3-2015

Enriching Surveys with Supplementary Data and Its Application to Studying Wage Regression

Denis H. Y. LEUNG
Singapore Management University, denisleung@smu.edu.sg

Ken YAMADA
Singapore Management University, kenyamada@smu.edu.sg

Biao ZHANG
University of Toledo

Citation

LEUNG, Denis H. Y.; YAMADA, Ken; and ZHANG, Biao. Enriching Surveys with Supplementary Data and Its Application to Studying Wage Regression. (2015). *Scandinavian Journal of Statistics*. 42, (1), 155-176.
Available at: https://ink.library.smu.edu.sg/soe_research/1562

Published in Scandinavian Journal of Statistics, Volume 42, Issue 1, 1 March 2015, Pages 155-179.
http://doi.org/10.1111/sjos.12100

Enriching Surveys with Supplementary Data and its Application to Studying Wage Regression

DENIS HENG YAN LEUNG and KEN YAMADA
School of Economics, Singapore Management University

BIAO ZHANG
Department of Mathematics and Statistics, University of Toledo

ABSTRACT. We consider the problem of supplementing survey data with additional information from a population. The framework we use is very general; examples are missing data problems, measurement error models and combining data from multiple surveys. We do not require the survey data to be a simple random sample of the population of interest. The key assumption we make is that there exists a set of common variables between the survey and the supplementary data. Thus, the supplementary data serve the dual role of providing adjustments to the survey data for model consistencies and also enriching the survey data for improved efficiency. We propose a semi-parametric approach using empirical likelihood to combine data from the two sources. The method possesses favourable large and moderate sample properties. We use the method to investigate wage regression using data from the National Longitudinal Survey of Youth Study.

Key words: empirical likelihood, inverse probability weighting, selection bias, supplementary data, surveys

1. Introduction

In many empirical studies, it is necessary to combine survey data with information from additional samples from the population. As an example, Tarozzi (2007) described a study on the poverty ratio using the 1999–2000 round of the Indian National Sample Survey, where because of changes in the expenditure questionnaire over the different rounds of the survey, supplementary data were utilized to calibrate the trend of poverty ratio estimate over time. In that study, the use of supplementary data was necessary for the researcher to derive meaningful and compatible time trend parameters. Sometimes, supplementary data are used to enhance the efficiency of a study. For example, in a study of the minimum wage using the UK Labour Force Survey, Skinner *et al.* (2002) found that data on the primary variable were only available on a subset of respondents. They used data from supplementary variables to impute the missing values of the primary variable.

Works on combining surveys with supplementary data fall into a few strands. In one strand, as in Tarozzi (2007), both the survey and the supplementary data are micro data collected from different surveys. The interest of the researcher is to combine the data from the survey of interest with the supplementary data. Often, the supplementary data are not directly compatible with the data from the survey of interest. In Tarozzi (2007), weights were used to reflect the non-compatibility between the data from the survey of interest and the supplementary data. Poverty rate was modelled as a parameter in a set of weighted moment conditions, and the supplementary data played a part in estimating the weights. Other examples of combining data from different surveys can be found in Arellano & Meghir (1992), Lusardi (1996) and Merkouris (2004).

In another strand, census data are treated as auxiliary information to supplement the micro data collected in a survey. Typically, the census data are used to form known moments of some of the variables in the survey data; the moments are then used to construct moment conditions that can be combined with the information given by the survey data. This approach has been used by Imbens & Lancaster (1994). The survey and census data are combined using generalized methods of moments (Hansen, 1982). Hellerstein & Imbens (1999) also used this approach but combined the survey and census data using a semi-parametric empirical likelihood (EL; Owen, 1988). In the context of finite population sampling, Chen & Qin (1993) showed that the use of the finite population size as supplementary information effectively improves the semi-parametric efficiency in an EL.

The third strand can be found in measurement error models. In there, the survey only collects data on mis-measured counterparts of the variables of interest. Data on supplementary variables allow the researcher to recover the true relationship between the variables of interest. A commonly used assumption in the literature is a linear correlation between the mis-measured variable(s) and the variable(s) of interest (see, e.g. Cook and Stefanski, 1994; Wang *et al.*, 1997; Liang *et al.*, 2007). Recently, Chen *et al.* (2005) proposed a method that allows arbitrary correlation between the mis-measured variables and the variables of interest.

The problem we study here is related to that in the missing data literature. In there, some of the survey data are missing in the outcome variable and/or the covariates, and the supplementary data provide additional information that allows the researcher to use those observations with missing values. Some recent works can be found in Pepe (1992) and Chen & Chen (2000) for data that are missing completely at random (Little & Rubin, 2002) and in Robins *et al.* (1994) and Chen *et al.* (2008) when data are missing at random.

We now introduce the framework of this paper. Consider the situation where we have n potential observations of random variables $(\mathbf{Z}, \mathbf{S})$, where in general, $\mathbf{Z}$ and $\mathbf{S}$ may be vector valued. These n observations may be collected from one or more surveys from the target population. Typically, $\mathbf{Z} = (Y, \mathbf{X}^T)^T$, where Y represents some outcome, $\mathbf{X}$ is a vector of covariates and $\mathbf{S}$ is a vector of additional variables. In practice, $\mathbf{S}$ may include design variables used for sampling or administrative records that are available for every unit in the population, as in the case of some high quality registries. The n observations are meant to represent a simple random sample of the target population.

In reality, for the ith observation, we get to observe $(\mathbf{z}_i, \mathbf{s}_i)$ or $(\mathbf{z}_i^*, \mathbf{s}_i)$, where $\mathbf{z}_i^*$ is a sub-vector of $\mathbf{z}_i$. Let R be a 0-1 variable such that $r_i = 1$ if the observation has $(\mathbf{z}_i, \mathbf{s}_i)$ or 0 otherwise. Without loss of generality, suppose the available data can be written as $\{(\mathbf{z}_1, \mathbf{s}_1, r_1 = 1), \ldots, (\mathbf{z}_m, \mathbf{s}_m, r_m = 1), (\mathbf{z}_{m+1}^*, \mathbf{s}_{m+1}, r_{m+1} = 0), \ldots, (\mathbf{z}_n^*, \mathbf{s}_n, r_n = 0)\}$. Hereafter, we call the first m observations the complete data and the remaining $n - m$ observations the incomplete data. Notice that, under this convention, if j and j' are two incomplete observations, it is possible that $\mathbf{z}_j^*$ and $\mathbf{z}_{j'}^*$ represent different sub-vectors of their respective complete observation counterparts. Hence, it accommodates the practical situation that the contents of the incomplete information may differ between observations. In the sequel, we use $\mathbf{Z}^*$ to represent any sub-vector of $\mathbf{Z}$, which may be different in length and contents for different observations. The vector $\mathbf{S}$ is observed for all observations. The researcher may not be directly interested in $\mathbf{S}$, but together with $\mathbf{Z}^*$, it helps to adjust for selection bias when data are missing in some observations. On the contrary, $\mathbf{Z}^*$ is a sub-vector of $\mathbf{Z}$ and hence contains information about the study. We assume the target population to be an infinite population. Our method is still valid under a finite population setting if the design weights of all observations are equal, as in the case of a simple random sampling situation. However, for more complex surveys, adjustments to the method are needed (e.g. Chen and Sitter, 1999; Wu & Rao, 2006).

Let the probability that an observation from the target population belonging to the complete data satisfy the following model:

$$\Pr(R = 1|\mathbf{Z}, \mathbf{S}) = \Pr(R = 1|\mathbf{Z}^*, \mathbf{S}) = w(\mathbf{Z}^*, \mathbf{S}, \boldsymbol{\eta}), \tag{1}$$

where w is a fully specified probability distribution function for given $\boldsymbol{\eta}$, an unknown vector of parameters. Model (1) allows for selection bias of the complete data, but the selection bias is only dependent on the observables (Little & Rubin, 2002). Selection on observables may arise by design, for example, in two-phase survey sampling, where in the first phase, a simple random sample is taken, and on the basis of values of the observations in the first phase, a non-random second stage sample is taken. Alternatively, it may be the result of missing data, where subjects with certain (observable) characteristics are more (or less) likely to contain missing information. While the former is easily handled by modelling (1), the latter requires the inherent missing at random assumption (Little & Rubin, 2002). Because we do not require the complete data to be a simple random sample of the target population, the use of the complete data without any adjustments may lead to inconsistent estimates of the parameters of interest.

Our interest is in an unknown p-dimensional population parameter $\boldsymbol{\beta}$ of the distribution of $\mathbf{Z}$ and suppose that the following set of $q \geq p$ moment conditions,

$$\mathrm{E}[\mathbf{U}(\mathbf{Z}, \boldsymbol{\beta}_0)] = \mathbf{0}, \tag{2}$$

is uniquely satisfied for $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, where the expectation is taken under the distribution of $\mathbf{Z}$ in the target population. We wish to estimate $\boldsymbol{\beta}_0$.

If we assume that (1) holds, then an estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CC}}$ obtained by solving the näive sample equivalence of the moment conditions (2), viz.

$$\sum_{i=1}^{n} r_i \mathbf{U}\left(\mathbf{z}_i, \hat{\boldsymbol{\beta}}_{\mathrm{CC}}\right) = \mathbf{0}, \tag{3}$$

will generally produce a biased estimate of $\boldsymbol{\beta}$. Conversely, the inverse probability (propensity score) weighting (IPW) estimator (Horvitz & Thompson, 1952) $\hat{\boldsymbol{\beta}}_{\mathrm{IPW}}$, obtained as the solution to

$$\sum_{i=1}^{n} \frac{r_i \mathbf{U}\left(\mathbf{z}_i, \hat{\boldsymbol{\beta}}_{\mathrm{IPW}}\right)}{w\left(\mathbf{z}_i^*, \mathbf{s}_i, \hat{\boldsymbol{\eta}}\right)} = \mathbf{0}, \tag{4}$$

where $\hat{\boldsymbol{\eta}}$ is a consistent estimate of $\boldsymbol{\eta}$ by modelling (1) using both the complete and the incomplete data and is consistent under (1) if w is correctly specified. The IPW estimator has been used widely, for example, in Lipsitz *et al.* (1999), Abowd *et al.* (2001) and Nevo (2003) for estimation of conditional means in the presence of missing data and Imbens (2000) and Hirano *et al.* (2003) for the evaluation of treatment effects with non-experimental data. A recent review of the IPW estimator can be found in Graham *et al.* (2012). It has been shown by many (e.g. Wooldridge, 2007) that it is better to use an estimate of $\boldsymbol{\eta}$ on the basis of the data even when it is known. We note that the sample moment conditions (4) utilize only the observations in the complete data.

In this paper, we take a different approach. We begin with the likelihood based on the observed data

$$L = \prod_{i=1}^{m} w\left(\mathbf{z}_i^*, \mathbf{s}_i, \boldsymbol{\eta}\right) dF\left(\mathbf{z}_i, \mathbf{s}_i\right) \prod_{j=m+1}^{n} \left\{1 - w\left(\mathbf{z}_j^*, \mathbf{s}_j, \boldsymbol{\eta}\right)\right\} dF\left(\mathbf{z}_j^*, \mathbf{s}_j\right), \tag{5}$$

where $F(\mathbf{z}, \mathbf{s})$ and $F(\mathbf{z}^*, \mathbf{s})$ stand for the unknown joint cumulative distribution functions of $(\mathbf{Z}, \mathbf{S})$ and $(\mathbf{Z}^*, \mathbf{S})$, respectively. Instead of using the full likelihood (5), which may be non-robust to misspecifications of $F(\mathbf{z}, \mathbf{s})$ and $F(\mathbf{z}^*, \mathbf{s})$, we take a semi-parametric approach, where except for (1) and (2), the distribution of the data is unspecified. For a fixed w, the proposed inference for $\boldsymbol{\beta}$ is formulated semi-parametrically via a two-sample EL (Owen, 2001, p. 223) based on moment conditions from both the complete and the incomplete data. There has been growing interest in the last decade in developing model-free inferential techniques to analyse data. One such technique is the EL, which is based on estimating an unknown multinomial likelihood supported on the observations, subject to some constraints that are assumed to hold and represent the only information available in the sample. Because no statistical model is imposed on the data other than (1) and (2), EL can be viewed as a semi-parametric method. Owen (1988) and Qin & Lawless (1994) showed how to construct semi-parametric likelihood ratios from such a method. An attractive feature of the EL is that it provides a unified framework for producing both point estimates and confidence regions for $\boldsymbol{\beta}$, without requiring a full parametric model for the data. Confidence regions can be obtained from a Wilks' theorem for the log-EL ratio, and in many situations, the confidence interval is Bartlett correctable (DiCiccio *et al.*, 1991; Chen & Cui, 2006).

Our method is related to some recent works in the literature. Chen & Chen (2000) suggested a method based on the regression estimate. Chen *et al.* (2003) used a two-sample EL, one based on moment conditions from the complete data with the other based on the incomplete data. However, both methods can only be applied to situations where the complete and incomplete data are both simple random samples. On the other hand, our proposed method adjusts for the selection bias by employing the biased sampling technique of Vardi (1985). Chen *et al.* (2008) proposed a method that also allows selection bias. Their method is a two-step approach where in the first step, moment conditions based on the incomplete data are used to create weights, which are then used in an IPW estimator in the second step. Earlier, Hellerstein & Imbens (1999) also considered a two-step approach; however, instead of using observations from an incomplete data set, they assumed that incomplete information comes from some known population moments.

The rest of this paper is organized as follows. The proposed method and its large sample results are given in Section 2. We separate the case where w is fully specified, that is, $\boldsymbol{\eta}$ known and the case where w is specified up to an unknown $\boldsymbol{\eta}$. Some finite sample Monte Carlo simulation results are reported in Section 3. In Section 4, the method is illustrated using data from the National Longitudinal Survey of Youth (NLS). Concluding remarks are given in Section 5. All proofs are given in the Appendix.

2. Main results

The incomplete data serve the dual role of calibrating the complete data for valid inferences and enriching the complete data for improved efficiency. Under moment conditions (2), if $(\mathbf{Z}^{*T}, \mathbf{S}^T)^T \equiv \mathbf{Z}$ or if $(\mathbf{Z}^{*T}, \mathbf{S}^T)^T$ is perfectly correlated with $\mathbf{Z}$, then the following set of sample moment conditions

$$\sum_{i=1}^{n} \frac{r_i \mathbf{U}(\mathbf{z}_i, \boldsymbol{\beta})}{w(\mathbf{z}_i^*, \mathbf{s}_i, \boldsymbol{\eta})} + \sum_{j=m+1}^{n} \frac{(1-r_j)\mathbf{U}\left(\mathrm{E}\left(\mathbf{z}_j|\mathbf{z}_j^*, \mathbf{s}_j\right), \boldsymbol{\beta}\right)}{1 - w\left(\mathbf{z}_j^*, \mathbf{s}_j, \boldsymbol{\eta}\right)} = \mathbf{0}, \tag{6}$$

would give an estimator of $\boldsymbol{\beta}$. Otherwise, the second component of (6) is not fully known. If (6) is replaced by

$$\sum_{i=1}^{n} \frac{r_i \mathbf{U}(\mathbf{z}_i, \boldsymbol{\beta})}{w\left(\mathbf{z}_i^*, \mathbf{s}_i, \boldsymbol{\eta}\right)} + \sum_{j=m+1}^{n} \frac{(1-r_j)\boldsymbol{\psi}_{opt}\left(\mathbf{z}_j^*, \mathbf{s}_j, \boldsymbol{\gamma}, \boldsymbol{\beta}\right)}{1 - w\left(\mathbf{z}_j^*, \mathbf{s}_j, \boldsymbol{\eta}\right)} = \mathbf{0}, \tag{7}$$

where $\boldsymbol{\psi}_{opt}(\mathbf{Z}^*, \mathbf{S}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \equiv \mathrm{E}[\mathbf{U}(\mathbf{Z}, \boldsymbol{\beta})|\mathbf{Z}^*, \mathbf{S}]$ and $\boldsymbol{\gamma}$ represents a vector of parameters that are known or to be estimated separately, then the solution of (7) gives a consistent and semi-parametrically efficient estimator of $\boldsymbol{\beta}$ as long as w is correctly specified (Robins *et al.*, 1994). However, $\boldsymbol{\psi}_{opt}(\mathbf{Z}^*, \mathbf{S}, \boldsymbol{\gamma}, \boldsymbol{\beta})$ is usually unknown and therefore (7) cannot be directly applied. A number of earlier works have attempted to estimate $\boldsymbol{\psi}_{opt}(\mathbf{Z}^*, \mathbf{S}, \boldsymbol{\gamma}, \boldsymbol{\beta})$ (e.g. Robins *et al.*, 1994; Chen *et al.*, 2008). However, the estimation of $\boldsymbol{\psi}_{opt}(\mathbf{Z}^*, \mathbf{S}, \boldsymbol{\gamma}, \boldsymbol{\beta})$ may be difficult in many situations. In this paper, we take a different approach. We do not directly estimate $\boldsymbol{\psi}_{opt}(\mathbf{Z}^*, \mathbf{S}, \boldsymbol{\gamma}, \boldsymbol{\beta})$. Instead we find a 'working' $\boldsymbol{\psi}(\mathbf{Z}^*, \mathbf{S}, \boldsymbol{\gamma}, \boldsymbol{\beta})$, which is defined as an arbitrary set of $q \geq p$ moment equations such that

$$\mathrm{E}\left[\boldsymbol{\psi}\left(\mathbf{Z}^*, \mathbf{S}, \boldsymbol{\gamma}_0, \boldsymbol{\beta}_0\right)\right] = \mathbf{0},$$

for some value $\boldsymbol{\gamma}_0$, and the true value $\boldsymbol{\beta}_0$ is defined in (2).

In practice, the following approach may be used to find $\boldsymbol{\psi}(\mathbf{Z}^*, \mathbf{S}, \boldsymbol{\gamma}, \boldsymbol{\beta})$. Let $\tilde{\boldsymbol{\beta}}$ be any consistent estimate of $\boldsymbol{\beta}$. A natural choice for $\tilde{\boldsymbol{\beta}}$ is the IPW estimate $\hat{\boldsymbol{\beta}}_{\mathrm{IPW}}$. Then on the basis of the complete data, $\boldsymbol{\psi}(\mathbf{Z}^*, \mathbf{S}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}})$ may be found as the projection of $\mathbf{U}(\mathbf{Z}, \tilde{\boldsymbol{\beta}})$ on the space spanned by $(\mathbf{Z}^*, \mathbf{S})$. Importantly, we point out that with the estimated $\tilde{\boldsymbol{\gamma}}$, as long as it converges to $\boldsymbol{\gamma}_0$ at the rate of $n^{-1/2}$, then all the large sample results remain unchanged. The reason can be appreciated by observing (34) in the Appendix, where it is inconsequential to replace $\boldsymbol{\gamma}_0$ with $\tilde{\boldsymbol{\gamma}}$ because the first factor in (34) has zero mean. Hence, we assume that $\boldsymbol{\gamma}$ is known in $\boldsymbol{\psi}(\mathbf{Z}^*, \mathbf{S}, \boldsymbol{\gamma}, \boldsymbol{\beta})$ in the rest of this paper, and we suppress $\boldsymbol{\gamma}$ to simplify exposition.

For simplicity, write $\mathbf{U}_i(\boldsymbol{\beta}) \equiv \mathbf{U}(\mathbf{z}_i, \boldsymbol{\beta}), i = 1, \ldots, m$, $\boldsymbol{\psi}_j(\boldsymbol{\beta}) \equiv \boldsymbol{\psi}(\mathbf{z}_j^*, \mathbf{s}_j, \boldsymbol{\beta}), j = m + 1, \ldots, n$, $w_i(\boldsymbol{\eta}) \equiv w(\mathbf{z}_i^*, \mathbf{s}_i, \boldsymbol{\eta}), i = 1, \ldots, n$. Throughout the paper, the symbol $\mathbf{0}$ is used to denote a zero vector or a null matrix. For any matrix $\mathbf{A}$, $\mathbf{A} \geq \mathbf{0}$ means that the matrix is positive semi-definite. For any matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \leq \mathbf{B}$ implies that $\mathbf{B} - \mathbf{A}$ is a non-negative definite matrix.

On the basis of the observed data, the likelihood (5) can be rewritten as

$$[\mathrm{E}(w_0)]^m[1 - \mathrm{E}(w_0)]^{n-m} \prod_{i=1}^{m} \frac{w_i(\boldsymbol{\eta})dF(\mathbf{z}_i, \mathbf{s}_i)}{\mathrm{E}(w_0)} \prod_{j=m+1}^{n} \frac{(1 - w_j(\boldsymbol{\eta}))dF\left(\mathbf{z}_j^*, \mathbf{s}_j\right)}{1 - \mathrm{E}(w_0)}, \tag{8}$$

where

$$\mathrm{E}(w_0) = \int w\left(\mathbf{z}^*, \mathbf{s}, \boldsymbol{\eta}_0\right) dF(\mathbf{z}, \mathbf{s}) = \int w\left(\mathbf{z}^*, \mathbf{s}, \boldsymbol{\eta}_0\right) dF(\mathbf{z}^*, \mathbf{s}).$$

The first and second terms of (8) are the binomial likelihood of the proportions of complete and incomplete data. The last two terms are the conditional likelihoods by conditioning on the complete and the incomplete observations, respectively. Note that the last two terms come from the biased sampling problem discussed in Vardi (1982, 1985). Following Vardi (1982, 1985), let $p_i = dF(\mathbf{z}_i, \mathbf{s}_i), i = 1, 2, ..., m$ and $q_j = dF(\mathbf{z}_j^*, \mathbf{s}_j), j = m + 1, ..., n$ be the jump sizes of the distributions. The semi-parametric log-EL is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\eta}) = \sum_{i=1}^{m} \log w_i(\boldsymbol{\eta}) + \sum_{j=m+1}^{n} \log\{1 - w_j(\boldsymbol{\eta})\} + \sum_{i=1}^{m} \log p_i + \sum_{j=m+1}^{n} \log q_j. \tag{9}$$

Note that for any function $\mathbf{h}(\mathbf{Z}^*, \mathbf{S})$,

$$\int \mathbf{h}(\mathbf{z}^*, \mathbf{s})dF(\mathbf{z}, \mathbf{s}) = \int \mathbf{h}(\mathbf{z}^*, \mathbf{s})dF(\mathbf{z}^*, \mathbf{s}) = \boldsymbol{\mu}, \tag{10}$$

and

$$\int \mathbf{U}(\mathbf{z}, \boldsymbol{\beta}) dF(\mathbf{z}) = \int \mathbf{U}(\mathbf{z}, \boldsymbol{\beta}) dF(\mathbf{z}, \mathbf{s}) = \mathbf{0}. \tag{11}$$

The log-EL (9) can be maximized with respect to the following constraints:

$$\sum_{i=1}^{m} p_i = 1, \qquad p_i \geq 0, \qquad \sum_{j=m+1}^{n} q_j = 1, \qquad q_j \geq 0, \tag{12}$$

$$\sum_{i=1}^{m} p_i \mathbf{U}_i(\boldsymbol{\beta}) = \mathbf{0}, \tag{13}$$

$$\sum_{i=1}^{m} p_i \left\{ \mathbf{h}_i(\boldsymbol{\eta}, \boldsymbol{\beta}) - \boldsymbol{\mu} \right\} = \mathbf{0}, \qquad \sum_{j=m+1}^{n} q_j \left\{ \mathbf{h}_j(\boldsymbol{\eta}, \boldsymbol{\beta}) - \boldsymbol{\mu} \right\} = \mathbf{0}, \tag{14}$$

where

$$\mathbf{h}_i(\boldsymbol{\eta}, \boldsymbol{\beta}) = \left(w_i(\boldsymbol{\eta}), \boldsymbol{\psi}_i^T(\boldsymbol{\beta}) \right)^T.$$

Note that the constraints $\sum_{i=1}^{m} p_i w_i(\boldsymbol{\eta}) = \sum_{j=m+1}^{n} q_j w_j(\boldsymbol{\eta}) = \mathrm{E}(w_0)$ are necessary. They reflect the fact that the complete data are not necessarily a simple random sample from the target population. The constraints $\sum_{i=1}^{m} p_i \boldsymbol{\psi}_i(\boldsymbol{\beta}) = \sum_{j=m+1}^{n} q_j \boldsymbol{\psi}_j(\boldsymbol{\beta}) = \int \boldsymbol{\psi}(\mathbf{z}^*, \mathbf{s}, \boldsymbol{\beta}) dF(\mathbf{z}^*, \mathbf{s})$, on the other hand, are optional. However, they can improve the estimation efficiency.

2.1. $\boldsymbol{\eta}$ known

First, we consider the case that $w(\boldsymbol{\eta})$ is completely known; that is, we use $\boldsymbol{\eta} = \boldsymbol{\eta}_0$. We only need to maximize the second part of the log-EL. Introducing Lagrange multipliers $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$ and $\boldsymbol{\nu}$, the values of p_i and q_j that maximize (9) subject to the constraints (12)–(14) give

$$p_i = \frac{1}{m} \frac{1}{1 + \boldsymbol{\lambda}_1^T \left\{ \mathbf{h}_i(\boldsymbol{\eta}_0, \boldsymbol{\beta}) - \boldsymbol{\mu} \right\} + \boldsymbol{\lambda}_2^T \mathbf{U}_i(\boldsymbol{\beta})}, \quad i = 1, 2, \ldots, m, \tag{15}$$

$$q_j = \frac{1}{n-m} \frac{1}{1 + \boldsymbol{\nu}^T \left\{ \mathbf{h}_j(\boldsymbol{\eta}_0, \boldsymbol{\beta}) - \boldsymbol{\mu} \right\}}, \quad j = m+1, \ldots, n, \tag{16}$$

which upon substituting back to the second part of (9), gives

$$\ell(\boldsymbol{\beta}, \boldsymbol{\mu}) = -\sum_{i=1}^{m} \log\left[1 + \boldsymbol{\lambda}_1^T \left\{ \mathbf{h}_i(\boldsymbol{\eta}_0, \boldsymbol{\beta}) - \boldsymbol{\mu} \right\} + \boldsymbol{\lambda}_2^T \mathbf{U}_i(\boldsymbol{\beta})\right] - \sum_{j=m+1}^{n} \log\left[1 + \boldsymbol{\nu}^T \left\{ \mathbf{h}_j(\boldsymbol{\eta}_0, \boldsymbol{\beta}) - \boldsymbol{\mu} \right\}\right], \tag{17}$$

with the restriction that the Lagrange multipliers satisfy the following equations:

$$\sum_{i=1}^{m} \frac{\mathbf{U}_i(\boldsymbol{\beta})}{1 + \boldsymbol{\lambda}_1^T \left\{ \mathbf{h}_i(\boldsymbol{\eta}_0, \boldsymbol{\beta}) - \boldsymbol{\mu} \right\} + \boldsymbol{\lambda}_2^T \mathbf{U}_i(\boldsymbol{\beta})} = \mathbf{0},$$

$$\sum_{i=1}^{m} \frac{\mathbf{h}_i(\boldsymbol{\eta}_0, \boldsymbol{\beta}) - \boldsymbol{\mu}}{1 + \boldsymbol{\lambda}_1^T \left\{ \mathbf{h}_i(\boldsymbol{\eta}_0, \boldsymbol{\beta}) - \boldsymbol{\mu} \right\} + \boldsymbol{\lambda}_2^T \mathbf{U}_i(\boldsymbol{\beta})} = \mathbf{0},$$

$$\sum_{j=m+1}^{n} \frac{\mathbf{h}_j(\boldsymbol{\eta}_0, \boldsymbol{\beta}) - \boldsymbol{\mu}}{1 + \boldsymbol{\nu}^T \left\{ \mathbf{h}_j(\boldsymbol{\eta}_0, \boldsymbol{\beta}) - \boldsymbol{\mu} \right\}} = \mathbf{0}.$$

Differentiating (17) with respect to $\boldsymbol{\mu}$ gives

$$\frac{\partial \ell(\boldsymbol{\beta},\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{m} \frac{\boldsymbol{\lambda}_1}{1+\boldsymbol{\lambda}_1^T\left\{\mathbf{h}_i(\boldsymbol{\eta}_0,\boldsymbol{\beta})-\boldsymbol{\mu}\right\}+\boldsymbol{\lambda}_2^T\mathbf{U}_i(\boldsymbol{\beta})} - \sum_{j=m+1}^{n} \frac{-\boldsymbol{\nu}}{1+\boldsymbol{\nu}^T\left\{\mathbf{h}_j(\boldsymbol{\eta}_0,\boldsymbol{\beta})-\boldsymbol{\mu}\right\}} = \mathbf{0}. \tag{18}$$

Using the constraints in (12) and combining (15), (16) and (18) give

$$m\boldsymbol{\lambda}_1 + (n-m)\boldsymbol{\nu} = \mathbf{0} \quad \text{or} \quad \boldsymbol{\nu} = -\frac{m}{n-m}\boldsymbol{\lambda}_1.$$

The Lagrange multipliers can be reparametrized as

$$\boldsymbol{\tau}_1 = \left\{\frac{m}{n}\left(1-\boldsymbol{\lambda}_1^T\boldsymbol{\mu}\right), \frac{m}{n}\boldsymbol{\lambda}_1^T\right\}^T, \quad \boldsymbol{\tau}_2 = \frac{m}{n}\boldsymbol{\lambda}_2.$$

Under the new parameters, and write $\mathbf{H}_i(\boldsymbol{\eta}_0,\boldsymbol{\beta}) = \left(1, \mathbf{h}_i^T(\boldsymbol{\eta}_0,\boldsymbol{\beta})\right)^T, i = 1,\ldots,n$, the constraints (13) and (14) can be rewritten as

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{r_i\mathbf{H}_i(\boldsymbol{\eta}_0,\boldsymbol{\beta})}{\boldsymbol{\tau}_1^T\mathbf{H}_i(\boldsymbol{\eta}_0,\boldsymbol{\beta})+\boldsymbol{\tau}_2^T\mathbf{U}_i(\boldsymbol{\beta})} - \frac{(1-r_i)\mathbf{H}_i(\boldsymbol{\eta}_0,\boldsymbol{\beta})}{1-\boldsymbol{\tau}_1^T\mathbf{H}_i(\boldsymbol{\eta}_0,\boldsymbol{\beta})}\right] \equiv \mathbf{g}_1(\boldsymbol{\tau}_1,\boldsymbol{\tau}_2,\boldsymbol{\beta},\boldsymbol{\eta}_0) = \mathbf{0}, \tag{19}$$

$$\sum_{i=1}^{n}\frac{r_i\mathbf{U}_i(\boldsymbol{\beta})}{\boldsymbol{\tau}_1^T\mathbf{H}_i(\boldsymbol{\eta}_0,\boldsymbol{\beta})+\boldsymbol{\tau}_2^T\mathbf{U}_i(\boldsymbol{\beta})} \equiv \mathbf{g}_2(\boldsymbol{\tau}_1,\boldsymbol{\tau}_2,\boldsymbol{\beta},\boldsymbol{\eta}_0) = \mathbf{0}. \tag{20}$$

The log-EL (17) becomes

$$\ell(\boldsymbol{\beta}) = -\sum_{i=1}^{n} r_i \log\left\{\boldsymbol{\tau}_1^T\mathbf{H}_i(\boldsymbol{\eta}_0,\boldsymbol{\beta})+\boldsymbol{\tau}_2^T\mathbf{U}_i(\boldsymbol{\beta})\right\} - \sum_{i=1}^{n}(1-r_i)\log\left\{1-\boldsymbol{\tau}_1^T\mathbf{H}_i(\boldsymbol{\eta}_0,\boldsymbol{\beta})\right\}. \tag{21}$$

Differentiating (21) with respect to $\boldsymbol{\beta}$ gives

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{n} r_i \frac{\boldsymbol{\tau}_1^T\frac{\partial \mathbf{H}_i(\boldsymbol{\eta}_0,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + \boldsymbol{\tau}_2^T\frac{\partial \mathbf{U}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}}{\boldsymbol{\tau}_1^T\mathbf{H}_i(\boldsymbol{\eta}_0,\boldsymbol{\beta})+\boldsymbol{\tau}_2^T\mathbf{U}_i(\boldsymbol{\beta})} + (1-r_i)\frac{\boldsymbol{\tau}_1^T\frac{\partial \mathbf{H}_i(\boldsymbol{\eta}_0,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}}{1-\boldsymbol{\tau}_1^T\mathbf{H}_i(\boldsymbol{\eta}_0,\boldsymbol{\beta})} \equiv \mathbf{g}_3(\boldsymbol{\tau}_1,\boldsymbol{\tau}_2,\boldsymbol{\beta},\boldsymbol{\eta}_0) = \mathbf{0}. \tag{22}$$

Define $\boldsymbol{\theta} = \left(\boldsymbol{\tau}_1^T, \boldsymbol{\tau}_2^T, \boldsymbol{\beta}^T\right)^T$ and let $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\tau}}_1^T, \hat{\boldsymbol{\tau}}_2^T, \hat{\boldsymbol{\beta}}^T\right)^T$ be the solution of $\mathbf{g} \equiv \left(\mathbf{g}_1^T, \mathbf{g}_2^T, \mathbf{g}_3^T\right)^T = \mathbf{0}$.

We have the following results:

Theorem 1. *Under conditions C1–C4 in the Appendix, then solving* $\mathbf{g} = \mathbf{0}$ *gives*

$$n^{1/2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right) \xrightarrow{d} MVN\left(\mathbf{0}, \left(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\right),$$

where

$$\mathbf{A} = \begin{pmatrix} \mathrm{E}\left(\frac{\mathbf{H}_0^T\mathbf{H}_0}{w_0(1-w_0)}\right) & \mathrm{E}\left(\frac{\mathbf{H}_0^T\mathbf{U}_0}{w_0}\right) \\ \mathrm{E}\left(\frac{\mathbf{U}_0^T\mathbf{H}_0}{w_0}\right) & \mathrm{E}\left(\frac{\mathbf{U}_0^T\mathbf{U}_0}{w_0}\right) \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{0} \\ \mathrm{E}\left(\frac{\partial \mathbf{U}_0}{\partial \boldsymbol{\beta}}\right) \end{pmatrix},$$

$\mathbf{U}_0 \overset{d}{=} \mathbf{U}_i(\boldsymbol{\beta}_0)$, $w_0 \overset{d}{=} w_i(\boldsymbol{\eta}_0)$, $\mathbf{H}_0 \overset{d}{=} \mathbf{H}_i(\boldsymbol{\eta}_0, \boldsymbol{\beta}_0)$ *and* $\overset{d}{=}$ *stands for equivalence in distributions.*

Furthermore, the difference of the asymptotic covariances of $\hat{\boldsymbol{\beta}}$ *and* $\hat{\boldsymbol{\beta}}_{IPW}$ *is positive semi-definite, where* $\hat{\boldsymbol{\beta}}_{IPW}$ *is the IPW estimator.*

In fact, using simple matrix algebra, the variance of $\hat{\boldsymbol{\beta}}$ can be shown to be

$$\mathrm{E}^{-1}\left(\frac{\partial \mathbf{U}_0}{\partial \boldsymbol{\beta}}\right)\left[\mathrm{E}\left(\frac{\mathbf{U}_0^T\mathbf{U}_0}{w_0}\right) - \mathrm{E}\left(\frac{\mathbf{U}_0^T\mathbf{H}_0}{w_0}\right)\mathrm{E}^{-1}\left(\frac{\mathbf{H}_0^T\mathbf{H}_0}{w_0(1-w_0)}\right)\mathrm{E}\left(\frac{\mathbf{H}_0^T\mathbf{U}_0}{w_0}\right)\right]\mathrm{E}^{-1}\left(\frac{\partial \mathbf{U}_0}{\partial \boldsymbol{\beta}^T}\right). \tag{23}$$

The first term in (23) is the variance of the IPW estimator, and the second term represents the reduction of variance due to using $\hat{\boldsymbol{\beta}}$. The reduction depends on two factors. First, the function w_0 and second, the correlation between $\mathbf{H}_0$ and $\mathbf{U}_0$. In particular, the reduction is high if the 'correlation' between $\mathbf{H}_0$ and $\mathbf{U}_0$ is high or if the proportion of incomplete data, that is, $(n-m)/n$ is high.

Note that the dimensions of $\mathbf{H}_i(\boldsymbol{\eta}_0, \boldsymbol{\beta}) = \left(1, w_i(\boldsymbol{\eta}_0), \boldsymbol{\psi}_i^T(\boldsymbol{\beta})\right)^T$ and $\mathbf{U}_i(\boldsymbol{\beta})$ are $p+2$ and p, respectively. By a property of positive matrices, it is shown in the Appendix that

$$-\mathrm{E}\left(\frac{\mathbf{U}_0^T\mathbf{H}_0}{w_0}\right)\mathrm{E}^{-1}\left(\frac{\mathbf{H}_0^T\mathbf{H}_0}{w_0(1-w_0)}\right)\mathrm{E}\left(\frac{\mathbf{H}_0^T\mathbf{U}_0}{w_0}\right) \leq -\mathrm{E}\left(\frac{\mathbf{U}_0^T\boldsymbol{\psi}_0}{w_0}\right)\mathrm{E}^{-1}\left(\frac{\boldsymbol{\psi}_0^T\boldsymbol{\psi}_0}{w_0(1-w_0)}\right)\mathrm{E}\left(\frac{\mathbf{U}_0^T\boldsymbol{\psi}_0}{w_0}\right),$$

where $\boldsymbol{\psi}_0 \overset{d}{=} \boldsymbol{\psi}_i(\boldsymbol{\beta}_0)$.

The following theorem motivates the form of $\boldsymbol{\psi}(\boldsymbol{\beta})$ that we should use.

Theorem 2. *For any measurable function* $\boldsymbol{\psi}(\boldsymbol{\beta})$*, let* $\boldsymbol{\psi}_i(\boldsymbol{\beta}) \overset{d}{=} \boldsymbol{\psi}_0$*, then*

$$\begin{aligned} &-\mathrm{E}\left(\frac{\mathbf{U}_0^T\boldsymbol{\psi}_0(1-w_0)}{w_0}\right)\mathrm{E}^{-1}\left(\frac{\boldsymbol{\psi}_0^T\boldsymbol{\psi}_0}{w_0}\right)\mathrm{E}\left(\frac{\mathbf{U}_0^T\boldsymbol{\psi}_0(1-w_0)}{w_0}\right), \\ &\leq -\mathrm{E}\left(\frac{\mathbf{U}_0^T\boldsymbol{\psi}_0}{w_0}\right)\mathrm{E}^{-1}\left(\frac{\boldsymbol{\psi}_0^T\boldsymbol{\psi}_0}{w_0(1-w_0)}\right)\mathrm{E}\left(\frac{\mathbf{U}_0^T\boldsymbol{\psi}_0}{w_0}\right). \end{aligned}$$

Therefore, according to the theorem, if the best guess of $\boldsymbol{\psi}_{opt}(\boldsymbol{\beta})$ is $\boldsymbol{\psi}_{opt,g}(\boldsymbol{\beta})$, then we should set $\boldsymbol{\psi}(\boldsymbol{\beta}) = (1-w(\boldsymbol{\eta}))\boldsymbol{\psi}_{opt,g}(\boldsymbol{\beta})$.

If the interest is in testing the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, an EL-ratio statistic can be based on

$$R(\boldsymbol{\beta}_0) = 2\left\{\max_{\boldsymbol{\tau}_1,\boldsymbol{\tau}_2,\boldsymbol{\beta}} \ell(\boldsymbol{\tau}_1,\boldsymbol{\tau}_2,\boldsymbol{\beta}) - \max_{\boldsymbol{\tau}_1,\boldsymbol{\tau}_2} \ell(\boldsymbol{\tau}_1,\boldsymbol{\tau}_2,\boldsymbol{\beta}_0)\right\},$$

whose large sample properties are summarized in the succeeding text.

Theorem 3. *Under conditions C1–C4 in the Appendix, under the hypothesis:* $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$*,*

$$R(\boldsymbol{\beta}_0) \xrightarrow{d} \chi^2(p),$$

where p *is the dimension of* $\boldsymbol{\beta}$*.*

2.2. $\boldsymbol{\eta}$ unknown

If $\boldsymbol{\eta}$ is unknown in the propensity score function, the full log-EL is

$$\ell_F(\boldsymbol{\beta},\boldsymbol{\eta}) = \sum_{i=1}^{n} [r_i \log w_i(\boldsymbol{\eta}) + (1-r_i)\log(1-w_i(\boldsymbol{\eta}))] \\ - \sum_{i=1}^{n} r_i \log\left\{\boldsymbol{\tau}_1^T \mathbf{H}_i(\boldsymbol{\eta},\boldsymbol{\beta}) + \boldsymbol{\tau}_2^T \mathbf{U}_i(\boldsymbol{\beta})\right\} - \sum_{i=1}^{n}(1-r_i)\log\left\{1-\boldsymbol{\tau}_1^T \mathbf{H}_i(\boldsymbol{\eta},\boldsymbol{\beta})\right\}. \tag{24}$$

One possible approach is to maximize (24) directly with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ simultaneously. However, this approach may encounter computational problems as $\ell_F(\boldsymbol{\beta},\boldsymbol{\eta})$ is very complicated. Therefore, in this paper, we take a different approach by maximizing the two terms in (24) separately; that is, (1) maximize the binomial log-likelihood with respect to $\boldsymbol{\eta}$ to obtain $\tilde{\boldsymbol{\eta}}$, and (2) for fixed $\boldsymbol{\eta}$ at $\tilde{\boldsymbol{\eta}}$, maximize the second term in (24), that is, maximize the log-EL. Let

$$\mathbf{g}_0 = \frac{1}{n}\sum_{i=1}^{n} r_i \frac{\partial \log w_i(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} + (1-r_i)\frac{\partial \log(1-w_i(\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}} = \frac{1}{n}\sum_{i=1}^{n} \frac{r_i - w_i(\boldsymbol{\eta})}{w_i(\boldsymbol{\eta})(1-w_i(\boldsymbol{\eta}))}\frac{\partial w_i}{\partial \boldsymbol{\eta}},$$

be the score function based on the binomial likelihood. Suppose $\hat{\boldsymbol{\theta}}^+ = \left(\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\eta}}^T\right)^T$ solves $\mathbf{g}^+ \equiv \left(\mathbf{g}_0^T, \mathbf{g}_1^T, \mathbf{g}_2^T, \mathbf{g}_3^T\right)^T = \mathbf{0}$, the following result for $\hat{\boldsymbol{\beta}}$ when $\boldsymbol{\eta}$ is unknown can be obtained.

Theorem 4. *Under conditions C1–C4 in the Appendix, solving* $\mathbf{g}^+ = \mathbf{0}$ *gives*

$$n^{1/2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} MVN\left(\mathbf{0}, \left(\tilde{\mathbf{B}}^T\mathbf{C}^{-1}\tilde{\mathbf{B}}\right)^{-1}\right),$$

where

$$\tilde{\mathbf{B}} = \begin{pmatrix} \mathbf{0} \\ \mathbf{B} \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} \mathbf{W}_{11} & -\mathbf{V}_{21}^T \\ -\mathbf{V}_{21} & \mathbf{A} \end{pmatrix},$$

with

$$\mathbf{W}_{11} = \mathrm{E}\left(\frac{\left(\frac{\partial w_0}{\partial \boldsymbol{\eta}}\right)^T\left(\frac{\partial w_0}{\partial \boldsymbol{\eta}}\right)}{w_0(1-w_0)}\right), \quad \mathbf{V}_{21}^T = -\left(\mathrm{E}\left(\frac{\left(\frac{\partial w_0}{\partial \boldsymbol{\eta}}\right)^T\mathbf{H}_0}{w_0(1-w_0)}\right), \mathrm{E}\left(\frac{\left(\frac{\partial w_0}{\partial \boldsymbol{\eta}}\right)^T\mathbf{U}_0}{w_0}\right)\right).$$

Using simple matrix algebra, $\left(\tilde{\mathbf{B}}^T\mathbf{C}^{-1}\tilde{\mathbf{B}}\right)^{-1}$ can be written as $\left(\mathbf{B}^T(\mathbf{A}+\mathbf{V}_{21}\mathbf{W}_{11}\mathbf{V}_{21}^T)^{-1}\mathbf{B}\right)^{-1} \geq \left(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1}$ from theorem 1. Notice from theorem 4, the forms of $\mathbf{W}_{11}$ and $\mathbf{V}_{21}$ suggest that the extra variation is due to the estimation of $\boldsymbol{\eta}$. In practice, $\boldsymbol{\eta}$ is likely to be unknown and needs to be estimated along with the parameter of interest, $\boldsymbol{\beta}$. Hence, the foregoing comparison allows us to assess the efficiency loss as a result of an unknown $\boldsymbol{\eta}$. We note in passing that our method of separating the estimation of the nuisance parameter $\boldsymbol{\eta}$ from the parameter of interest $\boldsymbol{\beta}$ is related to the pseudo-maximum likelihood estimation described in Gong & Samaniego (1981). They derived the asymptotic consistency and asymptotic variance formula of the pseudo-maximum likelihood estimate when the nuisance parameter is replaced by an estimate (Gong & Samaniego, 1981, theorem 2, eq. 2.6). Our expression $\left(\mathbf{B}^T(\mathbf{A}+\mathbf{V}_{21}\mathbf{W}_{11}\mathbf{V}_{21}^T)^{-1}\mathbf{B}\right)^{-1}$ generalizes their equation (2.6) in the current context, when the nuisance parameter estimate is assumed to be asymptotically consistent (cf. Gong and Samaniego, 1981, p. 865, Remarks).

3. Monte Carlo simulations

In this section, we give the results of a simulation study designed to evaluate the finite sample properties of the proposed estimator. We compared the proposed estimator (EL) with three other estimators:

(i) The maximum likelihood estimator $\hat{\boldsymbol{\beta}}_C$ 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 incomplete data.

(ii) The IPW estimator $\hat{\boldsymbol{\beta}}_{\mathrm{IPW}}$ obtained as the solution of

$$\sum_{i=1}^{m} \frac{1}{w_i(\hat{\boldsymbol{\eta}})} \mathbf{U}_i(\boldsymbol{\beta}) = \mathbf{0}. \tag{25}$$

This version of the IPW differs from the original IPW estimator in the use of an estimator $\hat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$ in $w(\boldsymbol{\eta})$. Our choice of using an estimator $\hat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$ is due to a well-known statistical advantage with estimated over the true propensity score (see, e.g. Imbens, 1992; Hirano *et al.*, 2003; Wooldridge, 2007).

(iii) The doubly robust estimator of Rotnitzky *et al.* (1998) (hereafter denoted as DR) $\hat{\boldsymbol{\beta}}_{\mathrm{DR}}$ that solves the moment condition (7) with $\boldsymbol{\psi}_{opt}(\boldsymbol{\beta})$ replaced by an estimate $\boldsymbol{\psi}_{\mathrm{DR}}(\boldsymbol{\beta})$.

If $\boldsymbol{\psi}_{\mathrm{DR}}(\boldsymbol{\beta}) \equiv \boldsymbol{\psi}_{opt}(\boldsymbol{\beta})$ and $w(\boldsymbol{\eta})$ is correctly specified, then $\hat{\boldsymbol{\beta}}$ attains the semi-parametric efficiency bound within the class of estimating functions generated by $\mathbf{U}(\boldsymbol{\beta})$ for estimating $\boldsymbol{\beta}$ (Newey, 1990). Furthermore, $\hat{\boldsymbol{\beta}}_{\mathrm{DR}}$ is consistent if *either* $w(\boldsymbol{\eta})$ or $\boldsymbol{\psi}_{\mathrm{DR}}(\boldsymbol{\beta})$ is correctly specified. This property is the so-called doubly robustness property. However, $\hat{\boldsymbol{\beta}}_{\mathrm{DR}}$ may suffer efficiency loss when $\boldsymbol{\psi}_{\mathrm{DR}}(\boldsymbol{\beta}) \neq \boldsymbol{\psi}_{opt}(\boldsymbol{\beta})$ even if $w(\boldsymbol{\eta})$ is correct.

Two models were used in our simulation study. In both models, $\mathbf{Z} = (Y, X)^T$ where Y is an univariate outcome variable and X is a univariate covariate; $Z^* = X$ and S represents a proxy variable for Y. In model 1, Y and S are given by the models

$$Y = \boldsymbol{\beta}^T(1, X) + e \quad \text{and} \quad S = Y + \tilde{e},$$

where $X, e, \tilde{e}$ are independently distributed as $N(0, 1)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$. For the complete data,

$$\mathbf{U}(\boldsymbol{\beta}) \equiv (1, X)^T \left\{Y - \boldsymbol{\beta}^T(1, X)\right\}. \tag{26}$$

For this model, we used

$$\boldsymbol{\psi}_{\mathrm{DR}}(\boldsymbol{\beta}) = (1, X)^T \left\{\boldsymbol{\gamma}^T(1, X, S) - \boldsymbol{\beta}^T(1, X)\right\}, \tag{27}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)^T$ for $\hat{\boldsymbol{\beta}}_{\mathrm{DR}}$ and $\boldsymbol{\psi}(\boldsymbol{\beta}) = \{1 - w(\boldsymbol{\eta})\}\boldsymbol{\psi}_{\mathrm{DR}}(\boldsymbol{\beta})$ for $\hat{\boldsymbol{\beta}}$. The initial estimate for $\boldsymbol{\gamma}$ is obtained by fitting an ordinary least squares on the model

$$\mathrm{E}(Y|S, X) = \boldsymbol{\gamma}^T(1, S, X). \tag{28}$$

As we pointed out in Section 2, (28) does not need to be optimal. The goal is to extract as much information as possible on $\boldsymbol{\beta}$ using S and X from the incomplete data. We note in passing that for a model like this, simpler methods are available for analysis of the data. For example, if we make the following additional assumptions: (i) that the selection bias is independent of S given

X, and (ii) that if the relationship between S and Y is known, then the model can be fitted using a weighted regression with the moment condition:

$$\sum_{i=1}^{m} t_i(1,x_i)^T\left\{y_i-\boldsymbol{\beta}^T(1,x_i)\right\}+\sum_{j=m+1}^{n} t_j(1,x_j)^T\left\{s_j-\boldsymbol{\beta}^T(1,x_j)\right\},$$

where t_i, t_j are appropriate inverse weights to deal with heteroscedasticity. In that case, the method in this paper is asymptotically equivalent to this simpler method. In practice, of course, there may be selection bias, and we would not know the true form of S, but the proposed method would remain consistent even if the form of $\boldsymbol{\psi}(\boldsymbol{\beta})$ is incorrect.

In model 2, Y and S are defined by

$$Y=\boldsymbol{\beta}^T(1,X)\times e,\quad S=Y+\tilde{e},$$

where $X\sim 0.25\chi^2(1)$ and $e\sim\exp(1)$, $\tilde{e}\sim N(0,1)$. The forms of $\mathbf{U}(\beta)$, $\boldsymbol{\psi}_{\mathrm{DR}}(\boldsymbol{\beta})$ and $\boldsymbol{\psi}(\boldsymbol{\beta})$ are the same as those for model 1.

To implement the estimator proposed in this paper, we adopted the algorithm of Chen *et al.* (2002) with the following modifications:

(i) For fixed $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, the Lagrange multipliers $(\hat{\boldsymbol{\tau}}_1,\hat{\boldsymbol{\tau}}_2)$ are solved using the constraint equations. However, in each iteration, we must ensure the conditions $\boldsymbol{\tau}_1^T\mathbf{H}_i+\boldsymbol{\tau}_2^T\mathbf{U}_i>0$ and $1-\boldsymbol{\tau}_1^T\mathbf{H}_i>0$ hold for each $i=1,\ldots,n$. If at least one of the two conditions is not satisfied, the step size is reduced by half and the conditions re-evaluated. The reduction is carried out until both conditions are satisfied.

(ii) Using $(\hat{\boldsymbol{\tau}}_1,\hat{\boldsymbol{\tau}}_2)$, the negative profile log-EL is minimized using an optimization algorithm such as `optim` or `nlm` function in `R`.

Throughout the simulation study, we used the value of $\boldsymbol{\beta}=(0,2)^T$. One thousand simulations of $n=1000$ each were carried out for each combination of the parameters. For each simulation run, we used the following model to generate the complete and the incomplete data:

$$w(\boldsymbol{\eta})=\frac{\exp\left\{\boldsymbol{\eta}^T(1,X,S)\right\}}{1+\exp\left\{\boldsymbol{\eta}^T(1,X,S)\right\}},\tag{29}$$

where $\boldsymbol{\eta}\equiv(\eta_1,\eta_2,\eta_3)^T$. Different values of $\boldsymbol{\eta}$ are used to simulate situations with different amount of complete data as a proportion of the total data available. Moreover, (29) allows us to model non-randomness in the data. For model 1, the values of $\boldsymbol{\eta}$ used are as follows: $(-1,0,0)^T$, $(-1,1/3,1/3)^T$, $(-1,2/3,2/3)^T$ and $(-1,1,1)^T$. These four sets of $\boldsymbol{\eta}$ induce a ratio of the observations in the complete data to the total number of observations, that is, $m/n\approx 0.7$, 0.5, 0.35 and 0.3, respectively. For model 2, the values of $\boldsymbol{\eta}$ are as follows: $(-1,0.5,0.5)^T$, $(-1,1,1)^T$, $(-1,1.5,1.5)^T$ and $(-1,2,2)^T$. These choices of $\boldsymbol{\eta}$ give, respectively, $m/n\approx 0.65$, 0.6, 0.55 and 0.45. Different choices of $\boldsymbol{\eta}$ are used for the two models because the distributions of X, S are different between the two models.

The simulation results for model 1 are reported in Table 1. For each method, the first row gives the mean (variance) of the estimate of β_1 based on the 1000 replications; the second row gives similar information for β_2. The estimates for β_1 are denoted by $\hat{\beta}_{\mathrm{C}1}, \hat{\beta}_{\mathrm{IPW}1}, \hat{\beta}_{\mathrm{DR}1}, \hat{\beta}_1$, respectively, for the infeasible complete case estimator, the inverse probability estimator, the doubly robustness estimator and the proposed EL estimator. The estimates for β_2 are similarly defined.

Table 1 shows that IPW, DR and EL all give estimates that are approximately unbiased for the underlying parameters. When compared with the infeasible estimator, all three methods

Table 1. *Mean (variance) of different estimators based on 1000 simulations with sample size* $\boldsymbol{n} = 1000$, $w(X,S,\boldsymbol{\eta}) = \frac{\exp\{\boldsymbol{\eta}^T(1,X,S)\}}{1+\exp\{\boldsymbol{\eta}^T(1,X,S)\}}$, $Y \sim N(\boldsymbol{\beta}^T(1,X),1)$, $S \sim N(Y,1)$, $\boldsymbol{\beta} = (\beta_1,\beta_2)^T = (0,2)^T$

Method	$\boldsymbol{\eta} = (-1,0,0)^T$	$\boldsymbol{\eta} = (-1,1/3,1/3)^T$	$\boldsymbol{\eta} = (-1,2/3,2/3)^T$	$\boldsymbol{\eta} = (-1,1,1)^T$
$\hat{\beta}_{\mathrm{C1}}$	−0.001 (0.00201)	−0.001 (0.00195)	0.001 (0.00194)	−0.001 (0.00190)
$\hat{\beta}_{\mathrm{C2}}$	2.001 (0.00101)	2.000 (0.00040)	1.999 (0.00094)	2.000 (0.00101)
$\hat{\beta}_{\mathrm{IPW1}}$	−0.002 (0.00645)	0.001 (0.01013)	0.016 (0.02141)	0.036 (0.03511)
$\hat{\beta}_{\mathrm{IPW2}}$	2.001 (0.00405)	1.999 (0.00451)	1.989 (0.00872)	1.977 (0.01402)
$\hat{\beta}_{\mathrm{DR1}}$	−0.001 (0.00500)	−0.001 (0.00661)	−0.002 (0.01374)	0.002 (0.14793)
$\hat{\beta}_{\mathrm{DR2}}$	2.000 (0.00258)	1.999 (0.00281)	2.001 (0.00589)	1.998 (0.08253)
$\hat{\beta}_1$	−0.001 (0.00506)	−0.001 (0.00651)	−0.002 (0.00999)	−0.010 (0.01557)
$\hat{\beta}_2$	2.000 (0.00262)	2.000 (0.00270)	2.000 (0.00373)	2.005 (0.00577)

Table 2. *Mean (variance) of different estimators based on 1000 simulations with sample size* $\boldsymbol{n} = 1000$, $w(X,S,\boldsymbol{\eta}) = \frac{\exp\{\boldsymbol{\eta}^T(1,X,S)\}}{1+\exp\{\boldsymbol{\eta}^T(1,X,S)\}}$, $Y \sim (\beta_1 + \beta_2 X)\exp(1)$, $S \sim N(Y,1)$, $\boldsymbol{\beta} = (\beta_1,\beta_2)^T = (0,2)^T$

Method	$\boldsymbol{\eta} = (-1,0.5,0.5)^T$	$\boldsymbol{\eta} = (-1,1,1)^T$	$\boldsymbol{\eta} = (-1,1.5,1.5)^T$	$\boldsymbol{\eta} = (-1,2,2)^T$
$\hat{\beta}_{\mathrm{C1}}$	−0.005 (0.00289)	0.002 (0.00243)	0.002 (0.00244)	−0.003 (0.00239)
$\hat{\beta}_{\mathrm{C2}}$	2.029 (0.08511)	1.996 (0.07714)	1.987 (0.07564)	2.021 (0.07251)
$\hat{\beta}_{\mathrm{IPW1}}$	−0.007 (0.00327)	0.002 (0.00268)	0.002 (0.00264)	−0.004 (0.00271)
$\hat{\beta}_{\mathrm{IPW2}}$	2.039 (0.09329)	1.993 (0.08188)	1.990 (0.07748)	2.026 (0.07711)
$\hat{\beta}_{\mathrm{DR1}}$	−0.007 (0.00419)	0.002 (0.00697)	−0.004 (0.01748)	−0.012 (0.03717)
$\hat{\beta}_{\mathrm{DR2}}$	2.036 (0.09112)	1.993 (0.08937)	1.992 (0.10325)	2.033 (0.12198)
$\hat{\beta}_1$	−0.008 (0.00336)	0.000 (0.00270)	−0.001 (0.00277)	−0.008 (0.00319)
$\hat{\beta}_2$	2.037 (0.09318)	1.994 (0.08252)	1.995 (0.08151)	2.040 (0.08736)

suffered substantial efficiency loss across the different scenarios. Table 1 also indicates that efficiency loss is a function of two factors: the proportion of incomplete data and the dependency of the function $w(\boldsymbol{\eta})$ on X and S so that the efficiency loss is directly proportional to the amount of incomplete data and the dependency of $w(\boldsymbol{\eta})$ on X and S. In fact, on the basis of the results in Table 1, efficiency seems to be more affected by the dependency of $w(\boldsymbol{\eta})$ on X and S. This result can be explained by observing that when $w(\boldsymbol{\eta})$ is highly dependent on X and S, some of the observations may have $w(\boldsymbol{\eta})$ close to 1 or 0, and these extreme values of $w(\boldsymbol{\eta})$ in turns induce high variance in the estimation of β_1 and β_2. When the non-randomness is not strong [$\boldsymbol{\eta} = (-1,0,0)^T$ and $(-1,1/3,1/3)^T$], both DR and EL perform similarly, and they are both better than the IPW estimator, as to be expected. When the non-randomness is moderately strong [$\boldsymbol{\eta} = (-1,2/3,2/3)^T$], DR is still better than IPW, but both are less efficient than EL. Under strong non-randomness [$\boldsymbol{\eta} = (-1,1,1)^T$], DR becomes much worse than the other two estimators, and EL remains the best among the three estimators.

The simulation results for model 2 are reported in Table 2. For this model, the efficiency loss of IPW, DR and EL compared with the infeasible estimator is not as severe as in model 1 (compare the first two rows with the rest of Table 2). The performance of IPW is similar to that of EL for all the scenarios studied. In fact, IPW is slightly better than EL, which could be explained by the fact that the efficiency loss is not that great in this model anyway and therefore, the extra step in estimating $\boldsymbol{\psi}_{opt}$ in the EL adds variance to its estimates. Another reason could be due to the moderate sample size used, so when w is too close to 0 or 1, the convergence rates of all existing estimators are slow. For larger sample size (available upon request), EL and IPW give almost identical results. The performance of DR is almost uniformly worse than IPW and EL.

4. Empirical illustration

Following the work of Mincer (1974), economists have often studied the effects of human capital on productivity via wage regression, where the natural logarithm of a measure of wage is regressed upon education, experience and ability. In a wage regression, the use of experience allows economists to study the influence of education on wage, adjusting for individual differences in human capital acquired on the job. In this section, we apply the methods discussed in this paper to investigate estimation of wage regression using data from the 1980 wave of the NLS. The NLS sampled 5255 young men in 1966 to represent the civilian population of men aged 14 to 24 years in the USA. The individuals selected into the NLS were followed longitudinally and were interviewed almost annually until 1981. The data set consists of detailed information about each individual, in particular, measures of ability. However, the data set suffers a high attrition rate such that by 1980, only 3438 (65.8%) of the men in the 1966 cohort were left in the study. To make matters worse, there is no evidence to suggest that attrition was completely at random.

In our application, we use the hourly wage as a measure of wage, and education is the highest grade completed. We use the following three variables as surrogates for human capital: education, experience and IQ test score. Experience is measured by age minus six minus years of education, and the wage regression equation is

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{ education} + \beta_2 \text{ experience} + \beta_3 \text{ experience}^2 + \beta_4 \text{ IQ} + \epsilon. \tag{30}$$

There are many evidences that show wage differences between race that cannot be accounted for by education, experience and ability; hence, following others, we focus our attention on the sample of 1784 white men in 1980. This sample accounts for about half of the 3438 men in the 1980 NLS because black men were deliberately oversampled. Among the 1784 white men, a sub-sample of 1041 have complete records of IQ score. However, all individuals in that cohort also took another ability test called the Knowledge of the World of Work (KWW), and the KWW test score is available for all 1784 men. Under the notations we defined in Section 1, then $Y = \log(\text{wage})$, $\mathbf{Z} = (\mathbf{Z}^{*T}, \text{ IQ})^T$, $\mathbf{Z}^* = (\text{education, experience, experience}^2)^T$, $S = \text{KWW}$ and $R = 1$ for the 1041 men with IQ information and $R = 0$ for the remaining $1784 - 1041 = 743$ men.

We use the four methods in Section 3 to analyse wage regression model (30) on the basis of the data. For the complete case method, CC, we only carry out the analysis using the 1041 observations with IQ score information. For IPW, the analysis is also based on 1041 observations but weighted by inverse probability of selection, which we assume is given by $w(\boldsymbol{\eta}) = \text{logit}(\eta_0 + \eta_1 \text{ KWW})$. For DR, we use all 1784 observations, and we assume that $\boldsymbol{\psi}_{\text{DR}}$ is the function: $\left(\mathbf{Z}^{*T}, \hat{\text{IQ}}\right)^T$ [log(wage) − $(\beta_0 + \beta_1\text{education} + \beta_2\text{experience} + \beta_3 \text{ experience}^2 + \beta_4 \hat{\text{IQ}}]$, where $\hat{\text{IQ}}$ is obtained from a linear regression of IQ on KWW. For EL, 1784 observations are used, with $\boldsymbol{\psi}(\boldsymbol{\beta}) = [1 - w(\boldsymbol{\eta})] \times \boldsymbol{\psi}_{\text{DR}}(\boldsymbol{\beta})$.

The results of the analysis are given in Table 3. The results show that the estimates of the regression coefficients using CC are quite different from the other methods, which support the hypothesis that the individuals with missing IQ may not be a simple random sample of the 1784 men. The other three methods give similar estimates. We use 100 bootstraps to estimate the variances of the estimates. Across all estimates, the variances using CC, IPW and DR are about the same in magnitude, but the corresponding variances using EL are about 40% lower.

Table 3. *Estimates (variance[1]) of wage regression analysis based on 1784 white men from the 1980 NLS*

	Intercept	Education	Experience	Experience^2	IQ
CC	0.289	0.0551	0.0781	−0.169	0.00393
	(4.39×10^{-2})	(4.82×10^{-5})	(4.02×10^{-4})	(5.04×10^{-3})	(8.29×10^{-7})
IPW	0.420	0.0529	0.0657	−0.126	0.00380
	(4.40×10^{-2})	(4.61×10^{-5})	(4.29×10^{-4})	(5.48×10^{-3})	(8.41×10^{-7})
DR	0.420	0.0528	0.0657	−0.126	0.00379
	(4.40×10^{-2})	(4.62×10^{-5})	(4.29×10^{-4})	(5.48×10^{-3})	(8.44×10^{-7})
EL	0.427	0.0547	0.0622	−0.119	0.00365
	(2.27×10^{-2})	(3.78×10^{-5})	(2.35×10^{-4})	(3.08×10^{-3})	(5.30×10^{-7})

NLS, National Longitudinal Survey of Youth; CC, complete case method; IPW, inverse probability weighting; DR, doubly robust estimator of Rotnitzky, Robins and Scharfstein (1998); EL, empirical likelihood.
[1] Based on 100 bootstrap samples.

5. Concluding remarks

In this paper, we consider the problem of combining data from multiple sources under the setting of sampling from an infinite population. We divide the data into those with complete information from all relevant variables and those with incomplete information on one or more of the variables. We assume that information is missing at random, in the sense of Little & Rubin (2002). Our method has two key features. First, unlike competitors such as the IPW and DR, which handle biased sampling using inverse probability weighting ($1/\hat{w}(\boldsymbol{\eta})$), the proposed method uses the direct weighting (14). Because $\hat{w}(\boldsymbol{\eta})$ is an estimated probability and therefore, ranges between 0 and 1, using $1/\hat{w}(\boldsymbol{\eta})$ may be unstable. Second, the method uses $\boldsymbol{\psi}(\boldsymbol{\beta})$ to capture information in the incomplete data, but it does not require $\boldsymbol{\psi}(\boldsymbol{\beta})$ to be unbiased for the optimal function $\boldsymbol{\psi}_{opt}(\boldsymbol{\beta})$.

A number of works have discussed the use of EL method for combining data from multiple surveys or supplementing survey data with information from known population quantities (Chen & Qin, 1993; Chen & Sitter, 1999; Wang & Rao, 2002; Wu, 2004; Rao & Wu, 2010). Except Wang & Rao (2002), these works focused on the situation of finite population sampling, and all assumed that the inclusion probability in the sample is known. Under finite population sampling, unless the inclusion probability is equal for all the units in the population, for example, simple random sampling (Chen & Qin, 1993), the EL formulation needs to be adjusted. Chen & Sitter (1999) introduced a pseudo-EL to handle more complex surveys from finite populations when there is auxiliary information from the finite population. The other works all follow the pseudo-EL approach. Wang & Rao (2002) considered survey data with some observations missing in the outcome, and they proposed imputing the missing outcome with a regression model on the basis of relevant covariates. They then used EL on the basis of the imputed data for constructing confidence intervals for model parameters. However, imputation removes the independent and identically distributed condition, and the resulting EL ratio statistic is no longer a standard chi-squared, as in our theorem 3. Wu (2004) considered combining two surveys from a single sampling frame. They used a two-sample pseudo-EL, and the EL weights were calibrated by equating summary statistics from the surveys. Rao & Wu (2010) extended the method of Wu (2004) for simultaneously combining data from multiple frames and incorporating known population values. They proposed a multiplicity pseudo-EL that does not require complete frame membership information from the units.

Our paper differs from these works in that we consider an infinite population and we allow more complicated parameters other than the mean or the distribution function of the outcome of a survey. There are pros and cons in our formulation. The infinite population assumption allows standard EL to be applied, but our method is only applicable to the simple random sampling situation, which may seem restrictive because most surveys are conducted in a more complex setting. Our method does allow more complicated parameters to be estimated under more general missingness mechanisms. An obvious follow-up to our paper is to extend the current work in more complex sampling schemes in a finite population setting.

For a simple random sample of n observations from a population, EL produces a non-parametric estimate of the cumulative distribution function of the data. In the absence of any auxiliary information, the EL weights are $1/n$, which are identical to the probability mass from the empirical cumulative distribution function of the data. In the presence of auxiliary information, EL adjusts the weights to reflect the auxiliary information. From this viewpoint, the calculation of EL weights is related to the concept of calibration (e.g. Särndal, 2007), which refers to a general class of methods for reweighting observations in a survey such that certain auxiliary information about the sample or population is taken into consideration. The goal of calibration is manifold and includes improving accuracy of estimates and harmonizing estimates from different surveys from the same population quantities. In its traditional form, calibration is model free, and non-response adjustment is carried out separately from the calibration process. But there has been work in introducing models in the method (e.g. Wu & Sitter, 2001) and building adjustment into the calibration process (e.g. Kott, 2006). Our method can be seen as a step in that direction.

Acknowledgements

We thank two anonymous referees and an associate editor for their detailed comments and suggestions. Denis Leung was partially supported by the Research Center at Singapore Management University. B. Zhang wishes to thank the National Security Agency for the support through grant H98230-11-1-0128.

Appendix. Proofs

We begin by summarizing some key notations used in the main text and introducing some new ones that are used in the proofs.

Let m be the number of observations with $(\mathbf{Z}, \mathbf{S})$ and $n - m$ be the number of observations with $(\mathbf{Z}^*, \mathbf{S})$. Let $\mathbf{H}_i(\boldsymbol{\eta}, \boldsymbol{\beta}) = (1, w_i(\boldsymbol{\eta}), \boldsymbol{\psi}_i^T(\boldsymbol{\beta}))^T, i = 1, \ldots, n$ and let

$$\mathbf{g}_1(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\beta}, \boldsymbol{\eta}) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{r_i \mathbf{H}_i(\boldsymbol{\eta}, \boldsymbol{\beta})}{\boldsymbol{\tau}_1^T \mathbf{H}_i(\boldsymbol{\eta}, \boldsymbol{\beta}) + \boldsymbol{\tau}_2^T \mathbf{U}_i(\boldsymbol{\beta})} - \frac{(1 - r_i)\mathbf{H}_i(\boldsymbol{\eta}, \boldsymbol{\beta})}{1 - \boldsymbol{\tau}_1^T \mathbf{H}_i(\boldsymbol{\eta}, \boldsymbol{\beta})} \right],$$

$$\mathbf{g}_2(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\beta}, \boldsymbol{\eta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{r_i \mathbf{U}_i(\boldsymbol{\beta})}{\boldsymbol{\tau}_1^T \mathbf{H}_i(\boldsymbol{\eta}, \boldsymbol{\beta}) + \boldsymbol{\tau}_2^T \mathbf{U}_i(\boldsymbol{\beta})},$$

$$\mathbf{g}_3(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\beta}, \boldsymbol{\eta}) = -\sum_{i=1}^{n} r_i \frac{\boldsymbol{\tau}_1^T \partial \mathbf{H}_i(\boldsymbol{\eta}, \boldsymbol{\beta})/\partial \boldsymbol{\beta} + \boldsymbol{\tau}_2^T \partial \mathbf{U}_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}}{\boldsymbol{\tau}_1^T \mathbf{H}_i(\boldsymbol{\eta}, \boldsymbol{\beta}) + \boldsymbol{\tau}_2^T \mathbf{U}_i(\boldsymbol{\beta})} + (1 - r_i) \frac{\boldsymbol{\tau}_1^T \partial \mathbf{H}_i(\boldsymbol{\eta}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}}{1 - \boldsymbol{\tau}_1^T \mathbf{H}_i(\boldsymbol{\eta}, \boldsymbol{\beta})}.$$

Write $\mathbf{g}_k \equiv \mathbf{g}_k(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\beta}, \boldsymbol{\eta}), k = 1, 2, 3$ for convenience and let $\mathbf{g} = (\mathbf{g}_1^T, \mathbf{g}_2^T, \mathbf{g}_3^T)^T$.

Let $\boldsymbol{\beta}_0$, $\boldsymbol{\eta}_0$, (0,1,0) and 0 be the true values of $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$, respectively. Define $\boldsymbol{\theta} = (\boldsymbol{\tau}_1^T, \boldsymbol{\tau}_2^T, \boldsymbol{\beta}^T)^T, \boldsymbol{\theta}^+ = (\boldsymbol{\theta}^T, \boldsymbol{\eta}^T)^T, \boldsymbol{\theta}_0 = ((0, 1, 0), 0, \boldsymbol{\beta}_0^T)^T, \boldsymbol{\theta}_0^+ = (\boldsymbol{\theta}_0^T, \boldsymbol{\eta}_0^T)^T$ and write $\mathbf{U}_0 \overset{d}{=} \mathbf{U}_i(\boldsymbol{\beta}_0)$, $\boldsymbol{\psi}_0 \overset{d}{=} \boldsymbol{\psi}_i(\boldsymbol{\beta}_0)$, $w_0 \overset{d}{=} w_i(\boldsymbol{\eta}_0)$, $\mathbf{H}_0 \overset{d}{=} \mathbf{H}_i(\boldsymbol{\eta}_0, \boldsymbol{\beta}_0)$ and $r \overset{d}{=} r_i$ where $\overset{d}{=}$ stands for equivalence in distributions.

The conditions needed to establish theorems 1–4 are the following:

(C1:) The propensity score $w_i(\boldsymbol{\eta})$ is twice continuously differentiable with respect to $\boldsymbol{\eta}$ in a neighbourhood of $\boldsymbol{\eta}_0$ and is uniformly bounded away between 0 and 1; furthermore, $m/n \to \rho \in (0,1)$ as $n \to \infty$.

(C2:) There exists an estimator $\tilde{\boldsymbol{\gamma}}$ that converges in mean square to $\boldsymbol{\gamma}_0$ within the parameter space Γ such that for sufficiently large m and n, $\mathrm{E}\{(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)^T\} \leq \mathbf{A}_0$ for a fixed positive definite matrix $\mathbf{A}_0$.

(C3:) Let $\boldsymbol{\xi}_0 = (\mathbf{U}_0^T, \mathbf{H}_0^T)^T$. It is assumed that $\mathrm{E}\left(\frac{\boldsymbol{\xi}_0\boldsymbol{\xi}_0^T}{w_0}\right)$ and $\mathrm{E}\left(\frac{\boldsymbol{\xi}_0\boldsymbol{\xi}_0^T}{1-w_0}\right)$ are positive definite; and the rank of $\mathrm{E}\left(\frac{\partial \mathrm{U}_0}{\partial \boldsymbol{\beta}}\right)$ is p, which is also the dimension of $\boldsymbol{\beta}$.

(C4:) $\frac{\partial^2 \mathbf{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}$ is continuous in a neighbourhood of $\boldsymbol{\beta}_0$ where $\left\|\frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\|$ is bounded; $\frac{\partial^2 \mathbf{H}(\boldsymbol{\eta},\boldsymbol{\beta})}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^T}$ is continuous in a neighbourhood of $(\boldsymbol{\eta}_0, \boldsymbol{\beta}_0)$, and in this neighbourhood $\left\|\frac{\partial \mathbf{H}(\boldsymbol{\eta},\boldsymbol{\beta})}{\partial \boldsymbol{\eta}}\right\|$ is bounded, $\mathrm{E}(||\mathbf{U}(\boldsymbol{\beta})||)^3 < \infty$ and $\mathrm{E}(||\mathbf{H}(\boldsymbol{\eta},\boldsymbol{\beta})||)^3 < \infty$.

We first prove two lemmas whose results will be applied in theorems 1–4. Let

$$\mathbf{q}_{n0} = n^{-1}\sum_{i=1}^{n}\frac{r_i - w_i(\boldsymbol{\eta}_0)}{w_i(\boldsymbol{\eta}_0)\{1-w_i(\boldsymbol{\eta}_0)\}}\frac{\partial w_i(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}} \quad \text{and} \quad \boldsymbol{\Lambda}_{\boldsymbol{\eta}} = \mathrm{E}\left[\frac{1}{w_0\{1-w_0\}}\frac{\partial w_0}{\partial \boldsymbol{\eta}}\frac{\partial w_0^T}{\partial \boldsymbol{\eta}}\right].$$

Lemma 1. *Under Condition C1,* $\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 = \boldsymbol{\Lambda}_{\boldsymbol{\eta}}^{-1}\mathbf{q}_{n0} + o_p\left(n^{-1/2}\right)$.

Proof. The binomial likelihood $\ell_B(\boldsymbol{\eta})$ corresponds to the first term of (24). Because $\hat{\boldsymbol{\eta}}$ is the maximizer of $\ell_B(\boldsymbol{\eta})$,

$$\frac{\partial \ell_B(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \sum_{i=1}^{n}\frac{r_i - w_i(\boldsymbol{\eta})}{w_i(\boldsymbol{\eta})\{1-w_i(\boldsymbol{\eta})\}}\frac{\partial w_i(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \mathbf{0}. \tag{31}$$

Applying Taylor's expansion of (31) at the true value $\boldsymbol{\eta}_0$,

$$\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 = \mathbf{B}_n^{-1}\mathbf{q}_{n0} + o_p\left(n^{-1}\right), \tag{32}$$

where

$$\mathbf{B}_n = n^{-1}\sum_{i=1}^{n}\left[\frac{r_i - w_i(\boldsymbol{\eta}_0)}{w_i(\boldsymbol{\eta}_0)\{1-w_i(\boldsymbol{\eta}_0)\}}\right]\left[\frac{\partial^2 w_i(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2} - \frac{\{1-2w_i(\boldsymbol{\eta}_0)\}}{w_i(\boldsymbol{\eta}_0)(1-w_i(\boldsymbol{\eta}_0))}\frac{\partial w_i(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}}\frac{\partial w_i^T(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}}\right],$$
$$+ n^{-1}\sum_{i=1}^{n}\left[\frac{1}{1-w_i(\boldsymbol{\eta}_0)}\frac{\partial w_i(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}}\frac{\partial w_i^T(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}}\right].$$

As $\mathbf{B}_n = \boldsymbol{\Lambda}_{\boldsymbol{\eta}} + o_P(1)$ and $\mathbf{q}_{n0} = O_P\left(n^{-1/2}\right)$, the lemma is established from (32). □

Lemma 2. *Under Conditions C1–C4,* $\hat{\boldsymbol{\tau}}_1$*,* $\hat{\boldsymbol{\tau}}_2$ *and* $(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)$ *are all of* $O_p\left(n^{-1/2}\right)$.

Proof. Because

$$\mathrm{E}\{\mathbf{U}_i(\boldsymbol{\eta}_0)\} = 0, \quad \mathrm{E}\{\mathbf{h}_i(\boldsymbol{\eta}_0) - \boldsymbol{\mu}\} = 0, \quad i = 1, \ldots, m,$$
$$\mathrm{E}\{\mathbf{h}_j(\boldsymbol{\eta}_0) - \boldsymbol{\mu}\} = 0, \quad j = m+1, \ldots, n,$$

therefore, $\mathbf{g}_k, k = 1, 2, 3$ are of $O_p\left(n^{-1/2}\right)$. Furthermore, using Condition C2, $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + O_p\left(n^{-1/2}\right)$. The lemma then follows similar derivations as those in Owen (1990) and Qin & Lawless (1994).

Using the results of lemmas 1 and 2, ignoring terms of $o_p\left(n^{-1/2}\right)$, we can replace $\hat{\boldsymbol{\eta}}$ by $\boldsymbol{\eta}_0$ and so on. □

Proof of theorem 1: We first show consistency of $\hat{\boldsymbol{\beta}}$. Using (10) and (11), it follows that from $\mathbf{g} = \mathbf{0}$ and Owen (1990) that in an $O_p\left(n^{-1/3}\right)$ neighbourhood, there is a unique and continuously differentiable implicit function

$$\boldsymbol{\phi}(\boldsymbol{\beta}) = O_p\left(n^{-1/3}\right),$$

as long as $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| \le O_p\left(n^{-1/3}\right)$. From Qin & Lawless (1994), for $\boldsymbol{\beta}$ in the set $\mathcal{B} = \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| = n^{-1/3}\}$, we can show

$$\ell(\boldsymbol{\beta}, \boldsymbol{\phi}(\boldsymbol{\beta})) > \ell(\boldsymbol{\beta}_0, \boldsymbol{\phi}(\boldsymbol{\beta}_0)), \quad \text{a.s.}$$

where $\ell(\boldsymbol{\beta}, \boldsymbol{\phi}(\boldsymbol{\beta}))$ is the profile log-EL likelihood (21). Because $\ell(\boldsymbol{\beta}, \boldsymbol{\phi}(\boldsymbol{\beta}))$ is continuously differentiable, it has a local minimum inside the ball with surface $\mathcal{B}$. Consistency of $\hat{\boldsymbol{\beta}}$ then follows from the fact that $\mathcal{B}$ degenerates to $\boldsymbol{\beta}_0$ as $n \to \infty$.

Differentiate $\mathbf{g}_k, k = 1, 2, 3$ with respect to $\boldsymbol{\theta}$, and evaluate at $\boldsymbol{\theta}_0^+$,

$$\begin{aligned}
\frac{\partial \mathbf{g}_1(\boldsymbol{\theta}_0^+)}{\partial \boldsymbol{\tau}_1} &= -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{r_i}{w_i(\eta_0)^2} + \frac{1-r_i}{(1-w_i(\eta_0))^2}\right]\mathbf{H}_i(\eta_0, \boldsymbol{\beta}_0)^T\mathbf{H}_i(\eta_0, \boldsymbol{\beta}_0),\\
\frac{\partial \mathbf{g}_1(\boldsymbol{\theta}_0^+)}{\partial \boldsymbol{\tau}_2} &= -\frac{1}{n}\sum_{i=1}^{n}\frac{r_i\mathbf{H}_i(\eta_0, \boldsymbol{\beta}_0)^T\mathbf{U}_i(\boldsymbol{\beta}_0)}{w_i(\eta_0)^2},\\
\frac{\partial \mathbf{g}_1(\boldsymbol{\theta}_0^+)}{\partial \boldsymbol{\beta}} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{r_i}{w_i(\eta_0)}\frac{\partial \mathbf{H}_i(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} - \frac{1-r_i}{1-w_i(\eta_0)}\frac{\partial \mathbf{H}_i(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right],\\
\frac{\partial \mathbf{g}_2(\boldsymbol{\theta}_0^+)}{\partial \boldsymbol{\tau}_1} &= -\frac{1}{n}\sum_{i=1}^{n}\frac{r_i\mathbf{U}_i(\boldsymbol{\beta}_0)^T\mathbf{H}_i(\eta_0, \boldsymbol{\beta}_0)}{w_i(\eta_0)^2},\\
\frac{\partial \mathbf{g}_2(\boldsymbol{\theta}_0^+)}{\partial \boldsymbol{\tau}_2} &= -\frac{1}{n}\sum_{i=1}^{n}\frac{r_i\mathbf{U}_i(\boldsymbol{\beta}_0)^T\mathbf{U}_i(\boldsymbol{\beta}_0)}{w_i(\eta_0)^2},\\
\frac{\partial \mathbf{g}_2(\boldsymbol{\theta}_0^+)}{\partial \boldsymbol{\beta}} &= \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{w_i(\eta_0)}\frac{\partial \mathbf{U}_i(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}},\\
\frac{\partial \mathbf{g}_3(\boldsymbol{\theta}_0^+)}{\partial \boldsymbol{\tau}_1} &= -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{r_i}{w_i(\eta_0)}\frac{\partial \mathbf{H}_i(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} - \frac{1-r_i}{1-w_i(\eta_0)}\frac{\partial \mathbf{H}_i(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right],\\
\frac{\partial \mathbf{g}_3(\boldsymbol{\theta}_0^+)}{\partial \boldsymbol{\tau}_2} &= -\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{w_i(\eta_0)}\frac{\partial \mathbf{U}_i(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}},\\
\frac{\partial \mathbf{g}_3(\boldsymbol{\theta}_0^+)}{\partial \boldsymbol{\beta}} &= \mathbf{0}.
\end{aligned} \tag{33}$$

It can easily be shown that

$$\frac{\partial g\left(\boldsymbol{\theta}_0^+\right)}{\partial \boldsymbol{\theta}} \to \mathbf{V} = \begin{pmatrix} \mathrm{E}\left(\dfrac{\mathbf{H}_0^T\mathbf{H}_0}{w_0(1-w_0)}\right) & \mathrm{E}\left(\dfrac{\mathbf{H}_0^T\mathbf{U}_0}{w_0}\right) & \mathbf{0} \\ \mathrm{E}\left(\dfrac{\mathbf{U}_0^T\mathbf{H}_0}{w_0}\right) & \mathrm{E}\left(\dfrac{\mathbf{U}_0^T\mathbf{U}_0}{w_0}\right) & \mathrm{E}\left(\dfrac{\partial \mathbf{U}_0}{\partial \boldsymbol{\beta}}\right) \\ \mathbf{0} & \mathrm{E}\left(\dfrac{\partial \mathbf{U}_0}{\partial \boldsymbol{\beta}}\right) & \mathbf{0} \end{pmatrix}.$$

Furthermore,

$$n^{1/2}\mathbf{g}\left(\boldsymbol{\theta}_0^+\right) = n^{-1/2} \begin{pmatrix} \sum_{i=1}^{n} \left[\dfrac{r_i}{w_i(\boldsymbol{\eta}_0)} - \dfrac{(1-r_i)}{(1-w_i(\boldsymbol{\eta}_0))}\right] \mathbf{H}_i(\boldsymbol{\eta}_0, \boldsymbol{\beta}_0) \\ \sum_{i=1}^{n} \dfrac{r_i}{w_i(\boldsymbol{\eta}_0)} \mathbf{U}_i(\boldsymbol{\beta}_0) \\ \mathbf{0} \end{pmatrix} \xrightarrow{d} MVN(\mathbf{0}, \mathbf{W}), \tag{34}$$

where

$$\mathbf{W} = \begin{pmatrix} \mathrm{E}\left(\dfrac{\mathbf{H}_0^T\mathbf{H}_0}{w_0(1-w_0)}\right) & \mathrm{E}\left(\dfrac{\mathbf{H}_0^T\mathbf{U}_0}{w_0}\right) & \mathbf{0} \\ \mathrm{E}\left(\dfrac{\mathbf{U}_0^T\mathbf{H}_0}{w_0}\right) & \mathrm{E}\left(\dfrac{\mathbf{U}_0^T\mathbf{U}_0}{w_0}\right) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Recalling that $\mathbf{H}_i(\boldsymbol{\eta}, \boldsymbol{\beta}) = (1, w_i(\boldsymbol{\eta}), \boldsymbol{\psi}_i^T(\boldsymbol{\beta}))^T, i = 1, \ldots, n$, where technically, $\boldsymbol{\psi}_i^T$ is a function of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. However, because the top element in (34) has mean zero, it is irrelevant whether we assume that $\boldsymbol{\gamma}$ is known or can be replaced by a root-n consistent estimator. This fact allows us to assume that $\boldsymbol{\gamma}$ is known and drop it from further consideration. On the basis of the foregoing results, we have

$$n^{1/2}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} MVN(\mathbf{0}, \boldsymbol{\Sigma}), \text{ where } \boldsymbol{\Sigma} = \left(\mathbf{V}\mathbf{W}^{-1}\mathbf{V}^T\right)^{-1}.$$

Let

$$\mathbf{A} = \begin{pmatrix} \mathrm{E}\left(\dfrac{\mathbf{H}_0^T\mathbf{H}_0}{w_0(1-w_0)}\right) & \mathrm{E}\left(\dfrac{\mathbf{H}_0^T\mathbf{U}_0}{w_0}\right) \\ \mathrm{E}\left(\dfrac{\mathbf{U}_0^T\mathbf{H}_0}{w_0}\right) & \mathrm{E}\left(\dfrac{\mathbf{U}_0^T\mathbf{U}_0}{w_0}\right) \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{0} \\ \mathrm{E}\left(\dfrac{\partial \mathbf{U}_0}{\partial \boldsymbol{\beta}}\right) \end{pmatrix}.$$

Then $\mathbf{V}$ and $\mathbf{W}$ can be written as

$$\mathbf{V} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Using the theory of inverse of block matrices, it can be shown that

$$(\mathbf{V}^{-1})^T\mathbf{W}\mathbf{V} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1} \end{pmatrix}.$$

Therefore,

$$n^{1/2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\xrightarrow{d} MVN\left(\mathbf{0},\left(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\right),$$

where

$$\begin{aligned}\left(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1}&=\mathrm{E}^{-1}\left(\frac{\partial\mathbf{U}_0}{\partial\boldsymbol{\beta}}\right)\left[\mathrm{E}\left(\frac{\mathbf{U}_0^T\mathbf{U}_0}{w_0}\right)\right.\\&\left.-\mathrm{E}\left(\frac{\mathbf{U}_0^T\mathbf{H}_0}{w_0}\right)\mathrm{E}^{-1}\left(\frac{\mathbf{H}_0^T\mathbf{H}_0}{w_0(1-w_0)}\right)\mathrm{E}\left(\frac{\mathbf{H}_0^T\mathbf{U}_0}{w_0}\right)\right]\mathrm{E}^{-1}\left(\frac{\partial\mathbf{U}_0}{\partial\boldsymbol{\beta}^T}\right)\end{aligned}$$

by using the inverse matrix formula and the fact that the first component of **B** is **0**. □

Proof of theorem 2: Note that

$$\mathrm{E}\left(\frac{\mathbf{H}_0^T\mathbf{H}_0}{w_0(1-w_0)}\right)=\begin{pmatrix}\mathrm{E}\left(\frac{1}{w_0(1-w_0)}\right) & \mathrm{E}\left(\frac{1}{1-w_0}\right) & \mathrm{E}\left(\frac{\boldsymbol{\psi}_0}{w_0(1-w_0)}\right)\\ \mathrm{E}\left(\frac{1}{1-w_0}\right) & \mathrm{E}\left(\frac{w_0}{1-w_0}\right) & \mathrm{E}\left(\frac{\boldsymbol{\psi}_0}{1-w_0}\right)\\ \mathrm{E}\left(\frac{\boldsymbol{\psi}_0}{w_0(1-w_0)}\right) & \mathrm{E}\left(\frac{\boldsymbol{\psi}_0}{1-w_0}\right) & \mathrm{E}\left(\frac{\boldsymbol{\psi}_0^T\boldsymbol{\psi}_0}{w_0(1-w_0)}\right)\end{pmatrix}.$$

By using the positive matrix property,

$$\mathrm{E}^{-1}\left(\frac{\mathbf{H}_0^T\mathbf{H}_0}{w_0(1-w_0)}\right)-\begin{pmatrix}\mathbf{0}\;\mathbf{0} & \mathbf{0}\\ \mathbf{0}\;\mathbf{0} & \mathbf{0}\\ \mathbf{0}\;\mathbf{0} & \mathrm{E}^{-1}\left(\frac{\boldsymbol{\psi}_0^T\boldsymbol{\psi}_0}{w_0(1-w_0)}\right)\end{pmatrix}\geq\mathbf{0}.$$

Therefore,

$$\mathrm{E}\left(\frac{\mathbf{U}_0^T\mathbf{H}_0}{w_0}\right)\mathrm{E}^{-1}\left(\frac{\mathbf{H}_0^T\mathbf{H}_0}{w_0(1-w_0)}\right)\mathrm{E}\left(\frac{\mathbf{H}_0^T\mathbf{U}_0}{w_0}\right)\geq\mathrm{E}\left(\frac{\mathbf{U}_0^T\boldsymbol{\psi}_0}{w_0}\right)\mathrm{E}^{-1}\left(\frac{\boldsymbol{\psi}_0^T\boldsymbol{\psi}_0}{w_0(1-w_0)}\right)\mathrm{E}\left(\frac{\boldsymbol{\psi}_0^T\mathbf{U}_0}{w_0}\right).$$

Next, we will show that

$$\begin{aligned}&-\mathrm{E}\left(\frac{\mathbf{U}_0^T\boldsymbol{\psi}_0(1-w_0)}{w_0}\right)\mathrm{E}^{-1}\left(\frac{\boldsymbol{\psi}_0^T\boldsymbol{\psi}_0(1-w_0)}{w_0}\right)\mathrm{E}\left(\frac{\boldsymbol{\psi}_0^T\mathbf{U}_0(1-w_0)}{w_0}\right)\\&\leq-\mathrm{E}\left(\frac{\mathbf{U}_0^T\boldsymbol{\psi}_0}{w_0}\right)\mathrm{E}^{-1}\left(\frac{\boldsymbol{\psi}_0^T\boldsymbol{\psi}_0}{w_0(1-w_0)}\right)\mathrm{E}\left(\frac{\boldsymbol{\psi}_0^T\mathbf{U}_0}{w_0}\right)\end{aligned}$$

Let

$$\boldsymbol{\Omega}_1=\mathrm{E}^{-1}\left(\boldsymbol{\psi}_0^T\boldsymbol{\psi}_0\frac{1-w_0}{w_0}\right)\frac{r-w_0}{w_0}\boldsymbol{\psi}_0,\quad\boldsymbol{\Omega}_2=\mathrm{E}^{-1}\left(\frac{\mathbf{U}_0^T\boldsymbol{\psi}_0}{w_0}\right)\frac{r-w_0}{w_0(1-w_0)}\boldsymbol{\psi}_0.$$

We can easily show that

$$\begin{aligned}\mathrm{var}(\boldsymbol{\Omega}_1)&=\mathrm{E}^{-1}\left(\boldsymbol{\psi}_0^T\boldsymbol{\psi}_0\frac{1-w_0}{w_0}\right),\\ \mathrm{var}(\boldsymbol{\Omega}_2)&=\mathrm{E}^{-1}\left(\frac{\mathbf{U}_0^T\boldsymbol{\psi}_0}{w_0}\right)\mathrm{E}\left(\frac{\boldsymbol{\psi}_0^T\boldsymbol{\psi}_0}{w_0(1-w_0)}\right)\mathrm{E}^{-1}\left(\frac{\boldsymbol{\psi}_0^T\mathbf{U}_0}{w_0}\right).\end{aligned}$$

Furthermore,

$$\mathrm{cov}(\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2) = \mathrm{E}^{-1}\left(\boldsymbol{\psi}_0^T \boldsymbol{\psi}_0 \frac{1-w_0}{w_0}\right) \mathrm{E}\left(\frac{\boldsymbol{\psi}_0^T \boldsymbol{\psi}_0}{w_0}\right) \mathrm{E}^{-1}\left(\frac{\mathbf{U}_0^T \boldsymbol{\psi}_0}{w_0}\right).$$

Because w_0 and $\boldsymbol{\psi}_0$ depend only on $\mathbf{Z}^*, \mathbf{S}$, by conditioning on $\mathbf{Z}^*, \mathbf{S}$, we can show that $\mathrm{E}(\mathbf{U}_0^T \boldsymbol{\psi}_0/w_0) = \mathrm{E}(\boldsymbol{\psi}_0^T \boldsymbol{\psi}_0/w_0)$, which implies that $\mathrm{var}(\boldsymbol{\Omega}_1) = \mathrm{cov}(\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2)$. Therefore

$$\begin{aligned} &\mathbf{0} \le \mathrm{var}(\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2) \equiv \mathrm{var}(\boldsymbol{\Omega}_1) - 2\mathrm{var}(\boldsymbol{\Omega}_1) + \mathrm{var}(\boldsymbol{\Omega}_2) \\ &\quad \Rightarrow \mathrm{var}(\boldsymbol{\Omega}_1) \le \mathrm{var}(\boldsymbol{\Omega}_2) \\ &\quad \Rightarrow -\mathrm{var}^{-1}(\boldsymbol{\Omega}_1) \le -\mathrm{var}^{-1}(\boldsymbol{\Omega}_2). \end{aligned}$$

However, if we set $\boldsymbol{\psi} = (1 - w_0)\boldsymbol{\psi}_0$, then

$$\mathrm{var}(\boldsymbol{\Omega}_2) = \mathrm{E}^{-1}\left(\frac{\mathbf{U}_0^T \boldsymbol{\psi}_0}{w_0}\right) \mathrm{E}\left(\frac{\boldsymbol{\psi}_0^T \boldsymbol{\psi}_0}{w_0(1-w_0)}\right) \mathrm{E}^{-1}\left(\frac{\boldsymbol{\psi}_0^T \mathbf{U}_0}{w_0}\right) = \mathrm{E}^{-1}\left(\boldsymbol{\psi}_0^T \boldsymbol{\psi}_0 \frac{1-w_0}{w_0}\right) = \mathrm{var}(\boldsymbol{\Omega}_1).$$

□

Proof of theorem 3: Write

$$R(\boldsymbol{\beta}_0) = 2\left\{\ell\left(\hat{\boldsymbol{\tau}}_1, \hat{\boldsymbol{\tau}}_2, \hat{\boldsymbol{\beta}}\right) - \ell\left(\tilde{\boldsymbol{\tau}}_1, \tilde{\boldsymbol{\tau}}_2, \boldsymbol{\beta}_0\right)\right\},$$

and let $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\tau}}_1, \hat{\boldsymbol{\tau}}_2, \hat{\boldsymbol{\beta}}\right)$, $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\tau}}_1, \tilde{\boldsymbol{\tau}}_2, \boldsymbol{\beta}_0)$. Expanding $\ell(\tilde{\boldsymbol{\theta}})$ at $\hat{\boldsymbol{\theta}}$ gives

$$\ell\left(\tilde{\boldsymbol{\theta}}\right) - \ell\left(\hat{\boldsymbol{\theta}}\right) = \frac{\partial \ell\left(\hat{\boldsymbol{\theta}}\right)}{\partial \boldsymbol{\theta}} + \frac{1}{2}\left(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\right)^T \frac{\partial^2 \ell\left(\hat{\boldsymbol{\theta}}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T}\left(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\right) + o_p(1).$$

Note that $\partial \ell\left(\hat{\boldsymbol{\theta}}\right)/\partial \boldsymbol{\theta} = \mathbf{0}$. Therefore,

$$R(\boldsymbol{\beta}_0) = -\left(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\right)^T \frac{\partial^2 \ell\left(\hat{\boldsymbol{\theta}}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T}\left(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\right) + o_p(1).$$

From theorem 1,

$$\frac{1}{n}\frac{\partial^2 \ell\left(\hat{\boldsymbol{\theta}}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T} \to \mathbf{V},$$

in probability and

$$n^{1/2}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = -\mathbf{V}^{-1}\mathbf{g}\left(\boldsymbol{\theta}_0^+\right) + o_p(1) = -\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{pmatrix}^{-1} \mathbf{g}\left(\boldsymbol{\theta}_0^+\right) + o_p(1), \tag{35}$$

where $\mathbf{A}$ and $\mathbf{B}$ are as defined in theorem 1. To find the asymptotic distribution of $\tilde{\boldsymbol{\theta}}$, we note that

$$\begin{pmatrix} \dfrac{\partial \mathbf{g}_1\left(\boldsymbol{\theta}_0^+\right)}{\partial \boldsymbol{\tau}_1} & \dfrac{\partial \mathbf{g}_2\left(\boldsymbol{\theta}_0^+\right)}{\partial \boldsymbol{\tau}_1} \\ \dfrac{\partial \mathbf{g}_1\left(\boldsymbol{\theta}_0^+\right)}{\partial \boldsymbol{\tau}_2} & \dfrac{\partial \mathbf{g}_2\left(\boldsymbol{\theta}_0^+\right)}{\partial \boldsymbol{\tau}_2} \end{pmatrix} \to \begin{pmatrix} \mathrm{E}\left(\dfrac{\mathbf{H}_0^T \mathbf{H}_0}{w_0(1-w_0)}\right) & \mathrm{E}\left(\dfrac{\mathbf{H}_0^T \mathbf{U}_0}{w_0}\right) \\ \mathrm{E}\left(\dfrac{\mathbf{U}_0^T \mathbf{H}_0}{w_0}\right) & \mathrm{E}\left(\dfrac{\mathbf{U}_0^T \mathbf{U}_0}{w_0}\right) \end{pmatrix} \equiv \mathbf{A}.$$

Expanding $\mathbf{g}_1, \mathbf{g}_2$ about $\boldsymbol{\theta}_0$ gives

$$n^{1/2}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \to -\begin{pmatrix} \mathbf{A}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{g}\left(\boldsymbol{\theta}_0^{+}\right) + o_p(1). \tag{36}$$

Combining (34), (35) and (36), and using the fact that

$$\mathbf{V}^{-1} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \\ \left(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{A}^{-1} & \left(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \end{pmatrix},$$

we have

$$n^{1/2}\left(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\right) = \mathbf{V}^{-1}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} \mathbf{g}\left(\boldsymbol{\theta}_0^{+}\right) + o_p(1) \xrightarrow{d} MVN\left(\mathbf{0}, \boldsymbol{\Omega}\right),$$

where

$$\boldsymbol{\Omega} = \mathbf{V}^{-1}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} \mathbf{W} \begin{pmatrix} \mathbf{0} & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \mathbf{V}^{-1}.$$

Using the Ogasawara–Takahashi theorem (Rao, 1973, p 188), to show $R(\boldsymbol{\beta}_0) \xrightarrow{d} \chi^2(p)$, where $\chi^2(p)$ is a chi-square variable with degrees of freedom equal p, which is also the dimension of $\boldsymbol{\beta}$, we only need to demonstrate that

$$\boldsymbol{\Omega}\mathbf{V}\boldsymbol{\Omega}\mathbf{V}\boldsymbol{\Omega} = \boldsymbol{\Omega}\mathbf{V}\boldsymbol{\Omega} \quad \text{and} \quad trace(\mathbf{V}\boldsymbol{\Omega}) = p.$$

We can write

$$\begin{aligned} \boldsymbol{\Omega} &= \mathbf{V}^{-1}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} \mathbf{W} \begin{pmatrix} \mathbf{0} & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \mathbf{V}^{-1} \\ &= \mathbf{V}^{-1}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \mathbf{V}^{-1} \\ &= \mathbf{V}^{-1}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{B}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \mathbf{V}^{-1} \\ &= \mathbf{V}^{-1}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B} \end{pmatrix} \mathbf{V}^{-1} \\ &= \mathbf{V}^{-1}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix}, \end{aligned} \tag{37}$$

and

$$\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix}. \tag{38}$$

Using (37) and (38),

$$\begin{aligned} \boldsymbol{\Omega}\mathbf{V}\boldsymbol{\Omega}\mathbf{V}\boldsymbol{\Omega} &= \mathbf{V}^{-1}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} \mathbf{V}\mathbf{V}^{-1} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} \mathbf{V}\boldsymbol{\Omega} \\ &= \mathbf{V}^{-1}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} \mathbf{V}\boldsymbol{\Omega} = \boldsymbol{\Omega}\mathbf{V}\boldsymbol{\Omega}. \end{aligned}$$

Also, because

$$\mathbf{V\Omega} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix},$$

and trace$(\mathbf{V\Omega}) = p$. □

Proof of theorem 4: By differentiating the binomial log-likelihood with respect to $\boldsymbol{\eta}$, we obtain

$$\mathbf{g}_0 = \frac{1}{n}\sum_{i=1}^{n} r_i \frac{\partial \log w_i(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} + (1-r_i)\frac{\partial \log(1-w_i(\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}} = \frac{1}{n}\sum_{i=1}^{n} \frac{r_i - w_i(\boldsymbol{\eta})}{w_i(\boldsymbol{\eta})(1-w_i(\boldsymbol{\eta}))}\frac{\partial w_i(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}.$$

Let $\hat{\boldsymbol{\theta}}^+ = \left(\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\eta}}^T\right)^T$ solve $\mathbf{g}^+ \equiv \left(\mathbf{g}_0^T, \mathbf{g}_1^T, \mathbf{g}_2^T, \mathbf{g}_3^T\right)^T = \mathbf{0}$. The full log-EL from (24) can be written as

$$\ell_F(\boldsymbol{\beta}, \boldsymbol{\eta}) = \ell_B(\boldsymbol{\eta}) + \ell(\boldsymbol{\beta}, \boldsymbol{\eta}).$$

Following a similar line of argument as in the proof of theorem 1, we can find an implicit function $\boldsymbol{\phi}^+(\boldsymbol{\beta}, \boldsymbol{\eta})$, such that

$$\ell\left(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\phi}^+(\boldsymbol{\beta}, \boldsymbol{\eta})\right) > \ell\left(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, \boldsymbol{\phi}^+(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\right), \quad \text{a.s.}$$

Furthermore,

$$\ell_B(\boldsymbol{\eta}) \geq \ell_B(\boldsymbol{\eta}_0).$$

Hence,

$$\ell_F(\boldsymbol{\beta}, \boldsymbol{\eta}) > \ell_F(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0), \quad \text{a.s.}$$

and consistency of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})$ follows.

Next, we need to evaluate $\partial \mathbf{g}_k/\partial \boldsymbol{\theta}, k = 0, 1, 2, 3$ at $\boldsymbol{\theta}_0^+$. Differentiate $\mathbf{g}_k, k = 0, 1, 2, 3$ with respect to $\boldsymbol{\eta}$ and evaluate at $\boldsymbol{\theta}_0^+$,

$$\begin{aligned}
\frac{\partial \mathbf{g}_0\left(\boldsymbol{\theta}_0^+\right)}{\partial \boldsymbol{\eta}} &= -\frac{1}{n}\sum_{i=1}^{n} \frac{r_i w_i(\boldsymbol{\eta}_0)[1-w_i(\boldsymbol{\eta}_0)] - [r_i - w_i(\boldsymbol{\eta}_0)][1-2w_i(\boldsymbol{\eta}_0)]}{[w_i(\boldsymbol{\eta}_0)(1-w_i(\boldsymbol{\eta}_0)]^2}\frac{\partial^2 w_i(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^T}, \\
\frac{\partial \mathbf{g}_1\left(\boldsymbol{\theta}_0^+\right)}{\partial \boldsymbol{\eta}} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{r_i}{w_i(\boldsymbol{\eta}_0)} - \frac{1-r_i}{1-w_i(\boldsymbol{\eta}_0)}\right]\frac{\partial \mathbf{H}_i(\boldsymbol{\eta}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\eta}} \\
&\quad - \left[\frac{r_i}{w_i(\boldsymbol{\eta}_0)^2} + \frac{1-r_i}{[1-w_i(\boldsymbol{\eta}_0)]^2}\right]\mathbf{H}_i(\boldsymbol{\eta}_0, \boldsymbol{\beta}_0)^T\frac{\partial w_i(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}}, \\
\frac{\partial \mathbf{g}_2\left(\boldsymbol{\theta}_0^+\right)}{\partial \boldsymbol{\eta}} &= -\frac{1}{n}\sum_{i=1}^{n}\frac{r_i \mathbf{U}_i(\boldsymbol{\beta}_0)}{w_i(\boldsymbol{\eta}_0)^2}\frac{\partial w_i(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}}, \\
\frac{\partial \mathbf{g}_3\left(\boldsymbol{\theta}_0^+\right)}{\partial \boldsymbol{\eta}} &= \mathbf{0}.
\end{aligned} \tag{39}$$

Furthermore,

$$\frac{\partial \mathbf{g}_0\left(\boldsymbol{\theta}_0^+\right)}{\partial \boldsymbol{\tau}_1} = \mathbf{0}, \quad \frac{\partial \mathbf{g}_0\left(\boldsymbol{\theta}_0^+\right)}{\partial \boldsymbol{\tau}_2} = \mathbf{0}, \quad \frac{\partial \mathbf{g}_0\left(\boldsymbol{\theta}_0^+\right)}{\partial \boldsymbol{\beta}} = \mathbf{0}. \tag{40}$$

Combining (33), (39) and (40), it can be shown that in probability

$$\frac{\partial \mathbf{g}^{+}\left(\boldsymbol{\theta}_0^{+}\right)}{\partial \boldsymbol{\theta}^{+}} \rightarrow \mathbf{V}^{+} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{0} & \mathbf{0} \\ \mathbf{V}_{21}^{T} & \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{B}^{T} & \mathbf{0} \end{pmatrix}, \tag{41}$$

where

$$\mathbf{V}_{11} = -\mathrm{E}\left(\frac{w_0 \frac{\partial w_0}{\partial \boldsymbol{\eta}}}{1 - w_0}\right), \quad \mathbf{V}_{21}^{T} = -\left(\mathrm{E}\left(\frac{\left(\frac{\partial w_0}{\partial \boldsymbol{\eta}}\right)^{T} \mathbf{H}_0}{w_0(1 - w_0)}\right), \mathrm{E}\left(\frac{\left(\frac{\partial w_0}{\partial \boldsymbol{\eta}}\right)^{T} \mathbf{U}_0}{w_0}\right)\right).$$

Note that the expression for $\mathbf{V}_{11}$ results from assuming a logistic propensity function. The results of this theorem still hold for other forms of propensity function, but the expression would be different. Also,

$$n^{1/2}\mathbf{g}^{+}\left(\boldsymbol{\theta}_0^{+}\right) = n^{-1}\begin{pmatrix} \sum_{i=1}^{n}\left[\frac{r_i}{w_i(\boldsymbol{\eta}_0)} - \frac{1 - r_i}{1 - w_i(\boldsymbol{\eta}_0)}\right]\frac{\partial w_i(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}} \\ \sum_{i=1}^{n}\left[\frac{r_i}{w_i(\boldsymbol{\eta}_0)} - \frac{(1 - r_i)}{1 - w_i(\boldsymbol{\eta}_0)}\right]\mathbf{H}_i(\boldsymbol{\eta}_0, \boldsymbol{\beta}_0) \\ \sum_{i=1}^{n}\frac{r_i}{w_i(\boldsymbol{\eta}_0)}\mathbf{U}_i(\boldsymbol{\beta}_0) \\ \mathbf{0} \end{pmatrix} \xrightarrow{d} MVN(\mathbf{0}, \mathbf{W}^{+}),$$

where

$$\mathbf{W}^{+} = \begin{pmatrix} \mathbf{W}_{11} & -\mathbf{V}_{21}^{T} & \mathbf{0} \\ -\mathbf{V}_{21} & \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{42}$$

with

$$\mathbf{W}_{11} = \mathrm{E}\left(\frac{\left(\frac{\partial w_0}{\partial \boldsymbol{\eta}}\right)^{T}\left(\frac{\partial w_0}{\partial \boldsymbol{\eta}}\right)}{w_0(1 - w_0)}\right).$$

Therefore,

$$n^{1/2}\left(\hat{\boldsymbol{\theta}}^{+} - \boldsymbol{\theta}_0^{+}\right) \xrightarrow{d} MVN(\mathbf{0}, \boldsymbol{\Sigma}^{+}), \text{ where } \boldsymbol{\Sigma}^{+} = \left(\mathbf{V}^{+}\left(\mathbf{W}^{+}\right)^{-1}\left(\mathbf{V}^{+}\right)^{T}\right)^{-1}.$$

We now derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}$. Write

$$\mathbf{D} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{0} \\ \mathbf{V}_{21} & \mathbf{A} \end{pmatrix}, \quad \tilde{\mathbf{B}} = \begin{pmatrix} \mathbf{0} \\ \mathbf{B} \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} \mathbf{W}_{11} & -\mathbf{V}_{21}^{T} \\ -\mathbf{V}_{21} & \mathbf{A} \end{pmatrix}. \tag{43}$$

Using (43), (41) and (42) can be rewritten as

$$\mathbf{V}^{+} = \begin{pmatrix} \mathbf{D} & \tilde{\mathbf{B}} \\ \tilde{\mathbf{B}}^{T} & \mathbf{0} \end{pmatrix}, \quad \mathbf{W}^{+} = \begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Using simple matrix algebra and the theory of generalized inverses, it follows that

$$\left(\mathbf{V}^{+}\left(\mathbf{W}^{+}\right)^{-1}\left(\mathbf{V}^{+}\right)^{T}\right)^{-1} = \begin{pmatrix} \mathbf{D}\mathbf{C}^{-1}\mathbf{D}^{T} & \mathbf{D}\mathbf{C}^{-1}\tilde{\mathbf{B}} \\ \tilde{\mathbf{B}}^{T}\mathbf{C}^{-1}\mathbf{D}^{T} & \tilde{\mathbf{B}}^{T}\mathbf{C}^{-1}\tilde{\mathbf{B}} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{B}}^{T}\mathbf{C}^{-1}\tilde{\mathbf{B}})^{-1} \end{pmatrix}. \tag{44}$$

From (44), therefore,

$$n^{1/2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} MVN\left(\mathbf{0}, \left(\tilde{\mathbf{B}}^{T}\mathbf{C}^{-1}\tilde{\mathbf{B}}\right)^{-1}\right).$$

□

References

Abowd, J., Crepon, B. & Kramarz, F. (2001). Moment estimation with attrition. *J. Amer. Statist. Assoc.* **96**, 1223–1231.

Arellano, M. & Meghir, C. (1992). Female labour supply and on-the-job search: an empirical model estimated using complementary data sets. *Rev. Econ. Stud.* **59**, 537–559.

Chen, J. & Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika* **80**, 107–116.

Chen, J. & Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statist. Sinica* **9**, 385–406.

Chen, J., Sitter, R. R. & Wu, C. (2002). Using empirical likelihood method to obtain range restricted weights in regression estimators for surveys. *Biometrika* **89**, 230–237.

Chen, S. X. & Cui, H. (2006). On Bartlett correction of empirical likelihood in the presence of nuisance parameters. *Biometrika* **93**, 215–220.

Chen, S. X., Leung, D. H.-Y. & Qin, J. (2003). Information recovery in a study with surrogate endpoints. *J. Amer. Statist. Assoc.* **98**, 1052–1062.

Chen, S. X., Leung, D. H.-Y. & Qin, J. (2008). Improving semiparametric estimation using surrogate data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70**, 803–823.

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

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

Cook, J. R. & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.* **89**, 1314–1328.

DiCiccio, T. J., Hall, P. & Romano, J. P. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.* **19**, 1053–1061.

Gong, G. & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: theory and applications. *Ann. Statist.* **9**, 861–869.

Graham, B. S., Pinto, C. & Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *Rev. Econ. Stud.* **79**, 1052–1079.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–1054.

Hellerstein, J. K. & Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *Rev. Econ. Statist.* **81**, 1–14.

Hirano, K., Imbens, G. W. & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–1190.

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

Imbens, G. W. (1992). An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica* **60**, 1187–1214.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87**, 706–710.

Imbens, G. W. & Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *Rev. Econ. Stud.* **61**, 655–680.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Surv. Methodol.* **32**, 133–142.

Liang, H., Wang, S. J. & Carroll, R. (2007). Partially linear models with missing response variables and error-prone covariates. *Biometrika* **94**, 185–198.

Lipsitz, S. R., Ibrahim, J. G. & Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *J. Amer. Statist. Assoc.* **94**, 1147–60.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data*, Wiley, New York.

Lusardi, A. (1996). Permanent income, current income and consumption: evidence from two panel data sets. *J. Bus. Econom. Statist.* **14**, 81–90.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *J. Amer. Statist. Assoc.* **99**, 1131–1139.

Mincer, J. (1974). *Schooling, experience, and earnings*, National Bureau of Economic Research, New York.

Nevo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *J. Bus. Econom. Statist.* **21**, 43–52.

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

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

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

Owen, A. (2001). *Empirical likelihood*, Chapman and Hall/CRC, Boca Raton.

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

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

Rao, C. R. (1973). *Linear Statistical inference and its applications*, Wiley, New York.

Rao, J. N. K. & Wu, C. (2010). Pseudo-empirical likelihood inference for multiple frame surveys. *J. Amer. Statist. Assoc.* **105**, 1494–1503.

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

Rotnitzky, A., Robins, J. M. & Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Amer. Statist. Assoc.* **93**, 1321–1339.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Surv. Methodol.* **33**, 99–119.

Skinner, C., Stuttard, N., Beissel-Durrant, G. & Jenkins, J. (2002). The measurement of low pay in the UK Labour Force Survey. *Oxford B. Econ. Statist.* **64**, 653–676.

Tarozzi, A. (2007). Calculating comparable statistics from incomparable surveys, with an application to poverty in India. *J. Bus. Econom. Statist.* **25**, 314–336.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10**, 616–620.

Vardi, Y. (1985). Empirical distribution in selection bias models. *Ann. Statist.* **13**, 178–203.

Wang, C., Wang, S. & Carroll, R. (1997). Estimation in choice-based sampling with measurement error and bootstrap analysis. *J. Econometrics* **77**, 65–86.

Wang, Q. & Rao, J. N. K. (2002). Empirical likelihood-based inference in linear errors-in-covariables models with validation data. *Biometrika* **89**, 345–358.

Wooldridge, J. (2007). Inverse probability weighted estimation for general missing data problems. *J. Econometrics* **141**, 1281–1301.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *Canad. J. Statist.* **32**, 15–26.

Wu, C. & Rao, J. N. K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *Canad. J. Statist.* **34**, 359–375.

Wu, C. & Sitter, R. R. (2001). A model calibration approach to using complete auxiliary information from survey data. *J. Amer. Statist. Assoc.* **96**, 185–193.

Denis Heng Yan Leung, School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903.
E-mail: denisleung@smu.edu.sg