The Annals of Statistics
2002, Vol. 30, No. 3, 896–924

EMPIRICAL LIKELIHOOD-BASED INFERENCE UNDER IMPUTATION FOR MISSING RESPONSE DATA[1]

BY QIHUA WANG AND J. N. K. RAO

Chinese Academy of Science and Carleton University

Inference under kernel regression imputation for missing response data is considered. An adjusted empirical likelihood approach to inference for the mean of the response variable is developed. A nonparametric version of Wilks' theorem is proved for the adjusted empirical log-likelihood ratio by showing that it has an asymptotic standard chi-squared distribution, and the corresponding empirical likelihood confidence interval for the mean is constructed. With auxiliary information, an empirical likelihood-based estimator is defined and an adjusted empirical log-likelihood ratio is derived. Asymptotic normality of the estimator is proved. Also, it is shown that the adjusted empirical log-likelihood ratio obeys Wilks' theorem. A simulation study is conducted to compare the adjusted empirical likelihood and the normal approximation methods in terms of coverage accuracies and average lengths of confidence intervals. Based on biases and standard errors, a comparision is also made by simulation between the empirical likelihood-based estimator and related estimators. Our simulation indicates that the adjusted empirical likelihood method performs competitively and that the use of auxiliary information provides improved inferences.

1. Introduction. For making statistical inference on the mean of a response variable, it is typically assumed that all the responses in the sample are available. This may not hold true in many practical situations and 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 the investigator to gather correct information and so forth. In fact, missing responses are common in opinion polls, market research surveys, mail enquiries, socioeconomic investigations, medical studies and other scientific experiments. In such circumstances, the usual inferential procedures for complete data sets cannot be applied directly. A common method is to impute a value for each missing response in order to achieve a complete data set and then apply standard statistical methods.

Received November 1999; revised September 2001.

[1]Suppported by grants from the Natural Sciences and Engineering Research Council of Canada and the National Natural Science Foundation of China.

AMS 2000 subject classifications. Primary 62G05; secondary 62E20.

Key words and phrases. Empirical likelihood, missing response, regression imputation.

Some commonly used imputation methods for missing responses (or nonresponses) include linear regression imputation [Yates (1933); Healy and Westmacott (1956)], kernel regression imputation [Cheng (1994)], nearest neighbor imputation [Chen and Shao (2000)] and ratio imputation [Rao (1996)]. Other usual imputation methods include random imputation [Rao and Shao (1992); Little and Rubin (1987)] and sequential imputation [Kong, Liu and Wong (1994)].

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

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

where all the X_i's are observed, and $\delta_i = 0$ if Y_i is missing; otherwise $\delta_i = 1$. Cheng (1994) applied kernel regression imputation to estimate the mean of Y, say θ. Let $m(x) = E[Y|X = x]$, let K be a kernel function and let h_n be a bandwidth sequence that decreases to 0 as n increases to ∞. Then $m(x)$ can be estimated by

$$\widehat{m}_n(x) = \frac{\sum_{i=1}^n \delta_i Y_i K((x - X_i)/h_n)}{\sum_{i=1}^n \delta_i K((x - X_i)/h_n)}. \tag{1.2}$$

Because $Em(X_i) = EY_i$, Cheng (1994) imputed missing Y_i by $\widehat{m}_n(X_i)$ and estimated θ by

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n \big(\delta_i Y_i + (1 - \delta_i)\widehat{m}_n(X_i)\big). \tag{1.3}$$

Under the assumption that some Y values may be missing at random (MAR), Cheng (1994) established the asymptotic normality of a trimmed version of $\widehat{\theta}_n$ and gave a consistent estimator of its asymptotic variance. This result can be used to perform interval estimation and hypothesis testing on θ. However, a competitive method of constructing confidence intervals for θ is the empirical likelihood method, introduced by Owen (1988). It has many advantages over normal approximation-based methods and the bootstrap for constructing confidence intervals [Hall (1992); Hall and La Scala (1990)]. For example, the empirical likelihood confidence intervals do not have a predetermined shape, whereas confidence intervals based on the asymptotic normality of an estimator have a symmetry implied by asymptotic normality. Also, empirical likelihood confidence intervals respect the range of the parameter: if the parameter is positive, then the confidence interval contains no negative values. Another preferred characteristic is that the empirical likelihood confidence interval is transformation respecting; that is, an empirical likelihood confidence interval for $\phi(\theta)$ is given by ϕ applied to each value in the confidence interval for θ. For the complete data setting, empirical likelihood methods have been studied extensively by many authors, including Owen (1988, 1990, 1991), DiCiccio, Hall and Romano (1991), Chen (1993, 1994),

Chen and Qin (1993), Chen and Hall (1993), Kolaczyk (1994), Kitamura (1997), Qin (1993), Qin and Lawless (1994) and Wang and Jing (1999). The original idea of empirical likelihood dates back at least to Hartley and Rao (1968), in a sample survey context, and Thomas and Grunkemeier (1975).

The main purpose of this article is to extend the empirical likelihood method to the missing response problem considered by Cheng (1994) and make inferences on the mean of response Y. Our main idea is to first impute the missing Y-values by the kernel regression imputation and then construct a complete data empirical likelihood for θ from the imputed data set as if they were indepedent and identically distributed (i.i.d.) observations. However, the imputed data are not i.i.d. As a consequence, the empirical log-likelihood ratio under imputation is asymptotically distributed as a scaled chi-square variable. Also, the empirical log-likelihood ratio cannot be applied directly to make statistical inference on θ. This motivates us to adjust the empirical log-likelihood ratio such that the adjusted log-likelihood ratio is asymptotically distributed as a standard chi-square variable. It should be noted that Adimari (1997) used the empirical likelihood method to make inference under random censorship and obtained an analogous result.

Another attractive feature of the empirical likelihood is that it can be used to make sharper inferences when some auxiliary information is available [Hartley and Rao (1968); Owen (1991); Zhang (1997)]. For our problem, some population characteristics of covariates X may be known in practice. For example, one may know the mean of X or that the distribution of X is symmetric about a known constant. By making effective use of the known auxiliary information on X and the empirical likelihood method, we propose an empirical likelihood-based estimator of θ, which has a smaller asymptotic variance than $\widehat{\theta}_n$, and some truncated versions of it. Also, we obtain an adjusted empirical likelihood ratio with auxiliary information and apply it to construct confidence intervals for θ.

The rest of this paper is organized as follows. In Section 2, an adjusted empirical log-likelihood ratio is derived and its asymptotic distribution is shown to be a standard chi-square. In Section 3, we define an empirical likelihood-based estimator of θ with auxiliary information and an adjusted empirical log-likelihood ratio, which is proved to be asymptotically distributed as a standard chi-square. In Section 4, a simulation study is conducted to compare the finite sample properties of the proposed empirical likelihood methods with normal approximation methods based on different estimators in terms of coverage accuracy and average length of confidence intervals. Based on bias and standard error, a comparison is also made between the empirical likelihood-based estimator and related estimators such as different trimmed versions of $\widehat{\theta}_n$. Proofs of theorems are relegated to the Appendix.

2. An adjusted empirical log-likelihood. Throughout this paper, we make the missing at random (MAR) assumption. The MAR assumption implies that δ and Y are conditionally independent given X. That is, $P(\delta = 1|Y, X) = P(\delta = 1|X)$. The MAR assumption may be reasonable in many practical situations

[see Little and Rubin (1987), Chapter 1]. Let $m(x) = E[Y|X = x]$ and $\widetilde{Y}_i = \delta_i Y_i + (1-\delta_i)m(X_i)$. Then, under MAR, $E\widetilde{Y}_i = \theta$ if θ is the true parameter. Hence, the problem of testing whether θ is the true parameter is equivalent to testing whether $E\widetilde{Y}_i = \theta$ for $i = 1, 2, \ldots, n$. By Owen (1991), this may be done using the empirical likelihood. Let $p_1, p_2, \ldots, p_n$ be nonnegative numbers summing to unity. Then the empirical log-likelihood ratio, evaluated at θ, is defined by

$$l_n(\theta) = -2 \max_{\sum_{i=1}^n p_i \widetilde{Y}_i = \theta, \sum_{i=1}^n p_i = 1} \sum_{i=1}^n \log(np_i). \tag{2.1}$$

If θ is the true parameter, $l_n(\theta)$ can be shown to be asymptotically distributed as a standard chi-square. However, the empirical log-likelihood ratio $l_n(\theta)$ cannot be used directly to make inferences on θ since it contains the unknown $m(\cdot)$ and hence θ is not identifiable. To resolve this problem, a natural way is to replace $m(\cdot)$ in $l_n(\theta)$ by the kernel regression estimator $\widehat{m}_n(\cdot)$ defined by (1.2). However, to avoid technical diffculties due to small values in the denominator of $\widehat{m}_n(\cdot)$, we define an estimated empirical log-likelihood ratio by replacing $m(\cdot)$ in $l_n(\theta)$ with a truncated version of $\widehat{m}_n(\cdot)$ instead of $\widehat{m}_n(\cdot)$. The truncation technique used here is slightly different from that used in Cheng (1994), but makes it easier to develop asymptotic theory.

Let

$$\widehat{g}_n(x) = (nh_n^d)^{-1} \sum_{i=1}^n \delta_i K\left(\frac{x - X_i}{h_n}\right) \quad \text{and} \quad \widehat{g}_{b_n}(x) = \max\{\widehat{g}_n(x), b_n\}$$

for some positive constant sequence b_n tending to zero. The truncated version of $\widehat{m}_n(x)$ is then defined by

$$\widehat{m}_{b_n}(x) = \frac{\widehat{m}_n(x)\widehat{g}_n(x)}{\widehat{g}_{b_n}(x)}.$$

Further, let $\widehat{Y}_{in} = \delta_i Y_i + (1 - \delta_i)\widehat{m}_{b_n}(X_i)$. Then the estimated empirical log-likelihood ratio is defined as

$$\hat{l}_n(\theta) = -2 \max_{\sum_{i=1}^n p_i \widehat{Y}_{in} = \theta, \sum_{i=1}^n p_i = 1} \sum_{i=1}^n \log(np_i). \tag{2.2}$$

By using the Lagrange multiplier method, when $\min_{1\le i\le n} \widehat{Y}_{in} < \theta < \max_{1\le i\le n} \widehat{Y}_{in}$, the optimal value of p_i satisfying (2.2) can be shown to be

$$p_i = \frac{1}{n} \frac{1}{1 + \lambda_n(\widehat{Y}_{in} - \theta)},$$

where λ_n is the solution of the equation

$$\frac{1}{n} \sum_{i=1}^n \frac{\widehat{Y}_{in} - \theta}{1 + \lambda_n(\widehat{Y}_{in} - \theta)} = 0. \tag{2.3}$$

This yields

$$\hat{l}_n(\theta) = 2\sum_{i=1}^{n} \log\bigl(1 + \lambda_n(\widehat{Y}_{in} - \theta)\bigr). \tag{2.4}$$

Compared to the standard empirical log-likelihood ratio, the main difference is that the $\widehat{Y}_{in}$'s in (2.4) are not independent and identically distributed. Hence, the asymptotic distribution of $\hat{l}_n(\theta)$ is not a standard chi-square. Actually, $\hat{l}_n(\theta)$ is asymptotically distributed as a scaled chi-square variable with one degree of freedom. Theorem 2.1 states this result.

THEOREM 2.1. *Under the assumptions listed in the Appendix* [*except* (C.A)], *if* θ *is the true parameter*, *we have*

$$\widehat{l}_n(\theta) \xrightarrow{\mathcal{L}} \frac{V(\theta)}{\widetilde{V}(\theta)}\chi_1^2, \tag{2.5}$$

where χ_1^2 *is a standard* χ^2 *variable with one degree of freedom*,

$$V(\theta) = E\left[\frac{\sigma^2(X)}{P(X)}\right] + \mathrm{Var}[m(X)] \tag{2.6}$$

and

$$\widetilde{V}(\theta) = E[P(X)\sigma^2(X)] + \mathrm{Var}[m(X)] \tag{2.7}$$

with $\sigma^2(X) = \mathrm{Var}(Y|X)$ *and* $P(X) = P(\delta = 1|X)$.

Clearly, (2.5) is equivalent to

$$l_{n,ad}(\theta) = r(\theta)\hat{l}_n(\theta) \xrightarrow{\mathcal{L}} \chi_1^2, \tag{2.8}$$

where $r(\theta) = \widetilde{V}(\theta)/V(\theta)$. If one can define a consistent estimator, say $r_n(\theta)$, of $r(\theta)$, an adjusted empirical log-likelihood ratio is then defined as

$$\hat{l}_{n,ad}(\theta) = r_n(\theta)\hat{l}_n(\theta) \tag{2.9}$$

with adjustment factor $r_n(\theta)$. It readily follows from (2.8) and (2.9) that $\hat{l}_{n,ad}(\theta) \xrightarrow{\mathcal{L}} \chi_1^2$ when θ is the true parameter.

We now provide a consistent estimator $r_n(\theta)$ of $r(\theta)$ under kernel regression imputation. Let $\widehat{f}_n(x) = (nh_n^d)^{-1}\sum_{i=1}^n K((x - X_i)/h_n)$ and $\widehat{f}_{b_n}(x) = \max\{\widehat{f}_n(x), b_n\}$. Write

$$\widehat{S}_n^2(x) = \frac{\sum_{i=1}^n K((x - X_i)/h_n)\delta_i Y_i^2}{\sum_{i=1}^n K((x - X_i)/h_n)\delta_i},$$

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

and

$$\widehat{V}_n(\theta) = n^{-1}\sum_{i=1}^{n}\left(\frac{\widehat{\sigma}^2_{b_n}(X_i)}{\widehat{P}_{b_n}(X_i)} + \widehat{m}^2_{b_n}(X_i)\right) - \theta^2, \tag{2.10}$$

where

$$\widehat{\sigma}^2_{b_n}(x) = \frac{\widehat{S}^2_n(x)\widehat{g}_n(x)}{\widehat{g}_{b_n}(x)} - \widehat{m}^2_{b_n}(x), \qquad \widehat{P}_{b_n}(x) = \frac{\widehat{P}_n(x)\widehat{f}_n(x)}{\widehat{f}_{b_n}(x)}.$$

Then, using Lemma A.2 and (A.46) from the Appendix, a consistent estimator of $r(\theta)$ is given by

$$r_n(\theta) = \frac{\widetilde{V}_n(\theta)}{\widehat{V}_n(\theta)},$$

where

$$\widetilde{V}_n(\theta) = n^{-1}\sum_{i=1}^{n}(\widehat{Y}_{in} - \theta)^2. \tag{2.11}$$

Using the adjustment factor $r_n(\theta)$, we get the following theorem.

THEOREM 2.2. *Under the assumptions listed in the Appendix* [*except* (C.A)], $\hat{l}_{n,ad}(\theta)$ *has an asymptotic* χ^2_1 *distribution if* θ *is the true parameter. That is*,

$$P\big(\hat{l}_{n,ad}(\theta) \le c_\alpha\big) = 1 - \alpha + o(1),$$

with $P(\chi^2_1 \le c_\alpha) = 1 - \alpha$.

REMARK 2.1. The results of Theorems 2.1 and 2.2 are valid for any suitable imputation method that leads to $\widehat{Y}_{in}$'s satisfying the properties in Lemmas A.1–A.5 of the Appendix after $r_n(\theta)$ is defined accordingly.

Clearly, Theorem 2.2 can be used to construct confidence intervals for θ. Let

$$I_{n,\alpha} = \{\theta' : \hat{l}_{n,ad}(\theta') \le c_\alpha\}.$$

Then, by Theorem 2.2, $I_{n,\alpha}$ gives an approximate confidence interval for θ with asymptotically correct coverage probability $1 - \alpha$; that is,

$$P(\theta \in I_{n,\alpha}) = 1 - \alpha + o(1).$$

Recalling the definition of $r(\theta)$ in (2.8), it is easy to see that $r(\theta) = 1$ when $P(\delta = 1|X = x) = 1$. In this case, the adjusted empirical log-likelihood ratio reduces to the standard empirical log-likelihood ratio. Actually, one can regard $r(\theta)$ as a measure of loss of information due to missing responses. Note from (2.6) and (2.7) that $\widetilde{V}(\theta) \le V(\theta)$ and hence $r(\theta) \le 1$.

3. Empirical likelihood-based methods with auxiliary information. We assume that auxiliary information on X of the form $EA(X) = 0$ is available, where $A(\cdot) = (A_1(\cdot), \ldots, A_r(\cdot))^\tau$, $r \geq 1$, 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

$$\prod_{i=1}^{n} \widetilde{p}_i \tag{3.1}$$

subject to $\sum_{i=1}^n \widetilde{p}_i = 1, \sum_{i=1}^n \widetilde{p}_i A(X_i) = 0$. 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

$$\widetilde{p}_i = \frac{1}{n} \frac{1}{1 + \zeta_n^\tau A(X_i)},$$

where ζ_n is the solution of the following equation:

$$\frac{1}{n} \sum_{i=1}^{n} \frac{A(X_i)}{1 + \zeta_n^\tau A(X_i)} = 0. \tag{3.2}$$

An empirical likelihood-based estimator (ELBE) of θ is then defined by

$$\widehat{\theta}_{n,AU} = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i Y_i + (1 - \delta_i)\widehat{m}_{b_n}(X_i)}{1 + \zeta_n^\tau A(X_i)}. \tag{3.3}$$

3.1. *Normal approximation-based confidence intervals.* Theorem 3.1 establishes the asymptotic normality of the ELBE estimator $\widehat{\theta}_{n,AU}$.

THEOREM 3.1. *Under all the conditions listed in the Appendix, if θ is the true parameter, we have*

$$\sqrt{n}(\widehat{\theta}_{n,AU} - \theta) \xrightarrow{\mathcal{L}} N\big(0, V_{AU}(\theta)\big),$$

where

$$V_{AU}(\theta) = V(\theta) - V'(\theta)$$

with

$$V'(\theta) = E\big((m(X) - \theta)A(X)\big)^\tau \big(EA(X)A^\tau(X)\big)^{-1} E\big((m(X) - \theta)A(X)\big)$$

and $V(\theta)$ given by (2.6).

REMARK 3.1. $\widehat{\theta}_{n,AU}$ remains asymptotically normal for any suitable imputation method that leads to $\widehat{Y}_{in}$'s satisfying Lemma A.1 and some properties such as (A.50) and (A.54).

By Cheng (1994) and Lemma A.1 of the Appendix, $V(\theta)$ is the asymptotic variance of $\widehat{\theta}_n$, the truncated version $\bar{\theta}_n$ and the truncated version of $\widehat{\theta}_n$ due to

Cheng (1994), say $\widetilde{\theta}_n$, where $\bar{\theta}_n$ and $\widetilde{\theta}_n$ are defined to be $\widehat{\theta}_n$ with $\widehat{m}_n(\cdot)$ replaced, respectively, by $\widehat{m}_{b_n}(\cdot)$ and

$$\widehat{m}_n(\cdot) I\left[\sum_{j=1}^{n} K\left(\frac{\cdot - X_j}{h_n}\right)\delta_j \geq c_n h_n^{-1} \log n\right]$$

with trimming coefficient c_n. It follows from Theorem 3.1 that the ELBE, $\widehat{\theta}_{n,AU}$, has smaller asymptotic variance and hence is asymptotically more efficient than $\widehat{\theta}_n$, $\widetilde{\theta}_n$ and $\bar{\theta}_n$.

REMARK 3.2. It is seen from the expression for $V'(\theta)$ in Theorem 3.1 that the amount of reduction from the asymptotic variance of $\widehat{\theta}_n$, $\widetilde{\theta}_n$ and $\bar{\theta}_n$ to that of $\widehat{\theta}_{n,AU}$, in the presence of auxiliary information on X, does not depend on the response probability function $P(x)$. This may be due to the fact that X is not missing and hence the amount of reduction due to use of auxiliary information on X does not depend on the missing rate of Y.

Let

$$V'_{n,AU} = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\left(\widehat{m}_{b_n}(X_i) - \widehat{\theta}_{n,AU}\right)A^{\tau}(X_i)\right)\right)\left(\frac{1}{n}\sum_{i=1}^{n} A(X_i)A^{\tau}(X_i)\right)^{-1}$$
$$\times \left(\frac{1}{n}\sum_{i=1}^{n}\left(\left(\widehat{m}_{b_n}(X_i) - \widehat{\theta}_{n,AU}\right)A(X_i)\right)\right).$$

By the "plug-in" method, we obtain consistent estimators of the asymptotic variance $V(\theta)$ of $\widehat{\theta}_n$, $\widetilde{\theta}_n$ and $\bar{\theta}_n$ as $V_n = \widehat{V}_n(\widehat{\theta}_n)$, $\widetilde{V}_n = \widehat{V}_n(\widetilde{\theta}_n)$ and $\bar{V}_n = \widehat{V}_n(\bar{\theta}_n)$, respectively, and of the asymptotic variance $V_{AU}(\theta)$ of $\widehat{\theta}_{n,AU}$ as

$$V_{n,AU} = \widehat{V}_n(\widehat{\theta}_{n,AU}) - V'_{n,AU}, \tag{3.4}$$

where $\widehat{V}_n(\cdot)$ is as defined in (2.10). The above results give the following normal approximation–based confidence intervals for θ: $\widehat{\theta}_n \pm u_{1-\alpha/2}V_n^{1/2}/\sqrt{n}$, $\widetilde{\theta}_n \pm u_{1-\alpha/2}\widetilde{V}_n^{1/2}/\sqrt{n}$, $\bar{\theta}_n \pm u_{1-\alpha/2}\bar{V}_n^{1/2}/\sqrt{n}$ and $\widehat{\theta}_{n,AU} \pm u_{1-\alpha/2}V_{n,AU}^{1/2}/\sqrt{n}$, where $u_{1-\alpha/2}$ is the $1-\alpha/2$ percentile point of the standard normal distribution.

REMARK 3.3. An important issue in the context of kernel regression is the selection of an appropriate bandwidth sequence h_n. For $\widehat{\theta}_{n,AU}$, a possible method is to choose both h_n and b_n as the values for which the mean square error is minimal. Unfortunately, the mean square error of $\widehat{\theta}_{n,AU}$ is difficult to calculate since $\widehat{\theta}_{n,AU}$ is obtained from both (3.2) and (3.3) and it is difficult to solve (3.2). However, it is noted that θ is a global mean functional and hence the $n^{1/2}$-rate asymptotic normality of $\widehat{\theta}_{n,AU}$ indicates that a proper choice of h_n and b_n specified in conditions (C.$h_n b_n$) and (C.h_n) of the Appendix depends only on the second-order term of the mean square error if $\widehat{\theta}_{n,AU}$ is uniformly square integrable. That

is, the $n^{1/2}$-rate asymptotic normality of $\widehat{\theta}_{n,AU}$ implies that $E(\widehat{\theta}_{n,AU}-\theta)^2 = V_{AU}(\theta)/n + r(h_n, b_n)$ with $r(h_n, b_n) = o(n^{-1})$ when $\widehat{\theta}_{n,AU}$ is uniformly square integrable. This shows that the selection of h_n and b_n might not be so critical in terms of the $n^{1/2}$-rate asymptotic normality and the mean square error of $\widehat{\theta}_{n,AU}$. This differs from nonparametric curve estimation in which the optimal choice of the smoothing parameter is required to achieve the optimal rate of convergence. If the remainder term $r(h_n, b_n)$ could be calculated, we suggest taking b_n as small as possible. One way of doing this is to start with some preliminary guess of b_n, use this to select h_n as the value for which the remainder term is minimal, then take a final b_n by minimizing the remainder term again.

3.2. *Adjusted empirical likelihood confidence intervals.* In what follows, we propose an adjusted empirical likelihood method to construct confidence intervals for θ when auxiliary information is available. The problem is to maximize

$$\prod_{i=1}^{n} \bar{p}_i \tag{3.5}$$

subject to $\sum_{i=1}^n \bar{p}_i = 1$, $\sum_{i=1}^n \bar{p}_i A(X_i) = 0$ and $\sum_{i=1}^n \bar{p}_i(\widehat{Y}_{in} - \theta) = 0$.

Let $h_{ni}(\theta) = (A^\tau(X_i), \widehat{Y}_{in} - \theta)^\tau$. Then, provided that the origin is inside the convex hull of points $h_{n1}(\theta), \ldots, h_{nn}(\theta)$, the method of Lagrange multipliers may be used to show that the solution is given by

$$\bar{p}_{in} = \frac{1}{n}\frac{1}{1+\eta_n^\tau h_{ni}(\theta)},$$

where η_n satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\frac{h_{ni}(\theta)}{1+\eta_n^\tau h_{ni}(\theta)} = 0. \tag{3.6}$$

Hence, the empirical log-likelihood ratio may be defined as

$$\hat{l}_{n,AU}(\theta) = -2\sum_{i=1}^{n}\log(n\bar{p}_{in}) = 2\sum_{i=1}^{n}\log\bigl(1+\eta_n^\tau h_{ni}(\theta)\bigr). \tag{3.7}$$

Let

$$V_{n1}(\theta) = \frac{1}{n}\sum_{i=1}^{n} A(X_i)A^\tau(X_i), \qquad V_{n2}(\theta) = \frac{1}{n}\sum_{i=1}^{n} A(X_i)(\widehat{Y}_{in} - \theta),$$

$$V_{n3}(\theta) = \frac{1}{n}\sum_{i=1}^{n} A(X_i)(\widehat{m}_{b_n}(X_i) - \theta)$$

and

$$V_{n1,AU}(\theta) = \begin{pmatrix} V_{n1}(\theta), & V_{n2}(\theta) \\ V_{n2}^\tau(\theta), & \widetilde{V}_n(\theta) \end{pmatrix} \quad \text{and} \quad V_{n2,AU}(\theta) = \begin{pmatrix} V_{n1}(\theta), & V_{n3}(\theta) \\ V_{n3}^\tau(\theta), & \widehat{V}_n(\theta) \end{pmatrix},$$

where $\widetilde{V}_n(\theta)$ and $\widehat{V}_n(\theta)$ are defined in (2.11) and (2.10), respectively. Further, let

$$W_{n1,AU}(\theta) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_{ni}(\theta)\right)^{\tau} V_{n1,AU}^{-1}(\theta)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_{ni}(\theta)\right)$$

and

$$W_{n2,AU}(\theta) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_{ni}(\theta)\right)^{\tau} V_{n2,AU}^{-1}(\theta)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_{ni}(\theta)\right).$$

It can be shown that

$$\hat{l}_{n,AU}(\theta) = W_{n1,AU}(\theta) + o_p(1).$$

Also, by the asymptotic normality of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_{ni}(\theta)$, it can be proved that $W_{n1,AU}(\theta)$ is asymptotically distributed as a weighted sum of independent chi-square variables, $\chi^2_{1,i}$, with weights w_i for $1 \le i \le r+1$ being the eigenvalues of $V_{1,AU}^{-1}(\theta)V_{2,AU}(\theta)$, where $V_{1,AU}(\theta) = P\lim_n V_{n1,AU}(\theta)$ and $V_{2,AU}(\theta) = P\lim_n V_{n2,AU}(\theta)$. This proves that $\hat{l}_{n,AU}(\theta) \xrightarrow{\mathcal{L}} \sum_{i=1}^{r+1} w_i \chi^2_{1,i}$. Unfortunately, this result cannot be applied to make statistical inference directly since the weights are unknown. Naturally, we hope to adjust $\hat{l}_{n,AU}(\theta)$, just like adjusting $\hat{l}_n(\theta)$ in Section 2, by examining the leading term in the asymptotic expansion of $\hat{l}_{n,AU}(\theta)$ such that the adjusted empirical log-likelihood ratio has asymptotically a standard chi-squared distribution.

By examining the leading term of $\hat{l}_{n,AU}(\theta)$, we define an adjusted empirical log-likelihood function $\hat{l}_{ad,AU}(\theta)$ as

$$\hat{l}_{ad,AU}(\theta) = \frac{W_{n2,AU}(\theta)}{W_{n1,AU}(\theta)}\hat{l}_{n,AU}(\theta). \tag{3.8}$$

THEOREM 3.2. *Under all the conditions listed in the Appendix, $\hat{l}_{ad,AU}(\theta)$ is asymptotically χ^2_{r+1} if θ is the true parameter.*

REMARK 3.4. The result of Theorem 3.1 is valid for any suitable imputation method that leads to $(A^{\tau}(X_i), \widehat{Y}_{in})^{\tau}$ satisfying Lemma A.5(b) and properties similar to those of Lemmas A.1–A.4.

Similarly to Theorem 2.1, the result of Theorem 3.2 can be used to construct a confidence interval for θ. It may be noted that $\hat{l}_{ad,AU}(\theta)$ reduces to $\hat{l}_{ad}(\theta)$ in Section 2 in the case of no auxiliary information.

REMARK 3.5. If the empirical maximum likelihood estimator (EMLE) of θ based on $\hat{l}_{n,AU}(\theta)$ exists, one may consider the statistic

$$\widetilde{l}_{n,AU}(\theta) = \hat{l}_{n,AU}(\theta) - \hat{l}_{n,AU}(\widetilde{\theta}_{n,AU}),$$

to construct a confidence interval, where $\widetilde{\theta}_{n,AU}$ is the EMLE. $\widetilde{l}_{n,AU}(\theta)$ can be proved to be asymptotically distributed as a scaled chi-square with one degree of freedom. Just like $\hat{l}_{n,AU}(\theta)$, $\widetilde{l}_{n,AU}(\theta)$ cannot be applied to make statistical inferences directly since the scaled chi-square contains an unknown scale parameter. However, we can adjust $\widetilde{l}_{n,AU}(\theta)$ such that the adjusted statistic has asymptotically a standard chi-square distribution. Unfortunately, we found that the EMLE sometimes was not available by simulation. It does not seem to be easy to prove the existence of the EMLE. Here, we do not investigate $\widetilde{l}_{n,AU}(\theta)$ further.

4. Simulation results. We considered five approaches for constructing confidence intervals in the presence of missing response (or nonresponse): the adjusted empirical likelihood method suggested in Section 2; normal approximation methods based on $\widetilde{\theta}_n$, $\bar{\theta}_n$ and $\widehat{\theta}_{n,AU}$; and the adjusted empirical likelihood method, introduced in Section 3, with auxiliary information. We do not consider $\widehat{\theta}_n$ because the denominator of $\widehat{m}_n(\cdot)$ becomes zero sometimes and hence it can behave poorly for small or moderate sample sizes.

A simulation study was conducted to compare the five methods in terms of coverage accuracies and average lengths of confidence intervals based on them. Also, we compared $\bar{\theta}_n$ with $\widetilde{\theta}_n$ and $\widehat{\theta}_{n,AU}$ in terms of their biases and standard errors.

The simulation used the model $Y = 3.2X^2 - 5.4X + \sqrt{|X|}\varepsilon$ with X simulated from the normal distribution with mean 1 and variance 1 or the corresponding truncated normal with truncation constant 4, and ε generated from the standard normal distribution. The kernel function of order 2 was taken as $K(t) = \frac{1}{2}$, $|t| \le 1$, and the bandwidth as $\frac{3}{2}n^{-1/3}$. For the calculation of $\widetilde{\theta}_n$ and $\bar{\theta}_n$, both b_n and c_n were taken as $n^{-1/6}\log n$. We considered the following three response probability functions $P(x) = P(\delta = 1|X = x)$ under the MAR assumption.

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

Case 2. $P(\delta = 1|X = x) = 0.9 - 0.2|x - 1|$ if $|x - 1| \le 4$, and $= 0.1$ elsewhere.

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

For each of the three cases, we generated 5000 Monte Carlo random samples of size $n = 30$, 60 and 100. For nominal confidence level $1 - \alpha = 0.95$, using the simulated samples, we evaluated the coverage probabilities and average lengths of the confidence intervals, which are reported in Tables 1 and 2. From the 5000 simulated values of $\widetilde{\theta}_n$, $\bar{\theta}_n$ and $\widehat{\theta}_{n,AU}$, we computed the biases and standard errors of the three estimators. These simulation results are reported in Table 3.

TABLE 1
Empirical coverages of the confidence intervals on θ under different missing functions $P(x)$ and sample sizes n when nominal level is 0.95

$P(x)$	n	AEL	AELA	NA($\tilde{\theta}_n$)	NA($\bar{\theta}_n$)	NA($\hat{\theta}_{n,AU}$)
$P_1(x)$	30	0.9234	0.9387	0.9122	0.9122	0.9002
	60	0.9411	0.9452	0.9356	0.9356	0.9298
	100	0.9466	0.9490	0.9445	0.9445	0.9413
$P_2(x)$	30	0.9130	0.9365	0.9075	0.9075	0.8861
	60	0.9362	0.9413	0.9298	0.9298	0.9186
	100	0.9439	0.9517	0.9409	0.9409	0.9386
$P_3(x)$	30	0.9129	0.9184	0.8728	0.8728	0.8242
	60	0.9318	0.9381	0.9220	0.9220	0.9050
	100	0.9421	0.9518	0.9307	0.9307	0.9232

Our simulation results for the case of normal X agree with those for the truncated normal X case. Our regularity conditions in the Appendix are satisfied for the latter case, and it is interesting that the results remain valid for the normal X case though it is not easy for us to prove the normal distribution to satisfy Assumption (C.gmb_n) for the polynomial model considered above.

For convenience, in what follows AEL and AELA denote the adjusted empirical likelihood confidence intervals based on Theorems 2.2 and 3.2, respectively, and NA($\tilde{\theta}_n$), NA($\bar{\theta}_n$) and NA($\hat{\theta}_{n,AU}$) denote the corresponding normal approximation confidence intervals defined in Section 3. The auxiliary information $EX = 1$ was used when we calculated the empirical coverages and average lengths of AELA and NA($\hat{\theta}_{n,AU}$) confidence intervals in Tables 1 and 2.

TABLE 2
Average lengths of the confidence intervals on θ under different missing functions $P(x)$ and sample sizes n when nominal level is 0.95

$P(x)$	n	AEL	AELA	NA($\tilde{\theta}_n$)	NA($\bar{\theta}_n$)	NA($\hat{\theta}_{n,AU}$)
$P_1(x)$	30	0.9000	0.7200	1.0624	1.0624	0.8047
	60	0.6400	0.5100	0.7733	0.7733	0.5851
	100	0.5200	0.3800	0.5855	0.5855	0.4529
$P_2(x)$	30	0.9400	0.8300	1.1465	1.1465	0.8652
	60	0.7300	0.6100	0.8687	0.8687	0.6340
	100	0.5700	0.4500	0.6102	0.6102	0.5039
$P_3(x)$	30	1.1200	0.9200	1.2156	1.2156	0.9288
	60	0.8100	0.6500	0.9375	0.9375	0.6953
	100	0.6300	0.5100	0.6588	0.6588	0.5488

From Tables 1 and 2, we have the following observations.

1. In the case where no auxiliary information is available, AEL performs better than NA($\widetilde{\theta}_n$) and NA($\bar{\theta}_n$) because the associated confidence intervals have uniformly higher coverage accuracies and shorter average lengths.
2. For the case when auxiliary information is available, we observe that the empirical coverage levels for the confidence intervals based on AELA are closer to the nominal level than those based on NA($\widehat{\theta}_{n,AU}$). Also, the average lengths of the confidence intervals based on AELA are uniformly shorter than those based on NA($\widehat{\theta}_{n,AU}$).
3. AELA obviously outperforms AEL, which does not use the auxiliary information, and hence also all the normal approximation methods in terms of coverage accuracies and average lengths of confidence intervals.
4. NA($\widehat{\theta}_{n,AU}$) has lower coverage accuracy, but shorter average length, than the other normal approximation methods. This could be explained by the fact that the estimator $\widehat{\theta}_{n,AU}$ has larger bias and smaller standard error (see Table 3).
5. NA($\widetilde{\theta}_n$) and NA($\bar{\theta}_n$) have the same coverage accuracy and average length of confidence interval for the cases considered here.
6. All the empirical coverage accuracies increase and the average lengths decrease as n increases. Also, the coverage accuracies and average lengths depend on the response probability function $P(x)$. In Case 1, all the methods generally perform better than in the other two cases. This could be explained by the fact that $EP_1(X) \approx 0.9$, $EP_2(X) \approx 0.74$ and $EP_3(X) = 0.6$, where $P_1(x)$, $P_2(x)$ and $P_3(x)$ are the response probability functions for Cases 1, 2 and 3, respectively; that is, Case 1 has the lowest missing rate and Case 3 has the highest missing rate.

TABLE 3
Biases and standard errors (SE) of $\widetilde{\theta}_n$, $\bar{\theta}_n$ and $\widehat{\theta}_{n,AU}$ under different missing functions $P(x)$ and different sample sizes n

		Bias			**SE**		
$\boldsymbol{P(x)}$	$\boldsymbol{n}$	$\widetilde{\boldsymbol{\theta}}_{\boldsymbol{n}}$	$\bar{\boldsymbol{\theta}}_{\boldsymbol{n}}$	$\widehat{\boldsymbol{\theta}}_{\boldsymbol{n,AU}}$	$\widetilde{\boldsymbol{\theta}}_{\boldsymbol{n}}$	$\bar{\boldsymbol{\theta}}_{\boldsymbol{n}}$	$\widehat{\boldsymbol{\theta}}_{\boldsymbol{n,AU}}$
$P_1(x)$	30	−0.0177	−0.0177	−0.0193	0.0928	0.0928	0.0614
	60	−0.0092	−0.0092	−0.0127	0.0418	0.0418	0.0311
	100	−0.0046	−0.0046	−0.0049	0.0254	0.0254	0.0139
$P_2(x)$	30	−0.1186	−0.1186	−0.1292	0.0974	0.0974	0.0711
	60	−0.0916	−0.0916	−0.0945	0.0505	0.0505	0.0342
	100	−0.0693	−0.0693	−0.0733	0.0311	0.0311	0.0214
$P_3(x)$	30	−0.0933	−0.0933	−0.1010	0.1232	0.1232	0.0848
	60	−0.0582	−0.0582	−0.0622	0.0592	0.0592	0.0435
	100	−0.0431	−0.0431	−0.0426	0.0369	0.0369	0.0250

From Table 3, we have the following observations.

(i) Biases and standard errors decrease as n increases. Also, the standard errors increase with the missing rate.

(ii) $\widetilde{\theta}_n$ and $\bar{\theta}_n$ have generally slightly smaller bias than $\widehat{\theta}_{n,AU}$. However, their standard errors (SE) are approximately 1.5 times as large as those of $\widehat{\theta}_{n,AU}$ irrespective of the choice of $P(x)$. This result is in agreement with the fact noted in Remark 3.2.

(iii) The bias and SD of $\widetilde{\theta}_n$ and of $\bar{\theta}_n$ are about the same under the cases considered here.

We have simulated other models such as $Y = 5\exp(-3X) + \varepsilon$, where X and ε have standard exponential and normal distributions, respectively, and the bivariate normal model that (X, Y) is distributed normally with mean vector $\mu = (1, 1)$ and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

The kernel function and bandwidth were taken to be the same as before. The above model with exponential X satisfies our condition (C.gmb_n), and the simulation results for the model are similar to those reported before. This can be seen from Table 4. We used auxiliary information $EX = 1$ when we calculated empirical coverages for AELA and NA($\widehat{\theta}_{n,AU}$) confidence intervals in Table 4.

Other bandwidths such as $Cn^{-1/3}$ for $C = (0.5, 1, 2)$ produced similar results, with $h_n = \frac{3}{2}n^{-1/3}$ generally performing better than the other bandwidths in terms of the coverage accuracy of confidence intervals. The choice of the trimming constants b_n and c_n seems not so sensitive with regard to coverage accuracy and interval length though it is important. When b_n and c_n were taken to be $n^{-1/7}$, n^{-1}, n^{-2} and n^{-10}, results similar to Tables 1–3 were obtained. When b_n and c_n were taken to be $n^{-1/10}$, $n^{-1/20}$ or $n^{-1/30}$, all the methods performed poorly, especially the normal approximation methods. For example, based on the model

TABLE 4

Empirical coverages (EC) and average lengths (AL) of confidence intervals on θ under different missing functions $P(x)$ and sample size $n = 60$ when nominal level is 0.95 and $Y = 5\exp(-3X) + \varepsilon$ with X and ε distributed as standard exponential and normal distributions, respectively

$P(x)$	n	AEL	AELA	NA($\widetilde{\theta}_n$)	NA($\bar{\theta}_n$)	NA($\widehat{\theta}_{n,AU}$)
$P_1(x)$	EC	0.9526	0.9480	0.9428	0.9428	0.9391
	AL	0.6400	0.5600	0.8274	0.8274	0.6195
$P_2(x)$	EC	0.9630	0.9382	0.9384	0.9384	0.9082
	AL	0.6900	0.5800	0.8743	0.8743	0.6237
$P_3(x)$	EC	0.9652	0.9357	0.9364	0.9364	0.9012
	AL	0.7500	0.6100	0.9056	0.9056	0.6578

used for Tables 1 and 2, the coverage probabilities of AEL, AELA, $\widetilde{\theta}_n$, $\bar{\theta}_n$ and $\widehat{\theta}_{n,AU}$ are 0.7751, 0.8845, 0.3511, 0.5720 and 0.6191 when $n = 30$, $b_n = n^{-1/10}$, $c_n = n^{-1/10}$, $P(x) = P_1(x)$ and the nominal level is 0.95.

5. Concluding remarks. Our paper considered empirical likelihood inference on the mean of response Y when Y is missing at random. Clearly, it is useful to extend these results to inference on the cumulative distribution function $F(t)$ of Y. To make the corresponding inference on $F(t)$, we can define the corresponding adjusted empirical likelihood functions and estimators of the distribution function of Y with Y and θ replaced by $I[Y \le t]$ and $F(t)$ in $\hat{l}_{n,ad}(\theta)$, $\hat{l}_{ad,AU}(\theta)$, $\widehat{\theta}_n$, $\widetilde{\theta}_n$ and $\bar{\theta}_n$, respectively, for any fixed t. Further, we can develop empirical likelihood inference for the quantiles of $F(t)$. We plan to study these problems in a separate paper.

APPENDIX

Assumptions and proofs of theorems. Denote by $f(\cdot)$ the probability density of X and let $\sigma^2(X) = \mathrm{Var}(Y|X)$. Define $P(x) = P(\delta = 1|X = x)$ and $g(x) = P(x)f(x)$. Let $\|a\| = \sum |a_i|$ for any vector a, where a_i is the ith component of a. To prove Theorems 2.1, 2.2, 3.1 and 3.2, the following assumptions are needed.

(C.P) (i) $P(x)$ has bounded partial derivatives up to order $k(> d)$ almost surely.
(ii) $\inf_x P(x) > 0$.

(C.f) $f(x)$ has bounded partial derivatives up to order $k(> d)$ almost surely.

(C.m) $m(x)$ has bounded partial derivatives up to order $k(> d)$ almost surely.

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

(C.gmb_n) $\sqrt{n}E[(1-\delta)|m(X)|I[g(X) < b_n]] \to 0$.

(C.K) (i) The kernel function K is a bounded kernel function with bounded support and bounded variation.
(ii) $K(\cdot)$ is a kernel of order $k(> d)$.

(C.h_n) (i) $nh_n^{2d}(b_n^2 \wedge (\log\log n)^{-1}) \to \infty$.
(ii) $nh_n^{2k}/b_n^2 \to 0$.

(C.$h_n b_n$) $h_n^k/b_n^2 \to 0$.

(C.A) $E\{A(X)A^\tau(X)\}$ is a positive definite matrix.

REMARK A.1. When $\sup_x |m(x)| < \infty$, (C.gmb_n) is implied by $P(g(X) < b_n) = o(n^{-1/2})$. Condition (C.gm$b_n$) is satisfied for the two cases studied in the simulation study: (i) truncated normal X and polynomial $m(X)$; (ii) standard exponential X and $m(X)$ proportional to $\exp(-aX)$ for $a > 0$ such that

$\sqrt{n}b_n^{1+a} \to 0$ [since it can be proved that $\sqrt{n}E[(1-\delta)|m(X)|I[g(X) < b_n]] = O(\sqrt{n}b_n^{1+a})$ in case (ii)]. In the simulation of Section 4, a was taken to be 3. This is to assure that the b_n used in our simulation satisfies $\sqrt{n}b_n^{1+a} \to 0$. Also, conditions (C.K), (C.h_n) and (C.$h_n b_n$) are standard conditions for nonparametric regression problems and they are satisfied for the dimension d of X, kernel $K(t)$, bandwidth h_n and truncation constant b_n used in the simulation study of Secton 4.

The following lemma is useful for proving Theorems 2.1, 2.2, 3.1 and 3.2.

LEMMA A.1. *Under the assumptions of Theorem* 2.1, *if* θ *is the true parameter*, *we have*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\widehat{Y}_{in}-\theta) \xrightarrow{\mathcal{L}} N\big(0, V(\theta)\big),$$

where

$$V(\theta) = E\left[\frac{\sigma^2(X)}{P(X)}\right] + \operatorname{Var}\big(m(X)\big)$$

with $\sigma^2(X) = \operatorname{Var}(Y|X)$.

REMARK A.2. Lemma A.1 gives the asymptotic normality result for the truncated version $\bar{\theta}_n$ of $\widehat{\theta}_n$, which is defined in Section 3.1. For the truncated version $\widetilde{\theta}_n$, Cheng (1994) obtained the same asymptotic result.

PROOF OF LEMMA A.1. Let

$$g_{b_n}(x) = \max\{g(x), b_n\} \quad \text{and} \quad m_{b_n}(x) = \frac{m(x)g(x)}{g_{b_n}(x)}.$$

It is easy to see that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\widehat{Y}_{in}-\theta) = \sqrt{n}(R_n + S_n + T_n + U_n), \tag{A.1}$$

where

$$R_n = n^{-1}\sum_{i=1}^{n}\big(m(X_i)-\theta\big),$$

$$S_n = n^{-1}\sum_{i=1}^{n}\delta_i\big(Y_i - m(X_i)\big),$$

$$T_n = n^{-1}\sum_{i=1}^{n}(1-\delta_i)\big(\widehat{m}_{b_n}(X_i) - m_{b_n}(X_i)\big)$$

and

$$U_n = -n^{-1}\sum_{i=1}^{n}(1-\delta_i)\big(m(X_i) - m_{b_n}(X_i)\big).$$

Recalling the definition of $m_{b_n}(\cdot)$, for any $\varepsilon > 0$ we have

$$\begin{aligned} P(\sqrt{n}|U_n| > \varepsilon) &\le P\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)|m(X_i)|I[g(X_i) < b_n] > \varepsilon\right) \\ &\le \varepsilon^{-1}\sqrt{n}E\big((1-\delta)|m(X)|I[g(X) < b_n]\big) \to 0 \end{aligned} \tag{A.2}$$

by condition (C.gmb_n), that is, $U_n = o_p(n^{-1/2})$.

Since R_n and S_n are means of independent and identically distributed random variables, the main task is to study T_n. Next, we show that T_n can be represented as a mean of i.i.d. random variables plus a remainder with order $o(n^{-1/2})$, that is,

$$T_n = n^{-1}\sum_{j=1}^{n}\delta_j\big(Y_j - m(X_j)\big)\frac{1-P(X_j)}{P(X_j)} + o_p(n^{-1/2}). \tag{A.3}$$

If (A.3) holds, then (A.1), (A.2) and (A.3) together prove Lemma A.1 by the central limit theorem and some direct computations.

Let

$$\eta_n(x) = (nh_n^d)^{-1}\sum_{j=1}^{n}\delta_j\big(Y_j - m(X_j)\big)K\left(\frac{x-X_j}{h_n}\right),$$

$$\zeta_n(x) = (nh_n^d)^{-1}\sum_{j=1}^{n}\delta_j\big(m(X_j) - m(x)\big)K\left(\frac{x-X_j}{h_n}\right),$$

$$\Delta_n(x) = \widehat{g}_n(x) - g(x)$$

and

$$\Delta_{b_n}(x) = \widehat{g}_{b_n}(x) - g_{b_n}(x).$$

Then

$$\begin{aligned} T_n &= n^{-1}\sum_{i=1}^{n}(1-\delta_i)\frac{\eta_n(X_i)}{g_{b_n}(X_i)} + n^{-1}\sum_{i=1}^{n}(1-\delta_i)\frac{\zeta_n(X_i)}{g_{b_n}(X_i)} \\ &\quad + n^{-1}\sum_{i=1}^{n}(1-\delta_i)\frac{m(X_i)\Delta_n(X_i)}{g_{b_n}(X_i)} - n^{-1}\sum_{i=1}^{n}(1-\delta_i)\frac{m(X_i)g(X_i)\Delta_{b_n}(X_i)}{g_{b_n}^2(X_i)} \\ &\quad - n^{-1}\sum_{i=1}^{n}(1-\delta_i)\frac{(\widehat{m}_n(X_i)\widehat{g}_n(X_i) - m(X_i)g(X_i))\Delta_{b_n}(X_i)}{g_{b_n}^2(X_i)} \\ &\quad + n^{-1}\sum_{i=1}^{n}(1-\delta_i)\frac{\widehat{m}_n(X_i)\widehat{g}_n(X_i)\Delta_{b_n}^2(X_i)}{g_{b_n}^2(X_i)\widehat{g}_{b_n}(X_i)} := \sum_{i=1}^{6}T_{ni}. \end{aligned} \tag{A.4}$$

To prove (A.3), we show that

$$T_{n1} = n^{-1}\sum_{j=1}^{n}\delta_j\big(Y_j - m(X_j)\big)\frac{1-P(X_j)}{P(X_j)} + o_p(n^{-1/2}), \tag{A.5}$$

$$T_{n3} + T_{n4} = o_p(n^{-1/2}) \tag{A.6}$$

and

$$T_{ni} = o_p(n^{-1/2}), \qquad i = 2, 5 \text{ and } 6. \tag{A.7}$$

(a) First, we prove (A.5). Observe that

$$\begin{aligned} T_{n1} &= n^{-1}\sum_{j=1}^{n}\delta_j\big(Y_j - m(X_j)\big)\frac{1}{h_n^d}\int\frac{(1-P(x))f(x)}{g_{b_n}(x)}K\left(\frac{x-X_j}{h_n}\right)dx \\ &\quad + n^{-1}\sum_{j=1}^{n}\delta_j\big(Y_j - m(X_j)\big)\left(\frac{1}{nh_n^d}\sum_{i=1}^{n}(1-\delta_i)\frac{K((X_i-X_j)/h_n)}{g_{b_n}(X_i)}\right. \\ &\quad \left. - h_n^{-d}\int\frac{(1-P(x))f(x)}{g_{b_n}(x)}K\left(\frac{x-X_j}{h_n}\right)dx\right) \\ &:= T_{n11} + T_{n12}. \end{aligned} \tag{A.8}$$

For T_{n11}, we have

$$\begin{aligned} T_{n11} &= n^{-1}\sum_{j=1}^{n}\delta_j\big(Y_j - m(X_j)\big)\int\frac{1-P(X_j+h_nu)}{P(X_j+h_nu)}K(u)\,du \\ &\quad + n^{-1}\sum_{j=1}^{n}\delta_j\big(Y_j - m(X_j)\big)\frac{1}{h_n^d}\int\frac{(1-P(x))(g(x)-g_{b_n}(x))}{g_{b_n}(x)P(x)} \\ &\quad \times K\left(\frac{x-X_j}{h_n}\right)dx \\ &:= \tau_{n1} + \tau_{n2}. \end{aligned} \tag{A.9}$$

Conditions (C.P)(i) and (C.P)(ii) imply that $\rho(x) = (1-P(x))/P(x)$ has partial derivatives up to order k almost surely. Applying Taylor's expansion for $\rho(x)$, and conditions (C.K) and $\sqrt{n}h_n^k \to 0$, which is implied by (C.h_n)(ii), we get

$$\tau_{n1} = n^{-1}\sum_{j=1}^{n}\delta_j\big(Y_j - m(X_j)\big)\frac{1-P(X_j)}{P(X_j)} + o_p(n^{-1/2}). \tag{A.10}$$

Noting that $|g(x) - g_{b_n}(x)| = 0$ when $g(x) \geq b_n$ and using arguments similar to those used in the proof of (A.10), we have

$$E[\sqrt{n}\tau_{n2}]^2 = E\left[\sigma^2(X)\frac{(1-P(X))^2}{P^2(X)}I[g(X) < b_n]\right] + O(h_n^{2k}) \tag{A.11}$$

by (C.K), (C.P)(i) and (C.f). Equation (A.11) proves

$$\tau_{n2} = o_p(n^{-1/2}) \tag{A.12}$$

by (C.P)(ii) and (C.Y).

Standard arguments can be used to prove

$$T_{n12} = o_p(n^{-1/2}).$$

This, together with (A.8), (A.9), (A.10) and (A.12), proves (A.5).

(b) Next we prove (A.6). Let

$$Q_n(X_i) = \frac{(1-\delta_i)m(X_i)[g_{b_n}(X_i)\widehat{g}_n(X_i) - g(X_i)\widehat{g}_{b_n}(X_i)]}{g^2_{b_n}(X_i)}.$$

It is easy to observe that

$$\begin{aligned} T_{n3} + T_{n4} &= \frac{1}{n}\sum_{i=1}^{n} Q_n(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} Q_n(X_i) I[g(X_i) < b_n, \widehat{g}_n(X_i) < b_n] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} Q_n(X_i) I[g(X_i) \geq b_n, \widehat{g}_n(X_i) < b_n] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} Q_n(X_i) I[g(X_i) < b_n, \widehat{g}_n(X_i) \geq b_n] \\ &:= J_{n1} + J_{n2} + J_{n3}. \end{aligned} \tag{A.13}$$

Note that $g_{b_n}(X_i) = b_n$, $|g_{b_n}(X_i)\widehat{g}_n(X_i) - g(X_i)\widehat{g}_{b_n}(X_i)| = b_n|\widehat{g}_n(X_i) - g(X_i)|$ and $|\widehat{g}_n(X_i) - g(X_i)| < 2b_n$ as $0 \leq g(X_i) < b_n$, $-b_n \leq \widehat{g}_n(X_i) < b_n$. Hence, we have

$$\begin{aligned} |J_{n1}| &\leq \frac{2}{n}\sum_{i=1}^{n}(1-\delta_i)|m(X_i)|I[g(X_i) < b_n] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} Q_n(X_i) I[g(X_i) < b_n, \widehat{g}_n(X_i) < -b_n] \\ &:= J_{n11} + J_{n12}. \end{aligned} \tag{A.14}$$

By the Markov inequality and condition (C.gmb_n), we get $\sqrt{n}|J_{n11}| \xrightarrow{p} 0$. On the other hand, for any $\varepsilon > 0$, $P(\sqrt{n}|J_{n12}| > \varepsilon) \leq P(\sup_x |\widehat{g}_n(x) - g(x)| > b_n) \to 0$ by the fact that $g(\cdot)$ is a nonnegative function and conditions (C.h_n)(i) and (C.$h_n b_n$). This together with (A.14) proves

$$\sqrt{n}|J_{n1}| \xrightarrow{p} 0. \tag{A.15}$$

Noting that $|g_{b_n}(X_i)\widehat{g}_n(X_i) - g(X_i)\widehat{g}_{b_n}(X_i)| \le b_n g(X_i)$ as $g(X_i) > b_n$ and $\widehat{g}_n(X_i) < b_n$ and using $g_{b_n}(X_i) \ge b_n$ and $g_{b_n}(X_i) \ge g(X_i)$, we get

$$
\begin{aligned}
|J_{n2}| &\le \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)|m(X_i)|I[g(X_i) \ge 2b_n, \widehat{g}_n(X_i) < b_n] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)|m(X_i)|I[g(X_i) < 2b_n] := J_{n21} + J_{n22}.
\end{aligned} \tag{A.16}
$$

By (C.gmb_n), it follows that $J_{n22} = o_p(n^{-1/2})$. For any $\varepsilon > 0$, $P(\sqrt{n}|J_{n21}| > \varepsilon) \le P(\sup_x |\widehat{g}_n(x) - g(x)| > b_n) \to 0$ by conditions (C.h_n)(i) and (C.$h_n b_n$). This proves

$$
J_{n2} = o_p(n^{-1/2}). \tag{A.17}
$$

Similarly to (A.17), we can prove

$$
J_{n3} = o_p(n^{-1/2}). \tag{A.18}
$$

Relations (A.13), (A.15), (A.16), (A.17) and (A.18) together prove (A.6).

(c) It remains to prove (A.7). By Taylor's expansion for $m(\cdot)$ and $g(\cdot)$, we get

$$
\begin{aligned}
&E[\zeta_n(X_i)|X_i] \\
&\quad = \int \big(m(X_i + h_n u) - m(X_i)\big) g(X_i + h_n u) K(u)\, du \\
&\quad = \int P_n(u) k(u)\, du + h_n^k \int R_n(\xi, u) K(u)\, du, \qquad 0 < \xi < 1,
\end{aligned} \tag{A.19}
$$

where $P_n(u)$ is a polynominal of degree $k-1$ on u and hence $\int P_n(u)K(u)\,du = 0$ by (C.K), and $R_n(\xi, u)$ is the kth remainder of the Taylor expansion and satisfies $\int R_n(\xi, u)K(u)\,du < \infty$ by (C.K). This together with (A.19) proves that

$$
E[\zeta_n(X_i)|X_i] \le c h_n^k, \qquad i = 1, 2, \ldots, n. \tag{A.20}
$$

By the derivative mean value theorem, it follows that

$$
\begin{aligned}
&E\big[\big(\zeta_n(X_i) - E[\zeta_n(X_i)|X_i]\big)^2 |X_i\big] \\
&\quad \le \frac{1}{n h_n^{2d}} \int \big(m(x) - m(X_i)\big)^2 K^2\Big(\frac{X_i - x}{h_n}\Big) g(x)\, dx \\
&\quad \le \frac{1}{n h_n^{2d-2}} \int \Big(\frac{\|x - X_i\|}{h_n}\Big)^2 K^2\Big(\frac{X_i - x}{h_n}\Big) g(x)\, dx \\
&\quad \le \frac{c}{n h_n^{d-2}} \int \|u\|^2 K^2(u) g(X_i - h_n u)\, du \\
&\quad \le \frac{c g(X_i)}{n h_n^{d-2}} + o\Big(\frac{1}{n h_n^{d-2}}\Big)
\end{aligned} \tag{A.21}
$$

by (C.P)(i) and (C.f). By some complicated calculations, it follows that

$$\begin{aligned}
&\frac{2}{n}\sum_{k\neq l}E\Big\{(1-\delta_k)(1-\delta_l)\\
&\qquad\times\Big[\frac{(\zeta_n(X_k)-E[\zeta_n(X_k)|X_k])(\zeta_n(X_l)-E[\zeta_n(X_l)|X_l])}{g_{b_n}(X_k)g_{b_n}(X_l)}\Big]\Big\}\\
&\le\frac{n(n-1)}{n^2}\frac{c}{nh_n^{2d}b_n^2}+\frac{n(n-1)(n-2)}{n^3}\frac{ch_n^{2k}}{b_n^2}.
\end{aligned}$$

This, together with (A.20) and (A.21), proves that

$$\begin{aligned}
E[\sqrt{n}T_{n2}]^2&=\frac{1}{n}E\Big[\sum_{i=1}^n(1-\delta_i)\frac{\zeta_n(X_i)-E[\zeta_n(X_i)|X_i]}{g_{b_n}(X_i)}\\
&\qquad+\sum_{i=1}^n(1-\delta_i)\frac{E[\zeta_n(X_i)|X_i]}{g_{b_n}(X_i)}\Big]^2\\
&\le\frac{2}{n}\sum_{i=1}^nE\Big[\frac{E[(\zeta_n(X_i)-E[\zeta_n(X_i)|X_i])^2|X_i]}{g_{b_n}^2(X_i)}\Big]\\
&\quad+\frac{2}{n}\sum_{k\neq l}E\Big\{(1-\delta_k)(1-\delta_l)\\
&\qquad\times\Big[\frac{(\zeta_n(X_k)-E[\zeta_n(X_k)|X_k])(\zeta_n(X_l)-E[\zeta_n(X_l)|X_l])}{g_{b_n}(X_k)g_{b_n}(X_l)}\Big]\Big\}\\
&\quad+2\sum_{i=1}^nE\Big[\frac{E^2(\zeta_n(X_i)|X_i)}{g_{b_n}^2(X_i)}\Big]\\
&\le\frac{c}{nh_n^{d-2}b_n}+\frac{cnh_n^{2k}}{b_n^2}+\frac{n(n-1)}{n^2}\frac{c}{nh_n^{2d}b_n^2}\\
&\quad+\frac{n(n-1)(n-2)}{n^3}\frac{ch_n^{2k}}{b_n^2}+o\Big(\frac{1}{b_n^2nh_n^{d-2}}\Big)\to 0
\end{aligned}\tag{A.22}$$

by (C.h_n). This proves the case of $i=2$ in (A.7).

Finally, we prove the case of $i = 5$ in (A.7). The case of $i = 6$ in (A.6) can be proved similarly. For T_{n5}, we have

$$\begin{aligned} -T_{n5} &= \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\frac{\eta_n(X_i)\Delta_{b_n}(X_i)}{g_{b_n}^2(X_i)} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\frac{\zeta_n(X_i)\Delta_{b_n}(X_i)}{g_{b_n}^2(X_i)} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\frac{m(X_i)\Delta_n(X_i)\Delta_{b_n}(X_i)}{g_{b_n}^2(X_i)} \\ &:= T_{n51} + T_{n52} + T_{n53}, \end{aligned} \tag{A.23}$$

where $\Delta_n(X_i) = \widehat{g}_n(X_i) - g(X_i)$. It is easy to observe that

$$|T_{n51}| \le \left(\frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\left|\frac{\eta_n(X_i)}{g_{b_n}^2(X_i)}\right|\right)\sup_x|\Delta_{b_n}(x)|. \tag{A.24}$$

Standard arguments can be used to prove that

$$\frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\left|\frac{\eta_n(X_i)}{g_{b_n}^2(X_i)}\right| = O_p\left(\frac{(nh_n^d)^{-1/2}}{b_n^2}\right) \tag{A.25}$$

and

$$\sup_x|\Delta_{b_n}(x)| \le \sup_x|\Delta_n(x)| = O_p((nh_n^d)^{-1/2}) + O_p(h_n^k). \tag{A.26}$$

Hence, by (A.24)–(A.26) we get

$$T_{n51} = o_p(n^{-1/2}), \tag{A.27}$$

by (C.h_n) and (C.h_nb_n).

Note that

$$|T_{n52}| \le \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\left|\frac{\zeta_n(X_i)}{g_{b_n}^2(X_i)}\right|\sup_x|\Delta_{b_n}(x)|. \tag{A.28}$$

By the proof of (A.7) with $i = 2$ and (A.26), we have

$$|T_{n52}| = o_p(n^{-1/2}). \tag{A.29}$$

For T_{n53}, we have

$$|T_{n53}| \le b_n^{-2}\left(\sup_x|\Delta_n(x)|\right)^2\frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)|m(X_i)|. \tag{A.30}$$

By (A.26) and condition (C.h_n), it follows that

$$T_{n53} = o_p(n^{-1/2}). \tag{A.31}$$

Equations (A.23), (A.27), (A.29) and (A.31) together prove (A.7) with $i = 5$. The case of $i = 6$ in (A.6) can be proved similarly. □

LEMMA A.2. *Under* (C.P), (C.f), (C.m), (C.Y) *and* (C.K), *if* $nh_n^d b_n \to \infty$, $h_n \to 0$ *and* θ *is the true parameter*, *then*

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_{in} - \theta)^2 \xrightarrow{p} \widetilde{V}(\theta), \tag{A.32}$$

where $\widetilde{V}(\theta) = E[P(X)\sigma^2(X)] + \mathrm{Var}(m(X))$ *with* $\sigma^2(X)$ *defined in Lemma* A.1.

PROOF. Using some arguments similar to those used in the proof of Lemma A.1, we can prove Lemma A.2. □

LEMMA A.3. *Let* $\widehat{Y}_{(n)} = \max_{1\le i\le n}|\widehat{Y}_{in}|$. *If* $EY^2 < \infty$, *then*

$$\widehat{Y}_{(n)} = o_p(n^{1/2}).$$

PROOF. By Owen (1990), $\max_{1\le i\le n}|Y_i| = o(n^{1/2})$ when $EY^2 < \infty$. Hence

$$\begin{aligned}\widehat{Y}_{(n)} &\le \max_{1\le i\le n}|Y_i| + \max_{1\le i\le n}|\widehat{m}_{b_n}(X_i)| \\ &\le \max_{1\le i\le n}|\widehat{m}_{b_n}(X_i) - m_{b_n}(X_i)| + \max_{1\le i\le n}|m_{b_n}(X_i)| + o_p(n^{1/2}).\end{aligned} \tag{A.33}$$

Clearly

$$Em_{b_n}^2(X) \le Em^2(X) \le E\big[E[Y^2|X]\big] = EY^2 < \infty. \tag{A.34}$$

Hence, by Lemma 3 of Owen (1990), we have

$$\max_{1\le i\le n}|m_{b_n}(X_i)| = o_p(n^{1/2}) \quad \text{and} \quad \max_{1\le i\le n}|m(X_i)| = o_p(n^{1/2}). \tag{A.35}$$

Standard arguments can be used to prove that $\max_{1\le i\le n}|\widehat{m}_{b_n}(X_i) - m_{b_n}(X_i)| = o_p(n^{1/2})$. This together with (A.33) and (A.35) proves Lemma A.3. □

LEMMA A.4. *Under the conditions of Lemmas* A.1 *and* A.2, *we have*

$$\lambda_n = O_p(n^{-1/2}).$$

PROOF. By Lemma A.1, it follows that

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_{in} = O_p(n^{-1/2}).$$

This, together with Lemma A.2, proves Lemma A.4 using the same arguments used in the proof of (2.14) in Owen (1990). □

LEMMA A.5. *Under assumptions* (C.P), (C.f), (C.m), (C.Y), (C.K), (C.h_n) *and* (C.$h_n b_n$), *if* θ *is the true parameter, we have the following*:

(a) $\min_{1\le i\le n}\widehat{Y}_{in} < \theta < \max_{1\le i\le n}\widehat{Y}_{in}$ *with probability tending to* 1 *when* $n\to\infty$.

(b) $(0^\tau,\theta)^\tau$ *is inside the convex hull of points* $(A^\tau(X_1),\widehat{Y}_{1n})^\tau,\dots,(A^\tau(X_n),\widehat{Y}_{nn}))^\tau$ *with probability tending to* 1.

PROOF. We first prove (a). It is easy to see that

$$P(\widehat{Y}_n<\theta)=P(Y<\theta,\delta=1)+P\big(\widehat{m}_n(X)<\theta,\delta=0\big) \ge P(Y<\theta,\delta=1), \tag{A.36}$$

where $\widehat{Y}_n=\delta Y+(1-\delta)\widehat{m}_{b_n}(X)$. By (C.Y) and Lemma 2 of Owen (1990), it follows that $P(Y<\theta)>0$. $P(Y<\theta)$ is continuous on θ from the left. Hence, there exists $\varepsilon_0>0$ such that $P(Y<\theta-\varepsilon_0)>0$. Let $I(A)$ be the indicator function of a certain event A. Then, by the MAR assumption (see Section 2) and (C.P)(ii), we get

$$\begin{aligned}P(Y<\theta-\varepsilon_0,\delta=1)&=E\{E[I(Y<\theta-\varepsilon_0,\delta=1)|X,Y]\}\\&=E\{I(Y<\theta-\varepsilon_0)P(\delta=1|X,Y)\}\\&=E\{I(Y<\theta-\varepsilon_0)P(\delta=1|X)\}\\&\ge P(Y<\theta-\varepsilon_0)\inf_x P(\delta=1|X=x)>0.\end{aligned}$$

This, together with (A.36), proves that

$$P(\widetilde{Y}<\theta-\varepsilon_0)\ge P(Y<\theta-\varepsilon_0,\delta=1)>0, \tag{A.37}$$

where $\widetilde{Y}=\delta Y+(1-\delta)m(X)$. Similarly, we have

$$P(\widetilde{Y}>\theta+\nu_0)\ge P(Y>\theta+\nu_0,\delta=1)>0 \tag{A.38}$$

for some $\nu_0>0$. Relations (A.37) and (A.38) then prove that

$$\begin{aligned}P\Big(\min_{1\le i\le n}\widehat{Y}_{in}\ge\theta\Big)&=P(\widehat{Y}_{1n}\ge\theta,\dots,\widehat{Y}_{nn}\ge\theta)\\&\le P(|\widehat{Y}_{1n}-\widetilde{Y}_1|+\widetilde{Y}_1\ge\theta,\dots,|\widehat{Y}_{nn}-\widetilde{Y}_n|+\widetilde{Y}_1\ge\theta)\\&\le P\Big(\widetilde{Y}_1\ge\theta-\varepsilon_0,\dots,\widetilde{Y}_n\ge\theta-\varepsilon_0,\\&\qquad\max_{1\le i\le n}|\widehat{m}_{b_n}(X_i)-m(X_i)|\le\varepsilon_0\Big)\\&\quad+P\Big(\max_{1\le i\le n}|\widehat{m}_{b_n}(X_i)-m(X_i)|>\varepsilon_0\Big)\\&\le P^n(\widetilde{Y}\ge\theta-\varepsilon_0)+o(1)\\&=\big(1-P(\widetilde{Y}<\theta-\varepsilon_0)\big)^n+o(1)\to0,\qquad n\to\infty.\end{aligned} \tag{A.39}$$

Similarly we have, by (A.38),

$$
\begin{aligned}
&P\Big(\max_{1\le i\le n}\widehat{Y}_{in}\le\theta+\nu_0\Big)\\
&\qquad\le P^n(\widetilde{Y}\le\theta+\nu_0)+o(1)\\
&\qquad=\big(1-P(\widetilde{Y}>\theta+\nu_0)\big)^n+o(1)\to 0,\qquad n\to\infty.
\end{aligned}
\tag{A.40}
$$

Together (A.38) and (A.40) prove (a).

Next, we prove (b). Let $\widehat{Z}_{in}=(A^\tau(X_i),\widehat{Y}_{in})^\tau$, $i=1,2,\ldots,n$, and let $\beta=(0^\tau,\theta)^\tau$. By Lemma 2 of Owen (1990) and arguments similar to those used in the proof of (a), we can prove that $\min_{1\le i\le n}\gamma'\widehat{Z}_{in}<\gamma'\beta<\max_{1\le i\le n}\gamma'\widehat{Z}_{in}$ with probability tending to 1 for any $\gamma\in\Omega$, where Ω is the set of unit vectors in R^{r+1}. This implies that there exists a vector $\alpha=(\alpha_1,\alpha_2,\ldots,\alpha_n)^\tau$ with $\sum_{i=1}^n\alpha_i=1$ and $\alpha_i\ge 0$, $i=1,2,\ldots,n$, such that $\gamma'\beta=\sum_{i=1}^n\alpha_i(\gamma'\widehat{Z}_{in})=\gamma'(\sum_{i=1}^n\alpha_i\widehat{Z}_{in})$ with probability tending to 1 for any $\gamma\in\Omega$. Since γ is arbitrary, it follows that $\beta=\sum_{1=1}^n\alpha_i\widehat{Z}_{in}$ with probability tending to 1. This proves (b). □

PROOFS OF THEOREMS 2.1 AND 2.2. By Lemma A.5(a) and the Lagrange multiplier method, (2.3) and (2.4) are obtained from (2.2). Applying Taylor's expansion to (2.4), we get

$$
\hat{l}_n(\theta)=2\sum_{i=1}^n\big\{\lambda_n(\widehat{Y}_{in}-\theta)-\tfrac{1}{2}[\lambda_n(\widehat{Y}_{in}-\theta)]^2\big\}+o_p(1)
\tag{A.41}
$$

by Lemmas A.2–A.4.

By (2.3), we get

$$
\begin{aligned}
0&=\sum_{i=1}^n\frac{(\widehat{Y}_{in}-\theta)}{1+\lambda_n(\widehat{Y}_{in}-\theta)}\\
&=\sum_{i=1}^n[(\widehat{Y}_{in}-\theta)]-\sum_{i=1}^n\lambda_n(\widehat{Y}_{in}-\theta)^2+\sum_{i=1}^n\frac{\lambda_n^2(\widehat{Y}_{in}-\theta)^3}{1+\lambda_n(\widehat{Y}_{in}-\theta)}.
\end{aligned}
$$

This implies

$$
\sum_{i=1}^n\lambda_n(\widehat{Y}_{in}-\theta)=\sum_{i=1}^n[\lambda_n(\widehat{Y}_{in}-\theta)]^2+o_p(1)
\tag{A.42}
$$

and

$$
\lambda_n=\left(\sum_{i=1}^n(\widehat{Y}_{in}-\theta)^2\right)^{-1}\sum_{i=1}^n(\widehat{Y}_{in}-\theta)+o_p(n^{-1/2})
\tag{A.43}
$$

by Lemmas A.2–A.4.

Equations (A.41), (A.42) and (A.43) together yield

$$\hat{l}_n(\theta) = \widetilde{V}_n^{-1}(\theta)\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\widehat{Y}_{in} - \theta)\right]^2 + o_p(1). \tag{A.44}$$

This, together with Lemmas A.1 and A.2, proves Theorem 2.1.

Recalling the definition of $\hat{l}_{n,ad}(\theta)$, by (A.44) we get

$$\hat{l}_{n,ad}(\theta) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\widehat{Y}_{in} - \theta}{\sqrt{\widehat{V}_n(\theta)}}\right)^2 + o_p(1). \tag{A.45}$$

Using arguments similar to those used in the proofs of Lemmas A.1 and A.2, we can prove

$$\widehat{V}_n(\theta) \xrightarrow{p} V(\theta), \tag{A.46}$$

where $V(\theta)$ is defined in Lemma A.1.

Hence, (A.45), (A.46) and Lemma A.1 together prove Theorem 2.2. □

PROOF OF THEOREM 3.1. By (C.A) and Lemma 2 of Owen (1990), the origin is inside the convex hull of $A(X_1), \ldots, A(X_n)$. Hence, the solution of (3.2) exists. Similarly to (A.43), by (3.2) we get

$$\zeta_n = \left(n^{-1}\sum_{i=1}^{n}A(X_i)A^{\tau}(X_i)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}A(X_i)\right) + o_p(n^{-1/2}). \tag{A.47}$$

Applying Taylor's expansion to (3.3), by (A.47) it follows that

$$\begin{aligned}\widehat{\theta}_{n,AU} = {} & \frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_{in} - \left(\frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_{in}A^{\tau}(X_i)\right)\left(\frac{1}{n}\sum_{i=1}^{n}A(X_i)A^{\tau}(X_i)\right)^{-1} \\ & \times\left(\frac{1}{n}\sum_{i=1}^{n}A(X_i)\right) + o_p(n^{-1/2}).\end{aligned} \tag{A.48}$$

Also, the law of large numbers implies that

$$\frac{1}{n}\sum_{i=1}^{n}A(X_i)A^{\tau}(X_i) \xrightarrow{p} EA(X)A^{\tau}(X). \tag{A.49}$$

Next, we prove

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_{in}A^{\tau}(X_i) \xrightarrow{p} E\big[\big(m(X) - \theta\big)A^{\tau}(X)\big]. \tag{A.50}$$

The left-hand side of (A.50) can be decomposed as

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_{in}A^{\tau}(X_i) &= \frac{1}{n}\sum_{i=1}^{n}\delta_i\big(Y_i - m(X_i)\big)A^{\tau}(X_i)\\ &\quad+\frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\big(\widehat{m}_{b_n}(X_i) - m_{b_n}(X_i)\big)A^{\tau}(X_i)\\ &\quad+\frac{\theta}{n}\sum_{i=1}^{n}A^{\tau}(X_i) + \frac{1}{n}\sum_{i=1}^{n}\big(m(X_i)-\theta\big)A^{\tau}(X_i)\\ &\quad-\frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\big(m(X_i) - m_{b_n}(X_i)\big)A^{\tau}(X_i)\\ &:= R_{n1}+R_{n2}+R_{n3}+R_{n4}+R_{n5}.\end{aligned} \tag{A.51}$$

By the law of large numbers and the MAR assumption, it follows that $R_{n1} \xrightarrow{p} E[\delta(Y - m(X))A^{\tau}(X)] = 0$, $R_{n3} \xrightarrow{p} 0$ and $R_{n4} \xrightarrow{p} E[(m(X)-\theta)A^{\tau}(X)]$. Using the same arguments as in proving (A.3), we can prove $R_{n2} \xrightarrow{p} 0$. By (A.2), we have $R_{n5} \xrightarrow{p} 0$. This proves (A.50). From (A.48)–(A.50), we get

$$\begin{aligned}\widehat{\theta}_{n,AU} &= \frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_{in} - E\big[\big(m(X)-\theta\big)A^{\tau}(X)\big]\big(E\big(A(X)A^{\tau}(X)\big)\big)^{-1}\\ &\quad\times\left(\frac{1}{n}\sum_{i=1}^{n}A(X_i)\right) + o_p(n^{-1/2}).\end{aligned} \tag{A.52}$$

Let $\bar{A}_n = \frac{1}{n}\sum_{i=1}^{n}A(X_i)$. By (A.1) and (A.2), it follows that

$$\begin{aligned}&\operatorname{Cov}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\widehat{Y}_{in}-\theta), \frac{1}{\sqrt{n}}\sum_{i=1}^{n}A(X_i)\right)\\ &\quad= \operatorname{Cov}(\sqrt{n}R_n, \sqrt{n}\bar{A}_n) + \operatorname{Cov}(\sqrt{n}S_n, \sqrt{n}\bar{A}_n)\\ &\qquad+\operatorname{Cov}(\sqrt{n}T_n, \sqrt{n}\bar{A}_n) + \operatorname{Cov}(\sqrt{n}U_n, \sqrt{n}\bar{A}_n).\end{aligned} \tag{A.53}$$

Standard arguments and calculations can be used to get $\operatorname{Cov}(\sqrt{n}R_n, \sqrt{n}\bar{A}_n) = E(A(X)(m(X)-\theta))$ and $\operatorname{Cov}(\sqrt{n}S_n, \sqrt{n}\bar{A}_n) = 0$. Using some of the arguments employed in the proof of Lemma A.1, we can prove $\operatorname{Cov}(\sqrt{n}T_n, \sqrt{n}\bar{A}_n) \to 0$. By (C.Y) and (C.A), the dominated convergence theorem can be used to prove that $\operatorname{Cov}(\sqrt{n}U_n, \sqrt{n}\bar{A}_n) = E((1-\delta)\|m(X)A(X)\|I[g(X) < b_n]) \to 0$. This proves

$$\operatorname{Cov}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\widehat{Y}_{in}-\theta), \frac{1}{\sqrt{n}}\sum_{i=1}^{n}A(X_i)\right) \longrightarrow E[A(X)(m(X)-\theta)]. \tag{A.54}$$

Hence, by (A.52), (A.54), Lemma A.1 and the central limit theorem, Theorem 3.1 is then proved. □

PROOF OF THEOREM 3.2. By Lemma A.1, the central limit theorem and (A.54), it follows that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_{ni}(\theta) \xrightarrow{\mathcal{L}} N\big(0, V_{2,AU}(\theta)\big),$$

where

$$V_{2,AU}(\theta) = \begin{pmatrix} V_1(\theta), & V_3(\theta) \\ V_3^{\tau}(\theta), & V(\theta) \end{pmatrix}$$

with $V_1(\theta) = EA(\theta)A^{\tau}(\theta)$, $V_3(\theta) = E[A(X)(m(X) - \theta)]$ and $V(\theta)$ defined in Lemma A.1.

By (A.46), (A.49) and (A.50), we get

$$V_{n2,AU}(\theta) \xrightarrow{p} V_{2,AU}(\theta). \tag{A.55}$$

By Lemma A.5(b), (3.6) and (3.7) can be obtained from (3.5). Hence, arguments similar to those used for Theorems 2.1 and 2.2 can be used to prove Theorem 3.2. □

Acknowledgments. The authors thank an Associate Editor and four referees for their constructive suggestions and comments that led to significant improvements. Thanks are also due to Dr. M. Mojirsheibani for useful discussion.

REFERENCES

ADIMARI, G. (1997). Empirical likelihood type confidence intervals under random censorship. *Ann. Inst. Statist. Math.* **49** 447–466.

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

CHEN, J. H. and SHAO, J. (2000). Nearest neighbor imputation for survey data. *J. Official Statist.* **16** 113–131.

CHEN, S. X. (1993). On the accuracy of empirical likelihood confidence regions for linear regression model. *Ann. Inst. Statist. Math.* **45** 621–637.

CHEN, S. X. (1994). Empirical likelihood confidence intervals for linear regression coefficients. *J. Multivariate Anal.* **49** 24–40.

CHEN, S. X. and HALL, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *Ann. Statist.* **21** 1166–1181.

CHENG, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Amer. Statist. Assoc.* **89** 81–87.

DICICCIO, T. J., HALL, P. and ROMANO, J. P. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.* **19** 1053–1061.

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

HALL, P. and LA SCALA, B. (1990). Methodology and algorithms of empirical likelihood. *Internat. Statist. Rev.* **58** 109–127.

HARTLEY, H. O. and RAO, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika* **55** 547–557.

HEALY, M. J. R. and WESTMACOTT, M. (1956). Missing values in experiments analysed on automatic computers. *Appl. Statist.* **5** 203–206.

KITAMURA, Y. (1997). Empirical likelihood methods with weakly dependent processes. *Ann. Statist.* **25** 2084–2102.

KOLACZYK, E. D. (1994). Empirical likelihood for generalized linear models. *Statist. Sinica* **4** 199–218.

KONG, A., LIU, J. S. and WONG, W. H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.* **89** 278–288.

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. *Ann. Statist.* **18** 90–120.

OWEN, A. (1991). Empirical likelihood for linear models. *Ann. Statist.* **19** 1725–1747.

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.

RAO, J. N. K. (1996). On variance estimation with imputed survey data (with discussion). *J. Amer. Statist. Assoc.* **91** 499–520.

RAO, J. N. K. and SHAO, J. (1992). Jacknife variance estimation with survey data under hot deck imputation. *Biometrika* **79** 811–822.

THOMAS, D. R. and GRUNKEMEIER, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Assoc.* **70** 865–871.

WANG, Q. H. and JING, B. Y. (1999). Empirical likelihood for partial linear models with fixed designs. *Statist. Probab. Lett.* **41** 425–433.

YATES, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empire Journal of Experimental Agriculture* **1** 129–142.

ZHANG, B. (1997). Empirical likelihood confidence intervals for M-functionals in the presence of auxiliary information. *Statist. Probab. Lett.* **32** 87–97.

INSTITUTE OF APPLIED MATHEMATICS
ACADEMY OF MATHEMATICS AND SYSTEM SCIENCE
CHINESE ACADEMY OF SCIENCE
BEIJING 100080
PEOPLE'S REPUBLIC OF CHINA
AND
DEPARTMENT OF PROBABILITY AND STATISTICS
PEKING UNIVERSITY
BEIJING 100871
PEOPLE'S REPUBLIC OF CHINA
E-MAIL: qhwang@amss.ac.cn

SCHOOL OF MATHEMATICS AND STATISTICS
CARLETON UNIVERSITY
OTTAWA K1S 5B6
CANADA
E-MAIL: jrao@math.carleton.ca