New estimating methods for surrogate outcome data

Bin Nan

University of Michigan

October 1, 2003

Abstract: Surrogate outcome data arise frequently in medical research. The true outcomes of interest are expensive or hard to ascertain, but measurements of surrogate outcomes (or more generally speaking, the correlates of the true outcomes) are usually available. In this paper we assume that the conditional expectation of the true outcome given covariates is known up to a finite dimensional parameter. When the true outcome is missing at random, the efficient score function for the parameter in the conditional mean model has a simple form, which is similar to the generalized estimating functions. There is no integral equation involved as in Robins, Rotnitzky and Zhao (1994) for general cases. We propose two estimating methods, parametric and nonparametric, to estimate the parameter by solving the efficient score equations. Simulation studies show the proposed estimators work well for reasonable sample sizes.

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

Running title: Surrogate Outcome

1 Introduction

It is not unusual in medical research that the outcome variables of interest are difficult or expensive to obtain. Often in these settings, surrogate outcome variables can be easily ascertained. Many real examples are described in the introductions of Pepe (1992) and Pepe, Reilly and Fleming (1994).

Suppose Y is the outcome of interest that is not always observable. Let S be a surrogate variable of Y, which is always available. The association of Y and a d-dimensional covariate vector $\mathbf{X}$ is the major interest. Pepe (1992) studied the model that both the conditional densities $f_\theta(Y|\mathbf{X})$ and $f_{\theta,\beta}(S|Y,\mathbf{X})$ are known up to the finite dimensional parameters θ and β. To avoid misspecification of the model $f_{\theta,\beta}(S|Y,\mathbf{X})$ that might result in an inconsistent estimate of θ, Pepe (1992) studied estimated likelihood method and later Pepe *et al.* (1994) proposed the mean score method by leaving the model $f(S|Y,\mathbf{X})$ unspecified. But the density $f_\theta(Y|\mathbf{X})$ still needs to be known and can be misspecified. Yet another disadvantage of their methods is that it is difficult to compute the conditional expectation of either the likelihood or the score function given $(S,\mathbf{X})$ when $(S,\mathbf{X})$ are continuous or highly dimensional. When both S and $\mathbf{X}$ are discrete, the mean score method of Pepe *et al.* (1994) is reduced to the Horvitz-Thompson type of inverse probability weighted method, see e.g. Horvitz and Thompson (1952).

Instead of modelling the whole density function $f_\theta(Y|\mathbf{X})$ parametrically, we very often like to relax the assumption and only assume that the conditional expectation of Y given $\mathbf{X}$ is known up to a parameter $\theta \in \mathrm{R}^d$, i.e.

$$E(Y|\mathbf{X}) = g(\mathbf{X};\theta) \ , \tag{1}$$

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

$$E(\epsilon|\mathbf{X}) = 0 \ . \tag{2}$$

Model (1) is semiparametric in the sense that there are three unknown functions in the underlying density function of $(S, Y, \mathbf{X})$: the conditional density function of S given $(Y, \mathbf{X})$; the conditional density function of Y, or equivalently ϵ, given $\mathbf{X}$; and the density function of $\mathbf{X}$. When we have complete data, i.e., Y is observable for all the subjects, the surrogate outcome S does not contribute to the estimation of θ. Chamberlain (1987), among others, showed that the asymptotically efficient estimator for θ for complete data can be obtained by solving the following estimating equation:

$$\sum_i \frac{\partial g(\mathbf{X}_i;\theta)/\partial\theta}{E(\epsilon^2|\mathbf{X}_i)}\ \epsilon_i = 0. \tag{3}$$

This equation has the same form as the quasi-likelihood estimating equation, see e.g. McCullagh (1983). Inevitably, the conditional variance $var(Y|\mathbf{X}) = E(\epsilon^2|\mathbf{X})$ needs to be specified/estimated. Carroll and Ruppert (1982) and Robinson (1987) showed that for linear models $g(\mathbf{X};\theta) = \mathbf{X}^T\theta$, substituting $E(\epsilon^2|\mathbf{X})$ by its kernel smoothing estimator into the above estimating equation yields the efficient estimator for θ. Extension to generalized linear models using smoothing techniques can be found in Newey (1993). Modelling $var(Y|\mathbf{X})$ parametrically is a useful alternative which avoids 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)), however, a surrogate outcome S is available for all the subjects. We call this type of data the surrogate outcome data as in Pepe (1992). Without specifying the joint distribution of $(S, Y, \mathbf{X})$, as what we will show later in this article, the semiparametrically efficient score equation for θ in model (1) for the surrogate outcome data actually has the same form as the efficient score function for complete data after certain "transformation" of Y. Thus the standard estimating methods using quasi-likelihood technique can be adopted with slight modifications to calculating the efficient estimators for the missing data problem. This problem is a special case of Robins *et al.*

(1994). Since we have not seen any specific result for this problem in a series of publications of Robins and colleagues since then, and clearly this special case is important enough to which a closer attention should be payed, we believe the results in this article are worth to be published.

In Section 2 we briefly introduce the efficient score function for the conditional mean regression model with surrogate outcome data. Our primary goal here in this paper is to study the estimating methods based on the efficient score function. Details are shown in Section 3. We show simulation results in Section 4.

2 The efficient score function

Suppose the association of Y and $\mathbf{X}$ can be modelled as in equation (1), or equivalently, in equation (2). 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., $pr(R = 1|S, Y, \mathbf{X}) = pr(R = 1|S, \mathbf{X}) \equiv \pi(S, \mathbf{X})$. We also assume that $\pi(S, \mathbf{X}) > \sigma > 0$ for some constant σ. We denote the observed data as

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

The following Theorem 1 gives us the efficient score function for θ.

Theorem 1. *The efficient score function l^*_θ for the observed data $(S, RY, \mathbf{X}, R)$ in the conditional mean model (1) is given by*

$$l^*_\theta = \frac{\partial g(\mathbf{X}; \theta)/\partial\theta}{E(\epsilon^{*2}|\mathbf{X})}\ \epsilon^* \ , \tag{4}$$

where

$$\epsilon^* = \frac{R}{\pi}Y - \frac{R - \pi}{\pi}E(Y|S, \mathbf{X}) - g(\mathbf{X}; \theta) \ . \tag{5}$$

Let $Y^* = (R/\pi)Y - \{(R-\pi)/\pi\}E(Y|S,\mathbf{X})$ be a kind of "transformation" to the response variable Y. Using the nested conditional expectation property, we can easily verify that $E(Y^*|\mathbf{X}) = E(Y|\mathbf{X}) = g(\mathbf{X};\theta)$. Hence by comparing the summand in equation (3) and the right hand side of equation (4) we see that the efficient score l^*_θ actually has the same form as that of the efficient score for the "full" data $(Y^*, \mathbf{X})$. So analyzing the observed data $(S, RY, \mathbf{X}, R)$ with the outcome Y missing at random and the availability of surrogate outcome S is actually equivalent to analyzing the "full" data $(Y^*, \mathbf{X})$ with the same conditional mean structure as that of $(Y, \mathbf{X})$. The interpretation of the parameter θ does not change at all, even though the scale of Y^* may not be the same as Y. The only complication is that $E(Y|S,\mathbf{X})$ in (5) may not be observable and need to be estimated.

Equation (4) can be obtained using the general method proposed by Robins, Rotnitzky and Zhao (1994) for problems with data missing at random. The beauty of equation (4) is that it has a simple clean form without any involvement of integral equations, which may not be the case for a variety of models with missing data. Many applications of Robins *et al.* (1994) can be found in literature. Among these applications, Holcroft, Rotnitzky and Robins (1997) studied efficient estimations for three-stage designs; Nan, Emond and Wellner (2002) developed the efficient score functions for Cox models with missing data where integral equations are involved. For those who are interested in applying Robins *et al.* (1994), we put a brief derivation of equation (4) in the Appendix as another application.

3 Two estimating methods

We develop estimators for θ based on the efficient score function (4) in this section. In some medical studies, the surrogate outcome might have been well investigated such that the functional form of $E(Y|S,\mathbf{X})$ could be estimated from previous studies, especially when S satisfies the surrogate criterion of Prentice (1989), i.e. $E(Y|S,\mathbf{X}) = E(Y|S)$. Then Y^*

can be obtained for each record in a new study, and the observed data can be treated as independent and identically distributed copies of $(Y^*, \mathbf{X})$. Thus the estimation using efficient score (4) can be a standard practice of the quasi-likelihood methods.

The more interesting case would be the situation that no previous study is available for estimating $E(Y|S, \mathbf{X})$. We propose two basic estimating methods in this section. As what we will discuss in Section 5, these two methods may be mixed.

3.1 Parametric method

The first method we discuss in this article is to specify the functional forms of $E(Y|S, \mathbf{X})$ and $E(\epsilon^{*2}|\mathbf{X})$ parametrically. Suppose that

$$E(Y|S, \mathbf{X}) = m(S, \mathbf{X}; \beta) \ , \tag{6}$$

and

$$E(\epsilon^{*2}|\mathbf{X}) = v(\mathbf{X}; \gamma) \ . \tag{7}$$

Here $m(\cdot, \cdot)$ and $v(\cdot)$ are known functions up to the finite dimensional parameters β and γ. A variety of choices and extensive discussion of modelling the functional forms of the conditional variances can be found in Carroll and Ruppert (1988). Then from (4) the efficient score equation for θ with cohort size n is

$$\sum_{i=1}^{n} \frac{\partial g(\mathbf{X}_i; \theta)/\partial\theta}{v(\mathbf{X}_i; \gamma)} \left\{ \frac{R_i}{\pi(S_i, \mathbf{X}_i)} Y_i \right. \\ \left. - \ \frac{R_i - \pi(S_i, \mathbf{X}_i)}{\pi(S_i, \mathbf{X}_i)} m(S_i, \mathbf{X}_i; \beta) - g(\mathbf{X}_i; \theta) \right\} = 0 \ . \tag{8}$$

The parameters β and γ can be estimated from the following equations based on models (6) and (7):

$$\sum_{i=1}^{n} R_i A(S_i, \mathbf{X}_i) \Big\{ Y_i - m(S_i, \mathbf{X}_i; \beta) \Big\} = 0, \tag{9}$$

and

$$\sum_{i=1}^{n} B(\mathbf{X}_i)\Big\{\epsilon_i^{*2} - v(\mathbf{X}_i;\gamma)\Big\} = 0, \tag{10}$$

where $A(S,\mathbf{X})$ and $B(\mathbf{X})$ are usually chosen as the vectors of covariates in corresponding models, though they can be arbitrary functions. Notice that the efficient score equation (8) and equation (10) use all the data including both completely observed and partially observed data, while equation (9) uses the completely observed data.

Model (6) looks like the imputation method of Chen (2000). But the estimating equation (8) 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 the data are missing completely at random (Little and Rubin, 1987, Ch. 1).

If both of the functional forms (6) and (7) are correctly specified, then both estimating functions in (9) and (10) are unbiased, and the root of equation (8) will be a semiparametrically efficient estimator of θ under the conditions described in the following Theorem 2. Usually specifying the functional form of $E(\epsilon^{*2}|\mathbf{X})$ could be difficult due to the "transformation", even though the form of $E(\epsilon^2|\mathbf{X})$ would be simple. However, no matter whether $m(S,\mathbf{X};\beta)$ and $v(\mathbf{X};\gamma)$ can be correctly specified, the estimator from equation (8) will always be consistent and asymptotically normally distributed whenever the probability π is correctly specified since the estimating function in (8) is unbiased. This can be guaranteed for two-phase sampling studies where π is determined by investigators. On the other hand, the estimator from equation (8) will always be consistent and asymptotically normally distributed whenever the functional form of $m(S,\mathbf{X};\beta)$ is correctly specified. Thus the estimating equation (8) has the so called double robustness feature discussed by Robins *et al.* (1994). In practice, the traditional residual diagnosis tools can be used to explore the appropriate forms of $m(S,\mathbf{X};\beta)$ and $v(\mathbf{X};\gamma)$. In the rest of the paper we shall assume that π is known. If π is unknown, then it needs to be estimated either parametrically or

nonparametrically from the observed data. Estimation of π when it is unknown is not a focus of this paper.

Remark 1: If the true functional forms of models (6) and (7) were known, then in principle the efficiency of the estimator from equation (8) could be improved because we had more information than just the mean structure, and the function (4) might no longer be the fully efficient score function. We do not go any further here along this line since the efficiency gain comes at a price that the resulting estimator may be inconsistent if the forms of models (6) and (7) are not correctly specified. See Newey (1993) for a discussion. We call the estimator obtained from equation (8) the "semiparametrically efficient estimator" when the functional forms of models (6) and (7) are correctly specified, since it asymptotically achieves the information bound that assumes full knowledge of the first moment (1) and Y missing at random.

We now outline an algorithm for obtaining the estimator by solving equation (8):

Algorithm 1:

Step 1: Estimate β from equation (9);

Step 2: For an initial value of $\hat{\theta}_n$, $\hat{\theta}_{(0)}$, estimate γ from equation (10);

Step 3: Calculate the predicted values $m(S_i, \mathbf{X}_i; \hat{\beta}_n)$ and $v(\mathbf{X}_i; \hat{\gamma}_n)$ for all $i \in \{1, \ldots, n\}$, then plug them into equation (8);

Step 4: Solve equation (8) to obtain $\hat{\theta}_n$; and use the root 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

$$\left(\sum_{i=1}^{n} \dot{l}^*_{\hat{\theta}_n,i} \right)^{-1} \left(\sum_{i=1}^{n} l^*_{\hat{\theta}_n,i} l^{*T}_{\hat{\theta}_n,i} \right) \left(\sum_{i=1}^{n} \dot{l}^*_{\hat{\theta}_n,i} \right)^{-1},$$

where $\dot{l}^*_\theta = \partial l^*_\theta / \partial \theta$. When (6) and (7) are correctly specified, the variance of $\hat{\theta}_n$ can be

estimated by $\left(\sum_{i=1}^{n} l^*_{\hat{\theta}_n,i} l^{*T}_{\hat{\theta}_n,i} \right)^{-1}$, which may be called as the model-based variance estimator.

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

$$\sum_{i=1}^{n} \frac{R_i}{\pi(S_i, \mathbf{X}_i)} \frac{\partial g(\mathbf{X}_i;\theta)/\partial\theta}{var(Y|\mathbf{X}_i)} \Big\{ Y_i - g(\mathbf{X}_i;\theta) \Big\} = 0 \ . \tag{11}$$

If the variance function in equation (7) is modelled without any parameter γ, a function of g for generalized linear models for instance, then the above algorithm does not need to be repeated.

The following Theorem 2 gives the asymptotic properties of the estimators calculated from Algorithm 1. A sketched proof is given in the Appendix.

Theorem 2. *Suppose that the score function (4) and the estimating functions for models (6) and (7) satisfy the regularity conditions in Foutz (1977), and the score function (4) has bounded second derivative to θ in a neighborhood of the true θ_0. Then the root of equation (8), $\hat{\theta}_n$, is a sequence of consistent and normally distributed estimators for θ_0, i.e., $n^{1/2}(\hat{\theta}_n - \theta_0)$ converges in distribution to a mean zero normal distribution with variance*

$$\Big\{ E(\dot{l}^*_{\theta_0}) \Big\}^{-1} \Big\{ E(l^*_{\theta_0} l^{*T}_{\theta_0}) \Big\} \Big\{ E(\dot{l}^*_{\theta_0}) \Big\}^{-1}$$

as $n \to \infty$. When models (6) and (7) are correctly specified, the estimator $\hat{\theta}_n$ is semiparametrically efficient with asymptotic variance $\Big\{ E(l^*_{\theta_0} l^{*T}_{\theta_0}) \Big\}^{-1}$.

3.2 Nonparametric method

In order to avoid misspecification of the functional forms of $E(Y|S, \mathbf{X})$ and $E(\epsilon^{*2}|\mathbf{X})$, nonparametric methods, such as kernel or spline smoothing techniques, can be used to estimate

$E(Y|S,\mathbf{X})$ and $E(\epsilon^{*2}|\mathbf{X})$. In this case, both $E(Y|S,\mathbf{X})$ and $E(\epsilon^{*2}|\mathbf{X})$ may be viewed as (infinitely dimensional) nuisance parameters in the efficient score function l^*_θ. Let $\eta = (\eta_1, \eta_2)$, where $\eta_1 = E(Y|S,\mathbf{X})$ and $\eta_2 = E(\epsilon^{*2}|\mathbf{X})$. We rewrite l^*_θ as $l^*_{\theta,\eta}$. Thus,

$$l^*_{\theta,\eta} = \frac{\partial g(\mathbf{X};\theta)/\partial\theta}{\eta_2}\left\{\frac{R}{\pi}Y - \frac{R-\pi}{\pi}\eta_1 - g(\mathbf{X};\theta)\right\} , \tag{12}$$

and θ can be estimated by solving the following estimating equation:

$$\sum_{i=1}^{n} l^*_{\theta,\hat{\eta}_n(\theta)}(S_i, R_iY_i, \mathbf{X}_i, R_i) = 0 . \tag{13}$$

The following algorithm can be used to obtain the estimator $\hat{\theta}_n$ for θ by solving the equation (13):

Algorithm 2:

Step 1: Estimate $\eta_1 = E(Y|S,\mathbf{X})$ via smoothing using all the fully observed records. Note that the observing probabilities for those records vary. So the i-th fully observed record should have weight $1/\pi(S_i,\mathbf{X}_i)$. For the records with missing data, the corresponding values of η_1 are predicted from the estimated model.

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

Step 3: Calculate Y_i^* and thus the residuals $\epsilon_i^* = Y_i^* - g(\mathbf{X}_i;\hat{\theta}_{(0)})$, $i = 1,\ldots,n$. Then estimate $\eta_2 = E(\epsilon^{*2}|\mathbf{X})$ using smoothing to the squares of residuals with respect to $\mathbf{X}_i$.

Step 4: Plug $\hat{\eta}_1$ and $\hat{\eta}_2(\hat{\theta}_{(0)})$ into equation (13) and solve the equation for $\hat{\theta}_n$.

Step 5: Use the root of equation (13) 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 $\left(\sum_{i=1}^n l^*_{\hat{\theta}_n,\hat{\eta}_n,i} l^{*T}_{\hat{\theta}_n,\hat{\eta}_n,i}\right)^{-1}$.

When $\mathbf{X}$ is discrete, $E(\epsilon^{*2}|\mathbf{X}) = var(Y^*|\mathbf{X})$ can be estimated from grouped data grouping on distinct values of $\mathbf{X}$ without using residuals, and thus no iteration in the above algorithm is needed. When both S and $\mathbf{X}$ are discrete, the two algorithms can be unified.

The asymptotic properties of the estimator calculated by Algorithm 2 can be argued by Theorem 6.21 of van der Vaart (2000), which is presented here as Theorem 3. Since the function (12) is the efficient score function, the root of equation (13) is a fully efficient regular estimator, i.e., $n^{1/2}(\hat{\theta}_n - \theta_0)$ converges in distribution to a mean zero normal distribution with variance $\left\{E(l^*_{\theta_0,\eta_0} l^{*T}_{\theta_0,\eta_0})\right\}^{-1}$ as $n \to \infty$, under certain regularity conditions that need to be specified for different models and with fairly slow rates of convergence (say $n^{-1/4}$) for $\hat{\eta}_1$ and $\hat{\eta}_2$. See van der Vaart (2000) for a nice discussion that follows Theorem 6.21.

Theorem 3. *Suppose that the class of functions* $\{\psi_{\theta,\eta} : \|\theta - \theta_0\| < \delta, d(\eta, \eta_0) < \delta\}$ *is Donsker for some* $\delta > 0$*, that the maps* $\theta \mapsto E\psi_{\theta,\eta}$ *are differentiable at* θ_0*, uniformly in* η *in a neighborhood of* η_0 *with nonsingular derivative matrices* $V_{\theta_0,\eta}$ *such that* $V_{\theta_0,\eta} \to V_{\theta_0,\eta_0}$ *as* $\eta \to \eta_0$*, and assume that the map* $(\theta, \eta) \mapsto \psi_{\theta,\eta}$ *is continuous in* $L_2(P)$ *at* (θ_0, η_0)*. If* $n^{-1/2} \sum \psi_{\hat{\theta}_n,\hat{\eta}_n} = o_p(1)$ *and* $(\hat{\theta}_n, \hat{\eta}_n) \to (\theta_0, \eta_0)$ *in probability for a point* (θ_0, η_0) *satisfying* $E\psi_{\theta_0,\eta_0} = 0$*, then*

$$n^{1/2}(\hat{\theta}_n - \theta_0) = -V^{-1}_{\theta_0,\eta_0} n^{1/2} E\psi_{\theta_0,\hat{\eta}_n} - V^{-1}_{\theta_0,\eta_0} n^{-1/2} \sum \psi_{\theta_0,\eta_0} + o_p(1 + n^{1/2}\|E\psi_{\theta_0,\hat{\eta}_n}\|).$$

In our cases $\psi_{\theta,\eta}$ is replaced by the efficient score $l^*_{\theta,\eta}$, and thus we have the nice results as discussed above, where we expect to have $n^{1/2} E\psi_{\theta_0,\hat{\eta}_n} = o_p(1)$ and the asymptotic normality becomes clear from Theorem 3.

4 Numerical examples

We conduct simulations to investigate the finite sample performance of the efficient estimators using "transformed" response variable, and compare it with the inverse probability weighted estimators using fully observed data only. 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 may be the true disease status and S may be 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 not binary when Y is observed only for a subsample of the subjects. Suppose that $pr(S = 1|Y, X) = pr(S = 1|Y)$. 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

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

Suppose we observe S and X for all subjects in the study, and observe Y for a subsample with 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 $q(1) = 0.9$. In the terminology of 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 θ_1 is of less interest. We only show the simulation results for estimating θ_2. We choose two different values for θ_2: 0 and $\log(2) \approx 0.693$. We also choose two sets of selection probabilities for each value of θ: $\pi(S, X) = 0.25$ for all S and X and $(\pi(0,0), \pi(0,1), \pi(1,0), \pi(1,1)) = (0.208, 0.625, 0.139, 0.417)$ 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 are chosen such that the numbers of selected subjects in all four cells of $(S, X) \in \{(0,0), (0,1), (1,0), (1,1)\}$ are expected to be the same.

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 θ_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 efficient estimator using all the observed data. The variance structure for completely observed data analysis is "$\mu(1-\mu)$". For the efficient estimating procedure, both the conditional expectation $E(Y|S,X)$ and the variance $var(Y^*|X) = E(\epsilon^{*2}|X)$ are estimated nonparametrically using grouped data. The results are listed in Table 1.

[Tables 1 is about here.]

Chen and Breslow (2003) (where they primarily argue that the efficient score can also be obtained 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 π is asymptotically fully efficient. This is verified from the simulation results listed in Table 1.

Table 1 shows that the 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 should not hold in general. Not surprisingly, both of these methods perform better than the inverse probability weighted methods using true selection probabilities. The variance estimators of the efficient estimating methods are valid, and they are close to the asymptotically optimal or minimum variances (based on theoretical calculations) among all regular estimators. The 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

$$E(Y|S,X) = \theta_0 + \theta_1 X + \theta_2 f(S), \tag{14}$$

and the model of interest is

$$E(Y|X) = \theta_0 + \theta_1 X, \tag{15}$$

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 $\theta_0 = \theta_1 = \theta_2 = 1$. The simulations are conducted using 1000 replications with cohort size $n =$ 200 and 1000, respectively, and the selection probability $\pi(S,X) = 0.5$, which means that the fully observed subsample is a simple random sample of the cohort. Here S does not satisfy the definition of surrogate outcome given by Prentice (1989). However, since it is correlated with the true outcome Y, we can use it in the same way as surrogate outcome discussed in this article to improve efficiency. Four estimating methods are simulated: (i) fitting linear regression model (15) 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 (14); and (iv) using Algorithm 2 with the functional form of $E(Y|S,X)$ being estimated via the generalized additive models with smoothing splines on S and X, which is the smoother $s()$ in Splus with default values for smoothing parameters. We do not need to model variances in Equations (8) and (9) 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 generalized additive models with spline smoothing, Algorithm 2 works equally well as Algorithm 1 with correctly specified parametric structure of $E(Y|S,X)$.

[Table 2 is about here.]

The simulations have been programmed and run in R to take advantages of existing regression functions in R.

5 Discussion

The assumption of fixed selection probabilities is not necessary. The proposed methods still work if we can estimate $\pi(S, \mathbf{X})$, see e.g. Robins *et al.* (1994). In two-phase sampling designs, $\pi(S, \mathbf{X})$ is determined by investigators. Thus obtaining the optimal $\pi(S, \mathbf{X})$ is an interesting problem and remains to be explored. We propose two algorithms to obtain the efficient estimator for θ in this paper. The two algorithms proposed in this article may be mixed, i.e., $E(Y|S, \mathbf{X})$ can be estimated parametrically, while $E(\epsilon^{*2}|\mathbf{X})$ is estimated nonparametrically. Algorithm 1 can easily handle continuous or highly dimensional covatiates and surrogate outcome. But we may face the problem of misspecifying $E(Y|S, \mathbf{X})$ and $E(\epsilon^{*2}|\mathbf{X})$, which may cause efficiency loss. Algorithm 2 is able to avoid the misspecification problem, and thus more robust. But we need to keep in mind that it inherits both the advantages and the disadvantages of smoothing techniques.

Actually, estimating equation (8) can be reduced to different forms when we choose different ways of estimating $E(Y|S, \mathbf{X})$. If we replace $E(Y|S, \mathbf{X})$ by the true observations of Y, then equation (8) becomes the estimating equation for full data. If S does not contribute, i.e., $E(Y|S, \mathbf{X}) = E(Y|\mathbf{X})$, equation (8) becomes the inverse-probability weighted Horvitz-Thompson estimating equation (11). If we replace $E(Y|S, \mathbf{X})$ by Y when $R = 1$, and view $E(Y|S, \mathbf{X})$ as an imputation of Y when $R = 0$, then equation (8) becomes the imputation method estimating equation, similar to the method discussed by Chen (2000). Intuitively, $E(Y|S, \mathbf{X})$ would be a better prediction of Y than $E(Y|\mathbf{X})$ even when S is treated as an extra covariate in the model of $E(Y|\mathbf{X})$. Hence we would expect that the estimating equation (8) yields more precise estimator than the inverse-probability weighted Horvitz-Thompson estimating method, no matter whether the functional form of $E(Y|S, \mathbf{X})$ is correctly specified.

We notice that both equations (8) and (13) can be viewed as the profile likelihood esti-

mating equations, which "profile out" the nuisance parameters. This is why we obtain the semiparametrically efficient estimators from both of the algorithms (when $E(Y|S,\mathbf{X})$ and $E(\epsilon^{*2}|\mathbf{X})$ are correctly specified for Algorithm 1). See Murphy and van der Vaart (2000) for nice discussions about the profile likelihood methods under semiparametric settings.

Acknowledgements

The author owes thanks to Norman Breslow, Margaret Pepe, and Jon Wellner for their kind help in this research.

Appendix: Proofs of Theoretical Results

Proof of Theorem 1. We apply the following Lemma 1, a specialized result of Robins *et al.* (1994), to prove Theorem 1 for the mean regression model with surrogate outcome data. Suppose θ is the parameter of interest as in model (1), and η is the nuisance parameter which is a vector of three unknown functions: the conditional density function of S given $(Y,\mathbf{X})$, the conditional density function of Y given $\mathbf{X}$, and the density function of $\mathbf{X}$. Let $\dot{\mathcal{Q}}_\eta$ be the nuisance tangent space for the full data model. We refer to Bickel, Klaassen, Ritov and Wellner (1993) for the definitions and properties of the tangent spaces. Let $\dot{\mathcal{Q}}_\eta^\perp$ be the orthogonal complement of $\dot{\mathcal{Q}}_\eta$ in $L_2^0(Q)$, the space of all functions with mean zero and finite second moment with respect to the distribution function Q of the underlying full data $(S,Y,\mathbf{X})$. Then we have:

Lemma 1. *Let l_θ^* be the efficient score for θ in the model of observed data $(S,RY,\mathbf{X},R)$, and l_θ^{*0} the efficient score for θ in the model of underlying full data $(S,Y,\mathbf{X})$. Then*

$$l_\theta^* = \frac{R}{\pi} D(S,Y,\mathbf{X}) - \frac{R-\pi}{\pi} E\{D(S,Y,\mathbf{X})|S,\mathbf{X}\} \ , \tag{16}$$

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

$$\Pi\left(\frac{1}{\pi}D - \frac{1-\pi}{\pi}E(D|S,\mathbf{X})\Big|\dot{\mathcal{Q}}_\eta^\perp\right) = l_\theta^{*0} \ . \tag{17}$$

Here Π *is a projection operator.*

Lemma 1 should hold for any regression models with data missing at random, and usually equation (17) is an integral equation. Alternative proof of Lemma 1 can be found in Nan (2001). In order to apply Lemma 1 to compute the efficient score l_θ^*, we need three ingredients: (1) the efficient score function for the full data model; (2) the characterization of the space $\dot{\mathcal{Q}}_\eta^\perp$; (3) the calculation of the projection of functions onto the space $\dot{\mathcal{Q}}_\eta^\perp$. All of them are from the full data model, which suggests that we can take advantages of possibly nicer structures of full data models to deal with missing data problems.

From Chamberlain (1987), Robins *et al.* (1994), van der Vaart (1998), and Nan, Emond and Wellner (2000), we know that for the underlying full data $(S, Y, \mathbf{X})$, the efficient score function for θ is:

$$l_\theta^{*0} = \frac{\partial g(\mathbf{X};\theta)/\partial\theta}{E(\epsilon^2|\mathbf{X})}\ \epsilon \ , \tag{18}$$

the space $\dot{\mathcal{Q}}_\eta^\perp$ has the following simple form:

$$\dot{\mathcal{Q}}_\eta^\perp = \Big\{h(\mathbf{X})\epsilon : \ E\{h^2(\mathbf{X})\epsilon^2\} < \infty\Big\} \ , \tag{19}$$

and we have the following projection:

$$\Pi(b|\dot{\mathcal{Q}}_\eta^\perp) = \frac{E\{b(S,Y,\mathbf{X})\epsilon|\mathbf{X}\}}{E(\epsilon^2|\mathbf{X})}\ \epsilon \ , \quad \text{for all } b \in L_2^0(Q) \ . \tag{20}$$

Now we are ready to prove Theorem 1. Let $D(Y, \mathbf{X}) = h(\mathbf{X})\epsilon$ and plug it into equation (17). Then from equations (18), (19) and (20) we obtain

$$\begin{aligned}
\frac{\partial g(\mathbf{X};\theta)/\partial\theta}{E(\epsilon^2|\mathbf{X})}\epsilon &= \frac{1}{E(\epsilon^2|\mathbf{X})}E\left[\frac{1}{\pi}h(\mathbf{X})\epsilon^2 - \epsilon\frac{1-\pi}{\pi}E\{h(\mathbf{X})\epsilon|S,\mathbf{X}\}\Big|\mathbf{X}\right]\epsilon \\
&= \frac{1}{E(\epsilon^2|\mathbf{X})}E\left\{\frac{1}{\pi}\epsilon^2 - \frac{1-\pi}{\pi}E^2(\epsilon|S,\mathbf{X})\Big|\mathbf{X}\right\}h(\mathbf{X})\epsilon \ .
\end{aligned}$$

Simplifying the above equality yields

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

Hence from equation (16) we obtain the efficient score l_θ^* for the observed data in the conditional mean model (1), which is give by

$$\begin{aligned} l_\theta^* &= \frac{R}{\pi}h(\mathbf{X})\epsilon - \frac{R-\pi}{\pi}E\{h(\mathbf{X})\epsilon|S,\mathbf{X}\} \\ &= h(\mathbf{X})\left\{\frac{R}{\pi}\epsilon - \frac{R-\pi}{\pi}E(\epsilon|S,\mathbf{X})\right\} \\ &= h(\mathbf{X})\left\{\frac{R}{\pi}Y - \frac{R-\pi}{\pi}E(Y|S,\mathbf{X}) - g(\mathbf{X};\theta)\right\} \\ &= \frac{\partial g(\mathbf{X};\theta)/\partial\theta}{E(\epsilon^{*2}|\mathbf{X})}\;\epsilon^* \end{aligned}$$

where

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

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

Proof of Theorem 2. Consistency can be proved by Foutz (1977) since θ can be viewed as an element of the finite dimensional parameter vector (θ,β,γ). Asymptotic normality and the limiting variance can be shown by Taylor expansion of equation (8) around θ_0. Efficiency is guaranteed when both models (6) and (7) are correctly specified since (8) is a semiparametrically efficient estimating equation.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

Carroll, R. J. and Ruppert D. (1982) Robust estimation in heteroscedastic linear models. *The Annals of Statistics* **10**, 429-441.

Carroll, R. J. and Ruppert D. (1988) *Transformation and Weighting in Regression.* Chapman and Hall.

Chamberlain, G. (1987) Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* **34**, 305-324.

Chen, Y.-H (2000) A robust imputation method for surrogate outcome data. *Biometrika* **87**, 711-716.

Chen, Y.-H and Chen, H. (2000) A unified approach to regression analysis under double-sampling designs. *J. Roy. Statist. Soc. B* **62**, 449-460.

Chen, J. and Breslow, N. E. (2003) Semiparametric efficient estimation for the auxiliary outcome problem with the conditional mean model. *Technical Report.*

Foutz, R. V. (1977) On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association* **72**, 147-148.

Holcroft, C. A., Rotnitzky, A. and Robins, J. M. (1997) Efficient estimation of regression parameters from multistage studies with validation of outcome and covariates. *Journal of Statistical Planning and Inference* **65**, 349-374.

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

Little, R. J. A. and Rubin, D. (2002) *Statistical Analysis with Missing Data, 2nd Ed.* John Wiley, New York.

McCullagh, P. (1983) Quasi-likelihood functions. *The Annals of Statistics* **11**, 59-67.

Murphy, S. A. and van der Vaart, A. W. (2000) On profile likelihood (with comments and a rejoinder by the authors). *Journal of the American Statistical Association* **95**, 449-485.

Nan, B. (2001) *Information Bounds and Efficient Estimation for Two-Phase Designs with Lifetime Data.* Ph. D. dissertation. University of Washington, Department of Biostatistics.

Nan, B., Emond, M. and Wellner, J. A. (2000) Information bounds for regression models with missing data. *Technical Report* **378**, Department of Statistics, University of Washington.

Nan, B., Emond, M. and Wellner, J. A. (2002) Information bounds for Cox regression models with missing data. *The Annals of Statistics*, to be published.

Newey, W. K. (1993) Efficient estimation of models with conditional moment restrictions. *Handbook of Statistics* (eds G. S. Maddala, C. R. Rao, and H. D. Vinod), vol. 11, pp. 419-454. Elsevier Science Publisher B.V.

Pepe, M. S. (1992) Inference using surrogate outcome data and a validation sample. *Biometrika* **79**, 355-365.

Pepe, M. S., Reilly, M. and Fleming, T. R. (1994) Auxiliary outcome data and the mean score method. *Journal of Statistical Planning and Inference* **42**, 137-160.

Prentice, R. L. (1989) Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine* **8**, 431-440.

Robins, J. M. and Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90**, 122-129.

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

Robinson, P. M. (1987) Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica* **55**, 875-891.

Rotnitzksy, A. and Robins, J. M. (1995) Semi-parametric estimation of models for means and covariances in the presence of missing data. *Scand. J. Statist.* **22**, 323-333.

van der Vaart, A. W. (1998) *Asymptotic Statistics.* Cambridge: Cambridge University Press.

van der Vaart, A. W. (2000) *Semiparametric Statistics.* Lectures on Probability Theory, Ecole d'Eté de Probabilités de St. Flour 1999 (ed P. Bernard), Berlin: Springer, to appear.

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 θ_2 in logistic models with 1000 replications.

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

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 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 θ_2 in logistic models with 1000 replications.

Methods	n	mean$(\hat{\theta}_{2,n})$	$s^2(\hat{\theta}_{2,n})$	mean[a] var$(\hat{\theta}_{2,n})$	Optimal[b] var$(\hat{\theta}_{2,n})$	95%CP[c]
(3) $\theta_2 = 0.693$, $(\pi(0,0), \pi(0,1), \pi(1,0), \pi(1,1)) = (0.25, 0.25, 0.25, 0.25)$						
IPW1	400	0.7011	0.2825	-	-	-
IPW2	400	0.7061	0.1918	-	-	-
Efficient	400	0.7060	0.1917	0.1600	0.1669	0.921
IPW1	1000	0.7009	0.0996	-	-	-
IPW2	1000	0.7000	0.0682	-	-	-
Efficient	1000	0.7000	0.0682	0.0658	0.0668	0.938
(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 θ_0 and θ_1 in linear models with 1000 replications.

Methods	mean $\hat{\theta}_{0,n}$	mean $\hat{\theta}_{1,n}$	$s^2(\hat{\theta}_{0,n})$	$s^2(\hat{\theta}_{1,n})$	mean[a] var$(\hat{\theta}_{0,n})$	mean[b] var$(\hat{\theta}_{1,n})$	95%CP[c] $\hat{\theta}_{0,n}$	95%CP[d] $\hat{\theta}_{1,n}$
				$n = 200$				
CC[e]	0.9955	0.9944	0.0201	0.0206	0.0192	0.0215	0.953	0.940
MSM[f]	0.9973	0.9973	0.0172	0.0174	0.0172	0.0192	0.945	0.930
CSM[g]	0.9960	0.9960	0.0149	0.0151	0.0150	0.0165	0.956	0.937
SM[h]	1.0040	0.9928	0.1644	0.1715	0.1761	0.3542	0.954	0.935
				$n = 1000$				
CC	1.0032	1.0021	0.0040	0.0040	0.0041	0.0045	0.952	0.934
MSM	1.0033	1.0019	0.0034	0.0034	0.0036	0.0037	0.940	0.943
CSM	1.0030	1.0027	0.0030	0.0030	0.0032	0.0033	0.950	0.937
SM	1.0038	1.0032	0.0031	0.0031	0.0033	0.0034	0.945	0.931

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 asymptotically normal distribution. [d] Coverage probability for $\hat{\theta}_{1,n}$, based on asymptotically normal distribution. [e] Complete-case. [f] Mis-specified model. [g] Correctly specified model. [h] Smoothing method.