Efficient restricted estimators for conditional mean models with missing data

BY Z. TAN

Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, U.S.A.
ztan@stat.rutgers.edu

SUMMARY
Consider a conditional mean model with missing data on the response or explanatory variables due to two-phase sampling or nonresponse. Robins et al. (1994) introduced a class of augmented inverse-probability-weighted estimators, depending on a vector of functions of explanatory variables and a vector of functions of coarsened data. Tsiatis (2006) studied two classes of restricted estimators, class 1 with both vectors restricted to finite-dimensional linear subspaces and class 2 with the first vector of functions restricted to a finite-dimensional linear subspace. We introduce a third class of restricted estimators, class 3, with the second vector of functions restricted to a finite-dimensional subspace. We derive a new estimator, which is asymptotically optimal in class 1, by the methods of nonparametric and empirical likelihood. We propose a hybrid strategy to obtain estimators that are asymptotically optimal in class 1 and locally optimal in class 2 or class 3. The advantages of the hybrid, likelihood estimator based on classes 1 and 3 are shown in a simulation study and a real-data example.

Some key words: Empirical likelihood; Generalized method of moments; Inverse weighting; Missing data; Nonparametric likelihood; Two-phase sampling.

1. INTRODUCTION
Regression analysis aims to study the relationship between a response variable and explanatory variables. However, there often exist missing data on some of these variables. In general, inference based on only subjects with complete data is inappropriate. It is interesting to develop methods to effectively handle missing data.

One type of missing data arises from two-phase sampling, which is widely used as a cost-effective design for data collection (e.g. Hausman & Wise, 1981; Manski & McFadden, 1981; Lawless et al., 1999). In phase one, a random sample is drawn from a population, and inexpensive variables are measured. Then a subsample is selected from the sample, with the selection probability depending on variables measured at phase one, and expensive variables are measured. Another type of missing data occurs due to nonresponse beyond the control of investigators. If the missing data mechanism is ignorable (Rubin, 1976), then the data structure is similar to that in two-phase sampling. Therefore, missing data under ignorability can be handled as in two-phase sampling, provided that the selection probability is correctly modelled.

It is important to distinguish two types of regression models, conditional mean or conditional density models, in which the conditional mean or conditional density is modelled for the response given explanatory variables. For conditional density models with missing data, efficient or approximately efficient estimators have been extensively studied. See Robins et al. (1995), Lawless et al. (1999), Chatterjee et al. (2003), Weaver & Zhou (2005), Song et al. (2009) and
Tan (2010), among others. Except for that of Robins et al. (1995) these estimators are based on the observed-data likelihood, which is analytically tractable because the conditional density is fully parameterized.

For conditional mean models with missing data, existing estimators are mainly based on estimating functions. Robins et al. (1994) introduced a class of augmented inverse-probability-weighted estimators consisting of all regular, asymptotically linear estimators up to the first order. Each estimator depends on a vector of full-data estimating functions indexed by \( \phi \) and a vector of augmentation terms indexed by \( h \). Optimizing efficiency over unrestricted \( \phi \) and \( h \) yields the semiparametric variance bound, but involves a complicated integral equation (Robins et al., 1994, § 4.2). Alternatively, Tsiatis (2004) studied two classes of restricted estimators, class 1 with \( \phi \) and \( h \) both restricted to finite-dimensional linear subspaces and class 2 with \( \phi \) restricted to a finite-dimensional linear subspace and \( h \) unrestricted, and proposed two corresponding estimators. The first estimator is asymptotically optimal in class 1, i.e., it is asymptotically equivalent to the optimal estimator in class 1. The second estimator is locally optimal in class 2, i.e., it is asymptotically equivalent to the optimal estimator in class 2 if a certain working model is correctly specified and remains consistent and asymptotically normal otherwise. However, the second estimator may lose much efficiency if the working model is misspecified (e.g., Chen et al., 2008).

We develop new restricted estimators for conditional mean models with missing data in three different directions. First, we introduce a third class of restricted estimators, class 3, with \( \phi \) unrestricted and \( h \) restricted to a finite-dimensional linear subspace, and derive the optimal estimating functions in closed form. Optimizing efficiency in class 3 is likely to be more useful than in class 2, although the comparison depends on how classes 2 and 3 are specified in individual problems. In fact, the optimal \( \phi \) for class 3, unlike the optimal \( h \) for class 2, involves the selection probability and hence is capable of stabilizing inverse probability weights.

Secondly, we recast the first estimator of Tsiatis (2006) using the generalized method of moments (Hansen, 1982), and derive a new estimator that is asymptotically optimal in class 1 by the methods of nonparametric likelihood (Tan, 2006) and of empirical likelihood (Qin & Lawless, 1994). The likelihood estimator is asymptotically equivalent to the first order to the first estimator of Tsiatis (2006), but is observed to yield smaller biases in our simulation study. This is related to higher order properties of empirical likelihood and generalized-method-of-moments estimators (Newey & Smith, 2004).

Thirdly, we propose a hybrid strategy to obtain estimators that are asymptotically optimal in class 1 and locally optimal in class 2 or 3. The idea is to incorporate into class 1 a locally optimal estimator in class 2 or 3 and then apply an asymptotically optimal estimator in class 1. Such estimators based on classes 1 and 2 are asymptotically equivalent to the second estimator of Tsiatis (2006) if the working model is correctly specified, but remain at least as efficient as the latter estimator if the working model is misspecified. Similarly, hybrid estimators based on classes 1 and 3 have advantages over estimators that are only locally optimal in class 3.

2. Data and models

For a population, let \( Y \) be a response variable and let \( X^* \) be a vector of explanatory variables. Consider a conditional mean model

\[
E(Y \mid X^*) = g(X^*; \theta_0),
\]

(1)

where \( g(\cdot; \theta) \) is a known function and \( \theta_0 \) is a \( d \times 1 \) unknown parameter. No other restriction is imposed on the conditional distribution of \( Y \) given \( X^* \).
The pair \((Y, X^*)\) may not always be observed. Let \(R\) be the nonmissingness indicator such that \(R = 1\) if \((Y, X^*)\) is fully observed and \(R = 0\) otherwise. Consider the following nonmissingness patterns. For missing response, \(Y\) is missing but \(X^*\) is fully observed if \(R = 0\). For missing regressors, a subvector \(X\) of \(X^*\) is missing but \((Y, V)\) is fully observed if \(R = 0\), where \(V\) is the subvector of \(X^*\) complementary to \(X\), i.e., \(X^* = (X, V)\). In addition, let \(V^\dagger\) be a vector of extraneous variables that are always observed. Such variables are often measured to potentially improve inference about \(\theta_0\) (Robins et al., 1994, § 2.2). Throughout, denote by \(Z\) the full data and by \(W\) the coarsened data. Then \(Z = (Y, X^*, V^\dagger)\), and \(W = (X^*, V^\dagger)\) in the case of missing response or \(W = (Y, V, V^\dagger)\) in the case of missing regressors.

Suppose that a simple random sample of size \(n\) is obtained. The observed data consist of independent and identically distributed copies, \(O_i\), of \(O = (Z1\{R = 1\}, W1\{R = 0\}, R)\) \((i = 1, \ldots, n)\) where \(1\{\cdot\}\) denotes an indicator function. The goal is to draw inference about \(\theta_0\) in model (1).

Assume that the missing data mechanism is ignorable (Rubin, 1976):

\[
\text{pr}(R = 1 \mid Z) = \text{pr}(R = 1 \mid W).
\]

This assumption is satisfied for two-phase sampling, but in general needs to be justified by external knowledge. Write \(\pi(W) = \text{pr}(R = 1 \mid W)\). Assume that there exists a constant \(\delta > 0\) such that

\[
\pi(W) > \delta.
\]

Consider a selection probability model

\[
\text{pr}(R = 1 \mid W) = \pi(W; \gamma_0),
\]

where \(\pi(\cdot; \gamma)\) is a known function and \(\gamma_0\) is a finite-dimensional unknown parameter. The score vector is \(s(\gamma) = \{R - \pi(W; \gamma))\{\partial \pi(W; \gamma)/\partial \gamma\}/\{\pi(W; \gamma)(1 - \pi(W; \gamma))\}\). Let \(\hat{\gamma}\) be the maximum likelihood estimator of \(\gamma_0\) and write \(\hat{\pi}(W) = \pi(W; \hat{\gamma})\).

Throughout, we say that two estimators are asymptotically equivalent up to the first order if their difference is \(o_P(n^{-1/2})\) as \(n \to \infty\).

3. REVIEW: SEMIPARAMETRIC THEORY

We present a brief review of basic results in Robins et al. (1994) for conditional mean models with missing data. See Van der Vaart (1998, § 25.5.3) and Tsiatis (2006) for textbook accounts. These results provide the background and framework for developing our work.

Under models (1) and (3), Robins et al. (1994) showed that each regular, asymptotically linear estimator of \(\theta_0\) is asymptotically equivalent, to the first order, to some estimator \(\hat{\theta}_{RRZ} = \hat{\theta}_{RRZ}(\phi, \hat{h})\), depending on \(d \times 1\) vectors \(\phi(X^*)\) and \(h(W)\). Write \(\epsilon(\theta) = Y - g(X^*; \theta)\). The estimator \(\hat{\theta}_{RRZ}\) is a solution to

\[
0 = \bar{E} \left[ \frac{R}{\hat{\pi}(W)} \phi(X^*)\epsilon(\theta) - \left\{ \frac{R}{\hat{\pi}(W)} - 1 \right\} h(W) \right],
\]

where \(\bar{E}\) denotes sample average. The influence function of \(\hat{\theta}_{RRZ}\) is

\[
\left[ E \left\{ \phi(X^*) \frac{\partial g}{\partial \theta^i} (X^*; \theta_0) \right\} \right]^{-1} \Pi_\perp \left[ \frac{R}{\pi(W)} \phi(X^*)\epsilon(\theta_0) - \left\{ \frac{R}{\pi(W)} - 1 \right\} h(W) \right] s(\gamma_0),
\]
where \( \Pi_1(\tau_1 | \tau_2) = \tau_1 - \text{cov}(\tau_1, \tau_2)\var^{-1}(\tau_2)\tau_2 \) for random vectors \( \tau_1 \) and \( \tau_2 \). By the nature of \( \Pi_1 \), \( \theta_{RRZ} \) has no greater asymptotic variance than the solution of \( 0 = \tilde{E}[\pi^{-1}(W)R\phi(X^*)\epsilon(\theta) - \{\pi^{-1}(W)R - 1\}h(W)] \), where the true \( \pi(W) \) is used instead of \( \hat{\pi}(W) \).

Robins et al. (1994) further established the following results. First, for fixed \( \phi \), the asymptotic variance of \( \hat{\theta}_{RRZ}(\phi, h) \) is minimized at \( \hat{\theta}_{RRZ}(\phi, h^\phi) \), where \( h^\phi(W) = E(\phi(X^*)\epsilon(\theta_0) | W) \). Secondly, the semiparametric variance bound for estimators of \( \theta_0 \) is attained at \( \hat{\theta}_{RRZ}(\phi_{eff}, h^\phi_{eff}) \), where \( \phi_{eff} \) is the unique solution to the equation

\[
\phi(X^*) = \frac{\partial}{\partial \theta} g(X^*; \theta_0) E \left\{ \frac{\epsilon^2(\theta_0)}{\pi(W)} | X^* \right\} + \frac{E \left\{ \frac{1-\pi(W)}{\pi(W)} \epsilon(\theta_0)h^\phi(W) | X^* \right\}}{E \left\{ \frac{\epsilon^2(\theta_0)}{\pi(W)} | X^* \right\}}.
\] (4)

The solution \( \phi_{eff} \) has no closed form, except in special cases such as that of missing response (Chen & Breslow, 2004) or of missing regressors with discrete \( W \) (Robins et al., 1994, § 5.2).

There are adaptive extensions of \( \hat{\theta}_{RRZ} \) for the situation where \( \phi(X^*) \) and \( h(W) \) depend on \( \theta_0, \gamma_0 \) and other unknown quantities, which are estimated under a working model with parameters \( \psi \). See Tsiatis (2006, § 10.3 and § 11.2). Let \( \hat{\psi} \) be an estimator such that \( \hat{\psi} \) converges to a constant \( \psi^* \) even though the working model may be misspecified and that \( \psi^* \) equals the true value of \( \psi \) if the working model is correctly specified. Let \( \hat{\theta} \) be a preliminary, consistent estimator of \( \theta_0 \). Then the following equations lead to asymptotically equivalent estimators of \( \theta_0 \) to the first order:

\[
0 = \tilde{E} \left[ \frac{R}{\hat{\pi}(W)} \phi(X^*; \hat{\theta}, \hat{\gamma}, \hat{\psi})\epsilon(\theta) - \left\{ \frac{R}{\hat{\pi}(W)} - 1 \right\} h(W; \hat{\theta}, \hat{\gamma}, \hat{\psi}) \right]
\]

and

\[
0 = \tilde{E} \left[ \frac{R}{\hat{\pi}(W)} \phi(X^*; \theta, \hat{\gamma}, \hat{\psi})\epsilon(\theta) - \left\{ \frac{R}{\hat{\pi}(W)} - 1 \right\} h(W; \theta, \hat{\gamma}, \hat{\psi}) \right].
\]

In fact, the two estimators are asymptotically equivalent to the solution of \( 0 = \tilde{E}[\hat{\pi}^{-1}(W)R\phi(X^*; \theta_0, \gamma_0, \psi^*)\epsilon(\theta) - \{\hat{\pi}^{-1}(W)R - 1\}h(W; \theta_0, \gamma_0, \psi^*)] \). Denote the first estimator by \( \hat{\theta}_{RRZ}(\phi(\cdot; \hat{\theta}, \hat{\gamma}, \hat{\psi}), h(\cdot; \hat{\theta}, \hat{\gamma}, \hat{\psi})) \) and the second by \( \hat{\theta}_{RRZ}(\phi(\cdot; \theta, \hat{\gamma}, \hat{\psi}), h(\cdot; \theta, \hat{\gamma}, \hat{\psi})) \) whenever this distinction is needed.

4. Restricted estimators

4.1. Optimality results

The optimal \( \phi_{eff} \) and \( h_{eff} = h^{\phi_{eff}} \) are derived by optimizing efficiency of \( \hat{\theta}_{RRZ}(\phi, h) \) over the entire infinite-dimensional linear spaces \( \mathcal{Z}_{\phi} \) of \( \phi(X^*) \) and \( \mathcal{Z}_h \) of \( h(W) \). In general, the resulting integral equation (4) is too complicated for practical use. Alternatively, an approximate approach is to optimize efficiency of \( \hat{\theta}_{RRZ}(\phi, h) \) for \( \phi \) restricted to a finite-dimensional linear subspace of \( \mathcal{Z}_{\phi} \) or \( h \) restricted to a finite-dimensional linear subspace of \( \mathcal{Z}_h \). Tsiatis (2006, Ch. 12) considered two classes of restricted estimators and derived the optimal estimators. For class 1, both \( \phi \) and \( h \) are restricted to finite-dimensional linear subspaces. For class 2, \( \phi \) is restricted to a finite-dimensional linear subspace but \( h \) is unrestricted in the entire linear space \( \mathcal{Z}_h \).

We consider a third class of restricted estimators and derive the optimal estimator. For class 3, \( h \) is restricted to a finite-dimensional linear subspace but \( \phi \) is unrestricted in the entire linear space \( \mathcal{Z}_{\phi} \). We recast Tsiatis’s (2006) equations (12.15) and (12.72) for classes 1 and 2 under our setting in Theorem 1 and present the new result for class 3 in Theorem 2.
**Theorem 1 (Tsiatis 2006).** Let $\Phi(X^*)$ be a $p \times 1$ ($p \geq d$) vector of functions of $X^*$ and $H(W)$ be a $k \times 1$ vector of functions of $W$. Write $H^1(W) = [H^1(W), \{\delta\pi(W; \gamma_0) / \delta\gamma^T\} / (1 - \pi(W))]^T$. Assume that the components of $\Phi(X^*)$ are linearly independent and those of $H^1(W)$ are linearly independent.

(a) For fixed $\Phi$ and $H$, consider class 1 of restricted estimators $\{\hat{\theta}_{RRZ}(a^T\Phi, b^T H) : a$ is a $p \times d$ matrix and $b$ is a $k \times d$ matrix of constants$\}$. The asymptotic variance of $\hat{\theta}_{RRZ}$ is minimized at $\hat{\theta}_{RRZ}^d(\alpha^T\Phi, \alpha^T \beta^T H)$, where $\alpha = A^{-1}D$, $\beta$ is the submatrix of the first $k$ rows in $\beta^T = B^{-1}C$, and

$$
A = \text{var}[\Pi \{\pi^{-1} R \Phi \varepsilon(\theta_0) \mid (\pi^{-1} R - 1) H^T\}], \\
B = \text{var}[(\pi^{-1} R - 1) H^T], \\
C = \text{cov}^T[(\pi^{-1} R \Phi \varepsilon(\theta_0) - (\pi^{-1} R - 1) H^T)].
$$

(b) For fixed $\Phi$, consider class 2 of restricted estimators $\{\hat{\theta}_{RRZ}(a^T\Phi, h) : a$ is a $p \times d$ matrix of constants and $h$ is a $d \times 1$ vector of functions$\}$. Take $H(W) = E[\Phi \varepsilon(\theta_0) \mid W]$. The asymptotic variance of $\hat{\theta}_{RRZ}$ is minimized at $\hat{\theta}_{RRZ}^d(\alpha^T\Phi, \alpha^T H)$, where $\alpha = A^{-1}D$ as in (a) and further $A = \text{var}[(\pi^{-1} R \Phi \varepsilon(\theta_0) - (\pi^{-1} R - 1) H^T)].$

Our formulae are simpler than those in Tsiatis (2006) and can be derived by the theory of combining estimating functions (McCullagh & Nelder, 1989) or the generalized method of moments (Hansen, 1982). The optimal estimator in class 1, $\hat{\theta}_{RRZ}(\alpha^T\Phi, \alpha^T \beta^T H)$, is asymptotically equivalent to the first order to the solution of

$$
0 = \alpha^T \hat{E} \left[ \frac{R}{\pi(W)} \Phi(X^*) \varepsilon(\theta) - \left\{ \frac{R}{\pi(W)} - 1 \right\} \beta^T H^T(W) \right].
$$

Here $\beta^T$ is the coefficient matrix for projecting $\pi^{-1} R \Phi \varepsilon(\theta)$ on $(\pi^{-1} R - 1) H^T = \{(\pi^{-1} R - 1) H^T, s^T(\gamma_0)^T\}^T$, and $\alpha$ is the coefficient matrix for optimally linearly combining $p$ estimating functions $\Pi \{\pi^{-1} R \Phi \varepsilon(\theta) \mid (\pi^{-1} R - 1) H^T\}$ into $d$ estimating functions, computed with $\varepsilon(\theta)$ replaced by $\varepsilon(\theta_0)$. See Tan (2006, Theorem 4) for an explanation for the inclusion of $s(\gamma_0)$ in $(\pi^{-1} R - 1) H^T$. The optimal estimator in class 2, $\hat{\theta}_{RRZ}(\alpha^T\Phi, \alpha^T H)$, is asymptotically equivalent to the first order to the solution of

$$
0 = \alpha^T \hat{E} \left[ \frac{R}{\pi(W)} \Phi(X^*) \varepsilon(\theta) - \left\{ \frac{R}{\pi(W)} - 1 \right\} E[\Phi \varepsilon(\theta_0) \mid W] \right].
$$

Here $\pi^{-1} R - 1) E[\Phi \varepsilon(\theta_0) \mid W]$ is the projection of $\pi^{-1} R \Phi \varepsilon(\theta)$ to the space $\{\pi^{-1} R - 1) H : H$ is a $p \times 1$ vector of functions of $W\}$, and $\alpha$ is the coefficient matrix for optimally linearly combining $p$ estimating functions $\pi^{-1} R \Phi \varepsilon(\theta) - (\pi^{-1} R - 1) E[\Phi \varepsilon(\theta_0) \mid W]$ into $d$ estimating functions, computed with $\varepsilon(\theta)$ replaced by $\varepsilon(\theta_0)$.

**Theorem 2.** For fixed $H$, consider class 3 of restricted estimators $\{\hat{\theta}_{RRZ}(\phi, b^T H) : \phi$ is a $d \times 1$ vector of functions and $b$ is a $k \times d$ matrix of constants$\}$. The asymptotic variance of $\hat{\theta}_{RRZ}$ is minimized at $\hat{\theta}_{RRZ}^d(\phi, \beta^T H)$, where

$$
\phi(X^*) = \phi_1(X^*; \theta_0) + \alpha^T_2 \phi_2(X^*; \theta_0)
$$

$$
\frac{\partial}{\partial \theta_0} g(X^*; \theta_0) = \frac{1}{E} \left\{ \frac{\pi^2(\theta_0)}{\pi(W)} \mid X^* \right\} + \alpha^T_2 \frac{1}{E} \left\{ \frac{1 - \pi(W)}{\pi(W)} \varepsilon(\theta_0) H^1(W) \mid X^* \right\},
$$

(5)
is the submatrix of the first \( k \) rows in \( \beta^\dagger \), and \( \alpha_2 \) and \( \beta^\dagger \) are determined as follows. Take 
\[
\Phi(X^*) = \begin{bmatrix} \phi_1^1(X^*; \theta_0), \phi_1^2(X^*; \theta_0) \end{bmatrix}^T.
\]
Then \((I_d, \alpha_2^\dagger)^T = A^{-1}D \) and \( \beta^\dagger = B^{-1}C \), where \( A \), \( B \), and \( D \) are defined as in Theorem 1(a), \( C = \text{cov}^T\{\pi^{-1}R\phi\varepsilon(\theta_0), (\pi^{-1}R - 1)H^\dagger\} \), and \( I_d \) is the \( d \times d \) identity matrix. Furthermore, \( \alpha_2 = \beta^\dagger \).

The key conclusion of Theorem 2 is that the optimal \( \phi \) is a linear combination of \( \phi_1 \) and \( \phi_2 \). The coefficients \( A^{-1}D \) and \( B^{-1}C \) can be derived as in Theorem 1(a), although they satisfy additional constraints. The optimal estimator in class 3, \( \hat{\theta}_{\text{RRZ}}(\phi, \beta^\dagger H) \), is asymptotically equivalent to the first order to the solution of 
\[
0 = \tilde{E} \left( \frac{R}{\pi(W)} \phi_1(X^*; \theta_0)\varepsilon(\theta) + \alpha_2^T \left[ \frac{R}{\pi(W)} \phi_2(X^*; \theta_0)\varepsilon(\theta) - \left\{ \frac{R}{\pi(W)} - 1 \right\} H^\dagger(W) \right] \right).
\]
For a special case, if \( H(W) = 0 \) and \( \hat{\sigma}(W) \) is replaced throughout by \( \pi(W) \) in \( \hat{\theta}_{\text{RRZ}} \), then the optimal estimator is \( \hat{\theta}_{\text{RRZ}}(\phi_1, 0) \).

Formula (5) appears similar to, but fundamentally differs from, (4). Formula (5) gives the optimal \( \phi \) for fixed \( H \) in closed form, whereas (4) only implicitly determines the overall optimal \( \phi_{\text{eff}} \). Moreover, (5) sheds new light on the nature of (4). If \( \phi \) satisfies (5) and \( h^\phi = \beta^\dagger H^\dagger \) simultaneously, then \( \phi \) satisfies (4). Therefore, (4) is a conjunction of (5) and \( h^\phi = \beta^\dagger H^\dagger \). The estimator \( \hat{\theta}_{\text{RRZ}}(\phi, \beta^\dagger H) \) attains semiparametric efficiency if it is optimal not only among class 3 of restricted estimators with \( H \) fixed, but also among class 2 of restricted estimators with \( \phi \) fixed.

### 4.2. Illustration

Consider a stratified case-control study with \( Y \) binary and, for simplicity, \( V^\dagger \) absent. By design, the selection probability \( \pi(W) \) is known with \( W = (Y, V) \). In this setting, Manski & Lerman (1977) proposed an estimator \( \hat{\theta}_{\text{ML}} \) solving 
\[
0 = \tilde{E} \{ \pi^{-1}R\phi_{\text{eff}}^\text{ML}(X^*; \theta)\varepsilon(\theta) \},
\]
where \( \phi_{\text{eff}}^\text{ML}(X^*; \theta) = \{ \partial g(X^*; \theta)/\partial \theta \}/[g(X^*; \theta)(1 - g(X^*; \theta))] \). Manski & McFadden (1981) proposed an estimator \( \hat{\theta}_{\text{MM}} \) maximizing the conditional loglikelihood \( \tilde{E}[R\log\{f(Y | X^*; \theta)\pi(Y, V)/\sum_{Y=0,1} f(y | X^*; \theta)\pi(y, V)\}] \), where \( f(y | X^*; \theta) = 1 - g(X^*; \theta) \) or \( g(X^*; \theta) \) if \( y = 0 \) or 1. The estimators \( \hat{\theta}_{\text{ML}} \) and \( \hat{\theta}_{\text{MM}} \), obtained by substituting \( \hat{\sigma}(W) \) for \( \pi(W) \), are asymptotically more efficient than \( \hat{\theta}_{\text{ML}} \) and \( \hat{\theta}_{\text{MM}} \), respectively (Cossets, 1981). We derive a new representation of \( \hat{\theta}_{\text{MM}} \) and apply Theorem 2 to obtain a new understanding of these estimators.

Firstly, we show in the Appendix that \( \hat{\theta}_{\text{MM}} \) is algebraically identical to \( \hat{\theta}_{\text{RRZ}}(\phi_1; \cdot; \theta, 0) \), with \( \phi_1 \) defined in Theorem 2. On the other hand, \( \hat{\theta}_{\text{ML}} \) is algebraically identical to \( \hat{\theta}_{\text{RRZ}}(\phi_{\text{eff}}^\text{ML}; \cdot; \theta, 0) \). Therefore, \( \hat{\theta}_{\text{MM}} \) and \( \hat{\theta}_{\text{ML}} \) are both of the form \( \hat{\theta}_{\text{RRZ}}(\phi, 0) \), but with different choices of \( \phi \). The estimator \( \hat{\theta}_{\text{MM}} \) is typically more efficient than \( \hat{\theta}_{\text{ML}} \), as observed in various simulation studies (e.g. Robins et al., 1994; Lawless et al., 1999).

Secondly, for the moment, replace \( \hat{\sigma}(W) \) in \( \hat{\theta}_{\text{RRZ}} \) by the known selection probability \( \pi(W) \). In this case, Theorem 2 holds with \( H^\dagger(W) \) replaced by \( H(W) \) throughout. The estimator \( \hat{\theta}_{\text{MM}} = \hat{\theta}_{\text{RRZ}}(\phi_1, 0) \) is optimal among class 3 of restricted estimators with \( H = 0 \). However, an estimator \( \hat{\theta}_{\text{RRZ}} \) is allowed outside this class if and only if \( (Y, V) \) from the phase one sample are used. Therefore, \( \hat{\theta}_{\text{MM}} \) is semiparametric efficient using only the phase two sample, a result previously shown in Robins et al. (1994, § 6.3).

Thirdly, suppose that \( V \) is discrete with \( J \) levels, 1, 2, . . . , and \( J \), model (3) is saturated, and model (1) is logistic with saturated main effects of \( V \), such as \( \logit\{g(X^*; \theta)\} = \sum_{j=1}^{J} \theta_1 j 1\{V = j\} + \sum_{j=1}^{J} \theta_2 j X^* 1\{V = j\} \), where \( \theta = (\theta_1^T, \theta_2^T)^T \), \( \theta_1 = (\theta_{11}, \ldots, \theta_{1J})^T \), and \( \theta_2 = \ldots \).
(θ^2_1, \ldots, θ^2_J)^T$. Then \( \hat{θ}_{MM} \) coincides with the nonparametric maximum likelihood estimator by Scott & Wild (1997, § 2.3), and hence is semiparametric efficient by the general results of Lawless et al. (1999) and Breslow et al. (2003). We provide a direct proof of this efficiency result in the Appendix.

5. PROPOSED METHODS

5.1. Summary

We develop unified methods to accommodate optimal restricted estimators in classes 1 and 2 or in classes 1 and 3 simultaneously. The strategy is to combine (a) and (c) or (b) and (c) of the following steps.

Step (a) For fixed \( Φ \), include in \( H \) a working estimator of \( E\{Φε(θ_0) \mid W\} \).

Step (b) For fixed \( H \), include in \( Φ \) working estimators of \( φ_1(X^*; θ_0) \) and \( φ_2(X^*; θ_0) \) in Theorem 2.

Step (c) For fixed \( Φ \) and \( H \), construct estimators that are asymptotically equivalent, to the first order, to the optimal estimator in class 1.

We defer the construction of estimators for step (c) until §§ 5.2–5.3. In § 5.1, we present the details of implementing (a) and (b), and then discuss the advantage of using the hybrid strategy provided that step (c) is fulfilled.

5.2. Steps and combinations

We posit working models to derive estimators required in steps (a) and (b). These models are not necessarily correct, but may provide reasonable approximations to the truth. To estimate \( E\{Φε(θ_0) \mid W\} \), the suggestion of Robins et al. (1994, § 2.7) is to specify a conditional mean model

\[
E\{Φε(θ_0) \mid W\} = l(W; η),
\]

where \( l(\cdot; η) \) is a \( p \times 1 \) known function and \( η \) is a finite-dimensional unknown parameter. Let \( \hat{θ} \) be a preliminary, consistent estimator of \( θ_0 \), e.g., \( \hat{θ}_{RRZ}(∂g/∂θ, 0) \), and let \( \tilde{η} \) be the least-squares estimator minimizing \( \hat{E}[R\{Φ(X^*)ε(\hat{θ}) - l(W; η)\}] \). Then \( l(W; \tilde{η}) \) serves as an estimator of \( E\{Φε(θ_0) \mid W\} \). Similarly, the unknown functions in \( φ_1(X^*; θ_0) \) and \( φ_2(X^*; θ_0) \) can be approximated by specifying

\[
E\left\{ \left( \frac{ε^2(θ_0)}{π(W)} \right) \bigg| X^* \right\} = m_1(X^*; v_1),
\]

\[
E\left\{ \frac{1 - π(W)}{π(W)} ε(θ_0) H^1(W) \bigg| X^* \right\} = m_2(X^*; v_2),
\]

where \( m_1(\cdot; v_1) \) is a known function, \( m_2(\cdot; v_2) \) is a \( k \times 1 \) known function and \( v_1 \) and \( v_2 \) are finite-dimensional unknown parameters. These models are similar to model (1), but \( v_1 \) and \( v_2 \) need to be estimated in the form of \( \hat{θ}_{RRZ} \). For example, define \( \tilde{v}_1 \) as a solution to

\[
0 = \hat{E}[\hat{r}^{-1}(W)R\{∂m_1(X^*; v_1)/∂v_1\}(\hat{r}^{-1}(W)ε^2(\hat{θ}) - m_1(X^*; v_1))] \]

and define \( \tilde{v}_2 \) similarly using \( \hat{θ}_{RRZ}(∂g/∂θ, 0) \). Then \( φ_1(X^*; θ_0) \) can be estimated by \( \{∂g(X^*; \hat{θ})/∂θ\}/m_1(X^*; \tilde{v}_1) \) and \( φ_2(X^*; θ_0) \) by \( m_2(X^*; \tilde{v}_2)/m_1(X^*; \tilde{v}_1) \).

In the case of missing regressors and no extraneous variables \( V^1 \), we suggest an alternative way to estimate \( φ_1(X^*; θ_0) \) and \( φ_2(X^*; θ_0) \). We specify and fit a conditional density model as a submodel to (1),

\[
p(y \mid X^*) = f(y \mid X^*; ψ),
\]

where \( f(y \mid X^*; ψ) \) is the unknown conditional density model and \( ψ \) is a vector of unknown parameters. The optimal estimator in class 1 coincides with the nonparametric maximum likelihood estimator by Scott & Wild (1997, § 2.3), and hence is semiparametric efficient by the general results of Lawless et al. (1999) and Breslow et al. (2003). We provide a direct proof of this efficiency result in the Appendix.
where \( p(y \mid X^*) \) is the conditional density of \( Y \) given \( X^* \), \( f(y \mid X^*; \psi) \) is a known conditional density function and \( \psi \) is a finite-dimensional unknown parameter. Let \( \hat{\psi} \) be an estimator of \( \psi \). Then we evaluate \( E[\pi^{-1}e^2(\theta_0) \mid X^*] \) as \( \int \hat{\pi}^{-1}(y, V) (y - g(X^*; \hat{\theta}))^2 f(y \mid X^*; \hat{\psi}) \, dy \) and evaluate \( E[\pi^{-1}(1 - \pi)e(\theta_0)H^1 \mid X^*] \) similarly under fitted model (7). For efficient estimation of \( \psi \), nonparametric and pseudolikelihood methods can be used (e.g. Lawless et al., 1999; Chatterjee et al., 2003; Weaver & Zhou, 2005; Song et al., 2009; Tan, 2010). On the other hand, inefficient estimators for \( \psi \) can be applied for simplicity. For the examples in \( \S \S \) 6–7 where \( Y \) is continuous, we specify model (7) as a homoscedastic normal regression model, i.e. \( Y \) given \( X^* \) is normal with mean \( g(X^*; \theta) \) and variance \( \sigma^2 \), with \( \psi = (\theta, \sigma^2) \). Given a preliminary estimator \( \hat{\theta} \), a simple estimator of \( \sigma^2 \) is \( \hat{\sigma} \). Next consider two steps for combining steps. The fixed-\( \Phi \) approach is to combine steps (a) and (c) for fixed \( \Phi \). We first choose \( \Phi \) and include in \( H \) a working estimator of \( E\{\Phi e(\theta_0) \mid W\} \) and possibly other functions. Given \( \Phi \) and \( H \), we then apply \( \hat{\theta}_{\text{GMM}} \) or \( \hat{\theta}_{\text{LIK}} \) in \( \S \S \) 5.2–5.3. The resulting estimator \( \hat{\theta} \), a hybrid estimator based on classes 1 and 2, has the following properties.

(i) The estimator \( \hat{\theta} \) is asymptotically optimal in class 1: it is asymptotically equivalent to the optimal estimator in class 1 for fixed \( \Phi \) and \( H \) even though the working model may be misspecified.

(ii) The estimator \( \hat{\theta} \) is locally optimal in class 2: it is asymptotically equivalent to the optimal estimator in class 2 for fixed \( \Phi \) if the working model is correctly specified.
In fact, $\hat{\theta}_{\text{GMM}}$ and $\hat{\theta}_{\text{LRZ}}$ are constructed to achieve property (i) in §§ 5.2–5.3. To show property (ii), suppose that model (6) is correctly specified. Then $l(W; \eta)$ is consistent for $E\{\Phi_\varepsilon(\theta_0) \mid W\}$, and $\hat{\theta}$ is asymptotically equivalent to the estimator of $\theta_0$ that would be obtained if $l(W; \eta)$ in $H$ were replaced by $E\{\Phi_\varepsilon(\theta_0) \mid W\}$. With this replacement, the optimal estimator in class 2 for fixed $\Phi$, i.e., $\hat{\theta}_{\text{RRZ}}^{\phi^T \Phi, \alpha^T E\{\Phi_\varepsilon(\theta_0) \mid W\}}$ by Theorem 1(b), belongs to class 1 and hence is identical to the optimal estimator in class 1 for fixed $\Phi$ and $H$, because class 1 is contained in class 2. Therefore, $\hat{\theta}$ is asymptotically equivalent to the optimal estimator in class 2 by property (i).

Alternatively, the fixed-$H$ approach is to combine steps (b) and (c) for fixed $H$. We first choose $H$ and include in $\Phi$ working estimators of $\phi_1(X^*; \theta_0)$ and $\phi_2(X^*; \theta_0)$ and possibly other functions. Given $\Phi$ and $H$, we then apply $\hat{\theta}_{\text{GMM}}$ or $\hat{\theta}_{\text{LRZ}}$ in §§ 5.2–5.3. The resulting estimator $\hat{\theta}$, a hybrid estimator based on classes 1 and 3, has property (i) and the following property.

(iii) The estimator $\hat{\theta}$ is locally optimal in class 3: it is asymptotically equivalent to the optimal estimator in class 3 for fixed $H$ if the working model is correctly specified.

Property (i) holds for both the fixed-$\Phi$ and fixed-$H$ approaches, depending on their respective choices of $\Phi$ and $H$. Property (iii) is similar to property (ii). To show property (iii), suppose that model (7) is correctly specified. Then $\hat{\theta}$ is asymptotically equivalent to the estimator of $\theta_0$ that would be obtained if $\phi_j(X^*; \theta, \gamma, \psi)$ in $\Phi$ were replaced by $\phi_j(X^*; \tilde{\theta}_0)$ $(j = 1, 2)$. With this replacement, the optimal estimator in class 3 for fixed $H$, i.e., $\hat{\theta}_{\text{RRZ}}^{\phi_1 + \alpha \phi_2, \beta^T H}$ by Theorem 2, belongs to class 1 and hence is identical to the optimal estimator in class 1 for fixed $\Phi$ and $H$. Therefore, $\hat{\theta}$ is asymptotically equivalent to the optimal estimator in class 3 by property (i).

The fixed-$\Phi$ approach requires $\Phi$ to be specified, whereas the fixed-$H$ approach requires $H$ to be specified. For simplicity, we suggest the following specifications: setting $\Phi$ to $\partial g(X^*; \theta) / \partial \theta$ in the fixed-$\Phi$ approach and $H$ to a vector of low-order polynomials of $W$ in the fixed-$H$ approach. See §§ 6–7 for examples.

In the remainder of § 5, we investigate two methods for constructing estimators required in step (c). Throughout, let $\Phi(X^*)$ be a $p \times 1$ $(p \geq d)$ vector of functions and $H(W)$ be a $k \times 1$ vector of functions, specified and fixed.

§ 5.3. Generalized method of moments

To accomplish step (c), a basic method is to replace $\alpha$ and $\beta$ by sample counterparts in $\hat{\theta}_{\text{RRZ}}^{\alpha^T \Phi, \alpha^T \beta^T H}$, the optimal estimator in class 1. Let $\tilde{\theta}$ be a preliminary, consistent estimator of $\theta_0$ and $\tilde{\theta}^T(W) = [H^T(W), \{\partial \pi(W; \tilde{\gamma}) / \partial \gamma^T\}/\{1 - \tilde{\pi}(W)\}]^T$. Consider the estimator (Tsiatis, 2006)

$$\hat{\theta}_{\text{GMM}} = \hat{\theta}_{\text{RRZ}}^{\hat{\alpha}^T \Phi, \hat{\alpha}^T \hat{\beta}^T H},$$

where $\hat{\alpha} = \hat{\alpha}^T \hat{\beta}$, $\hat{\beta}$ is the submatrix of the first $k$ rows in $\hat{\beta}^T = \hat{B}^{-T} \hat{C}$, and

$$\hat{\alpha} = \hat{\alpha}^T \hat{\beta}, \quad \hat{\beta} = \text{submatrix of first } k \text{ rows in } \hat{\beta}^T = \hat{B}^{-T} \hat{C}, \quad \hat{\alpha}^T \hat{\beta} = \text{submatrix of first } k \text{ rows in } \hat{\beta}^T = \hat{B}^{-T} \hat{C}.$$

Throughout, $\text{var}$ and $\text{cov}$ denote respectively sample variance and covariance and $\tilde{\Pi}^T(\tau_1 | \tau_2) = \tau_1 - \text{cov}(\tau_1, \tau_2) \tilde{\var}^{-1}(\tau_2) \tau_2$ for random vectors $\tau_1$ and $\tau_2$. A different but asymptotically equivalent estimator to $\hat{\theta}_{\text{GMM}}$ is obtained by replacing $\hat{\theta}$ by $\hat{A}$, $\hat{B}$, $\hat{C}$ and $\hat{D}$. See the discussion on adaptive extensions of $\theta_{\text{RRZ}}$ in § 3.
The estimator \( \hat{\theta}_{\text{GMM}} \) can be regarded as a generalized-method-of-moments estimator (Hansen, 1982). See the discussion after Theorem 1. We recast Tsiatis’s (2006) result on the asymptotic behaviour of \( \hat{\theta}_{\text{GMM}} \) in parallel to Theorem 1(a).

**Theorem 3 (Tsiatis 2006).** Under regularity conditions, \( \hat{\theta}_{\text{GMM}} \) is asymptotically equivalent to the first order to the optimal estimator in class 1,

\[
\hat{\theta}_{\text{GMM}} - \theta_0 = \mathcal{O}(n^{-1/2}),
\]

where \( A \) and \( D \) are defined as in Theorem 1(a), and \( \mathcal{V} = \mathcal{D}^\top \mathcal{A}^{-1} \mathcal{D} \).

The estimator \( \hat{\theta}_{\text{GMM}} \) has properties (i) and (ii) if a working estimator of \( E\{\Phi_\varepsilon(\theta_0)|W\} \) is included in \( H \) by step (a). See the discussion in §5.1. In this case, it is interesting to compare \( \hat{\theta}_{\text{GMM}} \) with the following estimator based on the optimal estimator in class 2 (Robins et al., 1994; Tsiatis, 2006),

\[
\hat{\theta}_{\text{RRZ}} \equiv \hat{\theta}_{\text{GMM},2} = \hat{\theta}_{\text{RRZ}}(\hat{\alpha}^\top, \hat{\alpha}^\top l(\cdot; \hat{\eta}), \hat{\varphi}),
\]

where \( \hat{\alpha} = \hat{\alpha} \equiv \hat{A}^{-1} \hat{D} \) as in \( \hat{\theta}_{\text{GMM}} \) except \( \hat{A} = \text{var}(\pi^{-1} R \Phi_\varepsilon(\hat{\theta}) - \pi^{-1} R - 1)l(W; \hat{\eta}) \). The estimator \( \hat{\theta}_{\text{GMM},2} \) has property (ii), but not property (i). Therefore, \( \hat{\theta}_{\text{GMM}} \) is asymptotically equivalent to \( \hat{\theta}_{\text{GMM},2} \) if model (6) is correctly specified, but always has no greater asymptotic variance than \( \hat{\theta}_{\text{GMM},2} \) if model (6) is misspecified. This difference shows an advantage of \( \hat{\theta}_{\text{GMM}} \) combined with step (a) over \( \hat{\theta}_{\text{GMM},2} \).

Similarly, there is an advantage of \( \hat{\theta}_{\text{GMM}} \) combined with step (b) over the following estimator based on the optimal estimator in class 3,

\[
\hat{\theta}_{\text{GMM},3} = \hat{\theta}_{\text{RRZ}} \left\{ \phi_1(\cdot; \hat{\theta}, \hat{\gamma}, \hat{\psi}) + \hat{\alpha}^\top_2 \phi_2(\cdot; \hat{\theta}, \hat{\gamma}, \hat{\psi}), \hat{\beta}^\top H \right\},
\]

where \( \hat{\alpha}_2 \) is the submatrix excluding the first \( d \) rows in \( \hat{A}^{-1} \hat{D} \) and \( \hat{A}, \hat{B}, \hat{C} \) and \( \hat{D} \) are sample counterparts of \( A, B, C \) and \( D \) in Theorem 2 with \( \phi_j(X^*; \theta_0) \) replaced by \( \phi_j(X^*; \hat{\theta}, \hat{\gamma}, \hat{\psi}) \) \((j = 1, 2) \). The estimator \( \hat{\theta}_{\text{GMM}} \) has property (i) and, if combined with step (b), property (iii). See the discussion in §5.1. In contrast, \( \hat{\theta}_{\text{GMM},3} \) has property (iii), but not property (i).

Incidentally, when model (3) is not always correctly specified, \( \hat{\theta}_{\text{GMM},2} \) is doubly robust but \( \hat{\theta}_{\text{GMM}} \) is not. See Robins & Rotnitzky (2001) for a discussion on double robustness. In fact, \( \hat{\theta}_{\text{GMM},2} \) remains consistent if either model (3) or (6) is correctly specified, whereas \( \hat{\theta}_{\text{GMM}} \) is in general inconsistent if model (3) is misspecified but model (6) is correctly specified. To overcome this limitation, an improved estimator can be derived as in Tan (2006) such that it is doubly robust and, if model (3) is correctly specified, asymptotically equivalent to \( \hat{\theta}_{\text{GMM}} \). The key idea is to replace \( \hat{B} \) by \( \hat{B} = \text{cov}(\hat{\pi}^{-1} R \hat{H}^\top, (\pi^{-1} R - 1) \hat{H}^\top) \). Because our focus here is efficient estimation with model (3) correctly specified, we leave further investigation of doubly robust estimation to future work.

We provide an interesting variant of \( \hat{\theta}_{\text{GMM}} \), which is constructed similarly as \( \hat{\theta}_{\text{GMM}} \) but without taking account of the variation of \( \hat{\pi} \) as in Tan (2006). Define

\[
\hat{\theta}_{\text{GMM}}^{(m)} = \hat{\theta}_{\text{RRZ}} \left\{ \hat{\alpha}_m^\top \Phi, \hat{\alpha}_m^\top \hat{\beta}_m^\top H \right\},
\]
where $\hat{\alpha}^{(m)} = \hat{A}^{(m)} - \hat{D}$, $\hat{\beta}^{(m)} = \hat{B}^{(m)} - \hat{C}^{(m)}$ and $\hat{A}^{(m)}$, $\hat{B}^{(m)}$, and $\hat{C}^{(m)}$ are defined as $\hat{A}$, $\hat{B}$ and $\hat{C}$ but $H$ is substituted for $\hat{H}$ throughout. By similar arguments as in Theorem 3, $\hat{\theta}^{(m)}_{GMM}$ has the asymptotic expansion

$$\hat{\theta}^{(m)}_{GMM} - \theta_0 = -\gamma^{(m)} - D^{T} A^{(m)-1} E \left( \prod_{1} \Pi \left\{ \frac{R}{\pi} \Phi \epsilon(\theta_0) \right\} \left( \frac{R}{\pi} - 1 \right) H^{\top} s(\gamma_0) \right) + o_p(n^{-1/2}),$$

where $A^{(m)}$ is defined as $A$ that $H$ is substituted for $H^{\top}$ and $\gamma^{(m)} = D^{T} A^{(m)-1} D$. The influence function of $\hat{\theta}^{(m)}_{GMM}$ involves the projection on $\{(\pi^{-1} R - 1) H^{\top}, s(\gamma_0)\}^{\top}$ jointly, whereas that of $\hat{\theta}^{(m)}_{GMM}$ involves the projection on $(\pi^{-1} R - 1) H$ and $s(\gamma_0)$ sequentially. The estimator $\hat{\theta}^{(m)}_{GMM}$ has property (ii) if an estimator of $E[\Phi \epsilon(\theta_0) \mid W]$ is included in $H$. However, $\hat{\theta}^{(m)}_{GMM}$ does not have either property (i) or, even if estimators of $\phi_1$ and $\phi_2$ are included in $\Phi$, property (iii). Nevertheless, if the true $\pi$ is used instead of $\hat{\pi}$ throughout, then $\hat{\theta}^{(m)}_{GMM}$ has property (i) and, if estimators of $\phi_1$ and $\phi_2^{(m)}$ are included in $\Phi$, property (iii), where $\phi_2^{(m)}$ is defined as $\phi_2$ except $H$ is substituted for $H^{\top}$.

5.4. Nonparametric and empirical likelihood

To accomplish step (c), we adopt the methods of nonparametric likelihood (Tan, 2006) and of empirical likelihood (Qin & Lawless, 1994) to derive an estimator that is asymptotically equivalent, to the first order, to the optimal estimator in class 1. The two methods are conceptually different, but lead to the same inferences for $\theta_0$.

To adopt the method of Tan (2006), write the likelihood as

$$L_1 \times L_2 = \left[ \prod_{i=1}^{n} \pi(W_i; \gamma)^{R_i} \right] \times \left[ \prod_{i=1}^{n} G_1(Z_i)^{R_i} G_0(W_i)^{1-R_i} \right],$$

where $G_1$ is the joint distribution of $Z$ and $G_0$ is that of $W$, subject to two types of constraints. Firstly, $G_0$ and the marginal distribution of $W$ under $G_1$ are identical, regardless of model (1). The resulting constraints are called inherent constraints. Secondly, the conditional distribution of $Y$ given $X^*$ under $G_1$ satisfies model (1). The resulting constraints are called modelling constraints.

Maximizing $L_1$ leads to the maximum likelihood estimator $\hat{\gamma}$. To maximize $L_2$, we choose to retain only finitely many constraints on $(G_1, G_0)$,

$$\int \hat{\pi}(w) \, dG_1 + \int \{1 - \hat{\pi}(w)\} dG_0 = 1,$$

$$\int \{1 - \hat{\pi}(w)\} H(w) \, dG_1 = \int \{1 - \hat{\pi}(w)\} H(w) dG_0,$$

$$\int \frac{\partial \pi}{\partial \gamma}(w; \hat{\gamma}) \, dG_1 = \int \frac{\partial \pi}{\partial \gamma}(w; \hat{\gamma}) \, dG_0,$$

$$\int \Phi(x^*) \{y - g(x^*; \theta)\} \, dG_1 = 0.$$  

Equations (8)–(10) give inherent constraints, whereas (11) gives modelling constraints. In addition, we require that $G_1$ be a nonnegative measure supported on $\{Z_i : R_i = 1, i = 1, \ldots, n\}$, and $G_0$ be a nonnegative measure supported on $\{W_i : R_i = 0, i = 1, \ldots, n\}$. The conditions $\int dG_1 = 1$ and $\int dG_0 = 1$ are not necessarily imposed, but can be accommodated by (8) and (9), depending on $H$. If 1 is included in $H$, then (9) contains $\int \{1 - \hat{\pi}(w)\} \, dG_1 = \int \{1 - \hat{\pi}(w)\} \, dG_0.$
and hence (8) yields \( \int dG_1 = 1 \). If \( \hat{\pi} / (1 - \hat{\pi}) \) is included in \( H \), then (9) contains \( \int \hat{\pi}(w) dG_1 = \int \hat{\pi}(w) dG_0 \) and hence (8) yields \( \int dG_0 = 1 \). We typically include 1 in \( H \) to incorporate the condition \( \int dG_1 = 1 \), but exclude \( \hat{\pi} / (1 - \hat{\pi}) \) in \( H \) to relax the condition \( \int dG_0 = 1 \). This relaxation is needed to handle the case where \( \pi(W) \) is not bounded away from 1.

Define the profile nonparametric likelihood function as

\[
pl_n(\theta) = \max_{G_1, G_0} L_2
\]

subject to constraints (8)–(11), and the nonparametric likelihood estimator of \( \theta_0 \) as

\[
\hat{\theta}_{\text{LIK}} = \arg\max pl_n(\theta).
\]

Theorem 4 presents a solution to the constrained maximization of \( L_2 \). Recall that \( \hat{H}^\dagger = [H^\top, (\partial \pi(\cdot; \hat{\gamma}) / \partial \gamma^\top)]^\top \) and hence (1) yields

\[
\hat{L}_{\text{LIK}} = (H^\top, \partial \pi(\cdot; \hat{\gamma}) / \partial \gamma^\top)^\top.
\]

**Theorem 4.** Assume that there exist no \( \lambda \) and \( \varrho \) such that \( \lambda^\top [1 - \hat{\pi}(W)] \hat{H}^\dagger(W_i) + \varrho^\top \Phi(X_i^\gamma) \epsilon_i(\theta) = 0 \) if \( R_i = 1 \) and \( \lambda^\top [1 - \hat{\pi}(W)] \hat{H}^\dagger(W_i) = 0 \) if \( R_i = 0 \) (i = 1, ..., n). Consider the following function

\[
\ell_n(\lambda, \varrho; \theta) = E \left[ R \log \left[ \hat{\pi}(W) + \lambda^\top [1 - \hat{\pi}(W)] \hat{H}^\dagger(W) + \varrho^\top \Phi(X^\gamma) \epsilon(\theta) \right] \right] + (1 - R) \log \left[ 1 - \hat{\pi}(W) - \lambda^\top [1 - \hat{\pi}(W)] \hat{H}^\dagger(W) \right],
\]

where the log of 0 or a negative number is \(-\infty\). For fixed \( \theta \), if \( \ell_n \) achieves a maximum at \( \lambda = \hat{\lambda}(\theta) \) and \( \hat{\gamma} = \hat{\gamma}(\theta) \), then \( L_2 \) achieves the constrained maximum at

\[
\hat{G}_1([Z_i]; \theta) = \frac{n^{-1}}{\hat{\pi}(W_i) + \hat{\lambda}^\top [1 - \hat{\pi}(W_i)] \hat{H}^\dagger(W_i) + \hat{\varrho}^\top \Phi(X_i^\gamma) \epsilon_i(\theta)} \quad (R_i = 1),
\]

\[
\hat{G}_0([W_i]; \theta) = \frac{n^{-1}}{1 - \hat{\pi}(W_i) - \hat{\lambda}^\top [1 - \hat{\pi}(W_i)] \hat{H}^\dagger(W_i)} \quad (R_i = 0).
\]

Consequently, \( \log pl_n(\theta) \) equals \(-n \ell_n(\hat{\lambda}(\theta), \hat{\gamma}(\theta)) - n \log(n)\).

We provide an alternative derivation of \( \hat{\theta}_{\text{LIK}} \) by using the method of empirical likelihood (Qin & Lawless, 1994). A prerequisite for this derivation is to recognize \( \hat{\pi}^{-1} R \Phi \epsilon(\theta) \) and \( (\hat{\pi}^{-1} R - 1) H \) as asymptotically unbiased estimating functions based on observed data. In contrast, the nonparametric likelihood derivation involves characterizing constraints (8)–(11) on the distributions of full and coarsened data (\( G_1, G_0 \)). Conceptually, the relationship between the two derivations is similar to that between observed- and full-data estimating functions in semiparametric theory of estimation with missing data (Van der Laan & Robins, 2003; Tsiatis, 2006). To our knowledge, the estimator \( \hat{\theta}_{\text{LIK}} \) and both of the derivations are new.

The empirical likelihood is

\[
L = \prod_{i=1}^{n} P([O_i]),
\]

where \( P \) is the joint distribution of \( O = (Z1[R = 1], W1[R = 0], R) \). Define the profile empirical likelihood function as \( \max_P L \), where \( P \) is a probability measure supported on
{O_i : i = 1, \ldots, n} and satisfies the constraints
\[
\int \left\{ \frac{r}{\hat{\pi}(w)} - 1 \right\} H(w) \, dP = 0, \\
\int \frac{r - \hat{\pi}(w)}{\hat{\pi}(w)[1 - \hat{\pi}(w)]} \frac{\partial \pi}{\partial \gamma}(w; \gamma) \, dP = 0, \\
\int \frac{r}{\hat{\pi}(w)} \Phi(x^*)[y - g(x^*; \theta)] \, dP = 0.
\]

For fixed \( \theta \), the constrained maximizer of \( L \) is, (Qin & Lawless, 1994)
\[
\hat{P}(O_i; \theta) = \frac{n^{-1}}{1 + \hat{\lambda}^T \{\hat{\pi}^{-1}(W_i)R_i - 1\} \hat{H}^\top(W_i) + \hat{\theta}^\top \hat{\pi}^{-1}(W_i)R_i \Phi(X_i^\top)e_i(\theta)},
\]
where \((\hat{\lambda}, \hat{\theta})\) is a maximizer of the function
\[
\bar{E} \left( \log \left[ 1 + \lambda^T \{\hat{\pi}^{-1}(W)R - 1\} \hat{H}^\top(W) + \phi^\top \hat{\pi}^{-1}(W)R \Phi(X^\phi)\varepsilon(\theta) \right] \right) \\
= \ell_n(\lambda, \phi) - \bar{E} \left[ R \log \hat{\pi}(W) + (1 - R) \log \{1 - \hat{\pi}(W)\} \right]
\]
and hence equivalently a maximizer of \( \ell_n(\lambda, \phi) \) in Theorem 4. Therefore, the profile empirical likelihood function of \( \theta \) is proportional to \( pl_n(\theta) \), and the empirical likelihood estimator is identical to \( \hat{\theta}_{\text{LIK}} \) as a maximizer of \( pl_n(\theta) \).

Theorem 5 summarizes the asymptotic properties of the profile likelihood and the likelihood estimator. The results extend the asymptotic theory for empirical likelihood (Qin & Lawless, 1994) by accommodating nuisance parameters \( \gamma \) due to missing data.

**Theorem 5.** In addition to Assumptions 1–2 in Newey & Smith (2004) for estimating functions \( \Phi(X^\phi)\varepsilon(\theta) \), assume that \( \hat{\gamma} - \gamma_0 = \text{var}^{-1}\{s(\gamma_0)\} \bar{E}\{s(\gamma_0)\} + o_p(n^{-1/2}) \), \( \partial \pi / \partial \gamma \) is uniformly bounded in a neighbourhood of \( \gamma_0 \), and \( E\{\|H(W)\|^d\} < \infty \) for some constant \( a > 2 \). Then \( \hat{\theta}_{\text{LIK}} \) is asymptotically equivalent to the first order to the optimal estimator in class 1,
\[
\hat{\theta}_{\text{LIK}} - \theta_0 = -\gamma^{-1} D^\top A^{-1} \bar{E} \left[ \Pi^\perp \left\{ \frac{R}{\pi} \Phi e(\theta_0) \mid \left( \frac{R}{\pi} - 1 \right) H^\top \right\} \right] + o_p(n^{-1/2}),
\]
where \( A, D, \) and \( \gamma \) are defined as in Theorem 3. Moreover, \( -n^{-1} \partial^2 \log pl_n(\theta) / (\partial\theta \partial\theta^T) \mid_{\theta = \hat{\theta}_{\text{LIK}}} \) is a consistent estimator of \( \gamma \). Finally, \( -2 \log \{pl_n(\theta_0) / pl_n(\hat{\theta}_{\text{LIK}})\} \) is asymptotically chi-squared with \( d \) degrees of freedom.

We compare our method with the empirical likelihood method of Chen et al. (2008), which is applicable in our general setting of missing data such as missing response or missing regressors, but for the just-determined case where \( \Phi \) is of dimension \( p = d \). Label the sample such that \( R_i = 1 \) for \( i = 1, \ldots, n_1 \) and \( R_i = 0 \) otherwise. Write \( p_i = \text{pr}(Z_i | R_i = 1) \), \( q_j = \text{pr}(W_j | R_j = 0) \) for \( j = n_1 + 1, \ldots, n \) and \( \mu = E \{[1 - \pi(W)]H(W)\} \). Their estimator \( \hat{\theta}_{\text{COL}} \) is a solution to
\[
\sum_{i=1}^{n_1} \hat{p}_i \frac{\Phi(X_i^\phi)e_i(\theta)}{\hat{\pi}(W_i)} = 0,
\]
where \((\hat{\theta}_1, \ldots, \hat{\theta}_n, \hat{\lambda}_1, \ldots, \hat{\lambda}_n, \hat{\mu})\) is a maximizer of \(\prod_{i=1}^{n_1} p_i \prod_{j=n_1+1}^{n} q_j\) subject to \(p_i > 0\), \(q_j > 0\), \(\sum_{i=1}^{n_1} p_i = 1\), \(\sum_{j=n_1+1}^{n} q_j = 1\) and

\[
\sum_{i=1}^{n_1} p_i \frac{(1 - \hat{\pi}(W_i)) H(W_i) - \mu}{\hat{\pi}(W_i)} = 0, \quad \sum_{j=n_1+1}^{n} q_j \frac{(1 - \hat{\pi}(W_j)) H(W_j) - \mu}{1 - \hat{\pi}(W_j)} = 0.
\]

Theorem 1 of Chen et al. (2008) suggests that

\[
\hat{\theta}_{CLQ} - \theta_0 = -D^T \hat{E} \left\{ \prod_{i=1}^{n_1} \left( \prod_{j=1}^{n} \left[ \frac{R}{\pi} \Phi \epsilon(\theta_0) \right] \left( \frac{R}{\pi} - \frac{1 - R}{1 - \pi} \right) \{(1 - \pi)H - \mu \} \right) s(\gamma_0) \right\} + o_p(n^{-1/2}),
\]

where \(D\) is defined as in Theorem 1(a). The estimator \(\hat{\theta}_{CLQ}\) is asymptotically at least as efficient as the estimator \(\hat{\theta}_{IPW}\) solving \(0 = \hat{E}(\pi^{-1} R \Phi \epsilon(\theta))\) with the true \(\pi\) used, but not necessarily as the estimator \(\hat{\theta}_{IPW}\) solving \(0 = \hat{E}(\hat{\pi}^{-1} R \Phi \epsilon(\theta))\) with \(\hat{\pi}\) used or as \(\hat{\theta}_{RZ}(\Phi, b^T H)\) for a \(k \times d\) matrix \(b\) of constants, i.e., \(\hat{\theta}_{CLQ}\) does not have property (i). If an estimator of \(E(\Phi \epsilon(\theta) \mid W)\) is included in \(H\) by step (a), then \(\hat{\theta}_{CLQ}\) has property (ii). In contrast, \(\hat{\theta}_{LIK}\) has property (i) and, if combined with step (a), property (ii). See the discussion in § 5.1. By property (i), \(\hat{\theta}_{LIK}\) is asymptotically at least as efficient as \(\hat{\theta}_{IPW}, \hat{\theta}_{IPW},\) and \(\hat{\theta}_{RZ}(\Phi, b^T H)\). Moreover, if \(1 - \hat{\pi}^{-1}\) is included in \(H\), then \(\hat{\theta}_{LIK}\) is asymptotically at least as efficient as \(\hat{\theta}_{CLQ}\).

We point out that a variant of \(\hat{\theta}_{LIK}\), in parallel to the variant \(\hat{\theta}_{GMM}\) of \(\hat{\theta}_{GMM}\), is asymptotically similar to \(\hat{\theta}_{CLQ}\) but algebraically more straightforward. See Tan (2006) for a related construction. Define \(\hat{\theta}_{(m)}\) as \(\arg\max\, pl_{(m)}(\theta)\), where \(pl_{(m)}(\theta) = \max_{G, \epsilon_0} L_2\) subject to constraints (8), (9) and (11), but not (10). In the just-determined case \(p = d, \hat{\theta}_{(m)}\) is equivalently a solution to

\[
\sum_{i=1}^{n_1} \hat{G}_{(m)}(Z_i) \Phi(X_i^T) \epsilon_i(\theta) \hat{\pi}(W_i) = 0,
\]

where \((\hat{G}_{(m)}, \hat{G}_{0}(m))\) is a maximizer of \(L_2\) subject to constraints (8)--(9), free of \(\theta\). A modification of Theorem 4 yields

\[
\hat{G}_{(m)}(Z_i) = \frac{n^{-1}}{\hat{\pi}(W_i) + \hat{\lambda}(m)^T (1 - \hat{\pi}(W_i)) H(W_i)} (R_i = 1),
\]

\[
\hat{G}_{0}(m)(W_i) = \frac{n^{-1}}{1 - \hat{\pi}(W_i) - \hat{\lambda}(m)^T (1 - \hat{\pi}(W_i)) H(W_i)} (R_i = 0),
\]

where \(\hat{\lambda}(m)\) is a maximizer of the function

\[
\ell_n(m)(\lambda) = \hat{E} \{R \log[\hat{\pi}(W)] + \hat{\lambda}^T (1 - \hat{\pi}(W)) H(W)\} + (1 - R) \log[1 - \hat{\pi}(W) - \hat{\lambda}^T (1 - \hat{\pi}(W)) H(W)].
\]

By similar arguments as in Theorem 5, \(\hat{\theta}_{(m)}\) has the asymptotic expansion

\[
\hat{\theta}_{(m)} - \theta_0 = -D^T \hat{E} \left\{ \prod_{i=1}^{n_1} \left( \prod_{j=1}^{n} \left[ \frac{R}{\pi} \Phi \epsilon(\theta_0) \right] \left( \frac{R}{\pi} - 1 \right) H \right) s(\gamma_0) \right\} + o_p(n^{-1/2}).
\]
The asymptotic expansion is similar to that of \( \hat{\theta}_{\text{CLQ}} \), but involves projection of \( \pi^{-1} R \Phi \varepsilon (\theta_0) \) on \( (\pi^{-1} R - 1) H \) instead of on \( (\pi^{-1} R - 1) \{ H - (1 - \pi)^{-1} \mu \} \). This difference reflects why the asymptotic expansion of \( \hat{\theta}^{(m)}_{\text{LIK}} \) holds under assumption (2), whereas that of \( \hat{\theta}_{\text{CLQ}} \) holds only under the assumption that \( \pi (W) \) is bounded away from both 0 and 1 (Chen et al., 2008, Condition 1), in addition to regularity conditions.

6. Simulation study

Lawless et al. (1999) conducted a simulation study to compare various estimators using only discrete phase-one data under condition density model (7). We adopt their simulation design, but consider condition mean model (1) with different disturbance distributions and accommodate continuous phase-one data.

Let \( X^* = (X, V) \) be jointly normal with means 0, variances 1, and correlation \( \rho \) and \( Y = \theta_0 + \theta_1 X + \theta_2 V + \varepsilon \), where \( \varepsilon \) is independent of \( X^* \). Consider the following specifications: surrogate covariate \( V \) with \( \rho = 0-9 \) and \( \theta_2 = 0 \); expensive covariate \( V \) with \( \rho = 0-3 \) and \( \theta_2 = 0-5 \); nonnull effect of \( X \) with \( \theta_0 = 1 \) and \( \theta_1 = 0 \); nonnull effect of \( X \) with \( \theta_0 = 1 \) and \( \theta_1 = 1 \); normal disturbance \( \varepsilon \) from \( N(0, 1) \); and nonnormal disturbance \( \varepsilon \) with variance 1 from a mixture of \( \text{Un} (-\sqrt{3}, \sqrt{3}) \) and \( N(0, 1) \) with proportions 0.9 and 0.1. Suppose that \( W = (Y, V) \) is measured for the phase-one sample and \( X \) is measured only for the phase-two sample, which is based on three strata, \( \{ Y \leq a_1 \} \), \( \{ a_1 < Y \leq a_2 \} \), and \( \{ Y > a_2 \} \). The breaks \( a_1 \) and \( a_2 \) are set to the 5% and 95% quantiles of \( Y \). The selection probabilities are set to \( \pi_1 = 0.25 \), \( \pi_2 = 0.014 \), and \( \pi_3 = 0.25 \), so that the numbers of units selected at the second phase are approximately the same, for example, 250 with \( n = 20,000 \), in the three strata.

Consider model (1) with \( g(X^*; \theta) = \theta_0 + \theta_1 X \) in the surrogate-covariate problem or \( \theta_0 + \theta_1 X + \theta_2 V \) in the expensive-covariate problem, and model (3) with \( \pi (W; \gamma) = \gamma_1 \{ Y \leq a_1 \} + \gamma_2 \{ a_1 < Y \leq a_2 \} + \gamma_3 \{ Y > a_2 \} \). Take model (7) as (1) with the assumption that \( \varepsilon \) is normal with constant variance. Therefore, model (7) is correctly specified in the normal-disturbance problem, but not in the nonnormal-disturbance problem. We compare six estimators labelled as follows.

**Estimator 1.** The nonparametric maximum likelihood estimator of Lawless et al. (1999) under model (7) with post-stratification by \( Y_{\text{DIS}} \) and \( V_{\text{DIS}} \), where \( Y_{\text{DIS}} \) is the discretization of \( Y \) into three levels, 1, 2 and 3, with the breaks \( (a_1, a_2) \) and \( V_{\text{DIS}} \) is the discretization of \( V \) into six levels, 1, \ldots , 6, with the breaks \( (-2, -1, 0, 1, 2) \).

**Estimator 2.** The estimator \( \hat{\theta}_{\text{IPW}} \) solving \( 0 = \tilde{E} \{ \pi^{-1} R \Phi \varepsilon (\theta) \} \) with \( \Phi = \partial g(X^*; \theta)/\partial \theta \).

**Estimator 3.** The estimator \( \hat{\theta}_{\text{GMM}} \) with \( \Phi = \partial g(X^*; \theta)/\partial \theta \) and \( H = (1, Y, Y^2, V, V^2, YV)^T \).

**Estimator 4.** The estimator \( \hat{\theta}_{\text{GMM}} \) with \( \Phi = (\phi_1, \phi_2)^T \) estimated from model (7) and \( H = (1, Y, Y^2, V, V^2, YV)^T \).

**Estimators 5–6.** The estimator \( \hat{\theta}_{\text{LIK}} \) with \( \Phi \) and \( H \) as in (3–4), respectively.

Estimators 3 and 5 are applications of the fixed-\( \Phi \) approach with \( l(W; \eta) = \eta^T H(W) \) in model (6). In fact, these estimators remain unchanged if \( H \) is replaced by \( \{(\tilde{\eta}^T H)^T, Y^2, V, V^2, YV\}^T \) or \( \{(\tilde{\eta}^T H)^T, V, V^2, YV\}^T \) for two-dimensional or three-dimensional \( \theta \). On the other hand, estimators 4 and 6 are applications of the fixed-\( H \) approach.

Figures 1 and 2 present the boxplots of the six estimators from 1000 Monte Carlo samples of size \( n = 20,000 \) as in Lawless et al. (1999). We summarize the findings as follows. Estimator 1 performs satisfactorily if model (7) is correctly specified, but yields considerable biases in some cases if model (7) is misspecified. If model (7) is correctly specified, then Estimator 1 is the
Fig. 1. Boxplots of estimators $\theta_0$ and $\theta_1$ in the surrogate-covariate problem for four specifications arranged in columns.

Fig. 2. Boxplots of estimators of $\theta_0$ and $\theta_1$ in the expensive-covariate problem for four specifications arranged in columns.

least variable among all the estimators except for the surrogate-covariate problem with normal or non-normal disturbances and $\theta_1 = 0$. In these cases, Estimators 4 and 6 yield smaller mean squared errors by a factor of about 1.5 than Estimator 1. Therefore, the estimator of Lawless et al. (1999), designed for conditional density models with discrete phase-one data, may suffer biases under conditional mean models or lose efficiency if continuous phase-one data are discretized.

Estimator 2 has negligible biases, but large variances. However, Estimators 3–6 have smaller variances than Estimator 2. Moreover, Estimators 4 and 6 in the fixed-$H$ approach are no more variable than Estimators 3 and 5 in the fixed-$\Phi$ approach. The efficiency gain of the fixed-$H$ over fixed-$\Phi$ approach is most evident for the estimation of $\theta_1$ in the surrogate-covariate problem with normal or nonnormal disturbances and $\theta_1 = 0$. In these cases, Estimators 4 and 6 have smaller
mean squared errors by a factor of 8–10 than those of Estimators 3 and 5. Therefore, the fixed-$H$ approach, designed to approximate the optimal estimator in class 3, can be substantially more efficient than the fixed-$\Phi$ approach, designed to approximate the optimal estimator in class 2, even though the working models may be misspecified in both approaches.

For the fixed-$H$ approach, the generalized-method-of-moments estimator 4 has appreciable biases in some settings, Estimator 6 yields no greater, sometimes much smaller, biases than Estimator 4. The biases of Estimators 4 and 6 are most noticeable in the expensive-covariate problem with normal or nonnormal disturbances and $\theta_1 = 1$. In these cases, the biases of Estimator 4 or Estimator 6 are, respectively, 20–40 or 8–16% of the square roots of mean squared errors. The likelihood method shows a marked improvement over the generalized method of moments, although the biases are not completely removed. Such results may be caused by high correlations of the estimating functions for Estimators 4 and 6.

7. REAL-DATA EXAMPLE

Addy et al. (1994) discussed an epidemiologic study of adolescent depression designed with two-phase sampling. The phase-one sample consisted of 3189 7th and 8th graders in a school district in southeastern U.S.A. All students were screened in classroom settings using self-administered questionnaires on demographic information, depressive feelings and family adaptability and cohesion. Then three strata were formed, based on the screening results. The first stratum consisted of 316 students with depression scores $\geq 30$. The second and third strata consisted of, respectively, 459 black and 2414 white students with depression scores $< 30$. At the second phase, a stratified sample of students was drawn, with the selection probability approximately 54%, 17%, or 8.6% in the first, second or third stratum. Each student and one parent were interviewed in the home using clinical protocols for diagnosis of mental disorders. Addy et al. (1994) studied a logistic regression model for clinical diagnosis of depression given race, gender, family cohesion and guardian status.

For illustration, we conduct a regression analysis taking self-reported depression score as the response variable, $Y$, and race, gender, and family cohesion as the explanatory variables, $X^*$, and ignoring data on family cohesion except for the second-phase sample. This set-up involves an always-observed continuous outcome, depression score and a partially observed regressor, cohesion score. In addition, this setup allows us to obtain an otherwise infeasible estimator based on phase-one data including family cohesion and to assess the proposed estimators against such an estimator.

In our analysis, model (1) is a linear regression model with the intercept and main effects of $X^* = (X, V)$, where $X$ encodes cohesion score, $V$ consists of race with white $= 0$ and black $= 1$ and gender with male $= 0$ and female $= 1$. Model (3) is specified with saturated effects of two binary factors, $1\{Y \geq 30\}$ and race, where $R$ is the selection indicator for the phase-two sample and $W = (Y, V)$. Model (7) is a homoscedastic normal regression model with the regression mean as in model (1), i.e. $\varepsilon$ is normal with constant variance. We compare seven estimators labelled as follows.

*Estimator 0.* The infeasible, ordinary least squares estimator based on phase-one data.

*Estimator 1.* The nonparametric maximum likelihood estimator of Lawless et al. (1999) under (7) with post-stratification by $1\{Y \geq 30\}$, race, and gender.

*Estimator 2.* The estimator $\hat{\theta}_{IPW}$ solving $0 = \hat{E}\{\hat{\pi}^{-1}R\Phi\varepsilon(\theta)\}$.

*Estimator 3.* The estimator $\hat{\theta}_{GMM}$ with $\Phi = \partial g(X^*; \theta)/\partial \theta$ and $H$ consisting of $\delta_{ij}, \delta_{ij}Y$, and $Y^2$, where $\delta_{ij} = 1\{\text{race} = i, \text{gender} = j\}$ for $i, j = 0, 1$. 

*Estimator 4.* The generalized-method-of-moments estimator (GMM).

*Estimator 5.* The fixed-$\Phi$ estimator.

*Estimator 6.* The fixed-$H$ estimator.
Estimator 4. The estimator $\hat{\theta}_{\text{GMM}}$ with $\Phi = (\phi_1^T, \phi_2^T)$ estimated from model (7) and $H$ as in (3).

Estimators 5–6. The estimator $\hat{\theta}_{\text{LIK}}$ with $\Phi$ and $H$ as in (3–4) respectively.

Estimators 1–6 are parallel to those in § 6. Estimators 3 and 5 are applications of the fixed-$\Phi$ approach, whereas Estimators 4 and 6 are applications of the fixed-$H$ approach.

Figure 3 presents the resulting estimates and 95% Wald confidence intervals. We observe the following comparisons of Estimates 1–10, using Estimate 0 as the gold standard. Estimate 1 is noticeably biased, which points to inappropriateness of the assumptions of normality and homoscedasticity in model (7). Estimate 2 is associated with large standard errors. Estimates 3–6 have considerably smaller standard errors than Estimate 2. Among Estimates 2–6, Estimates 4 and 6 have the smallest standard errors. The estimated variances of Estimators 4 and 6 for the coefficient of family cohesion are only four times greater than that of Estimator 0, even though the overall proportion of selection into the second-phase sample is 14%, less than 25%. The 95% confidence intervals for the coefficient of race exclude 0 for Estimates 0, 4 and 6, but not for other estimates. There is significant evidence at the 5% level for association of depressive feelings with race, after adjusting for gender and family cohesion.

8. Discussion

Consider a conditional mean model with missing data on the response or explanatory variables. Assume that a selection probability model is correctly specified. Our overall suggestion is to use hybrid estimators based on classes 1 and 2 in the fixed-$\Phi$ approach or classes 1 and 3 in the fixed-$H$ approach. In our simulation study, the fixed-$H$ approach appears to yield no greater, sometimes much smaller, variances than the fixed-$\Phi$ approach. In general, the relative performance of the fixed-$\Phi$ and -$H$ approaches depends on, respectively, the choices of $\Phi$ and $H$. For future work, it would be interesting to study iterative schemes of choosing $\Phi$ and $H$.

For class 1, the likelihood estimator is asymptotically equivalent to the first order to the generalized method of moments estimator. However, the likelihood estimator is expected to yield smaller biases than the latter estimator, as observed in the simulation study and supported by related theoretical and empirical results (e.g., Newey & Smith, 2004). A challenging issue is that even the likelihood estimator may deteriorate if the estimating functions from class 1 are highly correlated. This issue is applicable to estimating functions in general and warrants further research.

Acknowledgement

The author thanks the editor, the associate editor and two referees for valuable comments leading to substantial improvements in the article. This research was supported by the U.S. National Science Foundation.
APPENDIX

Proof of Theorem 2. First, we show that $A^{-1}D$ takes the form $(I_d, \alpha_i^2)\varepsilon$. Write $\tilde{A} = \text{var}(\pi^{-1}R\Phi(\theta_0))$ and $\tilde{C} = \text{cov}^2(\pi^{-1}R\Phi(\theta_0), (\pi^{-1}R - 1)H^1)$. By direct calculations,

$$
\tilde{A} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad D = \begin{pmatrix} A_{11} \\ A_{21} \end{pmatrix}, \quad \tilde{C} = (A_{21}, A_{22}),
$$

where $A_{11} = \text{var}(\pi^{-1}R\phi(\theta_0))$, etc. By matrix inversion, $A^{-1} = (\tilde{A} - \tilde{C}^\dagger B^{-1}\tilde{C})^{-1} = \tilde{A}^{-1} + \tilde{A}^{-1}\tilde{C}^\dagger (B - \tilde{C}\tilde{A}^{-1}\tilde{C}^\dagger)\tilde{A}^{-1} = \tilde{A}^{-1} + (0, I_m)^t(B - A_{22}^{-1})^{-1}(0, I_m)$ so that $A^{-1}D = (I_d, 0)^t + (0, I_m)^t(B - A_{22})^{-1}A_{21}$, where $m = \dim(H^1)$. Therefore, $a_2 = (B - A_{22})^{-1}A_{21}$.

Secondly, we show that $a_2 = B^{-1}C$. Note that $C = \tilde{C}(I_d, \alpha_i^2)^\varepsilon$. Therefore, $B^{-1}C = B^{-1}A_{21} + A_{22}(B - A_{22})^{-1}A_{21} = (B - A_{22})^{-1}A_{21}$, where the last equality follows by writing $A_{22}$ preceding $(B - A_{22})^{-1}$ as $B - (B - A_{22})$.

Finally, we show that $\hat{\theta}_{haz}(\phi, \beta^H)$ attains the minimum asymptotic variance among class 3 of restricted estimators. Let $\phi(X^\varepsilon)$ be a $d \times 1$ vector of functions and $\beta^H$ be the submatrix of the first $m$ rows in $B^{-1}C^\varepsilon$ where $C^\varepsilon = \text{cov}^2(\pi^{-1}R\phi(\theta_0), (\pi^{-1}R - 1)H^1)$. It is sufficient to prove that the asymptotic variance of $\hat{\theta}_{haz}(\phi, \beta^H)$ is no smaller than that of $\hat{\theta}_{haz}(\phi, \beta^H)$. The influence function of $\hat{\theta}_{haz}(\phi, \beta^H)$ is $E\{-\phi'(\partial g/\partial \theta)^\varepsilon\}[\pi^{-1}R\phi(\theta_0) - C^\dagger B^{-1}(\pi^{-1}R - 1)H^1]$. By direct calculations,

$$
cov\left\{ E\left( \frac{\partial g}{\partial \theta} \mid R = 1, X^\varepsilon \right) \right\} = \frac{1 - g}{\pi(1, V)} Y - \frac{g}{\pi(0, V)} (1 - Y)\left(\tilde{g}\frac{\hat{\pi}(1, V)}{\hat{\pi}(0, V)} + (1 - \hat{g})\frac{\hat{\pi}(0, V)}{\hat{\pi}(0, V)}\right).
$$

The algebraic equivalence follows by direct calculations.

Next, we show that if (a) $V$ is discrete, (b) model (3) is saturated and (c) model (1) is logistic with saturated main effects of $V$, then $\hat{\theta}_{haz}(\phi_1, 0)$ is semiparametric efficient. By assumption (b), $\partial g/\partial \gamma$ and hence $H^1$ spans all functions of $(Y, V)$. There is no loss of efficiency to consider class 3 of restricted estimators with $H = 0$ fixed. By using the fact that $\text{pr}(Y = 1 \mid X^\varepsilon) = g$ and $\text{pr}(Y = 0 \mid X^\varepsilon) = \tilde{g}$, we find

$$
\frac{\partial g}{\partial \theta} = E\left( \frac{(Y - g)^2}{\pi(1, V)} X^\varepsilon \right) = \frac{\tilde{g}}{\pi(1, V)} Y + \frac{\tilde{g}}{\pi(0, V)} (1 - Y).
$$

Therefore, the vector of estimating functions for $\hat{\theta}_{haz}(\phi_1, 0)$ is

$$
\frac{R}{\pi(Y, V)} \phi_1(Y - g)\bigg|_{\pi = \hat{\pi}} = R \left\{ \frac{1 - g}{\hat{\pi}(1, V)} + \frac{\tilde{g}}{\hat{\pi}(0, V)} (1 - Y) \right\} \left(\frac{\partial g/\partial \theta}{\pi(1, V)} + \frac{\tilde{g}}{\pi(0, V)}\right).
$$

The algebraic equivalence follows by direct calculations.
By assumption (c), $(\partial g/\partial \theta)/\{g(1-g)\}$ spans all functions of $V$ including the numerator in the display for $\phi_2$. Therefore, $\phi_2$ is a linear combination of $\phi_1$. By Theorem 2, the optimal $\phi$ is a linear combination of $\phi_1$ and $\phi_2$ and hence can be taken as $\phi_1$.

**Proof of Theorem 4.** Label the sample such that $R_i = 1$ for $i = 1, \ldots, n_1$ and $R_i = 0$ otherwise. Write $w_{i1} = G_1((Z_i))$ $(i = 1, \ldots, n_1)$ and $w_{0i} = G_0((W_i))$ $(i = n_1 + 1, \ldots, n)$. The problem of constrained maximization becomes

$$\max \sum_{i=1}^{n_1} \log w_{i1} + \sum_{i=n_1+1}^n \log w_{0i}$$

subject to $w_{i1} > 0$ $(i = 1, \ldots, n_1)$, $w_{0i} > 0$ $(i = n_1 + 1, \ldots, n)$, and the linear constraints

$$\sum_{i=1}^{n_1} \hat{\pi}(W_i) w_i + \sum_{i=n_1+1}^n \{1 - \hat{\pi}(W_i)\} w_{i1} = 1,$$

(A1)

$$\sum_{i=1}^{n_1} \{1 - \hat{\pi}(W_i)\} \hat{H}^T(W_i) w_{i1} = \sum_{i=n_1+1}^n \{1 - \hat{\pi}(W_i)\} \hat{H}^T(W_i) w_{0i},$$

$$\sum_{i=1}^{n_1} \Phi(X_i^\ast) (Y_i - g(X_i^\ast; \theta)) w_{i1} = 0.$$

The set of $(\lambda, \varrho)$’s such that $\ell_n$ is finite is nonempty and open. On this set, $\ell_n$ is strictly concave due to the strict concavity of the log function and the assumption of linear independence. If $\ell_n$ achieves a maximum at $(\hat{\lambda}, \hat{\varrho})$, then the first-order condition shows that $\hat{w}_{i1} = \hat{G}_1((Z_i))$ $(i = 1, \ldots, n_1)$ and $\hat{w}_{0i} = \hat{G}_0((W_i))$ $(i = n_1 + 1, \ldots, n)$ are positive and satisfy the linear constraints. Furthermore, by Jensen’s inequality, for any positive numbers $w_{i1}$ $(i = 1, \ldots, n_1)$ and $w_{0i}$ $(i = n_1 + 1, \ldots, n)$ satisfying the linear constraints,

$$\frac{1}{n} \sum_{i=1}^{n_1} \log \frac{w_{i1}}{w_{11}} + \frac{1}{n} \sum_{i=n_1+1}^n \log \frac{w_{0i}}{w_{0i}} \leq \log \left( \frac{1}{n} \sum_{i=1}^{n_1} \frac{w_{i1}}{w_{11}} + \frac{1}{n} \sum_{i=n_1+1}^n \frac{w_{0i}}{w_{0i}} \right) = \log(1) = 0.$$  

The simplification to log(1) follows by constraint (A1). The equality holds if and only if $w_{i1} = \hat{w}_{i1}$ $(i = 1, \ldots, n_1)$ and $w_{0i} = w_{0i}$ $(i = n_1 + 1, \ldots, n)$. Therefore, these weights uniquely maximize $L_2$ subject to the desired constraints.

**Proof of Theorem 5.** Write $\hat{\kappa} = \{1 - \hat{\pi}(W)\} \hat{H}^T(W)$ and $\eta(\theta) = \Phi(X^\ast)e(\theta)$. For fixed $\theta$, the gradient of $\ell_n(\lambda, \varrho)$ is

$$\frac{\partial \ell_n}{\partial \lambda} = \tilde{E} \left\{ \frac{R \hat{\kappa}}{\pi + \lambda \hat{\kappa} + \varrho^2 \eta} - \frac{(1 - R) \hat{\kappa}}{1 - \pi - \lambda \hat{\kappa}} \right\}, \quad \frac{\partial \ell_n}{\partial \varrho} = \tilde{E} \left( \frac{R \eta}{\pi + \lambda \hat{\kappa} + \varrho^2 \eta} \right).$$

The gradient of $\ell_n(\hat{\lambda}(\theta), \hat{\varrho}(\theta))$ equals $\partial \ell_n / \partial \theta |_{\lambda=\hat{\lambda}(\theta), \varrho=\hat{\varrho}(\theta)}$, where

$$\frac{\partial \ell_n}{\partial \theta} = \tilde{E} \left\{ \frac{R (\partial \eta^\ast / \partial \theta)}{\pi + \lambda \hat{\kappa} + \varrho^2 \eta} \right\} \varrho.$$

The first-order condition satisfied jointly by $\hat{\lambda} = \hat{\lambda}(\hat{\theta}_{\text{lik}}), \hat{\varrho} = \hat{\varrho}(\hat{\theta}_{\text{lik}})$ and $\hat{\theta}_{\text{lik}}$ is that $\partial \ell_n / \partial \lambda = 0, \partial \ell_n / \partial \varrho = 0$ and $\partial \ell_n / \partial \theta = 0$.

(a) By the assumptions on $\hat{\gamma}$ and $\partial \pi(\cdot; \gamma) / \partial \gamma$, it follows that $\hat{\pi}(\cdot)$ converges uniformly to $\pi(\cdot)$ in probability. By similar arguments as in Newey & Smith (2004), $\hat{\theta}_{\text{lik}}$ converges to $\theta_0$ in probability, and $(\hat{\lambda}, \hat{\varrho}, \hat{\theta}_{\text{lik}})$ jointly has the expansion

$$\left( \begin{array}{c} \hat{\lambda} \\ \hat{\varrho} \\ \hat{\theta}_{\text{lik}} - \theta_0 \end{array} \right) = \left( \begin{array}{cc} K & -L^T \\ -L & 0 \end{array} \right)^{-1} \left( \begin{array}{c} Q_n \\ 0 \end{array} \right) + o_p(n^{-1/2}),$$

(A2)
where \( K \) and \( L \) are the limits of \( K_n \) and \( L_n \) in probability, and

\[
K_n = \begin{bmatrix}
\tilde{E} \left( \frac{R \hat{k} \hat{k}^T}{\hat{\pi}^2} + \frac{(1 - R) \hat{k} \hat{k}^T}{(1 - \hat{\pi})^2} \right) \\
\tilde{E} \left( \frac{R \hat{\eta} \hat{\eta}^T}{\hat{\pi}^2} \right) \\
\tilde{E} \left( \frac{R \eta \eta^T}{\pi^2} \right)
\end{bmatrix}_{\theta = \theta_0}
\]

\[
L_n = \begin{bmatrix}
0 \\
\tilde{E} \left( \frac{R (\partial \eta / \partial \theta)^T}{\hat{\pi}} \right)
\end{bmatrix}_{\theta = \theta_0},
\]

\[
Q_n = \begin{bmatrix}
\tilde{E} \left( \frac{R \hat{k}}{\hat{\pi}^2} - \frac{(1 - R) \hat{k}}{1 - \hat{\pi}} \right) \\
\tilde{E} \left( \frac{R \eta}{\pi^2} \right)
\end{bmatrix}_{\theta = \theta_0}.
\]

Consequently, we obtain

\[
\hat{\theta}_{LK} - \theta_0 = -(L^T K^{-1} L)^{-1} L^T K^{-1} Q_n + o_p(n^{-1/2}) \tag{A3}
\]

\[
= -\mathcal{V}^{-1} D^T A^{-1} \tilde{E} [\tilde{\pi}^{-1} R \eta(\theta_0) - \beta^T (\tilde{\pi}^{-1} R - 1) \tilde{H}^1],
\]

by using twice the inverse formula

\[
K^{-1} = \begin{pmatrix}
K_{11} & K_{12} \\
K_{12} & K_{22}
\end{pmatrix}^{-1} = \begin{pmatrix}
* & * \\
-K_{22}^{-1} K_{12} K_{11}^{-1} & K_{22}^{-1}
\end{pmatrix},
\]

where \( K_{22}^{-1} = K_{22} - K_{12} K_{11}^{-1} K_{12} \). The claimed expansion follows because

\[
\tilde{E} [\tilde{\pi}^{-1} R \eta(\theta_0) - \beta^T (\tilde{\pi}^{-1} R - 1) \tilde{H}^1] = \tilde{E} [\pi^{-1} R \eta(\theta_0) - \beta^T (\pi^{-1} R - 1) H^1] + o_p(n^{-1/2})
\]

as in Tan (2006, Theorem 4).

(b) By direct calculations, \( -n^{-1} \partial \log p_n(\theta)/\partial \theta |_{\theta = \hat{\theta}_{LK}} = L_n^T K_n^{-1} L_n + o_p(1) \). The claim follows from the fact that \( \mathcal{V} = L^T K^{-1} L \) due to expansion of (A3).

(c) By a Taylor expansion of \( \tilde{\pi}^{-1} R \eta(\theta_0) - \beta^T (\tilde{\pi}^{-1} R - 1) \tilde{H}^1 \) about \( (0, 0) \),

\[
\begin{pmatrix}
\hat{\lambda}(\theta_0) \\
\hat{\delta}(\theta_0)
\end{pmatrix} = K^{-1} Q_n + o_p(n^{-1/2}).
\]

Next, a Taylor expansion of \( \log p_n(\theta_0) \) yields

\[
- \log p_n(\theta_0) = n \tilde{E} \{ R \log \tilde{\pi} + (1 - R) \log (1 - \tilde{\pi}) \} + \frac{n}{2} \{ \hat{\lambda}(\theta_0), \hat{\delta}(\theta_0) \} K \begin{pmatrix}
\hat{\lambda}(\theta_0) \\
\hat{\delta}(\theta_0)
\end{pmatrix} + o_p(1).
\]

Similarly, a Taylor expansion of \( \log p_n(\hat{\theta}_{LK}) \) yields

\[
- \log p_n(\hat{\theta}_{LK}) = n \tilde{E} \{ R \log \tilde{\pi} + (1 - R) \log (1 - \tilde{\pi}) \}
\]

\[
+ \frac{n}{2} \{ \hat{\lambda}, \hat{\delta}, \hat{\theta}_{LK} - \theta_0 \} \begin{pmatrix}
K \\
-L^T
\end{pmatrix}
\begin{pmatrix}
\hat{\lambda} \\
\hat{\delta}
\end{pmatrix}
+ o_p(1).
\]

By the three expansions together with (A2),

\[
-2 \{ \log p_n(\theta_0) - \log p_n(\hat{\theta}_{LK}) \} = n Q_n^T K^{-1} L (L^T K^{-1} L)^{-1} L^T K^{-1} Q_n + o_p(1)
\]

\[
= n^{1/2} Q_n^T (L^T K^{-1} L) n^{1/2} Q_n + o_p(1),
\]

where \( Q_n^T = (L^T K^{-1} L)^{-1} L^T K^{-1} Q_n \) converges to a multivariate normal of dimension \( d \) with variance matrix \( \mathcal{V}^{-1} = (L^T K^{-1} L)^{-1} \). Therefore, the claimed convergence holds.
REFERENCES


[Received January 2010. Revised December 2010]