Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2008)
70, *Part* 4, *pp.* 803–823

Improving semiparametric estimation by using surrogate data

Song Xi Chen,

Iowa State University, Ames, USA, and Peking University, Beijing, People's Republic of China

Denis H. Y. Leung

Singapore Management University, Singapore

and Jing Qin

National Institutes of Health, Bethesda, USA

[Received October 2006. Revised December 2007]

Summary. The paper considers estimating a parameter β that defines an estimating function $U(y, x, \beta)$ for an outcome variable y and its covariate x when the outcome is missing in some of the observations. We assume that, in addition to the outcome and the covariate, a surrogate outcome is available in every observation. The efficiency of existing estimators for β depends critically on correctly specifying the conditional expectation of U given the surrogate and the covariate. When the conditional expectation is not correctly specified, which is the most likely scenario in practice, the efficiency of estimation can be severely compromised even if the propensity function (of missingness) is correctly specified. We propose an estimator that is robust against the choice of the conditional expectation via an empirical likelihood. We demonstrate that the estimator proposed achieves a gain in efficiency whether the conditional score is correctly specified or not. When the conditional score is correctly specified, the estimator reaches the semiparametric variance bound within the class of estimating functions that are generated by U. The practical performance of the estimator is evaluated by using simulation and a data set that is based on the 1996 US presidential election.

Keywords: Empirical likelihood; Estimating equations; Missing values; Surrogate outcome

1. Introduction

Missing data are common in empirical studies. Statistical analysis in such situations is challenging. On one hand, every observation, whether it contains missing variables or not, carries some information. On the other hand, observations with missing variables must be handled delicately for valid inferences to be drawn. In this paper, we study the problem where the outcome variable of a study is missing in a subset of the sampled data. We assume that, apart from the outcome and the covariates, the study also collected information on a surrogate or proxy variable on every observation. Data of this nature are common in many disciplines. For example, in health sciences research, to evaluate the success of a treatment or procedure, it is often very difficult to observe the clinical outcome (e.g. cured *versus* not cured) in every study participant; therefore, a surrogate outcome (e.g. biomarkers) may be used for those participants without the true

Address for correspondence: Song Xi Chen, Department of Statistics, Iowa State University, Ames, IA 50011-1210, USA.
E-mail: songchen@iastate.edu

 1369–7412/08/70803

outcome (e.g. Wittes *et al.* (1989), Begg and Leung (2000), Leung (2001), Baker *et al.* (2005), Burzykowski *et al.* (2005) and Baker (2006)) and, in economics, proxy or surrogate outcomes are often used in surveys with missing responses (Chen *et al.*, 2005).

One difficulty in modelling data with missing outcomes is that the mechanism that leads to the missing data is often unknown or at best can only be approximated. For example, when there is a missing response in a survey, it is very difficult to ascertain the reason for the non-response. The non-response may be completely random, or it may depend on some (observed) variables or it may be related to the (unobserved) outcome. If the non-response is related to the unobserved outcome, then the identifiability of the solution may be called into question.

One solution to the identifiability problem is to use a surrogate outcome. We focus on situations with data missing at random (MAR), i.e. the probability of a missing outcome is independent of the (unobserved) outcome, given the surrogate and the covariates (Little and Rubin, 2002). Under data MAR, the model is identifiable if the surrogate and the covariates are always observed. Situations where the outcome is missing completely at random (MCAR) is a special case of MAR data and are also covered by the methods that are discussed herein.

Let Y be the outcome variable, X be the covariates of interest, S be a surrogate for Y and Z be an additional set of covariates that is not of direct interest. Suppose that S, X and Z are always observed but Y is missing in some observations. Let δ be an indicator variable that takes the value 1 if Y is observed and 0 otherwise. The sampled data consist of two parts; a part with (Y, S, X, Z) completely observed,

$$(\delta_1 = 1, y_1, x_1, s_1, z_1), \ldots, (\delta_m = 1, y_m, x_m, s_m, z_m),$$

and a part with missing Y,

$$(\delta_{m+1} = 0, ?, x_{m+1}, s_{m+1}, z_{m+1}), \ldots, (\delta_{m+n} = 0, ?, x_{m+n}, s_{m+n}, z_{m+n}).$$

Let $N = m + n$. We assume that Y is MAR in the sampled data, i.e.

$$P(\delta = 1 | y, x, s, z) = w(s, x, z, \theta),$$

where the form of w is known up to a parameter θ. For ease of discussion, in the next few sections, we assume that Z is null and we drop Z from the formulation of w. The results that we discuss also apply in the more general case where Z is non-null and, in Section 5, we apply the proposed method in a situation where Z is non-null.

Suppose that w can be estimated by $\hat{\theta}$ that maximizes the binomial log-likelihood:

$$l_{\mathrm{B}}(\theta) = \sum_{i=1}^{N} [\delta_i \log\{w(s_i, x_i, \theta)\} + (1 - \delta_i) \log\{1 - w(s_i, x_i, \theta)\}]. \tag{1}$$

The function w is a propensity score in the sense of Rosenbaum and Rubin (1983). The full log-likelihood based on the observed data is

$$l_{\mathrm{full}} = l_{\mathrm{B}}(\theta) + \sum_{i=1}^{m} \log\{f(y_i, x_i, s_i)\} + \sum_{j=m+1}^{N} \log\{f(x_i, s_i)\}. \tag{2}$$

If parametric models are postulated for $f(y, x, s)$ and $f(x, s)$, then making inferences is straightforward by maximizing the parametric likelihood. In practice, however, parametric models are often difficult to specify.

Suppose that $f(y|x) = f(y|x, \beta)$ is the conditional density of Y given X without considering S; then

$$U(y, x, \beta) = \frac{\partial \log\{f(y|x, \beta)\}}{\partial \beta}$$

is the conditional score of Y given X. Here, the parameter β is of primary interest. One way to utilize information in S is to consider the conditional density of S given X

$$f(s|x) = \int f(y|x, \beta)\, g(s|y, x)\, \mathrm{d}y. \tag{3}$$

However, in general, it is difficult to specify the law $[S|Y, X]$, especially when S is multivariate (Clayton *et al.*, 1998). When Y is MCAR, Pepe (1992) proposed an estimated likelihood method by replacing the unknown conditional density $g(s|y, x)$ in equation (3) by a kernel density estimate that is based on the completely observed data. Schenker and Taylor (1996) suggested the use of imputation (Rubin, 1987) for missing outcomes. Chen and Chen (2000) suggested a method that is based on the regression estimate. Chen *et al.* (2003) used a two-sample empirical likelihood (EL), one based on estimating equations from the complete observations and another based on the observations with missing outcomes. However, the methods of Chen and Chen (2000) and Chen *et al.* (2003) cannot be applied to the practically important case of data MAR because the structure of the likelihood is changed owing to the selection bias in the missingness under the assumption of data MAR. We propose a new approach that corrects for the selection bias by employing the biased sampling technique of Vardi (1985).

Instead of specifying $g(s|y, x)$, Robins *et al.* (1995) and Robins and Rotnitzky (1995) proposed the use of estimating equations for situations where Y can be MAR. In the framework that is considered here, their estimator (which is denoted by $\hat{\beta}_{\mathrm{RRZ}}$ hereafter) solves

$$\sum_{i=1}^{N} \left\{ \frac{\delta_i}{w(s_i, x_i, \hat{\theta})} U(y_i, x_i, \beta) - \frac{\delta_i - w(s_i, x_i, \hat{\theta})}{w(s_i, x_i, \hat{\theta})} \psi(s_i, x_i, \beta) \right\} = 0, \tag{4}$$

for a specific function ψ and a mean 0 estimating function U. If U is the score function for $f(y|x)$, then $\psi^* \equiv E\{U(y, x, \beta)|s, x\}$ corresponds to the conditional score function of Y given S and X. For a given unbiased estimating function $U(y, x, \beta)$, their estimator can attain the semiparametric efficiency bound within the class of estimating function that is generated by $U(y, x, \beta)$ (Newey, 1990) for estimating β if $\psi(s, x, \beta) = \psi^*(s, x, \beta)$. Furthermore, $\hat{\beta}_{\mathrm{RRZ}}$ is consistent if either w or ψ is correctly specified. This property is the so-called 'doubly robustness' property. However, $\hat{\beta}_{\mathrm{RRZ}}$ may suffer loss of efficiency when $\psi \neq \psi^*$, as shown in theorem 2 of Section 3.

The estimator (4) is a special case of a larger class of semiparametric efficient estimators that were developed by Robins *et al.* (1994). However, as Chen and Chen (2000) pointed out, the semiparametric efficient estimator that was suggested by Robins *et al.* (1994) is not practically feasible in general since the optimal estimating functions can only be obtained by solving a functional integral equation. The closed form optimal estimating equation (4) exists in the case that is considered here, i.e. $U(y, x, \beta)$ is the conditional score and S is a surrogate outcome. Recently, Chen and Breslow (2004) and Yu and Nan (2006) also discussed two situations that were similar to those considered here where closed form optimal estimating equations can be found.

Even though ψ^* is rarely known precisely, an estimate of $\psi^* \equiv E[U(y, x, \beta)|s, x]$ can be found, as follows. Let $\tilde{\beta}$ be a consistent estimate of β; $U(Y, X, \tilde{\beta})$ may be regressed on S and X to give a model

$$U(y, x, \hat{\beta}) = \psi(s, x, \gamma) + \varepsilon \tag{5}$$

with unknown parameter γ, using the complete data. Therefore, ψ is a working estimate of ψ^*. In general, ψ may be not be a perfect guess; hence $E[\psi]$ may be non-zero. However, the estimator

that is obtained from equation (4) is valid, albeit inefficient, since the estimating equation itself always has zero mean under the true parameter.

In this paper, we develop a set of weighted score equations by using EL weights obtained by leveraging the information that is contained in S and X. When ψ and w are correctly specified, our method is efficient within the class of estimating functions that are defined by $U(Y, X, \beta)$. Even when ψ is incorrectly specified, as long as w is correctly specified, it still achieves good efficiency. The rest of the paper is organized as follows. In Section 2, we use EL to combine unbiased estimating equations. Large sample results appear in Section 3. In Section 4, we report the results from a simulation study that compares the method proposed with existing methods. In Section 5, the method is applied to a real data set. Conclusions are given in Section 6. Proofs are given in Appendix A.

2. Method proposed

Suppose that $U(y, x, \beta)$ is an estimating function that captures the relationship between Y and X through a parameter β, and $\psi(s, x, \beta, \gamma)$ is a function of S and X. Without further explicit notation, we assume that X, β and γ may be vector valued.

Let $\tilde{\beta}$ be a consistent estimator of β. For example, $\tilde{\beta}$ may be the Horvitz and Thompson (1952) inversely weighted estimator $\hat{\beta}_{\mathrm{W}}$ that solves

$$\sum_{i=1}^{N} \frac{\delta_i \, U(y_i, x_i, \beta)}{w(s_i, x_i, \hat{\theta})} = 0 \tag{6}$$

where $\hat{\theta}$ is the binomial likelihood estimator that was given earlier.

By conditioning on the missingness status δ, the full likelihood based on the data can be written as

$$\prod_{i=1}^{N} W^{\delta_i}(1-W)^{1-\delta_i} \prod_{i=1}^{m} P(y_i, s_i, x_i | \delta_i = 1) \prod_{j=m+1}^{N} P(s_j, x_j | \delta_j = 0), \tag{7}$$

where $W = P(\delta = 1)$. Let $p_i = P(y_i, s_i, x_i | \delta_i = 1) = w(s_i, x_i, \theta)\, \mathrm{d}F(y_i, x_i, s_i)/W$ for $i = 1, 2, \ldots, m$ and $q_j = P(s_j, x_j | \delta_j = 0) = \{1 - w(s_j, x_j, \theta)\}\, \mathrm{d}F(x_j, s_j)/(1 - W)$ for $j = m+1, \ldots, N$. As indicated in Section 1, the mean of $\psi(s, x, \beta, \gamma)$ may not be 0. Therefore, likelihood (7) cannot be used directly for inferences. However,

$$E\left[\frac{\psi(s, x, \beta, \gamma) - \mu}{w(s, x, \theta)} \middle| \delta = 1\right] = 0,$$
$$E\left[\frac{\psi(s, x, \beta, \gamma) - \mu}{1 - w(s, x, \theta)} \middle| \delta = 0\right] = 0,$$

where $\mu = E[\psi(s, x, \beta, \gamma)]$. Therefore, with an appropriate initial estimate $\tilde{\gamma}$ to be discussed later, approximately

$$\begin{aligned} \sum_{i=1}^{m} \frac{\psi(s_i, x_i, \tilde{\beta}, \tilde{\gamma}) - \mu}{w(s_i, x_i, \hat{\theta})} &= 0, \\ \sum_{j=m+1}^{N} \frac{\psi(s_j, x_j, \tilde{\beta}, \tilde{\gamma}) - \mu}{1 - w(s_j, x_j, \hat{\theta})} &= 0, \end{aligned} \tag{8}$$

can be used for making inferences, as follows. A log-EL (Owen, 1990) for μ is

$$l(\mu)=\sum_{i=1}^{m}\log(p_i)+\sum_{j=m+1}^{N}\log(q_j),$$

subject to $\Sigma_{i=1}^{m}p_i=1,\ p_i\geqslant 0, \Sigma_{j=m+1}^{N}q_j=1,\ q_j\geqslant 0$ and

$$\begin{aligned}\sum_{i=1}^{m}p_i\frac{\psi(s_i,x_i,\tilde{\beta},\tilde{\gamma})-\mu}{w(s_i,x_i,\hat{\theta})}&=0,\\ \sum_{j=m+1}^{N}q_j\frac{\psi(s_j,x_j,\tilde{\beta},\tilde{\gamma})-\mu}{1-w(s_j,x_j,\hat{\theta})}&=0.\end{aligned}\tag{9}$$

To simplify the notation, we write $U_i(\beta)=U(y_i,x_i,\beta)$, $\eta=(\beta,\gamma)$, $\tilde{\eta}=(\tilde{\beta},\tilde{\gamma})$, $\psi_i(\eta)=\psi(s_i,x_i,\eta)$ and $w_i(\theta)=w(s_i,x_i,\theta)$. By introducing Lagrange multipliers λ and ν and following standard EL derivations for general estimating equations (Qin and Lawless, 1994), the optimal values of p_i and q_j that maximize the above log-EL satisfy

$$p_i=\frac{1}{m}\frac{1}{1+\lambda^{\mathrm{T}}\{\psi_i(\tilde{\eta})-\mu\}/w_i(\hat{\theta})},\qquad i=1,\ldots,m,\tag{10}$$

$$q_j=\frac{1}{n}\frac{1}{1+\nu^{\mathrm{T}}\{\psi_j(\tilde{\eta})-\mu\}/\{1-w_j(\hat{\theta})\}},\qquad j=m+1,\ldots,N,\tag{11}$$

with constraints

$$\sum_{i=1}^{m}\frac{\{\psi_i(\tilde{\eta})-\mu\}/w_i(\hat{\theta})}{1+\lambda^{\mathrm{T}}\{\psi_i(\tilde{\eta})-\mu\}/w_i(\hat{\theta})}=0,\tag{12}$$

$$\sum_{j=m+1}^{N}\frac{\{\psi_j(\tilde{\eta})-\mu\}/\{1-w_j(\hat{\theta})\}}{1+\nu^{\mathrm{T}}\{\psi_j(\tilde{\eta})-\mu\}/\{1-w_j(\hat{\theta})\}}=0.\tag{13}$$

Substituting equations (10) and (11) back to the log-EL gives

$$l(\mu)=-\log\left(\sum_{i=1}^{m}\left[1+\frac{\lambda^{\mathrm{T}}\{\psi_i(\tilde{\eta})-\mu\}}{w_i(\hat{\theta})}\right]\right)-\log\left(\sum_{j=m+1}^{n}\left[1+\frac{\nu^{\mathrm{T}}\{\psi_j(\tilde{\eta})-\mu\}}{1-w_j(\hat{\theta})}\right]\right).$$

Differentiating $l(\mu)$ with respect to μ and equating to 0 leads to

$$-\sum_{i=1}^{m}\frac{\lambda/w_i(\hat{\theta})}{1+\lambda^{\mathrm{T}}\{\psi_i(\tilde{\eta})-\mu\}/w_i(\hat{\theta})}-\sum_{j=m+1}^{N}\frac{\nu/\{1-w_j(\hat{\theta})\}}{1+\nu^{\mathrm{T}}\{\psi_j(\tilde{\eta})-\mu\}/\{1-w_j(\hat{\theta})\}}=0.\tag{14}$$

Let $(\hat{\mu},\hat{\lambda},\hat{\nu})$ be the solution of equations (12)–(14). Substituting it into equations (10) and (11) gives the EL weights $\hat{p}_i$. These weights can be used to reweight the original estimating equation (6) such that $\hat{\beta}$ solves

$$m^{-1}\sum_{i=1}^{m}\frac{1}{1+\hat{\lambda}^{\mathrm{T}}\{\psi_i(\tilde{\eta})-\hat{\mu}\}/w_i(\hat{\theta})}\frac{U_i(\beta)}{w_i(\hat{\theta})}=0.\tag{15}$$

We shall show that $\hat{\beta}$ is more efficient than $\hat{\beta}_{\mathrm{W}}$ in problem (6).

A heuristic understanding of our method is the following. Using the Lagrange multiplier

$$\hat{\lambda}=\left\{\sum_{i=1}^{m}\left(\frac{\psi_i(\tilde{\eta})-\hat{\mu}}{w_i(\hat{\theta})}\right)^{\mathrm{T}}\left(\frac{\psi_i(\tilde{\eta})-\hat{\mu}}{w_i(\hat{\theta})}\right)\right\}^{-1}\sum_{i=1}^{m}\frac{\psi_i(\tilde{\eta})-\hat{\mu}}{w_i(\hat{\theta})}+o_p(N^{-1/2}),$$

the EL estimating equation (15) becomes

$$m^{-1}\sum_{i=1}^{m}\frac{U_i(\beta)}{w_i(\hat{\theta})}-m^{-1}\sum_{i=1}^{m}\frac{U_i(\beta)}{w_i(\hat{\theta})}\left(\frac{\psi_i(\tilde{\eta})-\hat{\mu}}{w_i(\hat{\theta})}\right)^{\mathrm{T}}\hat{\lambda}+o_p(N^{-1/2}).$$

Hence, the estimator proposed is asymptotically equivalent to the solution from an estimating equation that regresses the inversely weighted estimating equation $m^{-1}\Sigma_{i=1}^{m}U_i(\beta)/w_i(\hat{\theta})$ on $m^{-1}\Sigma_{i=1}^{m}\{\psi_i(\tilde{\eta})-\mu\}/w_i(\hat{\theta})$. As a result, the variance of the EL estimating equation is smaller than that of the inversely weighted estimating equation $m^{-1}\Sigma_{i=1}^{m}U_i(\beta)/w_i(\hat{\theta})$. This result is similar to the case when Y and X are two random variables; then $\mathrm{var}(Y-AX)=\mathrm{var}(Y)-A\,\mathrm{var}(X)A^{\mathrm{T}}\leqslant\mathrm{var}(Y)$, where $A=\mathrm{cov}(Y,X)\,\mathrm{var}(X)^{-1}$. In contrast, the estimating function of Robins *et al.* (1995) is a difference between the inversely weighted estimating equation and $\Sigma_{i=1}^{m}\psi_i(\tilde{\eta})\{\delta_i-w_i(\hat{\theta})\}/w_i(\hat{\theta})$. It is known in survey sampling (Cochran, 1977; Cassel *et al.*, 1976) that difference estimation is not as efficient as regression estimation.

Using the EL formulation, the information about β is extracted by using $\psi(s,x,\tilde{\beta},\tilde{\gamma})$, where $\tilde{\beta}$ and $\tilde{\gamma}$ can be interpreted as summary statistics based on $\{(y_i,s_i,x_i)\}_{i=1}^{m}$ and $\{(s_j,x_j)\}_{j=1}^{N}$. When making inferences, we must determine ψ, $\tilde{\beta}$ and $\tilde{\gamma}$. Let ψ be an estimate of ψ^* by using equation (5) and $\tilde{\gamma}$ be an estimate that is based on that model. We shall show that the method works as long as $\tilde{\gamma}$ converges in mean square to some γ_0 within the parameter space of γ, i.e. there is a positive constant c_0 such that $E[\tilde{\gamma}-\gamma_0]^2\leqslant c_0n^{-1}$.

After finding $\hat{\beta}$, we could replace the initial estimate $\tilde{\beta}$ by $\hat{\beta}$ and repeat the estimation process. However, our analysis shows that the choice of the initial estimates $\tilde{\beta}$ and $\tilde{\gamma}$ have no influence on the asymptotic efficiency.

Our proposed estimator $\hat{\beta}$ is consistent as long as w is correctly specified. To appreciate this, we note that $\Sigma_{i=1}^{m}\hat{p}_i\,I\{(y_i,s_i,x_i)\leqslant t\}$ is a consistent estimate of $F(y,s,x|D=1)$ and

$$E\left[\frac{U(y_i,x_i,\beta)}{w(s_i,x_i,\theta)}|D=1\right]=0.$$

Since $\hat{\beta}$ solves problem (15) which can be regarded as a sample version of the above population equation, $\hat{\theta}$ is asymptotically unbiased and its variance converges to 0 as $\min(m,n)\to\infty$. Hence, $\hat{\beta}$ is consistent for β.

3. Main results

Let β_0, γ_0 and θ_0 be the true parameter values of β, γ and θ respectively. Define $\eta_0=(\beta_0,\gamma_0)$ and write $U_0=^{\mathrm{d}}U_i(\beta_0)$, $\psi_0=^{\mathrm{d}}\psi_i(\eta_0)$, $\mu_0=E[\psi_0]$ and $w_0=^{\mathrm{d}}w_i(\theta_0)$ where $=^{\mathrm{d}}$ denotes equivalence in distributions. Furthermore, let

$$A=E\left[\frac{U_0(\psi_0-\mu_0)^{\mathrm{T}}}{w_0}\right]E^{-1}\left[\frac{(\psi_0-\mu_0)(\psi_0-\mu_0)^{\mathrm{T}}}{w_0(1-w_0)}\right],$$

$$R=-E^{-1}\left[\frac{\partial U_0}{\partial\beta}\right](I_p,-A,A),$$

$$\zeta = \left(-\frac{U_0^{\mathrm{T}}}{w_0}, -\frac{(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0}, \frac{(\psi_0 - \mu_0)^{\mathrm{T}}}{1 - w_0} \right),$$

$$\Lambda_\theta = E\left[\frac{1}{w_0(1-w_0)} \frac{\partial w_0}{\partial \theta} \frac{\partial w_0^{\mathrm{T}}}{\partial \theta} \right];$$

the last quantity defines the asymptotic variance of the maximum likelihood estimator $\hat{\theta}$ based on the binomial likelihood (1).

Theorem 1. Under conditions 1–4 that are given in Appendix A,

$$N^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{\mathrm{d}} N(0, \Sigma_\beta^{(0)} - \Sigma_\beta^{(1)} - \Sigma_\beta^{(2)}), \tag{16}$$

where

$$\Sigma_\beta^{(0)} = E^{-1}\left[\frac{\partial U_0}{\partial \beta} \right] E\left[\frac{U_0 U_0^{\mathrm{T}}}{w_0} \right] E^{-1}\left[\frac{\partial U_0^{\mathrm{T}}}{\partial \beta} \right], \tag{17}$$

$$\Sigma_\beta^{(1)} = R\, E\left[\zeta \frac{\partial w_0^{\mathrm{T}}}{\partial \theta} \right] \Lambda_\theta^{-1}\, E\left[\frac{\partial w_0}{\partial \theta} \zeta^{\mathrm{T}} \right] R^{\mathrm{T}}, \tag{18}$$

$$\Sigma_\beta^{(2)} = E^{-1}\left[\frac{\partial U_0}{\partial \beta} \right] E\left[\frac{U_0(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0} \right] E^{-1}\left[\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0(1-w_0)} \right] E\left[\frac{(\psi_0 - \mu_0) U_0^{\mathrm{T}}}{w_0} \right] E^{-1}\left[\frac{\partial U_0^{\mathrm{T}}}{\partial \beta} \right].$$

We note that

(a) $\Sigma_\beta^{(0)}$ is the covariance matrix of $\hat{\beta}_{\mathrm{W}}$, the inverse weighted estimator by the true propensity score w_0 and
(b) both $\Sigma_\beta^{(1)}$ and $\Sigma_\beta^{(2)}$ are non-negative definite.

Hence, the covariance matrix $\Sigma_\beta^{(0)}$ can be reduced twice: once by $\Sigma_\beta^{(1)}$ and once by $\Sigma_\beta^{(2)}$. Therefore, the EL estimator proposed is more efficient than $\hat{\beta}_{\mathrm{W}}$, when the true propensity score is used to weight the estimating equation on the basis of the complete observations, unless $\Sigma_\beta^{(1)}$ and $\Sigma_\beta^{(2)}$ are zero matrices simultaneously.

The variance reduction that is offered by $\Sigma_\beta^{(2)}$ is a result of having the second constraint in expression (9) based on the observations with missing Y-values. If this constraint is removed from expression (9), $\Sigma_\beta^{(2)}$ will be 0. Therefore, it is worthwhile both to carry out weighting by propensity score and to have an extra estimating equation that is based on the covariate and the surrogate from the part of the sample with missing outcome. The variance reduction that is offered by $\Sigma_\beta^{(1)}$ is partly due to the use of $\hat{\theta}$ rather than the true parameter θ_0, as can be seen by the involvement of Λ_θ^{-1}. This reflects a known statistical advantage with estimation over the true propensity score (see, for example, Wooldridge (2004)).

We note that $\Sigma_\beta^{(2)}$ is essentially a weighted 'correlation' between U and ψ. The higher the value of this correlation, the larger the variance reduction is. This observation suggests that we find a function ψ that is highly correlated with U. The optimal choice for ψ is $(1-w)\ E[U(y, x, \beta)|s, x] = (1-w)\psi^*$. This choice can be justified by noting that

$$E\left[\frac{U_0(\psi^*-\mu_0)^{\mathrm{T}}}{w_0}\right]E^{-1}\left[\frac{(\psi^*-\mu_0)(\psi^*-\mu_0)^{\mathrm{T}}}{w_0(1-w_0)}\right]E\left[\frac{(\psi^*-\mu_0)U_0^{\mathrm{T}}}{w_0}\right]=E\left[\frac{1-w_0}{w_0}U_0E[U_0^{\mathrm{T}}|s,x]\right].$$

Hence

$$\Sigma_\beta^{(0)}-\Sigma_\beta^{(2)}=E^{-1}\left[\frac{\partial U_0}{\partial\beta}\right]\left(E\left[\frac{U_0U_0^{\mathrm{T}}}{w_0}\right]-E\left[\frac{1-w_0}{w_0}U_0E[U_0^{\mathrm{T}}|s,x]\right]\right)E^{-1}\left[\frac{\partial U_0^{\mathrm{T}}}{\partial\beta}\right], \tag{19}$$

which is the variance lower bound when the propensity score is known for a given U (see Robins *et al.* (1995) and Chen *et al.* (2005)). Owing to the different set-ups from previous works, our optimal choice of ψ has an extra factor of $1-w_0$.

We now give the properties of the estimator $\hat{\beta}_{\mathrm{RRZ}}$ that was proposed by Robins *et al.* (1995).

Theorem 2. Under conditions 1–4 that are given in Appendix A,

$$N^{1/2}(\hat{\beta}_{\mathrm{RRZ}}-\beta_0)\xrightarrow{\mathrm{d}}N(0,\Sigma_\beta^{(0)}-\tilde{\Sigma}_\beta^{(1)}-\tilde{\Sigma}_\beta^{(2)}), \tag{20}$$

where $\Sigma_\beta^{(0)}$ is defined in theorem 1,

$$\tilde{\Sigma}_\beta^{(1)}=E^{-1}\left[\frac{\partial U_0}{\partial\beta}\right]E\left[\frac{U_0-\psi_0}{w_0}\frac{\partial w_0^{\mathrm{T}}}{\partial\theta}\right]\Lambda_\theta^{-1}E\left[\frac{\partial w_0}{\partial\theta}\frac{(U_0-\psi_0)^{\mathrm{T}}}{w_0}\right]E^{-1}\left[\frac{\partial U_0^{\mathrm{T}}}{\partial\beta}\right]$$

and

$$\tilde{\Sigma}_\beta^{(2)}=E^{-1}\left[\frac{\partial U_0}{\partial\beta}\right]E\left[(1-w_0)\left(\frac{U_0\psi_0^{\mathrm{T}}}{w_0}+\frac{\psi_0U_0^{\mathrm{T}}}{w_0}-\frac{\psi_0\psi_0^{\mathrm{T}}}{w_0}\right)\right]E^{-1}\left[\frac{\partial U_0^{\mathrm{T}}}{\partial\beta}\right].$$

The estimator $\hat{\beta}_{\mathrm{RRZ}}$ reaches the semiparametric efficiency bound if $\psi=E[U(y,x,\beta)|s,x]$ and w is correctly specified. In this case, the asymptotic variance that is given by $\Sigma^{(0)}-\Sigma^{(2)}$ for the proposed estimator $\hat{\beta}$ is the same with $\tilde{\Sigma}^{(0)}-\tilde{\Sigma}^{(2)}$ of $\hat{\beta}_{\mathrm{RRZ}}$ and equals the semiparametric efficiency bound that is given in equation (19). However, when $\psi\neq E[U(y,x,\beta)|s,x]$, which is a likely scenario in practice, the efficiency of $\hat{\beta}_{\mathrm{RRZ}}$ can be severely compromised, even if the propensity function w is correctly specified. The reason is, whereas $\tilde{\Sigma}_\beta^{(1)}$ is always non-negative definite (indicating a gain in efficiency), there is no guarantee that $\tilde{\Sigma}_\beta^{(2)}$ is non-negative definite. Indeed, for some choices of ψ, $\hat{\beta}_{\mathrm{RRZ}}$ can be less efficient than the weighted estimator $\hat{\beta}_{\mathrm{W}}$ that solves problem (6); some examples of such cases are given in the next section. Although we are not suggesting that $\hat{\beta}$ is always better than $\hat{\beta}_{\mathrm{RRZ}}$, it is true that $\hat{\beta}$ always gains in efficiency over $\hat{\beta}_{\mathrm{W}}$, as long as $\Sigma_\beta^{(2)}$ is not 0, whereas no such guarantee can be said about $\hat{\beta}_{\mathrm{RRZ}}$.

4. Numerical study

We compared the estimator proposed with three other estimators in a simulation study:

(a) the maximum likelihood estimator $\hat{\beta}_{\mathrm{C}}$ assuming that all data are observed (this estimator is not feasible in practice; however, it sets a benchmark on how much information is contained in the sample if there were no missing data);
(b) the weighted estimator $\hat{\beta}_{\mathrm{W}}$ by solving problem (6) using only the complete observations (this is also the initial estimator $\tilde{\beta}$ that is used in obtaining the EL weights);
(c) the estimator $\hat{\beta}_{\mathrm{RRZ}}$.

Throughout the simulation study, the following model was used for generating missingness:

$$1-w(s,x,\theta)=P(\delta=0|y,s,x)=P(\delta=0|s,x)=\frac{1}{1+\exp(\theta_1+\theta_2 s+\theta_3 x)}, \tag{21}$$

for $\theta=(\theta_1,\theta_2,\theta_3)$. Two models for (Y,S,X) were studied. In model 1, Y and S were both normally distributed with means and variances respectively

$$E[Y|X]=\beta_1+\beta_2 X \quad \text{and} \quad E[S|Y,X]=1+2Y+X;$$
$$\text{var}(Y|X)=\text{var}(S|Y,X)=1$$

where $X\sim N(0,1)$. The estimating function corresponding to (Y,X) was

$$U(y,x)=\begin{pmatrix}1\\ x\end{pmatrix}(y-\beta_1-\beta_2 x).$$

Estimates of $E[U(y,x,\beta)|s,x]$ are required in the estimation process to obtain $\hat{\beta}_{\text{RRZ}}$ and $\hat{\beta}$. For this model, we used

$$\psi_{\text{RRZ}}(s_i,x_i,\beta)=E[U(y,x,\beta)|s,x]=\begin{pmatrix}1\\ x\end{pmatrix}(\gamma_1+\gamma_2 s+\gamma_3 x-\beta_1-\beta_2 x)$$

for $\hat{\beta}_{\text{RRZ}}$ and $\psi(s,x)=\{1-w(s,x,\theta)\}\,\psi_{\text{RRZ}}(s_i,x_i,\beta)$ for $\hat{\beta}$. The initial estimate for $\gamma=(\gamma_1,\gamma_2,\gamma_3)$ was obtained by fitting a linear regression

$$E[Y]=\gamma_1+\gamma_2 S+\gamma_3 X. \tag{22}$$

As mentioned in the Section 2, equation (22) does not need to be correct. The goal is to recover as much as possible the loss of information in the missing values of Y by using S and X.

In model 2, the outcome Y was a binary variable with

$$P(Y=1|X)=\frac{\exp(\beta_1+\beta_2 X)}{1+\exp(\beta_1+\beta_2 X)},$$

and S, conditioned on X and Y, was normal with unit variance and mean

$$E[S|Y,X]=1+2Y+X,$$

and $X\sim N(0,1)$. The estimating equations were

$$U(y,x)=\begin{pmatrix}1\\ x\end{pmatrix}\left\{y-\frac{\exp(\beta_1+\beta_2 x)}{1+\exp(\beta_1+\beta_2 x)}\right\},$$
$$\psi_{\text{RRZ}}(s,x)=\begin{pmatrix}1\\ x\end{pmatrix}\left\{\frac{\exp(\gamma_1+\gamma_2 s+\gamma_3 x)}{1+\exp(\gamma_1+\gamma_2 s+\gamma_3 x)}-\frac{\exp(\beta_1+\beta_2 x)}{1+\exp(\beta_1+\beta_2 x)}\right\},$$
$$\psi(s,x)=\{1-w(s,x,\theta)\}\,\psi_{\text{RRZ}}(s,x),$$

where $\gamma=(\gamma_1,\gamma_2,\gamma_3)$ was estimated by fitting a logistic regression based on the data with complete observations $\{y_i,s_i,x_i\}_{i=1}^m$.

For models 1 and 2, 2000 simulations were carried out for combinations of $\beta=(1,1)$ and $\beta=(1,2)$ and $\theta=(-1,0,0),(-1,0.2,0.2),(-1,0.35,0.35),(-1,0.5,0.5)$ in the missing probability function, with $N=1000$ in each simulation. The choices $\theta=(-1,0,0)$, $(-1,0.2,0.2)$, $(-1,0.35,0.35)$, $(-1,0.5,0.5)$ induced respectively approximately 75%, 60%, 47% and 45% missing outcomes in the data.

We considered two methods for variance estimation in each method:

(a) the asymptotic variance formulae in Section 3 and
(b) the bootstrap method.

Under data MCAR ($\theta = (-1, 0, 0)$) or weakly MAR ($\theta = (-1, 0.2, 0.2)$), both methods give similar variance estimates. However, under data strongly MAR ($\theta = (-1, 0.35, 0.35)$ and $\theta = (-1, 0.5, 0.5)$), the bootstrap method gives more reliable variance estimates. The better performance of the bootstrap method is because the asymptotic variance formula involves the quantity $\Sigma_{i=1}^{n}(d_i/w_i)^2\psi_i^{\mathrm{T}}\psi_i/n$, which can be unduly affected by values of w_i that are close to 0 or 1, when $\theta = (-1, 0.35, 0.35)$ or $\theta = (-1, 0.5, 0.5)$.

The simulation results are reported in Tables 1 and 2 for the case $\beta = (1, 2)$. The results for $\beta = (1, 1)$ follow the same pattern and hence are not reported. For each method, the first row is the mean and the variance based on the 2000 replications. The second row is the observed coverage for 95% nominal confidence interval and the bootstrap variance estimate. Table 1 shows that, when the outcome is data MCAR ($\theta = (-1, 0, 0)$), $\hat{\beta}_{\mathrm{RRZ}}$ and the estimator that is proposed in this paper, $\hat{\beta}$, are almost equivalent. However, when $\theta = (-1, 0.35, 0.35)$ and $\theta = (-1, 0.5, 0.5)$, then missingness depends strongly on (S, X) and in those cases $\hat{\beta}$ outperforms $\hat{\beta}_{\mathrm{RRZ}}$ and $\hat{\beta}_{\mathrm{W}}$. For model 2, $\hat{\beta}_{\mathrm{RRZ}}$ and $\hat{\beta}$ are much better than $\hat{\beta}_{\mathrm{W}}$ when the outcome is data MCAR. However, their gains in efficiency are reduced when the selection bias in missingness of the outcome variable is large, i.e. $\theta = (-1, 0.35, 0.35)$ and $\theta = (-1, 0.5, 0.5)$. Interestingly, in those cases, comparing with the unattainable estimator $\hat{\beta}_{\mathrm{C}}$ that is based on the full sample, missing data did not lead to much loss of information. The loss of efficiency of all three estimators when compared with $\hat{\beta}_{\mathrm{C}}$ is less severe in model 2 than the corresponding cases in model 1. Among the three estimators,

Table 1. Means (and variances in parentheses) of various estimators based on 2000 simulations with sample size $N = 1000$ each and bootstrap resample size 200†

Method	*Results for the following values of θ:*			
	$\theta = (-1, 0, 0)$	$\theta = (-1, 0.2, 0.2)$	$\theta = (-1, 0.35, 0.35)$	$\theta = (-1, 0.5, 0.5)$
$\hat{\beta}_{\mathrm{C1}}$	0.99880 (0.00095)	1.00085 (0.00099)	1.00104 (0.00100)	0.99962 (0.00104)
	94.75% (0.00099)	94.55% (0.00100)	94.50% (0.00100)	94.55% (0.00100)
$\hat{\beta}_{\mathrm{C2}}$	1.99870 (0.00097)	1.99954 (0.00095)	1.99997 (0.00097)	2.00120 (0.00106)
	94.65% (0.00100)	94.85% (0.00100)	94.15% (0.00100)	93.70% (0.00099)
$\hat{\beta}_{\mathrm{W1}}$	0.99934 (0.00160)	1.00114 (0.00318)	1.01477 (0.00642)	1.02475 (0.01338)
	94.15% (0.00159)	92.90% (0.00289)	89.75% (0.00522)	82.85% (0.00828)
$\hat{\beta}_{\mathrm{W2}}$	1.99736 (0.00372)	1.99422 (0.00746)	1.97474 (0.01444)	1.96273 (0.02547)
	94.75% (0.00382)	90.45% (0.00626)	84.95% (0.01009)	78.65% (0.01378)
$\hat{\beta}_{\mathrm{RRZ1}}$	0.99936 (0.00158)	1.00114 (0.00175)	1.00173 (0.00309)	0.99905 (0.04684)
	94.15% (0.00154)	93.95% (0.00168)	93.60% (0.00312)	94.10% (0.04397)
$\hat{\beta}_{\mathrm{RRZ2}}$	1.99758 (0.00154)	1.99981 (0.00272)	1.99867 (0.00955)	2.00236 (0.35024)
	94.35% (0.00156)	92.20% (0.00263)	92.80% (0.01004)	93.00% (0.32538)
$\hat{\beta}_1$	0.99931 (0.00158)	1.00217 (0.00180)	1.00487 (0.00293)	1.00077 (0.00567)
	94.35% (0.00155)	93.95% (0.00175)	94.85% (0.00291)	94.65% (0.00474)
$\hat{\beta}_2$	1.99747 (0.00156)	1.99872 (0.00267)	1.99235 (0.00536)	1.99500 (0.00961)
	94.25% (0.00159)	93.70% (0.00252)	94.65% (0.00494)	95.00% (0.00725)

†The second rows are the observed coverage for a 95% nominal confidence interval and bootstrap estimation of variance. The missing probability function is $P(\delta = 1 | S = s, X = x) = \exp(\theta_1 + \theta_2 s + \theta_3 x)/\{1 + \exp(\theta_1 + \theta_2 s + \theta_3 x)\}$; $Y \sim N(\beta_1 + \beta_2 X, 1)$, where $(\beta_1, \beta_2) = (1, 2)$; $S \sim N(1 + 2Y + X, 1)$.

Table 2. Means (and variances in parentheses) of various estimators based on 2000 simulations with sample size $N = 1000$ each and bootstrap resample size 200†

Method	*Results for the following values of* θ*:*			
	$\theta = (-1, 0, 0)$	$\theta = (-1, 0.2, 0.2)$	$\theta = (-1, 0.35, 0.35)$	$\theta = (-1, 0.5, 0.5)$
$\hat{\beta}_{C1}$	1.00334 (0.00934)	1.00075 (0.00898)	1.00315 (0.00843)	0.99887 (0.00877)
	93.60% (0.00893)	94.65% (0.00889)	94.95% (0.00888)	94.40% (0.00882)
$\hat{\beta}_{C2}$	2.00818 (0.01943)	2.01027 (0.00181)	2.08768 (0.01727)	2.00296 (0.01758)
	93.70% (0.01812)	94.0% (0.01816)	94.40% (0.01808)	94.05% (0.01803)
$\hat{\beta}_{W1}$	1.01157 (0.02583)	1.01425 (0.02112)	1.01159 (0.01893)	1.00393 (0.01757)
	94.90% (0.02852)	95.10% (0.02275)	94.75% (0.01990)	93.90% (0.01757)
$\hat{\beta}_{W2}$	2.02609 (0.07416)	2.03483 (0.06081)	2.02950 (0.05938)	2.02701 (0.05529)
	94.0% (0.07602)	93.35% (0.06384)	93.50% (0.05907)	93.40% (0.05714)
$\hat{\beta}_{RRZ1}$	1.00794 (0.02167)	1.00726 (0.01730)	1.00667 (0.01554)	1.00177 (0.01458)
	94.40% (0.02308)	4.85% (0.01823)	94.70% (0.01601)	93.10% (0.01446)
$\hat{\beta}_{RRZ2}$	2.02529 (0.05019)	2.02167 (0.04314)	2.02264 (0.04278)	2.02543 (0.03918)
	93.60% (0.05160)	93.25% (0.04399)	92.40% (0.04063)	92.20% (0.03891)
$\hat{\beta}_1$	1.00795 (0.02184)	1.00796 (0.01729)	1.00694 (0.01560)	1.00297 (0.01470)
	94.70% (0.02334)	94.60% (0.01838)	94.80% (0.01612)	93.30% (0.01463)
$\hat{\beta}_2$	2.02466 (0.05050)	2.02238 (0.04350)	2.02247 (0.04327)	2.02487 (0.04010)
	93.10% (0.05243)	93.50% (0.04468)	92.70% (0.04121)	92.40% (0.03977)

†The second rows are the observed coverage for a 95% nominal confidence interval and bootstrap estimation of variance. The missing probability function is $P(\delta = 1|S = s, X = x) = \exp(\theta_1 + \theta_2 s + \theta_3 x)/\{1 + \exp(\theta_1 + \theta_2 s + \theta_3 x)\}$; $P(Y = 1|X) = \{\exp(\beta_1 + \beta_2 X)\}/\{1 + \exp(\beta_1 + \beta_2 X)\}$, where $(\beta_1, \beta_2) = (1, 2)$; $S \sim N(1 + 2Y + X, 1)$.

the estimator that is proposed in this paper is the best. In certain cases, the relative efficiency of $\hat{\beta}_{RRZ}$ to $\hat{\beta}$ is less than 50%.

To illustrate the results of theorems 1 and 2 further, we compared the asymptotic relative efficiencies between the estimators in a modest set-up. Two models were used. The first model is a linear model that is similar to model 1 in the simulation study, except that $E[S|Y, X] = 2Y$ if $Y \geqslant 0$ and $E[S|Y, X] = Y$ if $Y < 0$ and $\theta \equiv (\theta_1, \theta_2, \theta_3) = (-2, \zeta, 0.5)$ in the missing function w, with ζ allowed to vary from 0 to 0.5. The second model is binary with Y as in model 2 in the simulation study and S is also binary with

$$P(S = 1|X) = \frac{\exp\{\beta_1 + \beta_2(X + \zeta)\}}{1 + \exp\{\beta_1 + \beta_2(X + \zeta)\}},$$

so ζ is a disturbance that makes S an imperfect surrogate. The value of ζ varied from -1.5 to 0 and $\theta = (-3, 3, 0)$ in w. Therefore, non-zero values of ζ in either model create situations where it would not be possible to find a simple ψ-function that is the same as $\psi_0(s, x, \beta) \equiv E[U(y, x, \beta)|s, x]$ under data MAR. In both models, we assumed $(\beta_1, \beta_2) = (1, 2)$ and we used the asymptotic formulae in theorems 1 and 2 to calculate

$$\mathrm{ARE}(\hat{\beta}, \hat{\beta}_W) = \mathrm{var}(\hat{\beta}_W)/\mathrm{var}(\hat{\beta}),$$
$$\mathrm{ARE}(\hat{\beta}, \hat{\beta}_{RRZ}) = \mathrm{var}(\hat{\beta}_{RRZ})/\mathrm{var}(\hat{\beta}),$$

for estimating β_1 and β_2. The results are given in Figs 1(a)–1(d). They show that $\hat{\beta}$ is always as efficient as $\hat{\beta}_W$ and $\hat{\beta}_{RRZ}$ in all the scenarios that were studied. The most noticeable features of these results is the poor performance of the method of Robins *et al.* (1995) under data MAR,

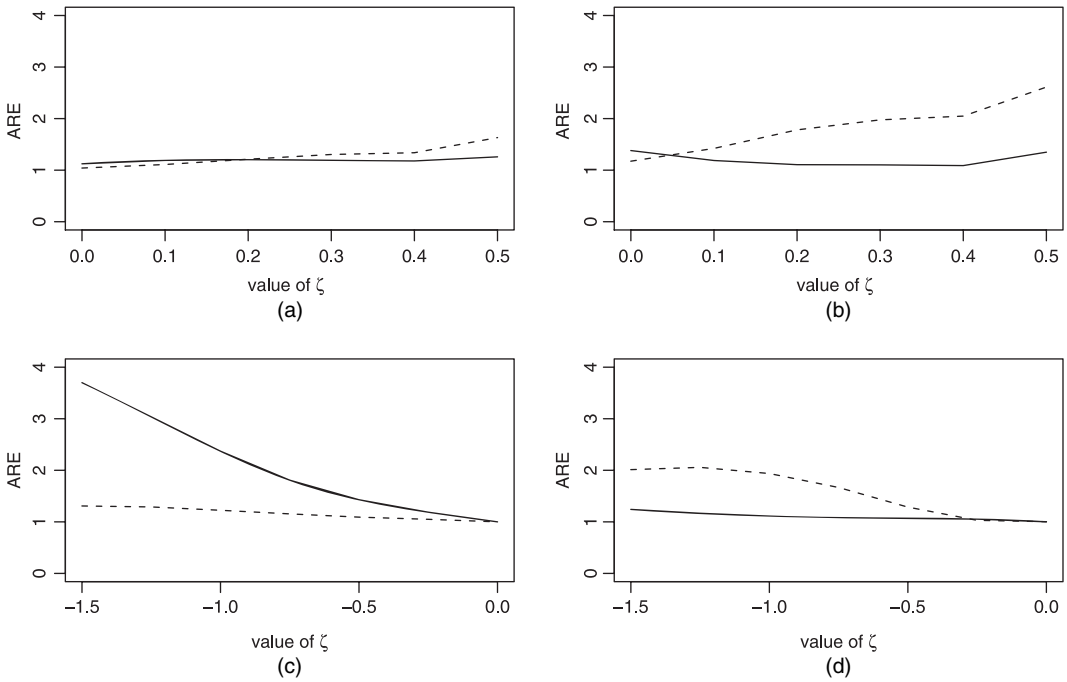

Fig. 1. Asymptotic relative efficiency (ARE) between three estimators (the missing probability function is $P(\delta = 1|S = s, X = x) = \exp(\theta_1 + \theta_2 s + \theta_3 x)/\{1 + \exp(\theta_1 + \theta_2 s + \theta_3 x)\}$; ———, ARE$(\hat{\beta}, \hat{\beta}_{\mathrm{W}})$; -------, ARE$(\hat{\beta}, \hat{\beta}_{\mathrm{RRZ}})$; for (a) and (b), $Y \sim N(\beta_1 + \beta_2 X, 1)$, $S \sim N(2Y, 1)$ if $Y \geqslant 0$ and $S \sim N(Y, 1)$ if $Y < 0$ and $\theta \equiv (\theta_1, \theta_2, \theta_3) = (-2, \zeta, 0.5)$; for (c) and (d), $P(Y = 1|X) = \exp(\beta_1 + \beta_2 X)/\{1 + \exp(\beta_1 + \beta_2 X)\}$, $P(S = 1|X) = \exp\{\beta_1 + \beta_2(X + \zeta)\}/[1 + \exp\{\beta_1 + \beta_2(X + \zeta)\}]$ and $\theta = (-3, 3, 0)$): (a) linear model, β_1; (b) linear model, β_2; (c) binary model, β_1; (d) binary model, β_2

when ζ is non-zero (Figs 1(b) and 1(d)). The poor performance of the method of Robins *et al.* (1995) results because ψ is very different from $E[U(y, x, \beta)|s, x]$ in those cases. This is a point that was made at the end of Section 3 that there is no guarantee that their estimator will always be better than the inversely weighted estimator. In both models, the disadvantage of using $\hat{\beta}_{\mathrm{RRZ}}$ is less pronounced for estimating β_1 than for β_2. This is because the set-ups of the models changed the distribution of X through S and, under data MAR, the changes affect β_2 more because it is the coefficient that is associated with X.

5. Application to election data

We applied the method proposed to a set of data from the National Election Study (Warren *et al.*, 1999; Lee and Kang, 2002; Lee, 2005). The US presidential election follows an electoral college system, not the usual popular vote. However, on two occasions (including the one between Bush and Gore in 2000), a candidate lost the election despite winning the popular vote. We assumed that the election followed a popular vote system. As argued in Lee (2005), this approach is reasonable for illustration for two reasons:

(a) since the election results by using the two systems were very close, the statistical conclusions should be similar by using either system;
(b) the sample size is not sufficiently large at the state level, which would be required if the electoral college system is used.

Table 3. Cross-tabulation of surrogate outcome (predicted voting choice) and true outcome (actual voting choice) for those who voted for Clinton or Bob Dole

Surrogate outcome	*True outcome*		*Total*
	Clinton	*Dole*	
Clinton	574	17	591
Dole	23	404	427
Total	597	421	1018

The data came from two surveys that were conducted before and after the election. There were three candidates: Clinton, Dole and Perot. We focused on the two main candidates: Clinton and Dole. A striking feature of the data set is the large proportion of observations (33%) with missing outcome, as represented by those who did not vote.

We used the responses from three questions to construct the surrogate outcome S. In the post-election survey, each non-voter was asked the question 'Who did you prefer (as the president)?'. If the answer is Clinton or Dole, then it is used as the surrogate outcome. If no answer was given, then we compared the average ratings (on a scale of 0–100) of Clinton and Dole by the non-voter in the pre- and post-election survey and took the candidate with the higher average rating as the surrogate outcome. If the average ratings were tied, then we looked at the political party trait of the non-voter. By carrying out this procedure, we arrived at $N = 1486$ respondents who either have a surrogate or the true outcome and with complete covariate information.

Voting patterns for the data that were available for analysis ($N = 1486$) are as follows: no vote, 474 or 32%; Clinton, 586 or 39%; Dole, 426 or 29%. Using the method that was described in the previous paragraph, out of the 1486 respondents, 929 have Clinton as the surrogate outcome and 557 have Dole as the surrogate outcome. One way to assess the quality of this surrogate is to compare its value with the true outcome for those who voted. The comparison is summarized in Table 3, which shows that the association between the true outcome and the surrogate outcome is highly significant ($p < 0.001$ by using a χ^2-test).

Alvarez and Nagler (1998) discussed several questions related to the National Election Study that may be of interest. We focused on the question of how voter's perception of the economy influenced the election outcome. In the pre-election survey, every respondent was asked whether the economy of the country had become better, stayed about the same or grown worse in the year leading up to the election. The answers from the respondents, along with the values of the true and surrogate outcome, are summarized in Table 4. Thus, voter's perception represents the X-variable in the model.

To model the probability of a missing outcome, we turned to previous works that studied voter turn-outs in US presidential elections (Riker and Ordshook, 1968; Filer and Kenney, 1980; Sanders, 2001). Sanders (2001) used the data set in this paper to model the probability of turn-out (Table 1 in Sanders (2001)) with the following variables: Age, Income, Race, Gender, Education (High school *versus* College *versus* others), Political Awareness and Efficacy (of the voter), Ideological and Character difference (between the voter and the candidate), Ideological and Character certainty (of the candidates by the voter), whether the voter was contacted (mobilized) by a political party before the election and whether the voter cared about the election.

Table 4. Cross-tabulation of surrogate outcome (predicted voting choice), true outcome (actual voting choice) and the covariate (perception on the economy) for all respondents, $N = 1486$ (excluding those who did not indicate perception of the economy)

True outcome	*Surrogate outcome*	*Results for the following perceptions of the economy:*		
		Better	*Same*	*Worse*
No vote	Clinton	117	168	52
	Dole	34	57	46
Clinton	Clinton	338	187	44
	Dole	6	9	2
Dole	Clinton	11	10	2
	Dole	94	222	87

These variables are the vector of Z that is discussed in Section 1. In addition to Z, we added S and X and modelled w by using a logistic regression

$$1 - w(s, x, z, \theta) = \frac{1}{1 + \exp(\theta_1 + \theta_2 s + \theta_3 x + \theta_4^{\mathrm{T}} z)}. \tag{23}$$

This example highlights the different roles that are played by Z and S. Whereas S is a surrogate for voting preference for those who did not vote, Z is used to model the act of voting. Both variables are necessary for combining the information from the voters and the non-voters to draw valid inferences.

A binary logistic regression was used to model the relationship between the true outcome (the choice of the President) and a single covariate (the perceived state of the economy). Let Y be the true outcome and $Y = 1$ represent 'Clinton is the choice' and $Y = 0$ represent 'Dole is the choice'; let X be the covariate and $X = -1, 0, 1$ if the respondent thought that the nation's economy had 'grown worse', 'stayed about the same' and 'become better' respectively. The model can be written as

$$P(Y = 1|X) = \frac{\exp(\beta_1 + \beta_2 X)}{1 + \exp(\beta_1 + \beta_2 X)}.$$

The surrogate outcome S is also a binary variable with $S = 1$ representing Clinton is the choice and $S = 0$ representing Dole is the choice. We assumed

$$\psi(s, x) = \{1 - w(s, x, z, \hat{\theta})\} \begin{pmatrix} 1 \\ x \end{pmatrix} \left\{ \frac{\exp(\gamma_1 + \gamma_2 s + \gamma_3 x)}{1 + \exp(\gamma_1 + \gamma_2 s + \gamma_3 x)} - \frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)} \right\},$$

where $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ is estimated by fitting a 'working' logistic regression based on respondents who voted and $\hat{\theta}$ was modelled as in equation (23).

The three methods that are considered in this paper were used to analyse the data. Table 5 gives the parameter estimates and the corresponding variances based on the bootstrap method and the asymptotic formulae in theorems 1 and 2. All methods show strong evidence ($\hat{\beta}_2/\mathrm{SE}(\hat{\beta}_2) \gg 0$)

Table 5. National Election Study data using three methods of analysis

Method	*Parameter estimates (variance†, variance‡) for the following parmeters:*	
	β_1	β_2
Weighted estimator	0.2989 (0.00839, 0.00642)	0.8004 (0.01682, 0.01578)
Robins *et al.* (1995)	0.2223 (0.00386, 0.00485)	0.8792 (0.00818, 0.01006)
Proposed estimator	0.2950 (0.00399, 0.00442)	0.7867 (0.00786, 0.00825)

†Variance estimate by using the asymptotic formulae in theorems 1 and 2.
‡Variance estimate by using 1000 bootstrap samples.

that voter's perception of the economy had a significant influence on voting behaviour. Using the weighted estimator, the odds ratio of voting for Clinton is

$$\frac{\exp(0.2989+0.8004)/\{1+\exp(0.2989+0.8004)\}}{\exp(0.2989-0.8004)/\{1+\exp(0.2989-0.8004)\}}=1.98$$

for someone who views the economy favourably against someone who views the economy negatively. The conclusions are similar by using the other two methods. Using either the method of Robins *et al.* (1995) and the method that is proposed in this paper, there are significant gains in efficiency over the weighted estimator. The bootstrap and the corresponding asymptotic formulae variances estimates are similar, as will be the case in most practical situations.

6. Concluding remarks

Surrogate outcome has become a popular means to enhance estimation efficiency when the true outcome is missing. This paper proposed a procedure that improves the efficiency of estimation in the surrogate outcome problem via Owen's (1990) EL. Two different decompositions of the observed likelihood were suggested. The first decomposition uses the binomial likelihood conditional on the observations with complete information on (Y, X, S) in equation (1). The parameter θ in the propensity function w can be easily estimated by maximizing the binomial likelihood. The second decomposition is conditional on the missingness status. As a result, two ELs can be constructed by linking the unbiased estimating equations. It is well known that the best estimating equation is not available in general, but simpler forms exist for the missing response data; see Chen and Breslow (2004) and Yu and Nan (2006). In practice, $U(Y, X, \beta)$ can be regressed on S and X by using a working non-linear model or a general additive model. We corrected the possible bias in the working model by using estimating equations (8) and then combined them by using EL. The resulting estimate has attractive theoretical properties as well as good finite sample performance. The method is especially useful when there is little information on the conditional density of S given (Y, X), since in that case the optimal conditional estimating function that is needed in methods such as the estimator of Robins *et al.* (1995) is not available. With some modifications, the method proposed may be generalized to other missing data situations, e.g. in measurement error problems.

Acknowledgements

We thank the Associate Editor and two referees for constructive comments and suggestions. Chen's research was supported by National Science Foundation grants SES-0518904 and DMS 06-04563. Leung's research was supported by the Singapore Management University Research Center. We thank Professor Myoung-Jae Lee of Korea University for providing us with the election data and for his valuable input regarding the data.

Appendix A

The conditions that are needed to establish theorems 1 and 2 are as follows.

Condition 1. The propensity score $w_i(\theta)$ is twice continuously differentiable with respect to θ in a neighbourhood of θ_0 and is uniformly bounded away from 0 and 1; furthermore, $m/N \to \rho \in (0, 1)$ as $N \to \infty$.

Condition 2. The initial estimator $\tilde{\gamma}$ converges in mean square to a γ_0 within the parameter space Γ such that, for sufficiently large m and N, $E[(\tilde{\gamma} - \gamma_0)(\tilde{\gamma} - \gamma_0)^{\mathrm{T}}] \leqslant A_0$ for a fixed positive definite matrix A_0.

Condition 3. Let $\xi_0 = (U_0^{\mathrm{T}}, (\psi_0 - \mu_0)^{\mathrm{T}})^{\mathrm{T}}$. It is assumed that $E[\xi_0 \xi_0^{\mathrm{T}}/w_0]$ and $E[\xi_0 \xi_0^{\mathrm{T}}/(1 - w_0)]$ are positive definite, and the rank of $E[\partial U_0/\partial \beta]$ is p, which is also the dimension of β.

Condition 4. $\partial^2 U(\beta)/\partial \beta \, \partial \beta^{\mathrm{T}}$ is continuous in a neighbourhood of β_0 where $\|\partial U(\beta)/\partial \beta\|$ is bounded; $\partial^2 \psi(\beta, \gamma)/\partial \gamma \, \partial \gamma^{\mathrm{T}}$ is continuous in a neighbourhood of (β_0, γ_0), and in this neighbourhood $\|\partial \psi(\beta, \gamma)/\partial \gamma\|$ is bounded, $E[\|U(\beta)\|]^2 < \infty$ and $E[\|\psi(\beta, \gamma)\|]^2 < \infty$.

Let

$$q_{N0} = N^{-1} \sum_{i=1}^{N} \frac{\delta_i - w_i(\theta_0)}{w_i(\theta_0)\{1 - w_i(\theta_0)\}} \frac{\partial w_i(\theta_0)}{\partial \theta},$$

$$\Lambda_\theta = E\left[\frac{1}{w_0(1 - w_0)} \frac{\partial w_0}{\partial \theta} \frac{\partial w_0^{\mathrm{T}}}{\partial \theta} \right].$$

We have the following result on the maximum likelihood estimator $\hat{\theta}$ for the parameter of the propensity score.

Lemma 1. Under condition 1, $\hat{\theta} - \theta_0 = \Lambda_\theta^{-1} q_{n0} + o_p(N^{-1/2})$.

Proof. Since $\hat{\theta}$ is the maximizer of the binomial likelihood (1),

$$\frac{\partial l_{\mathrm{B}}(\theta)}{\partial \theta} = \sum_{i=1}^{N} \frac{\delta_i - w_i(\theta)}{w_i(\theta)\{1 - w_i(\theta)\}} \frac{\partial w_i(\theta)}{\partial \theta} = 0. \tag{24}$$

By Taylor's expansion of equation (24) at the true value θ_0,

$$\hat{\theta} - \theta_0 = B_N^{-1} q_{N0} + o_p(N^{-1}) \tag{25}$$

where

$$B_N = N^{-1} \sum_{i=1}^{N} \left[\frac{\delta_i - w_i(\theta_0)}{w_i(\theta_0)\{1 - w_i(\theta_0)\}} \right] \left[\frac{\partial^2 w_i(\theta)}{\partial \theta^2} - \frac{\{1 - 2\, w_i(\theta_0)\}}{w_i(\theta_0)\{1 - w_i(\theta_0)\}} \frac{\partial w_i(\theta_0)}{\partial \theta} \frac{\partial w_i^{\mathrm{T}}(\theta_0)}{\partial \theta} \right]$$
$$+ N^{-1} \sum_{i=1}^{N} \frac{1}{1 - w_i(\theta_0)} \frac{\partial w_i(\theta_0)}{\partial \theta} \frac{\partial w_i^{\mathrm{T}}(\theta_0)}{\partial \theta}.$$

As $B_N = \Lambda_\theta + o_p(1)$ and $q_{N0} = O_p(N^{-1/2})$, the lemma is established from equation (25).

Lemma 2. Under conditions 1–4, $\hat{\lambda} = O_p(N^{-1/2})$, $\hat{\nu} = O_p(N^{-1/2})$ and $\hat{\mu} - \mu_0 = O_p(N^{-1/2})$.

Proof. The selection bias in the missingness of the outcome variable means that

$$E\left[\frac{\delta_i\{\psi_i(\eta_0) - \mu_0\}}{w_i(\theta_0)} \right] = 0, \qquad i = 1, \ldots, n,$$

$$E\left[\frac{(1-\delta_j)\{\psi_j(\eta_0)-\mu_0\}}{1-w_j(\theta_0)}\right]=0, \qquad j=m+1,\ldots,N.$$

Hence both $N^{-1}\Sigma_{i=1}^{m}\{\psi_i(\eta_0)-\mu_0\}/w_i(\theta_0)$ and $N^{-1}\Sigma_{j=m+1}^{N}\{\psi_j(\eta_0)-\mu_0\}/\{1-w_j(\theta_0)\}$ are $O_p(N^{-1/2})$. Note that $\tilde{\eta}=\eta_0+O_p(N^{-1/2})$ as assumed in condition 2. Lemma 2 then follows similar derivations to those in Owen (1990) and Qin and Lawless (1994).

A.1. Proof of theorem 1

Since $\hat{\theta}=\theta_0+O_p(N^{-1/2})$, then carrying out Taylor's expansions of equations (12)–(15) at $(\beta=\beta_0, \mu=\mu_0, \lambda=0)$ and ignoring terms of $o_p(N^{1/2})$ lead to

$$\sum_{i=1}^{m}\frac{\hat{\mu}-\mu_0}{w_i(\theta_0)}+\sum_{i=1}^{m}\frac{(\psi_i(\eta_0)-\mu_0)(\psi_i(\eta_0)-\mu_0)^{\mathrm{T}}}{w_i^2(\theta_0)}\hat{\lambda}=\sum_{i=1}^{m}\frac{\psi_i(\tilde{\eta})-\mu_0}{w_i(\hat{\theta})}, \tag{26}$$

$$\sum_{j=m+1}^{N}\frac{\hat{\mu}-\mu_0}{1-w_j(\theta_0)}+\sum_{j=m+1}^{N}\frac{(\psi_j(\eta_0)-\mu_0)(\psi_j(\eta_0)-\mu_0)^{\mathrm{T}}}{\{1-w_j(\theta_0)\}^2}\hat{\nu}=\sum_{j=m+1}^{N}\frac{\psi_j(\tilde{\eta})-\mu_0}{1-w_j(\hat{\theta})}, \tag{27}$$

$$\sum_{i=1}^{m}\frac{\hat{\lambda}}{w_i(\theta_0)}+\sum_{j=m+1}^{N}\frac{\hat{\nu}}{1-w_j(\theta_0)}=0, \tag{28}$$

$$-\sum_{i=1}^{m}\frac{\partial U_i^{\mathrm{T}}(\beta_0)/\partial\beta}{w_i(\theta_0)}(\hat{\beta}-\beta_0)+\sum_{i=1}^{m}\frac{U_i(\beta_0)(\psi_i^{\mathrm{T}}(\eta_0)-\mu_0)^{\mathrm{T}}}{w_i^2(\theta_0)}\lambda=\sum_{i=1}^{m}\frac{U_i(\beta_0)}{w_i(\hat{\theta})}. \tag{29}$$

Let

$$A_N=N^{-1}\begin{pmatrix}0 & A_{12}\\ A_{12}^{\mathrm{T}} & A_{22}\end{pmatrix},$$

where

$$A_{12}=N^{-1}\left(0,\quad \sum_{i=1}^{m}\frac{1}{w_i(\theta_0)},\quad \sum_{j=m+1}^{N}\frac{1}{1-w_j(\theta_0)}\right),$$

and

$$A_{22}=N^{-1}\begin{pmatrix}-\sum_{i=1}^{m}\frac{\partial U_i^{\mathrm{T}}(\beta_0)/\partial\beta}{w_i(\theta_0)} & \sum_{i=1}^{m}\frac{U_i(\beta_0)(\psi_i^{\mathrm{T}}(\eta_0)-\mu_0)^{\mathrm{T}}}{w_i^2(\theta_0)} & 0\\ 0 & \sum_{i=1}^{m}\frac{(\psi_i(\eta_0)-\mu_0)(\psi_i(\eta_0)-\mu_0)^{\mathrm{T}}}{w_i^2(\theta_0)} & 0\\ 0 & 0 & \sum_{j=m+1}^{N}\frac{(\psi_j(\eta_0)-\mu_0)(\psi_j(\eta_0)-\mu_0)^{\mathrm{T}}}{\{1-w_j(\theta_0)\}^2}\end{pmatrix}.$$

Furthermore, let

$$q_N=N^{-1}\left\{\sum_{i=1}^{m}\frac{U_i^{\mathrm{T}}(\beta_0)}{w_i(\hat{\theta})},\sum_{i=1}^{m}\frac{(\psi_i(\eta_0)-\mu_0)^{\mathrm{T}}}{w_i(\hat{\theta})},\sum_{j=m+1}^{N}\frac{(\psi_j(\eta_0)-\mu_0)^{\mathrm{T}}}{1-w_j(\hat{\theta})}\right\}^{\mathrm{T}}. \tag{30}$$

The four equations (26)–(29) can be written as

$$A_N((\hat{\mu}-\mu_0)^{\mathrm{T}},(\hat{\beta}-\beta_0)^{\mathrm{T}},\hat{\lambda}^{\mathrm{T}},\hat{\nu}^{\mathrm{T}})^{\mathrm{T}}=(0,q_N^{\mathrm{T}})^{\mathrm{T}}+o_p(N^{-1/2}). \tag{31}$$

It can be shown that

$$A_N \xrightarrow{\mathrm{p}} \Sigma =: \begin{pmatrix} 0 & \Sigma_{12} \\ \Sigma_{12}^{\mathrm{T}} & \Sigma_{22} \end{pmatrix} \qquad \text{as } N \to \infty, \tag{32}$$

where $\Sigma_{12} = (0, I_p, I_p)$ and

$$\Sigma_{22} = \begin{pmatrix} -E\left[\dfrac{\partial U_0}{\partial \beta}\right] & E\left[\dfrac{U_0(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0}\right] & 0 \\ 0 & E\left[\dfrac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0}\right] & 0 \\ 0 & 0 & E\left[\dfrac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^{\mathrm{T}}}{1 - w_0}\right] \end{pmatrix}.$$

Here I_p is a $p \times p$ identity matrix. Thus, equations (31) and (32) imply that

$$((\hat{\mu} - \mu_0)^{\mathrm{T}}, (\hat{\beta} - \beta_0)^{\mathrm{T}}, \hat{\lambda}^{\mathrm{T}}, \hat{\nu}^{\mathrm{T}})^{\mathrm{T}} = \Sigma^{-1}(0, q_N^{\mathrm{T}})^{\mathrm{T}} + o_p(N^{-1/2}). \tag{33}$$

Note that

$$\Sigma^{-1} = \begin{pmatrix} -D^{-1} & D^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ \Sigma_{22}^{-1}\Sigma_{12}^{\mathrm{T}}D^{-1} & \Sigma_{22}^{-1} - \Sigma_{22}^{-1}\Sigma_{12}^{\mathrm{T}}D^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{pmatrix}, \tag{34}$$

where

$$D = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^{\mathrm{T}} = E^{-1}\left[\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0}\right] + E^{-1}\left[\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^{\mathrm{T}}}{1 - w_0}\right].$$

Furthermore,

$$D^{-1}\Sigma_{12}\Sigma_{22}^{-1} = D^{-1}\left\{0, E^{-1}\left[\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0}\right], E^{-1}\left[\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^{\mathrm{T}}}{1 - w_0}\right]\right\}.$$

Let R be the second 'row' of Σ^{-1} after deleting the first 'column'. Then,

$$\begin{aligned} R = &-E^{-1}\left[\frac{\partial U_0}{\partial \beta}\right]\left(I_p, -E\left[\frac{U_0(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0}\right]E^{-1}\left[\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0(1 - w_0)}\right],\right. \\ &\left. E\left[\frac{U_0(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0}\right]E^{-1}\left[\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0}\right]D^{-1}E^{-1}\left[\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^{\mathrm{T}}}{1 - w_0}\right]\right) \\ = &-E^{-1}\left[\frac{\partial U_0}{\partial \beta}\right](I_p, -A, A), \end{aligned} \tag{35}$$

where

$$A = E\left[\frac{U_0(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0}\right]E^{-1}\left[\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0(1 - w_0)}\right].$$

This unique structure of R is instrumental in delivering a neat expression for the asymptotic covariance matrix of $\hat{\beta}$. From equation (33),

$$\hat{\beta} - \beta_0 = Rq_N + o_p(N^{-1/2}). \tag{36}$$

Applying Taylor's expansion on q_N,

$$q_N = q_N^{(1)} + q_N^{(2)} + o_p(N^{-1/2}) \tag{37}$$

where

$$\begin{aligned} q_N^{(1)} = N^{-1}&\left(\sum_{i=1}^{m}\frac{U_{i0}}{w_{i0}}, \sum_{i=1}^{m}\frac{\psi_{i0} - \mu_0}{w_{i0}}, \sum_{j=m+1}^{N}\frac{\psi_{j0} - \mu_0}{1 - w_{j0}}\right) + \left(-E\left[\frac{U_0\partial w_0^{\mathrm{T}}/\partial\theta}{w_0}\right], -E\left[\frac{(\psi_0 - \mu_0)\partial w_0^{\mathrm{T}}/\partial\theta}{w_0}\right],\right. \\ &\left. E\left[\frac{(\psi_0 - \mu_0)\partial w_0^{\mathrm{T}}/\partial\theta}{1 - w_0}\right]\right)^{\mathrm{T}}\Lambda_\theta^{-1}q_{N0}, \end{aligned}$$

$$q_N^{(2)} = N^{-1}(0, I_p, I_p)^{\mathrm{T}} E\left[\frac{\partial(\psi_0 - \mu_0)^{\mathrm{T}}}{\partial \eta}\right](\tilde{\eta} - \eta_0),$$

where q_{N0} is defined at the beginning of the appendix. Note that $q_N^{(1)}$ is a sample average of independent and identically distributed random vectors. Applying the standard multivariate central limit theorem and Slutsky's theorem, it can be shown that

$$N^{1/2} q_N^{(1)} \xrightarrow{\mathrm{d}} N(0, \Omega^{(1)}) \qquad \text{as } N \to \infty, \tag{38}$$

where

$$\Omega^{(1)} = \Omega^{(11)} - \Omega^{(12)}, \tag{39}$$

$$\Omega^{(11)} = \begin{pmatrix} E\left[\frac{U_0 U_0^{\mathrm{T}}}{w_0}\right] & E\left[\frac{U_0(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0}\right] & 0 \\ E\left[\frac{U_0(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0}\right] & E\left[\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0}\right] & 0 \\ 0 & 0 & E\left[\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^{\mathrm{T}}}{1 - w_0}\right] \end{pmatrix},$$

$$\Omega^{(12)} = E\left[\zeta \frac{\partial w_0^{\mathrm{T}}}{\partial \theta}\right] \Lambda_\theta^{-1} E\left[\frac{\partial w_0}{\partial \theta} \zeta^{\mathrm{T}}\right],$$

$$\zeta = \left(-\frac{U_0^{\mathrm{T}}}{w_0}, -\frac{(\psi_0 - \mu_0)^{\mathrm{T}}}{w_0}, \frac{(\psi_0 - \mu_0)^{\mathrm{T}}}{1 - w_0}\right).$$

Let

$$B = E\left[\frac{\partial(\psi_0 - \mu_0)}{\partial \eta}\right] \mathrm{var}(\tilde{\eta})\, E\left[\frac{\partial(\psi_0 - \mu_0)^{\mathrm{T}}}{\partial \eta}\right].$$

Then,

$$N \,\mathrm{var}(q_N^{(2)}) = \Omega^{(2)} =: \begin{pmatrix} 0 & 0 & 0 \\ 0 & B & B \\ 0 & B & B \end{pmatrix}. \tag{40}$$

From equation (35), $N \,\mathrm{var}(R q_N^{(2)}) = N R \Omega^{(2)} R^{\mathrm{T}} = 0$. Thus, $R q_N^{(2)} = o_p(N^{-1/2})$. Therefore, $\hat{\beta} - \beta_0 = R q_N^{(1)} + o_p(N^{-1/2})$. This result and equation (38) together give

$$N^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{\mathrm{d}} N(0, \Sigma_\beta) \qquad \text{as } N \to \infty, \tag{41}$$

where $\Sigma_\beta = R(\Omega^{(11)} - \Omega^{(12)}) R^{\mathrm{T}}$. After some matrix algebra, it can be shown that

$$R \Omega^{(11)} R^{\mathrm{T}} = \Sigma_\beta^{(0)} - \Sigma_\beta^{(2)}.$$

Clearly $R \Omega^{(12)} R^{\mathrm{T}} = \Sigma_\beta^{(1)}$. These results then imply the results of theorem 1.

A.2. Proof of theorem 2

Applying Taylor's expansion on equation (4) at $(\beta_0, \gamma_0, \theta_0)$ gives

$$E\left[\frac{\partial U^{\mathrm{T}}}{\partial \beta}\right](\hat{\beta}_{\mathrm{RRZ}} - \beta_0) = -r_{n1} + r_{n2} + o_p(N^{-1/2}) \tag{42}$$

where

$$r_{n1} = N^{-1} \sum_{i=1}^{N} \frac{\delta_i U_{i0} - (\delta_i - w_{i0})\psi_{i0}}{w_{i0}},$$

$$r_{n2} = E\left[\frac{U_0 - \psi_0}{w_0}\frac{\partial w_0^{\mathrm{T}}}{\partial \theta}\right](\hat{\theta} - \theta_0).$$

Standard derivations show that

$$\mathrm{var}(r_{n1}) =: N^{-1}\tilde{\Omega}_1 = N^{-1}\left\{E\left[\frac{U_0 U_0^{\mathrm{T}}}{w_0}\right] - E\left[(1-w_0)\left(\frac{U_0\psi_0^{\mathrm{T}}}{w_0} + \frac{\psi_0 U_0^{\mathrm{T}}}{w_0} - \frac{\psi_0\psi_0^{\mathrm{T}}}{w_0}\right)\right]\right\} \tag{43}$$

and

$$-\mathrm{cov}(r_{n1}, r_{n2}) - \mathrm{cov}(r_{n2}, r_{n1}) + \mathrm{var}(r_{n2}) =: N^{-1}\tilde{\Omega}_2 = -N^{-1}E\left[\frac{U_0 - \psi_0}{w_0}\frac{\partial w_0^{\mathrm{T}}}{\partial \theta}\right]\Lambda_\theta^{-1}E\left[\frac{\partial w_0}{\partial \theta}\frac{(U_0 - \psi_0)^{\mathrm{T}}}{w_0}\right]. \tag{44}$$

The central limit theorem and equations (43) and (44) together imply that

$$N^{1/2}(-r_{n1} + r_{n2}) \xrightarrow{\mathrm{d}} N(0, \tilde{\Omega}_1 + \tilde{\Omega}_2). \tag{45}$$

Theorem 2 is readily implied by equations (42) and (45).

References

Alvarez, R. M. and Nagler, J. (1998) Economics, entitlements, and social issues: voter choice in the 1996 presidential election. *Am. J. Polit. Sci.*, **42**, 1349–1363.

Baker, S. G. (2006) Surrogate endpoints: wishful thinking or reality? *J. Natn. Cancer Inst.*, **98**, 502–503.

Baker, S. G., Izmirlian, G. and Kipnis, V. (2005) Resolving paradoxes involving surrogate end points. *J. R. Statist. Soc.* A, **168**, 753–762.

Begg, C. B. and Leung, D. H. Y. (2000) On the use of surrogate end points in randomized trials. *J. R. Statist. Soc.* A, **163**, 15–28.

Burzykowski, T., Molenberghs, G. and Buyse, M. (2005) *The Evaluation of Surrogate Endpoints*. New York: Springer.

Cassel, C. M., Sarndal, C. E. and Wretman, J. H. (1976) Some results on generalized difference estimation and regression estimation for finite populations. *Biometrika*, **63**, 615–620.

Chen, J. and Breslow, N. E. (2004) Semiparametric efficient estimation for the auxiliary outcome problem with conditional mean model. *Can. J. Statist.*, **32**, 359–372.

Chen, S. X., Leung, D. and Qin, J. (2003) Information recovery in a study with surrogate endpoints. *J. Am. Statist. Ass.*, **98**, 1052–1062.

Chen, X., Hong, H. and Tamer, E. (2005) Measurement error models with auxiliary data. *Rev. Econ. Stud.*, **72**, 343–366.

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

Clayton, D., Spiegelhalter, D., Dunn, G. and Pickles, A. (1998) Analysis of longitudinal binary data from multiphase sampling. *J. R. Statist. Soc.* B, **60**, 71–87.

Cochran, W. G. (1977) *Sampling Techniques*, 3rd edn. New York: Wiley.

Filer, J. E. and Kenney, L. W. (1980) Voter turnout and the benefits of voting. *Publ. Choice*, **35**, 575–585.

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

Lee, M. J. (2005) Monotonicity conditions and inequality imputation for sample-selection and non-response problems. *Econometr. Rev.*, **24**, 175–194.

Lee, M. J. and Kang, S. J. (2002) Multinomial-choice and presidential election. *Unpublished manuscript*. Korea University, Seoul.

Leung, D. H.-Y. (2001) Statistical methods for clinical studies in the presence of surrogate end points. *J. R. Statist. Soc.* A, **164**, 485–503.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Values*, 2nd edn. Hoboken: Wiley.

Newey, W. (1990) Semiparametric efficiency bounds. *J. Appl. Econometr.*, **5**, 99–135.

Owen, A. (1990) Empirical likelihood ratio confidence regions. *Ann. Statist.*, **18**, 90–120.

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

Qin, J. and Lawless, J. F. (1994) Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, 300–325.

Riker, W. H. and Ordeshook, P. C. (1968) A theory of the calculus of voting. *Am. Polit. Sci. Rev.*, **62**, 25–42.

Robins, J. M. and Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Statist. Ass.*, **90**, 122–129.

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

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Ass.*, **90**, 106–121.
Rosenbaum, P. and Rubin, D. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.
Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
Sanders, M. S. (2001) Uncertainty and turnout. *Polit. Anal.*, **90**, 45–57.
Schenker, N. and Taylor, J. M. G. (1996) Partially parametric techniques for multiple imputation. *J. Computnl Statist. Data Anal.*, **22**, 425–446.
Vardi, Y. (1985) Empirical distributions in selection bias models (with comments). *Ann. Statist.*, **13**, 178–205.
Warren, E. M., Kinder, D. R. and Rosenstone, S. J. (1999) National election studies 1996. *Report*. Center for Political Studies, University of Michigan, Ann Arbor.
Wittes, J., Lakatos, E. and Probstfield, J. (1989) Surrogate endpoints in clinical trials: cardiovascular disease. *Statist. Med.*, **8**, 415–425.
Wooldridge, J. (2004) Inverse probability weighted estimation for general missing data problem. *Working Paper CWP05/04*. Institute for Fiscal Studies, London.
Yu, M. and Nan, B. (2006) A revisit of semiparametric regression models with missing data. *Statist. Sin.*, **16**, 1193–1212.