J. R. Statist. Soc. B (2007)
69, *Part* 1, *pp.* 101–122

Empirical-likelihood-based inference in missing response problems and its application in observational studies

Jing Qin

National Institute of Allergy and Infectious Diseases, Bethesda, USA

and Biao Zhang

University of Toledo, USA

[Received January 2006. Revised September 2006]

Summary. The problem of missing response data is ubiquitous in medical and social science studies. In the case of responses that are missing at random (depending on some covariate information), analyses focused only on the complete data may lead to biased results. Various debias methods have been extensively studied in the literature, particularly the weighting method that was motivated by Horvitz and Thompson's estimators. To improve efficiency, Robins, Rotnitzky and Zhao proposed augmented estimating equations based on corrected complete-case analyses. A nice feature of the augmented method is its 'double robustness', i.e. the estimator that is derived from the augmented method is asymptotically unbiased if either the underlying missing data mechanism or the underlying regression function is correctly specified. Furthermore, the augmented estimator can achieve full efficiency if both the missing data mechanism and the regression function are correctly specified. In general, however, it is very difficult to specify the regression function correctly, especially when the dimension of covariates is high—this is the so-called curse of dimensionality problem. The augmented estimator has much lower efficiency if the 'working regression model' is not close to the true regression model. In this paper, the empirical likelihood method is employed to seek a constrained empirical likelihood estimation of mean response with the assumption that responses are missing at random. The empirical-likelihood-based estimators enjoy the double-robustness property. Moreover, it is possible that the empirical-likelihood-based inference can produce asymptotically unbiased and efficient estimators even if the true regression function is not completely known. Simulation results indicate that the empirical-likelihood-based estimators are very robust to a misspecification of the propensity score and dominate other competitors in the sense of having smaller mean-square errors. Methods that are developed in this paper have a nice application in observational causal inferences. The propensity score is used to adjust for differences in pretreatment variables in the estimation of average treatment effects.

Keywords: Auxiliary information; Average treatment effect; Biased sampling; Causal inference; Empirical likelihood; Missing data; Observational studies; Propensity score; Survey sampling

1. Introduction

Missing data are very common practical problems in medical and social science studies. There are many different patterns for missingness, such as partially missing responses, or partially missing covariates or both. In this paper, we mainly discuss the problem of possibly missing

Address for correspondence: Biao Zhang, Department of Mathematics, University of Toledo, Toledo, OH 43606, USA.
E-mail: bzhang@utnet.utoledo.edu

 1369–7412/07/69101

responses. We assume that responses are missing at random, i.e., conditionally on the covariates and responses, the missing responses depend only on the covariates. For a more detailed discussion on data missing completely at random, missing at random and non-ignorable missingness, we refer readers to Rubin (1976) and Little and Rubin (2002). A very common way to deal with missing data is to use only those data with complete observations. However, this method may result in a loss of efficiency, or, more seriously, it may produce biased results if missingness is not completely at random. Various debias methods have been studied in the literature, particularly the weighting method that was motivated by Horvitz and Thompson's (1952) estimators. To improve efficiency, Robins *et al.* (1994, 1995) proposed augmented estimating equations based on corrected complete-case analyses. A nice feature of the augmented method is that it has the characteristic of being 'doubly robust', i.e. the estimator that is based on the augmented method is asymptotically unbiased if either the underlying missing data mechanism or the underlying regression function is correctly specified. Furthermore, the augmented estimator can achieve full efficiency if both the missing data mechanism and the regression function are correctly specified. In general, however, it is very difficult to specify the regression function correctly; this is especially true when the dimension of covariates is high. This has been called the curse of dimensionality problem. The augmented estimator is much less efficient if the 'working regression model' and the true regression model are not close to each other.

The regression method is the simplest method for producing an efficient estimator for the mean response when responses are missing completely at random. It originated from survey sampling in finite populations (Cochran, 1977). As an alternative method for estimating the mean response, Chen and Qin (1993) adapted the empirical likelihood method to finite population sampling. The empirical likelihood was introduced by Owen (1988, 1990, 1991, 2001) for constructing confidence intervals for the mean and other parameters. Chen and Qin (1993) showed that empirical likelihood can effectively use auxiliary information, such as the known mean value of a population. Chen and Sitter (1999) generalized the empirical likelihood method to complex sampling situations when auxiliary information, such as the stratum mean or stratum size, is available. When the mean value of some function of a random variable is known, Haberman (1984) obtained an efficient estimator for the underlying distribution function by minimizing the Kullback–Leibler divergence. Qin and Lawless (1994) demonstrated that the empirical likelihood method can be used to solve estimating equations when the number of estimating equations exceeds the number of parameters. Kitamura (1997) developed blockwise empirical likelihood methods for estimating equations and for smooth functions of means. Tripathi and Kitamura (2003) extended the empirical likelihood method to test conditional moment restrictions.

When responses are missing at random, the likelihood that is based only on complete data is biased. The bias function or weighting function depends on the missing data mechanism. Biased sampling problems have been discussed extensively in the last two decades. The length-biased sampling problem (the weighting function is proportional to the length of the measurement) was discussed by Vardi (1982). For general weighting functions, identifiability of the underlying parameters and distribution functions and the corresponding estimation problems were discussed by Vardi (1985), Gill *et al.* (1988), Qin (1993), Heckman, *et al.* (1998), Gilbert *et al.* (1999) and Gilbert (2000). Despite the missingness of the response variable, the covariate information is available for everyone. Wang and Rao (2002) proposed an empirical likelihood-based inference procedure for the mean response under kernel regression imputation. Their procedure requires a nonparametric kernel estimate of the regression function. If the dimension of the covariate vector is high, completely nonparametric estimation of the regression function is not generally practical owing to the curse of dimensionality. In this paper we propose a different

empirical likelihood method for estimating the mean response by maximizing the biased sampling likelihood subject to covariate moment constraints. Our main goal is to study, under three different set-ups for the regression function, the empirical-likelihood-based estimator for the mean response when responses are missing at random.

A natural application of data missing at random is the estimation of the treatment effect in an observational study. Recently, there has been a surge in econometric and epidemiologic works focusing on the estimation of average treatment effects under various sets of assumptions; see for example Rubin (1990), Korn and Baumrind (1998), Hahn (1998), Rosenbaum (2002), Breslow (2003), Hirano *et al.* (2003), van der Laan and Robins (2003), Imbens (2004) and Tan (2006). Estimation of the average effect of a binary treatment or policy on a scalar outcome is a basic goal of many empirical studies in economics and epidemiology. If assignment to the treatment is independent of potential outcomes and covariates, the average treatment effect can be estimated simply by making use of the mean difference. However, if the treatment assignment depends on covariates, such strategies may not be desirable and can even produce biased estimators. The estimation of average treatment effects in observational studies often requires adjustment for differences in pretreatment variables. In these studies, investigators have no control over the treatment assignment. As a result, large differences in observed covariates in treatment and control groups may exist, which may lead to biased estimators of the average treatment effects. Rosenbaum and Rubin (1983) proposed the propensity score method to estimate the average treatment effect by adjusting pretreatment variables. The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. A popular method for estimating the average difference between the treatment and control groups is to match the covariate analysis based on the propensity score. An alternative approach is to adjust for confounding factors by using estimated propensity scores to construct weights for each individual. If we treat the treatment responses as missing data for those individuals who are assigned to controls, then estimation of the average treatment effect can be accomplished by using the techniques that have been developed in one-sample missing-at-random data. Similarly, the average control effect can be estimated by treating the control responses as missing data for those individuals who are assigned to treatments.

This paper is organized as follows. In Section 2, we propose our constrained empirical likelihood estimator of mean response based on one-sample missing-at-random response data. The empirical-likelihood-based estimator proposed enjoys the double-robustness property, i.e. the estimator of the mean response is asymptotically unbiased if either the underlying propensity score or the underlying regression function is correctly specified. Moreover, it is possible that the empirical-likelihood-based inference can produce asymptotically unbiased and efficient estimators even when the true regression function is not completely known. In Section 3, the causal inference is treated as a two-sample missing data problem, where the treatment responses are missing for patients who chose controls and the control responses are missing for patients who chose treatments. An extensive comparison simulation study with competing methods is conducted in Section 4. A concluding remark is given in Section 5. Finally, proofs of the main theoretical results are provided in Appendix A.

2. Main results in a one-sample missing response problem

In this section we present our main results in a one-sample missing data problem. Denote the response variable, covariate vector and missing indicator variable as Y, X and D, where Y is missing if $D=0$, and X is always observable. Assume that

$$P(D=1|X=x)=w(x,\beta),$$

where $w(\cdot,\cdot)$ is a specified probability distribution function for given β, a $p\times 1$ unknown vector parameter. The most popular choice of $w(x,\beta)$ is the logistic regression function

$$\frac{\exp(\beta^{\mathrm{T}}x)}{1+\exp(\beta^{\mathrm{T}}x)}.$$

We are interested in estimating

$$\mu = E(Y) = \int y\, f(y,x)\,\mathrm{d}x\,\mathrm{d}y,$$

where $f(y,x)$ is the joint density function of (Y,X).

We denote the observed data by (y_i,x_i,d_i), $i=1,\ldots,n$. Then the likelihood is

$$L=\prod_{i=1}^{n}\{w(x_i,\beta)\ f(y_i,x_i)\}^{d_i}[\{1-w(x_i,\beta)\}\ g(x_i)]^{1-d_i},$$

where $g(x)$ is the marginal density function of X. Since the covariates x_i are observable for each subject, it is natural to estimate β by the maximum binomial likelihood estimator $\hat{\beta}$ which maximizes the binomial likelihood

$$L_{\mathrm{B}}(\beta)=\prod_{i=1}^{n} w(x_i,\beta)^{d_i}\{1-w(x_i,\beta)\}^{1-d_i}. \tag{2.1}$$

For notational simplicity, let $n_1=\Sigma_{i=1}^{n}d_i$ and let $(y_1,x_1),\ldots,(y_{n_1},x_{n_1})$ denote observations on (Y,X) corresponding to $d_1=\ldots=d_{n_1}=1$. Now, conditionally on $d_1=\ldots=d_{n_1}=1$, the conditional likelihood based on $(y_1,x_1),\ldots,(y_{n_1},x_{n_1})$ is given by

$$\begin{aligned} L_{\mathrm{C}} &= \prod_{i=1}^{n_1}\frac{w(x_i,\beta)\ f(y_i,x_i)}{\displaystyle\int\int w(x,\beta)\ f(y,x)\,\mathrm{d}x\,\mathrm{d}y} \\ &= \prod_{i=1}^{n_1}\frac{w(x_i,\beta)p_i}{\theta}, \end{aligned} \tag{2.2}$$

where

$$\theta=\int\int w(x,\beta)\ f(y,x)\,\mathrm{d}x\,\mathrm{d}y = E\{w(X,\beta)\}$$

and $p_i=f(y_i,x_i)$, $i=1,\ldots,n_1$, are non-negative jump sizes with total mass 1. Statistical inference on the basis of the conditional likelihood L_{C} may be viewed as a biased sampling problem of Vardi (1982, 1985). As it is well known that constrained empirical likelihood estimation is more efficient than unconstrained empirical likelihood estimation, we can impose some constraints to enhance the efficiency of statistical inference based on the conditional likelihood (2.2) by making use of the availability of all covariates $x_1,\ldots,x_n$.

For any q ($q\geqslant 1$) functionally independent known functions $a(x)=(a_1(x),\ldots,a_q(x))^{\mathrm{T}}$, their expectations $E\{a(X)\}=a=(a_1,\ldots,a_q)^{\mathrm{T}}$, if unknown, can be estimated by using the sample version from the complete data $x_1,\ldots,x_n$, denoted as $\hat{a}=(\bar{a}_1,\ldots,\bar{a}_q)^{\mathrm{T}}$, where

$$\bar{a}_k = n^{-1}\sum_{i=1}^{n} a_k(x_i)$$

for $k=1,\ldots,q$. In large survey sampling studies, we may know the expectation of $a=E\{a(X)\}$ approximately; see for example Imbens and Lancaster (1994). In that case, we can replace $\hat{a}$ by its survey value. In a similar manner, we can estimate $\theta=E\{w(X,\beta)\}$ by

$$\hat{\theta} = n^{-1} \sum_{i=1}^{n} w(x_i, \hat{\beta}).$$

Now, with β replaced by $\hat{\beta}$, θ replaced by $\hat{\theta}$ and a replaced by $\hat{a}$, we can maximize the conditional likelihood L_C in equation (2.2) subject to the constraints

$$\sum_{i=1}^{n_1} p_i = 1, \qquad \sum_{i=1}^{n_1} p_i \{w(x_i, \hat{\beta}) - \hat{\theta}\} = 0, \qquad \sum_{i=1}^{n_1} p_i \{a(x_i) - \hat{a}\} = 0. \tag{2.3}$$

The first two constraints are necessary as they reflect the fact that $f(y, x)$ is a density function ($\int\int f(y, x)\, \mathrm{d}x\, \mathrm{d}y = 1$) and the selection bias. The third constraint is to utilize auxiliary information on $a(X)$ and is thus optional. For fixed $(\hat{\beta}, \hat{\theta}, \hat{a})$, a Lagrange multiplier argument such as in Owen (1990) or Qin and Lawless (1994) shows that the maximum value of L_C is, subject to constraints (2.3), attained at

$$p_i = \frac{1}{n_1} \frac{1}{1 + \lambda_1 \{w(x_i, \hat{\beta}) - \hat{\theta}\} + \lambda_2^{\mathrm{T}} \{a(x_i) - \hat{a}\}}, \qquad i = 1, \ldots, n_1,$$

where λ_1 and λ_2 are Lagrange multipliers determined by the equations

$$\begin{aligned} \sum_{i=1}^{n_1} \frac{w(x_i, \hat{\beta}) - \hat{\theta}}{1 + \lambda_1 \{w(x_i, \hat{\beta}) - \hat{\theta}\} + \lambda_2^{\mathrm{T}} \{a(x_i) - \hat{a}\}} &= 0, \\ \sum_{i=1}^{n_1} \frac{a(x_i) - \hat{a}}{1 + \lambda_1 \{w(x_i, \hat{\beta}) - \hat{\theta}\} + \lambda_2^{\mathrm{T}} \{a(x_i) - \hat{a}\}} &= 0. \end{aligned} \tag{2.4}$$

Since the true value of the Lagrange multiplier λ_1 is non-zero in the current biased sampling problem, we reparameterize from (λ_1, λ_2) to (π_1, π_2) with $\pi_1 = \hat{\theta}\lambda_1 - 1$ and $\pi_2 = \hat{\theta}\lambda_2$. In doing so, the mass function p_i becomes

$$\begin{aligned} p_i &= p_i(\pi) \\ &= \frac{1}{n_1} \frac{\hat{\theta}\, w^{-1}(x_i, \hat{\beta})}{1 + \pi^{\mathrm{T}} r(x_i, \hat{\beta}, \hat{\theta}, \hat{a})}, \qquad i = 1, \ldots, n_1, \end{aligned} \tag{2.5}$$

where $\pi = (\pi_1, \pi_2^{\mathrm{T}})^{\mathrm{T}}$ is determined by the system of constraint equations

$$\begin{aligned} Q(\pi) &= Q(\pi, \hat{\beta}, \hat{\theta}, \hat{a}) \\ &= \sum_{i=1}^{n_1} \frac{r(x_i, \hat{\beta}, \hat{\theta}, \hat{a})}{1 + \pi^{\mathrm{T}} r(x_i, \hat{\beta}, \hat{\theta}, \hat{a})} = 0 \end{aligned} \tag{2.6}$$

with

$$r(x, \beta, \theta, a) = \begin{pmatrix} 1 - \theta\, w^{-1}(x, \beta) \\ w^{-1}(x, \beta)\{a(x) - a\} \end{pmatrix}.$$

One advantage of taking the transformations $\pi_1 = \hat{\theta}\lambda_1 - 1$ and $\pi_2 = \hat{\theta}\lambda_2$ is that the true values of π_1 and π_2 are 0, making our formulation similar to the standard empirical likelihood set-up of Owen (1990). Another advantage of introducing new Lagrange multipliers π_1 and π_2 is that the expression of p_i in equation (2.5) and the constraint equations (2.6) are analogous to their empirical likelihood counterparts in the usual unbiased sampling problem, so the optimization method of computing the empirical likelihood that was discussed in Owen (2001) can be applied to seek π_1 and π_2 satisfying constraint (2.6).

Let $\hat{\pi}$ denote a solution of $Q(\pi)=0$. Moreover, with $\hat{\pi}$ in place of π in equation (2.5), let $\hat{p}_i = p_i(\hat{\pi})$ for $i=1,\ldots,n_1$. Then we propose to estimate $\mu = E(Y) = \int y\, f(y,x)\,\mathrm{d}x\,\mathrm{d}y$ by the empirical-likelihood-based estimator

$$\begin{aligned}\hat{\mu}_{\mathrm{EL}} &= \sum_{i=1}^{n_1} \hat{p}_i y_i \\ &= \frac{1}{n_1}\sum_{i=1}^{n_1} \frac{\hat{\theta}\, w^{-1}(x_i,\hat{\beta})}{1+\hat{\pi}^{\mathrm{T}} r(x_i,\hat{\beta},\hat{\theta},\hat{a})} y_i \\ &= \frac{1}{n_1}\sum_{i=1}^{n} \frac{\hat{\theta}\, w^{-1}(x_i,\hat{\beta})}{1+\hat{\pi}^{\mathrm{T}} r(x_i,\hat{\beta},\hat{\theta},\hat{a})} d_i y_i. \end{aligned} \tag{2.7}$$

By contrast, Robins *et al.* (1994) proposed to estimate μ by

$$\hat{\mu}_{\mathrm{R}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{d_i y_i}{w(x_i,\hat{\beta})} - \frac{d_i - w(x_i,\hat{\beta})}{w(x_i,\hat{\beta})} m(x_i)\right\} \tag{2.8}$$

for some specified scalar function $m(x)$. Throughout this paper, let (β_0,θ_0) denote the true value of (β,θ) and let $\mu(x)=E(Y|X=x)$ and $\mu_2(x)=E(Y^2|X=x)$. Moreover, let (μ_0,a_0) denote the true value of (μ,a) in the one-sample problem of this section.

The choices of $a(x)$ in equation (2.7) and $m(x)$ in equation (2.8) play a crucial role in the estimation of μ. When $m(x)$ is chosen to be the regression function $\mu(x)$, Robins *et al.* (1995) and Hahn (1998) proved that $\hat{\mu}_{\mathrm{R}}$ achieves the semiparametric lower bound for the asymptotic variance of any regular estimator in a semiparametric missing data problem. In general, however, we do not know the form of $\mu(x)$; the best that we can do is to field a guess. On the basis of this guess about $\mu(x)$, we can specify a form of $a(x)$ in constraints (2.3). We present our results for three different forms of $a(x)$ along with some other results in the following subsections.

2.1. $a(x)=\mu(x)$; the guess is correct

If it happens that $a(x)=m(x)=\mu(x)=E(Y|X=x)$, the following theorem establishes the asymptotic equivalence of $\hat{\mu}_{\mathrm{EL}}$ and $\hat{\mu}_{\mathrm{R}}$ when $w(x,\beta)$ is also correctly specified. Write

$$\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$$

and

$$\hat{\mu} = n^{-1}\sum_{i=1}^{n} \mu(x_i).$$

Theorem 1. The empirical-likelihood-based estimator $\hat{\mu}_{\mathrm{EL}}$ and the estimator $\hat{\mu}_{\mathrm{R}}$ of Robins *et al.* (1994) have the following properties.

(a) Both $\hat{\mu}_{\mathrm{EL}}$ and $\hat{\mu}_{\mathrm{R}}$ are consistent estimators of μ_0 even if $w(x,\beta)$ is misspecified.
(b) Suppose that the propensity score function $w(x,\beta)$ is correctly specified. As $n\to\infty$, we can write $\hat{\mu}_{\mathrm{EL}} = \hat{\mu}_{\mathrm{R}} + O_p(n^{-1})$. Consequently, $\hat{\mu}_{\mathrm{EL}}$ is identical to $\hat{\mu}_{\mathrm{R}}$ apart from a term of order $O_p(n^{-1})$ and is thus asymptotically efficient.
(c) If $w(x,\beta)$ is correctly specified, then we have

$$\mathrm{var}(\hat{\mu}) \leqslant \mathrm{var}(\bar{y}) \leqslant \mathrm{var}(\hat{\mu}_{\mathrm{R}}) + o(n^{-1}) = \mathrm{var}(\hat{\mu}_{\mathrm{EL}}) + o(n^{-1}).$$

Proof. For part (a), note first that the $\hat{p}_i$ depend only on $(x_1, d_1), \ldots, (x_n, d_n)$. Conditionally on $(x_1, d_1), \ldots, (x_n, d_n)$, we have

$$\begin{aligned} E\{\hat{\mu}_{\mathrm{EL}} | (x_1, d_1), \ldots, (x_n, d_n)\} &= \sum_{i=1}^{n_1} \hat{p}_i \, E(y_i | x_i) \\ &= \sum_{i=1}^{n_1} \hat{p}_i \, \mu(x_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} \mu(x_i) \end{aligned}$$

by using the third constraint in expression (2.3). Therefore,

$$\begin{aligned} E(\hat{\mu}_{\mathrm{EL}}) &= \frac{1}{n} \sum_{i=1}^{n} E\{\mu(x_i)\} \\ &= \mu_0. \end{aligned}$$

To establish the consistency of $\hat{\mu}_{\mathrm{EL}}$ for μ_0 when $w_\beta(x) = w(x, \beta)$ is misspecified, suppose that $w(x)$ is the true propensity score so that $P(D = 1 | X = x) = w(x)$. Let β_0^* denote the value of β that minimizes the Kullback–Leibler discrepancy

$$D(w_\beta, w) = \int \log\{w(x)/w(x, \beta)\} \, w(x) \, \mathrm{d}x$$

with respect to β. According to White (1982), $\hat{\beta} \to \beta_0^*$ in probability under suitable regularity conditions. Write $\theta_0^* = E\{w(X, \beta_0^*)\}$ and let π_0^* be the value of π that satisfies $Q^*(\pi_0^*) = 0$, where

$$Q^*(\pi^*) = E[d_i \, r(x_i, \beta_0^*, \theta_0^*, a_0)\{1 + \pi^{\mathrm{T}} r(x_i, \beta_0^*, \theta_0^*, a_0)\}^{-1}].$$

Since $\hat{\pi}$ solves the equation

$$Q_n^*(\pi) = \sum_{i=1}^{n} \frac{d_i \, r(x_i, \beta_0^*, \theta_0^*, a_0)}{1 + \pi^{\mathrm{T}} \, r(x_i, \beta_0^*, \theta_0^*, a_0)} = 0,$$

it can be shown that $\hat{\pi} \to \pi_0^*$ in probability under suitable regularity conditions. An application of lemma 7.2.2A of Serfling (1980), page 253, gives

$$\begin{aligned} \hat{\mu}_{\mathrm{EL}} &= \sum_{i=1}^{n_1} \hat{p}_i \{y_i - \mu(x_i)\} + \sum_{i=1}^{n_1} \hat{p}_i \, \mu(x_i) \\ &= \frac{n}{n_1} \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\theta} \, d_i \, w^{-1}(x_i, \hat{\beta})\{y_i - \mu(x_i)\}}{1 + \hat{\pi}^{\mathrm{T}} r(x_i, \hat{\beta}, \hat{\theta}, \hat{a})} + \frac{1}{n} \sum_{i=1}^{n} \mu(x_i) \\ &\to \frac{\theta_0^*}{P(D = 1)} E\left[\frac{d_1 \, w^{-1}(x_1, \beta_0^*)\{y_1 - \mu(x_1)\}}{1 + \pi_0^{*\mathrm{T}} r(x_1, \beta_0^*, \theta_0^*, a_0)} \right] + E\{\mu(x_1)\} = \mu_0 \end{aligned}$$

in probability as $n \to \infty$, since $E[\{y_1 - \mu(x_1)\} | x_1] = 0$.

For part (b), the asymptotic expansions of $\hat{\beta}$, $\hat{\theta}$ and $\hat{\pi}$ in equations (A.1), (A.2) and (A.3) of Appendix A imply that $\hat{\beta} - \beta_0 = O_p(n^{-1/2})$, $\hat{\theta} - \theta_0 = O_p(n^{-1/2})$ and $\hat{\pi} = O_p(n^{-1/2})$. Furthermore, the asymptotic expansions (A.1), (A.2) and (A.3) with $a(x) = \mu(x)$ yield the following asymptotic expression for $\hat{\mu}_{\mathrm{EL}}$:

$$\begin{aligned}
\hat{\mu}_{\mathrm{EL}} &= \frac{1}{n_1}\sum_{i=1}^{n}\frac{\hat{\theta}\, w^{-1}(x_i,\hat{\beta})}{1+\hat{\pi}^{\mathrm{T}} r(x_i,\hat{\beta},\hat{\theta},\hat{\mu})}d_i\{y_i-\mu(x_i)\}+\hat{\mu}\\
&= \frac{1}{n_1}\sum_{i=1}^{n}\frac{\theta_0\, d_i\{y_i-\mu(x_i)\}}{w(x_i,\beta_0)}-\frac{1}{n_1}\sum_{i=1}^{n}\frac{\theta_0\, d_i\{y_i-\mu(x_i)\}}{w(x_i,\beta_0)}r^{\mathrm{T}}(x_i,\beta_0,\theta_0,\mu_0)\hat{\pi}\\
&\quad -\frac{1}{n_1}\sum_{i=1}^{n}\frac{\theta_0\, d_i\{y_i-\mu(x_i)\}}{w^2(x_i,\beta_0)}v^{\mathrm{T}}(x_i,\beta_0)(\hat{\beta}-\beta_0)\\
&\quad +\frac{1}{n_1}\sum_{i=1}^{n}\frac{d_i\{y_i-\mu(x_i)\}}{w(x_i,\beta_0)}(\hat{\theta}-\theta_0)+\hat{\mu}+O_p(n^{-1})\\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{d_i\{y_i-\mu(x_i)\}}{w(x_i,\beta_0)}+\mu(x_i)\right]+O_p(n^{-1})\\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{d_i\{y_i-\mu(x_i)\}}{w(x_i,\hat{\beta})}+\mu(x_i)\right]+O_p(n^{-1})\\
&= \hat{\mu}_{\mathrm{R}}+O_p(n^{-1}),
\end{aligned}$$

where $v(x,\beta)=\partial w(x,\beta)/\partial\beta$. Thus, $\hat{\mu}_{\mathrm{EL}}$ and $\hat{\mu}_{\mathrm{R}}$ are asymptotically equivalent. Moreover, $n^{1/2}(\hat{\mu}_{\mathrm{EL}}-\mu_0)\to N(0,\sigma_{\mathrm{R}}^2)$ and $n^{1/2}(\hat{\mu}_{\mathrm{R}}-\mu_0)\to N(0,\sigma_{\mathrm{R}}^2)$ in distribution with

$$\sigma_{\mathrm{R}}^2 = E\left\{\frac{\mu_2(x_1)-\mu^2(x_1)}{w(x_1,\beta_0)}\right\}+\mathrm{var}\{\mu(x_1)\}. \tag{2.9}$$

It has been shown by Robins *et al.* (1995) and Hahn (1998) that σ_{R}^2 is the lower bound for the asymptotic variance of any regular estimator in a semiparametric missing data problem.

For part (c), we have

$$\begin{aligned}
\mathrm{var}(\bar{y}) &= \frac{1}{n}\mathrm{var}(Y_1)\\
&= \frac{1}{n}\mathrm{var}\{E(Y|X)\}+\frac{1}{n}E\{\mathrm{var}(Y|X)\}\\
&= \frac{1}{n}\mathrm{var}\{\mu(X)\}+\frac{1}{n}E\{\mathrm{var}(Y|X)\}\\
&\geqslant \mathrm{var}\left\{\frac{1}{n}\sum_{i=1}^{n}\mu(X_i)\right\}=\mathrm{var}(\hat{\mu}).
\end{aligned}$$

Since $1/w(x,\beta_0)\geqslant 1$, we have

$$\begin{aligned}
\mathrm{var}(\bar{Y}) &\leqslant \frac{1}{n}\,\mathrm{var}\{\mu(X)\}+\frac{1}{n}\,E\left\{\frac{\mathrm{var}(Y|X)}{w(X,\beta_0)}\right\}\\
&= \mathrm{var}(\hat{\mu}_{\mathrm{R}})+o(n^{-1}).
\end{aligned}$$

Finally, using the fact that

$$\mathrm{var}(Y)=\mathrm{var}\{E(Y|X)\}+E\{\mathrm{var}(Y|X)\},$$

we have

$$\begin{aligned}
\mathrm{var}(\hat{\mu}_{\mathrm{EL}}) &= \mathrm{var}\left\{\frac{1}{n}\sum_{i=1}^{n}\mu(x_i)\right\}+E\left\{\sum_{i=1}^{n_1}\hat{p}_i^2\,\mathrm{var}(y_i|x_i)\right\}\\
&= \frac{1}{n}\,\mathrm{var}\{\mu(x_1)\}+E\left\{\sum_{i=1}^{n_1}\hat{p}_i^2\,\mathrm{var}(y_i|x_i)\right\}.
\end{aligned}$$

Using the asymptotic expansions of $\hat{\beta}, \hat{\theta}$ and $\hat{\pi}$ in equations (A.1), (A.2) and (A.3) of Appendix A gives

$$\begin{aligned}\sum_{i=1}^{n_1} \hat{p}_i^2 \operatorname{var}(y_i|x_i) &= \frac{1}{n_1^2} \sum_{i=1}^{n_1} \frac{\hat{\theta}^2}{w^2(x_i, \hat{\beta})} \frac{1}{\{1+\hat{\pi}^{\mathrm{T}} r(x_i, \hat{\beta}, \hat{\theta}, \hat{a})\}^2} \operatorname{var}(y_i|x_i) \\ &= \frac{1}{n_1^2} \sum_{i=1}^{n} \frac{\hat{\theta}^2}{w^2(x_i, \hat{\beta})} \frac{1}{\{1+\hat{\pi}^{\mathrm{T}} r(x_i, \hat{\beta}, \hat{\theta}, \hat{a})\}^2} d_i \operatorname{var}(y_i|x_i) \\ &= \frac{1}{n_1^2} \sum_{i=1}^{n} \frac{\theta_0^2}{w^2(x_i, \beta_0)} d_i \operatorname{var}(y_i|x_i) + O_p(n^{-3/2}) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} d_i \frac{\mu_2(x_i) - \mu^2(x_i)}{w^2(x_i, \beta_0)} + O_p(n^{-3/2}),\end{aligned}$$

which implies that

$$\operatorname{var}(\hat{\mu}_{\mathrm{EL}}) = \frac{1}{n} \operatorname{var}\{\mu(x_1)\} + \frac{1}{n} E\left\{ \frac{\mu_2(x_1) - \mu^2(x_1)}{w(x_1, \beta_0)} \right\} + o(n^{-1}).$$

This completes the proof of theorem 1.

2.2. $\mu(x)$ is a linear combination of $a(x)$

Suppose that the regression model can be written in the form of a generalized additive model given by

$$\mu(x) = E(Y|X=x) = c_0 + \sum_{k=1}^{q} c_k \, a_k(x)$$

for some known functions $a_k(x)$ and unknown coefficients c_k, $k = 0, 1, \ldots, q$. Recall that, when the dimension of x is high, the generalized additive model is commonly used to fit the regression function $\mu(x)$. If we can guess $a(x) = (a_1(x), \ldots, a_q(x))^{\mathrm{T}}$ correctly, the following theorem shows that the empirical-likelihood-based approach does not require knowledge of the coefficients $c_1, \ldots, c_q$, while still producing asymptotically efficient estimators of μ. By contrast, since the coefficients $c_0, c_1, \ldots, c_q$ are unknown, the estimator $\hat{\mu}_{\mathrm{R}}$ of Robins *et al.* (1994) with $m(x) = \mu(x)$ is not directly applicable in this case.

Theorem 2. The empirical-likelihood-based estimator $\hat{\mu}_{\mathrm{EL}}$ has the following properties.

(a) $\hat{\mu}_{\mathrm{EL}}$ is consistent even if $w(x, \beta)$ is misspecified.
(b) Suppose that the propensity score function $w(x, \beta)$ is correctly specified. As $n \to \infty$, we can write $\hat{\mu}_{\mathrm{EL}} = \hat{\mu}_{\mathrm{R}} + O_p(n^{-1})$. As a result, the empirical-likelihood-based estimator $\hat{\mu}_{\mathrm{EL}}$ is asymptotically efficient.

Proof. For part (a), conditionally on $(x_1, d_1), \ldots, (x_n, d_n)$, we have

$$\begin{aligned}E\{\hat{\mu}_{\mathrm{EL}}|(x_1, d_1), \ldots, (x_n, d_n)\} &= \sum_{i=1}^{n_1} \hat{p}_i \, \mu(x_i) \\ &= \sum_{i=1}^{n_1} \hat{p}_i \left\{ c_0 + \sum_{k=1}^{q} c_k \, a_k(x_i) \right\} \\ &= c_0 + \sum_{k=1}^{q} c_k \left\{ \sum_{i=1}^{n_1} \hat{p}_i \, a_k(x_i) \right\}\end{aligned}$$

$$= c_0 + \sum_{k=1}^{q} c_k \left\{ \frac{1}{n} \sum_{i=1}^{n} a_k(x_i) \right\}$$
$$= \frac{1}{n} \sum_{i=1}^{n} \mu(x_i),$$

which leads to $E(\hat{\mu}_{\mathrm{EL}}) = \mu_0$. The consistency of $\hat{\mu}_{\mathrm{EL}}$ for μ_0 can be established in a manner similar to the proof of part (a) of theorem 1, by noting that

$$\hat{\mu}_{\mathrm{EL}} = \sum_{i=1}^{n_1} \hat{p}_i \{ y_i - \mu(x_i) \} + \sum_{i=1}^{n_1} \hat{p}_i \, \mu(x_i)$$
$$= \frac{n}{n_1} \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\theta} d_i \, w^{-1}(x_i, \hat{\beta}) \{ y_i - \mu(x_i) \}}{1 + \hat{\pi}^{\mathrm{T}} \, r(x_i, \hat{\beta}, \hat{\theta}, \hat{a})} + \frac{1}{n} \sum_{i=1}^{n} \mu(x_i).$$

Part (b) can be proved by using the same argument as in the proof of part (b) of theorem 1. The proof is complete. □

Theorem 2 indicates that the empirical-likelihood-based estimator $\hat{\mu}_{\mathrm{EL}}$ is asymptotically equivalent to the estimator $\hat{\mu}_{\mathrm{R}}$ of Robins *et al.* (1994) in the case of $m(x) = \mu(x) = c_0 + \Sigma_{k=1}^{q} c_k \, a_k(x)$. Since $\hat{\mu}_{\mathrm{R}}$ involves unknown coefficients $c_0, c_1, \ldots, c_q$, it is not directly applicable without first estimating these coefficients. Nevertheless, the asymptotic equivalence of $\hat{\mu}_{\mathrm{EL}}$ and $\hat{\mu}_{\mathrm{R}}$ immediately implies that $\hat{\mu}_{\mathrm{EL}}$ is asymptotically efficient by using the results of Robins *et al.* (1995) and Hahn (1998). One attractive feature of $\hat{\mu}_{\mathrm{EL}}$ is that it avoids the problem of determining unknown coefficients $c_0, c_1, \ldots, c_q$, while still achieving full efficiency for estimating μ. This is a very desirable property that is obtained by employing the method of empirical likelihood.

2.3. Arbitrary $a(x)$

When we have no knowledge for specifying the true regression function $\mu(x) = E(Y|X = x)$ either completely or partially as discussed in Sections 2.1 and 2.2, we can field a guess for $\mu(x)$ as a linear combination of $a(x) = (a_1(x), \ldots, a_q(x))^{\mathrm{T}}$. We now study the asymptotic distribution of the empirical-likelihood-based estimator $\hat{\mu}_{\mathrm{EL}}$ in the case of correctly specified propensity score $w(x, \beta)$ but arbitrary auxiliary information function $a(x)$, along with the double-robustness property of $\hat{\mu}_{\mathrm{EL}}$. Write

$$s(x, \beta, \theta, a) = C^{\mathrm{T}} \, S^{-1} r(x, \beta, \theta, a) + \frac{K \, v(x, \beta)}{w(x, \beta)\{1 - w(x, \beta)\}},$$
$$C = E[r(x_1, \beta_0, \theta_0, a_0)\{\mu(x_1) - \mu_0\}],$$
$$D = -E\{r(x_1, \beta_0, \theta_0, a_0) v^{\mathrm{T}}(x, \beta_0)\},$$
$$S = E\{w(x_1, \beta_0) \, r(x_1, \beta_0, \theta_0, a_0) \, r^{\mathrm{T}}(x_1, \beta_0, \theta_0, a_0)\},$$
$$H = E\left[\frac{v(x_1, \beta_0)}{w(x_1, \beta_0)} \{\mu(x_1) - \mu_0\} \right],$$
$$S_B = E\left[\frac{v(x_1, \beta_0) \, v^{\mathrm{T}}(x_1, \beta_0)}{w(x_1, \beta_0)\{1 - w(x_1, \beta_0)\}} \right],$$
$$K = (C^{\mathrm{T}} S^{-1} D + H^{\mathrm{T}}) S_{\mathrm{B}}^{-1}. \tag{2.10}$$

Theorem 3. The empirical-likelihood-based estimator $\hat{\mu}_{\mathrm{EL}}$ and the estimator $\hat{\mu}_{\mathrm{R}}$ of Robins *et al.* (1994) have the following properties.

(a) Suppose that the propensity score $w(x, \beta)$ is correctly specified. Under suitable regularity conditions, we can write

$$\hat{\mu}_{\mathrm{EL}} - \mu_0 = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{d_i (y_i - \mu_0)}{w(x_i, \beta_0)} - \{d_i - w(x_i, \beta_0)\}\, s(x_i, \beta_0, \theta_0, a_0) \right] + o_p(n^{-1/2}). \tag{2.11}$$

As a result, as $n \to \infty$,

$$n^{1/2}(\hat{\mu}_{\mathrm{EL}} - \mu_0) \to N(0, \sigma^2)$$

in distribution, where σ^2 is equal to

$$\sigma^2 = \sigma_{\mathrm{R}}^2 + E\left[\frac{1 - w(x_1, \beta_0)}{w(x_1, \beta_0)} \{\mu(x_1) - \mu_0 - w(x_1, \beta_0)\, s(x_1, \beta_0, \theta_0, a_0)\}^2 \right]$$

with σ_{R}^2 given by equation (2.9).

(b) Let $q = 1$. Both $\hat{\mu}_{\mathrm{EL}}$ and $\hat{\mu}_{\mathrm{R}}$ have the double-robustness property, i.e. both $\hat{\mu}_{\mathrm{EL}}$ and $\hat{\mu}_{\mathrm{R}}$ are consistent estimators of μ_0 if either the propensity score $w(x, \beta)$ is correctly specified or $a(x) = \mu(x)$.

(c) Let $q = 1$. Suppose that $\mu(x) = c_1 + c_2\, a(x)$ for some constants c_1 and c_2 such that $a(x) \neq \mu(x)$. Then $\hat{\mu}_{\mathrm{EL}}$ is still consistent even if $w(x, \beta)$ is misspecified. Nevertheless, $\hat{\mu}_{\mathrm{R}}$ is not consistent in this case.

Proof. The proof of part (a) is given in Appendix A. For part (b), if the propensity score $w(x, \beta)$ is correctly specified, then the two terms in the asymptotic expansion of $\hat{\mu}_{\mathrm{EL}}$ in equation (2.11) of part (a) converge to 0 in probability, and thus $\hat{\mu}_{\mathrm{EL}}$ is a consistent estimator of μ_0. We need to show only that $\hat{\mu}_{\mathrm{EL}}$ is consistent if $a(x) = \mu(x) = E(Y|X = x)$. Note that the $\hat{p}_i$ depend on only $(x_1, d_1), \ldots, (x_n, d_n)$. Conditionally on $(x_1, d_1), \ldots, (x_n, d_n)$, we have

$$\begin{aligned} E\{\hat{\mu}_{\mathrm{EL}} | (x_1, d_1), \ldots, (x_n, d_n)\} &= \sum_{i=1}^{n_1} \hat{p}_i\, E(y_i | x_i) \\ &= \sum_{i=1}^{n_1} \hat{p}_i\, \mu(x_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} a(x_i) \end{aligned}$$

by using the constraint on auxiliary information in expression (2.3). Therefore,

$$\begin{aligned} E(\hat{\mu}_{\mathrm{EL}}) &= \frac{1}{n} \sum_{i=1}^{n} E\{a(x_i)\} \\ &= \mu_0. \end{aligned}$$

For part (c), applying the same argument as in part (b), we have

$$\begin{aligned} E\{\hat{\mu}_{\mathrm{EL}} | (x_1, d_1), \ldots, (x_n, d_n)\} &= \sum_{i=1}^{n_1} \hat{p}_i\, \mu(x_i) \\ &= \sum_{i=1}^{n_1} \hat{p}_i \{c_1 + c_2\, a(x_i)\} \end{aligned}$$

$$= c_1 + c_2 \frac{1}{n} \sum_{i=1}^{n} a(x_i)$$
$$= \frac{1}{n} \sum_{i=1}^{n} \mu(x_i),$$

which leads to $E(\hat{\mu}_{\mathrm{EL}}) = \mu_0$. As for $\hat{\mu}_{\mathrm{R}}$, we have

$$E\{\hat{\mu}_{\mathrm{R}} | (d_1, x_1), \ldots, (d_n, x_n)\} = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i\{\mu(x_i) - m(x_i)\}}{w(x_i, \hat{\beta})} + \frac{1}{n} \sum_{i=1}^{n} m(x_i),$$

which implies that $\hat{\mu}_{\mathrm{R}}$ is not a consistent estimator of μ_0 if $w(x, \beta)$ is misspecified. The proof is complete.

Part (a) of theorem 3 indicates that $\sigma^2 = \sigma_{\mathrm{R}}^2$ when

$$\mu(x) = \mu_0 + w(x, \beta_0)\, s(x, \beta_0, \theta_0, a_0)$$
$$= \mu_0 + C^{\mathrm{T}} S^{-1} \begin{pmatrix} w(x, \beta_0) - \theta_0 \\ a(x) - a_0 \end{pmatrix} + \frac{K\, v(x, \beta_0)}{1 - w(x, \beta_0)}.$$

In this case, the empirical-likelihood-based estimator $\hat{\mu}_{\mathrm{EL}}$ is asymptotically efficient. This result is closely related to that of part (b) of theorem 2.

Remark 1. For the choice of $a(x)$ in practice, we can plot the response y against each component of the covariate vector x. On the basis of the patterns in these graphs, we may naturally impose a linear or non-linear constraint by specifying a form of $a(x)$.

Remark 2. Part (a) of theorem 3 may be employed to construct Wald confidence intervals for μ. We may also construct empirical-likelihood-based confidence intervals for μ by imposing an additional constraint

$$\sum_{i=1}^{n_1} p_i(y_i - \mu) = 0$$

in expression (2.3). The empirical likelihood ratio statistic is twice the difference between the maximized empirical log-likelihood with constraints (2.3) and the maximized empirical log-likelihood with constraints (2.3) plus the aforementioned constraint. This statistic is, however, no longer asymptotically distributed as a standard χ^2-variable, but a weighted χ^2-variable. Wang and Rao (2002) showed that, in general, the empirical likelihood ratio statistic has an asymptotic weighted χ^2-distribution if the constraint equations involve estimated parameters. In practice, since the limiting weighted χ^2-distribution is complicated, we may construct a hybrid of bootstrap and empirical likelihood ratio confidence intervals for μ, as discussed in the simulation study of Section 4.

3. Main results in a two-sample missing response problem or causal inference

Let us begin with some notation. Let D_i denote a dummy variable such that $D_i = 1$ when treatment is given to the ith individual, and $D_i = 0$ otherwise. Let Y_{0i} and Y_{1i} denote potential outcomes when $D_i = 0$ and $D_i = 1$ respectively. The difference $Y_{1i} - Y_{0i}$ is called the treatment effect for the ith individual. Individual treatment effects cannot be observed. We can only observe D_i and

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}.$$

Let $\mu_1 = E(Y_{1i})$ and $\mu_0 = E(Y_{0i})$. Then the average treatment effect is defined as

$$\Delta = E(Y_{1i} - Y_{0i}) = \mu_1 - \mu_0$$

and the average treatment effect on the treated is defined as

$$\gamma = E(Y_{1i} - Y_{0i} | D_i = 1).$$

In causal inference, $w(x, \beta) = P(D = 1 | X = x)$ is the conditional probability of programme participation given some observed characteristics and is called the propensity score. The key assumption is that D and (Y_1, Y_0) are independent of each other given the covariate vector X. Under this assumption, the joint density of (Y_0, Y_1, D, X) is given by

$$f(y_0, y_1 | x)\, w(x, \beta)^d \{1 - w(x, \beta)\}^{1-d}\, g(x),$$

where $f(y_0, y_1 | x)$ and $g(x)$ represent the conditional density of (Y_0, Y_1) given X and the marginal density of X respectively. Furthermore, the joint density of observed (Y, D, X) is

$$\{f_1(y|x)\, g(x)\, w(x, \beta)\}^d [f_0(y|x)\, g(x) \{1 - w(x, \beta)\}]^{1-d},$$

where $f_1(\cdot|x) = \int f(y_0, \cdot\, | x)\, \mathrm{d}y_0$ and $f_0(\cdot|x) = \int f(\cdot, y_1 | x)\, \mathrm{d}y_1$.

Let $f_1(y, x) = f_1(y|x)\, g(x)$ and $f_0(y, x) = f_0(y|x)\, g(x)$ denote respectively the joint density of (Y_1, X) and (Y_0, X). Since $f_1(y, x)$ and $f_0(y, x)$ have common marginal density $g(x)$, the probability of selecting treatment $(D = 1)$ is given by

$$\begin{aligned}
\theta &= P(D = 1) \\
&= \int w(x, \beta)\, g(x)\, \mathrm{d}x \\
&= \int \int w(x, \beta)\, f_1(y, x)\, \mathrm{d}y\, \mathrm{d}x \\
&= \int \int w(x, \beta)\, f_0(y, x)\, \mathrm{d}y\, \mathrm{d}x.
\end{aligned}$$

Furthermore, the average treatment effect is equal to

$$\begin{aligned}
\Delta &= E(Y_1 - Y_0) \\
&= \int \int y\, f_1(y, x)\, \mathrm{d}y\, \mathrm{d}x - \int \int y\, f_0(y, x)\, \mathrm{d}y\, \mathrm{d}x \\
&= \mu_1 - \mu_0.
\end{aligned}$$

If parametric models are specified for $f_1(y|x)$ and $f_0(y|x)$, the statistical inference on the underlying parameters can be made on the basis of $f_1(y|x)^d f_0(y|x)^{1-d}$, which completely ignores the propensity score. This inference procedure also applies to estimating equations. However, Rosenbaum and Rubin (1983) and Imbens (2004) pointed out that, since the dimension of X is very high in practical applications, it is difficult to specify a model for $P(Y|X)$. Hahn (1998) proposed several efficient semiparametric methods to estimate the average treatment effect by using the estimated conditional expectations, such as $E(DY|X = x)$, $E\{(1 - D)Y|X = x\}$ and $E(D|X = x)$. Because of the curse of dimensionality problem, the small sample size performance of these methods is an open problem. As an alternative, the propensity score was introduced by Rosenbaum and Rubin (1983), which is a probability function of choosing a treatment for given covariate information. They advocated modelling the propensity score with a logistic regression model. Hirano *et al.* (2003) used a Horvitz and Thompson (1952) type of estimator with a nonparametric estimator of the propensity score. Under some regularity conditions, they

proved that this estimator achieves semiparametric efficiency; however, the small sample size performance of the estimator has not been assessed. In addition, the estimator may still suffer from the same curse of dimensionality problem for a high dimensional covariate x.

As discussed in Section 1, the estimation of the average treatment effect in observational studies can be treated as a two-sample missing data problem. To estimate μ_0 and μ_1, we can use data $\{(Y_{0i}, D_i, X_i),\ i=1,\ldots,n\}$ and $\{(Y_{1i}, D_i, X_i),\ i=1,\ldots,n\}$ separately, where Y_{0i} is missing if $D_i=1$ and Y_{1i} is missing if $D_i=0$. Throughout this section, let

$$\mu_1(x)=E(Y_1|X=x),$$

$$\mu_0(x)=E(Y_0|X=x),$$

$$\mu_{12}(x)=E(Y_1^2|X=x)$$

and

$$\mu_{02}(x)=E(Y_0^2|X=x).$$

According to the results in the one-sample case of unspecified $\mu_1(x)$ in Section 2.3 with $a_1(x)=(a_{11}(x),\ldots,a_{1q_1}(x))^{\mathrm{T}}$ and $a_1=(a_{11},\ldots,a_{1q_1})^{\mathrm{T}}=E\{a_1(X)\}$, the empirical-likelihood-based estimator of $\mu_1=\int\int y\, f_1(y,x)\,\mathrm{d}y\,\mathrm{d}x$ is given by

$$\begin{aligned}\hat{\mu}_{1,\mathrm{EL}}&=\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\hat{\theta}\,w^{-1}(x_i,\hat{\beta})}{1+\hat{\pi}_1^{\mathrm{T}}\,r_1(x_i,\hat{\beta},\hat{\theta},\hat{a}_1)}y_{1i}\\&=\frac{1}{n_1}\sum_{i=1}^{n}\frac{\hat{\theta}\,w^{-1}(x_i,\hat{\beta})}{1+\hat{\pi}_1^{\mathrm{T}}\,r(x_i,\hat{\beta},\hat{\theta},\hat{a}_1)}d_i y_{1i},\end{aligned}\tag{3.1}$$

where $\hat{a}_1=n^{-1}\Sigma_{i=1}^{n}a_1(x_i)$ and $\hat{\pi}_1$ is determined by the system of constraint equations

$$\sum_{i=1}^{n_1}\frac{r_1(x_i,\hat{\beta},\hat{\theta},\hat{a}_1)}{1+\pi_1^{\mathrm{T}}\,r_1(x_i,\hat{\beta},\hat{\theta},\hat{a}_1)}=0$$

with

$$r_1(x,\beta,\theta,a_1)=\begin{pmatrix}1-\theta\,w^{-1}(x,\beta)\\ w^{-1}(x,\beta)\{a_1(x)-a_1\}\end{pmatrix}.$$

In a similar manner, we can show that, in the case of unspecified $\mu_0(x)$ with $a_0(x)=(a_{01}(x),\ldots,a_{0q_0}(x))^{\mathrm{T}}$ and $a_0=(a_{01},\ldots,a_{0q_0})^{\mathrm{T}}=E\{a_0(X)\}$, the empirical-likelihood-based estimator of $\mu_0=\int\int y\, f_0(y,x)\,\mathrm{d}y\,\mathrm{d}x$ is given by

$$\begin{aligned}\hat{\mu}_{0,\mathrm{EL}}&=\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{(1-\hat{\theta})\{1-w(x_i,\hat{\beta})\}^{-1}}{1+\hat{\pi}_0^{\mathrm{T}}\,r_0(x_i,\hat{\beta},\hat{\theta},\hat{a}_0)}y_{0i}\\&=\frac{1}{n_0}\sum_{i=1}^{n}\frac{(1-\hat{\theta})\{1-w(x_i,\hat{\beta})\}^{-1}}{1+\hat{\pi}_0^{\mathrm{T}}\,r_0(x_i,\hat{\beta},\hat{\theta},\hat{a}_0)}(1-d_i)y_{0i},\end{aligned}\tag{3.2}$$

where $\hat{a}_0=n^{-1}\Sigma_{i=1}^{n}a_0(x_i)$ and $\hat{\pi}_0$ is determined by the system of constraint equations

$$\sum_{i=1}^{n_0}\frac{r_0(x_i,\hat{\beta},\hat{\theta},\hat{a}_0)}{1+\pi_0^{\mathrm{T}}\,r_0(x_i,\hat{\beta},\hat{\theta},\hat{a}_0)}=0$$

with

$$r_0(x,\beta,\theta,a_0)=\begin{pmatrix}1-(1-\theta)\{1-w(x,\beta)\}^{-1}\\ \{1-w(x,\beta)\}^{-1}\{a_0(x)-a_0\}\end{pmatrix}.$$

Note that the different auxiliary information functions $a_1(x)$ and $a_0(x)$ are used in the construction of empirical-likelihood-based estimators $\hat{\mu}_{1,\mathrm{EL}}$ and $\hat{\mu}_{0,\mathrm{EL}}$ for estimating μ_1 and μ_0. On the basis of equations (3.1) and (3.2), we propose to estimate the average treatment effect $\Delta=\mu_1-\mu_0$ by $\hat{\Delta}_{\mathrm{EL}}=\hat{\mu}_{1,\mathrm{EL}}-\hat{\mu}_{0,\mathrm{EL}}$. When the true regression functions $\mu_1(x)$ and $\mu_0(x)$ are correctly specified, the estimator of Δ of Robins *et al.* (1994) is defined as

$$\hat{\Delta}_{\mathrm{R}}=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{d_i\{y_{1i}-\mu_1(x_i)\}}{w(x_i,\hat{\beta})}-\frac{(1-d_i)\{y_{0i}-\mu_0(x_i)\}}{1-w(x_i,\hat{\beta})}+\mu_1(x_i)-\mu_0(x_i)\right].$$

Robins *et al.* (1995) and Hahn (1998) showed that $\hat{\Delta}_{\mathrm{R}}$ is an asymptotically efficient estimator of Δ. Throughout this section, let $(\mu_{10},\mu_{00},a_{10},a_{00})$ denote the true value of (μ_1,μ_0,a_1,a_0) and $\Delta_0=\mu_{10}-\mu_{00}$ denote the true value of Δ. For $k=0,1$, write

$$s_k(x,\beta,\theta,a_1)=(-1)^{k+1}C_k^{\mathrm{T}}S_k^{-1}\,r_k(x,\beta,\theta,a_k)+\frac{K_k\,v(x,\beta)}{w(x,\beta)\{1-w(x,\beta)\}},$$

$$e_k(x,\beta,\theta,a_k)=\mu_k(x)-\mu_{k0}+(-1)^k\,w^k(x,\beta_0)\{1-w(x,\beta_0)\}^{1-k}\,s_k(x,\beta,\theta,a_k),$$

$$C_k=E[r_k(x_1,\beta_0,\theta_0,a_{10})\{\mu_k(x_1)-\mu_{k0}\}],$$

$$D_k=(-1)^k\,E\{r_k(x_1,\beta_0,\theta_0,a_{k0})\,v^{\mathrm{T}}(x,\beta_0)\},$$

$$S_k=E[w^k(x_1,\beta_0)\{1-w(x_1,\beta_0)\}^{1-k}\,r_k(x_1,\beta_0,\theta_0,a_{k0})\,r_k^{\mathrm{T}}(x_1,\beta_0,\theta_0,a_{k0})],$$

$$H_k=E\left[\frac{v(x_1,\beta_0)\{\mu_k(x_1)-\mu_{k0}\}}{w^k(x_1,\beta_0)\{1-w(x_1,\beta_0)\}^{1-k}}\right],$$

$$K_k=(C_k^{\mathrm{T}}S_k^{-1}D_k+H_k^{\mathrm{T}})S_{\mathrm{B}}^{-1}.$$

The following theorem studies the asymptotic behaviour of $\hat{\Delta}_{\mathrm{EL}}$.

Theorem 4. The empirical-likelihood-based estimator $\hat{\Delta}_{\mathrm{EL}}$ and the estimator $\hat{\Delta}_{\mathrm{R}}$ of Robins *et al.* (1994) have the following properties.

(a) Let $q_1=q_0=1$. Both $\hat{\Delta}_{\mathrm{EL}}$ and $\hat{\Delta}_{\mathrm{R}}$ have the double-robustness property, i.e. both $\hat{\Delta}_{\mathrm{EL}}$ and $\hat{\Delta}_{\mathrm{R}}$ are consistent estimators of Δ_0 if either the propensity score $w(x,\beta)$ is correctly specified or $a_k(x)=\mu_k(x)$ for $k=0,1$.

(b) Suppose that the propensity score function $w(x,\beta)$ is correctly specified and $a_k(x)=\mu_k(x)$ for $k=0,1$. As $n\to\infty$, we can write $\hat{\Delta}_{\mathrm{EL}}=\hat{\Delta}_{\mathrm{R}}+O_p(n^{-1})$. Consequently, $\hat{\Delta}_{\mathrm{EL}}$ is identical to $\hat{\Delta}_{\mathrm{R}}$ apart from a term of order $O_p(n^{-1})$ and is thus asymptotically efficient. Moreover, we have $n^{1/2}(\hat{\Delta}_{\mathrm{EL}}-\Delta_0)\to N(0,\sigma^2_{\Delta_{\mathrm{R}}})$ and $n^{1/2}(\hat{\Delta}_{\mathrm{R}}-\Delta_0)\to N(0,\sigma^2_{\Delta_{\mathrm{R}}})$ as $n\to\infty$, where

$$\sigma^2_{\Delta_{\mathrm{R}}}=E\left\{\frac{\mu_{12}(x_1)-\mu_1^2(x_1)}{w(x_1,\beta_0)}\right\}+E\left\{\frac{\mu_{02}(x_1)-\mu_0^2(x_1)}{1-w(x_1,\beta_0)}\right\}+\mathrm{var}\{\mu_1(x_1)-\mu_0(x_1)\}.$$

(c) Suppose that the propensity score function $w(x,\beta)$ is correctly specified and

$$\mu_k(x)=c_{k0}+\sum_{j=1}^{q_k}c_{kj}\,a_{kj}(x)$$

for some unknown coefficients c_{kj}, $k=0,1$, $j=0,1,\ldots,q_k$. As $n\to\infty$, we can write $\hat{\Delta}_{\mathrm{EL}}=\hat{\Delta}_{\mathrm{R}}+O_p(n^{-1})$ so that $n^{1/2}(\hat{\Delta}_{\mathrm{EL}}-\Delta_0)\to N(0,\sigma^2_{\Delta_{\mathrm{R}}})$. As a result, the empirical-likelihood-based estimator $\hat{\Delta}_{\mathrm{EL}}$ is asymptotically efficient.

(d) Suppose that the propensity score $w(x, \beta)$ is correctly specified. Under suitable regularity conditions, we can write

$$\hat{\Delta}_{\mathrm{EL}} - \Delta_0 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{d_i(y_i-\mu_{10})}{w(x_i,\beta_0)} - \frac{(1-d_i)(y_{0i}-\mu_{00})}{1-w(x_i,\beta_0)} - \{d_i - w(x_i,\beta_0)\}\{s_1(x_i,\beta_0,\theta_0,a_{10}) - s_0(x_i,\beta_0,\theta_0,a_{00})\}\right] + o_p(n^{-1/2}).$$

As a result, $n^{1/2}(\hat{\Delta}_{\mathrm{EL}} - \Delta_0) \to N(0, \sigma_{\Delta}^2)$ in distribution as $n \to \infty$, where σ_{Δ}^2 is equal to

$$\sigma_{\Delta}^2 = \sigma_{\Delta_{\mathrm{R}}}^2 + E\left(\frac{[\{1-w(x_1,\beta_0)\}\, e_1(x_1,\beta_0,\theta_0,a_{10}) + w(x_1,\beta_0)\, e_0(x_1,\beta_0,\theta_0,a_{00})]^2}{w(x_1,\beta_0)\{1-w(x_1,\beta_0)\}}\right).$$

The proofs of parts (a), (b) and (c) of theorem 4 are similar to those of part (b) of theorem 3, part (b) of theorem 1 and part (b) of theorem 2, whereas part (d) of theorem 4 can be proved by using part (a) of theorem 3; therefore, the proof of theorem 4 is omitted here.

Part (d) of theorem 4 implies that, if

$$\mu_k(x) = \mu_{k0} + (-1)^k\, w^k(x,\beta_0)\{1-w(x,\beta_0)\}^{1-k}\, s_k(x,\beta_0,\theta_0,a_{k0})$$

for $k = 0, 1$, then $\sigma_{\Delta}^2 = \sigma_{\Delta_{\mathrm{R}}}^2$. In this case, the empirical-likelihood-based estimator $\hat{\Delta}_{\mathrm{EL}}$ is asymptotically efficient. This result is closely related to that of part (c) of theorem 4.

Finally, we remark that the empirical-likelihood-based estimator of the average treatment effect on the treated is given by

$$\hat{\gamma}_{\mathrm{EL}} = \frac{1}{n_1}\sum_{i=1}^{n}\frac{1}{1+\hat{\pi}_1^{\mathrm{T}}\, r(x_i,\hat{\beta},\hat{\theta},\hat{a}_1)} d_i y_{1i} - \frac{1}{n_0}\sum_{i=1}^{n}\frac{\hat{\theta}^{-1}(1-\hat{\theta})\, w(x_i,\hat{\beta})\{1-w(x_i,\hat{\beta})\}^{-1}}{1+\hat{\pi}_0^{\mathrm{T}}\, r_0(x_i,\hat{\beta},\hat{\theta},\hat{a}_0)}(1-d_i)y_{0i}.$$

The asymptotic distribution of $\hat{\gamma}_{\mathrm{EL}}$ can be derived similarly to that of $\hat{\Delta}_{\mathrm{EL}}$ in theorem 4.

4. Simulation and comparison study

In this section we conduct a simulation study to compare four relevant estimators of μ under the set-up of Section 2. These four estimators are the sample mean $\bar{y} = n^{-1}\Sigma_{i=1}^{n} y_i$ with no missing data, Horvitz and Thompson's estimator

$$\hat{\mu}_1 = n_1^{-1}\sum_{i=1}^{n}\hat{\theta} y_i / w(x_i,\hat{\beta}),$$

the estimator of Robins *et al.* (1994),

$$\hat{\mu}_{\mathrm{R}} = \frac{1}{n}\sum_{i=1}^{n}\frac{d_i y_i}{w(x_i,\hat{\beta})} - \frac{1}{n}\sum_{i=1}^{n}\frac{d_i - w(x_i,\hat{\beta})}{w(x_i,\hat{\beta})}a(x_i),$$

and the empirical-likelihood-based estimator $\hat{\mu}_{\mathrm{EL}} = \Sigma_{i=1}^{n_1}\hat{p}_i y_i$. Since the regression estimator $\hat{\mu}_{\mathrm{REG}} = n^{-1}\Sigma_{i=1}^{n} a(x_i)$ is very sensitive to the choice of $a(x)$, we do not include it. Corresponding to the six choices of $a(x) = x$, $2+3x$, x^2, $2+3x^2$, x^4, $2+3x^4$, the six estimators of Robins *et al.* (1994) are respectively denoted by $\hat{\mu}_{\mathrm{RRZ1}}, \ldots, \hat{\mu}_{\mathrm{RRZ6}}$. Since the empirical-likelihood-based estimator does not depend on the coefficients of the regression function, we denote the estimator as $\hat{\mu}_{\mathrm{EL1}}$, $\hat{\mu}_{\mathrm{EL2}}$ and $\hat{\mu}_{\mathrm{EL3}}$ corresponding to $a(x) = x, x^2, x^4$.

The propensity score that is used in our simulation study is the logistic regression function given by

Table 1. Biases and mean-square errors (in parentheses) of $\bar{y}$, $\hat{\mu}_1$, $\hat{\mu}_{\mathrm{RRZ}}$ and $\hat{\mu}_{\mathrm{EL}}$ based on 5000 simulations with sample size $n = 200$†

Estimator	*Results for the following distributions:*		
	$y\|x \sim N(2+3x, x^2)$	$y\|x \sim N(2+3x^2, x^2)$	$y\|x \sim N(2+3x^4, x^2)$
$\bar{y}$	0.00608 (0.04957)	−0.00343 (0.09519)	0.03556 (4.34703)
$\hat{\mu}_1$	0.00559 (0.05417)	−0.00002 (0.15875)	0.01614 (8.44175)
$\hat{\mu}_{\mathrm{RRZ1}}$	0.00568 (0.05333)	0.00002 (0.16550)	0.01621 (8.49143)
$\hat{\mu}_{\mathrm{RRZ2}}$	0.00581 (0.05279)	0.00001 (0.17662)	0.01627 (8.56966)
$\hat{\mu}_{\mathrm{RRZ3}}$	0.00467 (0.06627)	−0.00128 (0.12551)	0.01531 (8.16212)
$\hat{\mu}_{\mathrm{RRZ4}}$	0.00276 (0.13259)	−0.00385 (0.09738)	0.01359 (7.62220)
$\hat{\mu}_{\mathrm{RRZ5}}$	−0.00418 (0.64579)	0.00406 (0.33108)	0.02223 (6.14622)
$\hat{\mu}_{\mathrm{RRZ6}}$	−0.02378 (5.12497)	0.01218 (3.47708)	0.03433 (4.34628)
$\hat{\mu}_{\mathrm{EL1}}$	0.00598 (0.05272)	−0.01281 (0.10136)	−0.10715 (5.61036)
$\hat{\mu}_{\mathrm{EL2}}$	0.00735 (0.05291)	−0.00387 (0.09734)	−0.11617 (4.60149)
$\hat{\mu}_{\mathrm{EL3}}$	0.01073 (0.05341)	0.02769 (0.11645)	0.03449 (4.34736)

†The propensity score is $P(D=1|X=x) = \exp(1+0.5x)/\{1+\exp(1+0.5x)\}$ with $x \sim N(0,1)$. The estimates $\hat{\mu}_{\mathrm{RRZ1}}, \ldots, \hat{\mu}_{\mathrm{RRZ6}}$ correspond to $a(x) = x, 2+3x, x^2, 2+3x^2, x^4, 2+3x^4$, whereas $\hat{\mu}_{\mathrm{EL1}}$, $\hat{\mu}_{\mathrm{EL2}}$ and $\hat{\mu}_{\mathrm{EL3}}$ correspond to $a(x) = x, x^2, x^4$.

$$P(D=1|X=x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}. \tag{4.1}$$

Furthermore, the regression model is assumed to be $Y|X=x \sim N\{\mu(x), x^2\}$. Moreover, the auxiliary information function $a(x)$ is chosen to be $a(x) = x$. For $\mu(x) = 2+3x, 2+3x^2, 2+3x^4$, $(\beta_0, \beta_1) = (1, 0.5)$ and sample size $n = 200$, the simulation results are summarized in Table 1 based on 5000 simulations. We observe the following from Table 1.

(a) If the auxiliary information function $a(x)$ is identical to the true regression function $\mu(x)$, the empirical-likelihood-based estimators and the estimators of Robins *et al.* (1994) are almost equivalent, both of which are better than Horvitz and Thompson's (1952) estimator.
(b) If the true regression function $\mu(x)$ is not among the components of the auxiliary information function $a(x)$, the empirical-likelihood-based estimator is superior to the other three estimators in terms of mean-square error. In some cases, the empirical-likelihood-based estimator has a significant reduction in mean-square error, as compared with the estimators of Robins *et al.* (1994) and Horvitz and Thompson (1952).
(c) When the true polynomial regression function $\mu(x)$ of order 4 is approximated by a quadratic function via the auxiliary information function $a(x)$, the empirical-likelihood-based estimator $\hat{\mu}_{\mathrm{EL}}$ performs quite well.

To assess the robustness of the empirical-likelihood-based estimator $\hat{\mu}_{\mathrm{EL}}$, we use model (4.1) as our working model and instead generate D from a binary model with complementary log–log-link:

$$P(D=1|x) = 1 - \exp\{-\exp(\beta_0+\beta_1 x)\}$$

For $\mu(x) = 2+3x, 2+3x^2, 2+3x^4$, $(\beta_0, \beta_1) = (1, 0.5)$ and sample size $n = 200$, the simulation results are summarized in Table 2 based on 5000 simulations. Table 2 reveals that $\hat{\mu}_{\mathrm{EL}}$ has a much smaller mean-square error than $\hat{\mu}_1$ and $\hat{\mu}_{\mathrm{RRZ}}$ when $a(x) \neq \mu(x)$, suggesting that $\hat{\mu}_{\mathrm{EL}}$ is very

Table 2. Biases and mean-square errors (in parentheses) of $\bar{y}$, $\hat{\mu}_1$, $\hat{\mu}_{\mathrm{RRZ}}$ and $\hat{\mu}_{\mathrm{EL}}$ based on 5000 simulations with sample size $n = 200$†

Estimator	*Results for the following distributions:*		
	$y\|x \sim N(2+3x, x^2)$	$y\|x \sim N(2+3x^2, x^2)$	$y\|x \sim N(2+3x^4, x^2)$
$\bar{y}$	−0.00209 (0.04956)	0.00604 (0.09714)	−0.00699 (4.32924)
$\hat{\mu}_1$	−0.01972 (0.06077)	0.10874 (0.17601)	0.70216 (13.90812)
$\hat{\mu}_{\mathrm{RRZ1}}$	−0.01222 (0.05678)	0.11526 (0.18721)	0.70915 (14.02734)
$\hat{\mu}_{\mathrm{RRZ2}}$	−0.00218 (0.05340)	0.12390 (0.20431)	0.71842 (14.20072)
$\hat{\mu}_{\mathrm{RRZ3}}$	−0.05630 (0.08911)	0.07574 (0.13743)	0.66755 (13.41960)
$\hat{\mu}_{\mathrm{RRZ4}}$	−0.13443 (0.21989)	0.00535 (0.09946)	0.59362 (12.42681)
$\hat{\mu}_{\mathrm{RRZ5}}$	−0.28176 (1.62035)	−0.11148 (0.50203)	0.46696 (8.84080)
$\hat{\mu}_{\mathrm{RRZ6}}$	−0.81081 (12.96187)	−0.55635 (5.90534)	−0.00813 (4.33351)
$\hat{\mu}_{\mathrm{EL1}}$	−0.00162 (0.05079)	−0.00676 (0.09762)	−0.15881 (4.46274)
$\hat{\mu}_{\mathrm{EL2}}$	0.00444 (0.05095)	0.00530 (0.09858)	−0.09791 (4.38688)
$\hat{\mu}_{\mathrm{EL3}}$	0.01359 (0.05179)	0.03433 (0.11069)	−0.00808 (4.33150)

†The propensity score is $P(D=1|X=x) = 1-\exp\{-\exp(1+0.5x)\}$ with $x \sim N(0,1)$. The logistic regression model (4.1) is chosen as a working propensity score. The estimates $\hat{\mu}_{\mathrm{RRZ1}}, \ldots, \hat{\mu}_{\mathrm{RRZ6}}$ correspond to $a(x) = x, 2+3x, x^2, 2+3x^2, x^4, 2+3x^4$, whereas $\hat{\mu}_{\mathrm{EL1}}$, $\hat{\mu}_{\mathrm{EL2}}$ and $\hat{\mu}_{\mathrm{EL3}}$ correspond to $a(x) = x, x^2, x^4$.

robust to the misspecification of the propensity score. By comparing the results in Tables 1 and 2, it appears that the misspecified propensity score produces better results than the correctly specified propensity score. This is because the missing proportion with the complementary log–log-propensity-score is smaller than that with the logistic propensity score.

In summary, our simulation study indicates that the empirical-likelihood-based estimator $\hat{\mu}_{\mathrm{EL}}$ is comparable with the estimator $\hat{\mu}_{\mathrm{RRZ}}$ of Robins *et al.* (1994) when the regression function $\mu(x) = E(Y|X=x)$ is correctly specified and is more efficient and more robust than $\hat{\mu}_{\mathrm{RRZ}}$ when $\mu(x)$ is misspecified.

Finally, we consider constructing a hybrid of bootstrap and empirical likelihood ratio confidence intervals for μ under the model

$$P(D=1|x) = \frac{\exp(0.5+0.5x)}{1+\exp(0.5+0.5x)}, \qquad x \sim N(0,1), \quad y = 2+3x^k+\varepsilon, \quad k = 1,2,4,$$

where $\varepsilon \sim N(0,1)$. With 200 bootstrap replications resampled from the triplets (y_i, x_i, d_i), $i = 1, \ldots, n$, we calculate bootstrap empirical likelihood ratio confidence intervals for μ, along with bootstrap percentile confidence intervals for μ by repeatedly calculating point estimators of μ. Since calculations of bootstrap empirical likelihood ratio confidence intervals may not have a solution for small sample sizes, we have used a large sample size $n = 1000$ to ensure that the constraint equations (2.3) plus an additional constraint equation

$$\sum_{i=1}^{n_1} p_i(y_i - \mu) = 0$$

have solutions for each simulation. Note that, in observational studies, sample sizes are usually large. On the basis of 1000 simulations and by using a linear constraint $a(x) = x$, summarized in Table 3 are the simulation results on coverage probabilities and average lengths of the bootstrap empirical likelihood ratio and bootstrap percentile confidence intervals for μ.

Table 3. Simulation results on coverage probabilities and average lengths for the bootstrap methods

Confidence interval	*Results for the following values of k:*					
	k = 1		*k = 2*		*k = 4*	
	Coverage probability	*Average length*	*Coverage probability*	*Average length*	*Coverage probability*	*Average length*
Bootstrap empirical likelihood ratio	0.952	0.40295	0.952	0.76577	0.944	6.91335
Bootstrap percentile	0.949	0.40494	0.928	0.62997	0.906	4.50205

It is seen that the empirical likelihood ratio confidence intervals have better coverage probabilities than the bootstrap percentile confidence intervals. Nevertheless, the average length of the empirical likelihood ratio confidence intervals is longer than that of the bootstrap percentile confidence intervals.

5. Concluding remarks

In this paper we have proposed an empirical-likelihood-based inference procedure for estimation of a response mean when the response might be missing at random. As with the semiparametric estimation procedure that was proposed by Robins *et al.* (1994), the empirical-likelihood-based estimation procedure also enjoys the double-robustness property, i.e. the empirical-likelihood-based estimator is consistent when either the missingness mechanism is correctly specified or the regression function is correctly specified. Moreover, if the true regression function is a linear combination of some known functions, but with unknown coefficients, the empirical-likelihood-based estimator can still achieve full efficiency by incorporating these known functions into the auxiliary information functions. The advantage of empirical-likelihood-based estimation is that the true coefficients in the linear combination are irrelevant. By comparing with a variety of competitors, our simulation results indicate that the empirical-likelihood-based estimation procedure and the regression estimation procedure are virtually equivalent if the true regression function is correctly specified. Moreover, the empirical-likelihood-based estimator is appreciably more efficient in general than other commonly used estimators, especially when the true regression function is misspecified. In addition, the empirical-likelihood-based estimator is very robust to the misspecification of the propensity score. The methodology that is presented here might be extended to many situations in practice such as applications in which covariates are missing at random.

Acknowledgements

We are grateful to the Joint Editor, an Associate Editor and two referees for their careful reading and for some helpful comments and suggestions that have greatly improved our original submission.

Appendix A: Proof of part (a) of theorem 3

Since $\hat{\beta}$ maximizes the binomial likelihood $L_B(\beta)$ in equation (2.1), it is seen that $\hat{\beta}$ is a solution to the following system of score equations:

$$U(\beta) = \frac{\partial[\log\{L_{\mathrm{B}}(\beta)\}]}{\partial\beta} = \sum_{i=1}^{n} \frac{\{d_i - w(x_i,\beta)\}\, v(x_i,\beta)}{w(x_i,\beta)\{1-w(x_i,\beta)\}}.$$

Under suitable regularity conditions, it can be shown that $\hat{\beta}$ is $\sqrt{n}$ consistent. Now expanding $U(\hat{\beta})$ at β_0 along with a standard argument gives

$$\hat{\beta} - \beta_0 = \frac{1}{n} S_{\mathrm{B}}^{-1}\, U(\beta_0) + o_p(n^{-1/2}). \tag{A.1}$$

With $\theta_0 = E\{w(x_1,\beta_0)\}$ estimated by

$$\hat{\theta} = n^{-1}\sum_{i=1}^{n} w(x_i,\hat{\beta}),$$

the asymptotic expression (A.1) yields

$$\begin{aligned}\hat{\theta} &= \frac{1}{n}\sum_{i=1}^{n} w(x_i,\hat{\beta}) \\ &= \frac{1}{n}\sum_{i=1}^{n} w(x_i,\beta_0) + \left\{\frac{1}{n}\sum_{i=1}^{n} v^{\mathrm{T}}(x_i,\beta_0)\right\}(\hat{\beta}-\beta_0) + O_p(n^{-1}) \\ &= \bar{w} + \frac{1}{n} V^{\mathrm{T}} S_{\mathrm{B}}^{-1} U(\beta_0) + o_p(n^{-1/2}),\end{aligned} \tag{A.2}$$

where $\bar{w} = n^{-1}\Sigma_{i=1}^{n} w(x_i,\beta_0)$ and $V = E\{v(x_1,\beta_0)\}$. Write

$$D_1 = E\left[\begin{pmatrix} \theta_0\, w^{-1}(x,\beta_0) v^{\mathrm{T}}(x,\beta_0) \\ -w^{-1}(x,\beta_0)\{a(x_1)-a_0\}\, v^{\mathrm{T}}(x,\beta_0)\end{pmatrix}\right],$$

$$D_2 = \begin{pmatrix} -1 \\ 0\end{pmatrix},$$

$$D_3 = \begin{pmatrix} 0 \\ -I_q\end{pmatrix},$$

$$Q = Q(0,\beta_0,\theta_0,a_0) = \sum_{i=1}^{n} d_i\, r(x_i,\beta_0,\theta_0,a_0).$$

Then expanding $Q(\hat{\pi},\hat{\beta},\hat{\theta},\hat{a})$ in equation (2.6) at $(0,\beta_0,\theta_0,a_0)$ gives

$$\begin{aligned}0 &= Q(\hat{\pi},\hat{\beta},\hat{\theta},\hat{a}) \\ &= Q(0,\beta_0,\theta_0,a_0) + \frac{\partial Q(0,\beta_0,\theta_0,a_0)}{\partial\pi^{\mathrm{T}}}\hat{\pi} + \frac{\partial Q(0,\beta_0,\theta_0,a_0)}{\partial\beta^{\mathrm{T}}}(\hat{\beta}-\beta_0) + \frac{\partial Q(0,\beta_0,\theta_0,a_0)}{\partial\theta}(\hat{\theta}-\theta_0) \\ &\quad + \frac{\partial Q(0,\beta_0,\theta_0,a_0)}{\partial a^{\mathrm{T}}}(\hat{a}-a_0) + O_p(1) \\ &= Q - nS\hat{\pi} + nD_1(\hat{\beta}-\beta_0) + nD_2(\hat{\theta}-\theta_0) + nD_3(\hat{a}-a_0) + o_p(n^{1/2}),\end{aligned}$$

which implies that

$$\begin{aligned}\hat{\pi} &= \frac{1}{n}S^{-1}Q + S^{-1}D_1(\hat{\beta}-\beta_0) + S^{-1}D_2(\hat{\theta}-\theta_0) + S^{-1}D_3(\hat{a}-a_0) + o_p(n^{-1/2}) \\ &= \frac{1}{n}S^{-1}Q + S^{-1}(D_1 + D_2V^{\mathrm{T}})(\hat{\beta}-\beta_0) + S^{-1}\{D_2(\bar{w}-\theta_0) + D_3(\hat{a}-a_0)\} + o_p(n^{-1/2}) \\ &= \frac{1}{n}S^{-1}\left\{Q + DS_{\mathrm{B}}^{-1}U(\beta_0) - n\begin{pmatrix}\bar{w}-\theta_0 \\ \hat{a}-a_0\end{pmatrix}\right\} + o_p(n^{-1/2}).\end{aligned} \tag{A.3}$$

On the basis of the asymptotic expansions of $\hat{\beta}$, $\hat{\theta}$ and $\hat{\pi}$ in equations (A.1), (A.2) and (A.3) and using the fact that $n/n_1 = \theta_0^{-1} + o_p(1)$, we have the following asymptotic expression of $\hat{\mu}_{\mathrm{EL}}$ in equation (2.7):

$$
\begin{aligned}
\hat{\mu}_{\mathrm{EL}} - \mu_0 &= \sum_{i=1}^{n_1} \hat{p}_i (y_i - \mu_0) \\
&= \frac{1}{n_1} \sum_{i=1}^{n} \frac{\hat{\theta}\, w^{-1}(x_i, \hat{\beta})}{1 + \hat{\pi}^{\mathrm{T}} r(x_i, \hat{\beta}, \hat{\theta}, \hat{a})} d_i (y_i - \mu_0) \\
&= \frac{1}{n_1} \sum_{i=1}^{n} \frac{\theta_0 d_i (y_i - \mu_0)}{w(x_i, \beta_0)} - \frac{1}{n_1} \sum_{i=1}^{n} \frac{\theta_0 d_i (y_i - \mu_0)}{w(x_i, \beta_0)} r^{\mathrm{T}}(x_i, \beta_0, \theta_0, a_0) \hat{\pi} - \frac{1}{n_1} \sum_{i=1}^{n} \frac{\theta_0 d_i (y_i - \mu_0)}{w^2(x_i, \beta_0)} v^{\mathrm{T}}(x_i, \beta_0)(\hat{\beta} - \beta_0) \\
&\quad + \frac{1}{n_1} \sum_{i=1}^{n} \frac{d_i (y_i - \mu_0)}{w(x_i, \beta_0)} (\hat{\theta} - \theta_0) + O_p(n^{-1}) \\
&= \frac{1}{n_1} \sum_{i=1}^{n} \frac{\theta_0 d_i (y_i - \mu_0)}{w(x_i, \beta_0)} - \frac{n}{n_1} \theta_0 C^{\mathrm{T}} \hat{\pi} - \frac{n}{n_1} \theta_0 H^{\mathrm{T}} (\hat{\beta} - \beta_0) + o_p(n^{-1/2}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{d_i (y_i - \mu_0)}{w(x_i, \beta_0)} - \frac{1}{n} C^{\mathrm{T}} S^{-1} Q - \frac{1}{n} K\, U(\beta_0) + C^{\mathrm{T}} S^{-1} \begin{pmatrix} \bar{w} - \theta_0 \\ \hat{a} - a_0 \end{pmatrix} + o_p(n^{-1/2}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{d_i (y_i - \mu_0)}{w(x_i, \beta_0)} - \{ d_i - w(x_i, \beta_0) \}\, s(x_i, \beta_0, \theta_0, a_0) \right] + o_p(n^{-1/2}) \\
&= \frac{1}{n} \sum_{i=1}^{n} (Z_{1i} - Z_{2i}) + o_p(n^{-1/2}),
\end{aligned}
\tag{A.4}
$$

where

$$
Z_{1i} = \frac{d_i (y_i - \mu_0)}{w(x_i, \beta_0)}, \qquad Z_{2i} = \{ d_i - w(x_i, \beta_0) \}\, s(x_i, \beta_0, \theta_0, a_0), \qquad i = 1, \ldots, n. \tag{A.5}
$$

The asymptotic expansion in equation (A.4) establishes equation (2.11). It is seen from expression (A.5) that $E(Z_{11}) = E(Z_{21}) = 0$. Furthermore, it can be shown after some algebra that

$$
\begin{aligned}
\mathrm{var}(Z_1) &= E\left\{ \frac{\mu_2(x_1) - 2\mu_0\, \mu(x_1) + \mu_0^2}{w(x_1, \beta_0)} \right\}, \\
\mathrm{var}(Z_2) &= E[w(x_1, \beta_0)\{1 - w(x_1, \beta_0)\}\, s^2(x_1, \beta_0, \theta_0, a_0)], \\
\mathrm{cov}(Z_1, Z_2) &= E[\{\mu(x_1) - \mu_0\}\{1 - w(x_1, \beta_0)\}\, s(x_1, \beta_0, \theta_0, a_0)].
\end{aligned}
\tag{A.6}
$$

It now follows from equations (A.4) and (A.6) and the central limit theorem for sample means that $n^{1/2}(\hat{\mu} - \mu_0) \to N(0, \sigma^2)$ in distribution, where σ^2 is shown to be, after some algebra,

$$
\begin{aligned}
\sigma^2 &= \mathrm{var}(Z_{11} - Z_{21}) \\
&= \mathrm{var}(Z_{11}) + \mathrm{var}(Z_{21}) - 2\, \mathrm{cov}(Z_{11}, Z_{21}) \\
&= E\left\{ \frac{\mu_2(x_1) - \mu^2(x_1)}{w(x_1, \beta_0)} \right\} + \mathrm{var}\{\mu(x_1)\} + E\left[\frac{1 - w(x_1, \beta_0)}{w(x_1, \beta_0)} \{ \mu(x_1) - \mu_0 - w(x_1, \beta_0)\, s(x_1, \beta_0, \theta_0, a_0) \}^2 \right].
\end{aligned}
$$

The proof of part (a) of theorem 3 is complete.

References

Breslow, N. (2003) Are statistical contributions to medicine undervalued? *Biometrics*, **59**, 1–8.

Chen, J. and Qin, J. (1993) Empirical likelihood method in finite population and the effective usage of auxiliary information. *Biometrika*, **80**, 107–116.

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

Cochran, W. C. (1977) *Sampling Techniques*. New York: Wiley.

Gilbert, P. B. (2000) Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Ann. Statist.*, **28**, 151–194.

Gilbert, P. B., Lele, S. R. and Vardi, Y. (1999) Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika*, **86**, 27–43.
Gill, R. D., Vardi, Y. and Wellner, J. A. (1988) Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.*, **16**, 1069–1112.
Haberman, S. J. (1984) Adjustment by minimum discriminant information. *Ann. Statist.*, **12**, 971–988.
Hahn, J. (1998) On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, **66**, 315–331.
Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1998) Characterizing selection bias using experimental data. *Econometrica*, **66**, 1017–1098.
Hirano, K., Imbens, G. W. and Ridder, G. (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, **71**, 1161–1189.
Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Ass.*, **47**, 663–685.
Imbens, G. (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev. Econ. Statist.*, **86**, 4–29.
Imbens, G. and Lancaster, T. (1994) Combining micro and macro data in microeconometric model. *Rev. Econ. Stud.*, **61**, 655–680.
Kitamura, Y. (1997) Empirical likelihood methods with weakly dependent processes. *Ann. Statist.*, **25**, 2084–2102.
Korn, E. L. and Baumrind, S. (1998) Clinician preferences and the estimation of causal treatment differences. *Statist. Sci.*, **13**, 209–235.
van der Laan, M. J. and Robins, J. M. (2003) *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.
Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*, 2nd edn. New York: Wiley.
Owen, A. B. (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.
Owen, A. B. (1990) Empirical likelihood ratio confidence regions. *Ann. Statist.*, **18**, 90–120.
Owen, A. B. (1991) Empirical likelihood for linear models. *Ann. Statist.*, **19**, 1725–1747.
Owen, A. B. (2001) *Empirical Likelihood*. Boca Raton: Chapman and Hall–CRC.
Qin, J. (1993) Empirical likelihood in biased sample problems. *Ann. Statist.*, **21**, 1182–1196.
Qin, J. and Lawless, J. F. (1994) Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, 300–325.
Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994) Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Ass.*, **89**, 846–886.
Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Ass.*, **90**, 106–121.
Rosenbaum, P. R. (2002) *Observational Studies*. New York: Springer.
Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.
Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
Rubin, D. B. (1990) Comments on "On the application of probability theory to agricultural experiments. Essay on principles. Section 9": Neyman (1923) and causal inference in experiments and observational studies. *Statist. Sci.*, **5**, 472–480.
Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. New York: Wiley.
Tan, Z. (2006) A distributional approach for causal inference using propensity scores. *J. Am. Statist. Ass.*, to be published.
Tripathi, G. and Kitamura, Y. (2003) Testing conditional moment restrictions. *Ann. Statist.*, **31**, 2059–2095.
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, Q. H. and Rao, J. N. K. (2002) Empirical likelihood-based inference under imputation for missing response data. *Ann. Statist.*, **30**, 896–924.
White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.