New estimating methods for surrogate outcome data

Bin Nan

University of Michigan

June 21, 2004

Abstract: Surrogate outcome data frequently arise in medical research when true outcomes of interest are expensive or hard to ascertain, but measurements of surrogate outcomes (or more generally speaking, some correlates of the true outcomes) are available. In this paper we assume that the conditional expectation of the true outcome given covariates is known up to a finite dimensional Euclidean parameter. Without specifying the distribution of the data, this conditional mean model is semiparametric. Assuming the true outcome is missing at random, we derive the semiparametric efficient score function for the Euclidean parameter based on the general result of Robins, Rotnitzky and Zhao (1994). In contrast with the general cases in Robins, Rotnitzky and Zhao (1994), the efficient score function has a simple closed form that does not involve any integral equations. We propose two estimating methods, parametric and nonparametric, to estimate the Euclidean parameter using the efficient score function. Under regularity conditions given in the paper, these estimators are consistent and asymptotically Gaussian. Simulation studies show the proposed estimators work well for reasonable sample sizes.

Key words and phrases: Conditional mean model, missing at random, quasi-likelihood, semiparametric efficient score, surrogate outcome, tangent space.

Running title: Surrogate Outcome
1 Introduction

It is often in medical research that outcome variables of interest are difficult or expensive to obtain. Surrogate outcome variables, however, can sometimes be easily ascertained. Many real examples are described in the introductions of Pepe (1992) and Pepe, Reilly and Fleming (1994).

Let $Y$ be the outcome of interest that is not always observable. Let $S$ be a surrogate variable of $Y$ and always available. The association of $Y$ and a covariate vector $X$ is the major interest. Existing methods, such as those proposed by Pepe (1992) and Pepe, Reilly, and Fleming (1994) among others, require either both the conditional densities $f_\theta(Y|X)$ and $f_{\theta,\beta}(S|Y,X)$ or at least $f_\theta(Y|X)$ be known up to the finite dimensional parameter $(\theta, \beta)$ or $\theta$. As pointed out by the authors, any misspecification of the above parametric models can cause biased estimates.

Instead of modelling the conditional density function $f_\theta(Y|X)$ parametrically, we only assume that the conditional expectation of $Y$ given $X$ is known up to a parameter $\theta \in \mathbb{R}^d$, i.e.

$$E(Y|X) = g(X; \theta), \quad (1)$$

where $g(\cdot; \theta)$ is a known function. Let $\epsilon = Y - g(X; \theta)$, then

$$E(\epsilon|X) = 0. \quad (2)$$

Under model (1), the underlying joint density function of $(S,Y,X)$ can be written as

$$q(s,y,x; \theta, f_1, f_2, f_3) = f_1(s|y,x)f_2(y - g(x;\theta)|x)f_3(x), \quad (3)$$

where $f_1$ is the conditional density function of $S$ given $(Y,X)$; $f_2$ is the conditional density function of $Y$, or equivalently $\epsilon$, given $X$; and $f_3$ is the density function of $X$. Thus model (1) is semiparametric in the sense that functions $f_1$, $f_2$, and $f_3$ in (3) are unspecified and
they are infinite dimensional nuisance parameters. When we have complete data, i.e., \( Y \) is observable for all subjects, the surrogate outcome \( S \) does not contribute to the estimation of \( \theta \). Chamberlain (1987) and Robins, Rotnitzky, and Zhao (1994, Proposition 3.1), among others, showed that the asymptotically efficient estimator of \( \theta \) for complete data can be obtained by solving the following estimating equation:

\[
\sum_i \frac{\partial g(X_i; \theta)}{\partial \theta} \frac{\epsilon_i}{E(\epsilon^2 | X_i)} = 0.
\] (4)

This equation has the same form as the quasi-likelihood estimating equation, see e.g. McCullagh (1983). Inevitably, the conditional variance \( \text{var}(Y|X) = E(\epsilon^2|X) \) needs to be specified or estimated in order to calculate the estimator of \( \theta \), and correctly specified or consistently estimated in order to achieve efficiency. Carroll and Ruppert (1982) and Robinson (1987) showed that for a linear model, \( g(X; \theta) = X^T \theta \), substituting \( E(\epsilon^2|X) \) by its kernel smoothing estimator into the above estimating equation yields the efficient estimator for \( \theta \). Newey (1993) extended the smoothing method to the generalized linear models. Modelling \( E(\epsilon^2|X) \) parametrically is a useful alternative to avoid smoothing, but may lose efficiency if the model is incorrect.

In this paper, we are interested in the problem in which the outcome \( Y \) is missing at random (Little and Rubin (2002, Chap. 1)), a surrogate outcome \( S \), however, is available for all subjects. We call this type of data the surrogate outcome data as in Pepe (1992). Without specifying the joint distribution of \( (S,Y,X) \), as we show later in this article, the semiparametric efficient score equation for \( \theta \) in model (1) for the surrogate outcome data actually has the same form as that for the complete data after certain “transformation” of \( Y \). Thus standard estimating methods using quasi-likelihood technique can be adopted with slight modifications to the estimation of parameters for the missing data problem.

In Section 2, we use the general result of Robins, Rotnitzky, and Zhao (1994) to derive the efficient score function for the conditional mean regression model with surrogate outcome
data. We then propose two estimating methods based on the derived efficient score function and show their asymptotic properties in Section 3. We give simulation results in Section 4. Conclusion remarks and discussions are in Section 5. Technical proofs are in the Appendix.

2 The efficient score function

We consider the underlying model (1), where \( \theta \) is the parameter of interest. Let \( R \) be the observing indicator taking value 1 when \( Y \) is observed and 0 otherwise. We assume that \( Y \) is missing at random, i.e., \( pr(R = 1|S,Y,X) = pr(R = 1|S,X) \equiv \pi(S,X) \). We also assume that \( \pi(S,X) > \alpha > 0 \) for some constant \( \alpha \). Denote the observed data as

\[
(S, RY, X, R) \equiv \begin{cases} (S, Y, X) & \text{if } R = 1, \\ (S, X) & \text{if } R = 0. \end{cases}
\]

Then the density function for the observed data \((S, RY, X, R)\) is

\[
p(s, ry, x, r; \theta, f_1, f_2, f_3) = \left\{ \pi(s, x)q(s, y, x; \theta, f_1, f_2, f_3) \right\}^r \\
\left\{ (1 - \pi(s, x)) \int q(s, y, x; \theta, f_1, f_2, f_3) \nu(dy) \right\}^{1-r}, \tag{5}
\]

where \( r \in \{0, 1\} \), \( q \) is the density in (3), and \( \nu \) is a dominating measure.

Let \( \eta \) denote the nuisance parameter, thus \( \eta = (f_1, f_2, f_3) \). Let \( \hat{Q}_\eta \) be the nuisance tangent space, the closed linear span of score operators for one-dimensional parametric submodels for \( \eta \) passing through the true model, for the underlying full data model \((S,Y,X)\). We refer to Bickel, Klaassen, Ritov and Wellner (1993) for definitions and properties of tangent spaces.

The structure of \( \hat{Q}_\eta \) is given in the Appendix. Let \( \hat{Q}_\eta^\perp \) be the orthogonal complement of \( \hat{Q}_\eta \) in \( L^2_0(Q) \), here \( L^2_0(Q) \) is the space of all zero mean and square integrable functions with respect to the distribution function \( Q \) of the underlying full data \((S,Y,X)\). One way to obtain the efficient score for \( \theta \) is to project the score function for \( \theta \) to the orthogonal complement of the nuisance tangent space (see e.g Bickel, Klaassen, Ritov, and Wellner (1993)). Such a
calculation is difficult in general for missing data problems, but can be done through the full data model (see e.g. Robins, Rotnitzky, and Zhao (1994)). Restating Proposition 8.1.e1 and Proposition 8.2.e of Robins, Rotnitzky, and Zhao (1994) for our problem setting, we obtain the following lemma that shows a method of calculating the efficient score function for $\theta$ in the observed data model (5):

**Lemma 1.** Let $l^*_\theta$ be the efficient score for $\theta$ in the model of observed data $(S, RY, X, R)$, and $l^*_0$ the efficient score for $\theta$ in the model of underlying full data $(S, Y, X)$. Then

$$l^*_\theta = \frac{R}{\pi} D(S, Y, X) - \frac{R - \pi}{\pi} E\{D(S, Y, X)|S, X\} ,$$

where $\pi \equiv \pi(S, X)$ and the function $D(S, Y, X) \in \dot{Q}_{\eta}^\perp$ is the unique solution of the equation

$$\Pi \left( \frac{1}{\pi} D - \frac{1 - \pi}{\pi} E(D|S, X) \right) \dot{Q}_{\eta}^\perp = l^*_0 .$$

Here $\Pi$ is a projection operator.

We refer to Nan (2001) for an independent proof of Lemma 1 in a general two-stage design setting. In order to apply Lemma 1 to compute the efficient score $l^*_\theta$, we need three ingredients: (i) the efficient score $l^*_0$; (ii) the characterization of the space $\dot{Q}_{\eta}^\perp$; (iii) the calculation of the projection in equation (7). All of them are from the full data model, and their forms are given by the following two lemmas.

**Lemma 2.** For any $b \in L^0_2(Q)$,

$$\Pi(b|\dot{Q}_{\eta}^\perp) = \frac{E[b(S, Y, X)|X]|\epsilon}{E[\epsilon^2|X]} \epsilon .$$

Thus the efficient score for $\theta$ in the full data model is

$$l^*_0 = \Pi(\dot{l}^0|\dot{Q}_{\eta}^\perp) = \frac{\nabla \theta g(X; \theta)}{E[\epsilon^2|X]} \epsilon ,$$

where $\dot{l}^0$ is the usual score function for $\theta$ in the full data model.
Lemma 3. The orthogonal complement of $\hat{\mathcal{Q}}_\eta$ in $L^2_0(Q)$ is

$$\hat{\mathcal{Q}}_\eta^\perp = \left\{ h(X)\epsilon : E\{h^2(X)\epsilon^2\} < \infty \right\}.$$ (10)

Notice the equivalence between equation (4) and equation (9). Proofs of Lemmas 2 and 3 are scattered in Robins, Rotnitzky, and Zhao (1994), van der Vaart (1998), and in particular, Nan, Emond, and Wellner (2000). To make this article more self-contained, we gather those detailed proofs and list them in the Appendix.

Plugging results in Lemmas 2 and 3 into Lemma 1, we obtain the following Theorem 1 that gives us the efficient score function for $\theta$ in the observed data model (5). The detailed calculation is also in the Appendix.

**Theorem 1.** The efficient score function $l^*_\theta$ for the observed data $(S, RY, X, R)$ is given by

$$l^*_\theta = \frac{\partial g(X; \theta)}{\partial \theta} E(\epsilon^*^2 | X) \epsilon^*,$$ (11)

where

$$\epsilon^* = \frac{R}{\pi} Y - \frac{R - \pi}{\pi} E(Y|S, X) - g(X; \theta).$$ (12)

Let $Y^* = (R/\pi)Y - \{(R - \pi)/\pi\} E(Y|S, X)$ be a kind of “transformation” to the response variable $Y$. Using the nested conditional expectation property, we can easily verify that $E(Y^*|X) = E(Y|X) = g(X; \theta)$. Hence by comparing equation (9) and equation (11) we see that the efficient score $l^*_\theta$ actually has the same form as that of the efficient score for the “full” data $(Y^*, X)$. So analyzing the observed data $(S, RY, X, R)$ with the outcome $Y$ missing at random and the availability of surrogate outcome $S$ is actually similar to analyzing the “full” data $(Y^*, X)$ with the same conditional mean structure as that of $(Y, X)$. The interpretation of the parameter $\theta$ does not change at all, even though the scale of $Y^*$ may not be the same as $Y$. 
3 Two estimating methods

We develop estimators for $\theta$ based on the efficient score function (11) in this section. Since function (11) contains unknown quantities $E(Y|S, X)$ and $E(\epsilon^*|X)$, we need to either model or estimate them. In some medical studies, the surrogate outcome might have been well investigated such that the functional form of $E(Y|S, X)$ could be estimated from previous studies, especially when $S$ satisfies the surrogate criterion of Prentice (1989), i.e. $E(Y|S, X) = E(Y|S)$. Then $Y^*$ could be obtained for each record in a new study, and the observed data could be treated as independent and identically distributed copies of $(Y^*, X)$. Thus the estimation using efficient score (11) would be a standard practice of quasi-likelihood methods.

The more interesting case is that there is no previous study available for estimating $E(Y|S, X)$. We propose two basic estimating methods in this section. As what we will discuss in Section 5, these two methods may be mixed. We assume that the probability function $\pi(s, x)$ is known in this section. This is true for two-stage designs where missing data are caused by designs. We discuss the case when $\pi(s, x)$ is unknown in Section 5, which is related to the so-called double robustness.

3.1 Parametric method

The first proposed method is to model $E(Y|S, X)$ and $E(\epsilon^2|X)$. Suppose that

$$E(Y|S, X) = m(S, X; \beta), \quad (13)$$

and

$$E(\epsilon^2|X) = v(X; \gamma). \quad (14)$$

Here $m(\cdot, \cdot)$ and $v(\cdot)$ are known functions up to finite dimensional parameters $\beta$ and $\gamma$. Then from equation (11) we obtain the following estimating equation for $\theta$ for a cohort of $n$
subjects:
\[
\sum_{i=1}^{n} \frac{\partial g(X_i; \theta)}{v(X_i; \gamma)} \left\{ \frac{R_i}{\pi(S_i, X_i)} \frac{Y_i}{\pi(S_i, X_i)} \right. \\
\left. - \frac{R_i - \pi(S_i, X_i)}{\pi(S_i, X_i)} m(S_i, X_i; \beta) - g(X_i; \theta) \right\} = 0.
\] (15)

As we will show later in Theorem 2, the misspecification of \(m\) and \(v\) only affects the efficiency of the estimator obtained from equation (15). We refer to Carroll and Ruppert (1988) for a variety of choices of and extensive discussions on modelling the functional forms of \(v\).

Parameters \(\beta\) and \(\gamma\) are estimated from following equations based on models (13) and (14):
\[
\sum_{i=1}^{n} R_i A(S_i, X_i) \left\{ Y_i - m(S_i, X_i; \beta) \right\} = 0, \quad (16)
\]
and
\[
\sum_{i=1}^{n} B(X_i) \left\{ \epsilon_i^2 - v(X_i; \gamma) \right\} = 0, \quad (17)
\]
where \(A(S, X)\) and \(B(X)\) are usually chosen as, but not restricted to, the vector of covariates in corresponding models. Notice that equations (15) and (17) use all the data including both completely observed and partially observed data, while equation (16) only uses the completely observed data.

Model (13) looks like the imputation method of Chen (2000). But the estimating equation (15) is not the same as simply plugging the imputed data back to the estimating equation of the full data. It should be noted that the method of Chen (2000) was only developed for the situation where data are missing completely at random (Little and Rubin, 1987, Ch. 1).

We now outline an algorithm for obtaining the estimator \(\hat{\theta}_n\) by solving equation (15):

**Algorithm 1:**

**Step 1:** Estimate \(\beta\) from equation (16);
Step 2: For an initial value of \( \hat{\theta}_n; \hat{\theta}_{(0)} \), estimate \( \gamma \) from equation (17);

Step 3: Calculate the predicted values \( m(S_i; X_i; \hat{\beta}_n) \) and \( v(X_i; \hat{\gamma}_n) \) for all \( i \in \{1, \ldots, n\} \), then plug them into equation (15);

Step 4: Solve equation (15) to obtain \( \hat{\theta}_n \); and use the solution as a new initial value of \( \hat{\theta}_n \);

Step 5: Repeat previous steps from Step 2 until the values of \( \hat{\theta}_n \) converge. Calculate the variance estimator for \( \hat{\theta}_n \) by

\[
\left( \sum_{i=1}^{n} j_{\theta, i}^* \right)^{-1} \left( \sum_{i=1}^{n} j_{\theta, i}^* j_{\theta, i}^{*T} \right)^{-1} \left( \sum_{i=1}^{n} j_{\theta, i}^* \right)^{-1},
\]

where \( \hat{l}_\theta^* = \partial l_\theta^*/\partial \theta \). When both (13) and (14) are correctly specified, the above variance estimator is asymptotically equivalent to \( \left( \sum_{i=1}^{n} l_{\theta, i}^* l_{\theta, i}^{*T} \right)^{-1} \), which is usually called the model-based variance estimator.

In both Algorithm 1 and the following Algorithm 2 in the next subsection, the initial value of \( \hat{\theta}_{(0)} \) can be obtained using the following Horvitz-Thompson estimating equation for \( \theta \):

\[
\sum_{i=1}^{n} \frac{R_i}{\pi(S_i; X_i)} \frac{\partial g(X_i; \theta)}{\partial \theta} \left\{ Y_i - g(X_i; \theta) \right\} = 0. \tag{18}
\]

If the variance function in equation (14) is modelled without parameter \( \gamma \), e.g. a function of \( g \) as that in generalized linear models, then Step 2, the second part of Step 4, and the first part of Step 5 in the above algorithm should be omitted.

### 3.2 Nonparametric method

The second proposed method is to treat \( E(Y|S, X) \) and \( E(\epsilon^2|X) \) as infinite dimensional nuisance parameters and use a nonparametric method, such as kernel or spline smoothing technique, to estimate these two unknown functions. Let \( \eta = (\eta_1, \eta_2) \), where \( \eta_1 = E(Y|S, X) \) and \( \eta_2 = E(\epsilon^2|X) \). We rewrite \( l_\theta^* \) as \( l_{\theta, \eta}^* \). Thus,

\[
l_{\theta, \eta}^* = \frac{\partial g(X; \theta)}{\eta_2} \left\{ \frac{R}{\pi} Y - \frac{R}{\pi} \eta_1 - g(X; \theta) \right\}, \tag{19}
\]
and $\theta$ is estimated by solving the following estimating equation:

$$\sum_{i=1}^{n} l_{\hat{\theta}, \hat{\eta}}(\theta) = 0.$$ (20)

Following is the algorithm of calculating the estimator $\hat{\theta}_n$ by solving equation (20):

**Algorithm 2:**

**Step 1:** Estimate $\eta_1 = E(Y|S,X)$ via smoothing using all fully observed records. Note that the observing probabilities for those records vary. So the $i$-th fully observed record should have weight $1/\pi(S_i, X_i)$. Then calculate $\hat{\eta}_1$ for all records, including those with missing data.

**Step 2:** Choose an initial estimator of $\theta$, $\hat{\theta}^{(0)}$. This can be done similarly as in Algorithm 1.

**Step 3:** Calculate $Y_i^*$ and thus the residuals $\epsilon_i^* = Y_i^* - g(X_i; \hat{\theta}^{(0)})$, $i = 1, \ldots, n$. Then estimate $\eta_2 = E(\epsilon^2|X)$ using a smoothing method.

**Step 4:** Plug $\hat{\eta}_{1,n}$ and $\hat{\eta}_{2,n}(\hat{\theta}^{(0)})$ into equation (20) and solve the equation for $\hat{\theta}_n$.

**Step 5:** Use the root of equation (20) in Step 4 as a new initial value of $\hat{\theta}_n$ and repeat previous steps from Step 2 until $\hat{\theta}_n$ converges. The variance estimator for $\hat{\theta}_n$ is

$$\left(\sum_{i=1}^{n} l_{\hat{\theta}, \hat{\eta}}(\theta_i)\right)^{-1} \left(\sum_{i=1}^{n} l_{\hat{\theta}, \hat{\eta}}(\theta_i)^T \left(\sum_{i=1}^{n} l_{\hat{\theta}, \hat{\eta}}(\theta_i)^{-1}\right)^{-1},$$

where $l_{\hat{\theta}, \hat{\eta}} = \partial l_{\hat{\theta}, \hat{\eta}}/\partial \theta$. It is asymptotically equivalent to $\left(\sum_{i=1}^{n} l_{\hat{\theta}, \hat{\eta}}(\theta_i)^T \left(\sum_{i=1}^{n} l_{\hat{\theta}, \hat{\eta}}(\theta_i)\right)^{-1}$ if the smoothing models for $\eta_1 = E(Y|S,X)$ and $\eta_2 = E(\epsilon^2|X)$ are correct.

When $X$ is discrete, $E(\epsilon^2|X) = var(Y^*|X)$ can be estimated from grouped data grouping on distinct values of $X$ without using residuals, and thus the above algorithm does not need iteration. When both $S$ and $X$ are discrete, the two algorithms are unified.

### 3.3 Asymptotic properties

In this subsection, we give a set of regularity conditions that guarantees the desirable asymptotic properties of the proposed estimators. These conditions are reasonable for many prac-
tatical problems.

(C1) The surrogate outcome $S$ and covariates $X$ have finite support.

(C2) The parameter space of $\theta$, $\Theta$, is bounded.

(C3) $\sup_{S,X,\beta} |m(S,X;\beta)| < M_1 < \infty$; $\sup_{S,X} |\eta_1(S,X)| < M_1 < \infty$.

(C4) $0 < \sigma < v(X;\gamma) < M_2 < \infty$ for all $X$ and $\gamma$; $0 < \sigma < \eta_2(X) < M_2 < \infty$ for all $X$.

(C5) The function $g(X;\theta)$ is twice differentiable to $\theta$ with continuous second derivative for all $X$; and $E\{\dot{g}(X;\theta_0)\ddot{g}(X;\theta_0)^T\}$ is nonsingular. Here $\dot{g} = \partial g / \partial \theta$, $\ddot{g} = \partial \dot{g} / \partial \theta^T$, and $\theta_0$ is the true parameter which is an interior point in $\Theta$.

(C6) The true parameter $\theta_0$ is the unique root of $El^*_\theta = 0$.

(C7) The class of functions $l^*_{\theta_0,\eta}$ indexed by $\eta$ is asymptotically equicontinuous.

Note that Condition (C5) holds for all the mean functions of generalized linear models discussed by McCullagh and Nelder (1989). Condition (C6) is a usual assumption in the estimating equation theory for problems without missing data (Huber (2004), page 131). Condition (C7) is easy to check for parametric method, but not so easy for nonparametric method unless we restrict more on $\eta$ such as monotonicity. Simplifying (or verifying) Condition (C7) for specific models is an important future work.

**Theorem 2.** Suppose regularity conditions (C1)-(C7) hold. Then the estimator of $\theta$ in model (1) with distribution (5) obtained by either one of the two methods given in Subsections 3.1 and 3.2 is consistent and asymptotically Gaussian. In particular, if either $m$ and $v$ are correctly specified or $\eta_1$ and $\eta_2$ are consistently estimated, then the corresponding estimator achieves the information bound determined by (11).

The proof of Theorem 2 requires the empirical processes theory and is deferred to the Appendix. Note that we do not require consistent estimators for any nuisance parameters to
obtain the asymptotic normality for $\hat{\theta}_n$ in either method; but we do in order to achieve the information bound, the nuisance parameter estimators, however, can converge to the truth with arbitrary rates.

4 Numerical examples

We conduct simulations to investigate the finite sample performance of proposed estimators, and compare with inverse probability weighted estimators. We first look at a simple discrete case, where the information bound calculation can be done easily. Consider the setting of a binary outcome $Y$ and a binary surrogate outcome $S$. For example, $Y$ is the true disease status and $S$ is the result of a screening test. We are interested in estimating the association between $Y$ and a binary covariate $X$. Thus $S, Y, X \in \{0, 1\}$. Notice that the “transformed” response $Y^*$ is no longer binary when $Y$ is observed only for a proportion of subjects. Suppose $pr(S = 1|Y, X) = pr(S = 1|Y)$. Let $X \sim \text{Bernoulli}(p)$, $Y|X \sim \text{Bernoulli}(g(X; \theta))$, and $S|Y, X \sim S|Y \sim \text{Bernoulli}(q(Y))$. Here $p$ is a given constant, $q(Y)$ is a given function, and $g(X; \theta)$ is a logit function as follows:

$$g(X; \theta) = \frac{\exp(\theta_1 + \theta_2 X)}{1 + \exp(\theta_1 + \theta_2 X)}.$$

Suppose we observe $S$ and $X$ for all subjects in the study, and observe $Y$ for a subsample with the selection probability $pr(R = 1|S, X) = \pi(S, X)$. So $Y$ is missing at random. We further assume that $p = 0.25$, $\theta_1 = 0$, $q(0) = 0.3$, and $q(1) = 0.9$. Using terms in epidemiology, $1 - q(0) = 1 - pr(S = 1|Y = 0) = pr(S = 0|Y = 0)$ is called specificity, and $q(1) = pr(S = 1|Y = 1)$ is called sensitivity. Usually $\theta_1$ is of less interest. We only show the simulation results for estimating $\theta_2$. We choose two different values for $\theta_2$: 0 and $\log(2) \approx 0.693$. We also choose two sets of selection probabilities for each value of $\theta$: $\pi(S, X) = 0.25$ for all $S$ and $X$ and $(\pi(0, 0), \pi(0, 1), \pi(1, 0), \pi(1, 1)) = (0.208, 0.625, 0.139, 0.002)$. 


0.417) when $\theta = 0$, and $\pi(S, X) = 0.25$ for all $S$ and $X$ and $(\pi(0, 0), \pi(0, 1), \pi(1, 0), \pi(1, 1)) = (0.208, 0.833, 0.139, 0.357)$ when $\theta = \log(2)$, corresponding to non-stratified and stratified sampling designs. In both cases, we expect to observe $Y$ for a quarter of the subjects. The stratified sampling probabilities give us the same expected number of selected subjects across all four cells of $(S, X) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$.

Simulations are conducted using 1000 replications with cohort sizes $n = 400$ and 1000, and thus the numbers of expected fully observed subjects are 100 and 250, respectively. We estimate $\theta_2$ using weighted quasi-likelihood methods to obtain three types of estimators: the inverse probability weighted estimator using the completely observed data and the true selection probabilities; the inverse probability weighted estimator using the completely observed data and the estimated selection probabilities; and the proposed estimator using all the observed data. The variance structure for completely observed data analysis is “$\mu(1 - \mu)$”. For the proposed estimating procedure, both the conditional expectation $E(Y|S, X)$ and the variance $\text{var}(Y^*|X) = E(\epsilon^2|X)$ are estimated nonparametrically using grouped data. Hence the nuisance parameter estimators are consistent, which yield the semiparametric efficient estimator for $\theta_2$. The results are listed in Table 1.

[Tables 1 is about here.]

Chen and Breslow (2004) (where they obtained the same efficient score function (11) using the optimal estimating equation theory) pointed out an interesting special case that when $S$ is discrete and $X$ is binary, the inverse probability weighted estimator (a Horvitz-Thompson estimator) using empirically estimated $\pi$ is asymptotically efficient. This is verified from the simulation results listed in Table 1. They only studied the case when both $S$ and $X$ are discrete.
Table 1 shows that biases are close to zero in all simulations. The efficient estimator performs equally well with the inverse probability weighted methods using empirically estimated selection probabilities for the reason mentioned above, but notice that this does not hold in general. Both of these methods perform better than the inverse probability weighted methods using true selection probabilities. Variance estimators of efficient estimating methods are valid, and close to the asymptotic optimal variance calculated from the efficient score. Simulations also show that the stratification improves efficiency dramatically.

We then conduct simulations with continuous $X$ and $S$ to investigate model misspecification and the validity of handling continuous variables $X$ and $S$ via both Algorithms 1 and 2. Suppose the underlying true model is

$$E(Y|S, X) = \theta_0 + \theta_1 X + \theta_2 f(S),$$

(21)

and the model of interest is

$$E(Y|X) = \theta_0 + \theta_1 X,$$

(22)

here we choose $f(S) = S^{1/3} \sim N(0, 1)$. Let $X \sim N(0, 1)$, $\epsilon_0 = Y - E(Y|S, X) \sim N(0, 1)$, and $\theta_0 = \theta_1 = \theta_2 = 1$. Simulations are conducted using 1000 replications with cohort size $n = 200$ and $1000$, respectively, and the selection probability $\pi(S, X) = 0.5$. Here $S$ does not satisfy the definition of surrogate outcome given by Prentice (1989). Since it is correlated with the true outcome $Y$, however, we can use it in the same way as the surrogate outcome to improve efficiency. Four estimating methods are simulated: (i) fitting linear regression model (22) using fully observed data only (complete-case method); (ii) using Algorithm 1 with misspecified model $E(Y|S, X) = \theta_0 + \theta_1 X + \theta_2 S$; (iii) using Algorithm 1 with correctly specified model (21); and (iv) applying Algorithm 2 using a generalized additive model to estimate $E(Y|S, X)$ via smoothing splines on $S$ and $X$. We use smoother $s()$ in Splus with default values of smoothing parameters. We do not need to model variances in Equations
(15) and (16) for the above simulations since they are actually constants. The simulation results are listed in Table 2. When sample size is small \( n = 200 \), Algorithm 2 does not work very well. When we increase the sample size to \( n = 1000 \), which means that about 500 records are used to estimate \( E(Y|S, X) \) using a generalized additive model, Algorithm 2 works equally well as Algorithm 1 with the correct model for \( E(Y|S, X) \).

[Table 2 is about here.]

## 5 Discussion

The assumption of fixed selection probabilities is not necessary. Methods proposed in the paper still work if we can estimate \( \pi(S, X) \) consistently, see e.g. Robins, Rotnitzky, and Zhao (1994). The parametric method has the so-called double robustness property: if either \( \pi(S, X) \) or \( E(Y|S, X) \) is correctly specified, the estimating equation (15) is unbiased, which still yields a valid estimator for \( \theta \). In two-stage sampling designs, \( \pi(S, X) \) is determined by investigators. Thus obtaining the optimal \( \pi(S, X) \) is an interesting problem and remains to be explored.

Theorem 2 tells us that if \( m \) and \( v \) are correctly specified, then the estimating equation (15) yields an estimator that achieves the information bound determined by (11). Sometimes people still call it the semiparametric efficient estimator. The estimator, however, can be improved if one is willing to pay the price for the risk of misspecification of \( m \) and \( v \), because more assumptions are added to the original model. There is a nice discussion in Newey (1993) for the moment regression problem without missing data assuming the first two moments have known functional forms.

The two algorithms proposed in this article can be mixed, e.g., \( E(Y|S, X) \) can be estimated parametrically, while \( E(\epsilon^2|X) \) is estimated nonparametrically. Algorithm 1 can easily
handle continuous and high dimensional covariates. But misspecification of $E(Y|S, X)$ and $E(e^2|X)$ will cause efficiency loss. Algorithm 2 is trying to avoid the misspecification problem without compromising efficiency. But it usually needs large sample size.

Estimating equation (15) reduces to different forms when we choose different ways of estimating $E(Y|S, X)$. If we replace $E(Y|S, X)$ by true observations of $Y$, then equation (15) becomes the estimating equation for full data. If $S$ does not contribute, i.e., $E(Y|S, X) = E(Y|X)$, equation (15) becomes the inverse-probability weighted Horvitz-Thompson estimating equation (18). If we replace $E(Y|S, X)$ by $Y$ when $R = 1$, and by imputed value of $Y$ when $R = 0$, then equation (15) becomes the imputation method estimating equation, similar to the method discussed by Chen (2000). Intuitively, $E(Y|S, X)$ is a better prediction of $Y$ than $E(Y|X)$ even when $S$ is treated as an extra covariate in the same model as $E(Y|X)$. Hence in general we expect that the estimating equation (15) yields an more precise estimator than the inverse-probability weighted Horvitz-Thompson estimating method does, no matter whether the functional form of $E(Y|S, X)$ is correctly specified.

We use one dimensional $S$ in the paper. In fact, $S$ can be multiple dimensional and the same conclusion holds for the two estimating methods.

**Acknowledgements**

The author owes thanks to Norman Breslow, Margaret Pepe, and Jon Wellner for helpful discussions in this research.
Appendix: Proofs of Theoretical Results

Proof of Lemma 2:

In model (3), for any one-parameter family of conditional densities \( \{f_{1,\lambda}(s|y,x)\} \) with \( f_{1,0} = f_1 \), define
\[
a_1(s, y, x) = \frac{\partial}{\partial \lambda} \log f_{1,\lambda}(s|y,x) \bigg|_{\lambda=0}.
\]
Then the score operator for the nuisance parameter \( f_1 \) is \( \dot{l}_{f_1} a_1 = a_1(s, y, x) \). Similarly, we can define score operators for nuisance parameters \( f_2 \) and \( f_3 \) as \( a_2(y, x) \) and \( a_3(x) \), respectively.

Direct calculations show that the three components of the tangent space of model (3), \( \dot{Q}_1 \), \( \dot{Q}_2 \), and \( \dot{Q}_3 \) corresponding to \( f_1 \), \( f_2 \), and \( f_3 \), respectively, have the following structures:
\[
\dot{Q}_1 = [a_1(S, Y, X) : E[a_1|Y, X] = 0, Ea_1^2 < \infty],
\]
\[
\dot{Q}_2 = [a_2(Y, X) : E[a_2|X] = 0, E[\epsilon a_2|X] = 0, Ea_2^2 < \infty],
\]
\[
\dot{Q}_3 = [a_3(X) : Ea_3 = 0, Ea_3^2 < \infty].
\]

Here \([\cdot]\) denotes the closed linear span. It can be verified easily that these three spaces are mutually orthogonal. Thus the nuisance tangent space becomes: \( \dot{Q}_\eta = \dot{Q}_1 + \dot{Q}_2 + \dot{Q}_3 \), according to Bickel, Klaassen, Ritov, and Wellner (1993). The second restriction in equality (24) comes from the assumption that \( E[\epsilon|X] = 0 \). That \( \dot{Q}_2 \) contains the right side in (24) is difficult to prove. But the equality assumption works for our purpose. See the discussion in Bickel, Klaassen, Ritov, and Wellner (1993), pages 76-77.

For any \( b \in L_2^0(Q) \), let
\[
r_b = \frac{E[\epsilon b(S, Y, X)|X]}{E[\epsilon^2|X]} \epsilon.
\]

To prove (8), we will show that \( r_b \in \dot{Q}_\eta^\perp = (\dot{Q}_1 + \dot{Q}_2 + \dot{Q}_3)^\perp \) and that \( b - r_b \in (\dot{Q}_1 + \dot{Q}_2 + \dot{Q}_3) \).

For any \( a_1 \in \dot{Q}_1 \):
\[
\langle r_b, a_1 \rangle_{L_2^0(Q)} = E(r_b a_1) = E \left\{ \frac{E(\epsilon b|X)E(\epsilon a_1|X)}{E[\epsilon^2|X]} \right\}.
\]
\[
E \left\{ \frac{E(eb|X)E[E(a_1|Y,X)|X]}{E[e^2|X]} \right\} = E \left\{ \frac{E(eb|X)E[\epsilon E(a_1|Y,X)|X]}{E[e^2|X]} \right\} = 0
\]

by (23). For any \( a_2 \in \dot{Q}_2 \):

\[
\langle r_b, a_2 \rangle_{L_2^0(Q)} = E(r_b a_2) = E \left\{ \frac{E(eb|X)E(\epsilon a_2|X)}{E[e^2|X]} \right\} = 0
\]

by (24). And, for any \( a_3 \in \dot{Q}_3 \):

\[
\langle r_b, a_3 \rangle_{L_2^0(Q)} = E(r_b a_3) = E \left\{ \frac{E(eb|X)a_3(X)E(\epsilon|X)}{E[e^2|X]} \right\} = 0
\]

since \( E(\epsilon|X) = 0 \). Hence \( r_b \in \dot{Q}_\eta^+ \).

Rewrite \( b - r_b \) as \( b - r_b = b - E[b|X] - r_b + E[b|X] \). Since \( E\{b - E[b|X] - r_b|X\} = 0 \) and \( E\{(b - E[b|X] - r_b)\epsilon|X\} = 0 \), we know that \( b - E[b|X] - r_b \in \dot{Q}_2 \). The other part has zero mean since \( b \in L_2^0(Q) \), so \( E[b|X] \in \dot{Q}_3 \). Thus \( b - r_b \in (\dot{Q}_2 + \dot{Q}_3) \subset \dot{Q}_\eta \), which shows the desired result.

The score function of \( \theta \) in model (3) is \( \dot{l}_\theta^0 = -(f'_2/f_2)(\epsilon|X)\partial g(x;\theta)/\partial \theta \). Thus the efficient score \( l_\theta^0 \) in (9) can be obtained via direct calculation from (8) using the definition \( l_\theta^0 = \Pi(\dot{l}_\theta^0|\dot{Q}_\eta^+) \) (see e.g. Bickel, Klaassen, Ritov, and Wellner (1993)) and the fact that \( E[-\epsilon(f'_2/f_2)(\epsilon|X)|X] = 1 \).

**Proof of Lemma 3:**

Take \( a_1 \in \dot{Q}_1 \), \( a_2 \in \dot{Q}_2 \), and \( a_3 \in \dot{Q}_3 \). Then we have \( E[a_1 h(X)\epsilon|X] = 0 \), \( E[a_2 h(X)\epsilon|X] = 0 \), and \( E[a_3 h(X)\epsilon|X] = 0 \), as in the proof of Lemma 2, which shows \( \{h(X)\epsilon : E[\epsilon^2 h^2(X)] < \infty\} \subset \dot{Q}_\eta^+ \). Equation (8) shows the reverse inclusion, since

\[
E \left\{ \frac{E^2(eb|X)}{E^2(e^2|X)} e^2 \right\} \leq Eb^2 < \infty
\]
by the Cauchy inequality.

**Proof of Theorem 1:**

Let $D(Y, X) = h(X)\epsilon$ and plug it into equation (7). Then from equations (8), (9) and (10) we obtain

$$\frac{\partial g(X; \theta)}{\partial \theta} = \frac{1}{E(\epsilon^2 | X)} E \left[ \frac{1}{\pi} h(X)\epsilon^2 - \epsilon \frac{1 - \pi}{\pi} E\{h(X)\epsilon | S, X | X\} \right] \epsilon$$

$$= \frac{1}{E(\epsilon^2 | X)} E \left\{ \frac{1}{\pi} \epsilon^2 - \frac{1 - \pi}{\pi} E^2(\epsilon | S, X | X) \right\} h(X)\epsilon .$$

Simplifying the above equality yields

$$h(X) = \frac{\partial g(X; \theta)}{\partial \theta} \frac{E \left\{ \frac{1}{\pi} \epsilon^2 - \frac{1 - \pi}{\pi} E^2(\epsilon | S, X | X) \right\}}{E(\epsilon^2 | X)} .$$

Hence from equation (6) we obtain the efficient score $l_\theta^*$ for the observed data in the conditional mean model (1), which is give by

$$l_\theta^* = \frac{R}{\pi} h(X)\epsilon - \frac{R - \pi}{\pi} E\{h(X)\epsilon | S, X\}$$

$$= h(X) \left\{ \frac{R}{\pi} \epsilon - \frac{R - \pi}{\pi} E(\epsilon | S, X) \right\}$$

$$= h(X) \left\{ \frac{R}{\pi} Y - \frac{R - \pi}{\pi} E(Y | S, X) - g(X; \theta) \right\}$$

$$= \frac{\partial g(X; \theta)}{\partial \theta} \frac{E(\epsilon^2 | X)}{E(\epsilon^2 | X)} \epsilon^*$$

where

$$\epsilon^* = \frac{R}{\pi} Y - \frac{R - \pi}{\pi} E(Y | S, X) - g(X; \theta) ,$$

and it is easy to show that $E(\epsilon^2 | X) = E \left\{ \frac{1}{\pi} \epsilon^2 - \frac{1 - \pi}{\pi} E^2(\epsilon | S, X | X) \right\}$.
Proof of Theorem 2:

The proof does not need to distinguish parametric or nonparametric models for $E(Y|S, X)$ and $E(\epsilon^2|X)$. Hence we denote $m$ by $\eta_1$ and $v$ by $\eta_2$, and thus use the estimating function (19) in the proof. We also assume $d = 1$ for simplicity. The proof for multi-dimensional parameter $\theta$ is almost identical.

We will ignore measurability difficulties related to empirical processes and refer to van der Vaart and Wellner (1996, Part 1) for a complete treatment of the necessary weak convergence theory using outer integrals and outer probabilities. We will drop the word “outer” in this proof.

The proof of consistency follows the line of Hu (1998). We restate her Theorem 3.1.1 for our problem setting in the following Lemma 4 and also give her proof here for completeness. We define

$$\psi(\theta, \eta) = E_0 l^*_{\theta, \eta} = \int l^*_\theta dP_0, \quad \psi_n(\theta, \eta) = \frac{1}{n} \sum_{i=1}^n l^*_{\theta, \eta}(S_i, R_i Y_i, X_i, R_i),$$

here $P_0$ is the true distribution of observed data $(S, RY, X, R)$, and $l^*_{\theta, \eta}$ is the estimating function defined in (19) which is the same as that in the parametric method when $m$ and $v$ are replaced by $\eta_1$ and $\eta_2$. In the integral of the first equality above, parameters $\theta$ and $\eta$ are fixed.

**Lemma 4.** Suppose that $\theta_0$ is the unique solution to $\psi(\theta, \eta_0) = 0$ and $\|\hat{\eta}_n - \eta_0\| = o_p(1)$. If

$$\sup_{\theta \in \Theta, \|\eta - \eta_0\| < \delta_n} \frac{|\psi_n(\theta, \eta) - \psi(\theta, \eta_0)|}{1 + |\psi_n(\theta, \eta)| + |\psi(\theta, \eta_0)|} = o_p(1) \quad (26)$$

for every sequence $\{\delta_n\} \downarrow 0$, then $\hat{\theta}_n$ satisfying $\psi_n(\hat{\theta}_n, \hat{\eta}_n) = o_p(1)$ converges to $\theta_0$ in probability.

**Proof.** Since $\theta_0$ is the unique solution to $\psi(\theta, \eta_0) = 0$, this implies that for any fixed $\varepsilon > 0$,
there exists a \( \delta > 0 \) such that
\[
pr( |\hat{\theta}_n - \theta_0| > \varepsilon ) \leq pr( |\psi(\hat{\theta}_n, \eta_0)| > \delta ),
\]
and consistency of \( \hat{\theta}_n \) follows immediately if we can show \( |\psi(\hat{\theta}_n, \eta_0)| \to 0 \) in probability.

Taking \( \eta = \hat{\eta}_n \) in equation (26), we have
\[
|\psi(\hat{\theta}_n, \eta_0)| \leq |\psi_n(\hat{\theta}_n, \hat{\eta}_n)| + o_p(1) + |\psi_n(\hat{\theta}_n, \hat{\eta}_n)| \to 0 \text{ by equation (26)},
\]
which implies \( |\psi(\hat{\theta}_n, \eta_0)| = o_p(1) \).

We now verify those conditions in Lemma 4 for our problem. Suppose \( ||\hat{\eta}_{1,n} - \eta_{1,0}|| \to 0 \) and \( ||\hat{\eta}_{2,n} - \eta_{2,0}|| \to 0 \) in probability for some functions \( \eta_{1,0} \) and \( \eta_{2,0} \) satisfying Conditions (C3) and (C4), but not necessary to be \( E_0(Y|S,X) \) and \( E_0(\varepsilon^2|X) \). Here \( || \cdot || \) denotes the supremum norm. In other words, the estimators \( \hat{\eta}_{1,n} \) and \( \hat{\eta}_{2,n} \) do not need to be consistent.

Let \( \eta_0 = (\eta_{1,0}, \eta_{2,0}) \). Condition (C6) implies that \( \theta_0 \) is the unique solution to \( \psi(\theta, \eta_0) = 0 \). For both estimating methods proposed in this paper, we have \( \psi_n(\hat{\theta}_n, \hat{\eta}_n) = 0 \). Now we only need to verify equation (26).

For any \( \delta > 0 \), we consider all \( \eta \) in \( ||\eta - \eta_0|| < \delta \). Simple calculation shows that
\[
\psi(\theta, \eta) = E_0 \left\{ \frac{\hat{\xi}(X; \theta)}{\eta_2(X)} \left( g(X; \theta_0) - g(X; \theta) \right) \right\},
\]
and the law of large numbers yields \( \psi_n(\theta, \eta) = \psi(\theta, \eta) + o_p(1) \). Thus \( |\psi_n(\theta, \eta) | - |\psi(\theta, \eta_0)| = |\psi(\theta, \eta) - \psi(\theta, \eta_0)| + o_p(1) \), and
\[
|\psi(\theta, \eta) - \psi(\theta, \eta_0)| \leq E_0 \left\{ \left| \frac{\hat{\xi}(X; \theta)}{\eta_2(X)} \left( g(X; \theta_0) - g(X; \theta) \right) \left( \frac{1}{\eta_2} - \frac{1}{\eta_{2,0}} \right) \right| \right\}.
\]
From Conditions (C1), (C2), and (C5) we know that both \( g \) and \( \hat{\xi} \) are bounded for all \( \theta \) and \( X \); from Condition (C4) we have \( ||1/\eta_2 - 1/\eta_{2,0}|| \leq ||\eta_2 - \eta_{2,0}||/\sigma^2 \leq \delta/\sigma^2 \). Then we have
Thus we have

\[ |\psi(\theta, \eta) - \psi(\theta, \eta_0)| \leq M\delta, \]

where \( M \) is a finite constant. No matter how small \( \delta \) is, there is always an integer \( N \) such that \( \delta_n < \delta \) when \( n > N \). Let \( \delta \) be arbitrarily small, we thus have verified equation (26).

The rest of the proof is to show the asymptotic normality. Let \( \dot{\psi}_n(\theta, \eta) = \partial \psi_n(\theta, \eta) / \partial \theta \) and \( \dot{l}_{\theta, \eta}^* = \partial \dot{l}_{\theta, \eta}^* / \partial \theta \). Taylor expansion yields

\[
0 = \psi_n(\hat{\theta}_n, \hat{\eta}_n) = \psi_n(\theta_0, \hat{\eta}_n) + \psi_n(\hat{\theta}, \hat{\eta}_n)(\hat{\theta}_n - \theta_0),
\]

where \( \hat{\theta} \) is a point between \( \theta_0 \) and \( \hat{\theta}_n \). Rearrange the above equality,

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = \left\{ - \dot{\psi}_n(\hat{\theta}, \hat{\eta}) \right\}^{-1} \sqrt{n}\psi_n(\theta_0, \hat{\eta}_n). \tag{27}
\]

Since

\[
\dot{l}_{\theta, \eta}^* = \frac{\dot{g}(X; \theta)}{\eta_2(X)} \left\{ \frac{R}{\pi(S, X)} Y - \frac{R - \pi(S, X)}{\pi(S, X)} \eta_1(X, S) - g(X; \theta) \right\} - \frac{\dot{g}(X; \theta)^2}{\eta_2(X)},
\]

and \( \dot{g} \) is bounded from Conditions (C1), (C2), and (C5), we have

\[
|\dot{l}_{\theta, \eta}^* - \dot{l}_{\theta_0, \eta_0}^*| \leq \left| \frac{R}{\pi} \right| \cdot \left\| \dot{g}(X; \theta) \right\| \left( \hat{\eta}_2 \right) - \frac{\dot{g}(X; \theta_0)}{\eta_2,0} \right\| + \left| \frac{R - \pi}{\pi} \right| \cdot \left\| \dot{g}(X; \hat{\theta}) \hat{\eta}_1 \right\| \left( \hat{\eta}_2 \right) - \frac{\dot{g}(X; \theta_0) \eta_1,0}{\eta_2,0} \right\| + \left\| \dot{g}(X; \hat{\theta}) \right\| \left( \hat{\eta}_2 \right) - \frac{\dot{g}(X; \theta_0)^2}{\eta_2,0} \right\|.
\]

Since \( \hat{\theta} \to \theta_0 \) and \( \hat{\eta} \to \eta_0 \) in probability, \( g, \dot{g}, \) and \( \ddot{g} \) are continuous, and all the functions inside above norms are bounded from both below and above, simple algebra yields that

\[
|\dot{\psi}_n(\hat{\theta}, \hat{\eta}_n) - \dot{\psi}_n(\theta_0, \eta_0)| \leq \frac{1}{n} \sum_{i=1}^{n} \left| \frac{R_i}{\pi_i} Y_i \right| \cdot o_p(1) + \frac{1}{n} \sum_{i=1}^{n} \left| \frac{R_i - \pi_i}{\pi_i} \right| \cdot o_p(1) + o_p(1) + o_p(1) = o_p(1).
\]

Thus we have

\[
\dot{\psi}_n(\hat{\theta}, \hat{\eta}_n) = \psi_n(\theta_0, \eta_0) + o_p(1) \to E_0 \dot{l}_{\theta_0, \eta_0}^* = - \frac{\dot{g}(X; \theta_0)^2}{\eta_2,0}, \tag{28}
\]

in probability.
Similarly we have $\|l_{\theta_0, \hat{\eta}_n}^* - l_{\theta_0, \eta_0}^*\| = o_p(1)$. Since $\psi(\theta_0, \eta) = 0$ for all $\eta$, by Condition (C7) we obtain

$$\sqrt{n}\psi_n(\theta_0, \hat{\eta}_n) = \sqrt{n}\{\psi_n(\theta_0, \hat{\eta}_n) - \psi(\theta_0, \hat{\eta}_n)\}$$

$$= \sqrt{n}\{\psi_n(\theta_0, \eta_0) - \psi(\theta_0, \eta_0)\} + o_p(1)$$

$$\rightarrow N(0, E_0 l_{\theta_0, \eta_0}^{*2}) \quad (29)$$

in distribution. Combining (27), (28), and (29) we obtain

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow N(0, \Sigma)$$

in distribution, where

$$\Sigma = \{E_0 l_{\theta_0, \eta_0}^{*}\}^{-1}\{E_0 l_{\theta_0, \eta_0}^{*2}\}\{E_0 l_{\theta_0, \eta_0}^{*}\}^{-1}.$$

When $\eta_{0,1} = E_0(Y|S, X)$ and $\eta_{0,2} = E_0(\epsilon^2|X)$, we have $E_0 l_{\theta_0, \eta_0}^{*2} = -E_0 l_{\theta_0, \eta_0}^*$ and thus $\hat{\theta}_n$ achieve the information bound determined by (11).

References


Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, U.S.A.

E-mail: bnan@umich.edu

Phone: (734) 763-5538     Fax: (734) 763-2215
Table 1a. Simulation summary statistics for estimating $\theta_2$ in logistic models with 1000 replications.

<table>
<thead>
<tr>
<th>Methods</th>
<th>$n$</th>
<th>mean($\hat{\theta}_{2,n}$)</th>
<th>$s^2(\hat{\theta}_{2,n})$</th>
<th>mean$^a$ var($\hat{\theta}_{2,n}$)</th>
<th>Optimal$^b$ var($\hat{\theta}_{2,n}$)</th>
<th>95%CP$^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) $\theta_2 = 0$, $(\pi(0, 0), \pi(0, 1), \pi(1, 0), \pi(1, 1)) = (0.25, 0.25, 0.25, 0.25)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPW1$^d$</td>
<td>400</td>
<td>0.0065</td>
<td>0.2195</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IPW2$^e$</td>
<td>400</td>
<td>0.0064</td>
<td>0.1645</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Efficient$^f$</td>
<td>400</td>
<td>0.0064</td>
<td>0.1645</td>
<td>0.1483</td>
<td>0.1533</td>
<td>0.925</td>
</tr>
<tr>
<td>IPW1</td>
<td>1000</td>
<td>0.0101</td>
<td>0.0847</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IPW2</td>
<td>1000</td>
<td>0.0150</td>
<td>0.0601</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Efficient</td>
<td>1000</td>
<td>0.0150</td>
<td>0.0601</td>
<td>0.0602</td>
<td>0.0613</td>
<td>0.940</td>
</tr>
<tr>
<td>(2) $\theta_2 = 0$, $(\pi(0, 0), \pi(0, 1), \pi(1, 0), \pi(1, 1)) = (0.208, 0.625, 0.139, 0.417)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPW1</td>
<td>400</td>
<td>-0.0047</td>
<td>0.1690</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IPW2</td>
<td>400</td>
<td>-0.0038</td>
<td>0.1351</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Efficient</td>
<td>400</td>
<td>-0.0038</td>
<td>0.1351</td>
<td>0.1267</td>
<td>0.1288</td>
<td>0.935</td>
</tr>
<tr>
<td>IPW1</td>
<td>1000</td>
<td>0.0043</td>
<td>0.0656</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IPW2</td>
<td>1000</td>
<td>0.0014</td>
<td>0.0516</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Efficient</td>
<td>1000</td>
<td>0.0014</td>
<td>0.0516</td>
<td>0.0510</td>
<td>0.0515</td>
<td>0.947</td>
</tr>
</tbody>
</table>

Note: $^a$ Sample mean of variance estimators for $\hat{\theta}_{2,n}$. $^b$ Asymptotically efficient variance of $\hat{\theta}_{2,n}$. $^c$ Coverage probability, based on the asymptotically normal distribution. $^d$ Inverse probability weighted estimation based on completely observed data, using true selection probabilities. $^e$ Inverse probability weighted estimation based on completely observed data, using estimated selection probabilities. $^f$ Efficient estimation.
Table 1b. Simulation summary statistics for estimating $\theta_2$ in logistic models with 1000 replications.

<table>
<thead>
<tr>
<th>Methods</th>
<th>$n$</th>
<th>mean($\hat{\theta}_{2,n}$)</th>
<th>$s^2(\hat{\theta}_{2,n})$</th>
<th>mean$^a$ var($\hat{\theta}_{2,n}$)</th>
<th>Optimal$^b$ var($\hat{\theta}_{2,n}$)</th>
<th>95%CP$^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(3) $\theta_2 = 0.693, (\pi(0, 0), \pi(0, 1), \pi(1, 0), \pi(1, 1)) = (0.25, 0.25, 0.25, 0.25)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPW1</td>
<td>400</td>
<td>0.7011</td>
<td>0.2825</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IPW2</td>
<td>400</td>
<td>0.7061</td>
<td>0.1918</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Efficient</td>
<td>400</td>
<td>0.7060</td>
<td>0.1917</td>
<td>0.1600</td>
<td>0.1669</td>
<td>0.921</td>
</tr>
<tr>
<td>IPW1</td>
<td>1000</td>
<td>0.7009</td>
<td>0.0996</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IPW2</td>
<td>1000</td>
<td>0.7000</td>
<td>0.0682</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Efficient</td>
<td>1000</td>
<td>0.7000</td>
<td>0.0682</td>
<td>0.0658</td>
<td>0.0668</td>
<td>0.938</td>
</tr>
</tbody>
</table>

(4) $\theta_2 = 0.693, (\pi(0, 0), \pi(0, 1), \pi(1, 0), \pi(1, 1)) = (0.208, 0.833, 0.139, 0.357)$

| IPW1    | 400 | 0.6920                      | 0.1849                      | -                                | -                                 | -         |
| IPW2    | 400 | 0.7019                      | 0.1490                      | -                                | -                                 | -         |
| Efficient | 400 | 0.7019                      | 0.1490                      | 0.1356                           | 0.1378                            | 0.934     |
| IPW1    | 1000 | 0.7058                       | 0.0701                      | -                                | -                                 | -         |
| IPW2    | 1000 | 0.7021                      | 0.0558                      | -                                | -                                 | -         |
| Efficient | 1000 | 0.7021                       | 0.0558                      | 0.0549                           | 0.0551                            | 0.944     |

Note: See Note to Table 1a.
Table 2. Simulation summary statistics for estimating $\theta_0$ and $\theta_1$ in linear models with 1000 replications.

<table>
<thead>
<tr>
<th>Methods</th>
<th>mean $\hat{\theta}_{0,n}$</th>
<th>mean $\hat{\theta}_{1,n}$</th>
<th>$s^2(\hat{\theta}_{0,n})$</th>
<th>$s^2(\hat{\theta}_{1,n})$</th>
<th>mean$^a$ var($\hat{\theta}_{0,n}$)</th>
<th>mean$^b$ var($\hat{\theta}_{1,n}$)</th>
<th>95%CP$^c$ $\hat{\theta}_{0,n}$</th>
<th>95%CP$^d$ $\hat{\theta}_{1,n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC$^e$</td>
<td>0.9955</td>
<td>0.9944</td>
<td>0.0201</td>
<td>0.0206</td>
<td>0.0192</td>
<td>0.0215</td>
<td>0.953</td>
<td>0.940</td>
</tr>
<tr>
<td>MSM$^f$</td>
<td>0.9973</td>
<td>0.9973</td>
<td>0.0172</td>
<td>0.0174</td>
<td>0.0172</td>
<td>0.0192</td>
<td>0.945</td>
<td>0.930</td>
</tr>
<tr>
<td>CSM$^g$</td>
<td>0.9960</td>
<td>0.9960</td>
<td>0.0149</td>
<td>0.0151</td>
<td>0.0150</td>
<td>0.0165</td>
<td>0.956</td>
<td>0.937</td>
</tr>
<tr>
<td>SM$^h$</td>
<td>1.0040</td>
<td>0.9928</td>
<td>0.1644</td>
<td>0.1715</td>
<td>0.1761</td>
<td>0.3542</td>
<td>0.954</td>
<td>0.935</td>
</tr>
</tbody>
</table>

$n = 200$

<table>
<thead>
<tr>
<th>Methods</th>
<th>mean $\hat{\theta}_{0,n}$</th>
<th>mean $\hat{\theta}_{1,n}$</th>
<th>$s^2(\hat{\theta}_{0,n})$</th>
<th>$s^2(\hat{\theta}_{1,n})$</th>
<th>mean$^a$ var($\hat{\theta}_{0,n}$)</th>
<th>mean$^b$ var($\hat{\theta}_{1,n}$)</th>
<th>95%CP$^c$ $\hat{\theta}_{0,n}$</th>
<th>95%CP$^d$ $\hat{\theta}_{1,n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>1.0032</td>
<td>1.0021</td>
<td>0.0040</td>
<td>0.0040</td>
<td>0.0041</td>
<td>0.0045</td>
<td>0.952</td>
<td>0.934</td>
</tr>
<tr>
<td>MSM</td>
<td>1.0033</td>
<td>1.0019</td>
<td>0.0034</td>
<td>0.0034</td>
<td>0.0034</td>
<td>0.0037</td>
<td>0.940</td>
<td>0.943</td>
</tr>
<tr>
<td>CSM</td>
<td>1.0030</td>
<td>1.0027</td>
<td>0.0030</td>
<td>0.0030</td>
<td>0.0032</td>
<td>0.0033</td>
<td>0.950</td>
<td>0.937</td>
</tr>
<tr>
<td>SM</td>
<td>1.0038</td>
<td>1.0032</td>
<td>0.0031</td>
<td>0.0031</td>
<td>0.0033</td>
<td>0.0034</td>
<td>0.945</td>
<td>0.931</td>
</tr>
</tbody>
</table>

$n = 1000$

Note: $^a$ Sample mean of variance estimators for $\hat{\theta}_{0,n}$. $^b$ Sample mean of variance estimators for $\hat{\theta}_{1,n}$. $^c$ Coverage probability for $\hat{\theta}_{0,n}$, based on the asymptotically normal distribution. $^d$ Coverage probability for $\hat{\theta}_{1,n}$, based on the asymptotically normal distribution. $^e$ Complete-case. $^f$ Mis-specified model. $^g$ Correctly specified model. $^h$ Smoothing method.