The Canadian Journal of Statistics
Vol. 32, No. 4, 2004, Pages 1–14
La revue canadienne de statistique

Semiparametric efficient estimation for the auxiliary outcome problem with the conditional mean model

Jinbo CHEN and Norman E. BRESLOW

Key words and phrases: Auxiliary outcome; conditional mean model; Horvitz–Thompson estimator; missing at random; semiparametric efficient estimation.

MSC 2000: Primary: 62D05, 62J12; secondary: 62H12.

Abstract: The authors consider semiparametric efficient estimation of parameters in the conditional mean model for a simple incomplete data structure in which the outcome of interest is observed only for a random subset of subjects but covariates and surrogate (auxiliary) outcomes are observed for all. They use optimal estimating function theory to derive the semiparametric efficient score in closed form. They show that when covariates and auxiliary outcomes are discrete, a Horvitz–Thompson type estimator with empirically estimated weights is semiparametric efficient. The authors give simulation studies validating the finite-sample behaviour of the semiparametric efficient estimator and its asymptotic variance; they demonstrate the efficiency of the estimator in realistic settings.

Estimation semiparamétriquement efficace pour le problème du résultat auxiliaire dans le modèle à moyenne conditionnelle

Résumé : Les auteurs s'intéressent à l'estimation semiparamétriquement efficace de paramètres dans le modèle à moyenne conditionnelle pour une structure de données incomplète simple dans laquelle l'événement d'intérêt n'est observé que pour un sous-ensemble aléatoire de sujets alors que les covariables et les variables de substitution (auxiliaires) sont observées pour tous. Ils font appel à la théorie des fonctions d'estimation optimales pour déterminer le score semiparamétriquement efficace de façon explicite. Ils montrent que lorsque les covariables et les variables auxiliaires sont discrètes, un estimateur de type Horvitz–Thompson à poids estimés empiriquement est semiparamétriquement efficace. Les auteurs présentent des études de simulation validant le comportement à taille finie de l'estimateur semiparamétriquement efficace et de sa variance asymptotique; ils démontrent en outre l'efficacité de cet estimateur dans des contextes réalistes.

1. INTRODUCTION

Medical research frequently aims to study the association between an outcome variable and a set of covariates. Sometimes, it is feasible to obtain a crude outcome measure on a large sample, but the true outcome can be ascertained only for a subsample. Alonzo, Pepe & Lumley (2003), for example, analyzed data from the Great Smoky Mountains Study (Costelo et al. 1996) on the prevalence of depression among adolescents in western North Carolina. Subjects were first assessed using an inexpensive screening test; those who scored above a certain threshold and a subset of those who scored below were selected for definitive diagnosis of depression. Pepe, Reilly & Fleming (1994) described a setting where patients who received allogeneic bone marrow transplant for aplastic anemia may develop graft versus host disease (GVHD). Measuring chronic GVHD requires longitudinal follow-up, but acute GVHD can be readily ascertained when patients are still being treated at the hospital. Since young patients without a history of acute GVHD are thought to be at low risk of chronic GVHD, it is cost-effective to follow only a fraction of them, but all who do have such a history, for diagnosis of the chronic outcome. See Pepe, Reilly & Fleming (1994) for additional examples.

Variables such as the screening test result and acute GVHD are often called surrogate outcomes. Some authors, for example, Prentice (1989), have advocated using the surrogate outcome in place of the true outcome to make scientific inference, especially when the true outcome is im-

possible to measure. In the two examples mentioned, however, it was possible to measure the true outcome on a subsample. In this situation, it is desirable to use the available true outcomes to help answer a scientific question, for example, to evaluate the effect on outcome of a set of covariates. Because the surrogate and true outcomes are correlated, the relationship between the surrogate outcome and covariates is informative and may be used to strengthen the inference regarding the true outcome. Following Pepe, Reilly & Fleming (1994), henceforth we call the surrogate outcome the auxiliary outcome to emphasize that we are using the surrogate outcome as auxiliary information.

Let Y, S and X denote the true outcome, the auxiliary outcome and the covariates respectively. Y is a scalar and S can be a vector. The true relationship between (Y, S, X) may be complex and there may be insufficient information to specify a joint likelihood. We assume that the main scientific goal is to assess the regression relationship between Y and X specified by the conditional mean model

$$\mathrm{E}(Y \mid X) = \mu(X; \theta), \tag{1}$$

where μ is a known function and θ is a p-dimensional parameter of interest. Since the joint distribution of (Y, S, X) is otherwise unspecified, the nuisance parameter $\eta = (\eta_1, \eta_2, \eta_3)$ has three (possibly infinite-dimensional) components: $\eta_1 = \mathrm{P}(S \mid Y, X)$, the conditional distribution of the auxiliary; $\eta_2 = \mathrm{P}(\varepsilon \mid X)$, the mean zero conditional distribution of the residual $\varepsilon = Y - \mu(X; \theta)$; and $\eta_3 = \mathrm{P}(X)$, the marginal distribution of the covariates. Let R be a $0/1$ random variable with $R = 1$ indicating that Y is observed, otherwise $R = 0$. S and X are always observed. We assume that subjects are randomly sampled prospectively from an infinite population, and that the subset with $R = 1$ is a random subsample, usually called the validation subsample, selected using variable probability sampling (Lawless, Kalbfleisch & Wild 1999). Thus, the observations $\{Z_i = (R_i, R_i Y_i, S_i, X_i),\ i = 1, \ldots, n\}$ are identically and independently distributed. Let $\pi = \mathrm{P}(R = 1 \mid Y, S, X)$ be the probability of observing Y. We assume that $\pi = \pi(S, X)$, i.e., the true outcome is missing at random (MAR) as described in Little & Rubin (1987), and that there exists a $\sigma > 0$ so that $\pi(S, X) \geq \sigma$.

This article discusses semiparametric efficient estimation of θ. We start in Section 2 with a derivation of the efficient score function (Bickel, Klaassen, Ritov & Wellner 1993), which we denote as ℓ^*_θ. This also can be obtained as a special case of work by Holcroft, Rotnitzky & Robins (1997) and Rotnitzky & Robins (1995b). We believe our result is worth presenting separately, however, for two reasons. First, our derivation takes advantage of a connection between semiparametric efficient estimation and optimal estimating function theory which is both conceptually interesting and broadly applicable to problems of missing data. Second, for the auxiliary outcome problem, ℓ^*_θ may be obtained in simple closed form. In Section 3 we show that, when S and X are discrete, a Horvitz–Thompson estimator with empirically estimated weights is semiparametric efficient. We connect our approach with a class of estimators motivated by the mean-score method of Pepe, Reilly & Fleming (1994), showing that a semiparametric efficient estimator (SEE) may also be obtained by optimizing a class of Horvitz–Thompson estimators with estimated weights. Section 4 presents a data example and some simulation results evaluating the finite-sample performance of the SEE and its asymptotic variance estimator. Some discussion is provided in Section 5.

2. EFFICIENT SCORE FUNCTION FOR θ

A primary goal of semiparametric efficient estimation is to determine the efficient score function ℓ^*_θ or, equivalently, the efficient influence function $\tilde{\ell}_\theta = (\mathrm{E}\ell^*_\theta \ell^{*\top}_\theta)^{-1}\ell^*_\theta$. Estimators obtained as a consistent solution to estimated efficient score equations, or as "one-step"approximations to the solution starting from a consistent estimator, have an asymptotic normal distribution whose asymptotic variance attains the semiparametric efficiency bound $(\mathrm{E}\ell^*_\theta \ell^{*\top}_\theta)^{-1}$ (Bickel, Klaassen, Ritov & Wellner 1993). This is the minimum possible variance for regular, asymptotically linear (RAL) estimators; those that attain it are semiparametric efficient. (We restrict ourselves to RAL

estimators in this discussion rather than the broader class of regular estimators as discussed by Hajek and Lecam). Generally, ℓ_θ^* is defined to be the usual parametric score function for θ, with the nuisance parameters η fixed, minus its projection onto the "nuisance tangent space," the closed linear span of scores for one-dimensional parametric submodels for η passing through the true model. The efficient scores are often calculated by finding this orthogonal projection, either from first principles or by using operator theory applied to the "score operator" (Bickel, Klaassen, Ritov & Wellner 1993, §3.4; van der Vaart 1998, §25.4–§25.5).

An alternative approach is to first characterize the class of influence functions for all RAL estimators. Suppose there is a parametrization of the set $\mathcal{G} = \{G(h; Z, \theta) : h \in \mathcal{H}\}$ of all such influence functions in terms of θ and a class $\mathcal{H}$ of functions $h = h(Z)$. Suppose further that, for each fixed h and for θ in a neighborhood of the true value, $\mathrm{E}_\theta G(h; Z, \theta) = 0$. Then $\mathcal{G}$ is a set of generalized estimating functions indexed by $\mathcal{H}$ (see van der Vaart 1998, §25.9). Furthermore, it is the largest such set. If we can find h^* in $\mathcal{H}$ such that $G(h^*; Z, \theta)$ has minimum variance among all influence functions, then we will have found the efficient influence function and hence the efficient score. The theory of optimal estimating functions, as in Godambe (1960), Godambe & Heyde (1987) and Heyde (1988, 1997), is useful in finding h^*. A somewhat more general formulation is the following result of Newey & McFadden in Chapter 36 of Engle and McFadden (1994, Th. 5.3, page 2166), restated using our notation in the current context:

THEOREM 1. *Suppose $\mathcal{G} = \{G(h; Z, \theta) : h \in \mathcal{H}\}$ is a set of estimating functions indexed by a class $\mathcal{H}$ of functions $h = h(Z)$, such that the asymptotic variance of the estimator $\hat{\theta}_h$ that solves the equation $\sum_{i=1}^n G(h; Z_i, \theta) = 0$ can be expressed as $V(h) = D(h)^{-1}\mathrm{E}[m(h; Z)m(h; Z)^T][D(h)^{-1}]^T$, where D and m are functions of h and of h and Z, respectively. If there is an $h^* \in \mathcal{H}$ that satisfies*

$$D(h) = \mathrm{E}\{m(h; Z)m(h^*; Z)^T\} \text{ for all } h \in \mathcal{H}, \tag{2}$$

then $V(h) = V(h^*) + \mathrm{E}(UU^T)$ where $U = D^{-1}(h)m(h) - D^{-1}(h^*)m(h^*)$, so that $V(h^*)$ is the minimum variance.

Very often $D = -\mathrm{E}(\dot{G})$, where $\dot{f}$ denotes partial differentiation of f with respect to θ, and m is G itself, both evaluated at the true θ. In this case, $G(h^*; Z, \theta)$ identified by Theorem 1 is Godambe's "quasi-score". Of course, one must find an estimator based on the data alone that has the standardized version $[-\mathrm{E}(\dot{G})]^{-1}G$ of this optimal $G(h^*)$ as its influence function. This is often achieved by replacing h in the estimating equation in the theorem by a consistent estimate $\hat{h}^*$.

For a first application of Theorem 1, consider the "complete data problem" where (Y, X) is known for all subjects. Under standard regularity assumptions, the class of influence functions for all RAL estimators of θ in the semiparametric model (1) is the collection of functions of the form $\mathrm{E}\big\{h(X)\dot{\mu}^\top(X; \theta)\big\}^{-1}h(X)\varepsilon$, where h takes values in $\mathbb{R}^p$ (Chamberlain 1987; van der Vaart 1998, §§25.28, 25.66). Here it is further assumed that $h \in \mathcal{H}$ for which both $\mathrm{E}(h\dot{\mu}^\top)$ and $\mathrm{E}(hh^\top\varepsilon^2)$ exist and are nonsingular. Since their elements differ only by multiplication by a nonsingular constant matrix, this class of estimating functions is identical with the class $\mathcal{G} = \big\{G(h; Z) = h(X)\varepsilon : h \in \mathcal{H}\big\}$ that satisfies the conditions of the theorem with $D = -\mathrm{E}(\dot{G}) = \mathrm{E}\big\{h(X)\dot{\mu}^\top(X; \theta)\big\}$ and $m = G$. Applying criterion (2) leads immediately to $\mathrm{E}\big\{h(\dot{\mu}^\top - h^{*\top}\varepsilon^2)\big\} = 0$ for all $h \in \mathcal{H}$ and the conclusion $h^*(X) = \dot{\mu}(X; \theta)/\operatorname{var}(Y \mid X)$. The efficient score and the efficiency bound for the complete data problem are thus $G_C^* = h^*\varepsilon$ and

$$\{\operatorname{var}(G_C^*)\}^{-1} = \big[\mathrm{E}\big\{\dot{\mu}(X; \theta)\dot{\mu}^\top(X; \theta)/\operatorname{var}(Y \mid X)\big\}\big]^{-1},$$

respectively (Chamberlain 1987; van der Vaart 1998, §25.58).

Having characterized the set of influence functions for the complete data problem, results of Robins, Rotnitzky & Zhao (1994, Propositions 8.1(c2) and 8.3) and van der Vaart (1998,

Lemma 25.41 and Example 25.43) yield a similar characterization of influence functions for the corresponding missing-at-random problem. When π is known, the influence function of an arbitrary RAL estimator can be expressed in the form

$$G_1(h,\phi) = \frac{R}{\pi}\, h(X)\varepsilon - \frac{R-\pi}{\pi}\, \phi(S,X), \tag{3}$$

where ϕ is an arbitrary function of S and X and $h \in \mathcal{H}$. For fixed h, the choice of ϕ for which G_1 has minimum variance is $\phi(S,X) = h(X)\mathrm{E}(\varepsilon \mid S,X)$. When π is unknown, the influence functions of all RAL estimators of θ are those $G_1(h,\phi)$ which satisfy the additional restriction

$$\mathrm{E}\left\{ G_1(h,\phi)\, \frac{R-\pi}{\pi}\, d(S,X) \right\} = 0,$$

where $d(S,X)$ is an arbitrary real function of S and X. This follows from the fact that functions of the form $(R-\pi)d(S,X)/\pi$ constitute the score functions from one-dimensional parametric submodels of R (Robins, Rotnitzky & Zhao 1994; van der Vaart 1998, §25.43). Simple calculation shows that this restriction also leads to $\phi = h(x)\mathrm{E}(\varepsilon \mid S,X)$. Thus the influence function of any RAL estimator for θ when π is unknown, or of the optimal RAL estimator when π is known, takes the form

$$G(h;Z) = \frac{R}{\pi}\, h(X)\varepsilon - \frac{R-\pi}{\pi}\, h(X)\mathrm{E}(\varepsilon \mid S,X), \tag{4}$$

where $h\varepsilon$ is an influence function for the complete data problem. Expanding the class to include all $h \in \mathcal{H}$, we apply Theorem 1 in the same manner as before to determine the optimal influence function by identifying that h for which the corresponding G has minimum variance in the class of functions (4). The detailed calculation is in the Appendix.

THEOREM 2. *The efficient score function for θ is*

$$G^* = G(h^*,\phi^*) = h^*(X)\left\{ \frac{R}{\pi}\, Y - \frac{R-\pi}{\pi}\, \mathrm{E}(Y \mid S,X) - \mu(X;\theta) \right\} \equiv h^*(X)\varepsilon^*$$

with $h^*(X) = \dot{\mu}(X;\theta)\,\mathrm{var}^{-1}(\varepsilon^* \mid X)$ and $\phi^* = h^*(X)\mathrm{E}(\varepsilon \mid S,X)$, regardless of whether π is known or not. The semiparametric efficiency bound is $\mathrm{var}^{-1} G^*$ where

$$\mathrm{var}\, G^* = \mathrm{E}\left[\dot{\mu}(X;\theta) \left\{ \mathrm{var}(\varepsilon \mid X) + \mathrm{E}\left(\frac{1-\pi}{\pi}\, \mathrm{var}(\varepsilon \mid S,X) \;\middle|\; X \right) \right\}^{-1} \dot{\mu}^{\top}(X;\theta) \right]. \tag{5}$$

It is not surprising that G^* is the same whether π is known or not. Since Y is missing at random, the likelihood of R factorizes from the likelihood of (Y,S,X), whence the parameters π and (θ,η) are orthogonal (Little & Rubin 1987; Robins, Rotnitzky & Zhao 1994). When $\pi = 1$, the expression for the information bound in Theorem 2 reduces to that for the complete data case derived above. Otherwise the second term inside the inner braces in (5), $A(X) \equiv \mathrm{E}\{(1-\pi)\,\mathrm{var}(\varepsilon \mid S,X)/\pi \mid X\}$, quantifies the efficiency loss due to failure to observe the true outcomes for the nonvalidated subset. It may be used to evaluate the amount of information in S or to compare the quality of different auxiliary outcomes. When S is a perfect outcome, i.e., $S = Y$, $\mathrm{var}(\varepsilon \mid S,X) = 0$ so that $A(X) = 0$. When S is independent of Y given X, $\mathrm{var}(\varepsilon \mid S,X) = \mathrm{var}(\varepsilon \mid X)$, thus $A(X) = \mathrm{var}(\varepsilon \mid X)\mathrm{E}\{(1-\pi)/\pi \mid X\}$ and

$$\mathrm{var}\, G^* = \mathrm{E}\{\dot{\mu}(X;\theta)\mathrm{E}^{-1}(1/\pi \mid X)\,\mathrm{var}^{-1}(\varepsilon \mid X)\dot{\mu}^{\top}(X;\theta)\}.$$

The fact that $\mathrm{var}^{-1} G^*$ is the semiparametric efficiency bound in this case can be easily shown as follows. Observe $\mathrm{E}(\varepsilon \mid S,X) = \mathrm{E}(\varepsilon \mid X) = 0$, so that class (4) reduces to $G(h;Z) = Rh(X)\varepsilon/\pi$. A straightforward application of Theorem 1 leads to

$$h^*(X) = \dot{\mu}(X;\theta)\mathrm{E}^{-1}(\varepsilon^2/\pi \mid X) = \dot{\mu}(X;\theta)\mathrm{E}^{-1}(1/\pi \mid X)\,\mathrm{var}^{-1}(\varepsilon \mid X).$$

As another special case (Tenenbein 1970), suppose we are interested in estimating disease prevalence $\mathrm{P}(Y = 1)$ in the absence of covariates, i.e., Y is binary and $X \equiv 1$. A simple random sample of size n is taken to measure Y while a binary S is available for all N subjects. Straightforward calculation with the aid of Lemma 2(a) in Tenenbein (1970) leads to $A(X) = 1 - K$ where K, the square of the correlation coefficient between Y and S, was termed the "reliability" of S. Substitution of n/N for π in (5), and division by the sample size N, leads to the asymptotic variance for $\widehat{\mathrm{P}}(Y = 1)$ presented in formula (5.3) of Tenenbein (1970).

3. SEMIPARAMETRIC EFFICIENT ESTIMATION WHEN S AND X ARE DISCRETE

Since G^* involves the unknown nuisance quantities $\mathrm{var}(\varepsilon^* \mid X)$ and $\mathrm{E}(Y \mid S, X)$, we estimate θ using estimated score equations

$$\sum_{i=1}^{n} \widehat{G}_i^* = \sum_{i=1}^{n} \hat{h}^*(X_i) \left\{ \frac{R}{\pi_i} Y_i - \frac{R_i - \pi_i}{\pi_i} \widehat{\mathrm{E}}(Y \mid S_i, X_i) - \mu(X_i; \theta) \right\} = 0. \tag{6}$$

Here $h^*(X_i)$ and $\mathrm{E}(Y \mid S_i, X_i)$ in G^* are replaced by empirical estimates, which is feasible when S and X are discrete and of moderate dimensionality. Due to MAR, $\mathrm{E}(Y \mid S, X)$ can be consistently estimated from sample averages in a stratified validation subsample,

$$\widehat{\mathrm{E}}(Y \mid S, X) = \sum_{i=1}^{n} R_i Y_i I(S_i = S, X_i = X) \Big/ \sum_{i=1}^{n} R_i I(S_i = S, X_i = X).$$

For given θ, denote

$$\hat{\varepsilon}_i^* = \frac{R_i}{\pi_i} Y_i - \frac{R_i - \pi_i}{\pi_i} \widehat{\mathrm{E}}(Y \mid S_i, X_i) - \mu(X_i; \theta),$$

then $\mathrm{var}(\varepsilon^* \mid X)$ can be consistently estimated using

$$\widehat{\mathrm{var}}(\varepsilon^* \mid X) = \sum_{i=1}^{n} I(X_i = X) \hat{\varepsilon}_i^{*2} \Big/ \sum_{i=1}^{n} I(X_i = X).$$

Note that $\widehat{\mathrm{E}}(Y \mid S, X)$ is the predicted value from the regression of Y on S and X under a saturated model, and that $\widehat{\mathrm{var}}(\varepsilon^* \mid X)$ is likewise the predicted value from the saturated regression of ε^{*2} on X. In other words, when S and X are discrete, we can always specify correct models for the unknown functions.

Theorem 3 below states the asymptotic properties of $\hat{\theta}$. Theorem 4 notes that $\hat{\theta}$ may be obtained as a Horvitz–Thompson estimator with estimated weights. The proofs of these theorems are sketched in the Appendix.

THEOREM 3. *There exists a unique consistent solution $\hat{\theta}$ to equation* (6). *The asymptotic variance of $\hat{\theta}$ attains the semiparametric efficiency bound, which can be consistently estimated by*

$$\left(\sum_{i=1}^{n} \widehat{G}_i^* \widehat{G}_i^{*\top} \right)^{-1}.$$

THEOREM 4. *Equation* (6) *can be written as*

$$\sum_{i=1}^{n} \frac{R_i}{\hat{\pi}_i} \hat{h}^*(X_i) \varepsilon_i = 0 \tag{7}$$

with

$$\hat{\pi}_i = \frac{\sum_{j=1}^{n} I(R_j = 1, X_j = X_i, S_j = S_i)}{\sum_{j=1}^{n} I(X_j = X_i, S_j = S_i)}.$$

When X is multinomial and written as a vector of 0/1 indicators that index, for example, all possible combinations of levels of discrete covariates, any function of X including $\hat{h}^*(X)$ can be written as CX, where C is a constant matrix. Thus, when S is also discrete, $\hat{\theta}$ from the estimating equation

$$\sum_{i=1}^{n} \frac{R_i}{\hat{\pi}_i} X_i \varepsilon_i = 0, \tag{8}$$

with $\hat{\pi}$ as in Theorem 4, asymptotically achieves the semiparametric efficiency bound.

Returning to the problem of estimating disease prevalence considered at the end of Section 2, Tenenbein (1970, formula 4.1) derived the nonparametric maximum likelihood estimator of prevalence when Y was measured on a simple random sample. An easy calculation shows that his estimator equals that based on equation (8).

Equation (7) provides a connection between the SEE and the mean-score estimator of Pepe, Reilly & Fleming (1994). Suppose for now that the relationship between Y and X is specified by a fully parametric model $\mathrm{P}_\theta(Y \mid X)$ and let $\dot{\ell}_\theta = \dot{\mathrm{P}}_\theta(Y \mid X)/\mathrm{P}_\theta(Y \mid X)$ be the likelihood score. The mean score method for estimating θ replaces $\mathrm{E}(\dot{\ell}_\theta \mid S, X)$ by $\widehat{\mathrm{E}}(\dot{\ell}_\theta \mid S, X)$ in the observed data score function

$$\sum_{i=1}^{n} R_i \dot{\ell}_\theta(Y_i \mid X_i) + (1 - R_i)\mathrm{E}\big\{\dot{\ell}_\theta(Y \mid X_i) \mid S_i, X_i\big\}$$

and uses the result as an estimating function of θ. For the auxiliary outcome problem, substituting any unbiased estimating function of the form $h(X)\varepsilon = h(X)\{Y - \mu(X;\theta)\}$ for $\dot{\ell}_\theta$ leads, owing to the missing-at-random assumption, to a class of unbiased estimating functions

$$\sum_{i=1}^{n} R_i h(X_i)\varepsilon_i + (1 - R_i)\mathrm{E}\big\{h(X)\varepsilon \mid S_i,\ X_i\big\} = 0.$$

When $\mathrm{E}(Y \mid S, X)$ is replaced by the empirical estimate $\widehat{\mathrm{E}}(Y \mid S, X, R = 1)$ for S and X discrete, the resulting estimating equation may be shown to equal

$$\sum_{i=1}^{n} \frac{R_i}{\hat{\pi}_i} h(X_i)\varepsilon_i = 0, \tag{9}$$

with $\hat{\pi}$ as in Theorem 4. Denote the resulting estimator as $\hat{\theta}_{eh}$, which, for convenience, we still call the mean-score estimator. Denote the one solving the same equation but with known weights π_i as $\hat{\theta}_h$, which is the classical Horvitz–Thompson estimator.

A study of the asymptotic variance of $\hat{\theta}_{eh}$ provides insight into the efficiency superiority of the SEE. Note that $\hat{\pi}_i$ is the maximum likelihood estimate from a saturated regression model for π. Consequently, for given $h(X)$, a straightforward application of a result in Pierce (1982) gives that

$$\mathrm{Avar}\,\hat{\theta}_{eh} = A_{11}^{-1}\,\mathrm{var}\left\{\frac{Rh(X)\varepsilon}{\pi} - \frac{R - \pi}{\pi} h(X)\mathrm{E}(\varepsilon \mid S, X)\right\} A_{11}^{-1}, \tag{10}$$

where $A_{11} = \mathrm{E}\{h(X)\dot{\mu}^\top(X;\theta)\}$. Following exactly the same arguments as the proof of Theorem 2 in the Appendix, $h^*(X) = \dot{\mu}(X;\theta)\,\mathrm{var}^{-1}(\varepsilon^* \mid X)$ leads to the smallest $\mathrm{Avar}\,\hat{\theta}_{eh}$, with $\hat{\theta}_{eh^*}$ being the SEE. In the case of logistic regression, the original mean-score estimator of Pepe, Reilly and Fleming uses $h(X) = X$ whereas the SEE uses $h(X) = X\,\mathrm{var}(\varepsilon \mid X)/\,\mathrm{var}(\varepsilon^* \mid X)$.

By minimizing certain elements of $\mathrm{var}^{-1}\,G^*$, optimal sampling fractions at each (S, X) level could be calculated under the restriction of a fixed overall validation proportion or fixed budget for purposes of study design. Unfortunately, no closed-form solution for the optimal π exists. They are defined only by a system of equations of the same dimension as (S, X).

One could assume a parametric model for the joint distribution of (Y, S, X), for example, and then determine by numerical means the optimal sampling fractions for estimation of a specific regression coefficient at various values for the parameters.

4. EXAMPLES AND SIMULATION STUDIES

Clayton, Spiegelhalter, Dunn & Pickles (1998) simulated a study motivated by the Medical Research Council Multicenter Cognitive Function and Aging Study where the outcome was dementia (MRC-CFAS 1998). The study aimed to estimate both prevalence, the probability of having dementia at the first of two visits, and incidence, the probability of having dementia at the second visit for persons without dementia initially. Here we consider only prevalence. While cognitive function was measured on the entire cohort using the Mini-Mental State Examination (MMSE), dementia was diagnosed using a "gold standard"instrument on a subset of patients selected by stratified random sampling according to the MMSE at three levels and age at two.

TABLE 1: Study design and sampling fractions.

		Number of Subjects		Sampling
Age(years)	MMSE	Study cohort	Subsample	fraction
65–74	0–21	291	291	1
	22–25	950	220	0.232
	26–30	3,759	386	0.103
75+	0–21	1,037	496	0.478
	22–25	1,486	208	0.140
	26–30	2,477	179	0.072
Total		10,000	1,780	0.178

For each of the six strata, and using the first of five simulated data sets kindly provided by the authors, Table 1 shows the number of subjects in the study cohort, the number in the validation subsample and the corresponding sampling fraction. Readers are referred to the original paper for the detailed study design. We noted substantial differences between three of the observed sampling proportions shown in Table 1 and the theoretical proportions apparently employed in the simulation (see Table 2 of Clayton, Spiegelhalter, Dunn & Pickles 1998). Thus, for illustrative purposes, the six observed sampling proportions in Table 1 determined the true sampling weights. Data are presented in Table 2, where, within each age-sex-MMSE stratum, n_1^v (n_0^v) denotes to the number of diseased (nondiseased) in the validation subsample and n^{nv} denotes to the number in the nonvalidated subsample.

We fitted a logistic regression model for prevalence with the main effects for age in six groups and sex, and with the three categories of MMSE continuing to serve as the discrete auxiliary outcome. Thus 36 estimated or empirical weights were generated by the saturated model for π. The original mean-score estimator and the Horvitz–Thompson estimator were both obtained from equation (9) with $h(X) = X$, the former with estimated and the latter with true weights. The SEE was obtained by solving equation (7) with $\hat{h}^*(X)\varepsilon = X\{\mathrm{var}(\varepsilon \mid X)/\widehat{\mathrm{var}}(\varepsilon^* \mid X)\}\varepsilon$. The standard errors were calculated based on the asymptotic variance formulas for each estimator. The results shown in Table 3 also include those from a naive logistic regression analysis of just the data in the validation subsample, without consideration of the stratified sampling design.

For the estimation of the sex coefficient, the SEE method led to a 14.3% reduction in the variance compared with the mean-score, and both were more precise than the Horvitz–Thompson

estimation (35.7% and 25.0% reduction in variance, respectively). For the estimation of the age stratum coefficients, although the mean-score improved on the Horvitz–Thompson estimation, the SEE did not. Complete case analysis using the validation subsample alone yielded apparently biased estimates. Robins & Wang (1998) demonstrated more substantial gains in efficiency of SEE over the Horvitz–Thompson approach for a more complicated investigation of the effects of age and sex on the incidence of dementia that involved a nonmonotone missingness pattern.

TABLE 2: Subject distribution by age-sex-MMSE in the validated and nonvalidated subsamples.

		MMSE								
		0–21			22–25			26–30		
	Age(years)	n_1^v	n_0^v	n^{nv}	n_1^v	n_0^v	n^{nv}	n_1^v	n_0^v	n^{nv}
Males	65–69	12	38	0	1	35	137	0	101	876
	70–74	22	52	0	1	30	141	0	74	680
	75–79	13	25	53	3	28	198	0	52	528
	80–84	23	18	30	1	34	153	1	17	323
	85–89	21	17	37	1	6	35	0	6	77
	90+	12	10	23	0	2	23	0	3	27
Females	65–69	9	60	0	1	54	207	0	107	962
	70–74	28	70	0	5	93	245	1	103	855
	75–79	24	56	87	6	49	349	0	59	770
	80–84	79	56	145	3	48	330	0	30	409
	85–89	41	41	100	3	19	147	0	10	139
	90+	37	23	66	0	5	43	0	1	25

We performed simulation studies (i) to compare the relative efficiency of SEE versus the mean-score method; (ii) to investigate the performance of $(\sum_i \widehat{G}_i^* \widehat{G}_i^{*\top})^{-1}$ as the asymptotic variance estimator of the SEE in Theorem 3; (iii) to investigate (10) as an alternative variance estimator to that proposed by Pepe, Reilly & Fleming (1994) for the mean-score estimator; and (iv) to investigate how informative the auxiliary outcome must be to increase precision substantially. We assumed that X took three values $1/2/3$ with probabilities $0.4/0.5/0.1$, respectively. Y was binary generated from the logistic regression model

$$\log\{\mathrm{P}(Y=1\,|\,X)/\mathrm{P}(Y=0\,|\,X)\} = \theta_0 + \theta_1 X$$

with $\theta_0 = -2.5$. Two values for θ_1 were used, namely $0.0/1.0$. S was a nondifferentially misclassified version of Y: $\mathrm{P}(S=1\,|\,Y,X) = \mathrm{P}(S=1\,|\,Y)$, and we set sensitivity to $\mathrm{P}(S=1\,|\,Y=1) = 0.8$. Three values for the specificity $\mathrm{P}(S=0\,|\,Y=0)$, 0.55/0.75/0.90, were used, corresponding to whether the auxiliary outcome was weakly/moderately/strongly informative. We sampled S and X for 5000 subjects, and then sampled Y for subsets of size n^v=250/500/750 following a "balanced design" (Breslow & Cain 1988) in which the number of the second-phase subjects in cells defined by S and X were approximately equal. Each scenario was simulated 1000 times. The results are listed in Tables 4 and 5 corresponding to the two values of θ_1. "Evar" refers to the empirical variance, "$\overline{\mathrm{est(var)}}$" to the averaged estimated variance using the asymptotic variance formula, $\overline{\hat{\theta}_1}$ to the average of the estimators $\hat{\theta}_1$, and "CP" to the 90% coverage probability. The efficiency gain of the SEE over the mean-score estimator, obtained by subtracting one from the ratio of empirical variances for the mean-score and SEE, is shown in Table 6.

Both the SEE and the mean-score estimators were reasonably unbiased, whereas the naive complete data estimator was not. For SEE and mean-score, the empirical variances were usually close to the average estimated variances and the coverage probabilities were reasonably close to the nominal level. Exceptions occurred with the smallest validation samples, where standard errors were underestimated and coverage probabilities were below nominal level. In all scenarios, the SEE performed better than the mean-score method and sometimes resulted in very important efficiency gains. In other simulations (not reported here), we observed little advantage in the efficiency of the SEE over the mean-score method.

TABLE 3: Estimates and standard errors for age and sex effects.

	Age(years)					
	70–74	75–79	80–84	85–89	90+	Sex
Complete: validation sample only						
$\hat{\theta}$	0.833	1.070	2.189	2.424	2.936	0.099
sd	0.253	0.264	0.242	0.264	0.296	0.136
Horvitz–Thompson estimation (true weights)						
$\hat{\theta}$	1.086	1.658	2.711	3.201	3.883	0.331
sd	0.335	0.341	0.313	0.334	0.383	0.187
Mean score method (estimated weights)						
$\hat{\theta}$	1.081	1.732	2.648	3.182	3.742	0.283
sd	0.326	0.320	0.300	0.308	0.310	0.162
Semiparametric efficient estimation						
$\hat{\theta}$	1.152	1.810	2.766	3.275	3.809	0.289
sd	0.336	0.331	0.305	0.319	0.322	0.150

5. DISCUSSION

We studied semiparametric efficient estimation for regression parameters when the true outcome is missing at random but auxiliary outcomes are available for all subjects. Optimal estimating function theory was used to derive the efficient score function, which takes a simple, explicit form for this problem. When the auxiliary outcome and covariates are both discrete, we demonstrated that the optimality of the SEE is achieved partly by efficiently estimating the weights in the Horvitz–Thompson estimator and partly by minimizing the asymptotic variance of a class of Horvitz–Thompson estimators with efficiently estimated weights.

Our simulation studies suggested that the SEE can result in important efficiency gains over the original mean-score estimator of Pepe, Reilly & Fleming (1994), that the inverse of the estimated vector product of the efficient score function performs well as an estimator of the semiparametric efficiency bound, and that (10) performs well as an alternative asymptotic variance estimator for the mean-score method. Very often, limited resources only allow a small validation sample. Then, even moderately informative auxiliary outcomes can result in a meaningful efficiency gain. Standard methods for model checking apply. For example, one can group the X data and fit preliminary models with separate regression parameters for each group, or add X terms to check for linearity and the absence of interactions.

TABLE 4: Simulation results ($\theta_1 = 0$).

	$n^v = 250$			$n^v = 500$			$n^v = 750$		
Specificity	0.55	0.75	0.90	0.55	0.75	0.90	0.55	0.75	0.90
				Complete: validation sample only					
$\overline{\hat{\theta}_1}$	0.005	0.009	0.014	0.007	0.005	-0.031	-0.001	0.008	-0.133
Evar	0.094	0.060	0.038	0.048	0.029	0.018	0.027	0.020	0.013
est(var)	0.091	0.062	0.037	0.044	0.030	0.019	0.029	0.020	0.015
90% CP	0.891	0.901	0.894	0.889	0.906	0.902	0.915	0.899	0.717
				Semiparametric efficient estimation					
$\overline{\hat{\theta}_1}$	-0.003	0.007	0.003	0.006	0.003	-0.006	-0.002	0.006	0.008
Evar	0.129	0.077	0.052	0.051	0.034	0.026	0.029	0.022	0.019
est(var)	0.100	0.066	0.046	0.044	0.033	0.025	0.029	0.022	0.019
90% CP	0.859	0.872	0.852	0.871	0.877	0.874	0.899	0.895	0.893
				Mean score method					
$\overline{\hat{\theta}_1}$	0.001	0.017	-0.001	0.008	0.007	-0.010	-0.009	0.004	0.010
Evar	0.136	0.091	0.068	0.064	0.045	0.035	0.039	0.029	0.022
est(var)	0.124	0.085	0.059	0.059	0.043	0.032	0.040	0.029	0.022
90% CP	0.876	0.867	0.839	0.892	0.899	0.874	0.900	0.899	0.892

In the data example we provided, we observed that the mean-score method performed well for estimation of the age stratum effects. In fact, we observed the same phenomenon in many of the simulation studies, especially those involving relatively constant sampling fractions leading to "unbalanced"validation samples. The relative efficiency of the mean-score and SEE methods could be affected by many factors including the sampling scheme, the true regression association and the nuisance aspects of the probability distribution. However, we were unable to reach general conclusions about the circumstances when semiparametric efficient estimation resulted in important efficiency gains, and further work is warranted on this aspect.

The optimal estimating function approach we used for calculating the efficient score function for Euclidean parameters is a general approach applicable to many semiparametric problems. Theorem 1 is particularly useful for the general data-missing-at-random problem considered by Robins, Rotnitzky & Zhao (1994). Once the influence functions for RAL estimators are identified for the complete data problem, those for the problem with a monotone missingness pattern are identified by their equation (38). Application of Theorem 1 in this article, i.e., solving equation (2) for G^*, then leads to a straightforward calculation of the efficient score function.

We focused on the auxiliary outcome problem when both S and X are discrete, thus extending the work of Alonzo, Pepe & Lumley (2003) on various strategies for the estimation of disease prevalence in the absence of covariates. When (S, X) take many discrete values and the validation sample is small, fitting saturated models to $\mathrm{E}(Y \mid S, X)$ and $\mathrm{var}(\varepsilon^* \mid X)$ may not be feasible due to the lack of validation observations at some specific (S, X) values. In this case, it may be helpful to pool adjacent values for purposes of estimating the weights, provided that this still results in a correct model for π. As an example, for the dementia data considered in Section 4, one could pool adjacent age categories within the two broad age groups specified in Table 1. When (S, X) have continuous components, consistent estimation of $\mathrm{E}(Y \mid S, X)$ and $\mathrm{var}(\varepsilon^* \mid X)$ in equation (6) is even more challenging due to the lack of knowledge of their correct functional forms. Semiparametric efficient estimation may still be achieved, however, if they are indeed consistently estimated (van der Vaart 1998, §25.54). One compromise approach

is to specify working parametric models for these functions. If correctly specified, the estimator based on equation (6) is semiparametric efficient. If not, since (6) is still unbiased, the estimator is consistent though not efficient. See Robins, Rotnitzky & Zhao (1994) for full discussion of this aspect. Y. Chen (2000) proposed a robust imputation approach when the validation subsample of the outcome is obtained via simple random sampling. It uses a possibly incorrect "working model" for $\mathrm{E}(Y \mid S, X)$ to impute Y for subjects in the nonvalidated subsample, and is easily implemented regardless of the dimensionality of S and X. J. Chen (2002) extended the approach to the situation where the outcome is missing at random.

TABLE 5: Simulation results ($\theta_1 = 1.0$).

	$n^v = 250$			$n^v = 500$			$n^v = 750$		
Specificity	0.55	0.75	0.90	0.55	0.75	0.90	0.55	0.75	0.90
Complete: validation sample only									
$\overline{\hat{\theta}_1}$	0.887	0.735	0.531	0.884	0.734	0.523	0.880	0.745	0.520
Evar	0.032	0.028	0.024	0.014	0.014	0.012	0.010	0.009	0.007
est(var)	0.032	0.029	0.026	0.016	0.014	0.013	0.011	0.010	0.009
90% CP	0.804	0.509	0.104	0.769	0.286	0.007	0.674	0.150	0.000
Semiparametric efficient estimation									
$\overline{\hat{\theta}_1}$	1.016	1.013	1.018	1.011	1.006	1.006	1.003	1.015	1.003
Evar	0.033	0.024	0.020	0.015	0.014	0.011	0.011	0.009	0.007
est(var)	0.031	0.024	0.018	0.016	0.012	0.010	0.011	0.008	0.007
90% CP	0.889	0.889	0.863	0.901	0.893	0.873	0.884	0.893	0.913
Mean score method									
$\overline{\hat{\theta}_1}$	1.012	1.003	1.013	1.007	1.005	1.005	1.000	1.014	1.002
Evar	0.040	0.032	0.025	0.020	0.018	0.013	0.014	0.011	0.008
est(var)	0.042	0.031	0.023	0.021	0.016	0.012	0.014	0.011	0.009
90% CP	0.910	0.895	0.862	0.909	0.885	0.890	0.900	0.891	0.915

TABLE 6: Efficiency gain of the SEE over mean-score (%).

	Specificity	$\theta_1 = 0$	$\theta_1 = 1.0$
	0.55	5.4	21.2
$n^v = 250$	0.75	18.1	33.3
	0.90	30.7	25.0
	0.55	25.5	33.3
$n^v = 500$	0.75	32.4	28.6
	0.90	34.6	18.2
	0.55	34.5	27.3
$n^v = 750$	0.75	31.8	22.2
	0.90	15.8	14.2

APPENDIX: PROOFS FOR THEOREMS

Proof of Theorem 2. We apply Theorem 1 to the class of estimating functions G_1 in equation (3) indexed by h and ϕ. With $Q = h(X)\varepsilon$ and $Q^* = h^*(X)\varepsilon$, let $\dot{\ell}_\theta$ be the likelihood score for the complete data problem, and $U^m = (R, RY, S, X)$ be the observed data. Then the likelihood score for the auxiliary outcome problem is $\dot{\ell}^m_\theta = \mathrm{E}(\dot{\ell}_\theta \mid U^m)$ (van der Vaart 1998, Th. 25.40), so that $\mathrm{E}\{G(\dot{\ell}^m_\theta)^\top\} = \mathrm{E}\{h(X)\dot{\mu}^\top(X;\theta)\}$. Further,

$$\mathrm{E}GG^{*\top} = \mathrm{E}\left(\frac{1}{\pi}\, QQ^{*\top}\right) - \mathrm{E}\left(\frac{1-\pi}{\pi}\, Q\phi^{*\top}\right) - \mathrm{E}\left[\frac{1-\pi}{\pi}\, \phi\{\mathrm{E}(Q^* \mid S, X) - \phi^*\}^\top\right].$$

By Theorem 1, we search for h^* and ϕ^* such that for all (h, ϕ), $\mathrm{E}\{G(\dot{\ell}^m_\theta)^\top\} = \mathrm{E}(GG^{*\top})$. Since ϕ is arbitrary, we have $\mathrm{E}[(1-\pi)\,\phi\{\mathrm{E}(Q^* \mid S, X) - \phi^*\}^\top/\pi] = 0$, so that $\phi^* = \mathrm{E}(Q^* \mid S, X) = h^*(X)\mathrm{E}(\varepsilon \mid S, X)$. Substituting ϕ^* back into equation $\mathrm{E}\{G(\dot{\ell}^m_\theta)^\top\} = \mathrm{E}(GG^{*\top})$ leads to

$$\mathrm{E}\left\{\frac{1}{\pi}\, h^*(X)\varepsilon^2 - \frac{1-\pi}{\pi}\, h^*(X)\varepsilon\mathrm{E}(\varepsilon \mid S, X) - \dot{\mu}(X;\theta) \;\Big|\; X\right\} = 0$$

since $h(X)$ is arbitrary. Thus,

$$\begin{aligned} h^*(X) &= \dot{\mu}(X;\theta)\mathrm{E}\left\{\frac{1}{\pi}\,\varepsilon^2 - \frac{1-\pi}{\pi}\,\varepsilon\mathrm{E}(\varepsilon \mid S, X) \;\Big|\; X\right\}^{-1} \\ &= \dot{\mu}(X;\theta)\,\mathrm{var}^{-1}\left\{\frac{R}{\pi}\,\varepsilon - \frac{R-\pi}{\pi}\,\mathrm{E}(\varepsilon \mid S, X) \;\Big|\; X\right\}. \end{aligned}$$

This is the desired result. The fact that $\phi^* = \mathrm{E}(Q^* \mid S, X) = h^*(X)\mathrm{E}(\varepsilon \mid S, X)$ shows that ϕ^* already satisfies (3) and hence G^* is the same if π is completely unknown. When π is specified by a parametric model with an unknown Euclidean parameter α, let $T_\alpha = \{\dot{\pi}(\alpha)\}$ be the set spanned by the score functions of α. The influence functions for all RAL estimators are those $G_1(h, \phi)$ that satisfy $\mathrm{E}\{G_1^\top(h, \phi)\dot{\pi}(\alpha)\} = 0$. Since T_α is a subset of all functions $(R-\pi)d(S, X)/\pi$, this set of influence functions is larger than that when π is completely unknown, and smaller than that when π is completely known. Consequently, the efficient score function must be the same as in the other two situations.

Proof of Theorem 3. Let $e(S, X;\gamma)$ be a saturated model for $\mathrm{E}(Y \mid S, X)$, and $v(X;\tau)$ be a saturated model for $\mathrm{var}(\varepsilon^* \mid X)$. Let $\dot{e}(S, X;\gamma) = \partial e(S, X;\gamma)/\partial\gamma$ and $\dot{v}(X;\tau) = \partial v(X;\tau)/\partial\tau$. Then $\hat{\theta}$ can be obtained from the joint solution to the following three unbiased estimating equations:

$$\begin{aligned} &\sum_{i=1}^n \dot{\mu}(X_i;\theta)v^{-1}(X_i;\tau)\left[\frac{R_i}{\pi_i}\,\varepsilon_i - \frac{R_i-\pi_i}{\pi_i}\{e(S_i, X_i;\gamma) - \mu(X_i;\theta)\}\right] = 0, \\ &\sum_{i=1}^n R_i\dot{e}(S, X;\gamma)\{Y_i - e(S_i, X_i;\gamma)\} = 0, \\ &\sum_{i=1}^n \dot{v}(X_i;\tau)\left[\left[\frac{R_i}{\pi_i}\,\varepsilon_i - \frac{R_i-\pi_i}{\pi_i}\{e(S_i, X_i;\gamma) - \mu(X_i;\theta)\}\right]^2 - v(X_i;\tau)\right] = 0. \end{aligned}$$

Denote $\psi = (\theta, \tau, \gamma)$, and write this set of three estimating equations jointly as $\sum_{i=1}^n U_i(\psi) = 0$. Then existence and uniqueness of a consistent solution $\hat{\psi}$ follow from Foutz (1977) under the regularity conditions: (i) $\partial U_i(\psi)/\partial\psi$ exists and is continuous on an open set Ψ in the Euclidean space of the same dimension as ψ, with $\psi \in \Psi$; (ii) $n^{-1}\sum_{i=1}^n \partial U_i(\psi)/\partial\psi$ is nonsingular with probability going to 1 as n increases; (iii) $n^{-1}\sum_{i=1}^n \partial U_i(\psi)/\partial\psi$ converges to a fixed matrix $M(\psi)$ uniformly in an open neighbourhood of ψ. Asymptotic normality follows by direct Taylor expansion. In particular, $\hat{\theta}$ achieves the semiparametric efficiency bound

since the expected derivative matrix of the joint estimating function with respect to (θ, γ, τ) is block diagonal. Central to this proof is the fact that the saturated models $v(x;\tau)$ and $e(s,x;\gamma)$ are correct models when S and X are discrete.

Proof of Theorem 4. Let V denote the validation subsample and let $V(S_i, X_i)$ denote the collection of subjects in V with $S = S_i$ and $X = X_i$. Let $n(S_i, X_i)$ $[n^v(S_i, X_i)]$ denote the number of subjects in the whole [validation] sample with $S = S_i$ and $X = X_i$. For a general $h(X)$,

$$\begin{aligned}
\sum_{i=1}^{n} \frac{R_i - \pi_i}{\pi_i} h(X_i) \widehat{\mathrm{E}}(\varepsilon \mid S_i, X_i) &= \sum_{i=1}^{n} \frac{R_i - \pi_i}{\pi_i} h(X_i) \frac{\sum_{j \in V(S_i, X_i)} \varepsilon_j}{n^v(S_i, X_i)} \\
&= \sum_{i=1}^{n} \sum_{j \in V(S_i, X_i)} \frac{(R_i - \pi_i) h(X_i) \varepsilon_j}{\pi_i n^v(S_i, X_i)} \\
&= \sum_{j \in V} h(X_j) \varepsilon_j \sum_{i=1}^{n} \frac{(R_i - \pi_i) I(S_j = S_i, X_j = X_i)}{\pi_i n^v(S_i, X_i)} \\
&= \sum_{j \in V} \frac{h(X_j) \varepsilon_j}{\pi_j n^v(S_j, X_j)} \sum_{i=1}^{n} (R_i - \pi_i) I(S_j = S_i, X_j = X_i) \\
&= \sum_{j \in V} h(X_j) \varepsilon_j \frac{n^v(S_j, X_j) - \pi_j n(S_j, X_j)}{n^v(S_j, X_j) \pi_j} \\
&= \sum_{i=1}^{n} \frac{R_i}{\pi_i} h(X_i) \varepsilon_i - \sum_{i=1}^{n} \frac{R_i n(S_i, X_i)}{n^v(S_i, X_i)} h(X_i) \varepsilon_i.
\end{aligned}$$

Substituting this back into equation (6) with $h(X)$ replaced by $\hat{h}^*(x)$, leads to the theorem.

ACKNOWLEDGEMENTS

This work was supported in part by grant 5-R01-CA40644 from the United States Public Health Service. We would like to thank Professor Andrea Rotnitzky, whose discussion with Jinbo Chen helped to improve the manuscript. We are deeply grateful to Professor Margaret S. Pepe for her helpful comments.

REFERENCES

T. A. Alonzo, M. S. Pepe & T. S. Lumley (2003). Estimating disease prevalence in two-phase studies. *Biostatistics (Oxford)*, 4, 313–326.

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

N. E. Breslow & K. C. Cain (1988). Logistic regression for two-stage case-control data. *Biometrika*, 5, 11–20.

G. Chamberlain (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Economics*, 34, 305–334.

J. Chen (2002). *Semiparametric Efficient and Inefficient Estimation for the Auxiliary Outcome Problem with the Conditional Mean Model*. Doctoral dissertation, Department of Biostatistics, University of Washington, Seattle.

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

D. Clayton, D. Spiegelhalter, G. Dunn & A. Pickles (1998). Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society Series B*, 60, 71–87.

E. J. Costelo, A. Angold, B. J. Burns, D. K. Stangl, D. T. Tweed, A. Erkanli & C. M. Worthman (1996). The Great Smoky Mountains Study of Youth: prevalence and correlates of DSM-III-R disorders. *Archives of General Psychiatry*, 53, 1129–1136.

R. F. Engle & D. L. McFadden, eds. (1994). *Handbook of Econometrics, Volume 4*. Elsevier, Amsterdam.

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

V. P. Godambe (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31, 1208–1212.

V. P. Godambe & C. C. Heyde (1987). Quasi-likelihood and optimal estimation. *International Statistical Review*, 55, 231–244.

C. C. Heyde (1988). Fixed sample and asymptotic optimality for classes of estimating functions. *Contemporary Mathematics*, 80, 241–247.

C. C. Heyde (1997). *Quasi-likelihood and its Application: a General Approach to Optimal Parameter Estimation*. Springer-Verlag, New York.

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

J. F. Lawless, J. D. Kalbfleisch & C. J. Wild (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society Series B*, 61, 413–438.

R. J. A. Little & D. B. Rubin (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

MRC-CFAS (1998). Cognitive function and dementia in six areas of England and Wales: The distribution of MMSE and prevalence of GMS organicity level in the MRC-CFA study. *Psychological Medicine*, 28, 319–335.

B. Nan (2001). *Information Bounds and Efficient Estimates for Two-Phase Designs with Life-Time Data*. Doctoral dissertation, Department of Biostatistics, University of Washington, Seattle.

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

D. A. Pierce (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *The Annals of Statistics*, 10, 475–478.

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

J. M. Robins, F. Hsieh & W. Newey (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society Series B*, 57, 409–424.

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

J. M. Robins & N. Wang (1998). Discussion on the papers by Forster and Smith and Clayton et al. *Journal of the Royal Statistical Society Series B*, 60, 91–93.

A. Rotnitzky & J. M. Robins (1995a). Semiparametric estimation of models for means and covariances in the presence of missing data. *Scandinavian Journal of Statistics*, 22, 323–333.

A. Rotnitzky & J. M. Robins (1995b). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82, 805–820.

E. Schisterman & A. Rotnitzky (2001). Estimation of the mean of a k-sample u-statistic with missing outcomes and auxiliaries. *Biometrika*, 88, 713–725.

A. Tenenbein (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association*, 65, 1350–1361.

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

Received 21 August 2003

Accepted 13 May 2004

Jinbo CHEN: chenjin@mail.nih.gov

Biostatistics Branch, Division of Cancer Epidemiology and Genetics
National Cancer Institute, Rockville, MD 20852, USA Rockville, MD 20852, USA

Norman E. BRESLOW: norm@u.washington.edu

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA