

Improving semiparametric estimation using surrogate data

Song Xi Chen¹, Denis H.Y. Leung² and Jing Qin³

¹ Department of Statistics, Iowa State University

Email: songchen@iastate.edu

² School of Economics, Singapore Management University

Email: denisleung@smu.edu.sg

³ National Institute of Allergy and Infectious Disease, National Institute of Health

Email: jingqin@niaid.nih.gov

ABSTRACT

This 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 depend 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 estimation efficiency 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 proposed estimator achieves efficiency gain 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 generated by U . The practical performance of the estimator is evaluated using simulation and a dataset based on the 1996 U.S. 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 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 *vs* not cured) in every study participant, therefore, a surrogate outcome (*e.g.*, biomarkers) may be used for those participants without the true outcome (*e.g.*, Wittes, Lakatos and Probstfield, 1989; Begg and Leung, 2000; Leung, 2001; Baker, Izmirlian and Kipnis, 2005; Burzykowski, Molenberghs and Buyse, 2005; Baker, 2006) and in economics, proxy or surrogate outcomes are often used in surveys with missing response (Chen, Hong and Tamer, 2005)

One difficulty in modeling data with missing outcome is 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 missing at random (MAR), *i.e.*, probability of a missing outcome is independent of the (unobserved) outcome, given the surrogate and the covariates (Little and Rubin, 2002). Under 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 and are also covered by the methods 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 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 consists of two parts; a part with (Y, S, X, Z) completely observed:

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

and a part with missing Y :

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

Let $N = m + n$. We assume 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 Z is null and we drop Z from the formulation of w . The results we discuss also applies 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 w can be estimated by $\hat{\theta}$ that maximizes the binomial log-likelihood:

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

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

$$\ell_{full} = \ell_B(\theta) + \sum_{i=1}^m \log \{f(y_i, x_i, s_i)\} + \sum_{j=m+1}^N \log \{f(x_j, s_j)\}. \quad (2)$$

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

$$f(s|x) = \int f(y|x, \beta)g(s|y, x)dy, \quad (3)$$

However, in general, it is hard 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 (3) by a kernel density estimate based on the completely observed data. Schenker and Taylor (1996) suggested using imputation (Rubin, 1987) for missing outcomes. Chen and Chen (2000) suggested a method based on the regression estimate. Chen, Leung and Qin (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 MAR case because the structure of the likelihood is changed due to the selection bias in the missingness under the 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, Rotnitzky and Zhao (1995) and Robins and Rotnitzky (1995) proposed using estimating equations for situations where Y can be MAR. In the framework considered here, their estimator (denoted by $\hat{\beta}_{RRZ}$ hereafter) solves

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

for a specific function ψ and a mean zero 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 generated

by $U(y, x, \beta)$ (Newey, 1990) for estimating β if $\psi(s, x, \beta) = \psi^*(s, x, \beta)$. Furthermore, $\hat{\beta}_{RRZ}$ is consistent if either w or ψ is correctly specified. This property is the so-called “doubly robustness” property. However, $\hat{\beta}_{RRZ}$ may suffer efficiency loss when $\psi \neq \psi^*$, as shown in Theorem 2 of Section 3.

The estimator (4) is a special case of a larger class of semi-parametric efficient estimators developed by Robins, Rotnitzky and Zhao (1994). However, as Chen and Chen (2000) pointed out that the semiparametric efficient estimators suggested by Robins *et al.* (1994) is practically not 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 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 similar situations as considered here where closed form optimal estimating equations can be found.

Eventhough 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

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

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 obtained from equations (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 using EL weights obtained by leveraging the information contained in S and X . When ψ and w are correctly specified, our method is efficient within the class of estimating functions 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 the 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 proposed method to existing methods. In Section 5, the method is applied to a real dataset. Conclusions are given in Section 6. Proofs are given in the Appendix.

2. PROPOSED METHOD

Suppose $U(y, x, \beta)$ 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, $\hat{\beta}_W$, that solves

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

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

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

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

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

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

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

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

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

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

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

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

To simplify notations, 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

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

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

with constraints

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

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

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

$$\ell(\mu) = -\log \sum_{i=1}^m [1 + \lambda^T \{\psi_i(\tilde{\eta}) - \mu\} / w_i(\hat{\theta})] - \log \sum_{j=m+1}^N [1 + \nu^T \{\psi_j(\tilde{\eta}) - \mu\} / \{1 - w_j(\hat{\theta})\}].$$

Differentiating $\ell(\mu)$ with respect to μ and equating to zero leads to

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

Let $(\hat{\mu}, \hat{\lambda}, \hat{\nu})$ be the solution of (12) – (14). Substitute them to (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

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

We will show that $\hat{\beta}$ is more efficient than $\hat{\beta}_W$ in (6).

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

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

the EL estimating equation (15) becomes

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

Hence, the proposed estimator is asymptotically equivalent to the solution from an estimating equation that regresses the inversely weighted estimating equation $m^{-1} \sum_{i=1}^m U_i(\beta)/w_i(\hat{\theta})$ on $m^{-1} \sum_{i=1}^m \{\psi_i(\tilde{\eta}) - \mu\}/w_i(\hat{\theta})$. As a result, the variance of the EL estimating equation is smaller than that of the inversely weighted estimating equation $m^{-1} \sum_{i=1}^m U_i(\beta)/w_i(\hat{\theta})$. This result is similar to the case when Y and X are two random variables, then $\text{Var}(Y - AX) = \text{Var}(Y) - A\text{Var}(X)A^T \leq \text{Var}(Y)$, where $A = \text{Cov}(Y, X)[\text{Var}(X)]^{-1}$. In contrast, the estimating function of Robins *et al.* (1995) is a difference between the inversely weighted estimating equation and $\sum_{i=1}^m \psi_i(\tilde{\eta})\{\delta_i - w_i(\hat{\theta})\}/w_i(\hat{\theta})$. 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 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^* using (5) and $\tilde{\gamma}$ be an estimate based on that model. We will 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 exists a positive constant c_0 such that $E(\tilde{\gamma} - \gamma_0)^2 \leq c_0 n^{-1}$.

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 $\sum_{i=1}^m \hat{p}_i I\{(y_i, s_i, x_i) \leq t\}$ is a consistent estimate of $F(y, s, x|D = 1)$ and

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

Since $\hat{\beta}$ solves (15) which can be regarded as a sample version of the above population equation, $\hat{\theta}$ is asymptotically unbiased and its variance converges to zero as $\min\{m, n\} \rightarrow \infty$. 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 $U_0 \stackrel{d}{=} U_i(\beta_0)$, $\psi_0 \stackrel{d}{=} \psi_i(\eta_0)$, $\mu_0 = E(\psi_0)$ and $w_0 \stackrel{d}{=} w_i(\theta_0)$ where $\stackrel{d}{=}$ denotes equivalence in distributions. Furthermore, let

$$A = E\left(\frac{U_0(\psi_0 - \mu_0)^T}{w_0}\right) E^{-1}\left(\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^T}{w_0(1 - w_0)}\right), \quad R = -E^{-1}\left(\frac{\partial U_0}{\partial \beta}\right) (I_p, -A, A),$$

$$\zeta = \left(-\frac{U_0^T}{w_0}, -\frac{(\psi_0 - \mu_0)^T}{w_0}, \frac{(\psi_0 - \mu_0)^T}{1 - w_0}\right), \quad \Lambda_\theta = E\left(\frac{1}{w_0\{1 - w_0\}} \frac{\partial w_0}{\partial \theta} \frac{\partial w_0^T}{\partial \theta}\right),$$

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

Theorem 1: Under Conditions C1-C4 given in the Appendix,

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} N\left(0, \Sigma_\beta^{(0)} - \Sigma_\beta^{(1)} - \Sigma_\beta^{(2)}\right), \quad (16)$$

where

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

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

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

We note that (i) $\Sigma_\beta^{(0)}$ is the covariance matrix of $\hat{\beta}_W$, the inverse weighted estimator by the true propensity score w_0 ; (ii) 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 proposed EL estimator is more efficient than $\hat{\beta}_W$, when the true propensity score is used to weight the estimating equation based on the complete observations, unless $\Sigma_\beta^{(1)}$ and $\Sigma_\beta^{(2)}$ are zero matrices simultaneously.

The variance reduction offered by $\Sigma_\beta^{(2)}$ is a result of having the second constraint in (9) based on the observations with missing Y values. If this constraint is removed from (9), $\Sigma_\beta^{(2)}$ will be zero. Therefore, it is worthwhile to carry out both weighting by propensity score and having an extra estimating equation based on the covariate and the surrogate from the part of the sample with missing outcome. The variance reduction 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 estimated over the true propensity score (see, *e.g.*, 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. This observation suggests finding 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

$$\begin{aligned} & E\left(\frac{U_0(\psi^* - \mu_0)^T}{w_0}\right) E^{-1}\left(\frac{(\psi^* - \mu_0)(\psi^* - \mu_0)^T}{w_0(1 - w_0)}\right) E\left(\frac{(\psi^* - \mu_0)U_0^T}{w_0}\right) \\ &= E\left\{\frac{1 - w_0}{w_0}U_0E(U_0^T|s, x)\right\}. \end{aligned}$$

Hence

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

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.*, 2007). Due 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 $\hat{\beta}_{RRZ}$ proposed by Robins *et al.* (1995).

Theorem 2: Under Conditions C1-C4 given in the Appendix,

$$\sqrt{N}(\hat{\beta}_{RRZ} - \beta_0) \xrightarrow{d} N\left(0, \Sigma_{\beta}^{(0)} - \tilde{\Sigma}_{\beta}^{(1)} - \tilde{\Sigma}_{\beta}^{(2)}\right), \quad (20)$$

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

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

and

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

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

4. NUMERICAL STUDY

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

1. The maximum likelihood estimator $\hat{\beta}_C$ assuming 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.

2. The weighted estimator $\hat{\beta}_W$ by solving (6) using only the complete observations. This is also the initial estimator $\tilde{\beta}$ used in obtaining the empirical likelihood weights.
3. The estimator $\hat{\beta}_{RRZ}$.

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

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

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, as

$$\begin{aligned} E(Y|X) &= \beta_1 + \beta_2 X & \text{and} & & E(S|Y, X) &= 1 + 2Y + X, \\ \text{Var}(Y|X) &= \text{Var}(S|Y, X) &= & & 1 \end{aligned}$$

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 $\hat{\beta}_{RRZ}$ and $\hat{\beta}$.

For this model, we used:

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

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

$$E(Y) = \gamma_1 + \gamma_2 S + \gamma_3 X. \quad (22)$$

As mentioned in the Section 2, (22) does not need to be correct. The goal is to recover as much as possible the information loss 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, 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

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

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 $(1, 2)$ and $\theta = (-1, 0, 0), (-1, 0.2, 0.2), (-1, 0.35, 0.35)$ and $(-1, 0.5, 0.5)$ in the missing probability function, with $N=1000$ in each simulation. The choices $\theta = (-1, 0, 0), (-1, 0.2, 0.2), (-1, 0.35, 0.35)$ and $(-1, 0.5, 0.5)$ induced, respectively, approximately 75%, 60%, 47% and 45% missing outcomes in the data.

We considered two methods for variance estimation in each method: (1) the asymptotic variance formulae in Section 3 and (2) the bootstrap method. Under MCAR ($\theta = (-1, 0, 0)$) or weakly MAR ($\theta = (-1, 0.2, 0.2)$), both methods give similar variance estimates. However, under strongly MAR ($\theta = (-1, 0.35, 0.35), (-1, 0.5, 0.5)$), the bootstrap method gives more reliable variance estimates. The better performance of the bootstrap method is due to the fact that the asymptotic variance formula involves the quantity $\sum_{i=1}^n (d_i/w_i)^2 \psi_i^T \psi_i/n$, which can be unduly affected by values of w_i close to 0 or 1, when $\theta = (-1, 0.35, 0.35)$ or $(-1, 0.5, 0.5)$.

The simulation results are reported in Tables 1-2 for the case of $\beta = (1, 2)$. 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 MCAR ($\theta = (-1, 0, 0)$), $\hat{\beta}_{RRZ}$ and the estimator proposed in this paper, $\hat{\beta}$, are almost equivalent. On the other hand, when $\theta = (-1, 0.35, 0.35)$ and $(-1, 0.5, 0.5)$, then missingness depends strongly on (S, X) and in those cases, $\hat{\beta}$ outperforms $\hat{\beta}_{RRZ}$ and $\hat{\beta}_W$. For Model 2, $\hat{\beta}_{RRZ}$ and $\hat{\beta}$ are much better than $\hat{\beta}_W$ when the outcome is MCAR. However, their efficiency gains are reduced when the selection bias in missingness of the outcome variable is large, *i.e.*, $\theta = (-1, 0.35, 0.35)$ and $(-1, 0.5, 0.5)$. Interestingly, in those cases, comparing to the unattainable estimator $\hat{\beta}_C$ based on the full sample, missing data did not lead to much loss of information. The efficiency loss of all three estimators when compared to $\hat{\beta}_C$ is less severe in Model 2 than the corresponding cases in Model 1. Among the three estimators, the estimator 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 further illustrate the results of Theorems 1 and 2, we compared the asymptotic relative efficiencies between the estimators in a modest setup. Two models were used. The first model is a linear model that is similar to Model 1 in the simulation study, except $E(S|Y, X) = 2Y$ if $Y \geq 0$ 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 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:

$$ARE(\hat{\beta}, \hat{\beta}_W) = \frac{Var(\hat{\beta}_W)}{Var(\hat{\beta})}, \quad ARE(\hat{\beta}, \hat{\beta}_{RRZ}) = \frac{Var(\hat{\beta}_{RRZ})}{Var(\hat{\beta})},$$

for estimating β_1, β_2 . The results are given in Figures 1a - 1d. They show that $\hat{\beta}$ is always as efficient as $\hat{\beta}_W$ and $\hat{\beta}_{RRZ}$ in all scenarios studied. The most noticeable features of these results is the poor performance of RRZ under MAR, when ζ is non-zero (Figures 1b and 1d). The poor performance of RRZ results because ψ is very different from $E[U(y, x, \beta)|s, x]$ in those cases. This is a point made at the end of Section 3 that there is no guarantee that the RRZ estimator will always be better than the inversely weighted estimator. In both models, the disadvantage of using $\hat{\beta}_{RRZ}$ 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 MAR, the changes affect β_2 more because it is the coefficient associated with X .

5. APPLICATION TO ELECTION DATA

We applied the proposed method to a set of data from the National Election Study (Warren, Kinder and Rosenstone, 1999; Lee and Kang, 2003; Lee, 2005). The U.S. 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 the election followed a popular vote system. As argued in Lee (2005), this approach is reasonable for illustration purposes because of two reasons: (1) Since the election results using the two systems were very close, the statistical conclusions should be similar using either system; (2) The sample size is not large enough at the state level, which would be required if the electoral college system is used.

The data came from two surveys 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 dataset 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 available for analysis ($N = 1486$) are: No Vote (474 or 32%), Clinton (586 or 39%) and Dole (426 or 29%). Using the method 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 to 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 ($p < 0.001$ using a Chi-squared test).

Alvarez and Nagler (1998) discussed a number of 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 gotten better, stayed about the same or gotten 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 turnouts in US presidential elections (Riker and Ordshook, 1968; Filer and Kenney, 1980; Sanders, 2001). Sanders (2001) used the dataset in this paper to model the probability of turnout (Table 1 in Sanders, 2001) with the following variables: Age, Income, Race, Gender, Education (High school *vs.* College *vs.* 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. These variables are the vector of Z discussed in Section 1. In addition to Z , we added S and X and modeled w using a logistic regression

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

This example highlights the different roles played by Z and S . While S is 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 (choice of the president) and a single covariate (perceived state of the economy). Let Y be the true outcome and $Y = 1$ represents “Clinton is the choice” and $Y = 0$ represents “Dole is the choice”; X be the covariate and $X = -1, 0, 1$ if the respondent thought the nation’s economy had “gotten worse”, “stayed about the same” and “gotten 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 $S = 1$ represents “Clinton is the choice” and $S = 0$ represents “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 modeled as in (23).

The three methods considered in this paper were used to analyze 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 ($\hat{\beta}_2/SE(\hat{\beta}_2) \gg 0$) that voter’s perception on the economy had a significant impact on voting behaviour. Using the weighted estimator, the odds ratio of voting Clinton is $\{\exp(.2989 + .8004)/[1 + \exp(.2989 + .8004)]\}/\{\exp(.2989 - .8004)/[1 + \exp(.2989 - .8004)]\} = 1.98$ for someone who views the economy favourably against someone who views the economy negatively. The conclusions are similar using the other two methods. Using either Robins *et al.*’s (1995) method and the method 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 estimation efficiency in the surrogate outcome problem via Owen’s (1990) empirical likelihood. 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 (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 empirical likelihoods 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” nonlinear model or a general additive model. We corrected the possible bias in the “working” model using estimating equations (8) and then combined them using empirical likelihood. 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 needed in methods such as Robins *et al.*’s (1995) estimator

is not available. With some modifications, the proposed method may be generalized to other missing data situations, for example in measurement error problems.

Appendix

The conditions needed to establish Theorems 1 and 2 are the following:

- C1: The propensity score $w_i(\theta)$ is twice continuously differentiable with respect to θ in a neighborhood of θ_0 and is uniformly bounded away from 0 and 1; furthermore, $m/N \rightarrow \rho \in (0, 1)$ as $N \rightarrow \infty$.
- C2: 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 , $E\{(\tilde{\gamma} - \gamma_0)(\tilde{\gamma} - \gamma_0)^T\} \leq A_0$ for a fixed positive definite matrix A_0 .
- C3: Let $\xi_0 = (U_0^T, (\psi_0 - \mu_0)^T)^T$. It is assumed that $E\left(\frac{\xi_0 \xi_0^T}{w_0}\right)$ and $E\left(\frac{\xi_0 \xi_0^T}{1-w_0}\right)$ are positive definite; and the rank of $E\left(\frac{\partial U_0}{\partial \beta}\right)$ is p , which is also the dimension of β .
- C4: $\frac{\partial^2 U(\beta)}{\partial \beta \partial \beta^T}$ is continuous in a neighborhood of β_0 where $\left\|\frac{\partial U(\beta)}{\partial \beta}\right\|$ is bounded; $\frac{\partial^2 \psi(\beta, \gamma)}{\partial \gamma \partial \gamma^T}$ is continuous in a neighborhood of (β_0, γ_0) , and in this neighborhood $\left\|\frac{\partial \psi(\beta, \gamma)}{\partial \gamma}\right\|$ is bounded, $E(\|U(\beta)\|)^2 < \infty$ and $E(\|\psi(\beta, \gamma)\|)^2 < \infty$.

Let

$$q_{N0} = N^{-1} \sum_{i=1}^N \frac{\delta_i - w_i(\theta_0)}{w_i(\theta_0)\{1 - w_i(\theta_0)\}} \frac{\partial w_i(\theta_0)}{\partial \theta} \quad \text{and} \quad \Lambda_\theta = E\left[\frac{1}{w_0\{1 - w_0\}} \frac{\partial w_0}{\partial \theta} \frac{\partial w_0^T}{\partial \theta}\right].$$

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

Lemma 1: Under Condition C1, $\hat{\theta} - \theta_0 = \Lambda_\theta^{-1} q_{n0} + o_p(N^{-1/2})$.

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

$$\frac{\partial \ell_B(\theta)}{\partial \theta} = \sum_{i=1}^N \frac{\delta_i - w_i(\theta)}{w_i(\theta)\{1 - w_i(\theta)\}} \frac{\partial w_i(\theta)}{\partial \theta} = 0. \tag{A.1}$$

By Taylor's expansion of (A.1) at the true value θ_0 ,

$$\hat{\theta} - \theta_0 = B_N^{-1} q_{N0} + o_p(N^{-1}) \quad (\text{A.2})$$

where

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

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

Lemma 2: Under Conditions C1 - C4, $\hat{\lambda} = O_p(N^{-1/2})$, $\hat{\nu} = O_p(N^{-1/2})$ and $\hat{\mu} - \mu_0 = O_p(N^{-1/2})$.

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

$$\begin{aligned} E\{\delta_i\{\psi_i(\eta_0) - \mu_0\}/w_i(\theta_0)\} &= 0, \quad i = 1, \dots, n \\ E\{(1 - \delta_j)\{\psi_j(\eta_0) - \mu_0\}/\{1 - w_j(\theta_0)\}\} &= 0, \quad j = m + 1, \dots, N. \end{aligned}$$

Hence both $N^{-1} \sum_{i=1}^m \{\psi_i(\eta_0) - \mu_0\}/w_i(\theta_0)$ and $N^{-1} \sum_{j=m+1}^N \{\psi_j(\eta_0) - \mu_0\}/\{1 - w_j(\theta_0)\}$ are $O_p(N^{-1/2})$. Note that $\tilde{\eta} = \eta_0 + O_p(N^{-1/2})$ as assumed in Condition C2. The lemma then follows similar derivations as those in Owen (1990) and Qin and Lawless (1994).

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

$$\sum_{i=1}^m \frac{\hat{\mu} - \mu_0}{w_i(\theta_0)} + \sum_{i=1}^m \frac{\{\psi_i(\eta_0) - \mu_0\}\{\psi_i(\eta_0) - \mu_0\}^T \hat{\lambda}}{w_i^2(\theta_0)} = \sum_{i=1}^m \frac{\psi_i(\tilde{\eta}) - \mu_0}{w_i(\hat{\theta})}, \quad (\text{A.3})$$

$$\sum_{j=m+1}^N \frac{\hat{\mu} - \mu_0}{1 - w_j(\theta_0)} + \sum_{j=m+1}^N \frac{\{\psi_j(\eta_0) - \mu_0\}\{\psi_j(\eta_0) - \mu_0\}^T \hat{\nu}}{\{1 - w_j(\theta_0)\}^2} = \sum_{j=m+1}^N \frac{\psi_j(\tilde{\eta}) - \mu_0}{1 - w_j(\hat{\theta})}, \quad (\text{A.4})$$

$$\sum_{i=1}^m \frac{\hat{\lambda}}{w_i(\theta_0)} + \sum_{j=m+1}^N \frac{\hat{\nu}}{1 - w_j(\theta_0)} = 0, \quad (\text{A.5})$$

$$-\sum_{i=1}^m \frac{\partial U_i^T(\beta_0)/\partial \beta}{w_i(\theta_0)} (\hat{\beta} - \beta_0) + \sum_{i=1}^m \frac{U_i(\beta_0)(\psi_i^T\{\eta_0\} - \mu_0)^T \lambda}{w_i^2(\theta_0)} = \sum_{i=1}^m \frac{U_i(\beta_0)}{w_i(\hat{\theta})}. \quad (\text{A.6})$$

Let

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

where

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

and

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

Furthermore, let

$$q_N = N^{-1} \left(\sum_{i=1}^m \frac{U_i^T(\beta_0)}{w_i(\hat{\theta})}, \sum_{i=1}^m \frac{\{\psi_i(\eta_0)-\mu_0\}^T}{w_i(\hat{\theta})}, \sum_{j=m+1}^N \frac{\{\psi_j(\eta_0)-\mu_0\}^T}{1-w_j(\hat{\theta})} \right)^T. \quad (\text{A.7})$$

The four equations (A.3) to (A.6) can be written as

$$A_N \left((\hat{\mu} - \mu_0)^T, (\hat{\beta} - \beta_0)^T, \hat{\lambda}^T, \hat{\nu}^T \right)^T = (0, q_N^T)^T + o_p(N^{-1/2}). \quad (\text{A.8})$$

It can be shown that

$$A_N \xrightarrow{p} \Sigma =: \begin{pmatrix} 0 & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} \quad \text{as } N \rightarrow \infty, \quad (\text{A.9})$$

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

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

Here I_p is a $p \times p$ identity matrix. Thus, (A.8) and (A.9) imply that

$$\left((\hat{\mu} - \mu_0)^T, (\hat{\beta} - \beta_0)^T, \hat{\lambda}^T, \hat{\nu}^T \right)^T = \Sigma^{-1} (0, q_N^T)^T + o_p(N^{-1/2}). \quad (\text{A.10})$$

Note that,

$$\Sigma^{-1} = \begin{pmatrix} -D^{-1} & D^{-1} \Sigma_{12} \Sigma_{22}^{-1} \\ \Sigma_{22}^{-1} \Sigma_{12}^T D^{-1} & \Sigma_{22}^{-1} - \Sigma_{22}^{-1} \Sigma_{12}^T D^{-1} \Sigma_{12} \Sigma_{22}^{-1} \end{pmatrix}, \quad (\text{A.11})$$

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

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

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

$$\begin{aligned} R &= -E^{-1}\left(\frac{\partial U_0}{\partial \beta}\right)\left[I_p, -E\left(\frac{U_0(\psi_0 - \mu_0)^T}{w_0}\right)E^{-1}\left(\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^T}{w_0(1 - w_0)}\right),\right. \\ &\quad \left.E\left(\frac{U_0(\psi_0 - \mu_0)^T}{w_0}\right)E^{-1}\left(\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^T}{w_0}\right)D^{-1}E^{-1}\left(\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^T}{1 - w_0}\right)\right] \\ &= -E^{-1}\left(\frac{\partial U_0}{\partial \beta}\right)(I_p, -A, A), \end{aligned} \quad (\text{A.12})$$

where $A = E\left(\frac{U_0(\psi_0 - \mu_0)^T}{w_0}\right)E^{-1}\left(\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^T}{w_0(1 - w_0)}\right)$. This unique structure of R is instrumental in delivering a neat expression for the asymptotic covariance matrix of $\hat{\beta}$. From (A.10),

$$\hat{\beta} - \beta_0 = R q_N + o_p(N^{-1/2}). \quad (\text{A.13})$$

Applying Taylor’s expansion on q_N ,

$$q_N = q_N^{(1)} + q_N^{(2)} + o_p(N^{-1/2}) \quad (\text{A.14})$$

where

$$\begin{aligned} q_N^{(1)} &= N^{-1}\left(\sum_{i=1}^m \frac{U_{i0}}{w_{i0}}, \sum_{i=1}^m \frac{\psi_{i0} - \mu_0}{w_{i0}}, \sum_{j=m+1}^N \frac{\psi_{j0} - \mu_0}{1 - w_{j0}}\right) \\ &\quad + \left(-E\left(\frac{U_0 \frac{\partial w_0^T}{\partial \theta}}{w_0}\right), -E\left(\frac{(\psi_0 - \mu_0) \frac{\partial w_0^T}{\partial \theta}}{w_0}\right), E\left(\frac{(\psi_0 - \mu_0) \frac{\partial w_0^T}{\partial \theta}}{1 - w_0}\right)\right)^T \Lambda_\theta^{-1} q_{N0}, \\ q_N^{(2)} &= N^{-1}(0, I_p, I_p)^T E\left(\frac{\partial(\psi_0 - \mu_0)^T}{\partial \eta}\right)(\tilde{\eta} - \eta_0), \end{aligned}$$

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

$$\sqrt{N}q_N^{(1)} \xrightarrow{d} N(0, \Omega^{(1)}) \quad \text{as } N \rightarrow \infty, \quad (\text{A.15})$$

where

$$\begin{aligned}
\Omega^{(1)} &= \Omega^{(11)} - \Omega^{(12)}, \\
\Omega^{(11)} &= \begin{pmatrix} E\left(\frac{U_0 U_0^T}{w_0}\right) & E\left(\frac{U_0(\psi_0 - \mu_0)^T}{w_0}\right) & 0 \\ E\left(\frac{U_0(\psi_0 - \mu_0)^T}{w_0}\right) & E\left(\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^T}{w_0}\right) & 0 \\ 0 & 0 & E\left(\frac{(\psi_0 - \mu_0)(\psi_0 - \mu_0)^T}{1 - w_0}\right) \end{pmatrix}, \\
\Omega^{(12)} &= E\left(\zeta \frac{\partial w_0^T}{\partial \theta}\right) \Lambda_\theta^{-1} E\left(\frac{\partial w_0}{\partial \theta} \zeta^T\right), \\
\zeta &= \left(-\frac{U_0^T}{w_0}, -\frac{(\psi_0 - \mu_0)^T}{w_0}, \frac{(\psi_0 - \mu_0)^T}{1 - w_0}\right).
\end{aligned} \tag{A.16}$$

Let $B = E\left(\frac{\partial(\psi_0 - \mu_0)}{\partial \eta}\right) \text{Var}(\tilde{\eta}) E\left(\frac{\partial(\psi_0 - \mu_0)^T}{\partial \eta}\right)$. Then,

$$N\text{Var}(q_N^{(2)}) = \Omega^{(2)} =: \begin{pmatrix} 0 & 0 & 0 \\ 0 & B & B \\ 0 & B & B \end{pmatrix}. \tag{A.17}$$

From (A.12), $N\text{Var}(Rq_N^{(2)}) = NR\Omega^{(2)}R^T = 0$. Thus, $Rq_N^{(2)} = o_p(N^{-1/2})$. Therefore, $\hat{\beta} - \beta_0 = Rq_N^{(1)} + o_p(N^{-1/2})$. This result and (A.15) together give

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \Sigma_\beta) \quad \text{as } N \rightarrow \infty, \tag{A.18}$$

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

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

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

Proof of Theorem 2: Applying Taylor's expansion on (4) at $(\beta_0, \gamma_0, \theta_0)$ gives

$$E\left(\frac{\partial U^T}{\partial \beta}\right) (\hat{\beta}_{RRZ} - \beta_0) = -r_{n1} + r_{n2} + o_p(N^{-1/2}) \tag{A.19}$$

where

$$\begin{aligned}
r_{n1} &= N^{-1} \sum_{i=1}^N \frac{\delta_i U_{i0} - (\delta_i - w_{i0}) \psi_{i0}}{w_{i0}} \quad \text{and} \\
r_{n2} &= E\left(\frac{U_0 - \psi_0}{w_0} \frac{\partial w_0^T}{\partial \theta}\right) (\hat{\theta} - \theta_0).
\end{aligned}$$

Standard derivations show that

$$\text{Var}(r_{n1}) =: N^{-1}\tilde{\Omega}_1 = N^{-1}\left[E\left(\frac{U_0U_0^T}{w_0}\right) - E\left\{(1-w_0)\left(\frac{U_0\psi_0^T}{w_0} + \frac{\psi_0U_0^T}{w_0} - \frac{\psi_0\psi_0^T}{w_0}\right)\right\}\right] \quad (\text{A.20})$$

and

$$\begin{aligned} & -\text{Cov}(r_{n1}, r_{n2}) - \text{Cov}(r_{n2}, r_{n1}) + \text{Var}(r_{n2}) \\ =: & N^{-1}\tilde{\Omega}_2 = -N^{-1}E\left(\frac{U_0 - \psi_0}{w_0} \frac{\partial w_0^T}{\partial \theta}\right) \Lambda_\theta^{-1} E\left(\frac{\partial w_0}{\partial \theta} \frac{(U_0 - \psi_0)^T}{w_0}\right). \end{aligned} \quad (\text{A.21})$$

The Central Limit Theorem and (A.20) and (A.21) together imply that

$$\sqrt{N}(-r_{n1} + r_{n2}) \xrightarrow{d} N(0, \tilde{\Omega}_1 + \tilde{\Omega}_2). \quad (\text{A.22})$$

Theorem 2 is readily implied by (A.19) and (A.22).

Acknowledgement: We thank the Associate Editor and two referees for constructive comments and suggestions. Chen's research was supported by NSF 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.

References

- Alvarez, R.M. and Nagler, J. (1998). Economics, entitlements, and social issues: voter choice in the 1996 presidential election. *American Journal of Political Science* **42**, 1349-1363.
- Baker, S.G., Izmirlan, G. and Kipnis, V. (2005). Resolving paradoxes involving surrogate endpoints. *Journal of the Royal Statistical Society A*, **168**, 753-762.
- Baker, S.G. (2006). Surrogate endpoints: wishful thinking or reality? (Editorial). *Journal of the National Cancer Institute*, **98(8)**, 502-503
- Begg, C. and Leung, D. (2000). On the use of surrogate endpoints in randomized trials. *Journal of the Royal Statistical Society A*, **163**, 15-27.
- 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. *Canadian Journal of Statistics*, **32**, 359-372.
- Chen, S.X., Leung, D., Qin, J. (2003). Information recovery in a study with surrogate endpoints. *Journal of the American Statistical Association* **98**, 1052-1062.
- Chen, X., Hong, H. and Tamer, E. (2005). Measurement error models with auxiliary data. *Review of Economic Studies*, **72**, 343-366.
- Chen, Y.H. and Chen, H. (2000). A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society B*, **62**, 449-460.

- Clayton, D. Spiegelhalter, Dunn and Pickles (1998). Analysis of longitudinal binary data from multiphase sampling. *JRSSB*, **60**, 71-87.
- Cochran, W. G. (1977). *Sampling Techniques*. Third Edition. Wiley: New York.
- Filer, J.E. and Kenney, L.W. (1980). Voter turnout and the benefits of voting. *Public Choice*, **35**, 575-585.
- Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663-685.
- Lee, M. J. (2005). Monotonicity conditions and inequality imputation for sample-selection and non-response problems. *Econometric Reviews*, **24(2)**, 175-194.
- Lee, M. J. and Kang, S. J. (2002). Multinomial-choice and presidential election. Unpublished manuscript.
- Leung, D. (2001). Statistical methods in surrogate endpoints. *Journal of the Royal Statistical Society, Series A*, **164**, 485-503.
- Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, New York: Wiley.
- Newey, W. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, **5(2)**, 99-135.
- Owen, A. (1990). Empirical likelihood ratio confidence regions. *Annals of Statistics*, **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. *Annals of Statistics*, **22**, 300-25.

- Riker, W.H. and Ordeshook, P.C. (1968). A theory of the calculus of voting. *American Political Science Review*, **62**, 25-42.
- Robins, J. M., Hsieh, F. and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society Series B*, **57**, 409-424.
- Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, **90**, 122–129.
- Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846–866.
- Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **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. *Political Analysis*, **90**, 45-57.
- Schenker, N. and Taylor, J.M.G. (1996). Partially parametric techniques for multiple imputation. *Journal of Computational Statistics and Data Analysis*, **22**, 425-446.
- Warren, E. M., Kinder, D. R., Rosenstone, S. J. (1999). *National Election Studies 1996*, Center for Political Studies, University of Michigan, U.S.A.
- Vardi, Y. (1985). Empirical distributions in selection bias models (Com: p204-205). *The Annals of Statistics*, **13**, 178-203.

- Wittes, J., Lakatos, E. and Probstfield, J. (1989). Surrogate endpoints in clinical trials: Cardiovascular disease. *Statistics in Medicine*, **8**, 415-425.
- Wooldridge, J. (2004). Inverse probability weighted estimation for general missing data problem. CeMMAP Working Paper Number CWP05/04. Institute for Fiscal Studies, London, U.K..
- Yu, M. and Nan, B. (2006). A revisit of semiparametric regression models with missing data. *Statistica Sinica*, **16**, 1193-1212.

Table 1. Mean (Variance) of different estimators based on 2000 simulations with sample size $N = 1000$ each and bootstrap resample size 200. The second row is 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 $(\beta_1, \beta_2) = (1, 2)$; $S \sim N(1 + 2Y + X, 1)$.

Method	$\theta = (-1, 0, 0)$	$\theta = (-1, 0.2, 0.2)$	$\theta = (-1, 0.35, 0.35)$	$\theta = (-1, 0.5, 0.5)$
$\hat{\beta}_{C1}$	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)
$\hat{\beta}_{C2}$	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)
$\hat{\beta}_{W1}$	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)
$\hat{\beta}_{W2}$	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)
$\hat{\beta}_{RRZ1}$	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)
$\hat{\beta}_{RRZ2}$	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)

Table 2. Mean (Variance) of different estimators based on 2000 simulations with sample size $N = 1000$ each and bootstrap resample size 200. The second row is 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 $(\beta_1, \beta_2) = (1, 2)$; $S \sim N(1 + 2Y + X, 1)$.

Method	$\theta = (-1, 0, 0)$	$\theta = (-1, 0.2, 0.2)$	$\theta = (-1, 0.35, 0.35)$	$\theta = (-1, 0.5, 0.5)$
$\hat{\beta}_{C1}$	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)

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

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

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 on the economy)

True outcome	Surrogate outcome	Perception on 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

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

Method	Parameter estimate (Variance ¹ , Variance ²)	
	β_1	β_2
Weighted estimator	0.2989 (0.00839, 0.00642)	0.8004 (0.01682, 0.01578)
RRZ	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)

¹Variance estimate using asymptotic formulae in Theorems 1 and 2

²Variance estimate using 1000 bootstrap samples

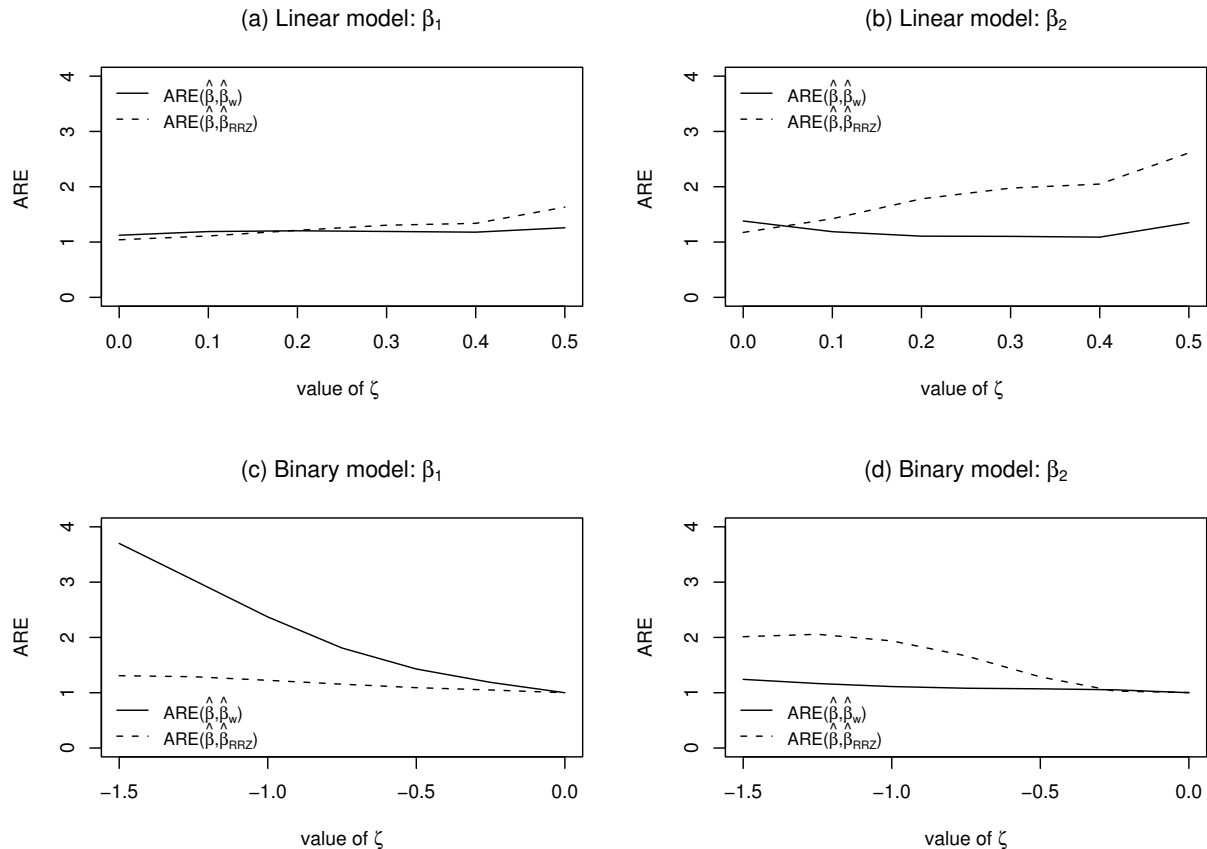


Figure 1: Asymptotic relative efficiency (ARE) between three estimators. The missing probability function is $P(\delta = 1|S = s, X = x) = \exp(\theta_1 + \theta_2 s + \theta_3 x) / \{1 + \exp(\theta_1 + \theta_2 s + \theta_3 x)\}$. (a) and (b): $Y \sim N(\beta_1 + \beta_2 X, 1)$, $S \sim N(2Y, 1)$ if $Y \geq 0$ and $S \sim N(Y, 1)$ if $Y < 0$ and $\theta \equiv (\theta_1, \theta_2, \theta_3) = (-2, \zeta, 0.5)$. (c) and (d): $P(Y = 1|X) = \exp(\beta_1 + \beta_2 X) / [1 + \exp(\beta_1 + \beta_2 X)]$, $P(S = 1|X) = \exp(\beta_1 + \beta_2(X + \zeta)) / [1 + \exp(\beta_1 + \beta_2(X + \zeta))]$, $\theta = (-3, 3, 0)$.