Ann. Inst. Statist. Math.
Vol. 56, No. 3, 403–414 (2004)
©2004 The Institute of Statistical Mathematics

LIKELIHOOD-BASED IMPUTATION INFERENCE FOR MEAN FUNCTIONALS IN THE PRESENCE OF MISSING RESPONSES*

QI-HUA WANG

Academy of Mathematics and System Science, Chinese Academy of Science, Beijing 100080, China and Heilongjiang University, Harbin 150008, China, e-mail: qhwang@amss.ac.cn

(Received February 25, 2003; revised September 2, 2003)

Abstract. This paper considers a semiparametric model which parameterizes only the conditional density of a response given covariates and allows the marginal distribution of the covariates to be completely arbitrary when responses are missing. Different estimators with asymptotic normality for the mean of the response variable are derived, respectively, in the two cases where auxiliary information is available or not. The resulting asymptotic behaviors show that the use of auxiliary information improves inference via empirical likelihood approach.

Key words and phrases: Asymptotic efficiency, missing response, resampling imputation.

1. Introduction

In the study of the association between a response variable Y and various covariates and the inference on the mean of a response variate, some responses may be missing for various reasons such as unwillingness of some sampled units to supply the desired information, loss of information caused by uncontrollable factors, failure on the part of investigator to gather correct information, and so forth. In fact, missingness of responses is common in opinion polls, market research surveys, mail enquires, socio-economic investigations, medical studies and other scientific experiments.

Let X be a d-dimensional vector of factors and Y be a response variable influenced by X. In practice, one often obtains a random sample of incomplete data

$$(X_i, Y_i, \delta_i), \qquad i = 1, 2, \dots, n, \tag{1.1}$$

where all the X_i's are observed and $\delta_i = 0$ if Y_i is missing, otherwise $\delta_i = 1$. In such circumstances, most statistical packages will drop observations with missing response data from the analysis. This can result in a much reduced effective sample size, serious loss of efficiency and considerably biased estimator when a substantial proportion of observations is missing. An alternative strategy is to impute the missing responses with predicted values from a regression of the missing response on the available covariable and then apply standard methods to the complete data set as if they were true observations. Cheng (1994) explored the usefulness of a basic nonparametric estimation scheme for the missing data (1.1) by focusing on the case where some Y may be missing at random

*This research was supported by National Natural Science Foundation of China (Key grant 10231030, Special grant 10241001).

(MAR). In Cheng's paper, he applied kernel regression imputation to estimate the mean of Y, and established the asymptotic normality of a trimmed version of this estimator. Recently, Wang and Rao (2002*a*) used kernel regression imputation to develop empirical likelihood based inference for the mean of Y. Wang and Rao (2001, 2002*b*) considered linear regression models and used linear regression imputation to develop empirical likelihood based inference for the mean of Y and the regression coefficients, respectively. It should be pointed out that use of regression imputation to treat missing responses has in recently years attracted interest. We refer the readers to Little and Rubin (1987) for an excellent account of the regression imputation methods for missing responses.

In this paper, we consider a semiparametric model which parameterizes only the relation of interest, $f(y \mid x,\theta)$, and allows the marginal distribution $G(\cdot)$ of X to be completely arbitrary. Throughout this paper, we make MAR assumption. The MAR assumption implies that δ and Y are conditionally independent given X (see, e.g., Little and Rubin (1987)). That is, $P(\delta=1 \mid Y,X)=P(\delta=1 \mid X)$. We are interested in inferences on the mean of the response Y. First, we use the observed data to estimate θ by likelihood and semiparametric likelihood approaches, respectively, in the two cases where auxiliary information is available or not. Then, based on the estimators of θ, we impute the missing responses with predicted values and define four semiparametric estimators of μ respectively. It should be noted that the method of empirical likelihood developed by Owen (1988, 1990) provides a means of determining nonparametric confidence regions for statistical functionals. However, this paper just uses it to develop sharper inferences when some auxiliary information is available.

The rest of this paper is organized as follows. In Section 2, we define a likelihood based imputation estimator and a weighted one by incorporating auxiliary information via empirical likelihood technique. In Section 3, we define a semiparametric empirical likelihood based imputation estimator and a weighted one by incorporating a different form of auxiliary information via a semiparametric empirical likelihood technique.

2. Likelihood based imputation estimators

Let the complete data $(X_1,Y_1),\ldots,(X_n,Y_n)$ be randomly drawn from the semiparametric population $f(y \mid x,\theta)dG(x)$. Let δ_i be the indicator whether or not Y_i was observed for $i=1,2,\ldots,n$. Under MAR assumption, one obtains a sample of incomplete data (X_i,Y_i,δ_i), $i=1,2,\ldots,n$. Based on the incomplete data set, in this section, we first define a resampling parametric imputation estimator for θ and then construct two semiparametric estimators for the mean of the response using regression imputation.

Based on the observed data $\{(X_i,Y_i,\delta_i)\}_{i=1}^n$, the likelihood function for θ is

$$\prod_{i=1}^n f^{\delta_i}(Y_i \mid X_i,\theta).$$

Let $\widetilde{\theta}_n$ be the MLE of θ satisfying

$$\sum_{i=1}^n \delta_i \frac{\partial \log f(Y_i \mid X_i,\theta)}{\partial\theta}=0. \tag{2.1}$$

2.1 *Estimation for μ under regression imputation*

Let $m(X_i,\theta)=\int yf(y \mid X_i,\theta)dy$ and $\widetilde{Y}_{i,n}=m(X_i,\widetilde{\theta}_n)$. Note that $E(m(X,\theta))=\mu$. Hence, we impute the missing Y_i by $\widetilde{Y}_{i,n}$ and construct the following semiparametric estimator of μ

$$\widehat{\mu}_{n,1} = \frac{1}{n}\sum_{i=1}^{n}(\delta_i Y_i + (1-\delta_i)\widetilde{Y}_{i,n}).$$

THEOREM 2.1. *Under Conditions* (C.Y) *and* (C.f), *we have*

$$\sqrt{n}(\widehat{\mu}_{n,1} - \mu) \xrightarrow{\mathcal{L}} N(0, \sigma_1^2)$$

where $\sigma_1^2 = S_2^{\mathrm{T}} S_1^{-1} S_2 + 2S_2^{\mathrm{T}} S_1^{-1} S_3 + S_4 + S_5$ *with*

$$S_1 = E\left[\delta \frac{\partial \log f(Y \mid X, \theta)}{\partial \theta} \frac{\partial \log f(Y \mid X, \theta)}{\partial \theta^T}\right],$$
$$S_2 = E\left[(1-\delta) Y \frac{\partial \log f(Y \mid X, \theta)}{\partial \theta}\right],$$
$$S_3 = E\left[\delta Y \frac{\partial \log f(Y \mid X, \theta)}{\partial \theta}\right],$$
$$S_4 = E[\delta(Y - E[Y \mid X])^2], \qquad S_5 = \mathrm{Var}(E[Y \mid X]).$$

Let

$$\widehat{S}_{n1} = \frac{1}{n}\sum_{i=1}^{n}\delta_i \left(\frac{\partial \log f(Y_i \mid X_i, \theta)}{\partial \theta}\bigg|_{\theta=\widetilde{\theta}_n}\right)\left(\frac{\partial \log f(Y_i \mid X_i, \theta)}{\partial \theta^{\mathrm{T}}}\bigg|_{\theta=\widetilde{\theta}_n}\right),$$
$$\widehat{S}_{n2} = \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i) Y_i \left(\frac{\partial \log f(Y_i \mid X_i, \theta)}{\partial \theta}\bigg|_{\theta=\widetilde{\theta}_n}\right),$$
$$\widehat{S}_{n3} = \frac{1}{n}\sum_{i=1}^{n}\delta_i Y_i \left(\frac{\partial \log f(Y_i \mid X_i, \theta)}{\partial \theta}\bigg|_{\theta=\widetilde{\theta}_n}\right),$$
$$\widehat{S}_{n4} = \frac{1}{n}\sum_{i=1}^{n}\delta_i \left(Y_i - \int y f(y \mid X_i, \widetilde{\theta}_n)\right)^2,$$
$$\widehat{S}_{n5} = \frac{1}{n}\sum_{i=1}^{n}\left(\int y f(y \mid X_i, \widetilde{\theta}_n) dy\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\int y f(y \mid X_i, \widetilde{\theta}_n)\right)^2.$$

Then, the asymptotic variance σ_1^2 can be estimated consistently by

$$\widehat{\sigma}_{1n}^2 = \widehat{S}_{n2}\widehat{S}_{n1}^{-1}\widehat{S}_{n2} + 2\widehat{S}_{n2}\widehat{S}_{n1}^{-1}\widehat{S}_{n3} + \widehat{S}_{n4} + \widehat{S}_{n5}.$$

Most standard statistical software will drop incomplete cases from the analysis. The resulting 'complete case' estimator is then defined as $\widetilde{\mu}_{n1} = \sum_{i=1}^{n}\delta_i Y_i / \sum_{i=1}^{n}\delta_i$. However, $\widetilde{\mu}_{n1}$ is asymptotically biased since δ can depend on X under MAR assumption.

2.2 *Weighting estimation for μ with auxiliary information*

We assume that auxiliary information on X of the form $EA(X) = 0$ is available, where $A(\cdot) = (A_1(\cdot), \ldots, A_r(\cdot))^{\mathrm{T}}$, $r \geq 1$, is a known vector (or scalar) function; for example, when the mean or median of X is known in the scalar X case.

To use the auxiliary information, we first maximize

$$\prod_{i=1}^{n} p_i \tag{2.2}$$

subject to $\sum_{i=1}^{n} p_i = 1$, $\sum_{i=1}^{n} p_i A(X_i) = 0$. Provided that the origin is inside the convex hull of $A(X_1), \ldots, A(X_n)$, by the method of Lagrange multipliers, we get

$$p_i = \frac{1}{n} \frac{1}{1 + \zeta_n^{\top} A(X_i)},$$

where ζ_n is the solution of the following equation:

$$\frac{1}{n} \sum_{i=1}^{n} \frac{A(X_i)}{1 + \zeta_n^{\top} A(X_i)} = 0. \tag{2.3}$$

An empirical likelihood-based weighted estimator of μ is then defined by

$$\widehat{\mu}_{n,2} = \sum_{i=1}^{n} p_i (\delta_i Y_i + (1 - \delta_i) \widetilde{Y}_{i,n}). \tag{2.4}$$

THEOREM 2.2. *Under Assumptions* (C.Y), (C.f) *and* (C.A), *we have*

$$\sqrt{n}(\widehat{\mu}_{n,2} - \mu) \xrightarrow{\mathcal{L}} N(0, \sigma_2^2)$$

where $\sigma_2^2 = \sigma_1^2 - A_1^{\top} A_2^{-1} A_1$ *with* $A_1 = E[(E[Y \mid X] - \mu) A(X)]$ *and* $A_2 = EA(X)A^{\top}(X)$.

Clearly, $\widehat{\mu}_{n,2}$ has smaller asymptoic variance than $\widehat{\mu}_{n,1}$. This shows that the use of auxiliary information on X of the form $EA(X) = 0$ by the weighting method improves inference.

σ_2^2 can be estimated consistently by $\widehat{\sigma}_{2n}^2 = \widehat{\sigma}_{1n}^2 - A_{n1}^{\top} A_{n2}^{-1} A_{n1}$, where $\widehat{\sigma}_{1n}^2$ is defined in Subsection 2.1 and

$$A_{n1} = \frac{1}{n} \sum_{i=1}^{n} \left(\int y f(y \mid X_i, \widetilde{\theta}_n) dy - \widehat{\mu}_{n,2} \right) A(X_i)$$

and

$$A_{n2} = \frac{1}{n} \sum_{i=1}^{n} A(X_i) A^{\top}(X_i).$$

As pointed out by a referee that Hellerstein and Imbens (1999) also uses empirical likelihood weights in a regression context, and the Horvitz-Thompson estimator deals with the missing data by weighting the complete observations, with weights equal to the estimated propensity score. That is, the estimator is defined by

$$\widehat{\mu}_{HT} = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i Y_i}{\widehat{\Delta}_n(X_i)},$$

where $\widehat{\Delta}_n(\cdot)$ is an estimator of $\Delta(x) = P(\delta = 1 \mid X = x)$. Clearly, $\widehat{\mu}_{HT}$ is a nonparametric estimator and does not use the assumed parametric structure and the auxiliary information. This implies that the estimator may not use information sufficiently. It can be shown that $\widehat{\mu}_{HT}$ is asymptotically normal with asymptotic variance

$$\sigma_{HT}^2 = E\left[\frac{\sigma^2(X)}{\triangle(X)}\right] + \mathrm{Var}(m(X,\theta)),$$

when $\widehat{\triangle}_n(x)$ is defined by kernel smoothing the participation indicator against covariate values, where $\sigma^2(x) = E[(Y - E[Y \mid X])^2 \mid X = x]$. It seems not easy to compare $\widehat{\mu}_{HT}$ with $\widehat{\mu}_{n,1}$ and $\widehat{\mu}_{n,2}$ by comparing their asymptotic variances. We will compare them by simulation.

It should be pointed out that $\widehat{\theta}_{HT}$ has a disadvantages, requiring a high dimension smoothing technique to compute the propensity score when the propensity score is unknown completely. That is, $\widehat{\theta}_{HT}$ has the so called "curse of dimension" problem.

3. Semiparametric likelihood based imputation estimators with auxiliary information

In some cases, auxiliary information with the form $E\psi(X,\theta) = 0$ is available. For example, this kind of information arises naturally in microeconometric models (Imbens and Lancaster (1994)). To use the auxiliary information, we maximize the following semiparametric likelihood with data (X_i, Y_i, δ_i), $i = 1, 2, \ldots, n$,

$$L_0(\theta) \equiv \prod_{i=1}^{n} (f(Y_i \mid X_i;\theta)dG(X_i))^{\delta_i}(dG(X_i))^{1-\delta_i} = \prod_{i=1}^{n}\widetilde{p}_i \prod_{i=1}^{n} f^{\delta_i}(Y_i \mid X_i,\theta)$$

subject to $\sum_{i=1}^{n}\widetilde{p}_i = 1$ and $\sum_{i=1}^{n}\widetilde{p}_i\psi(X_i,\theta) = 0$, where $\widetilde{p}_i = dG(X_i)$.

If 0 is in the convex of $\psi(X_1,\theta), \ldots, \psi(X_n,\theta)$, by Lagrange multiplier, we have

$$\widetilde{p}_i = \frac{1}{n(1 + \lambda_n^{\mathsf{T}}\psi(X_i,\theta))},$$

where λ_n satisfies

$$\sum_{i=1}^{n} \frac{\psi(X_i,\theta)}{1 + \lambda_n^{\mathsf{T}}\psi(X_i,\theta)} = 0. \tag{3.1}$$

Clearly,

$$\log L_0(\theta) = -\sum_{i=1}^{n}\log(1 + \lambda_n^{\mathsf{T}}\psi(X_i,\theta)) + \sum_{i=1}^{n}\delta_i \log f(Y_i \mid X_i,\theta) - n\log n. \tag{3.2}$$

Let $\widetilde{\theta}_{n,AU}$ be the MLE satisfying

$$\frac{\partial \log L_0(\theta)}{\partial\theta} = 0.$$

3.1 *Estimation for μ under regression imputation*

Let $\widetilde{Y}_{i,AU}$ be $\widetilde{Y}_i$ with $\widetilde{\theta}_n$ replaced by $\widetilde{\theta}_{n,AU}$. We impute the missing Y_i by $\widetilde{Y}_{i,AU}$ and define the estimator of μ to be

$$\widehat{\mu}_{n,3} = \frac{1}{n}\sum_{i=1}^{n}(\delta_i Y_i + (1-\delta_i)\widetilde{Y}_{i,AU}).$$

THEOREM 3.1. *Under Assumptions* (C.Y), (C.f) *and* (C.ψ), *we have*

$$\sqrt{n}(\widehat{\mu}_{n,3}-\mu)\xrightarrow{\mathcal{L}} N(0,\sigma_3^2),$$

where $\sigma_3^2=S_2^{\mathrm{T}}(S_1+S_6)^{-1}S_2+2S_2^{\mathrm{T}}(S_1+S_6)^{-1}S_3+S_4+S_5$ *with* S_i *defined in Theorem* 2.1 *for* $i=1,2,3,4,5$ *and*

$$S_6=E\left(\frac{\partial\psi(X,\theta)}{\partial\theta^{\mathrm{T}}}\right)(E\psi(X,\theta)\psi^{\mathrm{T}}(X,\theta))^{-1}E\left(\frac{\partial\psi(X,\theta)}{\partial\theta}\right).$$

σ_3^2 can be estimated consistently by $\widehat{\sigma}_{3n}^2=S_{2n}^{\mathrm{T}}(S_{1n}+S_{6n})^{-1}S_{2n}+2S_{2n}^{\mathrm{T}}(S_{1n}+S_{6n})^{-1}S_{3n}+S_{4n}+S_{5n}$, where S_{in} are $\widehat{S}_{in}$ defined in Section 2 with θ_n replaced by $\widetilde{\theta}_{n,AU}$ for $i=1,2,3,4,5$ and $S_{6n}=\Gamma_{n1}^{\mathrm{T}}\Gamma_{n2}^{-1}\Gamma_{n1}$ with

$$\Gamma_{n1}=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\psi(X_i,\theta)}{\partial\theta}\bigg|_{\theta=\widetilde{\theta}_{n,AU}}$$

and

$$\Gamma_{n2}=\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\widetilde{\theta}_{n,AU})\psi^{\mathrm{T}}(X_i,\widetilde{\theta}_{n,AU}).$$

3.2 *Weighting estimation for* μ *under regression imputation*

It is easy to see that $\widehat{\mu}_{n,3}$ has less asymptotic variance than $\widehat{\mu}_{n,1}$ and hence is asymptotically more efficient. This shows that the use of this form $E\psi(X,\theta)=0$ of auxiliary information also improves inference. However, this improved estimator uses the auxiliary information only by $\widetilde{\theta}_{n,AU}$, and hence it does not use the information sufficiently. The following weighted estimator provides further improved inference.

Let

$$\widehat{p}_i=\frac{1}{n(1+\widehat{\lambda}_n^{\mathrm{T}}\psi(X_i,\widetilde{\theta}_{n,AU}))},$$

where $\widehat{\lambda}_n$ satisfies

$$\sum_{i=1}^{n}\frac{\psi(X_i,\widetilde{\theta}_{n,AU})}{1+\widehat{\lambda}_n^{\mathrm{T}}\psi(X_i,\widetilde{\theta}_{n,AU})}=0.$$

We can define a weighted semiparametric empirical likelihood based imputation estimator as follows

$$\widehat{\mu}_{n,4}=\sum_{i=1}^{n}\widehat{p}_i(\delta_iY_i+(1-\delta_i)\widetilde{Y}_{i,AU}).$$

This estimator uses the auxiliary information not only by $\widetilde{\theta}_{n,AU}$ but also by the weights. Hence, it provides further improved inference. This can be seen by the following Theorem 3.2.

THEOREM 3.2. *Under the assumptions of Theorem* 3.1, *we have*

$$\sqrt{n}(\widehat{\mu}_{n,4}-\mu)\xrightarrow{\mathcal{L}} N(0,\sigma_4^2),$$

where

$$\sigma_4^2 = \sigma_3^2 - B_1^{\mathrm{T}} B_2^{-1} B_1,$$

σ_3^2 *is as defined in Theorem* 3.1 *and*

$$B_1 = E[(E[Y \mid X] - \mu)\psi(X,\theta)], \qquad B_2 = E[(\psi(X,\theta))(\psi^{\mathrm{T}}(X,\theta))].$$

It is clear that $\widehat{\mu}_{n,4}$ has smaller asymptotic variance than $\widehat{\mu}_{n,3}$ and hence $\widehat{\mu}_{n,1}$. And the asymptotic variance can be estimated consistently by $\widehat{\sigma}_{n4}^2 = \widehat{\sigma}_{n3}^2 - \widehat{B}_n^{\mathrm{T}} \Gamma_{n2}^{-1} \widehat{B}_n$ with $\widehat{\sigma}_{3n}^2$ and Γ_{n2} defined before and

$$\widehat{B}_n = \frac{1}{n} \sum_{i=1}^{n} \left[\left(\int y f(y \mid X_i, \widetilde{\theta}_{n,AU}) - \widehat{\mu}_{n,4} \right) \psi(X_i, \widetilde{\theta}_{n,AU}) \right].$$

4. Simulation

From the results derived in Sections 2 and 3, $\widehat{\mu}_{n,2}$, $\widehat{\mu}_{n,3}$ and hence $\widehat{\mu}_{n,4}$ have smaller asymptotic variances than $\widehat{\mu}_{n,1}$. In this section, we compare these estimators with the Horvitz-Thompson estimator $\widehat{\mu}_{HT}$ in terms of their biases and standard errors (SE) via small sample simulation study.

The simulation used the two models which were used in Imbens and Lancaster (1994):

Model 1:

$$f(y \mid x, \theta) = (\sqrt{2\pi})^{-1} \exp\{-(y - \theta_1 - \theta_2 x)^2/2\}, \qquad (\theta_1, \theta_2) = (1.0, 0.5).$$

Model 2:

$$f(y \mid x, \theta) = \exp(-\theta_1 - \theta_2 x) \exp\{-y \exp(-\theta_1 - \theta_2 x)\}, \qquad y > 0, \qquad (\theta_1, \theta_2) = (1.0, 1.0).$$

Based on the above two models, X was simulated from normal $N(1,1)$, respectively. Based on each of the two models, we considered the following three response probability functions $P(x) = P(\delta = 1 \mid X = x)$ under the MAR assumption.

Case 1: $P(\delta = 1 \mid X = x) = 0.8 + 0.2|x - 1|$ if $|x - 1| \le 1$, and $= 0.95$ elsewhere.

Case 2: $P(\delta = 1 \mid X = x) = 0.9 - 0.2|x - 1|$ if $|x - 1| \le 4$, and $= 0.10$ elsewhere.

Case 3: $P(\delta = 1 \mid X = x) = 0.6$ for all x.

For the above three cases, the mean response rates are $EP_1(X) \approx 0.90$, $EP_2(X) \approx 0.74$ and $EP_3(X) = 0.60$, where $P_1(x)$, $P_2(x)$ and $P_3(x)$ are the response probability functions for Cases 1, 2 and 3, respectively. For each model, in the above three cases, we generated, respectively, 5000 Monte Carlo random samples of size $n = 30$, 60 and 120. From the 5000 simulated values of $\widehat{\mu}_{n,i}$ for $i = 1,2,3,4$, the bias and SE of these estimators were calculated. We assume that the mean μ_x of X is known and $A(X) = X - \mu_x$ when we use auxiliary information on X of the form $EA(X) = 0$. When the auxiliary information of the form $E\psi(X,\theta) = 0$ were used, we took function ψ's as

$$\psi_1 = \frac{\theta_1 + \theta_2 X}{\theta_1 + \theta_2 \mu_x} - 1, \qquad \psi_2 = \exp(\theta_1) \left\{ \exp(\theta_2 X) - \exp\left(\mu_x \theta_2 + \frac{\theta_2^2}{2} \right) \right\},$$

Table 1. Biases and standard errors (SE) of $\widehat{\mu}_{n,i}$ for $i = 1, 2, 3, 4$ under Model 1 with different missing functions $P(x)$ and different sample sizes n.

		Bias			SE		
n	Estimators	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$
	$\widehat{\mu}_{n,1}$	0.0134	0.0150	−0.0171	0.2163	0.2513	0.2847
	$\widehat{\mu}_{n,2}$	−0.0149	−0.0157	−0.0194	0.2027	0.2308	0.2522
30	$\widehat{\mu}_{n,3}$	0.0152	−0.0154	0.0182	0.2033	0.2272	0.2558
	$\widehat{\mu}_{n,4}$	−0.0146	0.0149	0.0187	0.1947	0.2189	0.2374
	$\widehat{\mu}_{HT}$	0.0187	0.0192	0.0191	0.2353	0.2675	0.3251
	$\widehat{\mu}_{n,1}$	0.0115	0.0129	0.0156	0.1634	0.1825	0.2084
	$\widehat{\mu}_{n,2}$	−0.0133	0.0121	−0.0162	0.1592	0.1747	0.1891
60	$\widehat{\mu}_{n,3}$	−0.0127	−0.0134	0.0158	0.1537	0.1752	0.1868
	$\widehat{\mu}_{n,4}$	0.0131	0.0137	0.0169	0.1325	0.1548	0.1752
	$\widehat{\mu}_{HT}$	0.0121	0.0132	0.0179	0.1735	0.1997	0.2312
	$\widehat{\mu}_{n,1}$	−0.0065	0.0094	−0.0087	0.1141	0.1306	0.1504
	$\widehat{\mu}_{n,2}$	0.0076	0.0082	0.0093	0.0982	0.1195	0.1352
120	$\widehat{\mu}_{n,3}$	−0.0070	0.0089	0.0095	0.1037	0.1203	0.1363
	$\widehat{\mu}_{n,4}$	0.0081	−0.0095	−0.0102	0.0758	0.0946	0.1129
	$\widehat{\mu}_{HT}$	−0.0055	−0.0076	0.0093	0.1198	0.1456	0.1712

Table 2. Biases and standard errors (SE) of $\widehat{\mu}_{n,i}$ for $i = 1, 2, 3, 4$ under Model 2 with different missing functions $P(x)$ and different sample sizes n.

		Bias			SE		
n	Estimators	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$
	$\widehat{\mu}_{n,1}$	−0.0442	−0.0458	−0.0481	0.5429	0.5697	0.5947
	$\widehat{\mu}_{n,2}$	0.0439	−0.0462	−0.0477	0.5233	0.5369	0.5538
30	$\widehat{\mu}_{n,3}$	−0.0458	0.0454	0.0484	0.5217	0.5359	0.5531
	$\widehat{\mu}_{n,4}$	0.0451	−0.0445	0.0491	0.4896	0.5093	0.5369
	$\widehat{\mu}_{HT}$	0.0455	0.0468	0.0501	0.5735	0.6100	0.6537
	$\widehat{\mu}_{n,1}$	−0.0302	0.0316	−0.0339	0.3783	0.3933	0.4112
	$\widehat{\mu}_{n,2}$	−0.0317	−0.0324	−0.0355	0.3591	0.3761	0.3907
60	$\widehat{\mu}_{n,3}$	−0.0310	0.0329	−0.0342	0.3576	0.3758	0.3874
	$\widehat{\mu}_{n,4}$	0.0325	−0.0332	0.0358	0.3249	0.3540	0.3795
	$\widehat{\mu}_{HT}$	−0.0324	−0.0309	0.0342	0.3928	0.4241	0.4578
	$\widehat{\mu}_{n,1}$	−0.0157	−0.0179	−0.0177	0.2676	0.2976	0.3201
	$\widehat{\mu}_{n,2}$	0.0164	−0.0172	0.0185	0.2465	0.2864	0.2993
120	$\widehat{\mu}_{n,3}$	−0.0152	0.0182	−0.0187	0.2496	0.2896	0.2935
	$\widehat{\mu}_{n,4}$	0.0168	0.0188	0.0192	0.2367	0.2667	0.2707
	$\widehat{\mu}_{HT}$	0.0159	−0.0174	0.0182	0.2711	0.3198	0.3475

for the above two models, respectively, and assume that the mean μ_x of X is known. To calculate $\widehat{\mu}_{HT}$, the propensity score estimator $\widehat{\triangle}_n(x)$ was taken to be the nonparametric kernel estimator given by

$$\widehat{\triangle}_n(x) = \frac{\sum_{i=1}^n \delta_i K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}$$

where $K(u) = -\frac{15}{8}u^2 + \frac{9}{8}$ if $|u| \le 1$, 0 otherwise, and $h_n = n^{-1/3}$.

We reported the estimated biases and SE for $\widehat{\mu}_{n,i}$, $i = 1,2,3,4$ and $\widehat{\mu}_{HT}$ in Tables 1 and 2.

From Tables 1 and 2, we observe that $\widehat{\mu}_{n,2}$, $\widehat{\mu}_{n,3}$ and $\widehat{\mu}_{n,4}$ have smaller SE than $\widehat{\mu}_{n,1}$ and $\widehat{\mu}_{HT}$, and $\widehat{\mu}_{n,1}$ behaves better than $\widehat{\mu}_{HT}$ in terms of the bias and SE. It is also noted that $\widehat{\mu}_{n,4}$ performs better than $\widehat{\mu}_{n,3}$ in terms of SE. Clearly, SE and bias decrease when missing rate decreases (response probability increases) or when the sample size increases.

Appendix: Assumptions and proofs of theorems

The following assumptions are needed to prove our theorems.

(C.Y) $EY^2 < \infty$.

(C.A) $EA(X)A^{\top}(X)$ is a positive definite matrix and $E\|A(X)\|^2 < \infty$.

(C.ψ) i) $E\psi(X,\theta)\psi^{\top}(X,\theta)$ is a positive matrix and $E\|\psi(X,\theta)\|^2 < \infty$.

ii) The second absolute moment of every component of $\frac{\partial\psi(X,\theta)}{\partial\theta}$ is finite.

(C.f) $f(x,\theta)$ satisfies the regular conditions given in Theorem 2.3 of Lehmann ((1983), Chapter 6) on the asymptotic normality of the maximum likelihood estimator in fully parametric model.

Proof of Theorem 2.1. By Taylor's expansion and some standard arguments, it can be used

$$\text{(A.1)} \qquad \widetilde{\theta}_n - \theta = S_1^{-1}\left(\frac{1}{n}\sum_{i=1}^n \delta_i \frac{\partial \log f(Y_i \mid X_i;\theta)}{\partial\theta}\right) + o_p(n^{-1/2})$$

where S_1 is defined in Theorem 2.1.

By two terms Taylor's expansion and (A.1), we get

$$\begin{aligned}\text{(A.2)} \quad \widehat{\mu}_{n,1} - \mu &= \frac{1}{n}\sum_{i=1}^n(\delta_i Y_i - \delta_i E[Y_i \mid X_i]) + \frac{1}{n}\sum_{i=1}^n(E[Y_i \mid X_i] - \mu) \\ &\quad + \frac{1}{n}\sum_{i=1}^n(1-\delta_i)\int y(f(y \mid X_i,\widetilde{\theta}_n) - f(y \mid X_i,\theta))dy \\ &= \frac{1}{n}\sum_{i=1}^n(\delta_i Y_i - \delta_i E[Y_i \mid X_i]) + \frac{1}{n}\sum_{i=1}^n(E[Y_i \mid X_i] - \mu) \\ &\quad + \left(\frac{1}{n}\sum_{i=1}^n(1-\delta_i)\int y\frac{\partial}{\partial\theta^{\top}}f(y \mid X_i,\theta)dy\right)[(\widetilde{\theta}_n - \theta)] + o_p(n^{-1/2})\end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\delta_i Y_i - \delta_i E[Y_i \mid X_i]) + \frac{1}{n}\sum_{i=1}^{n}(E[Y_i \mid X_i] - \mu)$$
$$+ S_2 S_1^{-1}\frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{\partial \log f(Y_i \mid X_i,\theta)}{\partial\theta} + o_p(n^{-1/2}).$$

Hence, by central limit theorem and some simple calculations, we can get the result of Theorem 2.1 by noting

$$\mathrm{Cov}(\delta Y - \delta E[Y \mid X], E[Y \mid X] - \mu) = 0,$$
$$\mathrm{Cov}\left(E[Y \mid X] - \mu, \delta\frac{\partial \log f(Y \mid X,\theta)}{\partial\theta}\right) = 0$$

and

$$\mathrm{Cov}\left(\delta Y - \delta E[Y \mid X], \delta\frac{\partial \log f(Y_i \mid X_i,\theta)}{\partial\theta}\right) = S_2^{\mathrm{T}} S_1^{-1} S_3 \tag{A.3}$$

under MAR assumption.

Proof of Theorem 2.2. By Lemma 2 of Owen (1990), the origin is inside the convex hull of $A(X_1),\ldots,A(X_n)$. Hence, the solution of (2.3) exists. Applying Taylor's expansion to (2.3), together with the fact that $\frac{1}{n}\sum_{i=1}^{n} A(X_i) = O_p(n^{-1/2})$, it follows that

$$\zeta_n = \left(\frac{1}{n}\sum_{i=1}^{n} A(X_i)A^{\mathrm{T}}(X_i)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} A(X_i)\right) + o_p(n^{-1/2}). \tag{A.4}$$

Hence, by Taylor's expansion and (A.4), we get

$$\widehat{\mu}_{n,2} = \frac{1}{n}\sum_{i=1}^{n}(\delta_i Y_i + (1-\delta_i)\widetilde{Y}_{i,n}) + \left[\frac{1}{n}\sum_{i=1}^{n}(\delta_i Y_i + (1-\delta_i)\widetilde{Y}_{i,n})A^{\mathrm{T}}(X_i)\right] \tag{A.5}$$
$$\times\left(\frac{1}{n}\sum_{i=1}^{n} A(X_i)A^{\mathrm{T}}(X_i)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} A(X_i)\right) + o_p(n^{-1/2}).$$

By law of large numbers, Taylor's expansion and the fact $\widetilde{\theta}_n - \theta = O_p(n^{-1/2})$, we get

$$\frac{1}{n}\sum_{i=1}^{n}\{\delta_i Y_i + (1-\delta_i)\widetilde{Y}_{i,n}\}A^{\mathrm{T}}(X_i) \tag{A.6}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\delta_i(Y_i - E[Y_i \mid X_i])A^{\mathrm{T}}(X_i)$$
$$+ \frac{\mu}{n}\sum_{i=1}^{n}A^{\mathrm{T}}(X_i) + \frac{1}{n}\sum_{i=1}^{n}(E[Y_i \mid X_i] - \mu)A^{\mathrm{T}}(X_i)$$
$$+ \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\left(\int yf(y \mid X_i,\widetilde{\theta}_n) - \int yf(y \mid X_i,\theta)dy\right)A^{\mathrm{T}}(X_i)$$
$$\to E[\delta(Y - E[Y \mid X])A^{\mathrm{T}}(X)] + E[(E[Y \mid X] - \theta)A^{\mathrm{T}}(X)].$$

By MAR assumption, we have $E[\delta(Y-E[Y\mid X])A^{\top}(X)]=0$. This together with (A.5) and (A.6) proves

$$\widehat{\mu}_{n,2}-\mu=(\widehat{\mu}_{n,1}-\mu)-A_1^{\top}A_2^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}A(X_i)\right)+o_p(n^{-1/2}), \tag{A.7}$$

where A_1 and A_2 are defined in Theorem 2.2. Let K_n be the second term at the right hand side of the equality in (A.7). Let T_{n1}, T_{n2} and T_{n3} be the first, second and third terms at the right hand side of the last equality in (A.2). By MAR assumption, we have $\mathrm{Cov}(\sqrt{n}K_n,\sqrt{n}T_{ni})=0$ for $i=1,3$ and $\mathrm{Cov}(\sqrt{n}K_n,\sqrt{n}T_{n2})=A_1^{\top}A_2^{-1}A_1$. Hence, by (A.2), (A.7), Theorem 2.1 and central limit theorem, Theorem 2.2 is then proved.

Proof of Theorem 3.1. Taylor's expansion and some standard arguments can be used to prove

$$(\widetilde{\theta}_{n,AU}-\theta)=(S_1+S_6)^{-1}\left(I_n+\frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{\partial\log f(Y_i\mid X_i,\theta)}{\partial\theta}\right)+o_p(1), \tag{A.8}$$

where

$$I_n=\frac{1}{n}\sum_{i=1}^{n}\left(-\frac{1}{1+\lambda_n^{\top}\psi(X_i,\theta)}\lambda_n^{\top}\frac{\partial\psi(X_i,\theta)}{\partial\theta}\right).$$

Similar to (A.2), we have

$$\begin{aligned}\widehat{\mu}_{n,3}-\mu &= \frac{1}{n}\sum_{i=1}^{n}(\delta_iY_i-\delta_iE[Y_i\mid X_i])+\frac{1}{n}\sum_{i=1}^{n}(E[Y_i\mid X_i]-\mu)\\ &\quad+\left(\frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\int y\frac{\partial}{\partial\theta^{\top}}f(y\mid X_i,\theta)dy\right)(\widetilde{\theta}_{n,AU}-\theta)\\ &= \frac{1}{n}\sum_{i=1}^{n}(\delta_iY_i-\delta_iE[Y_i\mid X_i])+\frac{1}{n}\sum_{i=1}^{n}(E[Y_i\mid X_i]-\mu)\\ &\quad+S_2(S_1+S_6)^{-1}\left(I_n+\frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{\partial\log f(Y_i\mid X_i,\theta)}{\partial\theta}\right).\end{aligned} \tag{A.9}$$

Applying Taylor's expansion to (3.1), together with the fact $\frac{1}{n}\sum_{i=1}^{n}\psi(Y_i,\theta)=O_p(n^{-1/2})$, it follows

$$\lambda_n=\left(\sum_{i=1}^{n}\psi(X_i,\theta)\psi^{\top}(X_i,\theta)\right)^{-1}\sum_{i=1}^{n}\psi(X_i,\theta)+o_p(n^{-1/2}). \tag{A.10}$$

By Owen (1988), we have $\max_{1\le i\le n}\|\psi(X_i,\theta)\|=o(n^{1/2})$ and $\lambda_n=O_p(n^{-1/2})$. This implies $\max_{1\le i\le n}|\lambda^{\top}\psi(X_i,\theta)|=o_p(1)$. Hence, Taylor's expansion together with (A.10) can be applied to prove

$$\begin{aligned}I_n &= -\Gamma_{n1}^{\top}\Gamma_{n2}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\theta)\right)\\ &\quad+\frac{1}{n}\sum_{i=1}^{n}\left\{\psi^{\top}(X_i,\theta)\Gamma_{n2}^{-1}\Psi_n\Psi_n^{\top}\Gamma_{n2}^{-1}\frac{\partial\psi(X_i,\theta)}{\partial\theta}\right\}+o_p(n^{-1/2}),\end{aligned} \tag{A.11}$$

where $\Gamma_{n1} = \frac{1}{n}\sum_{i=1}^n \frac{\partial\psi(X_i,\theta)}{\partial\theta}$, $\Gamma_{n2} = \frac{1}{n}\sum_{j=1}^n \psi(X_j,\theta)\psi(X_j,\theta)$ and $\Psi_n = \frac{1}{n}\sum_{j=1}^n \psi(X_j,\theta)$.

Note that

$$\psi_n = O_p(n^{-1/2}), \qquad \Gamma_{n1} = O_p(1), \qquad \Gamma_{n2} = O_p(1)$$

and

$$\Omega_n := \frac{1}{n}\sum_{i=1}^n \frac{\partial\psi(X_i,\theta)}{\partial\theta}\psi^{\mathrm{T}}(X_i,\theta) = O_p(1).$$

We have

$$\begin{aligned}(\mathrm{A.12}) \qquad I_n &= -E\left(\frac{\partial\psi(X,\theta)}{\partial\theta}\right)E^{-1}(\psi(X,\theta)\psi^{\mathrm{T}}(X,\theta))\left(\frac{1}{n}\sum_{i=1}^n\psi(X_i,\theta)\right)\\ &\quad + \mathrm{tr}(\Gamma_{n2}^{-1}\Psi_n\Psi_n^{\mathrm{T}}\Gamma_{n2}^{-1}\Omega_n) + o_p(n^{-1/2})\\ &= -E\left\{\frac{\partial\psi(X,\theta)}{\partial\theta}\right\}E^{-1}\{\psi(X,\theta)\psi^{\mathrm{T}}(X,\theta)\}\left\{\frac{1}{n}\sum_{i=1}^n\psi(X_i,\theta)\right\}\\ &\quad + o_p(n^{-1/2}).\end{aligned}$$

Hence, central limit theorem can be used to prove Theorem 3.1 by (A.9), (A.12) and some calculations.

Proof of Theorem 3.2. Similar to (A.7), we have

$$\widehat{\mu}_{n,4} - \mu = \frac{1}{n}\sum_{i=1}^n(\delta_iY_i + (1-\delta_i)\widetilde{Y}_{i,AU}) - B_1^{\mathrm{T}}B_2^{-1}\left(\frac{1}{n}\sum_{i=1}^n\psi(X_i,\theta)\right) + o_p(n^{-1/2}).$$

By Theorem 3.1, similar arguments to that used in the proof of Theorem 2.2 can be used to prove Theorem 3.2.

References

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random, *Journal of the American Statistical Association*, **89**, 81–87.

Hellerstein, J. K. and Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting, *Review of Economics and Statistics*, **81**, 1–14.

Imbens, G. W. and Lancaster, T. (1994). Combining micro and micro data in microeconometric models, *Review of Economic Studies*, **61**, 655–680.

Lehmann, E. L. (1983). *Theory of Point Estimation*, Wiley, New York.

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

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

Owen, A. (1990). Empirical likelihood ratio confidence regions, *The Annals of Statistics*, **18**, 90–120.

Wang, Q. H. and Rao, J. N. K. (2001). Empirical likelihood for linear regression models under imputation for missing responses, *Canadian Journal of Statistics*, **29**(4), 597–608.

Wang, Q. H. and Rao, J. N. K. (2002*a*). Empirical likelihood-based inference under imputation for missing response data, *The Annals of Statistics*, **30**(3), 896–924.

Wang, Q. H. and Rao, J. N. K. (2002*b*). Empirical likelihood-based inference in linear models with missing data, *Scandinavian Journal of Statistics*, **29**(3), 563–576.