Statistica Sinica **16**(2006), 1193-1212

A REVISIT OF SEMIPARAMETRIC REGRESSION MODELS WITH MISSING DATA

Menggang Yu and Bin Nan

Indiana University and University of Michigan

Abstract: The theoretical results in Robins, Rotnitzky and Zhao (1994) and Robins and Rotnitzky (1992) are revisited for semiparametric regression models with missing data. The main results provide a more relevant format for the calculations of efficient score functions. The intuition behind those abstract results and major steps of their proofs are discussed. The surrogate outcome regression problem is studied as a new application. Beyond the derivation of its efficient score function, an estimating method based on the efficient score function is proposed. A set of regularity conditions is given that provides desirable large sample properties for the proposed method.

Key words and phrases: Double robustness, efficient score, projection, missing at random, score function, semiparametric regression models, surrogate outcome, tangent space.

1. Introduction

Improving efficiency of estimators in semiparametric regression models with missing data has been an interesting and active research subject. Robins, Rotnitzky and Zhao (1994) (hereafter RRZ) provided profound calculations of efficient score functions and information bounds using score projection for models with data missing at random (MAR, a terminology of Little and Rubin (1987)). Part of their calculations can also be found in Robins and Rotnitzky (1992) (hereafter RR). Their basic idea is to bridge the model with missing data (observed model) and the corresponding model without missing data (full model) if certain properties of the full model are known or easily obtained. The results are fundamental and can be applied to a variety of regression models, but the proofs are difficult and short on intuition. We think it useful to revisit these important results.

The purpose of this study is to explain RRZ and RR using the score projection method discussed by Bickel, Klaassen, Ritov and Wellner (1993) (hereafter BKRW), and to explicate their results by introducing new regression model: the mean regression with response variable sometimes missing and a surrogate variable always available, termed the surrogate outcome regression model. The desired outcome is that a wider audience of statisticians interested in semiparametric models with missing data will better understand the recent developments in the area, and apply the results in RRZ and RR.

We begin by introducing the general semiparametric model with data MAR, with the surrogate outcome regression model as a special case. In Section 3, we introduce three theorems for the calculation of efficient score functions for data MAR. The most general result is Theorem 3.1 for arbitrary missingness; Theorem 3.2 is a direct consequence of Theorem 3.1 for monotone missingness; Theorem 3.3 is for two-phase sampling designs that are special cases of monotone missingness, and widely used in practice. We provide some insights and intuition behind these theorems, especially Theorem 3.1, to illustrate the ideas and major steps of the proofs. For rigorous proofs, see Yu and Nan (2005). In Section 4, we discuss the surrogate outcome regression problem. We provide detailed derivation of the efficient score function for this model to illustrate the application of the theoretical results in Section 3. Beyond the calculation of the efficient score, we also propose a nonparametric estimating method, and give a set of regularity conditions under which the estimator for the parameter of interest is asymptotically normal. Simulations show that the proposed estimator works well. The method for the surrogate outcome regression is a new addition to the literature.

2. Missing Data Models

In this section, we first discuss missing data models in general, and then introduce the surrogate outcome regression problem as an example of two-phase designs.

2.1. A general model

Suppose the underlying full data are i.i.d. copies of the m-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_m)$. Then we can focus on a single observation for the derivation of the efficient score function. We denote the model for $\boldsymbol{X}$ as $\mathcal{Q} = \{Q_{\theta,\eta}\}$, where $Q_{\theta,\eta}$ is a distribution function, θ is the finite dimensional parameter of interest, and η is an infinite-dimensional nuisance parameter or a vector of several infinite-dimensional nuisance parameters.

Let $\boldsymbol{R} = (R_1, \ldots, R_m)$ be a random vector with $R_j = 1$ if X_j is observed, and $R_j = 0$ if X_j is missing, $j = 1, \ldots, m$. Let $\boldsymbol{r} = (r_1, \ldots, r_m)$ be the realized value of $\boldsymbol{R}$. For some $\boldsymbol{R}$, we observe data $\boldsymbol{X}_{(R)} = (R_1 * X_1, \ldots, R_m * X_m)$, where we define

$$R_j * X_j \equiv \begin{cases} X_j, & R_j = 1, \\ \text{Missing}, & R_j = 0, \end{cases} \qquad j = 1, \ldots, m.$$

Throughout the paper we assume that the data are MAR, i.e.,

$$\pi(\boldsymbol{r}) \equiv P(\boldsymbol{R} = \boldsymbol{r} | \boldsymbol{X}) = P(\boldsymbol{R} = \boldsymbol{r} | \boldsymbol{X}_{(r)}) \equiv \pi(\boldsymbol{r}, \boldsymbol{X}_{(r)}). \tag{2.1}$$

Thus the observed data are i.i.d. copies of $(\boldsymbol{R}, \boldsymbol{X}_{(R)})$. We also assume that the probability of observing full data $\boldsymbol{X}$ is bounded away from zero,

$$\pi(\mathbf{1}_m) \geq \sigma > 0, \tag{2.2}$$

where $\mathbf{1}_m$ is a vector of 1's with length m and σ is a constant. Here $\boldsymbol{R} = \mathbf{1}_m$ means that we observe full data $\boldsymbol{X} \equiv \boldsymbol{X}_{(\mathbf{1}_m)}$. We use $\pi(\boldsymbol{r})$ to denote the conditional probability in (2.1) for ease of notation. It is obvious that

$$\sum_r \pi(\boldsymbol{r}) = 1. \tag{2.3}$$

The missing probability $\pi(\boldsymbol{r})$ can be either known or unknown, in either case the efficient score function for θ has the same form. This becomes clear in (2.4) below, where $\pi(\boldsymbol{r})$ is factored out from other parts of the density. To be general, we denote $\pi(\boldsymbol{r})$ as another parameter. The induced model for the observed data $(\boldsymbol{R}, \boldsymbol{X}_{(R)})$ is denoted as $\mathcal{P} = \{P_{\theta,\eta,\pi}\}$, where $P_{\theta,\eta,\pi}$ is a distribution function with an additional nuisance parameter π.

Let $q_{\theta,\eta}$ be the density function of the probability measure $Q_{\theta,\eta}$, and $p_{\theta,\eta,\pi}$ the density function of the probability measure $P_{\theta,\eta,\pi}$. By the MAR assumption in (2.1), we have:

$$p_{\theta,\eta,\pi}(\boldsymbol{r}, \boldsymbol{x}_{(r)}) = \pi(\boldsymbol{r}) \int q_{\theta,\eta}(\boldsymbol{x}) \prod_{j=1}^{m} \Big(d\mu_j(x_j)\Big)^{1-r_j} , \tag{2.4}$$

where the μ_j are dominating measures for x_j, $j = 1, \ldots, m$.

Our goal is to derive the efficient score function for θ in model $\mathcal{P}$ under different missing patterns: arbitrary missingness, monotone missingness, and a two-phase sampling design where some random variables in $\boldsymbol{X}$ are always observed, and others are either observed or missing simultaneously. Arbitrary missingness means that the pattern of 1's and 0's in vector $\boldsymbol{r}$ is arbitrary. Monotone missingness means that $\boldsymbol{r} \in \{\mathbf{1}_j : \ j = 1, \ldots, m\}$, where

$$\mathbf{1}_j = (\underbrace{1, \ldots, 1}_{j}, \underbrace{0, \ldots, 0}_{m-j}), \quad j = 1, \ldots, m. \tag{2.5}$$

A natural example of monotone missingness is a longitudinal study with dropouts. Sometimes monotone missingness can be obtained by rearranging the order of the random variables $X_1, \ldots, X_m$. For example, if we put all the fully observed random variables in front of the variables with missing data in a two-phase sampling design, then the data structure becomes monotone missing with $\boldsymbol{r} \in \{\mathbf{1}_t,\ \mathbf{1}_m\}$, where t is a fixed integer. It is clear to see that a two-phase sampling design is a simple case of monotone missingness.

2.2. An example: the surrogate outcome regression

It is common in medical research that outcome variables of interest are difficult, or expensive, to ascertain. Surrogate outcome variables (or their correlates), however, can sometimes be readily obtained. Many examples are described in the introductions of Pepe (1992) and Pepe, Reilly and Fleming (1994).

Let Y be an outcome of interest that is not always observable. Let S be a surrogate variable of Y that is always available. The association of Y and a covariate vector $\mathbf{Z}$ is of major interest. Existing methods often require that the conditional density $f_\theta(Y|\mathbf{Z})$ be known up to the finite dimensional parameter θ. However, any misspecification of the conditional density function may cause biased estimates.

Instead of modeling the conditional density function $f_\theta(Y|\mathbf{Z})$ parametrically, we only assume

$$E(Y|\mathbf{Z}) = g(\mathbf{Z};\theta)\ , \tag{2.6}$$

where $g(\cdot\ ;\theta)$ is a known function, $\theta \in \mathrm{R}^d$. Let $\epsilon = Y - g(\mathbf{Z};\theta)$, then

$$E(\epsilon|\mathbf{Z}) = 0\ . \tag{2.7}$$

Under (2.6), the underlying joint density function of $(S, Y, \mathbf{Z})$ can be written as

$$q_{\theta,f_1,f_2,f_3}(s, y, \mathbf{z}) = f_1(s|y,\mathbf{z}) f_2(y - g(\mathbf{z};\theta)|\mathbf{z}) f_3(\mathbf{z})\ , \tag{2.8}$$

where f_1 is the conditional density function of S given $(Y, \mathbf{Z})$, f_2 is the conditional density function of Y, equivalently ϵ, given $\mathbf{Z}$, and f_3 is the density function of $\mathbf{Z}$. Thus the model at (2.6) is semiparametric in the sense that functions f_1, f_2, and f_3 in (2.8) are unspecified, in effect they are infinite dimensional nuisance parameters.

Let R be 1 when Y is observed and 0 otherwise. We assume that Y is missing at random, i.e., $P(R = 1|S, Y, \mathbf{Z}) = P(R = 1|S, \mathbf{Z}) \equiv \pi(S, \mathbf{Z})$. We also assume that $\pi(S, \mathbf{Z}) \geq \sigma > 0$ for some constant σ. Denote the observed data as

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

Then the density function for the observed data $(S, R*Y, \mathbf{Z}, R)$ is

$$p_{\theta,f_1,f_2,f_3}(s, r*y, \mathbf{z}, r) = \Big\{\pi(s,\mathbf{z}) q_{\theta,f_1,f_2,f_3}(s,y,\mathbf{z})\Big\}^r \times\Big\{(1-\pi(s,\mathbf{z}))\int q_{\theta,f_1,f_2,f_3}(s,y,\mathbf{z})d\nu(y)\Big\}^{1-r}, \quad (2.9)$$

where $r \in \{0,1\}$, q is the density in (2.8), and ν is a dominating measure. Clearly this is a two-phase sampling design problem, and the above density function is a special case of that in (2.4).

3. Main Results

In this section, we place the fundamental results of the efficient score calculations of RRZ and RR in a more relevant format. Detailed proofs are in Yu and Nan (2005), Section 4.

Let $\dot{l}_\theta^0$ and $\dot{l}_\theta$ be the score functions for θ in models $\mathcal{Q}$ and $\mathcal{P}$, respectively, the partial derivatives of the corresponding log likelihood functions with respect to θ. Let l_θ^{*0} and l_θ^* be the efficient score functions for θ in models $\mathcal{Q}$ and $\mathcal{P}$, respectively. Here the superscript "0" stands for the full data.

Let $\dot{\mathcal{Q}}_\eta$ be the tangent space for the nuisance parameter η in model $\mathcal{Q}$, and $\dot{\mathcal{Q}}_\eta^\perp$ the orthogonal complement of $\dot{\mathcal{Q}}_\eta$ in $L_2^0(Q)$. Let $\dot{\mathcal{P}}_{\eta,\pi}$, $\dot{\mathcal{P}}_\eta$, and $\dot{\mathcal{P}}_\pi$ be the tangent spaces for nuisance parameters (η,π), η, and π, respectively, in model $\mathcal{P}$, and $\dot{\mathcal{P}}_{\eta,\pi}^\perp$, $\dot{\mathcal{P}}_\eta^\perp$, and $\dot{\mathcal{P}}_\pi^\perp$ be their orthogonal complements in $L_2^0(P)$. Here $L_2^0(\cdot)$ is the space of all zero mean and square integrable functions with respective to a probability measure that is either Q or P here. We refer to BKRW for definitions and detailed discussions on tangent spaces.

The goal is to obtain l_θ^*, the efficient score function for θ in model $\mathcal{P}$. When l_θ^* is obtained, the information matrix $I_\theta = E\{l_\theta^* l_\theta^{*\mathrm{T}}\}$ can be computed at given values of θ and η. Its inverse is the lower bound of the asymptotic variance matrices of all regular asymptotically linear estimators for θ in model $\mathcal{P}$. BKRW introduced two approaches to the calculation of efficient score functions: the nonparametric approach via derivatives of functions (BKRW, Section 3.3), and the semiparametric approach via score projections (BKRW, Section 3.4). Usually the second approach is more convenient for semiparametric models, and it is the approach used by RRZ and RR.

According to the score projection method of BKRW, efficient score functions l_θ^{*0} and l_θ^* can be written as

$$l_\theta^{*0} = \dot{l}_\theta^0 - \mathbf{\Pi}(\dot{l}_\theta^0|\dot{\mathcal{Q}}_\eta) = \mathbf{\Pi}(\dot{l}_\theta^0|\dot{\mathcal{Q}}_\eta^\perp), \quad l_\theta^* = \dot{l}_\theta - \mathbf{\Pi}(\dot{l}_\theta|\dot{\mathcal{P}}_{\eta,\pi}) = \mathbf{\Pi}(\dot{l}_\theta|\dot{\mathcal{P}}_{\eta,\pi}^\perp). \quad (3.1)$$

Here $\mathbf{\Pi}$ denotes projection. Model (2.4) often becomes complicated even though the density $q_{\theta,\eta}$ is simple (which implies the calculation of the first projection in

(3.1) is easy), and the calculation of l^*_θ directly from (2.4) through the second projection above can be extremely difficult. RRZ and RR are able to relate l^*_θ to the full data efficient score function l^{*0}_θ based on the following conditional expectations, making the calculation of l^*_θ possible.

Definition 3.1.

1. For $g^0 \in L^0_2(Q)$, define the conditional expectation map $\mathbf{A}$ from $L^0_2(Q)$ to $L^0_2(P)$ by

$$\mathbf{A}(g^0) \equiv E\{\, g^0(\boldsymbol{X}) \,|\, \boldsymbol{R}, \boldsymbol{X}_{(R)}\,\} = \sum_r I(\boldsymbol{R} = \boldsymbol{r}) E\{g^0(\boldsymbol{X}) \,|\, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{X}_{(r)}\} \ .$$

2. The transpose of $\mathbf{A}$, denoted as $\mathbf{A}^{\mathrm{T}}$, is a map from $L^0_2(P)$ to $L^0_2(Q)$ satisfying

$$E_P(g \cdot \mathbf{A}h^0) = E_Q(\mathbf{A}^{\mathrm{T}} g \cdot h^0)$$

for all $g \in L^0_2(P)$ and all $h^0 \in L^0_2(Q)$.

Apparently $\mathbf{A}$ is linear and continuous, its transpose is often called the adjoint of $\mathbf{A}$. A use of $\mathbf{A}$ and its transpose is clearly seen in the following nice connection between models $\mathcal{P}$ and $\mathcal{Q}$.

Lemma 3.1.

1. $\mathbf{A}(\dot{l}^0_\theta) = \dot{l}_\theta$ *and* $\mathbf{A}(\dot{l}^0_\eta) = \dot{l}_\eta$.
2. *The transpose of* $\mathbf{A}$ *is given by* $\mathbf{A}^{\mathrm{T}}(g) = E(g \,|\, \boldsymbol{X})$ *for* $g \in L^0_2(P)$.
3. *The composite map* $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ *from* $L^0_2(Q)$ *to* $L^0_2(Q)$ *is given by* $\mathbf{A}^{\mathrm{T}}\mathbf{A}(g^0) = E\{\mathbf{A}(g^0) \,|\, \boldsymbol{X}\} = \sum_r \pi(\boldsymbol{r}) E\{g^0(\boldsymbol{X}) \,|\, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{X}_{(r)}\}$.

Assertions 1 and 2 of Lemma 3.1 are the content of BKRW (see BKRW Section 4.6, p.144; Section 6.6, pp.271-272.) The proof of the lemma is given in Yu and Nan (2005) (see their proof of Lemma 2.1 that amounts to straightforward calculations of conditional expectations). The geometric meaning of the mapping $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is that it projects a function g^0 in $L^0_2(Q)$ to $L^0_2(P)$, then projects the projection in $L^0_2(P)$ back to the space $L^0_2(Q)$. Notice that $\mathbf{A}^{\mathrm{T}}$ usually is not the inverse of $\mathbf{A}$, thus $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ does not usually return g^0. The inverse of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is actually more important. From the Theorem 3.1 below we see that $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$ connects the two efficient score functions l^*_θ and l^{*0}_θ. Explicit calculation of $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$ is the key to obtaining the efficient score function l^*_θ, as seen in Theorems 3.2 and 3.3. The invertibility of both $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is shown by Yu and Nan (2005) in their Lemmas 4.1 and 4.2.

3.1. Arbitrary missingness

We first take

$$\mathcal{N}(\mathbf{A}^{\mathrm{T}}) \equiv \big\{\, a(\boldsymbol{R}, \boldsymbol{X}_{(R)}) : E(\, a \,|\, \boldsymbol{X}\,) = 0, \ a \in L^0_2(P)\big\}, \tag{3.2}$$

as the space of square integrable functions of observed data with conditional mean 0 given full data $\boldsymbol{X}$. By rearranging the material of Proposition 8.1 in RRZ to emphasize the calculation of efficient score function, we obtain the following theorem.

Theorem 3.1. *The efficient score function for θ in model $\mathcal{P}$ has the form*

$$l_\theta^* = \mathbf{U}(h^0) - \mathbf{\Pi}\Big(\mathbf{U}(h^0)\,\Big|\,\mathcal{N}(\mathbf{A}^{\mathrm{T}})\Big) = \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}(h^0)\,, \tag{3.3}$$

where $\mathbf{U}(h^0) \equiv \{I(\boldsymbol{R} = \mathbf{1}_m)/\pi(\mathbf{1}_m)\}\, h^0$, and h^0 is the unique function in $\dot{\mathcal{Q}}_\eta^\perp$ satisfying

$$\mathbf{\Pi}\Big((\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}(h^0)\,\Big|\,\dot{\mathcal{Q}}_\eta^\perp\Big) = l_\theta^{*0}. \tag{3.4}$$

Note that $\mathbf{U}(h^0)$ yields a Horvitz-Thompson inverse probability weighted estimating function (see e.g., Horvitz and Thompson (1952)) for the missing data problem when $h^0 \in \dot{\mathcal{Q}}_\eta^\perp$. The indicator $I(\boldsymbol{R} = \mathbf{1}_m)$ in the estimating function $\mathbf{U}(h^0)$ shows that only completely observed data are used in h^0, and weighted by $1/\pi(\mathbf{1}_m)$. Since it throws away all partially observed data, the Horvitz-Thompson inverse probability weighted estimating function $\mathbf{U}(h^0)$ usually yields an inefficient estimator. Theorem 3.1 gives the most efficient estimating function. Also note that (3.4) is usually an integral equation.

It is helpful to think about the intuition behind the theorem. From (3.1), we know that $l_\theta^* = \dot{l}_\theta - \mathbf{\Pi}(\dot{l}_\theta|\dot{\mathcal{P}}_{\eta,\pi})$. Since $\dot{l}_\theta = \mathbf{A}(\dot{l}_\theta^0)$ by assertion 1 of Lemma 3.1, and it can be shown that $\mathbf{\Pi}(\dot{l}_\theta|\dot{\mathcal{P}}_{\eta,\pi}) = \mathbf{\Pi}(\dot{l}_\theta|\dot{\mathcal{P}}_\eta) = \mathbf{A}(a^0)$ for some $a^0 \in \dot{\mathcal{Q}}_\eta$ (see Yu and Nan (2005), Section 4), we have $l_\theta^* = \mathbf{A}(\dot{l}_\theta^0 - a^0)$. Knowledge of a^0 at this stage would allow computation of l_θ^*. This is not always possible due to the complicity of model $\mathcal{P}$. All we know now is $l_\theta^* = \mathbf{A}(g^0)$ and $\mathbf{\Pi}(g^0|\dot{\mathcal{Q}}_\eta^\perp) = l_\theta^{*0}$ (by (3.1) and $a^0 \in \dot{\mathcal{Q}}_\eta$), where $g^0 = \dot{l}_\theta^0 - a^0 \in L_2^0(Q)$. The form of g^0, given as $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}(h^0)$ in Theorem 3.1, is determined by the following argument.

If we have the class of all estimating functions for regular asymptotically linear estimators for θ, then the optimal estimating function should be equivalent to the efficient score function (possibly multiplied by a constant), see e.g., BKRW. We can show that functions in such a class have the form $\mathbf{U}(h^0) + a$ where $h^0 \in \dot{\mathcal{Q}}_\eta^\perp$ and $a \in \mathcal{N}(\mathbf{A}^{\mathrm{T}})$, see Proposition 5.1, labeled as Proposition 4.2 in Yu and Nan (2005). Similar to the argument in van der Vaart (1998), page 383, $a = -E\{\mathbf{U}(h^0)|\mathcal{N}(\mathbf{A}^{\mathrm{T}})\}$ may minimize the variance of $\mathbf{U}(h^0) + a$ for a given h^0. Then we should be able to find the efficient score function by minimizing the variance of $\mathbf{U}(h^0) - E\{\mathbf{U}(h^0)|\mathcal{N}(\mathbf{A}^{\mathrm{T}})\}$ over all $h^0 \in \dot{\mathcal{Q}}_\eta^\perp$. This heuristic yields the first equality in (3.3) for some h^0.

Now we combine the two arguments above. Since $l_\theta^* = \mathbf{A}(g^0)$, and $\mathbf{U}(h^0) - l_\theta^* = \mathbf{U}(h^0) - \mathbf{A}(g^0) \in \mathcal{N}(\mathbf{A}^{\mathrm{T}})$ for the h_0 that determines the efficient score function in the form of $\mathbf{U}(h^0) - E\{\mathbf{U}(h^0)|\mathcal{N}(\mathbf{A}^{\mathrm{T}})\}$, we have $\mathbf{A}^{\mathrm{T}}\{\mathbf{U}(h^0) - \mathbf{A}(g^0)\} = 0$. Since $\mathbf{A}^{\mathrm{T}}\mathbf{U}(h^0) = h^0$ (which can be verified by straightforward conditional expectation calculation) and $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is invertible, we have $h^0 = \mathbf{A}^{\mathrm{T}}\mathbf{A}(g^0)$, and thus $g^0 = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}(h^0)$. As a result, the corresponding efficient score function is $\mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}(h^0)$.

The above idea is simple. But a proof requires care (see Section 4 of Yu and Nan (2005)). It is worth mention that this simple heuristic idea actually establishes the major steps of the proof of Theorem 3.1.

Without the middle expression in (3.3), Theorem 3.1 would look like a rewrite of the two equalities in (3.1), replacing $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}(h^0)$ by $\dot{l}_\theta^0 - a^0$ as in the above discussion. It is the middle expression in (3.3) that makes the whole argument meaningful. In addition to introducing the map $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$ into the efficient score formulation, it also gives us a hint about the functional form of l_θ^*. It can even be applied to calculate the efficient score l_θ^* directly if more knowledge about the structure of the space $\mathcal{N}(\mathbf{A}^{\mathrm{T}})$ is available, which is the case for two-phase designs when π is given (see Subsection 3.3).

An explicit form of $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$, however, is not available for arbitrary missingness patterns. We will see in the next subsection that the explicit form of $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$ does exist for monotone missingness. From Theorem 3.1 we see that for any specific full data model $\mathcal{Q}$, in addition to an explicit form of $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$, we also need the following three ingredients in order to derive the efficient score function via $l_\theta^* = \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}(h^0)$: (1) the efficient score function l_θ^{*0}; (2) the characterization of space $\dot{\mathcal{Q}}_\eta^\perp$; and (3) the calculation of projecting functions in $L_2^0(Q)$ to space $\dot{\mathcal{Q}}_\eta^\perp$. Very often the efficient score function l_θ^{*0} is ready to use. The other two, however, usually need more work (see e.g., Subsection 4.1 for the surrogate outcome regression).

3.2. Monotone missingness

We know that for monotone missingness, we have $\boldsymbol{r} \in \{\mathbf{1}_j : \ j = 1, \ldots, m\}$. If $\boldsymbol{r} = \mathbf{1}_k$, then $\boldsymbol{X}_{(r)} = (X_1, \ldots, X_k)$. Instead of using the whole vector $\boldsymbol{R}$ or $\boldsymbol{r}$, we can actually work on the individual indicator for each of the random variables in $\boldsymbol{X}$. Let R_k be the kth element of $\boldsymbol{R}$ and $R_0 = 1$ for convenience. Then $R_k = 1$ implies $R_j = 1$ whenever $j \le k$. As in RRZ, we define

$$\pi_k = P(R_k = 1 \,|\, R_{k-1} = 1, \boldsymbol{X}_{(\mathbf{1}_{k-1})}) \quad \text{and} \quad \bar{\pi}_k = \prod_{j=1}^{k} \pi_j \ . \qquad (3.5)$$

Let $\pi_0 = 1$ and $\bar{\pi}_0 = 1$. Then we have the following result for monotone missingness, corresponding to Proposition 8.2 in RRZ.

Theorem 3.2. *When data are missing in monotone patterns, the efficient score function for θ in model $\mathcal{P}$ has the form*

$$l_\theta^* = \frac{R_m}{\bar{\pi}_m} h^0 - \sum_{k=1}^{m} \frac{R_k - \pi_k R_{k-1}}{\bar{\pi}_k} E(h^0 \,|\, \boldsymbol{X}_{(\mathbf{1}_{k-1})}) \ , \tag{3.6}$$

where h^0 is the unique function in $\dot{\mathcal{Q}}_\eta^\perp$ satisfying

$$\mathbf{\Pi}\Big(\frac{1}{\bar{\pi}_m} h^0 - \sum_{k=1}^{m} \frac{1-\pi_k}{\bar{\pi}_k} E(h^0 \,|\, \boldsymbol{X}_{(\mathbf{1}_{k-1})}) \,\Big|\, \dot{\mathcal{Q}}_\eta^\perp \Big) = l_\theta^{*0} \ . \tag{3.7}$$

Notice that $I(\boldsymbol{R} = \mathbf{1}_m) = R_m$, and it can be easily shown that $\pi(\mathbf{1}_m) = \bar{\pi}_m$. So we see that the leading term on the right hand side of (3.6) is equal to $\mathbf{U}(h^0)$ in (3.3), which is an inverse sampling probability weighted estimating function using completely observed data.

Theorem 3.2 is obtained directly from Theorem 3.1 by replacing $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$ by its explicit form for the monotone missingness problem, which becomes the key part of the proof of Theorem 3.2. This explicit form can be derived by using the Neumann series and mathematical induction, see the proof of Proposition 4.3 in Yu and Nan (2005) for details.

3.3. Two-phase sampling designs

Consider the two-phase sampling scheme where we have either $\boldsymbol{R} = \mathbf{1}_m$ or $\boldsymbol{R} = \mathbf{1}_t$ for a known integer $t < m$. Hence $\pi(\mathbf{1}_t) = 1 - \pi(\mathbf{1}_m)$, which means that $\pi(\mathbf{1}_m)$ is a function of $\boldsymbol{X}_{(\mathbf{1}_t)}$, the set of always observed variables.

Theorem 3.3. *For two-phase sampling designs, the efficient score function for θ in model $\mathcal{P}$ has the form*

$$l_\theta^* = \frac{I(\boldsymbol{R} = \mathbf{1}_m)}{\pi(\mathbf{1}_m)} h^0 - \frac{I(\boldsymbol{R} = \mathbf{1}_m) - \pi(\mathbf{1}_m)}{\pi(\mathbf{1}_m)} E(h^0 \,|\, \boldsymbol{X}_{(\mathbf{1}_t)}) \ , \tag{3.8}$$

where h^0 is the unique function in $\dot{\mathcal{Q}}_\eta^\perp$ satisfying

$$\mathbf{\Pi}\left(\frac{1}{\pi(\mathbf{1}_m)} h^0 - \frac{1-\pi(\mathbf{1}_m)}{\pi(\mathbf{1}_m)} E(h^0 \,|\, \boldsymbol{X}_{(\mathbf{1}_t)}) \,\middle|\, \dot{\mathcal{Q}}_\eta^\perp \right) = l_\theta^{*0} \ . \tag{3.9}$$

Remark. When $\pi \equiv \pi(\mathbf{1}_m)$ is given, the space $\mathcal{N}(\mathbf{A}^{\mathrm{T}})$ for a two-phase design is well characterized. Let $a(\boldsymbol{R}, \boldsymbol{X}_{(R)}) \equiv a(R_m, \boldsymbol{X}_{(R)}) \in \mathcal{N}(\mathbf{A}^{\mathrm{T}})$. It is clear that $R_m = 1$ corresponds to $\boldsymbol{R} = \mathbf{1}_m$ and 0 otherwise. Since $a(R_m, \boldsymbol{X}_{(R)}) = R_m a(1, \boldsymbol{X}) + (1 - R_m) a(0, \boldsymbol{X}_{(\mathbf{1}_t)})$ and $E(a|\boldsymbol{X}) = 0$, we have $\pi a(1, \boldsymbol{X}) + (1 -$

$\pi)a(0,\boldsymbol{X}_{(\mathbf{1}_t)}) = 0$, which yields $a(1,\boldsymbol{X}) = -(1-\pi)\pi^{-1}a(0,\boldsymbol{X}_{(\mathbf{1}_t)})$. Hence $a(R_m,\boldsymbol{X}_{(R)}) = -(R_m-\pi)\pi^{-1}a(0,\boldsymbol{X}_{(\mathbf{1}_t)})$. So we obtain $\mathcal{N}(\mathbf{A}^{\mathrm{T}}) = \{-(R_m-\pi)\pi^{-1}\phi(\mathbf{X}_{(\mathbf{1}_t)}) : \phi \in L_2(Q)\}$.

The projection $\mathbf{\Pi}\Big(\mathbf{U}(h^0)|\mathcal{N}(\mathbf{A}^{\mathrm{T}})\Big)$ can thus be easily calculated. Since the projection is in $\mathcal{N}(\mathbf{A}^{\mathrm{T}})$, we let $\mathbf{\Pi}\Big(\mathbf{U}(h^0)|\mathcal{N}(\mathbf{A}^{\mathrm{T}})\Big) = (R_m-\pi)\pi^{-1}\phi^*(\mathbf{X}_{(\mathbf{1}_t)})$, and then obtain

$$\begin{aligned}
0 &= E\left[\left\{\mathbf{U}(h^0) - \frac{R_m-\pi}{\pi}\phi^*(\mathbf{X}_{(\mathbf{1}_t)})\right\}\cdot\frac{R_m-\pi}{\pi}\phi(\mathbf{X}_{(\mathbf{1}_t)})\right]\\
&= E\left[\left\{\frac{R_m-R_m\pi}{\pi^2}h^0 - \frac{R_m-2\pi R_m+\pi^2}{\pi^2}\phi^*(\mathbf{X}_{(\mathbf{1}_t)})\right\}\cdot\phi(\mathbf{X}_{(\mathbf{1}_t)})\right]\\
&= E\left[\frac{1-\pi}{\pi}\left\{E(h^0|\mathbf{X}_{(\mathbf{1}_t)}) - \phi^*(\mathbf{X}_{(\mathbf{1}_t)})\right\}\phi(\mathbf{X}_{(\mathbf{1}_t)})\right]
\end{aligned}$$

for all $\phi(\mathbf{X}_{(\mathbf{1}_t)}) \in L_2(Q)$. Hence $\phi^*(\mathbf{X}_{(\mathbf{1}_t)}) = E(h^0|\mathbf{X}_{(\mathbf{1}_t)})$ for the nontrivial case that $\pi \neq 1$. The same functional form of l_θ^* can be obtained directly from the first equality in (3.3), and the corresponding equation to (3.4) for a two-phase design can be determined by direct calculation, using the property that $l_\theta^* \perp \dot{\mathcal{P}}_\eta$.

4. Surrogate Outcome Regression

In this section, we study surrogate outcome regression with the conditional mean model introduced in Subsection 2.2. The goal is two-fold: to illustrate the steps of deriving the efficient score function for a missing data problem using the theoretical results in Section 3, and to develop an estimating method for surrogate outcome regression. The latter is itself an interesting statistical methodological problem with a broad range of applications.

4.1. The efficient score function

Since the surrogate outcome regression model we are interested in belongs to the family of two-phase designs with the nuisance parameter $\eta = (f_1, f_2, f_3)$, Theorem 3.3 applies directly. To do this, we first need to obtain the three ingredients listed at the end of Subsection 3.1. They are given in the following lemmas. We give detailed proofs here in order to clearly illustrate the application of Theorem 3.3 to the surrogate outcome regression model.

Lemma 4.1. *For any $b \in L_2^0(Q)$, we have*

$$\Pi(b|\dot{\mathcal{Q}}_\eta^\perp) = \frac{E\{b(S,Y,\mathbf{Z})\epsilon|\mathbf{Z}\}}{E(\epsilon^2|\mathbf{Z})}\epsilon\ . \tag{4.1}$$

Proof. In (2.8), for any one-parameter family of conditional densities $\{f_{1,\lambda}(s|y,\mathbf{z})\}$ with $f_{1,0}=f_1$, define

$$a_1(s,y,\mathbf{z})=\frac{\partial}{\partial\lambda}\log f_{1,\lambda}(s|y,\mathbf{z})\Big|_{\lambda=0} .$$

Then the score for the nuisance parameter f_1 is $\dot{l}_{f_1}a_1=a_1(s,y,\mathbf{z})$. Similarly, we can define the scores for nuisance parameters f_2 and f_3 as $a_2(y,\mathbf{z})$ and $a_3(\mathbf{z})$, respectively.

Direct calculations show that the three components of the tangent space of (2.8), $\dot{\mathcal{Q}}_1$, $\dot{\mathcal{Q}}_2$, and $\dot{\mathcal{Q}}_3$, corresponding to f_1, f_2, and f_3, respectively, are

$$\dot{\mathcal{Q}}_1=[a_1(S,Y,\mathbf{Z}):E(a_1|Y,\mathbf{Z})=0,Ea_1^2<\infty]\ , \tag{4.2}$$

$$\dot{\mathcal{Q}}_2=[a_2(Y,\mathbf{Z}):E(a_2|\mathbf{Z})=0,E(\epsilon a_2|\mathbf{Z})=0,Ea_2^2<\infty], \tag{4.3}$$

$$\dot{\mathcal{Q}}_3=[a_3(\mathbf{Z}):Ea_3=0,Ea_3^2<\infty]\ . \tag{4.4}$$

Here $[\cdot]$ denotes the closed linear span. It can be verified easily that these three spaces are mutually orthogonal. Thus the nuisance tangent space becomes: $\dot{\mathcal{Q}}_\eta=\dot{\mathcal{Q}}_1+\dot{\mathcal{Q}}_2+\dot{\mathcal{Q}}_3$, according to BKRW. The second restriction in (4.3) comes from the assumption that $E(\epsilon|\mathbf{Z})=0$. That $\dot{\mathcal{Q}}_2$ contains the right side in (4.3) is difficult to prove, as in the constrained models considered by BKRW (see the discussion in BKRW, pp.76-77). We assume equality as in RRZ, where they considered the mean regression model with missing covariates.

For any $b\in L_2^0(Q)$, let

$$r_b=\frac{E\{\epsilon b(S,Y,\mathbf{Z})|\mathbf{Z}\}}{E(\epsilon^2|\mathbf{Z})}\epsilon\ .$$

To prove (4.1), we show that $r_b\in\dot{\mathcal{Q}}_\eta^\perp=(\dot{\mathcal{Q}}_1+\dot{\mathcal{Q}}_2+\dot{\mathcal{Q}}_3)^\perp$ and that $b-r_b\in(\dot{\mathcal{Q}}_1+\dot{\mathcal{Q}}_2+\dot{\mathcal{Q}}_3)$. For any $a_1\in\dot{\mathcal{Q}}_1$,

$$\begin{aligned}\langle r_b,a_1\rangle_{L_2^0(Q)}=E(r_ba_1)&=E\left\{\frac{E(\epsilon b|\mathbf{Z})E(\epsilon a_1|\mathbf{Z})}{E(\epsilon^2|\mathbf{Z})}\right\}\\&=E\left\{\frac{E(\epsilon b|\mathbf{Z})E\{E(\epsilon a_1|Y,\mathbf{Z})|\mathbf{Z}\}}{E(\epsilon^2|\mathbf{Z})}\right\}\\&=E\left\{\frac{E(\epsilon b|\mathbf{Z})E\{\epsilon E(a_1|Y,\mathbf{Z})|\mathbf{Z}\}}{E(\epsilon^2|\mathbf{Z})}\right\}=0\end{aligned}$$

by (4.2). For any $a_2\in\dot{\mathcal{Q}}_2$, $\langle r_b,a_2\rangle_{L_2^0(Q)}=E(r_ba_2)=E\{E(\epsilon b|\mathbf{Z})E(\epsilon a_2|\mathbf{Z})/E(\epsilon^2|\mathbf{Z})\}$ $=0$ by (4.3). And, for any $a_3\in\dot{\mathcal{Q}}_3$, $\langle r_b,a_3\rangle_{L_2^0(Q)}=E(r_ba_3)=E\{E(\epsilon b|\mathbf{Z})a_3(\mathbf{Z})$ $E(\epsilon|\mathbf{Z})/E(\epsilon^2|\mathbf{Z})\}=0$ since $E(\epsilon|\mathbf{Z})=0$. Hence $r_b\in\dot{\mathcal{Q}}_\eta^\perp$.

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

Lemma 4.2. *The efficient score for θ in the full data model is*

$$l_\theta^{*0} = \Pi(\dot{l}_\theta^0|\dot{\mathcal{Q}}_\eta^\perp) = \frac{\partial g(\mathbf{Z};\theta)/\partial\theta}{E(\epsilon^2|\mathbf{Z})}\epsilon \ , \tag{4.5}$$

where $\dot{l}_\theta^0$ is the usual score function for θ in the full data model.

Proof. The score function of θ in (2.8) is $\dot{l}_\theta^0 = -(f_2'/f_2)(\epsilon|\mathbf{z})\partial g(\mathbf{z};\theta)/\partial\theta$. Thus the efficient score l_θ^{*0} in (4.5) can be obtained via direct calculation from (4.1) using the definition $l_\theta^{*0} = \Pi(\dot{l}_\theta^0|\dot{\mathcal{Q}}_\eta^\perp)$ in (3.1) and the fact that $E\{-\epsilon(f_2'/f_2)(\epsilon|\mathbf{Z})|\mathbf{Z}\} = 1$. The latter can be shown by differentiating both sides of (2.7), i.e., $\int \epsilon f_2(\epsilon|\mathbf{Z}) d\epsilon = 0$.

Notice that the full data efficient score function in (4.5) has been established by Chamberlain (1987) and RRZ.

Lemma 4.3. *The orthogonal complement of $\dot{\mathcal{Q}}_\eta$ in $L_2^0(Q)$ is*

$$\dot{\mathcal{Q}}_\eta^\perp = \left\{ h(\mathbf{Z})\epsilon : \ E\{h^2(\mathbf{Z})\epsilon^2\} < \infty \right\} . \tag{4.6}$$

Proof. Take $a_1 \in \dot{\mathcal{Q}}_1$, $a_2 \in \dot{\mathcal{Q}}_2$, and $a_3 \in \dot{\mathcal{Q}}_3$. Then we have $E\{a_1 h(\mathbf{Z})\epsilon|\mathbf{Z}\} = E\{h(\mathbf{Z})\epsilon E(a_1|Y,\mathbf{Z})|\mathbf{Z}\} = 0$, $E\{a_2 h(\mathbf{Z})\epsilon|\mathbf{Z}\} = h(\mathbf{Z})E\{a_2\epsilon|\mathbf{Z}\} = 0$, and $E\{a_3 h(\mathbf{Z})\epsilon|\mathbf{Z}\} = a_3 h(\mathbf{Z})E\{\epsilon|\mathbf{Z}\} = 0$, as in the proof of Lemma 4.1. This shows that $\{h(\mathbf{Z})\epsilon : \ E\{\epsilon^2 h^2(\mathbf{Z})\} < \infty\} \subset \dot{\mathcal{Q}}_\eta^\perp$. Equation (4.1) shows the reverse inclusion, since

$$E\left\{\frac{E^2(\epsilon b|\mathbf{Z})}{E^2(\epsilon^2|\mathbf{Z})}\ \epsilon^2\right\} \le Eb^2 < \infty$$

by the Cauchy inequality.

Note that the above proofs of the three lemmas for the complete data model Q slightly extend those in RRZ, van der Vaart (1998) and, in particular, Nan, Emond, and Wellner (2000) for the conditional mean regression model without the random variable S. For different regression models, calculations differ. The basic principle, however, is the same. For a quick view of more examples, we refer to Yu and Nan (2005), Section 5.

Plugging the results of Lemmas 4.1–4.3 into Theorem 3.3, we obtain the following result on the efficient score function for θ in the observed data model (2.9).

Theorem 4.1. *The efficient score function* l_θ^* *for the observed data* $(S, R*Y, \mathbf{Z}, R)$ *is given by*

$$l_\theta^* = \frac{\partial g(\mathbf{Z};\theta)/\partial\theta}{E(\epsilon^{*2}|\mathbf{Z})}\ \epsilon^*, \tag{4.7}$$

where

$$\epsilon^* = \frac{R}{\pi}Y - \frac{R-\pi}{\pi}E(Y|S,\mathbf{Z}) - g(\mathbf{Z};\theta)\ . \tag{4.8}$$

Proof. Let $h^0 = h(\mathbf{Z})\epsilon$ in (3.9). Then from equations (4.1), (4.5) and (4.6) in Lemmas 4.1−4.3, we obtain

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

Simplifying the above equality yields

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

Hence from (3.8) we obtain the efficient score l_θ^* for the observed data in the conditional mean model (2.6). It is given by

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

where $\epsilon^* = (R/\pi)Y - [(R-\pi)/\pi]E(Y|S,\mathbf{Z}) - g(\mathbf{Z};\theta)$, and it is easy to show that $E(\epsilon^{*2}|\mathbf{Z}) = E\{\epsilon^2/\pi - [(1-\pi)/\pi]E^2(\epsilon|S,\mathbf{Z})\Big|\mathbf{Z}\}$.

Let $Y^* = (R/\pi)Y - \{(R-\pi)/\pi\}E(Y|S,\mathbf{Z})$ be a kind of "transformation" to the response variable Y. Using the nested conditional expectation property, we can easily verify that $E(Y^*|\mathbf{Z}) = E(Y|\mathbf{Z}) = g(\mathbf{Z};\theta)$. Hence by comparing (4.5) and (4.7), we see that the efficient score l_θ^* actually has the same form as that of the efficient score for the "full" data $(Y^*, \mathbf{Z})$. Analyzing the observed data $(S, R * Y, \mathbf{Z}, R)$ with the outcome Y missing at random, and the availability of a surrogate outcome S, is actually similar to analyzing the "full" data $(Y^*, \mathbf{Z})$ with the same conditional mean structure as that of $(Y, \mathbf{Z})$. The interpretation of the parameter θ does not change at all, even though the scale of Y^* may not be the same as that of Y.

4.2. Estimation

When we have complete data, the surrogate outcome S does not contribute to the estimation of θ, as seen in Lemma 4.2. Chamberlain (1987) and RRZ (Proposition 3.1, p.852), among others, showed that the asymptotically efficient estimator of θ for complete data can be obtained by solving the estimating equation

$$\sum_i \frac{\frac{\partial}{\partial\theta} g(\mathbf{Z}_i;\theta)}{E(\epsilon^2|\mathbf{Z}_i)}\ \epsilon_i = 0\ . \tag{4.9}$$

This equation has the same form as the quasi-likelihood estimating equation, see e.g., McCullagh (1983). Inevitably, the conditional variance $\mathrm{Var}\,(Y|\mathbf{Z}) = E(\epsilon^2|\mathbf{Z})$ needs to be specified or estimated in order to calculate the estimator for θ, and correctly specified or consistently estimated in order to achieve efficiency. Carroll and Ruppert (1982) and Robinson (1987) showed that for a linear model, $g(\mathbf{Z};\theta) = \mathbf{Z}^T\theta$, substituting $E(\epsilon^2|\mathbf{Z})$ by its kernel smoothing estimator in the above estimating equation yields the efficient estimator for θ. Newey (1993) extended the smoothing method to generalized linear models.

For the surrogate outcome regression problem, we propose an estimating method for θ based on the efficient score function (4.7) in this subsection. Since (4.7) contains unknown quantities $E(Y|S,\mathbf{Z})$ and $E(\epsilon^{*2}|\mathbf{Z})$, we need to either model or estimate them. In some biomedical studies, the surrogate outcome might have been investigated to the degree that the functional form of $E(Y|S,\mathbf{Z})$ could be estimated from previous studies, especially when S satisfies the surrogate criterion of Prentice (1989): $E(Y|S,\mathbf{Z}) = E(Y|S)$. Then Y^* could be obtained for each record in a new study, and the observed data could be treated as independent and identically distributed copies of $(Y^*, \mathbf{Z})$. Thus estimation using the efficient score (4.7) would be a standard practice of quasi-likelihood methods.

The more interesting case occurs when there is no previous study that allows the estimation of $E(Y|S,\mathbf{Z})$. Following the ideas for the estimation with full data from equation (4.9), we can estimate both $E(Y|S,\mathbf{Z})$ and $E(\epsilon^{*2}|\mathbf{Z})$ nonparametrically via smoothing. We assume the probability function $\pi(s,\mathbf{z})$ is known, which is the case for two-phase designs where missing data are caused by design.

Let $\eta = (\eta_1, \eta_2)$, where $\eta_1 = E(Y|S, \mathbf{Z})$ and $\eta_2 = E(\epsilon^{*2}|\mathbf{Z})$. We rewrite l^*_θ as $l^*_{\theta,\eta}$. Thus,

$$l^*_{\theta,\eta} = \frac{\frac{\partial}{\partial\theta} g(\mathbf{Z};\theta)}{\eta_2} \left\{ \frac{R}{\pi} Y - \frac{R-\pi}{\pi}\eta_1 - g(\mathbf{Z};\theta) \right\} \ , \tag{4.10}$$

and θ is estimated by solving

$$\sum_{i=1}^{n} l^*_{\theta,\hat{\eta}_n(\theta)}(S_i, R_i * Y_i, \mathbf{Z}_i, R_i) = 0 \ . \tag{4.11}$$

Here is an algorithm for solving (4.11).

Step 1: Estimate $\eta_1 = E(Y|S, \mathbf{Z})$ via smoothing, using all fully observed records. Note that the observing probabilities for those records vary, so the ith fully observed record should have weight $1/\pi(S_i, \mathbf{Z}_i)$. Calculate $\hat{\eta}_{1,n}$ for all records, including those with missing data.

Step 2: Choose an initial value of the estimator for θ, $\hat{\theta}_{(0)}$.

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

Step 4: Plug $\hat{\eta}_{1,n}$ and $\hat{\eta}_{2,n}(\hat{\theta}_{(0)})$ into (4.11) and solve for $\hat{\theta}_n$.

Step 5: Use the root of (4.11) in Step 4 as a new initial value of $\hat{\theta}_n$ and repeat Steps 2−4 until $\hat{\theta}_n$ converges.

The variance estimator for $\hat{\theta}_n$ is

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

where $\dot{l}^*_{\theta,\eta} = \partial l^*_{\theta,\eta}/\partial\theta$. It is asymptotically equivalent to $\left(\sum_{i=1}^{n} l^*_{\hat{\theta}_n,\hat{\eta}_n,i} l^{*T}_{\hat{\theta}_n,\hat{\eta}_n,i} \right)^{-1}$, and thus $\hat{\theta}_n$ is semiparametrically efficient, if the smoothing estimates for $\eta_1 = E(Y|S, \mathbf{Z})$ and $\eta_2 = E(\epsilon^{*2}|\mathbf{Z})$ are consistent.

In the above algorithm, the initial value of $\hat{\theta}_{(0)}$ can be obtained using the following Horvitz-Thompson estimating equation for θ:

$$\sum_{i=1}^{n} \frac{R_i}{\pi(S_i, \mathbf{Z}_i)} \frac{\frac{\partial}{\partial\theta} g(\mathbf{Z}_i;\theta)}{\mathrm{Var}\,(Y|\mathbf{Z}_i)} \left\{ Y_i - g(\mathbf{Z}_i;\theta) \right\} = 0 \ .$$

When $\mathbf{Z}$ is discrete, $E(\epsilon^{*2}|\mathbf{Z}) = \mathrm{Var}\,(Y^*|\mathbf{Z})$ can be estimated from data grouped on distinct values of $\mathbf{Z}$ without using residuals, and thus the above algorithm does not need iteration.

The estimator obtained from the above algorithm has the following nice asymptotic properties, under mild regularity conditions.

Theorem 4.2. *Suppose the regularity conditions given in the Appendix hold. Then the estimator of θ in* (2.6), *with distribution* (2.9), *obtained by solving equation* (4.11) *is consistent and asymptotically Gaussian. In particular, if both η_1 and η_2 are consistently estimated, then the corresponding estimator achieves the information bound determined by the efficient score in equation* (4.7).

The proof of Theorem 4.2 requires modern empirical process theory and can be found in Appendix A of Yu and Nan (2005), where the corresponding result is labeled as Theorem 6.2. Note that we do not require consistent estimators for any nuisance parameters to obtain the asymptotic normality for $\hat{\theta}_n$ mainly because $El^*_\theta = 0$ for any η_1 and η_2. We do, however, need consistent estimators for both η_1 and η_2 in order to achieve the information bound. Their estimators, however, can converge to the truth without restrictions on rates. But a slower rate, which happens when applying a smoothing technique, may require a larger sample size to achieve a stable estimator for θ, as shown by the following simulation study.

4.3. A simulation study

We conducted simulations with continuous Z and S to investigate the validity of handling continuous variables Z and S via smoothing. Suppose the underlying true model is

$$E(Y|S,Z) = \theta_0 + \theta_1 Z + \theta_2 f(S), \tag{4.12}$$

and the model of interest is

$$E(Y|Z) = \theta_0 + \theta_1 Z, \tag{4.13}$$

where $f(S) = S^{1/3} \sim N(0,1)$. Let $Z \sim N(0,1)$, $\epsilon_0 = Y - E(Y|S,Z) \sim N(0,1)$, and $\theta_0 = \theta_1 = \theta_2 = 1$. Simulations are conducted using 1,000 replications with cohort size $n = 200$ and $1,000$, respectively, and the selection probability $\pi(S,Z) = 0.5$. In other words, we observe Y for half of the subjects. Here S does not satisfy the definition of surrogate outcome given by Prentice (1989). Since it is correlated with the true outcome Y, however, we can use it in the same way as a surrogate outcome to improve efficiency. The following estimating methods are simulated: (i) fitting the linear regression model (4.13) using fully observed data only (complete-case method); (ii) using the proposed method with $\eta_1 = E(Y|S,Z)$ estimated using the true linear regression model (4.12) (without smoothing); and (iii) using the proposed method with $\eta_1 = E(Y|S,Z)$ estimated by a generalized additive model via smoothing splines on S and Z. We use smoother $s()$ in Splus with default values of the smoothing parameters. We do not estimate the variance $\eta_2 = E(\epsilon^{*2}|Z)$ for the above simulations since they are actually constant. The simulation results are listed in Table 1. When sample size is small ($n = 200$), the algorithm does not work very well. When we increase the sample size to $n = 1,000$, which means that about 500 records are used to estimate $E(Y|S,Z)$ using a generalized additive model, the algorithm works as well as it would if the model for $E(Y|S,Z)$ were correctly specified.

Table 1. Simulation summary statistics for estimating θ_0 and θ_1 in linear models with 1,000 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
CSM[f]	0.9960	0.9960	0.0149	0.0151	0.0150	0.0165	0.956	0.937
SM[g]	1.0040	0.9928	0.1644	0.1715	0.1761	0.3542	0.954	0.935
				$n = 1,000$				
CC	1.0032	1.0021	0.0040	0.0040	0.0041	0.0045	0.952	0.934
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

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

5. Concluding Remarks

In practice it may not always be possible to obtain a useable form of the efficient score function for developing an efficient estimator for θ. It then becomes important to obtain the class of all estimating functions for regular estimators. The following proposition gives us the desirable class of estimating functions, see also Proposition 8.1 in RRZ.

Proposition 5.1. $\dot{\mathcal{P}}^{\perp}_{\eta,\pi} = \{\mathbf{V}(h^0, a) : \; h^0 \in \dot{\mathcal{Q}}^{\perp}_{\eta}, a \in \mathcal{N}(\mathbf{A}^{\mathrm{T}})\}$, *where* $\mathbf{V}(h^0,a)$ *is the map from* $L_2^0(Q) \times L_2^0(P)$ *to* $L_2^0(P)$ *given by*

$$\mathbf{V}(h^0,a) \equiv \mathbf{U}(h^0) + a - \mathbf{\Pi}\{\mathbf{U}(h^0) + a \,|\, \dot{\mathcal{P}}_{\pi}\} = \mathbf{\Pi}\{\mathbf{U}(h^0) + a \,|\, \dot{\mathcal{P}}_{\pi}^{\perp}\}.$$

The proof of the proposition is referred to Yu and Nan (2005), where the corresponding result is labeled as Proposition 4.2. Proposition 5.1 characterizes the space $\dot{\mathcal{P}}_{\eta,\pi}^{\perp}$ that contains all the estimating functions (or influence functions) for regular asymptotically linear estimators in model $\mathcal{P}$, including the efficient score function. Since $a \in \mathcal{N}(\mathbf{A}^{\mathrm{T}})$ and $\dot{\mathcal{P}}_{\pi} \subset \mathcal{N}(\mathbf{A}^{\mathrm{T}})$ (see Lemma 4.4 in Yu and Nan (2005)), we know that any estimating function in model $\mathcal{P}$ is an inverse probability weighted estimating function in model $\mathcal{Q}$, plus a term that has expectation zero given full data $\boldsymbol{X}$. This property gives us the flexibility to develop workable estimating methods for a missing data problem if an explicit efficient estimating function is hard to get.

From the remark following Theorem 3.3 we know that, for a two-phase sampling design where π is given, the second term in any such estimating function has the form $-(R_m - \pi(\mathbf{1}_m))\pi(\mathbf{1}_m)^{-1}\phi(\boldsymbol{X}_{(\mathbf{1}_t)})$, where ϕ has finite second moment. This type of estimating function has the so-called double robustness property, i.e., the estimating function

$$\frac{R_m}{\pi(\mathbf{1}_m)} h^0(\boldsymbol{X}) - \frac{R_m - \pi(\mathbf{1}_m)}{\pi(\mathbf{1}_m)} \phi(\boldsymbol{X}_{(\mathbf{1}_t)})$$

for an arbitrary full data estimating function h^0 is unbiased if either π or $\phi(\boldsymbol{X}_{(\mathbf{1}_t)}) = E(h^0 \,|\, \boldsymbol{X}_{(\mathbf{1}_t)})$ is correctly specified. The details follow. If π is correctly specified, then for any h^0 satisfying $E(h^0) = 0$, and any $\phi(\mathbf{X}_{(\mathbf{1}_t)}) \in L_2(Q)$, we have

$$E\left\{\frac{R_m}{\pi}h^0 - \frac{R_m - \pi}{\pi}\phi(\mathbf{X}_{(\mathbf{1}_t)})\right\} = E\left\{h^0 - 0 \cdot \phi(\mathbf{X}_{(\mathbf{1}_t)})\right\} = E(h^0) = 0.$$

If $\phi(\mathbf{X}_{(\mathbf{1}_t)}) = E(h^0|\mathbf{X}_{(\mathbf{1}_t)})$ is correctly specified for a full data estimating function h^0, then for any π we have

$$\begin{aligned}
& E\left\{\frac{R_m}{\pi}h^0 - \frac{R_m - \pi}{\pi}\phi(\mathbf{X}_{(\mathbf{1}_t)})\right\} \\
&= E\left\{\frac{P(R_m = 1|\mathbf{X})}{\pi}h^0 - \frac{P(R_m = 1|\mathbf{X}) - \pi}{\pi}E(h^0|\mathbf{X}_{(\mathbf{1}_t)})\right\} \\
&= E\left\{\frac{P(R_m = 1|\mathbf{X}_{(\mathbf{1}_t)})}{\pi}h^0 - \frac{P(R_m = 1|\mathbf{X}_{(\mathbf{1}_t)}) - \pi}{\pi}E(h^0|\mathbf{X}_{(\mathbf{1}_t)})\right\} \\
&= E\left\{\frac{P(R_m = 1|\mathbf{X}_{(\mathbf{1}_t)})}{\pi}E(h^0|\mathbf{X}_{(\mathbf{1}_t)}) - \frac{P(R_m = 1|\mathbf{X}_{(\mathbf{1}_t)}) - \pi}{\pi}E(h^0|\mathbf{X}_{(\mathbf{1}_t)})\right\} \\
&= E\left\{E(h^0|\mathbf{X}_{(\mathbf{1}_t)})\right\} \;=\; E(h^0) \;=\; 0.
\end{aligned}$$

Note that an unbiased estimating function does not necessarily yield an asymptotically normally distributed estimator. Usually more work needs to be done to obtain desirable asymptotic properties, as we have seen in the case of the surrogate outcome regression model.

Acknowledgements

The authors would like to thank Mary Emond and Jon Wellner for their generous help in this research.

Appendix. Regularity Conditions for Theorem 4.2

We give a set of regularity conditions that guarantee the desirable asymptotic properties of the proposed estimators. They are reasonable for many practical problems.

(C1) The surrogate outcome S and covariates $\mathbf{Z}$ have finite support.

(C2) The parameter space of θ, Θ, is compact.

(C3) $\sup_{S,\mathbf{Z}} |\eta_1(S,\mathbf{Z})| < M_1 < \infty$.

(C4) $0 < \sigma < \eta_2(\mathbf{Z}) < M_2 < \infty$ for all $\mathbf{Z}$.

(C5) The function $g(\mathbf{Z};\theta)$ is twice differentiable in θ, with continuous second derivative for all $\mathbf{Z}$; and $E\{\dot{g}(\mathbf{Z};\theta_0)\dot{g}(\mathbf{Z};\theta_0)^T\}$ is nonsingular. Here $\dot{g} = \partial g/\partial\theta$, $\ddot{g} = \partial\dot{g}/\partial\theta^T$, and θ_0 is the true parameter, an interior point in Θ.

(C6) The true parameter θ_0 is the unique root of $E_0 l_\theta^{*0} = 0$ for an arbitrary variance function. Here E_0 denotes the expectation with respect to the true probability measure of the underlying full data.

(C7) The parameters η_1 and η_2 and their estimators belong to Donsker classes.

Note Conditions (C1)$-$(C4) are common assumptions for regression models. Condition (C5) holds for all the mean functions of the generalized linear models discussed by McCullagh and Nelder (1989). Condition (C6) is a usual assumption for the proof of consistency in estimating equation theory for problems without missing data (Huber (2004), p.131). Condition (C7) holds when both η_1 and η_2 are estimated using generalized additive models (see e.g., Hastie and Tibshirani (1990)) where each component in the generalized additive models is a nice function of its parameters, e.g., (the sum of) Lipschitz or bounded monotone functions. An example of the latter situation estimates each component using polynomial splines. We refer to van der Vaart and Wellner (1996) for general discussions on Donsker properties.

References

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

Carroll, R. J. and Ruppert D. (1982). Robust estimation in heteroscedastic linear models. *Ann. Statist.* **10**, 429-441.

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

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**, 663-685.

Huber, P. J. (2004). *Robust Statistics.* John Wiley, New York.

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

McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**, 59-67.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.*, 2nd edition. Chapman and Hall, London.

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.

Newey, W. K. (1993) Efficient estimation of models with conditional moment restrictions. *Handbook of Statistics* **11** (Edited by G. S. Maddala, C. R. Rao and H. D. Vinod), 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. *J. Statist. Plann. Inference* **42**, 137-160.

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

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS Epidemiology - Methodological Issues*, (Edited by N. Jewell, K. Dietz and V. Farewell), 297-331. Birkhäuser, Boston.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846-866.

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

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

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.

Yu, M. and Nan, B. (2005). Semiparametric regression models with missing data: the mathematical review and a new application. Technical Report, Department of Biostatistics, University of Michigan. Downloadable from the second author's web site: http://www.sph.umich.edu/∼bnan/research/

The Division of Biostatistics, Department of Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, U.S.A.

E-mail: meyu@iupui.edu

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

E-mail: bnan@umich.edu

(Received May 2003; accepted July 2005)