ϕ, all evaluated at the truth. Taking an element $A^* \in \mathcal{N}_\phi^{O,\perp}$, we can estimate ϕ_0 by $\hat{\phi}$ which solves $\sum_i A^*\left(O_i;\hat{\phi}\right) = 0$. Under regularity conditions, $\hat{\phi}$ is an RAL estimator with the expansion $n^{1/2}\left(\hat{\phi}-\phi_0\right) = n^{-1/2}E(A^*S_\phi)^{-1}\sum_i A^*(O_i;\phi_0) + o_p(1)$, with influence function $E(A^*S_\phi)^{-1}A^*(O;\phi_0) \in IF_\phi$ (Bickel et al., 1993). This motivates derivation of the ortho-complement nuisance tangent space, since estimators can be constructed by identifying and solving empirical versions of the elements in the space.

It is shown in the appendix that the ortho-complement nuisance tangent spaces of the selection bias parameter ζ_0 and the population mean outcome ϕ_0 are given by

$$\mathcal{N}_\zeta^{O,\perp} = \left\{ \begin{array}{c} N_1^{O,\perp}\left(C-C^\dagger\right) = R\left\{C-C^\dagger\right\}/\pi\left(L\right) + (1-R)a_c\left(O\right) - RE\left[(1-R)a_c\left(O\right)|L\right]/\pi\left(L\right): \\ a_c = E\left[C-C^\dagger|R=0,X,Z\right] \end{array} \right\} \tag{1.20}$$

and

$$\mathcal{N}_\phi^{O,\perp} = \left\{ \begin{array}{c} N_1^{O,\perp}\left(k(Y-\phi_0)+C-C^\dagger\right) + E\left\{\nabla_\zeta N_1^{O,\perp}\left(k(Y-\phi_0)+C-C^\dagger\right)|_{\zeta_0}\right\} N_1^{O,\perp}\left(D-D^\dagger\right): \\ k \text{ is any constant} \end{array} \right\} \tag{1.21}$$

respectively, where $C = c\left(L\right)$ is any arbitrary function of $L = (Y,X,Z)$ and $C^\dagger = E\left[C|Z,X\right] + E\left[C|Y,X\right] - E\left[C|X\right]$. D and $D^\dagger$ are defined similarly. The estimator $\hat{\phi}^{\text{DR}}$ of the population mean outcome given in Proposition 3 corresponds to the element

$$N_1^{O,\perp}\left(k(Y-\phi_0)\right) + E\left\{\nabla_\zeta N_1^{O,\perp}\left(k(Y-\phi_0)\right)|_{\zeta_0}\right\} N_1^{O,\perp}\left(D_{\text{DR}} - D_{\text{DR}}^\dagger\right) \in \mathcal{N}_\phi^{O,\perp},$$

for a particular choice of D_{DR} corresponding to the element $N_1^{O,\perp}\left(D_{\text{DR}} - D_{\text{DR}}^\dagger\right) \in \mathcal{N}_\zeta^{O,\perp}$ of $\hat{\zeta}_{\text{DR}}$. The influence function of the efficient RAL estimator for ϕ_0 is given in the following result.

Proposition 4. Under Assumption 1, the efficient influence function of all RAL estimators of ϕ_0 is proportional to

$$N_1^{O,\perp}\left(k(Y-\phi_0)\right) - M + E\left\{\nabla_\zeta\left[N_1^{O,\perp}\left(k(Y-\phi_0)\right) - M\right]|_{\zeta_0}\right\} U \tag{1.22}$$

up to a normalizing constant, where $M = \Pi\left(N_1^{O,\perp}\left(k(Y-\phi_0)\right) \mid \left\{N_1^{O,\perp}\left(C-C^\dagger\right) : C\right\}\right)$, $\Pi(\cdot)$ is the projection operator and $U = N_1^{O,\perp}\left(D_{\text{eff}} - D_{\text{eff}}^\dagger\right)$ denote the influence function of the efficient RAL estimator of ζ_0. The estimator of ϕ_0 with the efficient influence function (1.22) is

$$\hat{\phi}^{\text{EFF}} = \mathbb{P}_n\left\{\boldsymbol{G}^{\text{DR}}\left(R,X,Y,Z,\hat{\zeta}_{\text{EFF}},\boldsymbol{u}\right) - \Pi\left(\boldsymbol{G}^{\text{DR}}\left(R,X,Y,Z,\hat{\zeta}_{\text{EFF}},\boldsymbol{u}\right) \mid \left\{N_1^{O,\perp}\left(C-C^\dagger\right) : C\right\}\right)\right\}, \tag{1.23}$$

where $\hat{\zeta}_{\text{EFF}}$ is the efficient estimator of ζ_0 and $\boldsymbol{u}(X,Y) = Y$.

Finding the closed form expression for (1.22) may be difficult in general. However, in the special case where both Z and Y are binary variables, $C - C^\dagger$ can be expressed as

$$b(X)\left\{Y - E(Y|X)\right\}\left\{Z - E(Z|X)\right\},$$

indexed by some function $b(X)$ so that

$$\begin{aligned} N_1^{O,\perp}\left(C-C^\dagger\right) &= b(X) \times \left\{R\left\{Y-E(Y|X)\right\}\left\{Z-E(Z|X)\right\}/\pi(L) + \right. \\ &\quad (1-R)E\left[\left\{Y-E(Y|X)\right\}\left\{Z-E(Z|X)\right\}|X,R=0,Z\right] \\ &\quad \left. -RE\left[(1-R)E\left[\left\{Y-E(Y|X)\right\}\left\{Z-E(Z|X)\right\}|X,R=0,Z\right]|L\right]/\pi(L)\right\} \\ &= b(X)W \end{aligned}$$

Let the efficient influence function for RAL estimators of the selection bias parameter ζ_0 be $IF_\zeta^{\text{EFF}} = d^*(X)W$. Then by Theorem 5.3 of Newey and McFadden (1993),

$$E\left\{d^*(X)W\left(\zeta\right)^2 d(X)\right\} = E\left\{\nabla_\zeta d(X)W(\zeta)\right\} \text{ for all } d(X)$$

solving which yields

$$d^*(X) = E\left\{W\left(\zeta\right)^2|X\right\}^{-1} E\left\{\nabla_\zeta W\left(\zeta\right)|X\right\}.$$

Let $b^*(X)W$ be the projection of $H = N_1^{O,\perp}\left(k(Y-\phi_0)\right)$ onto the space $\left\{N_1^{O,\perp}\left(C-C^\dagger\right) : C\right\}$. It follows by the property of projection that $E\left\{[H - b^*(X)W]b(X)W\right\} = 0$ for all $b(X)$. Therefore $E\left\{HW - b^*(X)W^2|X\right\} = 0$ which yields $IF_\phi^{\text{EFF}} = H - b^*(X)W = H - E\left\{HW|X\right\}E\left\{W^2|X\right\}^{-1}W$.

In practice, the conditional expectations in $b^*(X)$ and $d^*(X)$ usually need parametric modeling, especially when the covariate X is high-dimensional or continuous. However, even if they are misspecified, the models remain functions of the covariate X and therefore the resulting estimators $\hat{\zeta}$ and $\hat{\phi}$ are still consistent and have influence functions that belong to the spaces (1.20) and (1.21) respectively. If the posited parametric models contain the true model, then the resulting estimators attain the local semiparametric efficiency bounds as described in Proposition 4.

Instead of directly solving empirical versions of the respective efficient influence functions to obtain $\hat{\zeta}_{\text{EFF}}$ and $\hat{\phi}_{\text{EFF}}$, we use a result due to Bickel et al. (1993) which states that starting with an initial $\sqrt{n}$-consistent estimator, the efficient estimator can be obtained by a one-step update in the direction of the estimated efficient influence function. Therefore, we can first obtain $\hat{\zeta}_{\text{EFF}}$ by the formula

$$\hat{\zeta}_{\text{EFF}} = \hat{\zeta} - \left\{ \sum_i \left[\widehat{\nabla_\zeta IF_\zeta^{\text{EFF}}}; \hat{\zeta} \right] \right\}^{-1} \sum_i \left[\widehat{IF_\zeta^{\text{EFF}}}; \hat{\zeta} \right], \tag{1.24}$$

where $\hat{\zeta}$ is an initial estimate of any $\sqrt{n}$-consistent estimator of ζ_0 and $\widehat{IF_\zeta^{\text{EFF}}}$ is the empirical version of IF_ζ^{EFF} with estimated conditional expectations evaluated at $\hat{\zeta}$. Subsequently, we obtain $\hat{\phi}_{\text{EFF}}$ by a further one-step update

$$\hat{\phi}_{\text{EFF}} = \hat{\phi} - \left\{ \sum_i \left[\widehat{\nabla_\phi IF_\phi^{\text{EFF}}}; \hat{\phi}, \hat{\zeta}_{\text{EFF}} \right] \right\}^{-1} \sum_i \left[\widehat{IF_\phi^{\text{EFF}}}; \hat{\phi}, \hat{\zeta}_{\text{EFF}} \right], \tag{1.25}$$

where $\hat{\phi}$ is an initial estimate of any $\sqrt{n}$-consistent estimator of ϕ_0.

1.6 Simulation study

In order to investigate the finite-sample performance of the proposed estimators, we carried out a simulation study involving i.i.d. data (Y, Z, X_1, X_2). For each of sample sizes

$n = 1000, 2000, 5000$, we generated 1000 simulation replicates as followed,

$$X_1 \sim \text{Bernoulli}(p = 0.4), \quad X_2 \sim \text{Bernoulli}(p = 0.6)$$
$$P(Z = 1|X_1, X_2) = \text{expit}(0.2 - 0.5X_1 + 0.8X_2)$$
$$Y|Z, X_1, X_2 \sim N(0.7 - 0.5X_1 + 0.5X_2, 1)$$
$$P(R = 1|X_1, X_2, Y, Z) = \text{expit}(-1.2 + 2.5Z + 0.3X_1 + 0.8ZX_1 + Y).$$

Estimation was then based on the observed data (X_1, X_2, Z, R, RY). Under the above data generating mechanism, Z satisfies **(IV.1)** and **(IV.2)**, with the true value of $\phi_0 = E(Y) = 0.8$. The selection bias model is $\alpha(x, y, z) = \zeta y$ with true value $\zeta_0 = 1$. The model is identified since the missing data mechanism follows the separable logistic regression model described in Example 2. For IPW estimation, we specified the correct extended propensity score and conditional p.m.f. $f(Z = 1|X_1, X_2; \xi)$, and solved (7)-(9) with $h_1 = (Z, X_1, ZX_1)^T$, $g = Y$ and $h_2 = Z$. For OR estimation, we let $(q_1, q_2) = (Z, Y)$ in (13) and specified the outcome regression models

$$E\left[Y \exp(-\zeta Y)|R = 1, X_1, X_2, Z\right] = \mu\left(Z, X_1, X_2; \theta\right)$$
$$E\left[\exp(-\zeta Y)|R = 1, X_1, X_2, Z\right] = \mu\left(Z, X_1, X_2; \theta'\right)$$

where $\mu\left(z, x_1, x_2; \theta\right)$ is saturated in terms of (z, x_1, x_2) and obtained $\left(\hat{\theta}, \hat{\theta}'\right)$ via ordinary least squares using complete-cases only. DR estimation was carried out by combining the above estimators as described in the previous section.

To study the performance of the proposed estimators in situations where some models may be misspecified, we also evaluated the estimators within submodels in which either the extended propensity score model or the complete-case outcome density model was misspecified by replacing them with the models

$$P(R = 1|X_1, X_2, Y, Z) = \text{expit}(\omega + \zeta Y)$$
$$Y|R = 1, Z, X_1, X_2 \sim N\left(\theta_0 + \theta_1 X_1, \sigma^2\right)$$

respectively.

In each simulated sample, we evaluated the standard error of the estimator using the sandwich estimator given in (12) and (13). The coverage rate for the true value $\phi_0 = 0.8$

across 1000 simulations was calculated based on Wald 95% confidence intervals. Bias is given as the difference between mean Monte Carlo estimates and ϕ_0. We solved the estimating equations using the R package BB (Varadhan and Gilbert, 2009) and evaluated the Jacobi ans at estimated parameters with the R package numeric (Gilbert and Varadhan, 2012). Results for the simulation are presented in Table 1.1.

Table 1.1: Estimation of the average response $\phi_0 = 0.8$ from 1000 simulation replicates. ASE refers to the median standard error obtained using sandwich estimator. Coverages are based on 95% Wald confidence intervals. The bias in estimation of selection bias parameter $\zeta_0 = 1$ is also included.

Misspecified extended propensity score						
Estimator	n	Bias $\hat{\phi}$	MCSE	ASE	% Cov	Bias $\hat{\zeta}$
IPW	1000	-0.179	0.086	0.071	39.5	0.419
	2000	-0.175	0.063	0.051	14.6	0.367
	5000	-0.173	0.041	0.033	2.5	0.339
OR	1000	-0.007	0.071	0.067	94.9	0.151
	2000	-0.006	0.048	0.046	95.4	0.077
	5000	-0.003	0.029	0.029	96.0	0.031
DR	1000	-0.007	0.072	0.062	92.8	0.149
	2000	-0.006	0.048	0.044	92.7	0.077
	5000	-0.003	0.029	0.028	94.4	0.031
Misspecified complete-case outcome density						
Estimator	n	Bias $\hat{\phi}$	MCSE	ASE	% Cov	Bias $\hat{\zeta}$
IPW	1000	0.001	0.069	0.060	95.3	0.072
	2000	0.004	0.043	0.042	95.7	0.033
	5000	0.004	0.025	0.026	95.8	0.007
OR	1000	0.068	0.050	0.049	70.9	-0.495
	2000	0.065	0.035	0.035	52.8	-0.484
	5000	0.064	0.022	0.022	18.9	-0.485
DR	1000	-0.003	0.069	0.061	94.2	0.082
	2000	-0.005	0.047	0.044	94.9	0.055
	5000	-0.002	0.028	0.027	95.0	0.020
Both models correct						
Estimator	n	Bias $\hat{\phi}$	MCSE	ASE	% Cov	Bias $\hat{\zeta}$
DR	1000	-0.005	0.069	0.066	94.1	0.126
	2000	-0.005	0.048	0.047	95.1	0.074
	5000	-0.003	0.029	0.029	96.6	0.030

Under correct model specification, all estimators have negligible bias for ϕ_0 and ζ_0 that diminishes with increasing sample size, with empirical coverage near the nominal 95% level. In agreement with our theoretical results, the IPW and OR estimators are biased with poor empirical coverages when $\lambda(X, Z)$ or $f(Y|R = 1, X, Z; \theta)$ is misspecified, respectively. The DR estimator performs well in terms of bias and coverage when either

model is misspecified but the other is correct.

1.7 Applications

To illustrate the proposed IV approach, we obtained data from a household survey in Mochudi, Botswana to estimate HIV seroprevalence among adults adjusting for selective missingness of HIV test results. The data consist of 4997 adults between the ages of 16 and 64 who were contacted for the survey, out of whom 4045 (81%) had complete information on HIV testing. Of those who did not have HIV test results $(R = 0)$, 111 (2%) agreed to participate in the HIV test but their final HIV outcomes are unknown, and 841 (17%) refused to participate in the HIV testing component. It is likely that refusal to participate in the survey when contact is established presents a possible source of selection bias.
Fully available individual characteristics from the survey include participant gender (X). Candidate IVs include interviewer gender (Z_1), age (Z_2) and years of experience (Z_3). These interviewer characteristics are likely to influence the response rates of individuals who were contacted for the survey, but are unlikely to directly influence an individual's HIV status, given that interviewer deployment was determined at random prior to the survey. We implemented the proposed IPW, OR and DR estimators by making use of interviewer gender, age and years of experience as IVs. For IPW estimation, the missingness propensity score is specified as a linear main effects model with logistic link

$$\text{logit}\, P(R = 1|X, Y, \boldsymbol{Z}) = \omega_0 + \omega_1 X + \omega_2 Z_1 + \omega_3 Z_2 + \omega_4 Z_3 + \zeta Y \tag{A1}$$

where Y indicates HIV serostatus as our outcome of interest and the selection bias function is specified as $\alpha(x, y, z) = \zeta y$. The posited missing data mechanism belongs to the separable logistic class, therefore the average HIV prevalence can be identified by Proposition 2. For OR estimation, we specified the regression model

$$\text{logit}\, P(Y = 1|R = 1, X, \boldsymbol{Z}) = \theta_0 + \theta_1 X + \theta_2 Z_1 + \theta_3 Z_2 + \theta_4 Z_3. \tag{A2}$$

Finally, the doubly robust estimator is implemented by incorporating both of the above two models. Since more than one IV was available, estimating equations $\boldsymbol{U}^{\text{IPW}}$, $\boldsymbol{U}^{\text{OR}}$

and $\boldsymbol{U}^{\text{DR}}$ were solved using the generalized method of moments (GMM) package in R (Chaussé, 2010). Standard errors were obtained using the proposed sandwich estimator. For comparison, we also carried out standard complete-case analysis and standard IPW estimation assuming MAR given (x, z) under the propensity score model

$$\text{logit}\, P(R = 1|X, \boldsymbol{Z}) = \omega_0 + \omega_1 X + \omega_2 Z_1 + \omega_3 Z_2 + \omega_4 Z_3. \tag{A3}$$

Results from the analysis are presented in table 1.2.

Table 1.2: Estimation for HIV seroprevalence (ϕ) and magnitude of selection bias (ζ) in Mochudi, Botswana with 95% Wald confidence intervals.

Estimator	$\hat{\phi}$	$\hat{\zeta}$	$\hat{\zeta}$ p-val
CC	0.214 (0.202, 0.227)	-	-
MAR IPW	0.213 (0.201, 0.226)	-	-
IV IPW	0.260 (0.175, 0.341)	-1.601 (-2.992, -0.210)	0.02
IV OR	0.241 (0.175, 0.307)	-0.757 (-1.889, 0.376)	0.19
IV DR	0.258 (0.174, 0.342)	-1.121 (-2.433, 0.191)	0.09

IV-based point estimates of HIV seroprevalence are $12.6 - 21.5\%$ higher than the crude estimate of 0.214 (95% CI: 0.202-0.227) based on complete-cases only. Standard IPW produced similar estimates as complete-case analysis. The negative point estimates of the selection bias parameter ζ suggest that HIV-infected persons are less likely to participate in the HIV testing component of the survey, although this difference is statistically significant at 0.05 α-level only for IPW. The larger confidence intervals of the three IV estimators of ϕ_0 compared to those of the CC and MAR estimators are a more accurate reflection of the amount of uncertainty involving inferences about ϕ_0, since the CC and MAR estimators do not take into account the uncertainty about the underlying MNAR mechanism by assuming MCAR and MAR respectively. $\hat{\phi}^{\text{IPW}}$ and $\hat{\phi}^{\text{DR}}$ are close to each other. This comparison is useful as an informal goodness of fit test in that their similarity suggests that the missingness propensity score may be specified nearly correctly (Robins and Rotnitzky, 2001). In addition, by incorporating all possible pairwise interaction terms in the outcome logistic regression model (A2) and therefore allowing it to be more flexible, the OR point estimate $\hat{\phi}^{\text{OR}}$ increases to 0.246 (95% CI: 0.179-0.314) and becomes closer to $\hat{\phi}^{\text{IPW}}$ and $\hat{\phi}^{\text{DR}}$.

1.8 Discussion

In this paper, we have considered a pernicious form of selection bias which can arise from outcome missing not at random. We have argued that under fairly reasonable assumptions this problem can be made more tractable with the aid of an IV, and proposed a general framework for establishing identifiability of parametric, semiparametric and nonparametric models. We have proposed IPW and OR estimators which are consistent and asymptotically normal if the selection bias and the IV models are correctly specified, when either the extended propensity score or the outcome regression model is correctly specified respectively. We also constructed a DR estimator that remains consistent if either of the two models is correct, which gives the analyst two chances, instead of only one, to get correct inferences about the magnitude of selection bias and the mean outcome in the underlying population of interest.

Several interesting extensions could be explored in the future, including analogous methods for longitudinal data, as well as for dependent censoring of a survival outcome. It may also be of interest to extend the approach to a regression framework with covariate missing not at random.

On Inverse Probability Weighting for Nonmonotone Missing at Random Data

BaoLuo Sun[1] and Eric J. Tchetgen Tchetgen[1,2]

Departments of Biostatistics[1] and Epidemiology[2], Harvard T.H. Chan School of Public Health.

2.1 Introduction

Missing data is a major complication which occurs frequently in empirical research. Non-response in sample surveys, dropout or non-compliance in clinical trials and data excision by error or to protect confidentiality are but a few examples of ways in which full data is unavailable and our ability to make accurate inferences may be compromised. Missingness could also be introduced into a study by design, e.g. multi-stage sampling plans in order to reduce the cost associated with measurements for all subjects. In many practical situations, the missing data pattern is nonmonotone, that is, there is no nested pattern of missingness such that observing variable X_k implies that variable X_j is also observed, for any $j < k$. Nonmonotone missing data patterns may occur, for instance, when individuals who dropped out of a longitudinal study re-enter at later time points or in a cross-sectional regression analysis in which the outcome and covariates may be missing in patterns that are arbitrary across persons. The missing data process is said to be missing-completely-at-random (MCAR) if it is independent of both observed and unobserved variables in the full data, and missing-at-random (MAR) if, conditional on the observed variables, the process is independent of the unobserved ones (Rubin, 1976; Little and Rubin, 2002). A missing data process which is neither MCAR nor MAR is said to be missing-not-at-random (MNAR).

While complete-case (CC) analysis is the easiest to implement and often used in practice, the method is generally known to produce biased estimates when the missingness mechanism is not MCAR (Little and Rubin, 2002), although in regression settings, a CC analysis remains unbiased provided the missingness process does not depend on the outcome given observed covariates included in the regression model (Little and Rubin, 2002; Little and Zhang, 2011). Other commonly used procedures include last-observation-carried-forward analysis most commonly used in longitudinal studies and other single imputation techniques. However, such ad-hoc approaches typically provide valid inferences only under restrictive and often unrealistic conditions (Molenberghs et al., 2004; Siddiqui and Ali, 1998; Little and Rubin, 2002). More principled methods to appropriately account for missing data include parametric likelihood or Bayesian inference (Little and

Rubin, 2002; Horton and Laird, 1999; Ibrahim and Chen, 2000; Ibrahim et al., 2002, 2005) and parametric multiple imputation (MI) inference (Rubin, 1977; Schafer, 1999) which is widely utilized through its incorporation into mainstream statistical software (Horton and Lipsitz, 2001)

Inverse probability weighting (IPW) (Horvitz and Thompson, 1952; Little and Rubin, 2002; Robins et al., 1994; van der Laan and Robins, 2003; Tsiatis, 2006; Li et al., 2013; Seaman and White, 2013) is another method to reduce selection bias from missing data or unequal sampling fractions. IPW estimation does not require specification of the full-data likelihood, but the missingness mechanism needs to be modeled. The development of coherent models and practical estimation procedures for the response mechanism of nonmonotone missing data is challenging, even under the assumption that the data is MAR. To the best of our knowledge, and as discussed in the seminal missing data book of Tsiatis (2006), there currently is not available, a general approach to model an arbitrary nonmonotone missing data generating process strictly imposing MAR only. This represents an important gap in the missing data literature, which has essentially restricted the use of IPW estimation to monotone missing data settings.

There has been some debate in literature about the plausibility of the MAR assumption with nonmonotone missing data, and it has been argued that MNAR may be a more natural mechanism under such settings (Robins and Gill, 1997; Little and Rubin, 2002). Methods based on nonmonotone MNAR generally require and can be sensitive to additional parametric assumptions for the full data and missingness mechanism (Troxel et al., 1998; Ibrahim et al., 2001), or for just the missingness mechanism (Rotnitzky et al., 1998). An analysis assuming MAR may be preferable to one assuming MCAR even if the missingness mechanism is strictly MNAR (Little and Rubin, 2002; Molenberghs et al., 2014), and in some empirical settings yield more accurate predictions of the missing values than those based on MNAR for nonmonotone missing data (Rubin et al., 1995). The analytic simplifications with methods based on MAR for the often nuisance missingness mechanism benefit the main focus of inquiry (Schafer and Graham, 2002; Schafer, 2003). In addition, estimation under the MAR assumption provides a principled framework for anchoring inference in the presence of incomplete data (Molenberghs et al., 2014). Such

inference can and should subsequently be supplemented with sensitivity analyses to assess the extent to which a violation of MAR might lead to bias (Scharfstein et al., 2003). In this paper, we propose a class of models for arbitrary nonmonotone MAR data patterns. In order to estimate the missingness mechanism required for IPW estimation, we present two approaches: unconstrained maximum likelihood estimation (UMLE) and constrained Bayesian estimation (CBE). The first approach is easily implemented in standard software, say using existing procedures in SAS or R. However, despite this appealing feature, as we illustrate in the simulation studies, UMLE has a major drawback, in that the estimator may not be defined in finite samples, even if all regression models are correctly specified. This problematic feature of the approach is mainly due to certain natural restrictions of the model. In addition to UMLE, we introduce a CBE approach (Gelfand et al., 1992) which largely resolves any convergence difficulty and is easily implemented in standard Bayesian software packages. As IPW may be inefficient in practice, we improve its asymptotic efficiency by recovering available information from incomplete cases through implementing an augmented IPW (AIPW) estimator which is optimal within a very large class of AIPW estimators. The approach, which combines the proposed estimators of the nonmonotone missing data process with ideas originating from the seminal work of Robins et al. (1994) and further developed by van der Laan and Robins (2003) and Tsiatis (2006), holds appeal in the fact that it leverages available information from incomplete cases without having to specify a model of the full data distribution. We present a simulation study to investigate the finite-sample properties of both constrained and unconstrained inferences in the context of logistic regression with nonmonotone missing outcome and covariates, followed by an analysis of preterm delivery on a cohort of women in Botswana to illustrate an application of the methods.

2.2 Notation and Assumptions

Let $L = (L_1, ..., L_K)'$ be a random K-vector representing the complete data. Let R be the scalar random variable encoding the different missing data patterns. For missing data pattern $R = m$, where $1 \leq m \leq 2^K$, we only observe $L_{(m)} \subseteq L$. For each of n individuals,

we observe an independently and identically distributed realization of $\left(R_i, L_{(R)i}\right)$, $i = 1, 2, ..., n$, and we suppress the subject index i when not essential. We reserve $R = 1$ to denote complete cases. Let $\mathbb{P}_n$ denote the empirical measure $\mathbb{P}_n f(O) = n^{-1} \sum_i f(O_i)$.

We assume that the missing data process is MAR (Rubin, 1976; Robins et al., 1994). IPW methodology essentially requires for unbiasedness that MAR holds for all persons in the population and so, more specifically, we shall assume everywhere MAR (Seaman et al., 2013), also sometimes called missing-always-at-random (Mealli and Rubin, 2015) such that $\forall i, \gamma$,

$$\Pr\{R_i = m | l_i; \gamma\} = \Pr\{R_i = m | l_i^*; \gamma\}, \quad \forall m, l_i, l_i^* \text{ such that } l_{(m)i} = l^*_{(m)i} \tag{2.1}$$

where (l_i, l_i^*) represents a pair of possible values of L_i, so that the conditional probability of having missing data pattern m, which we denote by $\pi_m\left(l_{(m)}\right)$, depends only on the observed variables for that pattern. The finite or infinite dimensional parameter indexing the missing data mechanism is denoted by γ. Throughout, we also make the positivity assumption that $\forall i$,

$$\pi_1(l_i) > \sigma > 0 \quad \forall l_i \text{ in the support of } L_i, \tag{2.2}$$

for a fixed positive constant σ. That is, the probability of being a complete case is bounded away from zero with probability 1. Assumption (2.2) is necessary for identification of the full data law and smooth functionals of the latter (Robins et al., 1994), and ensures finite asymptotic variance of the IPW and AIPW estimators.

A key implication of assumptions (2.1) and (2.2) is that the missing data process is non-parametrically identified. We note that for likelihood-based methods the weaker assumption of realized MAR (Seaman et al., 2013) already implies that if separate parameters index the missing data mechanism and the full data distribution, efficient estimation of the parameters of the missing data process can be obtained by maximizing its partial likelihood, ignoring the part of the likelihood corresponding to the full data.

2.3 Estimation of Missing Data Mechanism

Although the missingness mechanism is in principle nonparametrically identified under assumptions (2.1) and (2.2), in practice estimation typically entails specifying parametric models as dictated by the curse of dimensionality, since L is typically of moderate to high dimension (Robins and Ritov, 1997). To motivate our discussion of nonmonotone missing data models, we briefly review strategies for modeling some common missing data structures. In the simple case of two missing data patterns, i.e. $R = 1, 2$, the probability of being a complete case is $1 - \pi_2\left(L_{(2)}\right)$ and the parameters γ of a model $\pi_2\left(L_{(2)};\gamma\right)$ can be estimated by maximizing the likelihood function

$$\prod_i \left\{1 - \pi_2\left(L_{(2)};\gamma\right)\right\}^{\mathbb{1}(R_i=1)} \left\{\pi_2\left(L_{(2)};\gamma\right)\right\}^{1-\mathbb{1}(R_i=1)}.$$

The two-missing-data-pattern scenario arises in familiar settings such as in regression analysis with incomplete data only on the outcome for a subset of the sample.

When $M > 2$ the missing data is said to be monotone if for some ordering of the variables in L, the k^{th} variable is observed only if the $k-1^{\text{th}}$ variable was observed, and therefore one can sort the missing data patterns in such a way that $L_{(m+1)} \subset L_{(m)}$ for $m = 1, ..., M-1$. Some of the earliest works in this area include weighting methods to adjust for nonresponse in panel studies with monotone missing data patterns (Little and David, 1983). In general, any monotone response mechanism can be modeled using a discrete hazard function (Robins et al., 1994; Tsiatis, 2006) by defining

$$\lambda_m(L_{(m)}) = \begin{cases} \Pr(R = m | R \leq m, L), & m \neq 1. \\ 1, & m = 1. \end{cases}$$

The discrete hazard $\lambda_m(\cdot)$ is a function of $L_{(m)}$ only since

$$\frac{\Pr(R = m|L)}{\Pr(R \leq m|L)} = \frac{\pi_m(L_{(m)})}{1 - \sum_{j>m} \pi_j(L_{(j)})}$$

and $L_{(j)} \subset L_{(m)}$ for all $j > m$ by the monotone missing data structure. Defining

$$K_m(L_{(m)}) = \Pr(R < m|L) = \prod_{j \geq m} \left\{1 - \lambda_j(L_{(j)})\right\}, \quad m \neq 1,$$

the conditional probability for each missing data pattern is

$$\pi_m(L_{(m)}) = \begin{cases} K_{m+1}(L_{(m+1)})\lambda_m(L_{(m)}), & m < M. \\ \lambda_m(L_{(m)}), & m = M. \end{cases}$$

and in particular the complete case probability is

$$\pi_1(L) = K_2(L_{(2)}) = \Pr(R < 2|L) = \prod_{j \geq 2} \left\{1 - \lambda_j(L_{(j)})\right\}$$

To estimate the hazard functions $\lambda_m(L_{(m)})$, in practice we may run a series of logistic regressions of the indicator variable $\mathbb{1}(R = m)$ on $L_{(m)}$ among individuals with $R \leq m$, $m = 2, ..., M$. Alternatively, one may pool information by allowing $\lambda_m(L_{(m)})$ to share parameters across m.

2.3.1 The failure of standard polytomous regression

For nonmonotone missing data patterns, the nesting of patterns $L_{(m+1)} \subset L_{(m)}$ is no longer available, and building coherent models for the conditional probabilities of the various missing data patterns is challenging even under assumptions (2.1) and (2.2) (Robins et al., 1994; Robins and Gill, 1997; Tsiatis, 2006). A straightforward approach to model $\pi_m(L_{(m)})$ using standard polytomous regression for the multinomial missing data process will often have the unintended consequence of imposing more restrictive conditions than what MAR assumption (2.1) strictly entails (Robins and Gill, 1997). We illustrate this using an example from Robins and Gill (1997), which we adapt to a general bivariate pattern (Little and Rubin, 2002, pp. 18-19). Suppose the full data is bivariate $L = (L_1, L_2)$ and one encodes the missing data patterns as follows: $R = 1$ if L is observed; $R = 2$ if one only observes $L_{(2)} = L_1$; $R = 3$ if one only observes $L_{(3)} = L_2$; and $R = 4$ if neither variable is observed. In general, the MAR assumption (2.1) for this scenario is $\forall\gamma$,

$$\Pr\{R = m|L; \gamma\} = \Pr\{R = m|L_{(m)}; \gamma\}, \quad m = 1, 2, 3, 4.$$

A standard polytomous logistic regression for R corresponds to

$$\Pr\{R = m|L; \gamma\} = \frac{\exp(\gamma_{0m} + \gamma_{1m}L_1 + \gamma_{2m}L_2)}{1 + \sum_{k=2}^{4} \exp(\gamma_{0k} + \gamma_{1k}L_1 + \gamma_{2k}L_2)}, \quad m = 2, 3, 4. \tag{2.3}$$

By the MAR assumption, since for $R = 4$ neither variable is observed, the probability $\Pr\{R = 4|L\}$ depends on neither L_1 nor L_2 so that $\gamma_{1j} = \gamma_{2j} = 0$ for $j = 2, 3, 4$. Therefore assuming model (2.3) under MAR implies MCAR. In general, it can be shown using a similar argument that the missing data pattern probabilities modeled using polytomous logistic regression can at most depend on the intersection of the sets of observed variables $L_{(m)}$, $m = 2, 3, ..., M$ (i.e. the set of fully observed variables), which is strictly stronger than the MAR assumption (2.1). This suggests that standard polytomous regression is ill-suited as modeling strategy for nonmonotone missing data process under MAR.

As a remedy, Robins and Gill proposed a large class of models for the missing data mechanism, which they call the randomized monotone missingness (RMM) processes, that are guaranteed to be MAR for a non-monotone missing data mechanism without necessarily being MCAR (Robins and Gill, 1997). This class of models does not span the space of all MAR models and therefore it is indeed possible to test whether the proposed class of models includes the true missing data mechanism. However, estimation of the missing data mechanism within this class is complex and computationally demanding, even for small to moderate sample sizes and number of different missing data patterns, and no software is currently available to implement the approach, which has limited its widespread adoption. In this paper we take a different direction and propose a class of models for nonmonotone missing data that spans the entire MAR model (with the class of RMM processes being a possible submodel) and therefore, with enough data such that non-parametric models can be used reliably, in principle one would not be able to reject MAR based on the observed data.

2.3.2 Proposed nonmonotone missing data model

Our approach involves modelling the conditional probability for each missing data pattern separately as

$$\Pr\{R = m|L\} = \pi_m(L_{(m)}), \quad m = 2, ..., M. \tag{2.4}$$

The probability of observing complete data is

$$\Pr\{R=1|L\} = \pi_1(L) = 1 - \sum_{m=2}^{M} \pi_m(L_{(m)}), \tag{2.5}$$

which depends on the union set of observed variables $\bigcup_{m=2}^{M} L_{(m)}$. To ground ideas, consider as a parametric submodel of (2.4) the series of simple logistic models

$$\begin{aligned}\pi_m\left(L_{(m)};\gamma_m\right) &= \left\{1+\exp\left[-\gamma_m\left(1,L_{(m)}\right)^T\right]\right\}^{-1}, \quad m=2,...,M,\\ \pi_1(L;\gamma) &= 1-\sum_{m=2}^{M}\left\{1+\exp\left[-\gamma_m\left(1,L_{(m)}\right)^T\right]\right\}^{-1}, \quad \gamma=(\gamma_2,...,\gamma_M).\end{aligned} \tag{2.6}$$

By assumption (2.2), model (2.6) must satisfy the constraint

$$1-\sum_{m=2}^{M}\pi_m\left(L_{(m)};\gamma_m\right) > \sigma \quad \text{with probability 1.} \tag{2.7}$$

Consider the UMLE estimator of γ, defined as the value which maximizes the unconstrained log-likelihood function corresponding to missing data model (2.6).

$$\sum_{i=1}^{N}\left\{\left[\sum_{m=2}^{M}\mathbb{1}(R_i=m)\log\pi_m\left(L_{(m)i};\gamma_m\right)\right] + \mathbb{1}(R_i=1)\log\left[1-\sum_{k=2}^{M}\pi_k\left(L_{(k)i};\gamma_k\right)\right]\right\} \tag{2.8}$$

with corresponding score equation

$$\mathbb{P}_n\left\{\left[\frac{\mathbb{1}(R=1)}{\pi_1\left(L_{(1)}\right)} - \frac{\mathbb{1}(R=m)}{\pi_m\left(L_{(m)}\right)}\right]\pi_m(1-\pi_m)\left(1,L_{(m)}\right)^T\right\} = 0 \tag{2.9}$$

for the parameters γ_m for missing data pattern m, where γ_m and $\left(1,L_{(m)}\right)^T$ have the same dimension.

It may be in practice that maximizing (2.8) fails to converge. This could happen if there is at least one individual for whom the empirical version of constraint (2.7) is not satisfied in the process of finding the maximum, in which case the fitted complete case probability may be near zero or possibly negative, a real possibility especially at small or moderate sample sizes. Thus, we have referred to (2.8) as an unconstrained log-likelihood function, as it does not naturally impose constraint (2.7).

Note that even if the missingness mechanism were known, constraint (2.7) which depends

on $\bigcup_{m=2}^{M} L_{(m)}$ can only be observed for complete case individuals. In fact, only complete cases need to satisfy the constraint in order to ensure that the UMLE can be computed in practice. Thus, one could in principle attempt to maximize the observed data log-likelihood (2.8) together with the observable constraints

$$\mathbb{1}(R_i = 1) \sum_{k=2}^{M} \pi_k \left(L_{(k)i}; \gamma_k \right) < 1 - \sigma^* \quad \text{for } i = 1, 2, ..., N, \tag{2.10}$$

where σ^* is a user-specified small positive constant. Still, this is potentially computationally prohibitive, since there are as many constraints as complete case observations.

Instead, in addition to UMLE, we develop a constrained Bayesian estimation approach where samples are drawn from the unconstrained posterior conditional distribution for γ and only those draws that fall into the constrained parameter space (2.10) are retained (Gelfand et al., 1992). An additional appeal of this approach is that the posterior credible intervals of γ are guaranteed to satisfy constraint (2.10), which is useful if one wishes to perform hypothesis testing to identify significant predictors in the missing data regression models. Constrained Bayesian estimation has been used previously in several other settings, for instance to estimate risk ratio and relative excess risk regressions (Chu and Cole, 2010, 2011); however, to the best of our knowledge, it has not been used in the current context. To implement the approach, we specify a diffuse prior distribution $g(\gamma)$ for $\gamma = (\gamma_2, ..., \gamma_M)$ under model (2.6) and incorporate constraint (2.10) in the posterior distribution of γ. Under the constrained Bayesian model, the posterior distribution of γ is proportional to

$$f(\gamma|data) \propto f(data|\gamma) g(\gamma) = \prod_{i=1}^{N} \left\{ \prod_{m=2}^{M} \left\{ \pi_m \left(L_{(m)i}; \gamma_m \right) \right\}^{\mathbb{1}(R_i = m)} \times \Omega(\gamma, L_i)^{\mathbb{1}(R_i = 1)} \right\} g(\gamma) \tag{2.11}$$

where

$$\Omega(\gamma, L_i) = \left\{ \left[1 - \sum_{k=2}^{M} \pi_k \left(L_{(k)i}; \gamma_k \right) \right] \times \mathbb{1} \left[\sum_{k=2}^{M} \pi_k \left(L_{(k)i}; \gamma_k \right) < 1 - \sigma^* \right] \right\}.$$

We define the CBE estimator of γ as the posterior mode (or mean) from distribution (2.11). We note that in practice there may be some missing data patterns that are sparsely ob-

served. In such cases, a simple approach entails combining across patterns with small event probabilities and estimating the missingness process under an additional assumption that the probability of any pattern within the combined set only depends on the intersection set of variables observed for all patterns in the combined set. Although the suggested approach to handle sparse patterns may introduce some bias, we do not anticipate the magnitude of this bias to be significant provided the combined set of patterns remains relatively rare compared to other more prominent missing data patterns. If the combination of sparse patterns gives rise to a monotone missing data pattern in the overall data set, then the standard approach of modeling variationally independent discrete hazards described earlier may be used. The probabilities of the missing data patterns are not variationally independent because of the nesting of patterns, i.e. probability of pattern m depends on hazards from m to M while that of pattern $m+1$ depends on hazards $m+1$ to M. The proposed approach subsumes monotone nonresponse patterns as a special case. However, some care is needed to ensure that the parameterization of models for each pattern respects their natural nesting in this setting. Nonetheless it will lead to a complicated estimation procedure without any apparent benefit in bias reduction or efficiency. Therefore in practice, existing discrete hazard function models should be used to construct weights with monotone missing data patterns.

2.4 IPW Inference

Suppose we observe n i.i.d. realizations of the vector L, and we wish to make inferences about the parameter β_0 which is the unique solution of the full data population estimating equation

$$E\{M(L;\beta_0)\} = 0 \tag{2.12}$$

where expectation is taken over the distribution of the complete data L. Note that we do not require a model for the distribution of the full data L; in fact, estimation is possible under certain weak regularity conditions (van der Vaart, 1998) as long as full data unbiased estimating functions exist. In the presence of missing data, the estimating function

in (2.12) may only be evaluated for complete cases, which may be a highly selective subsample even under MAR. This motivates the use of IPW estimating functions of complete cases to form the following population estimating equation

$$E\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L)}M(L;\beta_0)\right\}=0. \tag{2.13}$$

The unbiasedness of the above estimating equation holds by straightforward iterated expectations. We note that the IPW estimator $\hat{\beta}_{ipw}$ which solves empirical versions of (2.13) is inefficient especially when the fraction of complete-cases is small, since incomplete cases are discarded except in that they may be included in the estimation of the weights $\pi_1(L;\hat{\gamma})$. In the next section we will describe a strategy to recover information from incomplete cases by augmenting estimating function (2.13) to gain efficiency.

The IPW estimating equations framework encompasses a great variety of settings under which investigators may wish to account for non-monotone missing data. This includes IPW of the full data score equation, where the score function is such an unbiased estimating function, given a model $f(L|\beta)$ for the law of the full data, in which case (2.13) reduces to

$$E\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L)}\frac{\partial \log f(L|\beta)}{\partial\beta}\bigg|_{\beta_0}\right\}=0. \tag{2.14}$$

Note that equation (2.14) does not necessarily correspond to the observed data score equation, and will therefore generally not achieve the efficiency bound for the model. Estimation can also be extended to classes of semiparametric models which specify only certain marginal relationships in L and in which scientific interest focuses on some low dimensional functional $\beta=\beta(F_L)$ of the distribution F_L of the full data L. For instance, in many health related applications it is common to specify a model $g(X,\beta)$ for the conditional mean of the outcome response Y given a set of covariates $X=(X_1,X_2,...,X_P)^T$. Here $L=(Y,X)$ and either the outcome or any covariate may be missing. Then the parameter of interest can be identified by the population IPW estimating equation

$$E\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L)}\left[Y-g(X,\beta_0)\right]h(X)\right\}=0,$$

where $h(X)$ is a user-specified function of X of the same dimension as β_0. Regression parameters in semiparametric models for right censored failure time data can likewise be identified by similar IPW population estimating equations, e.g. Cox proportional hazards regression and Aalen's additive hazards regression. Analogous estimating equations are also available for longitudinal and clustered data. In all cases empirical estimating equations are obtained by replacing population expectations with their empirical counterparts, and $\pi_1(L)$ with a consistent estimator.

To fix ideas, let $\pi_1(L;\hat{\gamma}) = 1 - \sum_{m=2}^{M}\left\{1+\exp\left[-\hat{\gamma}_m\left(1, L_{(m)}\right)^T\right]\right\}^{-1}$ where $\hat{\gamma} = (\hat{\gamma}_2, ...\hat{\gamma}_M)$ is either the UMLE (assuming it can be computed) or CBE estimate. Then, an estimate for the parameter of interest β_0 is given by the solution $\hat{\beta}_{ipw}$ to the inverse probability weighted estimating equation

$$\mathbb{P}_n\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L;\hat{\gamma})}M(L;\beta)\right\} = 0. \tag{2.15}$$

Subject to standard regularity conditions and assuming that the missing data model given in (2.6) is correctly specified, we show in the supplementary material that $\hat{\beta}_{ipw}$ is consistent and asymptotically normal

$$\sqrt{n}(\hat{\beta}_{ipw} - \beta_0) \xrightarrow{d} N\left(0, E\{\nabla_\beta\Gamma(\beta_0,\gamma_0)\}^{-1}\mathrm{Var}\left[\Gamma(\beta_0,\gamma_0) - W(\beta_0,\gamma_0)\right]E\{\nabla_\beta\Gamma(\beta_0,\gamma_0)\}^{-1^T}\right) \tag{2.16}$$

where $\Gamma(\beta,\gamma) = \{\mathbb{1}(R=1)/\pi_1(L;\gamma)\}M(L;\beta)$, S_{γ_0} is the score function (2.9) for the missing data mechanism evaluated at the truth and

$$W(\beta_0,\gamma_0) = E\left[\Gamma(\beta_0,\gamma_0)S_{\gamma_0}^T\right]E\left[S_{\gamma_0}S_{\gamma_0}^T\right]^{-1}S\gamma_0.$$

The asymptotic variance in (2.16) can be consistently estimated by replacing the terms under expectation with empirical averages evaluated at $\left(\hat{\beta}_{ipw}, \hat{\gamma}\right)$

$$\widehat{E}\{\nabla_\beta\Gamma(\hat{\beta},\hat{\gamma})\}^{-1}\widehat{\mathrm{Var}}\left[\Gamma(\hat{\beta},\hat{\gamma}) - \widehat{W}(\hat{\beta},\hat{\gamma})\right]\widehat{E}\{\nabla_\beta\Gamma(\hat{\beta},\hat{\gamma})\}^{-1^T}. \tag{2.17}$$

Although the posterior mode (or mean) is asymptotically efficient by the Bernstein-von Mises Theorem (van der Vaart, 1998), in finite sample the CBE estimate may not necessarily correspond to the solution of the score function (2.9). For inference under the con-

strained Bayesian approach, we therefore apply a finite-sample correction to the variance estimate

$$\widehat{E}\{\nabla_\beta \Gamma(\hat{\beta},\hat{\gamma})\}^{-1}\widehat{\mathrm{Var}}\left[\Gamma(\hat{\beta},\hat{\gamma}) - \widehat{W}(\hat{\beta},\hat{\gamma}) + \widehat{E}\{W(\hat{\beta},\hat{\gamma})\}\right]\widehat{E}\{\nabla_\beta \Gamma(\hat{\beta},\hat{\gamma})\}^{-1^T} \tag{2.18}$$

so that the term in $\widehat{\mathrm{Var}}[\cdot]$ has mean zero empirically. The correction term $\widehat{E}\{W(\hat{\beta},\hat{\gamma})\}$ is expected to vanish as sample size increases. A conservative, albeit more easily implementable, estimate of the asymptotic variance in (2.16) is obtained by the standard sandwich variance formula (Robins et al., 1994)

$$\widehat{E}\{\nabla_\beta \Gamma(\hat{\beta},\hat{\gamma})\}^{-1}\widehat{\mathrm{Var}}\left[\Gamma(\hat{\beta},\hat{\gamma})\right]\widehat{E}\{\nabla_\beta \Gamma(\hat{\beta},\hat{\gamma})\}^{-1^T}. \tag{2.19}$$

2.4.1 Improved IPW estimator via augmentation

The efficiency of the IPW estimator introduced in the previous section, which only makes direct use of complete cases, can be improved by incorporating information from individuals with missing data via augmentation of the IPW estimating equation (Robins et al., 1994; van der Laan and Robins, 2003; Tsiatis, 2006). The approach is based on a result due to Robins et al. (1994) who show that under assumptions (1) and (2), all regular and asymptotically linear (RAL) estimators based on observed data, of a functional β_0, can be shown to be asymptotically equivalent to an estimator solving

$$\mathbb{P}_n\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L)}U(L;\beta) + A\left(R, L_{(R)}\right)\right\} = 0. \tag{2.20}$$

$U(L;\beta)$ is an element of $\mathbb{U}^F$, the set of all full data estimating equations of β_0, and $A\left(R, L_{(R)}\right)$ is an element of the space $\mathbb{A}$ spanned by all scores of the missing data mechanism which are of the form

$$\left\{\sum_{r\neq 1}\left[\frac{\mathbb{1}(R=1)}{\pi_1(L)} - \frac{\mathbb{1}(R=r)}{\pi_r\left(L_{(r)}\right)}\right]t_r\left(L_{(r)}\right)\right\},$$

where $t_r\left(L_{(r)}\right)$ is an arbitrary q-dimensional function of the observed data $L_{(r)}$ corresponding to missing data pattern $R = r$ (Robins et al., 1994). The class of estimating equations obtained by varying $U(L)$ over $\mathbb{U}^F$ and $A\left(R, L_{(R)}\right)$ over $\mathbb{A}$ is referred to as

augmented estimating equations, since it entails augmenting a standard IPW estimating equation by an arbitrary score function of the missingness process (Robins et al., 1994; Tsiatis, 2006). In principle, one can therefore construct an efficient estimator by identifying the optimal full data estimating function $U_{\text{opt}} \in \mathbb{U}^F$ paired with the optimal choice of augmentation $A_{\text{opt}} \in \mathbb{A}$ to use in equation (2.20). Unfortunately the optimal index leading to a semiparametric efficient estimator is generally not available in closed form and often computationally prohibitive in most problems of interest. Instead, we take a more practical approach to improve efficiency by using a restricted class of estimators (Tsiatis, 2006).

We illustrate the approach using an example with two levels of missingness. Suppose the full data $L_i = (Y_i, \boldsymbol{X}_i)$ is independent and identically distributed for $i = 1, 2, ..., n$, where Y is the binary response variable and $\boldsymbol{X} = (X_1, X_2)^T$ are two univariate covariates. For a subsample of individuals, only (Y, X_1) was observed. Let the missing data indicator be $R_i = 1$ if the i-th individual is a complete-case and $R_i = 2$ if we only observe $L_{(2)i} = (Y_i, X_{1i})$. Suppose we assume the substantive model to be

$$\Pr(Y = 1|\boldsymbol{X}) = \left[1 + \exp(\boldsymbol{\beta}^T \boldsymbol{X})\right]^{-1} = \mu(\boldsymbol{X}, \boldsymbol{\beta}),$$

and we are interested in estimating $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$, then the class of all augmented IPW estimators (AIPW) will be any estimator that solves

$$\mathbb{P}_n \left\{ \frac{\mathbb{1}(R = 1)}{\pi_1(L)} \boldsymbol{h}_{3\times 1}(\boldsymbol{X}, \boldsymbol{\beta}) \left[Y - \mu(\boldsymbol{X}, \boldsymbol{\beta})\right] + \left[\frac{\mathbb{1}(R = 1)}{\pi_1(L)} - \frac{\mathbb{1}(R = 2)}{\pi_2\left(L_{(2)}\right)} \right] \boldsymbol{f}_{3\times 1}(Y, X_1) \right\} = 0.$$

The functions $\boldsymbol{h}_{3\times 1}(\boldsymbol{X}, \boldsymbol{\beta})$ and $\boldsymbol{f}_{3\times 1}(Y, X_1)$ are any arbitrary functions of $\boldsymbol{X}$ and (Y, X_1) respectively, where the subscripts denote their dimensions. The optimal AIPW estimator in terms of asymptotic variance corresponds to a specific choice which we denote as $\boldsymbol{h}^{\text{opt}}_{3\times 1}(\boldsymbol{X}, \boldsymbol{\beta})$ and $\boldsymbol{f}^{\text{opt}}_{3\times 1}(Y, X_1)$. The optimal choice $(\boldsymbol{h}^{\text{opt}}_{3\times 1}, \boldsymbol{f}^{\text{opt}}_{3\times 1})$ is only available in closed form in special simple settings, and typically require solving complicated integral equations for each observation (Robins et al., 1994; Tsiatis, 2006). This will generally be the case for nonmonotone nonresponse, and therefore we consider a more practical approach, which we introduce here in the simple case with two levels of missingness, in the interest

of simplifying the presentation. The supplement includes a detailed description of the approach for general nonmonotone patterns.

The proposed approach entails approximating $\boldsymbol{h}^{\text{opt}}_{3\times1}(\boldsymbol{X},\boldsymbol{\beta})$ and $\boldsymbol{f}^{\text{opt}}_{3\times1}(Y,X_1)$ with a linear combination of basis functions. For instance, the choice of basis functions $\boldsymbol{J}^{h}_{6\times1}(\boldsymbol{X}) = \{1, X_1, X_2, X_1^2, X_2^2, X_1X_2\}^T$ and $\boldsymbol{J}^{f}_{6\times1}(Y,X_1) = \{1, Y, X_1, Y^2, X_1^2, YX_1\}^T$ allows for quadratic relationships in (X_1, X_2) and (Y, X_1) respectively. The approximations to $(\boldsymbol{h}^{\text{opt}}_{3\times1}, \boldsymbol{f}^{\text{opt}}_{3\times1})$ are $\boldsymbol{h}^{*}_{3\times1} = A_{3\times6}\boldsymbol{J}^{h}_{6\times1}$ and $\boldsymbol{f}^{*}_{3\times1} = B_{3\times6}\boldsymbol{J}^{f}_{6\times1}$ respectively, where $A_{3\times6}$ and $B_{3\times6}$ are arbitrary constant matrices. We can then consider the class of augmented estimators

$$\mathbb{P}_n\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L)}\boldsymbol{h}^{*}_{3\times1}(\boldsymbol{X})\left[Y-\mu(\boldsymbol{X},\boldsymbol{\beta})\right]+\left[\frac{\mathbb{1}(R=1)}{\pi_1(L)}-\frac{\mathbb{1}(R=2)}{\pi_2\left(L_{(2)}\right)}\right]\boldsymbol{f}^{*}_{3\times1}(Y,X_1)\right\}=0. \quad \text{(A1)}$$

It is possible to estimate the unique constant matrices $A_{3\times6}$ and $B_{3\times6}$ in the class of estimators (A1) which give the optimal efficiency in the class. This estimator is guaranteed to be more efficient asymptotically compared to the simple IPW estimator typically used by analysts which solves

$$\mathbb{P}_n\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L)}(1,X_1,X_2)^T\left\{Y-\mu(\boldsymbol{X},\boldsymbol{\beta})\right\}\right\}=0.$$

An appeal of the proposed approach of approximating the optimal functions with linear combinations of basis functions is that it does not require specification of the full data law beyond the substantive model of interest as well as assumptions (1) and (2) to estimate the weights $\pi_1(L;\hat{\gamma})$.

2.5 Simulation

In this section we report a simulation study to investigate the finite-sample properties of the proposed estimators. Independent and identically distributed $(Y, \boldsymbol{X})$ is generated where $\boldsymbol{X} = (X_1, X_2, X_3)$ follow the truncated normal distributions $X_1 \sim \text{N}(\mu = 1, \sigma = 0.5)$, $X_2 \sim \text{N}(\mu = X_1U, \sigma = 0.5)$ and $X_3 \sim \text{N}(\mu = X_1X_2, \sigma = 0.5)$ on the support $\boldsymbol{X} \in [0,2]^3$ with $U \sim \text{Unif}(0,1)$. The binary outcome variable Y is then generated with the

substantive model

$$\text{logit}\Pr(Y=1|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3. \tag{2.21}$$

The generated full data is then induced with missing values following missing data model (2.6) under three scenarios.

2.5.1 MAR Mechanism which depends on both the outcome and covariates

Scenario 1 concerns nonmonotone missing data with three patterns $L_{(1)} = (Y, \boldsymbol{X})$, $L_{(2)} = (Y, X_1)$ and $L_{(3)} = (X_2, X_3)$ generated under the MAR mechanism

$$\begin{aligned}\text{logit}\Pr\{R=2|Y,\boldsymbol{X}\} &= \{\gamma_{20} + \gamma_{21}Y + \gamma_{22}X_1\}\\ \text{logit}\Pr\{R=3|Y,\boldsymbol{X}\} &= \{\gamma_{30} + \gamma_{31}X_2 + \gamma_{32}X_3\},\end{aligned} \tag{2.22}$$

with the probability of being a complete case $\Pr\{R=1|Y,\boldsymbol{X}\} = 1 - \Pr\{R=2|Y,\boldsymbol{X}\} - \Pr\{R=3|Y,\boldsymbol{X}\}$. This mechanism may be reasonable when, for example, certain combinations of variables in a survey data increase the risks of personal identification and are withheld from the analysts due to confidentiality concerns (Molenberghs et al., 2014).

2.5.2 MAR Mechanism which depends on covariates only

In scenario 2, the missing data model is independent of the outcome Y given covariates $\boldsymbol{X}$ in the substantive model of interest. Nonmonotone missing data patterns $L_{(1)} = (Y, \boldsymbol{X})$, $L_{(2)} = (X_1)$ and $L_{(3)} = (X_2, X_3)$ are generated under the MAR mechanism

$$\begin{aligned}\text{logit}\Pr\{R=2|Y,\boldsymbol{X}\} &= \{\gamma_{20} + \gamma_{21}X_1\}\\ \text{logit}\Pr\{R=3|Y,\boldsymbol{X}\} &= \{\gamma_{30} + \gamma_{31}X_2 + \gamma_{32}X_3\}.\end{aligned} \tag{2.23}$$

2.5.3 MNAR Mechanism

Scenario 3 concerns a MNAR missing data model, which has been argued as a more natural mechanism with nonmonotone missing data (Robins and Gill, 1997; Little and Rubin, 2002). Missing data with three observed patterns $L_{(1)} = (Y, \boldsymbol{X})$, $L_{(2)} = (Y, X_1)$ and

$L_{(3)} = (X_2, X_3)$ is generated under the MNAR mechanism

$$\begin{aligned}
&\text{logit}\Pr\{R = 2|Y, \boldsymbol{X}\} = \{\gamma_{20} + \gamma_{21}Y + \gamma_{22}X_1 + \gamma_{23}X_2 + \gamma_{24}X_3\} \\
&\text{logit}\Pr\{R = 3|Y, \boldsymbol{X}\} = \{\gamma_{30} + +\gamma_{31}Y + \gamma_{32}X_1 + \gamma_{33}X_2 + \gamma_{34}X_3\}, \qquad (2.24)
\end{aligned}$$

i.e. the missing data mechanism for each pattern depends on all the variables in the substantive model, although only $L_{(m)}$ is observed for $m = 1, 2, 3$. The missing data process is generated from a multinomial distribution with the probabilities in (2.22)-(2.24) for the three scenarios respectively and only the corresponding observed data for the sampled pattern contributes to estimation. We perform 1000 replicates each with sample sizes $n = 500, 1000$. Each simulation replicate has approximately 30% to 40% of complete cases.

For IPW and AIPW estimators, the missing data models are specified as (2.22) in scenarios 1 and 3, and (2.23) in scenario 2. The choice of basis functions for AIPW estimation includes linear and quadratic terms. The parameters γ of the missing data models are estimated using both the UMLE and CBE to construct the weights, and the substantive model is correctly specified as (2.21). The UMLE estimator of γ is implemented using the R function optim with the quasi-Newton method BFGS. We obtain the CBE estimator of γ as the posterior mean of distribution (2.11) with independent diffuse priors $\gamma \sim N(0, 10^2)$ and $\sigma^* = 10^{-8}$. Adaptive Gibbs sampling (Gilks et al., 1995) was implemented through "BRugs" , the R interface to the OpenBUGS MCMC software (Lunn et al., 2009). More details on the implementation as well as the sample OpenBUGS code for estimation of missing data model in scenario 1 are included in the supplementary materials.

For comparison, two likelihood-based MI methods are also included: full predictive distribution sampling assuming multivariate normality (NORM) (Schafer, 1997) and multivariate imputation by chained equations (MICE) (van Buuren and Oudshoorn, 2000; van Buuren and Groothuis-Oudshoorn, 2011) based on the variables $(Y, \boldsymbol{X})$ in the substantive and missing data models. For NORM the imputed Y value is dichotomized to the binary $\mathbb{1}(Y > 0)$. The imputation model for outcome Y in MICE is the correctly specified substantive model (2.21), while the continuous variables $\boldsymbol{X}$ are imputed via predictive mean

matching (Rubin, 1986). The results for both MI methods are based on 10 imputed data sets in each simulation replicate. Lastly we also implement unweighted complete-case (CC) regression to evaluate the magnitude of selection bias, and carry out MLE based on the full data (Full MLE) to assess the extent of efficiency loss due to missing data. The results of the simulation study for scenarios 1-3 are shown in tables 2.1-2.3 respectively.

Table 2.1: Absolute value of empirical bias (|Bias|), mean square root of estimated variance (MSE) and Monte Carlo standard error (MCSE) for estimation of $\boldsymbol{\beta}$ in the substantive model with MAR mechanism† which depends on both outcome and covariates (scenario 1). The true value of $\boldsymbol{\beta}$ is $(\beta_0, \beta_1, \beta_2, \beta_3) = (-2.5, 0.7, 0.8, 1.0)$. All entries are original values multiplied by 1000.

		β_0			β_1			β_2			β_3		
n	Method	|Bias|	MSE	MCSE	|Bias|	MSE	MCSE	|Bias|	MSE	MCSE	|Bias|	MSE	MCSE
500	Full MLE	29	297	309	19	266	266	7	291	296	14	260	265
	CC	262	450	465	151	406	403	113	464	467	117	409	423
	IPW	81	470	472	53	419	405	20	482	482	28	422	432
	AIPW	96	400	434	56	350	369	24	460	483	32	408	432
	NORM	119	349	355	42	308	316	124	369	421	100	334	398
	MICE	52	394	401	65	346	351	55	461	467	24	413	432
1000	Full MLE	9	209	214	2	187	187	3	205	205	7	183	187
	CC	301	313	328	119	282	288	109	324	339	111	285	288
	IPW	28	330	328	14	292	280	8	340	347	17	296	297
	AIPW	42	281	299	21	243	251	11	328	347	19	288	297
	NORM	193	238	245	32	212	223	119	260	302	99	229	265
	MICE	19	276	284	45	241	249	58	321	340	11	287	294

†True value of γ in the missing data model is $(\gamma_{20}, \gamma_{21}, \gamma_{22}, \gamma_{30}, \gamma_{31}, \gamma_{32}) = (-0.8, -1.5, 0.2, -1.2, 0.3, 0.3)$.

Table 2.2: Absolute value of empirical bias (|Bias|), mean square root of estimated variance (MSE) and Monte Carlo standard error (MCSE) for estimation of $\boldsymbol{\beta}$ in the substantive model with MAR mechanism† which depends on the covariates only (scenario 2). The true value of $\boldsymbol{\beta}$ is $(\beta_0, \beta_1, \beta_2, \beta_3) = (-2.5, 0.7, 0.8, 1.0)$. All entries are original values multiplied by 1000.

		β_0			β_1			β_2			β_3		
n	Method	|Bias|	MSE	MCSE	|Bias|	MSE	MCSE	|Bias|	MSE	MCSE	|Bias|	MSE	MCSE
500	Full MLE	36	298	290	8	266	255	22	292	292	5	260	267
	CC	90	531	540	23	475	472	53	543	569	16	477	481
	IPW	108	544	567	29	486	487	67	553	596	19	487	499
	AIPW	109	534	570	29	478	489	68	539	599	20	477	500
	NORM	236	385	442	76	343	382	61	382	472	108	341	405
	MICE	63	542	541	6	487	481	25	550	570	36	487	491
1000	Full MLE	11	209	214	2	187	188	11	205	217	8	183	184
	CC	33	367	370	6	331	334	12	377	403	24	332	353
	IPW	33	379	389	7	338	342	10	387	415	23	338	359
	AIPW	33	374	390	6	335	343	9	381	416	24	335	359
	NORM	298	271	304	83	245	274	106	263	339	103	235	302
	MICE	8	372	377	15	338	348	12	375	410	36	336	369

†True value of γ in the missing data model is $(\gamma_{20}, \gamma_{21}, \gamma_{30}, \gamma_{31}, \gamma_{32}) = (-0.8, 0.2, -1.2, 0.3, 0.3)$.

In scenario 1, the CC estimator has substantial bias irrespective of sample size since the missing data model involves both the outcome and covariates. Among the estimators

Table 2.3: Absolute value of empirical bias (|Bias|), mean square root of estimated variance (MSE) and Monte Carlo standard error (MCSE) for estimation of $\boldsymbol{\beta}$ in the substantive model with MNAR mechanism† (scenario 3). The true value of $\boldsymbol{\beta}$ is $(\beta_0, \beta_1, \beta_2, \beta_3) = (-2.5, 0.7, 0.8, 1.0)$. All entries are original values multiplied by 1000.

		β_0			β_1			β_2			β_3		
n	Method	|Bias|	MSE	MCSE	|Bias|	MSE	MCSE	|Bias|	MSE	MCSE	|Bias|	MSE	MCSE
500	Full MLE	28	298	303	20	266	264	0	291	285	11	260	255
	CC	309	403	401	206	366	355	127	415	421	95	367	366
	IPW	21	412	404	78	373	348	88	423	428	62	373	370
	AIPW	20	368	387	72	323	329	91	414	430	63	366	373
	NORM	234	333	338	20	301	301	75	356	386	47	315	346
	MICE	75	366	372	55	323	322	33	416	415	55	370	374
1000	Full MLE	22	210	213	3	187	187	16	205	202	5	183	181
	CC	328	282	276	177	255	258	143	291	291	84	257	257
	IPW	38	289	279	51	260	251	101	296	296	48	261	263
	AIPW	47	259	266	37	226	236	102	291	295	50	257	263
	NORM	253	236	228	8	213	215	31	248	265	82	219	242
	MICE	88	256	257	26	226	233	48	291	289	50	262	266

†True value of γ in the missing data model is $(\gamma_{20}, \gamma_{21}, \gamma_{22}, \gamma_{23}, \gamma_{24}, \gamma_{30}, \gamma_{31}, \gamma_{32}, \gamma_{33}, \gamma_{34}) = (-1.0, -1.5, 0.3, -0.1, -0.2, -1.4, 0.4, 0.4, -0.8, 0.1)$.

which account for missing data, the NORM estimator is generally the most efficient, although it also has the greatest bias, since it places strong assumptions on the full data law by specifying multivariate normality. Compared to the NORM estimator, the MICE estimator is less biased, but also less efficient. The imputation model for the binary outcome is correctly specified as the substantive model, and predictive mean matching is less vulnerable to misspecification than explicit models for the distribution of the missing values conditional on the observed ones (Andridge and Little, 2010). The IPW and AIPW estimators, which place no further restrictions on the full data law beyond the substantive model of interest, are generally the least biased. The efficiency of the IPW and CC estimator is similar, though in instances where CC estimator is biased such differences of efficiency are not meaningful. By incorporating incomplete cases in the estimation, the AIPW estimator achieves efficiency which is on par with that of MICE. The asymptotic relative efficiency comparing AIPW to IPW estimates varies between 0.69–0.95 based on estimated variances. For the current data generating mechanism, the estimated variances of the IPW, AIPW, NORM and MICE estimators are generally biased downwards compared to empirical variances in finite samples, but show improvement at the larger sample size $n = 1000$.

The proportion of simulation replicates for which the UMLE converged increased slightly

with a doubling of sample size (96.5% and 99.4% for $n = 500, 1000$) under scenario 1. The smallest estimated complete-case probabilities for non-convergence cases hover around zero, suggesting that, as we have previously hypothesized, lack of convergence of the UMLE approach may be due mostly to empirical complete case probabilities that effectively violate the positivity assumption (2.2), which may occur by chance particularly in small samples, even when the assumption holds in the population and the missing data model is correctly specified. In contrast, the CBE is guaranteed to produce an estimate for the complete case probabilities within the parameter space of the model. Since CBE and UMLE produce similar estimates for the weights (when the latter converges), only the results from CBE estimation of the weights are shown for the IPW and AIPW estimators in tables 1 to 3.

Under scenario 2, the NORM estimator similarly has the greatest bias while being the most efficient. The bias and efficiency of AIPW and MICE estimators are similar. The CC estimator also has low bias, since in this case the missing data mechanism depends on only the covariates in the regression model (Little and Rubin, 2002; Little and Zhang, 2011; White and Carlin, 2010), and is in fact slightly more efficient than the AIPW and MICE estimators. Lastly, under the MNAR mechanism in scenario 3, all four estimators IPW, AIPW, NORM and MICE are biased. However, their bias is smaller than that of the CC estimator, as the missing data mechanism depends at least in part on some of the observed variables in each missing data pattern. Therefore, assuming MAR in accounting for missing data is able to mitigate some but not all selection bias.

2.6 Application

The empirical application concerns a study of the association between maternal exposure to highly active antiretroviral therapy (HAART) during pregnancy and birth outcomes among HIV-infected women in Botswana. A detailed description of the study cohort has been presented elsewhere (Chen et al., 2012). The entire study cohort consists of 33148 obstetrical records abstracted from 6 sites in Botswana for 24 months. Our current analysis focuses on the subset of women who were known to be HIV positive ($n = 9711$). The birth

outcome of interest is preterm delivery, defined as delivery < 37 weeks gestation. 6.7% of the outcomes are unobserved. The data also contains a number of predictors of interest with unobserved values (Table 2.4): maternal hypertension in pregnancy (6.5% missing), whether CD4^+ cell count is less than 200 μL (53.4% missing) and whether a woman continued HAART from before pregnancy or not. Our goal is to correlate these factors with preterm delivery. We applied the proposed IPW and AIPW estimators in logistic regression as well as performed CC analysis. We also provide results for MICE and imputation assuming multivariate normality (NORM) for comparison (Table 2.5).

Table 2.4: Tabulation of non-monotone missing data patterns as a percentage of total data ($n = 9711$). Missing variables are indicated by 0. Complete-cases are given in the first pattern $R = 1$.

R	Preterm Delivery	Hypertension	Low CD4^+	Cont. HAART	% of data
1	1	1	1	1	43.7
2	0	1	1	1	2.0
3	1	0	1	1	0.7
4	0	0	1	1	0.2
5	1	1	0	1	44.9
6	0	1	0	1	2.9
7	1	0	0	1	4.0
8	0	0	0	1	1.6

Table 2.5: Analysis for outcome preterm delivery with estimated odds ratios from logistic regression. Wald 95% confidence intervals for IPW /AIPW estimators are based on estimated asymptotic variances. The standard errors for MICE and NORM are estimated by Rubin's formula (Rubin, 1987) with $M = 50$ imputed samples.

Method	Hypertension	Low CD4^+	Cont. HAART
CC	1.29 (1.06, 1.57)	1.12 (0.89, 1.40)	1.31 (1.04, 1.65)
IPW	1.55 (1.19, 2.01)	1.12 (0.84, 1.50)	1.52 (1.18, 1.97)
AIPW	1.41 (1.22, 1.61)	1.08 (0.88, 1.34)	1.46 (1.28, 1.66)
MICE	1.34 (1.17, 1.54)	1.03 (0.77, 1.39)	1.22 (1.09, 1.36)
NORM	1.35 (1.19, 1.54)	1.08 (0.91, 1.29)	1.21 (1.09, 1.35)
Analysis combining missing data patterns $R = 3, 4$			
IPW	1.55 (1.20, 2.01)	1.13 (0.85, 1.52)	1.52 (1.18, 1.97)
AIPW	1.40 (1.22, 1.61)	1.11 (0.90, 1.37)	1.46 (1.28, 1.66)

The IPW estimator of the logistic regression for preterm delivery uses to estimate the weight a missing data model of the form given by (2.6), which includes the main effects of observed variables $L_{(m)}$ for each missing data pattern $m = 2, ..., 8$. Given the fairly large sample size $(n = 9711)$, the results for IPW are similar using UMLE and CBE to estimate the missing data process, consistent with findings from both the simulation study and

Table 2.6: Posterior medians with 95% credible intervals from constrained Bayesian estimation (CBE) of missing data model parameters γ for each missing data pattern $R = m$, $m = 2, 3, ..., 8$. Asterisk denotes exclusion of zero from the credible interval.

R	intercept	Preterm delivery	Hypertension	Low CD4+ Count	Cont. HAART
2	-4.24(-4.45,-4.05)∗		1.06(0.78, 1.35)∗	0.58(0.24, 0.90)∗	-0.33(-0.68, 0.03)
3	-5.12(-5.46,-4.81)∗	0.94(0.45, 1.39)∗		-0.55(-1.54, 0.19)	-0.26(-0.86, 0.30)
4	-6.31(-6.95,-5.77)∗			-1.08(-4.13, 0.73)	-0.82(-2.72, 0.47)
5	-0.51(-0.56,-0.45)∗	0.16(0.07, 0.26)∗	-0.13(-0.24,-0.03)∗		1.25(1.16, 1.34)∗
6	-4.04(-4.21,-3.87)∗		1.06(0.82, 1.30)∗		0.72(0.48, 0.96)∗
7	-3.47(-3.63,-3.33)∗	1.27(1.08, 1.48)∗			-0.93(-1.23,-0.64)∗
8	-4.04(-4.21,-3.87)∗				-0.34(-0.77, 0.05)

asymptotic theory. Hence, only results for CBE are presented for the IPW estimator in Table 2.5. The results of CBE for the missing data model parameters γ are shown in Table 2.6, and suggest that assuming MAR and a correctly specified missing data model, preterm delivery, maternal hypertension and continued HAART treatment are the main variables influencing the missing data process as shown by the exclusion of zero from the 95% credible intervals of their respective parameters γ. In particular, the dependence of the missing data process on the outcome variable preterm delivery suggests that unweighted CC estimates should differ from adjusted estimates. More specifically, the positive posterior median estimates of γ associated with the outcome variable in missing data patterns $R = 3, 5, 7$ suggest that women with preterm delivery are less likely to be observed with complete data.

MICE specifies a univariate imputation model for each of the incomplete variables preterm delivery, maternal hypertension and low CD4^+ (the variable continued HAART treatment is fully observed in the sample and not imputed). The binary variables preterm delivery, hypertension and low CD4^+ are imputed using logistic regressions, to provide a total of $M = 50$ imputed data sets for linear regression before pooling the results in the final analysis. The imputed values for missing variables L in NORM are dichotomized to the binary values $\mathbb{1}(L > 0)$. In a separate analysis, the two sparsely observed missing data patterns $R = 3, 4$ with 75 and 15 samples respectively are combined into one pattern. The probability of observing this combined pattern depends on the set of covariates $L_{(3)} \cap L_{(4)}$, i.e. low CD4^+ and continued HAART treatment.

The IPW and AIPW estimated odds ratio for preterm delivery associated with maternal

hypertension and continued HAART treatment increased by approximately 15% respectively compared to CC estimates. The point estimates of the effect for low CD4^+ are similar between CC and IPW. The observed ARE of AIPW compared to IPW differs across different coefficients: 0.28 for maternal hypertension, 0.53 for low CD4^+ and 0.24 for continued HAART treatment. The observed ARE of AIPW compared to MICE are 1.00 for maternal hypertension, 0.51 for low CD4^+ and 1.25 for continued HAART treatment. The analysis which combines missing data patterns $R = 3, 4$ for IPW/AIPW gives similar results to the original analysis. Point estimates from MICE show that the odds ratio for preterm delivery associated with maternal hypertension increased marginally by about 4%, but the odds ratios associated with low CD4^+ and continued HAART treatment decreased by 8% and 7% respectively. Results from MICE and NORM are similar, although the latter tends to produce smaller standard errors, in agreement with theory and simulation study.

Differences between MICE / NORM and IPW / AIPW estimates may reflect differences of modeling assumptions, since the former relies on model assumptions about full data univariate conditional or multivariate laws while the latter relies on a model for the missing data mechanism. In the current application, neither the conditional distribution of covariates in the full data nor the missing data model is of primary scientific interest. Although model compatibility of the conditional laws specified in MICE may be an issue (White et al., 2011; van Buuren, 2007), simulation studies suggest that this may not be a serious problem in certain practical settings (van Buuren et al., 2006). In general, more efficient estimators can be obtained by specifying a full data model, and the NORM estimates indeed have the smallest standard errors among the methods being compared. However, in this particular application, the proposed AIPW estimator produces standard errors which are comparable to those of MICE, while at the same time entirely avoiding the need to model the full data law. This is in agreement with simulation study results which show similar efficiency between the AIPW and MICE estimators.

2.7 Discussion

We have proposed a simple yet general class of missing data models for nonmonotone MAR mechanisms which makes no assumption about the full data distribution. Our models are explicit in their dependence on only the observed variables, and the proposed IPW estimator can easily be implemented using existing software. The paper makes two important contributions, first we describe a simple UMLE approach to estimate the missing data mechanism that is straightforward to implement although that may suffer from convergence issues in small samples. Our second contribution offers a remedy to failure of UMLE by introducing a constrained Bayesian estimator which circumvents any potential convergence difficulty encountered with UMLE. Another contribution shows that AIPW can achieve substantial gains in efficiency over simple IPW estimators by recovering information from incomplete cases, while avoiding having to model the full data distribution. Assuming no model misspecification, the proposed IPW / AIPW estimators corrects the bias of CC analysis and may be used whenever one has available a full data estimating equation and the nonmonotone MAR missing data mechanism potentially depends also on the outcome. The constrained Bayesian estimator is guaranteed to produce valid probability weights for subsequent estimation of a full data regression or other functionals of interest. In addition, constrained Bayesian estimation of the missing data model parameters is able to elucidate important variables that influence the missingness process by studying the properties of the Monte Carlo approximations to their posterior distributions (e.g. posterior medians and 95% credible intervals, as illustrated in the application). Constrained Bayesian estimation under a parametric model for the missing data process also allows for sensitivity analysis under a unified framework to explore the possibility that the process is MNAR, which is part of future work.

Lastly, Robins and Gill have argued that the class of RMM models represents the most general plausible physical mechanism for generating non-monotone missing data (Robins and Gill, 1997). Therefore, they have effectively argued that any model within our class that is not RMM may be difficult to motivate scientifically. We emphasize that the perspective we have presented is completely agnostic as to whether a particular submodel

of MAR may be more scientifically meaningful than another; in fact, RMM, like any other submodel of MAR, can be accommodated by the proposed approach, but would require placing additional constraints while sampling from the posterior, to ensure that one remains within the submodel. This will necessarily result in a more complicated fitting procedure, with little apparent benefit for bias reduction or efficiency gain. This is because, as well established in the missing data literature, it is generally advisable for efficiency considerations in IPW estimation under MAR, that one estimates the probability of a complete-case using as richly parameterized a regression as empirically feasible (Robins et al., 1994). This implies that even if RMM is correctly specified, one would generally benefit from including correlates of the full data estimating equation into a model for the missing data mechanism, even if such variables do not necessarily correlate with the missing data process. We believe such efficiency considerations trump any concern for scientific interpretation of the model for the missing data process, particularly since after all, the missing data process is technically a nuisance parameter not of primary scientific interest.

A Multinomial Regression Approach to Model Outcome Heterogeneity

BaoLuo Sun[1], Tyler VanderWeele[1,2] and Eric J. Tchetgen Tchetgen[1,2]

Departments of Biostatistics[1] and Epidemiology[2], Harvard School of Public Health

3.1 Introduction

Categorical outcomes are of common occurrence in epidemiologic practice. A standard modeling approach to evaluate risk factors in such settings involves fitting by maximum likelihood, a polytomous logistic regression for the multinomial outcome (Agresti, 2002). In empirical studies, an important form of outcome heterogeneity arises when a given risk factor affects certain categories of the outcome but not necessarily others. This form of outcome heterogeneity, also sometimes called *etiologic heterogeneity* (Begg and Zabor, 2012), has in recent years drawn considerable interest in medicine and other health sciences (Troester and Swift-Scanlan, 2009; Begg et al., 2014; Wang et al., 2015). In this paper, we establish that standard polytomous logistic regression is often ill-suited to model this type of outcome heterogeneity, in the sense that the approach may understate the degree to which such heterogeneity may be present. Specifically, standard polytomous regression will often a priori rule out the possibility of outcome heterogeneity from its parameter space, because under the model a risk factor for a given category of the outcome must necessarily be a risk factor for all other categories of the outcome. In the following sections we demonstrate how this phenomenon is manifested with a certain paradox that arises in the context of using polytomous logistic regression in the presence of outcome heterogeneity, and propose an alternative general multinomial regression approach with constrained Bayesian estimation of the regression parameters. We investigate the finite-sample properties of the proposed estimators in a simulation study and illustrate the new methodology in an application evaluating risk factors for death from cardiovascular heart disease (CHD), stroke and cancer in the original cohort of the Framingham Heart Study.

3.2 Methods

3.2.1 A paradox from using polytomous logistic regression

To describe the paradox, suppose that the outcome Y takes one of three possible values $k = 0, 1, 2$, where $Y = 0$ denotes disease-free persons, whereas $Y = 1$ denotes individuals with the given disease of the first subtype, while $Y = 2$ denotes diseased persons with

the second subtype. Also suppose that two continuous risk factors (X_1, X_2) are known to be associated with diseased individuals, i.e.

$$\Pr\{Y \neq 0 | x_1, x_2\} = \pi_0(x_1, x_2) \tag{3.1}$$

varies both in (x_1, x_2), where x_j denotes a possible value of X_j. A standard approach to model such data entails positing a polytomous logistic regression, such as say

$$\Pr\{Y = k|X\} = \frac{\exp\{\beta_{k0} + \beta_{k1}X_1 + \beta_{k2}X_2\}}{1 + \exp\{\beta_{10} + \beta_{11}X_1 + \beta_{12}X_2\} + \exp\{\beta_{20} + \beta_{21}X_1 + \beta_{22}X_2\}}, \quad k = 1, 2 \tag{3.2}$$

where $X = (1, X_1, X_2)$, and

$$\Pr\{Y = 0|X\} = 1 - \Pr\{Y = 1|X\} - \Pr\{Y = 2|X\}. \tag{3.3}$$

Now, suppose also that, reflecting the presence of outcome heterogeneity, the first risk factor X_1 only affects the first disease subtype, while X_2 only affects the second disease subtype, i.e.

$$\Pr\{Y = 1|x_1, x_2\} = \pi_1(x_1), \text{ for all } x_2 \text{ and each } x_1 \tag{3.4}$$

and

$$\Pr\{Y = 2|x_1, x_2\} = \pi_2(x_2), \text{ for all } x_1 \text{ and each } x_2. \tag{3.5}$$

Then, for equation 3.4 to hold under the polytomous regression model, it must be that

$$\beta_{12} = \beta_{22} = 0, \tag{3.6}$$

so that the right-hand side of equation $3.2)$ does not depend on X_2. Likewise, for equation 3.5 to hold under the polytomous regression model, it must be that

$$\beta_{11} = \beta_{21} = 0, \tag{3.7}$$

so that the right-hand side of equation 3.2 does not depend on X_2. However, both equa-

tions 3.6 and 3.7 would imply that

$$\Pr\{Y \neq 0|X\} = \frac{\exp\{\beta_{10}\} + \exp\{\beta_{20}\}}{1 + \exp\{\beta_{10}\} + \exp\{\beta_{20}\}}$$

depends neither on X_1 nor on X_2, which contradicts the fact that X is a risk factor for Y as given by equation 3.1, giving rise to the paradox.

The above paradox stems from the fact that a standard polytomous logistic regression of the form given in expression 3.2 cannot simultaneously encode assumption 3.2, and assumptions 3.4 and 3.5. This is because such models do not have a specific parameter or set of parameters which can be set to a value that solely implies either assumption 3.4 or 3.5, without also implying that Y is altogether independent of either X_2 or X_1, respectively. Note that incorporating interactions and nonlinearities in X_1 and X_2 would in principle make the regression model somewhat more flexible, however, this would not necessarily resolve the above paradox, unless a genuine nonparametric model were used in place of a parametric model. Even under a nonparametric polytomous regression framework, it is unclear whether one could easily encode assumption 3.4. Note also that this form of paradox will become even more ubiquitous when multiple risk factors are being considered, in which case nonparametric regression may no longer be practical. We may conclude that polytomous logistic regression is generally ill-suited to either detect or model outcome heterogeneity of the type described above. In the next section, we describe a simple alternative approach which circumvents this difficulty. Before doing so, we briefly note that in the special case where the outcome is rare for all levels of X, the paradox may not be as relevant since expression 3.2 can then be approximated by

$$\Pr\{Y = k|X\} \approx \exp\{\beta_{k0} + \beta_{k1}X_1 + \beta_{k2}X_2\}. \tag{3.8}$$

3.2.2 A general multinomial regression approach to model outcome heterogeneity

The proposed approach involves modeling each category of the outcome (other than a reference level) with a separate binary regression model. To fix ideas, let us reconsider the example from the previous section. Suppose that instead of equation 3.2, one posits the following pair of logistic regressions

$$\text{logit}\,\Pr\{Y = 1|x_1, x_2\} = \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 \tag{3.9}$$

$$\text{logit}\,\Pr\{Y = 2|x_1, x_2\} = \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 \tag{3.10}$$

and as before $\Pr\{Y = 0|x_1, x_2\}$ is given by equation 3.3. The intercept β_{k0} may be interpreted as the log-odds that $Y = k$, $k = 1, 2$ when $x_1 = x_2 = 0$. The regression coefficient β_{kj} corresponds to a difference in the log-odds of the event $I\{Y = k\}$ versus its complement $I\{Y \neq k\}$ per unit increment in X_j conditional on the value of the other covariate, i.e. β_{kj} captures the association between X_j and the risk of disease subtype k. The degree of outcome heterogeneity as it relates to X_1 is therefore measured by the difference in the regression coefficients β_{11} and β_{21}, the associations of X_1 with disease subtype 1 and 2, respectively. Likewise, the degree of outcome heterogeneity as it relates to X_2 can be captured by comparing β_{12} and β_{22}. Notably, the hypothesis corresponding to equations 3.4 and 3.5 is readily encoded without imposing further restriction by setting $\beta_{12} = \beta_{21} = 0$.
In contrast to the interpretation of the parameters in separate logistic models 3.9 and 3.10, β_{kj} in polytomous regression 3.2 corresponds to a difference in the log-odds of the event $I\{Y = k\}$ versus the baseline event $I\{Y = 0\}$ per unit increment in X_j while holding the value of the other covariate constant, i.e. the referent event is different. However, β_{kj} in the two models should be approximately the same when all the outcome types are rare compared to the baseline level $Y = 0$, for all values of (x_1, x_2), since in this special case both models can be approximated by 3.8.
For inference, one could in principle estimate $\beta_k = (\beta_{k0}, \beta_{k1}, \beta_{k2})'$ by separately maximizing the likelihood function for the corresponding logistic regression in equations 3.9 and

3.10, with binary outcome $I\{Y = k\}$. However, such a strategy has two potential limitations that make it unattractive. First, the approach is potentially inefficient, since it does not respect the multinomial nature of Y and therefore, does not make use of all available information in estimating β_k separately. A second concern, is that although the logit link function in equations 3.9 and 3.10 guarantees that the resulting estimate of the predicted probability $\Pr\{Y = k|X_1, X_2\}$ for each person in the sample, falls within the unit interval $(0, 1)$, it does not ensure that the resulting estimate of $\Pr\{Y = 0|X_1, X_2\}$ given by equation 3.3 also falls within the unit interval.

In order to resolve these limitations, we propose that the collection of regression parameters be jointly estimated using the following constrained Bayesian approach, which ensures model coherence, and maximizes efficiency. The approach basically entails specifying a prior distribution $\pi(\beta)$ for the vector of unknown parameters $\beta = (\beta_1', \beta_2')'$, which combined with the observed data likelihood, gives rise to a posterior distribution proportional to

$$\pi(\beta) \prod_i f(Y_i|X_i;\beta)\, I\{\Pr\{Y = 1|X_i;\beta_1\} + \Pr\{Y = 2|X_i;\beta_2\} < 1\} \tag{3.11}$$

where $f(k|X_i;\beta) = \Pr\{Y = k|X_i;\beta\}$, and the indicator function ensures that posterior samples are restricted to values of β for which the multinomial model is well defined, i.e. $0 < \Pr\{Y = 0|X_i;\beta\} = 1 - \Pr\{Y = 1|X_i;\beta_1\} - \Pr\{Y = 2|X_i;\beta_2\} < 1$. The posterior mode (or mean) provides an efficient estimate of β and 95% credible intervals can likewise be obtained from the resulting posterior sample. For posterior computation, we may specify the diffuse priors $\beta_{kj} \sim \mathrm{N}(0, 10^2)$. Adaptive Gibbs sampling (Gilks et al., 1995) may be implemented through BRugs, the R interface to the OpenBUGS MCMC software (Lunn et al., 2009). Sample OpenBUGS code for posterior estimation in the simulation study is included in the Appendix. One may then assess convergence by visually inspecting the trace plots as well as through the Gelman-Rubin convergence statistic (Gelman and Rubin, 1992).

The approach is easily extended to handle a multinomial outcome $K > 3$ levels. As

before, we simply define $K-1$ logistic regression models

$$\text{logit}\Pr\{Y=k|x\}=\beta_k'X,\ k=1,...,K-1$$

where X is a vector of J risk factors, with first component set to 1 for the intercept. The density $\pi(\beta)$ is again a diffuse prior for $\beta=\{\beta_{jk}:j=1,...,J;k=1,...,K-1\}$. The posterior distribution for the general case is proportional to

$$\pi(\beta)\prod_i f(Y_i|X_i;\beta)\,I\left\{\sum_{k>0}\Pr\{Y=k|X_i;\beta_k\}<1\right\}$$

where the indicator function constrains the posterior sampling space so that $0<\Pr\{Y=0|X_i;\beta\}<1$. Implementation of the approach is as described above.

3.3 Simulation

This section reports a simulation study to evaluate the finite-sample properties of the proposed constrained Bayesian estimator compared with the polytomous and separate logistic estimators. Full data consists of n independent and identically distributed (Y_i,X_{1i},X_{2i}), $i=1,...,n$ where Y denotes the categorical outcome and (X_1,X_2) the two risk factors. The vector (Z_1,Z_2) is generated from a bivariate standard normal distribution with correlation coefficient $\rho=-0.3$ and $X_1=\Phi(Z_1),X_2=\Phi(Z_2)$ where $\Phi(\cdot)$ is the CDF of the standard normal distribution. The categorical outcome is generated as

$$\begin{aligned}
\Pr(Y=1)&=\{1+\exp[-(\beta_{10}+\beta_{11}X_1+\beta_{12}X_2)]\}^{-1}\\
\Pr(Y=2)&=\{1+\exp[-(\beta_{20}+\beta_{21}X_1+\beta_{22}X_2)]\}^{-1}\\
\Pr(Y=3)&=\{1+\exp[-(\beta_{30}+\beta_{31}X_1+\beta_{32}X_2)]\}^{-1}\\
\Pr(Y=0)&=1-\Pr(Y=1)-\Pr(Y=2)-\Pr(Y=3),
\end{aligned}$$

with true parameter values $(\beta_{10},\beta_{11},\beta_{12})=(-1.1,0.3,0.0)$, $(\beta_{20},\beta_{21},\beta_{22})=(-0.9,0.0,-0.4)$ and $(\beta_{30},\beta_{31},\beta_{32})=(-1.1,0.3,-0.3)$ for $n=200,500$ with 1000 simulation replicates at each sample size. Table 3.1 shows the results of polytomous logistic

regression based on the model

$$\Pr\{Y = k|X\} = \frac{\exp\{\alpha_{k0} + \alpha_{k1}X_1 + \alpha_{k2}X_2\}}{1 + \sum_{j=1}^{3}\exp\{\alpha_{j0} + \alpha_{j1}X_1 + \alpha_{j2}X_2\}} \qquad (3.12)$$
$$k = 1, 2, 3$$

where $X = (1, X_1, X_2)$ and $Y = 0$ is the referent level. Separate logistic (SL) regression estimates are based on the model

$$\text{logit}\Pr\{Y = k|X\} = \gamma_{k0} + \gamma_{k1}X_1 + \gamma_{k2}X_2, \qquad (3.13)$$
$$k = 1, 2, 3$$

while the constrained Bayesian (CB) estimates are the Monte Carlo mean values of the posterior distribution which is proportional to

$$\pi(\eta)\prod_{i=1}^{n}\left\{\prod_{k=0}^{3}\{\Pr\{Y_i = k|X_i;\eta_k\}\}^{I(Y_i=k)}\, I\left(\sum_{j=1}^{3}\Pr\{Y_i = j|X_i;\eta_j\} < 1\right)\right\} \qquad (3.14)$$

where

$$\text{logit}\Pr\{Y_i = k|X_i;\eta_k\} = \eta_{k0} + \eta_{k1}X_1 + \eta_{k2}X_2,$$
$$k = 1, 2, 3$$

The OpenBUGS code for fitting model 3.14 can be found in the Appendix. The results for SL and CB analyses are included in Table 3.2.

The simulation results from polytomous logistic regression show that the mean estimated odds ratios for covariates X_1 and X_2 differ from one (OR $\neq 1.0$) across each of the three comparison groups. Based on model 3.12, this implies that (X_1, X_2) are risk factors for each of the three levels of outcome in Y, and therefore appears to contradict the outcome heterogeneity of risk factors (X_1, X_2) with Y under the true model. In order to be a coherent model for outcome heterogeneity in the present data generating mechanism, polytomous logistic regression must depend neither on X_1 nor X_2 (OR $= 1.0$), as argued in the paradox previously described. Model 3.12 is therefore a misspecified model for the full data law incorporating outcome heterogeneity, and the odds ratio estimates suggest

Table 3.1: Simulation Results Based on Polytomous Logistic Regression.

		$Y = 1$ versus $Y = 0$			$Y = 2$ versus $Y = 0$			$Y = 3$ versus $Y = 0$		
n	Variable	OR	RMSE[a]	MCSE	OR	RMSE	MCSE	OR	RMSE	MCSE
200	Intercept	1.49	1.26	1.11	3.05	3.42	2.97	0.75	0.79	0.71
	X_1	1.79	1.55	1.40	2.48	2.78	2.42	0.53	0.55	0.48
	X_2	1.46	1.31	1.17	3.14	3.71	3.25	0.61	0.66	0.57
500	Intercept	1.32	0.58	0.54	2.44	1.32	1.28	0.61	0.33	0.33
	X_1	1.54	0.69	0.66	1.97	1.08	0.99	0.45	0.25	0.25
	X_2	1.35	0.61	0.61	2.42	1.32	1.22	0.49	0.27	0.26

Abbreviations: OR, mean estimated odds ratio; RMSE, square root of mean estimated variance; MCSE, Monte-Carlo standard error.
[a] Estimated variance of odds ratio is derived from estimated variance of coefficient estimates using the delta method.

that it is unable to differentiate between risk factors that influence a particular outcome category and those risk factors that do not. Estimates of variance tend to be conservative compared to the empirical variance in finite samples.

The odds ratio estimates from SL and CB regressions are consistent for the true odds ratios with vanishing biases as sample size increases. The CB regression estimator appears to be slightly less biased than the SL estimator in finite samples. The variance estimator based on the posterior distribution in the CB approach performs well, while those for SL tend to be conservative in finite samples. In addition, the asymptotic relative efficiency comparing SL and CB estimates (i.e. $\text{Var}(\widehat{OR}_{CB})/\text{Var}(\widehat{OR}_{SL})$) varies between 0.59–0.88 and 0.81–0.90 for sample sizes $n = 200$ and $n = 500$ respectively. This is in agreement with theory since the CB estimation incorporates all available information from the data by simultaneously estimating all parameters.

Even though 3.13 is the correct model, the SL estimate $\widehat{\Pr}(Y = 0|X_1, X_2)$ is negative for at least one sample in 11.2% and 0.15% of the simulation replicates when $n = 200$ and $n = 500$ respectively. Therefore fitted probabilities for the reference outcome sometimes do not lie in the unit interval, which occurs despite the absence of model misspecification. Estimation with the proposed CB approach ensures that estimates $\widehat{\Pr}(Y = 0|X_1, X_2)$ all lie within the unit interval.

Table 3.2: Simulation Results Based on Separate Logistic and Constrained Bayesian Regressions.

			$Y = 1$ versus $Y \neq 1$			$Y = 2$ versus $Y \neq 2$			$Y = 3$ versus $Y \neq 3$		
n	Method	Variable	$\lvert\text{Bias}\rvert_{OR}$	RMSE[a]	MCSE	$\lvert\text{Bias}\rvert_{OR}$	RMSE	MCSE	$\lvert\text{Bias}\rvert_{OR}$	RMSE	MCSE
200	SL	Intercept	0.03	0.20	0.19	0.30	1.22	1.17	0.19	0.85	0.80
		X_1	0.06	0.26	0.25	0.22	0.95	0.93	0.11	0.58	0.55
		X_2	0.02	0.20	0.19	0.30	1.26	1.18	0.19	0.68	0.62
	CB	Intercept	0.04	0.19	0.19	0.09	0.93	0.93	0.19	0.75	0.76
		X_1	0.02	0.21	0.22	0.21	0.81	0.82	0.15	0.54	0.53
		X_2	0.01	0.18	0.17	0.24	1.06	1.00	0.16	0.60	0.59
500	SL	Intercept	0.01	0.11	0.10	0.11	0.57	0.53	0.07	0.42	0.39
		X_1	0.02	0.14	0.14	0.08	0.44	0.41	0.05	0.29	0.29
		X_2	0.02	0.12	0.12	0.07	0.57	0.53	0.06	0.32	0.32
	CB	Intercept	0.01	0.11	0.10	0.04	0.51	0.50	0.08	0.39	0.40
		X_1	0.01	0.13	0.13	0.08	0.40	0.40	0.06	0.28	0.30
		X_2	0.01	0.11	0.11	0.08	0.54	0.53	0.05	0.30	0.32

Abbreviations: $\lvert\text{Bias}\rvert_{OR}$, mean absolute bias of estimated odds ratio; RMSE, square root of mean estimated variance; MCSE, Monte-Carlo standard error; SL, separate logistic; CB, Constrained Bayesian.
[a] Estimated variance of odds ratio is derived from estimated variance of coefficient estimates using the delta method.

3.4 Empirical Illustration

The empirical application concerns a cohort study of community health in Framingham, Massachusetts (Dawber et al., 1963). Categories of the multinomial outcome Y are different causes of death in the present analysis, with 261 (6.2%) subjects who died from CHD, 164 (3.9%) subjects who died from stroke, 539 (12.9%) subjects who died from cancer and 3218 (76.9%) subjects who survived by the last examination taken in the years between 1979 and 1982. Our goal is to investigate the associations of different risk factors including gender (female coded as 1), age in years, body mass index (BMI), serum cholesterol (mg/100 mL) and high blood pressure (systolic blood pressure >140 mm/Hg or diastolic blood pressure >90 mm/Hg), taken at baseline during the first examination in the years 1948 to 1953, with the separate causes of death. There are 4060 (97%) subjects with complete information on the outcome and risk factors, and 122 (3%) subjects with missing values are excluded from the analysis. The results of polytomous logistic regression are shown in Table 3.3, while the results from separate logistic and constrained Bayesian logistic regressions of Y on the risk factors are shown in Table 3.4.

Table 3.3: Estimated Odds Ratio of Mortality from Various Causes by Risk Factors Based on Polytomous Logistic Regression.

	CHD versus Survival		Stroke versus Survival		Cancer versus Survival	
Variable	OR	95% CI	OR	95% CI	OR	95% CI
Age	1.11	1.09, 1.13∗	1.18	1.15, 1.21∗	1.10	1.09, 1.10∗
Female	0.19	0.14, 0.26∗	0.57	0.41, 0.81∗	0.52	0.43, 0.63∗
Choles.	1.01	1.00, 1.01∗	1.00	0.99, 1.00	1.00	0.99, 1.00
BMI	1.03	0.99, 1.07	1.00	0.96, 1.04	0.99	0.97, 1.02
High BP	2.25	1.02, 1.55∗	2.21	1.55, 3.16∗	1.26	1.02, 1.55∗

Abbreviations: OR, odds ratio; CI, confidence interval; BP, blood pressure.
∗Denotes significance with $P < 0.05$.

The results from polytomous logistic regression suggest that increasing values in age and serum cholesterol, as well as male gender and high blood pressure, are significantly associated with greater risks of death from at least one of the three causes (CHD, stroke and cancer) relative to survival rates by the end of the follow-up period. Only BMI is not significantly associated with the risks of death from any cause. Based on a main effects polytomous logistic model, the results suggest that age, serum cholesterol, gender and high blood pressure are significant risk factors for all causes of death.

Estimation using the separate logistic method suggests that increasing values in age is significantly associated with greater risks of death from CHD, stroke and cancer. On the other hand, the risk factors gender and high blood pressure show more heterogeneity. Being female is significantly associated with lower risk of death from CHD and cancer, but not stroke. High blood pressure is a significant risk factor for greater risk of death from CHD and stroke, but not from cancer. 18 persons have negative estimated probabilities of surviving through the follow-up period under the separate logistic method. Results from the constrained Bayesian and separate logistic methods for age, gender and high blood pressure are similar. The estimated asymptotic relative efficiency of the constrained Bayesian estimator compared to the separate logistic estimator varies between 0.61 to 0.98. More efficient estimation from the constrained Bayesian method identifies serum cholesterol as a statistically significant risk factor for greater risk of death from CHD, but

Table 3.4: Estimated Odds Ratio of Mortality from Various Causes by Risk Factors Based on Separate Logistic and Constrained Bayesian Regressions.

		CHD versus non-CHD		Stroke versus non-Stroke		Cancer versus non-Cancer	
Method	Variable	OR	95% CI	OR	95% CI	OR	95% CI
SL	Age	1.08	1.06, 1.10∗	1.15	1.12, 1.17∗	1.08	1.07, 1.10∗
	Female	0.23	0.17, 0.31∗	0.82	0.58, 1.15	0.63	0.52, 0.76∗
	Choles.	1.01	1.00, 1.01∗	1.00	0.94, 1.00	1.00	0.99, 1.00
	BMI	1.03	0.99, 1.07	1.00	0.96, 1.04	0.99	0.97, 1.02
	High BP	2.03	1.53, 2.69∗	1.92	1.35, 2.73∗	1.08	0.88, 1.33
CB	Age	1.06	1.05, 1.08∗	1.13	1.11, 1.15∗	1.08	1.07, 1.09∗
	Female	0.25	0.19, 0.34∗	0.92	0.67, 1.28	0.66	0.55, 0.80∗
	Choles.	1.01	1.00, 1.01∗	0.99	0.99, 1.00∗	0.99	0.99, 1.00∗
	BMI	1.03	1.01, 1.06∗	0.99	0.96, 1.03	1.00	0.98, 1.02
	High BP	1.88	1.43, 2.48∗	1.78	1.26, 2.52∗	1.05	0.86, 1.29

Abbreviations: SL, separate logistic; CB, constrained Bayesian; OR, odds ratio; CI, confidence interval (or credible interval for constrained Bayesian regression); BP, blood pressure.
∗Denotes significance with $P < 0.05$ or exclusion of one from 95% credible interval.

is paradoxically significantly associated with lower risks of death from stroke and cancer. These apparent "protective" associations could be essentially due to competing risk from death by CHD. Higher BMI is found to be significantly associated with death from CHD, but not from stroke or cancer.

Using constrained Bayesian estimation, it appears that high blood pressure is associated with increased CHD and Stroke mortality but not cancer, whereas outcome heterogeneity is entirely understated by polytmous logistic regression. Likewise, using constrained Bayesian estimation, being a female is associated with lower CHD and cancer mortality but not stroke, another level of outcome heterogeneity undetected by polytomous logistic regression. We see then that the problem described in this paper with polytomous logistic regression is not simply theoretical; it can and does arise in practice. Continued use of this standard approach might perpetuate lack of detection of scientifically relevant outcome heterogeneity in epidemiological practice.

3.5 Discussion

Polytomous regression is the standard approach in the analysis of data from clinical or observational studies with polytomous outcome. However, a peculiar feature of this model is that its parameterization cannot encode or detect simple outcome heterogeneity, whereby certain risk factors contribute exclusively to the occurrence of some outcomes, but not others. We propose an alternative approach to polytomous logistic regression, which involves modeling each category of the outcome (other than a reference level) with a separate binary regression model. By doing so, our multinomial regression readily encodes a broad range of outcome heterogeneity of practical interest. In order to ensure coherent inferences and maximize efficiency, the collection of regression parameters are jointly estimated, which is straightforward to implement in standard software for Bayesian estimation. The constrained Bayesian approach should form a part of the standard statistical methods for assessing outcome heterogeneity.

References

AGRESTI, A. (2002). *Categorical Data Analysis*. Wiley.

ANDRIDGE, R. R. and LITTLE, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review* **78(1)** 40–64.

BEGG, C., SESHAN, V., ZABOR, E., FURBERG, H., ARORA, A., SHEN, R., MARANCHIE, J., NIELSEN, M., RATHMELL, W., SIGNORETTI, S., TAMBOLI, P., KARAM, J., CHOUEIRI, T., HAKIMI, A. and HSIEH, J. (2014). Genomic investigation of etiologic heterogeneity: methodologic challenges. *BMC Medical Research Methodology* **14** 138.

BEGG, C. B. and ZABOR, E. C. (2012). Detecting and exploiting etiologic heterogeneity in epidemiologic studies. *American Journal of Epidemiology* .

BICKEL, P., KLAASSEN, C. A., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins series in the mathematical sciences, Johns Hopkins University Press.

CHAUSSÉ, P. (2010). Computing generalized method of moments and generalized empirical likelihood with R. *Journal of Statistical Software* **34** 1–35.
URL `http://www.jstatsoft.org/v34/i11/`

CHEN, H. Y. (2007). A semiparametric odds ratio model for measuring association. *Biometrics* **63** 413–421.

CHEN, J. Y., RIBAUDO, H. J., SOUDA, S., PAREKH, N., OGWU, A., LOCKMAN, S., POWIS, K., DRYDEN-PETERSON, S., CREEK, T., JIMBO, W., MADIDIMALO, T., MAKHEMA, J.,

ESSEX, M. and SHAPIRO, R. L. (2012). Highly active antiretroviral therapy and adverse birth outcomes among hiv-infected women in botswana. *The Journal of Infectious Diseases* **206(11)** 1695–1705.

CHU, H. and COLE, S. R. (2010). Estimation of risk ratios in cohort studies with common outcomes: A bayesian approach. *Epidemiology* **21(6)** 855–862.

CHU, H. and COLE, S. R. (2011). Estimating the relative excess risk due to interaction: A bayesian approach. *Epidemiology* **22(2)** 242–248.

DAS, M., NEWEY, W. K. and VELLA, F. (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies* **70(1)** 33–58.

DAWBER, T. R., KANNEL, W. B. and LYELL, L. P. (1963). An approach to longitudinal studies in a community: The framingham study. *Annals of the New York Academy of Sciences* **107** 539–556.

GELFAND, A. E., SMITH, A. F. M. and LEE, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association* **87(418)** 523–532.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7(4)** 457–511.

GILBERT, P. and VARADHAN, R. (2012). *numDeriv: Accurate Numerical Derivatives*. R package version 2012.9-1.
URL `http://CRAN.R-project.org/package=numDeriv`

GILKS, W., BEST, N. and TAN, K. (1995). Adaptive rejection metropolis sampling within gibbs sampling. *Applied Statistics* **44(4)** 455–472.

GILL, R. D., VAN DER LAAN, M. J. and ROBINS, J. M. (1997). Coarsening at random: Characterizations, conjectures, counter-examples. In *Lecture Notes in Statistics* (D. Lin and T. Fleming, eds.). Springer-Verlag.

HECKMAN, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47(1)** 153–161.

HECKMAN, J. J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources* **32(3)** 441–462.

HORTON, N. J. and LAIRD, N. M. (1998). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research* **8** 37–50.

HORTON, N. J. and LAIRD, N. M. (1999). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research* **8** 37–50.

HORTON, N. J. and LIPSITZ, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician* **55(3)** 244–254.

HORVITZ, D. and THOMPSON, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47(260)** 663–685.

IBRAHIM, J. G. and CHEN, M.-H. (2000). Power prior distributions for regression models. *Statistical Science* **15** 46–60.

IBRAHIM, J. G., CHEN, M.-H. and LIPSITZ, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika* **88** 551–564.

IBRAHIM, J. G., CHEN, M.-H. and LIPSITZ, S. R. (2002). Bayesian methods for generalized linear models with covariates missing at random. *Canadian Journal of Statistics* **30** 55–78.

IBRAHIM, J. G., CHEN, M.-H., LIPSITZ, S. R. and HERRING, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association* **100** 332–346.

KENWARD, M. and CARPENTER, J. (2007a). Multiple imputation: Current perspectives. *Statistical Methods in Medical Research* **16** 199–218.

KENWARD, M. and CARPENTER, J. (2007b). Sensitivity analysis after multiple imputation under missing at random: A weighting approach. *Statistical Methods in Medical Research* **16** 259–275.

LI, L., SHEN, C., LI, X. and ROBINS, J. M. (2013). On weighting approaches for missing data. *Statistical Methods in Medical Research* **22** 14–30.

LIPSITZ, S., IBRAHIM, J. and ZHAO, L. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association* **94** 1147–1160.

LITTLE, R. and DAVID, M. (1983). Weighting adjustments for non-response in panel surveys. *Bureau of the Census technical report* .

LITTLE, R. J. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley.

LITTLE, R. J. and 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.

LUNCEFORD, J. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* **23** 2937–2960.

LUNN, D., SPIEGELHALTER, D., THOMAS, A. and BEST, N. (2009). The bugs project: Evolution, critique and future directions. *Statistics in Medicine* **28(25)** 3049–3067.

MANSKI, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *The Econometrics Journal* **27(3)** 313–333.

MEALLI, F. and RUBIN, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* .

MIAO, W., DING, P. and GENG, Z. (2014). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association* **Submitted**.

MOLENBERGHS, G., FITZMAURICE, G., KENWARD, M., TSIATIS, A. and VERBEKE, G. (2014). *Handbook of Missing Data Methodology*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press.

MOLENBERGHS, G., THIJS, H., JANSEN, I. and BEUNCKENS, C. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics* **5(3)** 445–464.

MORENO-BETANCUR, M. and CHAVANCE, M. (2013). Sensitivity analysis of incomplete longitudinal data departing from the missing at random assumption: Methodology and application in a clinical trial with drop-outs. *Statistical Methods in Medical Research* .

NEUGEBAUER, R. and VAN DER LAAN, M. (2005). Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference* **129** 405–426.

NEWEY, W. and MCFADDEN, D. (1993). Large sample estimation and hypothesis testing. In *Handbook of Econometrics* (D. McFadden and R. Engler, eds.), vol. 4. North-Holland.

NEWEY, W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal* **12(S1)** S217–S229.

NEWEY, W. K., POWELL, J. and WALKER, J. (1990). Semiparametric estimation of selection models: some empirical results. *The American Economic Review* **80(2)** 324–328.

POTTHOFF, R. F., TUDOR, G. E., PIEPER, K. S. and HASSELBLAD, V. (2006). Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical Research* **15** 213–234.

PUHANI, P. (2000). The heckman correction for sample selection and its critique. *Journal of Economic Surveys* **14(1)** 53–68.

ROBINS, J., ROTNITZKY, A. and SCHARFSTEIN, D. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (E. Halloran and D. Berry, eds.). Springer-Verlag.

ROBINS, J. M. and GILL, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine* **16** 39–56.

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.

ROBINS, J. M. and ROTNITZKY, A. (2001). Comment on "inference for semiparametric models: Some questions and an answer". *Statistica Sinica* **11** 920–936.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89(427)** 846–866.

ROTNITZKY, A., ROBINS, J. M. and SCHARFSTEIN, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93** 1321–1339.

ROTNITZKY, A., SCHARFSTEIN, D. O., SU, T. and ROBINS, J. M. (2001). Methods for conducting sensitivity analysis of trials with potentially non-ignorable competing causes of censoring. *Biometrics* **57** 103–113.

ROY, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics* **59** 829–836.

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

RUBIN, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association* **72** 538–543.

RUBIN, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics* **4** 87–94.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

RUBIN, D. B., STERN, H. S. and VEHOVAR, V. (1995). Handling "don't know" survey responses: The case of the slovenian plebiscite. *Journal of the American Statistical Association* **90** 822–828.

SCHAFER, J. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall.

SCHAFER, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research* **8** 3–15.

SCHAFER, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica* **57** 19–35.

SCHAFER, J. L. and GRAHAM, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods* **7** 147–177.

SCHARFSTEIN, D. O., DANIELS, M. J. and ROBINS, J. M. (2003). Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes. *Biostatistics* **4(4)** 495–512.

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association* **94** 1096–1146.

SEAMAN, S., GALATI, J., JACKSON, D. and CARLIN, J. (2013). What is meant by missing at random? *Statist. Sci.* **28** 257–268.

SEAMAN, S. R. and WHITE, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research* **22** 278–295.

SIDDIQUI, O. and ALI, M. W. (1998). A comparison of the random-effects pattern mixture model with last-observation-carried-forward (locf) analysis in longitudinal clinical trials with dropouts. *Journal of Biopharmaceutical Statistics* **8(4)** 545–563.

TCHETGEN TCHETGEN, E. (2009). A simple implementation of doubly robust estimation in logistic regression with covariates missing at random. *Epidemiology* **20(3)** 391–394.

TCHETGEN TCHETGEN, E. J., ROBINS, J. M. and ROTNITZKY, A. (2010). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika* **97** 171–180.

TCHETGEN TCHETGEN, E. J. and WIRTH, K. (2013). A general instrumental variable framework for regression analysis with outcome missing not at random. *Harvard University Biostatistics Working Paper Series* **Working Paper 165**.

TROESTER, M. A. and SWIFT-SCANLAN, T. (2009). Challenges in studying the etiology of breast cancer subtypes. *Breast Cancer Research* **11** 104.

TROXEL, A. B., LIPSITZ, S. R. and HARRINGTON, D. P. (1998). Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. *Biometrika* **85** 661–672.

TSIATIS, A. (2006). *Semiparametric Theory and Missing Data*. Springer.

VAN BUUREN, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* **16** 219–242.

VAN BUUREN, S., BRAND, J., OUDSHOORN, C. and RUBIN, D. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* **76(12)** 1049–1064.

VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software* **45** 1–67.

VAN BUUREN, S. and OUDSHOORN, C. (2000). Multivariate imputation by chained equations: Mice v1.0 users manual. *Leiden: TNO Prevention and Health* .

VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer.

VAN DER VAART, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

VANSTEELANDT, S., ROTNITZKY, A. and ROBINS, J. M. (2007). Estimation of regression models for the mean of repeated outcomes under non-ignorable non-monotone nonresponse. *Biometrika* **94** 841–860.

VARADHAN, R. and GILBERT, P. (2009). BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software* **32** 1–26.

WANG, M., KUCHIBA, A. and OGINO, S. (2015). A meta-regression method for studying etiological heterogeneity across disease subtypes classified by multiple biomarkers. *American Journal of Epidemiology* **182** 263–270.

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.

WHITE, I. R. and CARLIN, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine* **29** 2920–2931.

WHITE, I. R., ROYSTON, P. and WOOD, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* **30(4)** 377–399.

WINSHIP, C. and MARE, R. (1992). Models for sample selection bias. *Annual Review of Sociology* **18** 327–350.

WOOLDRIDGE, J. (2007). Inverse probability weighted m-estimation for general missing data problems. *Journal of Econometrics* **141** 1281–1301.

WOOLDRIDGE, J. (2010). *Economic Analysis of Cross Section and Panel Data*. MIT press.

WU, M. and CARROLL, R. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **44** 175–188.

Appendix A

Proofs of Instrumental Variable Identification and Consistency of Estimators

A.1 Theorem 1

Proof of Theorem 1. The proof is based on contradiction. By the exclusion restriction assumption **(IV.1)** the decomposition of the joint distribution for (Z, Y, R) is

$$P_{\theta_i,\eta_i,\xi_i}(z, y, r) = P_{\theta_i}(r|z, y)P_{\eta_i}(z)P_{\xi_i}(y), \quad i = 1, 2, ..., n$$

The only quantities we can identify from the observed data are the joint distribution $P(z, y, R = 1)$ and IV distribution $P(z)$. Suppose we have two sets of candidates satisfying the same observed quantities:

$$P_{\theta_1}(z, y, R = 1) = P_{\theta_2}(z, y, R = 1)$$
$$P_{\eta_1}(z) = P_{\eta_2}(z)$$

Substituting the above observed quantities into the joint distribution gives

$$P_{\theta_1}(R = 1|z, y)P_{\xi_1}(y) = P_{\theta_2}(R = 1|z, y)P_{\xi_2}(y)$$

or equivalently

$$\frac{P_{\theta_1}(R = 1|z, y)}{P_{\theta_2}(R = 1|z, y)} = \frac{P_{\xi_2}(y)}{P_{\xi_1}(y)}$$

This contradicts with the requirement that the ratios are unequal. Therefore the condition that the ratios are unequal is equivalent to ruling out the possiblity that we can have two sets of candidates satisfying the same observed quantities. □

A.2 Examples 1-3

Proof of Example 1. For binary outcome Y and binary instrument Z, let $P(R = 1|Z, Y; \theta) = \text{expit}\,[\theta_0 + \theta_1 Z + \theta_2 Y + \theta_3 ZY]$ and $P(Y = 1; \xi) = \exp(\xi)$. We show that for every (θ, ξ), there exists $(\tilde{\theta}, \tilde{\xi}) \neq (\theta, \xi)$ such that

$$\frac{P(R = 1|Z, Y; \theta)}{P(R = 1|Z, Y; \tilde{\theta})} = \frac{P(Y; \tilde{\xi})}{P(Y; \xi)} \tag{A}$$

Let $\frac{P(Y=0;\tilde{\xi})}{P(Y=0;\xi)} = \exp(\rho_0)$ for some $\rho_0 \neq 0$, then $\frac{P(Y;\tilde{\xi})}{P(Y;\xi)} = \exp(\rho_0 + \rho_1 Y)$ where

$$\rho_1 = \log\left\{\exp(-\rho_0 - \xi) + [\exp(\xi) - 1]/\exp(\xi)\right\}.$$

Equality (A) then holds by choosing $(\tilde{\theta}, \tilde{\xi})$ such that

$$\begin{aligned}
\tilde{\theta}_0 &= \theta_0 - \rho_0 - \log(\alpha_0) \\
\tilde{\theta}_1 &= \theta_1 + \log(\alpha_0) - \log(\alpha_1) \\
\tilde{\theta}_2 &= \theta_2 - \rho_1 + \log(\alpha_0) - \log(\alpha_2) \\
\tilde{\theta}_3 &= \theta_3 + \log(\alpha_1) + \log(\alpha_2) - \log(\alpha_0) - \log(\alpha_3) \\
\tilde{\xi} &= \xi + \rho_0 + \rho_1,
\end{aligned}$$

where $\alpha_0 = 1+\exp(\theta_0)-\exp(\theta_0-\rho_0)$, $\alpha_1 = 1+\exp(\theta_0+\theta_1)-\exp(\theta_0+\theta_1-\rho_0)$, $\alpha_2 = 1+\exp(\theta_0+\theta_2)-\exp(\theta_0+\theta_2-\rho_0-\rho_1)$ and $\alpha_3 = 1+\exp(\theta_0+\theta_1+\theta_2+\theta_3)-\exp(\theta_0+\theta_1+\theta_2+\theta_3-\rho_0-\rho_1)$. For example, choose $(\rho_0, \rho_1) = (0.3, -0.38)$ and equality (A) holds for $(\theta_0, \theta_1, \theta_2, \theta_3, \xi) = (0.3, 0.6, 0.1, 0.7, -0.2)$ and $(\tilde{\theta}_0, \tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3, \tilde{\xi}) = (-0.3, 0.41, 0.91, 1.37, -0.28)$.

Next, we consider the missingness mechanism $P(R = 1|Z, Y; \theta) = \text{expit}\,[\theta_0 + \theta_1 Z + \theta_2 Y]$, where the interaction effect between (Z, Y) is absent. Under this mechanism, we have

$\theta_3 = \tilde{\theta}_3 = 0$ and therefore $\alpha_1\alpha_2 = \alpha_0\alpha_3$ which implies the equality

$$\exp(\rho_0 + \rho_1) = \frac{\exp(\theta_2 + \rho_0)}{\exp(\theta_2 + \rho_0) + [1 - \exp(\rho_0)]}. \tag{B}$$

Since $\exp(\rho_0 + \rho_1 Y)$ is the ratio of the two probability mass distributions for Y, ρ_0 and $\rho_0 + \rho_1$ should be of opposite signs. Based on (B), if $\exp(\rho_0) > 1$ then $\exp(\rho_0 + \rho_1) > 1$ and similarly if $\exp(\rho_0) < 1$ then $\exp(\rho_0 + \rho_1) < 1$, which implies that the only possibility is $\rho_0 = \rho_1 = 0$ and hence $(\tilde{\theta}, \tilde{\xi}) = (\theta, \xi)$. □

Proof of Example 2. Consider the case where Z and Y are both continuous random variables. Suppose two sets of candidates in the separable logistic missing data mechanism has the following relationship

$$\frac{\text{expit}(q_1(z) + h_1(y))}{\text{expit}(q_2(z) + h_2(y))} = g(y)$$

for some function $g(\cdot)$, i.e. the ratio is a function of y only. Taking derivative with respect to Z on both sides (assuming IV relevance **(IV.2)** holds) gives

$$\frac{\frac{\partial}{\partial z}\text{expit}(q_1(z) + h_1(y))}{\text{expit}(q_1(z) + h_1(y))} = \frac{\frac{\partial}{\partial z}\text{expit}(q_2(z) + h_2(y))}{\text{expit}(q_2(z) + h_2(y))}$$

or equivalently

$$\frac{\partial q_1(z)/\partial z}{\partial q_2(z)/\partial z} = \frac{1 + \exp(q_1(z) + h_1(y))}{1 + \exp(q_2(z) + h_2(y))} \tag{A}$$

Taking derivatives with respect to Y on both sides leads to

$$\frac{\partial q_1(z)/\partial z}{\partial q_2(z)/\partial z} \exp(q_2(z) - q_1(z)) = \frac{\partial h_1(y)/\partial y}{\partial h_2(y)/\partial y} \exp(h_1(y) - h_2(y))$$

The left hand side of the above equation depends only on Z but the right hand side depends only on Y, so it must be that

$$\frac{\partial q_1(z)/\partial z}{\partial q_2(z)/\partial z} \exp(q_2(z) - q_1(z)) = c_1$$

for some constant c_1. Substituting the above expression into equality (A) leads to

$$c_1 \left\{\exp(-q_2(z)) + \exp(h_2(y))\right\} = \exp(-q_1(z)) + \exp(h_1(y))$$

and therefore

$$c_1 \exp(-q_2(z)) + c_2 = \exp((-q_1(z)), \quad c_1 \exp(h_2(y)) - c_2 = \exp((h_1(y))$$

for some constant c_2. Substituting the above equalities into the ratio of propensity scores

$$\frac{\text{expit}(q_1(z) + h_1(y))}{\text{expit}(q_2(z) + h_2(y))} = 1 + c_2 \exp(-h_1(y)) = g(y)$$

Note that $g(y)$ is the ratio of two candidate densities of Y, and so it must be that $c_2 = 0$ and the two sets of candidates are equivalent, leading to a contradiction. Therefore the ratio

$$\frac{\text{expit}(q_1(z) + h_1(y))}{\text{expit}(q_2(z) + h_2(y))}$$

is either a constant or depends on z, which by Corollary 1 leads to identifiability of this class of missing data models.

Consider the case where Z is a binary random variable, and assume two sets of candidates in the separable logistic missing data mechanism has the following relationship

$$\frac{\text{expit}(\eta_1 z + h_1(y))}{\text{expit}(\eta_2 z + h_2(y))} = g(y).$$

The above relationship holds for $z = 0, 1$, therefore

$$\frac{\text{expit}(h_1(y))}{\text{expit}(h_2(y))} = \frac{\text{expit}(\eta_1 + h_1(y))}{\text{expit}(\eta_2 + h_2(y))}$$

and

$$g(y) = 1 + \frac{\exp(\eta_2) - \exp(\eta_1)}{\exp(\eta_2) - \exp(\eta_1 + \eta_2)} \exp[-h_2(y)].$$

Since $g(y)$ is the ratio of two densities, we must have $\eta_1 = \eta_2$ and $g(y) = 1$, leading to a contradiction. The proof for Y or Z as discrete variables is similar to the above proof for binary Z. □

Proof of Example 3. Suppose two set of candidates in the separable probit missing data mechanism has the following relationship

$$\frac{\Phi(q_1(z) + h_1(y))}{\Phi(q_2(z) + h_2(y))} = g(y)$$

for some function $g(\cdot)$, i.e. the ratio is a function of y only. Taking derivatives with respect to Z on both sides (assuming inclusion restriction **(IV.2)** holds) gives

$$\frac{\frac{\partial q_1(z)}{\partial z}\phi(q_1(z)+h_1(y))}{\Phi(q_1(z)+h_1(y))} = \frac{\frac{\partial q_2(z)}{\partial z}\phi(q_2(z)+h_2(y))}{\Phi(q_2(z)+h_2(y))}$$

or equivalently

$$\frac{\phi(q_1(z)+h_1(y))}{\phi(q_2(z)+h_2(y))} = \frac{\frac{\partial q_2(z)}{\partial z}}{\frac{\partial q_1(z)}{\partial z}}g(y)$$

$$\{q_2(z)+h_2(y)\}^2 - \{q_1(z)+h_1(y)\}^2 = 2\left\{\log\frac{\frac{\partial q_2(z)}{\partial z}}{\frac{\partial q_1(z)}{\partial z}} + \log g(y)\right\} \tag{A}$$

The right hand side of the above equation does not include any interaction term between z and y, therefore

$$q_1(z)h_1(y) = q_2(z)h_2(y)$$
$$\frac{q_1(z)}{q_2(z)} = \frac{h_2(y)}{h_1(y)} = c$$

for some constant c. Substitute $q_2(z) = q_1(z)/c$ and $h_2(y) = ch_1(y)$ into equality (A) leads to

$$\left\{\frac{1}{c^2} - 1\right\} q_1^2(z) + \{c^2-1\}h_1^2(y) = 2\{-\log c + \log g(y)\}$$

The right hand side does not depend on z, so $c = 1$ and $q_1(z) = q_2(z)$, $h_1(y) = h_2(y)$, leading to a contradiction. Therefore the ratio

$$\frac{\Phi(q_1(z)+h_1(y))}{\Phi(q_2(z)+h_2(y))}$$

is either a constant or depends on z, which by Corollary 1 leads to identifiability of this class of missing data models.

□

A.3 Propositions 1-4

Proof of Proposition 1. Let $(\eta_0, \omega_0, \xi_0)$ denote the true values of the parameters for parametric models $\eta(x, y, z; \zeta), P(r|Y = 0, x, z; \omega)$ and $q(z|x; \xi)$ which are assumed to be cor-

rectly specified. It is clear that $\hat{\xi}_{\text{MLE}}$ has a probability limit equal to ξ_0. Consider estimating function for (1.11) which under the law of iterated expectations equals to

$$\begin{aligned} & E\left\{E\left\{\left[\frac{R}{\pi(\zeta_0,\omega_0)}-1\right]\boldsymbol{h}_1(X,Z)\right\}\middle|X,Y,Z\right\} \\ =& E\left\{E\left\{\left[\frac{\pi(\zeta_0,\omega_0)}{\pi(\zeta_0,\omega_0)}-1\right]\boldsymbol{h}_1(X,Z)\right\}\right\}=0 \end{aligned}$$

Under the law of iterated expectations, the estimating function for (1.12) equals

$$\begin{aligned} & E\left\{\frac{R}{\pi(\zeta_0,\omega_0)}g(Y,X)\{h_2(Z,X)-E[h_2(Z,X)|X;\xi_0]\}\right\} \\ =& E\left\{g(Y,X)\{h_2(Z,X)-E[h_2(Z,X)|X;\xi_0]\}\right\} \\ =& E\left\{E[g(Y,X)|X]\{h_2(Z,X)-E[h_2(Z,X)|X;\xi_0]\}\right\} \quad \text{by } \textbf{(IV.1)} \\ =& E\left\{E[g(Y,X)|X]\{E[h_2(Z,X)|X;\xi_0]-E[h_2(Z,X)|X;\xi_0]\}\right\} \\ =& 0. \end{aligned}$$

Therefore (η_0,ω_0) are the probability limits of the solutions to estimating equations (1.11) and (1.12). The IPW estimator is also unbiased,

$$E\left\{\frac{RY}{\pi(\zeta_0,\omega_0)}\right\}=E\{Y\}=\phi_0,$$

by taking iterated expectations with respect to (X,Y,Z). The consistency and asymptotic normality of $\hat{\phi}^{\text{IPW}}$ can be established under standard regularity conditions for GMM estimators (Newey and McFadden, 1993) , typically by placing moment restrictions on the vector of estimating functions. In particular, we require that the probability of observing the outcome is bounded away from zero, a necessary assumption for identification of a full data functional (Robins et al., 1994).

$$\pi(x,y,z)>\sigma>0 \quad \text{with probability 1} \tag{A.1}$$

for a non-zero positive constant $\sigma>0$. □

Proof of Proposition 2. Let (η_0,θ_0,ξ_0) denote the true values of the parameters for parametric models $\eta(x,y,z;\zeta)$, $f(y|R=1,x,z;\theta)$ and $q(z|x;\xi)$ which are assumed to be correctly specified. The probability limits of the MLEs $(\hat{\theta}_{\text{MLE}},\hat{\xi}_{\text{MLE}})$ are (θ_0,ξ_0). Under true

parameter values, the expectation of the estimating function for (1.15) is

$$
\begin{aligned}
&E\left\{\{q_1(X,Z)-E\left[q_1(X,Z)|X;\xi_0\right]\}\left\{(1-R)E\left(q_2(X,Y)|R=0,X,Z;\zeta_0,\theta_0\right)+Rq_2(X,Y)\right\}\right\}\\
=&E\{E(\cdot|R=0,X,Z)\times\Pr(R=0|X,Z)\}+E\{E(\cdot|R=1,X,Z)\times\Pr(R=1|X,Z)\}\\
=&E\left(\{q_1(X,Z)-E[q_1(X,Z)|X;\xi_0]\}\,E[q_2(X,Y)|X,Z]\right)\\
=&E\left(\{q_1(X,Z)-E[q_1(X,Z)|X;\xi_0]\}\,E[q_2(X,Y)|X]\right) \qquad \text{by } \textbf{(IV.1)}\\
=&E\left(\{E[q_1(X,Z)|X;\xi_0]-E[[q_1(X,Z)|X;\xi_0]\}\,E[q_2(X,Y)|X]\right)\\
=&0,
\end{aligned}
$$

so that ζ_0 is the probability limit of the solution $\hat{\zeta}$ of (1.15). The OR estimator is unbiased since

$$
\begin{aligned}
&E\left\{RY+(1-R)E(Y|R=0,X,Z;\zeta_0,\theta_0)\right\}\\
=&E\left\{E\{RY+(1-R)E(Y|R=0,X,Z)|R=0,X,Z\}\times\Pr(R=0|X,Z)\right\}\\
&+E\left\{E\{RY+(1-R)E(Y|R=0,X,Z)|R=1,X,Z\}\times\Pr(R=1|X,Z)\right\}\\
=&E\left\{E\{Y|R=0,X,Z\}\times\Pr(R=0|X,Z)\right\}+E\left\{E\{Y|R=1,X,Z\}\times\Pr(R=1|X,Z)\right\}\\
=&E\left\{E\{Y|X,Z\}\right\}\\
=&E\{Y\}=\phi_0.
\end{aligned}
$$

The consistency and asymptotic normality of $\hat{\phi}^{\text{OR}}$ can be established under standard regularity conditions for GMM estimators (Newey and McFadden, 1993) . A necessary condition is that the probability of observing the outcome is bounded away from zero (A.1). □

Proof of Proposition 3. Under model $\mathcal{M}_{\text{IPW}}$, let ξ_0 denote the true value for parametric model $q(z|x;\xi)$ and it is clear that $\hat{\xi}_{\text{MLE}}$ has a probability limit equal to ξ_0. Let superscript asterisks denote possibly misspecified models. Let θ^* denote the probability limit of estimation under model $f^*(y|R=1,x,z;\theta)$ and let $\rho(X,Z)=\int \boldsymbol{u}(x,y)\frac{\exp[-\eta(x,y,z;\zeta)]f(y|R=1,x,z;\theta)}{\int\exp[-\eta(x,y,z)]f^(y|R=1,x,z;\theta)\mathrm{d}\mu(y)}\mathrm{d}\mu(y)$. Then at true parameter values (ζ_0,ω_0),

$$
\begin{aligned}
&E\left\{\boldsymbol{G}^{\text{DR}}\left(R,X,Y,Z;\zeta_0,\omega_0,\theta^*,\boldsymbol{u}\right)|X,Y,Z\right\}\\
=&\boldsymbol{u}(X,Y)-\rho^*(X,Z;\zeta_0,\theta^*)+\rho^*(X,Z;\zeta_0,\theta^*)=\boldsymbol{u}(X,Y),
\end{aligned}
$$

and therefore the estimating function for (1.18), under iterated expectations with respect to (X, Y, Z) at $(\xi_0, \zeta_0, \omega_0)$, is

$$
\begin{aligned}
& E\left\{ \left[\boldsymbol{v}(X,Z) - E\left(\boldsymbol{v}(X,Z)|X\right)\right]\left\{\boldsymbol{u}(X,Y)\right\} \right\} \\
=& E\left\{ \left[\boldsymbol{v}(X,Z) - E\left(\boldsymbol{v}(X,Z)|X\right)\right]\left\{E\left(\boldsymbol{u}(X,Y)|X,Z\right)\right\} \right\} \\
=& E\left\{ \left[\boldsymbol{v}(X,Z) - E\left(\boldsymbol{v}(X,Z)|X\right)\right]\left\{E\left(\boldsymbol{u}(X,Y)|X\right)\right\} \right\} \qquad \text{by } \textbf{(IV.1)} \\
=& E\left\{ \left[E\left(\boldsymbol{v}(X,Z)|X\right) - E\left(\boldsymbol{v}(X,Z)|X\right)\right]\left\{E\left(\boldsymbol{u}(X,Y)|X\right)\right\} \right\} \\
=& \mathbf{0}
\end{aligned}
$$

In addtion, under iterated expectations with respect to (X, Y, Z),

$$E\left\{\boldsymbol{G}^{\text{DR}}\left(R, X, Y, Z, \zeta_0, \omega_0, \theta^*, \boldsymbol{u} = Y\right)\right\} = E\{Y\}.$$

Under model $\mathcal{M}_{\text{OR}}$, let ω^* denote the probability limit of estimation under model $P^*(r|Y = 0, x, z; \omega)$. Then at true parameter values (ζ_0, θ_0),

$$
\begin{aligned}
& E\left\{\boldsymbol{G}^{\text{DR}}\left(R, X, Y, Z; \zeta_0, \omega^*, \theta_0, \boldsymbol{u}\right)|X, Z\right\} \\
=& E\left\{ \frac{R}{\pi(\zeta_0, \omega^*)}\{\boldsymbol{u}(X,Y) - \rho(X,Y)\} + \rho(X,Y) \middle| X, Z \right\} \\
=& E\left\{ \frac{R\{1 - \pi(\zeta_0, \omega^*)\}}{\pi(\zeta_0, \omega^*)}\{\boldsymbol{u}(X,Y) - \rho(X,Y)\} \middle| X, Z \right\} + E\left\{\rho(X,Y) + R\{\boldsymbol{u}(X,Y) - \rho(X,Y)\}|X, Z\right\} \\
=& E\left\{ R\left\{\mathrm{e}^{-\{\lambda(X,Y;\omega^*) + \eta(X,Y,Z;\zeta_0)\}}\right\}\{\boldsymbol{u}(X,Y) - \rho(X,Y)\} \middle| X, Z \right\} + E\left\{\boldsymbol{u}(X,Y)|X, Z\right\} \\
=& \mathrm{e}^{-\lambda(X,Y;\omega^*)}\left\{E[\boldsymbol{u}(X,Y)\mathrm{e}^{-\eta(X,Y,Z;\zeta_0)}|R = 1, X, Z] - E[\boldsymbol{u}(X,Y)\mathrm{e}^{-\eta(X,Y,Z;\zeta_0)}|R = 1, X, Z]\right\} Pr(R = 1|X, Z) \\
& + E\left\{\boldsymbol{u}(X,Y)|X, Z\right\} \\
=& E\left\{\boldsymbol{u}(X,Y)|X, Z\right\} \qquad \text{(S1)}
\end{aligned}
$$

The estimating function for (1.18), under iterated expectations with respect to (X, Z) at

$(\xi_0, \zeta_0, \theta_0)$, is

$$
\begin{aligned}
=&E\left\{ \left[\boldsymbol{v}(X,Z) - E\left(\boldsymbol{v}(X,Z)|X\right)\right]\left\{E\left(\boldsymbol{u}(X,Y)|Z,X\right)\right\} \right\} \\
=&E\left\{ \left[\boldsymbol{v}(X,Z) - E\left(\boldsymbol{v}(X,Z)|X\right)\right]\left\{E\left(\boldsymbol{u}(X,Y)|X\right)\right\} \right\} \qquad \text{by } \textbf{(IV.1)} \\
=&E\left\{ \left[E\left(\boldsymbol{v}(X,Z)|X\right) - E\left(\boldsymbol{v}(X,Z)|X\right)\right]\left\{E\left(\boldsymbol{u}(X,Y)|X\right)\right\} \right\} \\
=&\mathbf{0}.
\end{aligned}
$$

In addition, under iterated expectations with respect to (X, Z) and with similar reasoning given in (S1),

$$E\left\{\boldsymbol{G}^{\text{DR}}\left(R, X, Y, Z, \zeta_0, \omega^*, \theta_0, \boldsymbol{u} = Y\right)\right\} = E\{Y\}.$$

The consistency and asymptotic normality of $\hat{\phi}^{\text{DR}}$ can be established under standard regularity conditions for GMM estimators (Newey and McFadden, 1993) . A necessary condition is that the probability of observing the outcome is bounded away from zero (A.1). □

Proof of Proposition 4. Let $(L, R) = (X, Z, Y, R)$ denote the complete data. Suppose we observe $O = (R, X, Z, YR)$. Furthermore, assume that Z is a valid missing data IV, such that (i) Y is independent of Z given X, and (ii) R given (X, Y, Z) follows a model logit $\Pr\{R = 1|X, Z, Y\} = \alpha_0\left(X, Z\right) + \alpha_y\left(Y, X, Z\right)$ with $\alpha_0\left(X, Z\right)$ unrestricted and $\alpha_y\left(Y, X, Z\right)$ known, and $\alpha_y\left(0, X, Z\right) = 0$. Throughout, we assume that $\Pr\{R = 1|X, Z, Y\} > \sigma > 0$ w.p.1 for some constant σ. Let $\mathcal{N}_1$ and $\mathcal{N}_2$ denote the tangent space of the full data and the missing data model respectively, such that $\mathcal{N} = \mathcal{N}_1 \oplus \mathcal{N}_2$ is the tangent space in the full data model. Rotnitzky et al. (1998) established that the observed data tangent space is given by $\mathcal{N}^O = \overline{\mathcal{N}_1^O + \mathcal{N}_2^O}$, where $\mathcal{N}_j^O = \overline{R\left(g \circ \Pi_j\right)}$ where $R\left(\cdot\right)$ is the range of the operator $g : \Omega^{(L,R)} \to \Omega^{(O)}$ is the conditional expectation operator $g\left(\cdot\right) = E\left[\cdot|O\right]$, $\Omega^{(L,R)}$ and $\Omega^{(O)}$ are the spaces of all random functions of (C, L) and O respectively. Π_j is the Hilbert space projection operator from $\Omega^{(L,R)}$ onto $\mathcal{N}_j$ and $\overline{\mathcal{S}}$ is the close linear span of the set $\mathcal{S}$. We wish to characterize the orthocomplement to the tangent space in the observed data

model $\mathcal{N}^{O,\perp} = \mathcal{N}_1^{O,\perp} \cap \mathcal{N}_2^{O,\perp}$. Rotnitzky et al. (1998) showed that

$$\begin{aligned}\mathcal{N}_1^{O,\perp} &= \left\{N_1^{O,\perp} = Rm(L)/\pi(L) + N_{car} : m(L) \in \mathcal{N}_1^{\perp} \text{ and } N_{car} \in \mathcal{N}_{car}\right\} \\ &\quad \text{where} \\ \mathcal{N}_{car} &= \left\{N_{car} = (1-R)a(O) - RE\left[(1-R)a(O)\,|L\right]/\pi(L) : \text{ for any } a(O) \in \Omega^{(O)}\right\}.\end{aligned}$$

Thus we need to characterize $\mathcal{N}_1^{\perp}$. By the exclusion restriction, all scores of $f(L)$ may be written as

$$\mathcal{N}_1 = \{s(L) = s_1(Y|X) + s_2(Z|X) + s_3(X) : E(S_1|X) = E(S_2|X) = E(S_3) = 0\}.$$

Therefore

$$\mathcal{N}_1^{\perp} = \left\{C - C^{\dagger} : C = c(Y,X,Z) \text{ arbitrary}, C^{\dagger} = E[C|Z,X] + E[C|Y,X] - E[C|X]\right\},$$

a result given by Bickel et al. (1993) and Tchetgen Tchetgen et al. (2010). Therefore, we have that $\mathcal{N}_1^{O,\perp}$ consists of functions

$$R\left\{C - C^{\dagger}\right\}/\pi(L) + (1-R)a(O) - RE\left[(1-R)a(O)\,|L\right]/\pi(L)$$

for arbitrary functions $C = c(L)$ and $A = a(O)$. Also, Rotnitzky et al. (1998) establish that $\mathcal{N}_2^{O,\perp} = \left\{b(O) : b(O) \in \mathcal{N}_2^{\perp}\right\}$ and therefore,

$$\mathcal{N}^{O,\perp} = \left\{N_1^{O,\perp} \in \mathcal{N}_1^{O,\perp} : E\left[N_2 N_1^{O,\perp}\right] = 0, N_2 \in \mathcal{N}_2\right\}.$$

Note that $\mathcal{N}_2 = \{N_2 = (R - \pi(L))\, g(X,Z) \text{ for all } g\}$, which leads to the following result.

Lemma 1.

$$\mathcal{N}^{O,\perp} = \left\{ \begin{array}{c} N_1^{O,\perp}(a_c) = R\left\{C - C^{\dagger}\right\}/\pi(L) + (1-R)a_c(O) - RE\left[(1-R)a_c(O)\,|L\right]/\pi(L) : \\ a_c = E\left[C - C^{\dagger}|R=0,X,Z\right] \end{array} \right\}$$

Proof. $N_1^{O,\perp}(a_c)$ is clearly in $\mathcal{N}_1^{O,\perp}$, it suffices to show that the unique solution to the equa-

tion $E\left[N_1^{O,\perp^*} N_2\right] = 0$, for all $N_2 \in \mathcal{N}_2$ is given by $N_1^{O,\perp^*} = N_1^{O,\perp}(a_c)$. In this vein

$$
\begin{aligned}
0 &= E\left[N_1^{O,\perp^*} N_2\right] \\
&= E\left[\left\{ \begin{array}{c} R\left\{C - C^\dagger\right\}/\pi(L) + (1-R)a^*(O) \\ -RE\left[(1-R)a^*(O)\,|L\right]/\pi(L) \end{array} \right\}(R - \pi(L))\, g(X,Z)\right] = 0 \text{ for all } g \\
&\Leftrightarrow\ 0 = E\left[\left\{ \begin{array}{c} R\left\{C - C^\dagger\right\}/\pi(L) + (1-R)a^*(O) \\ -RE\left[(1-R)a^*(O)\,|L\right]/\pi(L) \end{array} \right\}(R - \pi(L))\,|X,Z)\right] \\
&\Leftrightarrow\ 0 = E\left[(1-\pi(L))\left\{C - C^\dagger\right\}|X,Z\right] - E\left[(1-\pi(L))\pi(L)\, a^*(O)\,|X,Z\right] \\
&\quad -E\left[(1-\pi(L))\, E\left[(1-R)a^*(O)\,|L\right]|X,Z\right] \\
&\Leftrightarrow\ 0 = E\left[(1-\pi(L))\left\{C - C^\dagger\right\}|X,Z\right] - E\left[(1-\pi(L))\, a^*(O)\,|X,Z\right] \\
&\Leftrightarrow\ 0 = E\left[\left[E\left[\left\{C - C^\dagger\right\}|X, R=0, Z\right] - a^*(O)\right](1-R)|X,Z\right]
\end{aligned}
$$

Upon writing $a^*(O) = a_1^*(L)\,R + a_2^*(X,Z)\,(1-R)$, we have that $a_2^*(X,Z) = E\left[\left\{C - C^\dagger\right\}|X, R=0, Z\right] = a_c$, proving the result. □

Therefore the ortho-complement to the tangent space in a model where (i) and (ii) hold is given by $\mathcal{N}^{O,\perp}$. Next, we consider the goal of estimating a full data functional $\phi = \phi(F_L) = E(Y)$ in the missing data model given by (i) and (ii). Let $IF_{\phi,1} = Y - \phi$ denote the full data influence function in the nonparametric model which does not assume (i). Then, in the model that assumes (i) and (ii) hold we have that

$$
\widetilde{\mathcal{N}}_1^{\perp} = \left\{ \begin{array}{c} k \cdot IF_{\phi,1} + C - C^\dagger : \text{for all constants } k \text{ and} \\ C = c(Y,X,Z) \text{ arbitrary}, C^\dagger = E[C|Z,X] + E[C|Y,X] - E[C|X] \end{array} \right\}
$$

Similar to Lemma 1, we get the following set of influence functions for ϕ in the model given by (i) and (ii)

Lemma 2.

$$
\widetilde{\mathcal{N}}^{O,\perp} = \left\{ \begin{array}{c} \widetilde{N}_1^{O,\perp}(a_{c,\phi}) = R\left\{k \cdot IF_{\phi,1} + C - C^\dagger\right\}/\pi(L) \\ +(1-R)a_c(O) - RE\left[(1-R)a_c(O)\,|L\right]/\pi(L) : \\ a_{c,\phi} = E\left[k \cdot IF_{\phi,1} + C - C^\dagger | R=0, X, Z\right], \text{for arbitrary } C = c(Y,X,Z) \text{ and constant } k \end{array} \right\}
$$

The proof is similar to that of Lemma 1. Next, lets suppose that (ii) does not hold, and instead, we have (iii) a parametric model $\alpha_y(Y,X,Z;\gamma)$ with unknown p-dimensional parameter γ. Let $F_t(R,L)$ denote the complete data submodel indexed by t such that

$F_0(R,L) = F(R,L)$. Under the submodel, we have that Let $\phi(\gamma_t, t)$ denote the solution to

$$0 = E_t\left\{\widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi(t), \gamma_t, t)\right\} \text{ for all } t \text{ in the model}$$

and therefore

$$\begin{aligned}
0 &= \nabla_t E_t\left\{\widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi(t), \gamma_t, t)\right\} \\
&= E\left\{\widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi(\gamma))\, S\right\} + E\left\{\nabla_t \widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi(t), \gamma_t, t)\right\} \\
&= E\left\{\widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi(\gamma))\, S\right\} + E\left\{\nabla_\phi \widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi)\right\} \nabla_t \phi(t) \\
&\quad + E\left\{\nabla_\gamma \widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi, \gamma)\right\} \nabla_t \gamma_t \\
&\quad + E\left\{\nabla_t \widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi, \gamma, t)\right\}
\end{aligned}$$

Now since $\widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi, \gamma, t)$ is orthogonal to all nuisance parameters in the model where (ϕ, γ) is known $E\left\{\nabla_t \widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi, \gamma, t)\right\} = 0$, therefore, we get

$$\nabla_t \phi(t) = -E\left\{\nabla_\phi \widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi)\right\}^{-1} \times \left(E\left\{\widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi(\gamma))\, S\right\} + E\left\{\nabla_\gamma \widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi, \gamma)\right\} \nabla_t \gamma_t\right)$$

Note that by Lemma 1

$$\nabla_t \gamma_t = E\left(N_1^{O,\perp}(a_d)\, S\right)$$

where $N_1^{O,\perp}(a_d) \in \mathcal{N}^{O,\perp}$ with $a_d = E\left[D - D^\dagger | R = 0, X, Z\right]$ with D an arbitrary p-dimensional function of L . Therefore, we conclude that

$$\begin{aligned}
\nabla_t \phi(t) &= -E\left\{\nabla_\phi \widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi)\right\}^{-1} \\
&\quad \times E\left\{\left[\widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi(\gamma)) + E\left\{\nabla_\gamma \widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi, \gamma)\right\} N_1^{O,\perp}(a_c)\right] S\right\}
\end{aligned}$$

proving that the orthocomplement to the nuisance tangent space in the model given by (i) and (iii) is given by

$$\widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi(\gamma)) + E\left\{\nabla_\gamma \widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi, \gamma)\right\} N_1^{O,\perp}(a_d)$$

Now, we note that $\widetilde{N}_1^{O,\perp}(a_{c,\phi}; \phi(\gamma))$ can be written $\widetilde{N}_1^{O,\perp}(a_c; \phi(\gamma)) + \widetilde{N}_1^{O,\perp}(a_\phi; \phi(\gamma))$ where
$a_c = E\left[C - C^\dagger | R = 0, X, Z\right]$ and $a_\phi = E\left[k \cdot IF_{\phi,1} | R = 0, X, Z\right]$
$a_{c,\phi} = E\left[k \cdot IF_{\phi,1} + C - C^\dagger | R = 0, X, Z\right]$.

Let
$M = \widetilde{N}_1^{O,\perp}(a_{c^*};\phi) = \Pi\left(\widetilde{N}_1^{O,\perp}(a_\phi;\phi(\gamma)) \mid \left\{\widetilde{N}_1^{O,\perp}(a_c;\phi(\gamma)) : c\right\}\right),$
and let $U = N_1^{O,\perp}(a_{d^*})$ denote the efficient influence function of γ. Then we have that the efficient influence function of ϕ is given by

$$\widetilde{N}_1^{O,\perp}(a_\phi;\phi) - M + E\left\{\nabla_\gamma\left[\widetilde{N}_1^{O,\perp}(a_\phi;\gamma) - M(\gamma)\right]\right\}U$$

since $\widetilde{N}_1^{O,\perp}(a_\phi;\phi) - M$ is in the tangent space of the model, and so is U.
In the special case where Z and Y are binary, $C - C^\dagger$ can be written

$$b(X)\{Y - E(Y|X)\}\{Z - E(Z|X)\}$$

for some function b, so that

$$\begin{aligned}\widetilde{N}_1^{O,\perp}(a_c;\phi(\gamma)) &= b(X)\times\{R\{Y-E(Y|X)\}\{Z-E(Z|X)\}/\pi(L)+\\ &\quad (1-R)E[\{Y-E(Y|X)\}\{Z-E(Z|X)\}|X,R=0,Z]\\ &\quad -RE[(1-R)E[\{Y-E(Y|X)\}\{Z-E(Z|X)\}|X,R=0,Z]|L]/\pi(L)\}\\ &= b(X)\times W\end{aligned}$$

Therefore, letting $H = \widetilde{N}_1^{O,\perp}(a_\phi;\phi(\gamma))$

$$\begin{aligned}M &= \Pi\left(\widetilde{N}_1^{O,\perp}(a_\phi;\phi(\gamma)) \mid \left\{\widetilde{N}_1^{O,\perp}(a_c;\phi(\gamma)) : c\right\}\right)\\ &= E\{HW|X\}E\left\{W^2|X\right\}^{-1}W\end{aligned}$$

and $U = N_1^{O,\perp}(a_{d^*})$ solves

$$E\left\{N_1^{O,\perp}(a_{d^*})N_1^{O,\perp}(a_d)\right\} = E\left\{\nabla_\gamma N_1^{O,\perp}(a_d;\gamma)\right\} \text{ for all } D.$$

one can verify that $N_1^{O,\perp}(a_{d^*}) = D^*(X)\times W(\gamma)$ where

$$D^*(X) = E\left\{W(\gamma)^{\otimes 2}|X\right\}^{-1}E\{\nabla_\gamma W(\gamma)|X\}$$

□

Appendix B

Proofs of results for nomonotone MAR IPW

B.1 Restrictions imposed by polytomous logistic regression model

Suppose there are M missingness patterns, each with observed variables $L_{(m)}$, $m = 1, ..., M$. Choosing pattern j as the baseline category, we model the other missingness pattern probabilities as

$$\Pr\{R = m|L\} = \frac{\exp({\gamma_m}' L_{(m)})}{1 + \sum_{k \in \{1,...,M\} \setminus \{j\}} \exp({\gamma_k}' L_{(k)})} \quad \text{for} \quad m \in \{1, ..., M\} \setminus \{j\}.$$

Let $L_I = \bigcap_{m \in \{1,...,M\} \setminus \{j\}} L_{(m)}$. Then by the MAR assumption, each of the above probabilities $\Pr\{R = m|L\}$ depends on $L_{(m)}$ respectively. But they can only depend on L_I. If not, then the probability for one of the missing data patterns h will depend on variables $L_{(h)} \setminus L_I$ that another pattern does not have. This is not possible due to the linked nature of the terms in the denominator of the probability expression.

B.2 Asymptotic results for IPW estimator

The consistency of $\hat{\beta}$ can be established under general conditions for 2-step estimators (Newey and McFadden, 1993) to show uniform convergence of estimating equation (2.15) in β, where we make use of the fact that $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}$. Typically one would need to impose moment assumptions on $\pi_1(L; \gamma)$ and $M(L; \beta)$ (Wooldridge, 2007).

To investigate the asymptotic distribution of $\hat{\beta}$, under suitable regularity conditions expand (2.15) around the true values β_0 and subsequently γ_0,

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta_0) &= -\left[\frac{1}{n}\sum_{i=1}^{n}\nabla_\beta\Gamma_i(\beta^*,\hat{\gamma})\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Gamma_i(\beta_0,\hat{\gamma})\\ &= -\left[\frac{1}{n}\sum_{i=1}^{n}\nabla_\beta\Gamma_i(\beta^*,\hat{\gamma})\right]^{-1}\times\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Gamma_i(\beta_0,\gamma_0)+\left(\frac{1}{n}\sum_{i=1}^{n}\nabla_\gamma\Gamma_i(\beta_0,\gamma^*)\right)\sqrt{n}(\hat{\gamma}-\gamma_0)\right]\end{aligned}$$

where β^* and γ^* are the mean values and $\Gamma(\beta,\gamma)=\{\mathbb{1}(R_1=1)/\pi_1(L;\gamma)\}M(L;\beta)$. When $\hat{\gamma}$ is the maximum likelihood estimator or a Bayes point estimator satisfying conditions in the Bernstein-von Mises Theorem, it is an asymptotically linear estimator with the influence function

$$\sqrt{n}(\hat{\gamma}-\gamma_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{E}\left[S_{\gamma_0}S_{\gamma_0}^T\right]^{-1}S_i\gamma_0+o_p(1) \tag{B.1}$$

where S_γ is the score function with respect to the missing data model parameters γ. Substituting the influence function representation into previous expansion gives

$$\begin{aligned}&\sqrt{n}(\hat{\beta}-\beta_0)\\ &=-\mathrm{E}\{\nabla_\beta\Gamma(\beta_0,\gamma_0)\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\Gamma_i(\beta_0,\gamma_0)+\mathrm{E}\{\nabla_\gamma\Gamma(\beta_0,\gamma_0)\}\mathrm{E}\left[S_{\gamma_0}S_{\gamma_0}^T\right]^{-1}S_i\gamma_0\right\}+o_p(1).\end{aligned} \tag{B.2}$$

In addition, from the assumption that the parameters governing full data and the missing data process are separable, under standard regularity conditions we have for observed data $\boldsymbol{O}$

$$\begin{aligned}&\mathrm{E}[\Gamma(\beta,\gamma)]=\int\Gamma(\beta,\gamma)f(\boldsymbol{O};\beta,\gamma)\,d\boldsymbol{O}=0\\ &\frac{\partial}{\partial\gamma}\mathrm{E}[\Gamma(\beta,\gamma)]=\int\frac{\partial}{\partial\gamma}\Gamma(\beta,\gamma)f(\boldsymbol{O};\beta,\gamma)\,d\boldsymbol{O}+\int\Gamma(\beta,\gamma)\frac{\partial}{\partial\gamma}f(\boldsymbol{O};\beta,\gamma)\,d\boldsymbol{O}=0\\ &\implies\mathrm{E}\{\nabla_\gamma\Gamma(\gamma,\beta)\}=-\int\Gamma(\beta,\gamma)\frac{\frac{\partial}{\partial\gamma}f(\boldsymbol{O};\beta,\gamma)}{f(\boldsymbol{O};\beta,\gamma)}f(\boldsymbol{O};\beta,\gamma)\,d\boldsymbol{O}=-\mathrm{E}[\Gamma(\beta,\gamma)S_\gamma].\end{aligned}$$

Substituting the above equality to (B.2)

$$\sqrt{n}(\hat{\beta}-\beta_0)=$$
$$-\mathrm{E}\{\nabla_\beta\Gamma(\beta_0,\gamma_0)\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\Gamma_i(\beta_0,\gamma_0)-\mathrm{E}[\Gamma(\beta_0,\gamma_0)S_{\gamma_0}^T]\mathrm{E}\left[S_{\gamma_0}S_{\gamma_0}^T\right]^{-1}S_i\gamma_0\right\}+o_p(1).$$

An application of Slutsky's theorem shows that

$$\sqrt{n}(\hat{\beta}-\beta_0)\xrightarrow{d}N\left(0,\mathrm{E}\{\nabla_\beta\Gamma(\beta_0,\gamma_0)\}^{-1}\mathrm{Var}\left[\Gamma(\beta_0,\gamma_0)-W(\beta_0,\gamma_0)\right]\mathrm{E}\{\nabla_\beta\Gamma(\beta_0,\gamma_0)\}^{-1^T}\right) \tag{B.3}$$

where

$$W(\beta_0,\gamma_0)=\mathrm{E}[\Gamma(\beta_0,\gamma_0)S_{\gamma_0}^T]\mathrm{E}\left[S_{\gamma_0}S_{\gamma_0}^T\right]^{-1}S\gamma_0.$$

The sandwich estimator is consistent for $\mathrm{E}\{\nabla_\beta\Gamma(\beta_0,\gamma_0)\}^{-1}\mathrm{E}\left[\Gamma(\beta_0,\gamma_0)^{\otimes 2}\right]\mathrm{E}\{\nabla_\beta\Gamma(\beta_0,\gamma_0)\}^{-1^T}$. In the Hilbert space of mean-zero random functions, $\mathrm{E}\left[\Gamma(\beta_0,\gamma_0)S_{\gamma_0}^T\right]\mathrm{E}\left[S_{\gamma_0}S_{\gamma_0}^T\right]^{-1}S\gamma_0$ is the projection of $\Gamma(\beta_0,\gamma_0)$ onto the linear subspace spanned by elements of S_{γ_0}. Therefore by Pythagorean Theorem

$$\mathrm{E}\left[\Gamma(\beta_0,\gamma_0)^{\otimes 2}\right]-\mathrm{E}\left[\left\{\Gamma(\beta_0,\gamma_0)-\mathrm{E}\left[\Gamma(\beta_0,\gamma_0)S_{\gamma_0}^T\right]\mathrm{E}\left[S_{\gamma_0}S_{\gamma_0}^T\right]^{-1}S\gamma_0\right\}^{\otimes 2}\right]$$

is positive semi-definite and the sandwich estimator provides conservative estimate for the true asymptotic variance.

B.3 Implementation and sample OpenBUGS code for simulation study

We obtain the BCE estimator of γ as the posterior median of distribution (11) with diffuse priors $\gamma_j\sim N(0,10^2)$ and $\sigma^*=10^{-8}$. Adaptive Gibbs sampling (Gilks et al. 1995) was implemented through BRugs, the R interface to the OpenBUGS MCMC software (Lunn et al. 2009). We assessed convergence by visually inspecting the trace plots as well as through the Gelman-Rubin convergence statistic (Gelman and Rubin 1992), and included an adaptive phase of 10^4 iterations out of a total of 2×10^4 iterations.

In the OpenBUGS code for estimation of the missing data model (29) in scenario 1 of the simulation study, individuals are assigned to their respective missing data patterns,

$R_k = 1, k = 1, 2, 3$, with $R_1 = 1$ denoting complete-cases. For a person with missing data pattern $R_h = 1$, the encoding follows that $R_j = 0$ for $j \neq h$. The first part of the code describes the contribution of each missing data probability of the n individuals to the likelihood function corresponding to missing data model (29). Then the diffuse prior distributions for the parameters in the missing data model are specified as independent $N(0, 10^2)$. Finally, constraints (10) are imposed on the n_c complete-cases with user-defined σ^*, where the input dataset is ordered such that the first n_c individuals are complete-cases. Posterior mean, median and 95% credible intervals can be obtained directly from Markov-chain Monte-Carlo sampling in R through BRugs, an interface to the OpenBUGS software.

```
Model <- function() {

for (i in 1:n){
z[i]   <- 1
z[i]    ~ dbern(p[i])
p[i]   <- L[i]
L[i]   <-  R1[i]*pi1[i]+R2[i]*pi2[i]+R3[i]*pi3[i]

#Probability for each missing data pattern

logit(pi2[i])<- g[1]+g[2]*Y[i]+g[3]*X1[i]
logit(pi3[i])<- g[4]+g[5]*X2[i]+g[6]*X3[i]
pi1[i]  <- 1-pi2[i]-pi3[i]
}

#Priors for parameters in missing data model
for (j in 1:6) {
        g[j] ~ dnorm(0, 0.01)
}
```

```
# implementing the constraints for complete-cases
for (k in 1:n_c){
          ones[k] <- 1
          ones[k] ~ dbern(C[k])
          C[k] <- step(pi1[k]-sigma_star)
}
}
```

B.4 Augmented inverse probability weighted (AIPW) estimators

We consider the restricted augmentation space $\mathbb{A}^* \subset \mathbb{A}$ formed by the span of a finite vector of linearly independent functions

$$\left\{ \left[\frac{\mathbb{1}(R=1)}{\pi_1(L)} - \frac{\mathbb{1}(R=r)}{\pi_r\left(L_{(r)}\right)} \right] t^*_{rk}\left(L_{(r)}\right) : r; k = 1, ..., K_r \right\},$$

where for each r, $t^*_r(L_{(r)})$ is a K_r-vector of user defined functions of $L_{(r)}$, $r = 1, ..., M$. It is recommended to include in $\mathbb{A}^*$ scores corresponding to the model used to estimate the missing data mechanism, which leads to simplification in estimating the asymptotic variance of the resulting estimator (Robins et al. 1994; Tsiatis 2006). Specifically, under model (2.6), $\mathbb{A}^*$ includes the score functions given by (2.9).

Similarly, we consider a restricted linear subspace $\mathbb{U}^{F*} \subset \mathbb{U}^F$ spanned by l linearly independent full-data estimating equations, where $l > q$. The resulting class of restricted augmented estimating equations is given by

$$\mathbb{P}_n \left\{ \frac{\mathbb{1}(R=1)}{\pi_1(L;\hat{\gamma})} C_1 U^*(L;\beta) + C_2 A^*(R, L_{(R)};\hat{\gamma}) \right\} = 0 \tag{S4}$$

for any choice of constant matrices C_1 of dimensions $q \times l$ and C_2 of dimensions $q \times k$ where $k = \sum_{r \geq 2} K_r$. $U^*(L;\beta)$ is a l-dimensional vector of basis functions spanning $\mathbb{U}^{F*}$ and $A^*(R, L_{(R)};\gamma)$ is a k-dimensional vector of basis functions spanning $\mathbb{A}^*$. Using a result due to Tsiatis (2006) one can show that the optimal choice of (C_1, C_2) within the class (S4)

is given by the solution to

$$\left[C_1^{opt}, C_2^{opt}\right] \begin{bmatrix} U_{11} & U_{12} \\ U_{12}^T & U_{22} \end{bmatrix} = [H_1, H_2]$$

where

$$\begin{aligned} U_{11} &= E\left\{\frac{U^*(\beta)U^*(\beta)^T}{\pi_1(L)}\right\}^{l\times l} \\ U_{12} &= E\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L)}U^*(\beta)A^{*T}\right\}^{l\times k} \\ U_{22} &= E\left\{A^*A^{*T}\right\}^{k\times k} \\ H_1 &= \left(-E\left\{\frac{\partial U^*(\beta)}{\partial \beta}\right\}^T\right)^{q\times l} \\ H_2 &= 0^{q\times k} \end{aligned}$$

The matrices (U_{11}, U_{12}, H_1) that involve full data L can be estimated from the complete cases only by standard inverse probability weighted empirical averages and the matrix U_{22} by an empirical average of the observed data. Constrained Bayesian estimation of the missing data process involves centering A^* so that it has mean zero empirically. Then the optimal AIPW estimator $\hat{\beta}_{opt}$ in the restricted class of estimating equations is given by the solution to

$$\mathbb{P}_n\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L;\hat{\gamma})}\widehat{C}_1^{opt}(\beta)U^*(L;\beta) + \widehat{C}_2^{opt}(\beta)A^*(R, L_{(R)};\hat{\gamma})\right\} = 0, \tag{S5}$$

and a consistent estimator for the asymptotic variance of $\hat{\beta}_{opt}$ is given by

$$\left\{\widehat{H}_1(\hat{\beta}_{opt})\widehat{U}^{11}(\hat{\beta}_{opt})\widehat{H}_1^T(\hat{\beta}_{opt})\right\}^{-1} \tag{S6}$$

where

$$\widehat{U}^{11} = \left(\widehat{U}_{11} - \widehat{U}_{12}\widehat{U}_{22}^{-1}\widehat{U}_{12}^T\right)^{-1}.$$

(Tsiatis 2006). Finding the solution $\hat{\beta}_{opt}$ to (S5) involves estimating the matrices for each value of β, which can be computationally intensive. Instead, an estimator asymptotically equivalent to $\hat{\beta}_{opt}$ is obtained by the simple one-step update of a standard IPW estimator

$\hat{\beta}_{ipw}$:

$$\hat{\beta}^*_{opt} = \hat{\beta}_{ipw} + \widetilde{IF}_\beta \left(\hat{\beta}_{ipw} \right), \tag{S7}$$

where

$$\widetilde{IF}_\beta \left(\hat{\beta}_{ipw} \right) = \left\{ -\sum_i \partial \left[\frac{\mathbb{1}(R_i = 1)}{\pi_1(L_i; \hat{\gamma})} M(L_i; \hat{\beta}_{ipw}) \right] / \partial \beta \right\}^{-1} \times \left\{ \sum_i \left[\frac{\mathbb{1}(R_i = 1)}{\pi_1(L_i; \hat{\gamma})} \widehat{C}_1^{opt}(\hat{\beta}_{ipw}) U^*(L_i; \hat{\beta}_{ipw}) + \widehat{C}_2^{opt}(\hat{\beta}_{ipw}) A^*(R_i, L_{(R),i}; \hat{\gamma}) \right] \right\}$$

and $\hat{\beta}_{ipw}$ is the standard IPW solution to (15). It is straightforward to show that under standard regularity conditions and in the absence of model misspecification, the influence function of $\hat{\beta}^*_{opt}$ is identical to that of $\hat{\beta}_{opt}$ (van der Vaart 1998).

The asymptotic efficiency of the optimal restricted AIPW estimator in relation to the semiparametric efficiency bound for a given full data semiparametric model of interest depends on how close the span of $\mathbb{A}^*$ and $\mathbb{U}^{F*}$ is to $\mathbb{A}$ and $\mathbb{U}^F$ respectively. One can show that as one suitably enriches the span of $\mathbb{A}^*$ and $\mathbb{U}^{F*}$ with elements of $\mathbb{A}$ and $\mathbb{U}^F$ so that the former two vector spaces increasingly become dense in the latter two subspaces respectively, the asymptotic variance of $n^{1/2} \left(\hat{\beta}_{opt} - \beta_0 \right)$ nearly attains the semiparametric local efficiency bound for the semiparametric model of the full data and only other restriction that data are MAR (Newey 1993).

Appendix C

Implementation of Constrained Bayesian Estimation for Outcome Heterogeneity

The OpenBUGS code for posterior computation in the simulation study is shown below. For the i^{th} individual with outcome $Y = k$, the encoding follows that $Y_k[i] = 1$ and $Y_j[i] = 0$ for $j \neq k$. The first part of the code describes each individual's contribution to the observed data likelihood. Diffuse prior distributions for the parameters in the model are specified as independent $N(0, 10^2)$. The final part of the code imposes constraints on the sampling space so that $0 < \Pr\left\{Y = 0|X_i; \hat{\beta}\right\} < 1$ for all individuals $i = 1, 2, ..., n$. Posterior mean, median and 95% credible intervals can be obtained directly from Markov-chain Monte-Carlo sampling in R through BRugs, an interface to the OpenBUGS software.

```
Model <- function() {

    for (i in 1:N){
        z[i]   <- 1
        z[i]    ~ dbern(p[i])
        p[i]   <- L[i]
        L[i]   <-  Y0[i]*pi0[i]+Y1[i]*pi1[i]+Y2[i]*pi2[i]+Y3[i]*pi3[i]

        #Probability for each of the outcomes Y=1,2,3.

        logit(pi1[i])<-beta[1]+beta[2]*X1[i]+beta[3]*X2[i]
        logit(pi2[i])<-beta[4]+beta[5]*X1[i]+beta[6]*X2[i]
        logit(pi3[i])<-beta[7]+beta[8]*X1[i]+beta[9]*X2[i]
        pi0[i]  <- 1-pi1[i]-pi2[i]-pi3[i]
    }
    #Priors for parameters in missing data model
    for (j in 1:9) {
```

```
        beta[j] ~ dnorm(0, 0.01)
    }

    # implementing the constraints
    for (k in 1:N){
        ones[k] <- 1
        ones[k] ~ dbern(C[k])
        C[k] <- step(pi0[k])
    }
}
```