Statistica Sinica **6**(1996), 63-78

KERNEL ESTIMATION OF DISTRIBUTION FUNCTIONS AND QUANTILES WITH MISSING DATA

P. E. Cheng and C. K. Chu*

*Academia Sinica, National Dong Hwa University and National Tsing Hua University**

Abstract. A distribution-free imputation procedure based on nonparametric kernel regression is proposed to estimate the distribution function and quantiles of a random variable that is incompletely observed. Assuming the baseline missing-at-random model for nonrespondence, we discuss consistent estimation via estimating the conditional distribution by the kernel method. A strong uniform convergence rate comparable to that of density estimation is proved. We derive asymptotic normality for estimating the cdf and the quantile via establishing the mean square consistency and the asymptotically optimal bandwidth selection. A simulation study compares the proposed nonparametric method with the naive pairwise deletion method and a linear regression method under a parametric linear model.

Key words and phrases: Incomplete data, missing-at-random, nonparametric regression, conditional distribution function, quantile estimation, strong uniform consistency, asymptotic normality.

1. Introduction

Consider statistical inference with a basic pattern of incomplete data:

$$(X_i, Y_i, \delta_i), \quad i = 1, \ldots, N, \tag{1.1}$$

where $\delta_i = 1$ if Y_i is observed, otherwise $\delta_i = 0$ and Y_i is missing. Missing data of form (1.1) naturally arises from the double sampling scheme proposed by Neyman (1938), and extensively discussed by Cochran (1963). It also occurs in some longitudinal studies where follow-up records may be missing for a variety of reasons. Outpatients may miss the second medical examination, students may miss a follow-up test and survey interviewers often miss unit responses. Such common phenomena generate data with various missing patterns, among which the basic form (1.1) is usually termed fragmentary or monotone. Readers are referred to Little and Rubin (1987) for examples and the background history of parametric statistical inference with missing data.

To estimate the mean of Y with the missing data (1.1), Cheng and Wei (1986) utilized the Nadaraya-Watson (1964) kernel regression estimate to substitute for

each missing value of Y (see Müller (1988) and Härdle (1990) for properties and practical examples of kernel estimator). This idea parallels the notion of the Horvitz-Thompson (1952) estimate, incorporating continuous covariates. By this method of mean imputation together with an empirical assessment of the dependence of the missing mechanism on the covariate X, asymptotic normality for estimating the mean of Y was characterized (Cheng (1994)). For data of form (1.1), Titterington and Mill (1983) considered nonparametric estimation of the joint density of (X, Y), utilizing random imputations to generate empirical versions of the joint density. Both approaches were based on nonparametric estimation of the mean of a conditional distribution. The former assumed that data are missing at random (MAR), which is comparable with the notion that the missing mechanism is ignorable. The latter provided analyses when the data are missing completely at random (MCAR, see Rubin (1976) for elaborate definitions). It is well known that MAR fails to hold when missing (or censoring) occurs entirely over certain intervals, or when information of the background demographic factors, or a follow-up validation sample indicates that the missing mechanism is truly nonignorable. Nevertheless, the MAR assumption has been widely used as a baseline model for nonrespondence among many plausible parametric models, because the problem of model sensitivity is often a difficult issue in practice.

The goal of this study is to complement the theory developed in Cheng (1994) by establishing more precise asymptotic properties via estimating the cdf of Y when the data are MAR. A Glivenko-Cantelli type theorem with a uniform convergence rate is proved in Section 2. The rate is also valid for estimating the individual sample quantile. The corresponding asymptotic normality results are derived in Section 3. Remarks on the asymptotic minimum mean squared error and the associated optimal bandwidth selection are also addressed. The proofs are given for scalar-valued X, and the extension to vector-valued X is briefly remarked. Section 4 presents a simulation study which shows that the proposed nonparametric scheme could obtain more satisfactory bias performance, compared with the pairwise deletion method and a standard linear regression method, under an MAR data model. In terms of variance performance, our method can be inferior to the linear regression method when a parametric linear model holds. For an improvement over this preliminary study, one might consider the idea of multiple imputation (cf. Rubin (1987); Efron (1994)). In addition, methods with nonignorable missing data certainly merit further study.

2. Estimation of Distribution Function

Assume the missing data pattern (1.1), where both X and Y are continuous scalar-valued random variables. Suppose all the X_i's are observed, and the binary indicator variable $\delta_i = 1$, if Y_i is observed, otherwise $\delta_i = 0$. The assumption

of an ignorable missing mechanism (essentially, MAR) will be imposed in the sequel. Specifically, it assumes that δ and Y are conditionally independent given X, i.e.

$$P(\delta = 1|Y, X) = P(\delta = 1|X) = p(X). \tag{2.1}$$

It is intuitively clear that the MAR assumption (2.1) may be practically justified by the nature of the experiment when it is speculated that missing a value of Y depends mainly on the covariate X, but not on the values of Y (Little and Rubin (1987) p.14). Following Cheng and Wei (1986), we consider the method of nonparametric mean imputation based on the well known Nadaraya-Watson kernel regression estimator. To estimate the cdf of Y, $G(y) = P(Y \leq y)$, without completely observed Y in forming the usual empirical cdf, we will make use of the extra X values as follows. For each real value y, and each X_i with $\delta_i = 0$, construct an estimate of the conditional distribution function $G(y|X_i)$ by

$$\widehat{G}(y|X_i) = \sum_{j \neq i} W_j(X_i) I(Y_j \leq y), \tag{2.2}$$

where $W_j(X_i) = K_b(X_j - X_i)\delta_j / \sum_{j \neq i} K_b(X_j - X_i)\delta_j$ with $K_b(X_j - X_i) = b^{-1}K((X_j - X_i)/b)$. Here K is a kernel function that integrates to 1 on the real line, and $b = b(N)$ is the so-called bandwidth sequence which decreases toward 0 as N increases. An obvious estimate of $G(y)$ is given by

$$\widehat{G}(y) = N^{-1} \sum_{i=1}^{N} \{\delta_i I(Y_i \leq y) + (1 - \delta_i)\widehat{G}(y|X_i)\}. \tag{2.3}$$

The aim of this section is to show that the proposed $\widehat{G}$ obtains strong uniform consistency. A standard condition in the literature of nonparametric regression is that X has compact support where the pdf f of X is bounded away from 0, because the ratio weights $W_j(X_i)$ could yield unstable estimates $\widehat{G}(y|X_i)$ at the tail of f. For a uniform property, we shall impose this condition throughout this section. A few terminologies are now defined. Let $g(x) = f(x)p(x)$, and let $\sigma^2(y|X) = \mathrm{Var}(I[Y \leq y]|X) = G(y|X) - G^2(y|X)$ be the conditional variance, where $G(y|x) = P(Y \leq y|x)$. Assume that $P(\delta = 1) = E[p(X)] = p$ is a constant strictly between 0 and 1, excluding the trivial case $p = 1$. In the sequel the following conditions are assumed:

(1) $g(x) \geq 2c > 0$, where c is arbitrarily small, and g has bounded first derivative g' within the compact support of X;

(2) the kernel K is a finite-valued (or Lipschitz continuous), symmetric pdf with compact support, say $[-1, 1]$;

(3) the bandwidth sequence b is such that $b \to 0$, $Nb^5 \simeq \log N$ as $N \to \infty$;

(4) $\frac{\partial}{\partial x} G(y|x)$, denoted by $G'(y|x)$, is bounded in both x and y.

Condition (1) implies that missing Y entirely over any subinterval of X is disallowed, which is intuitively necessary for a purely nonparametric approach. The effect of the smoothness of g is however less crucial; and this will be explained in the simulation study of Section 4. Conditions (3) and (4) are used in deriving the specific uniform convergence rate in the Glivenko-Cantelli type theorem below, although it suffices to require that $Nb/\log N \to \infty$ as $N \to \infty$ in (3).

Theorem 2.1. *Assume* (2.1) *and conditions* (1) *to* (4). *Then almost surely*

$$\limsup_{N\to\infty}(Nb/\log N)^{1/2}\sup_{y\in\mathbb{R}}|\widehat{G}(y)-G(y)| \le (||K||/c)^{1/2}+C/c, \tag{2.4}$$

where $C \ge \sup_{x,y}|G'(y|x)|\sup_x|g'(x)|(\int_{-1}^1 u^2K(u)du)$, *and* $||K|| = \sup_u|K(u)|$.

The proof of Theorem 2.1 is based on three lemmas given below. Express $\widehat{G}(y)-G(y)$ as the sum of the following four terms:

$$Q_N(y) = N^{-1}\sum_{i=1}^N\{I(Y_i\le y)-G(y)\}, \tag{2.5}$$

$$R_N(y) = N^{-1}\sum_{i=1}^N(1-\delta_i)\{G(y|X_i)-I(Y_i\le y)\}, \tag{2.6}$$

$$S_N(y) = N^{-1}\sum_{i=1}^N(1-\delta_i)\Big\{\sum_j W_j(X_i)[I(Y_j\le y)-G(y|X_j)]\Big\}, \tag{2.7}$$

$$T_N(y) = N^{-1}\sum_{i=1}^N(1-\delta_i)\Big\{\sum_j W_j(X_i)[G(y|X_j)-G(y|X_i)]\Big\}. \tag{2.8}$$

We first note that conditioned on all the X_i, (2.6) is a weighted empirical distribution function of independent summands with mean zero. It follows by a theorem of Singh (1975) that a logarithmic upper inequality holds for $\sup_y|R_N(y)|$ via an exponential bound that is independent of the distribution of X_i. This, together with the classical law of the iterated logarithm, implies the first lemma below.

Lemma 2.1. *The classical* LIL *applies to* $\sup_y|Q_N(y)|$. *Further, assume* (2.1); *then for some positive constant* C, $\limsup_{N\to\infty}(N/\log N)^{1/2}\sup_y|R_N(y)| \le C$ *wp* 1.

The rate in Lemma 2.1 may not apply to (2.7) and (2.8); however, we will show that a slower uniform convergence rate can be obtained also via Singh's theorem. A useful preliminary fact is first explained here. Define $\widehat{g}(X_i) = \sum_{j\ne i}K_b(X_j-X_i)\delta_j/(N-1)$, and write $W_j(X_i) = K_b(X_j-X_i)\delta_j/((N-1)\widehat{g}(X_i))$.

A standard proof, using Bernstein's inequality (Serfling (1980) p.95) and Conditions (1) to (3), in the literature of kernel density estimation (cf. Cheng and Cheng (1990) for Lipschitz-continuous kernels, and Härdle, Janssen and Serfling (1988) for step function kernels) asserts that

$$\sum_{N\geq 1}(N^2/b)P[\sup_x|\widehat{g}(x)-E[\widehat{g}(x)]|>a(\log N/(Nb))^{1/2}]<\infty,$$

for some $a>1$. On the other hand, it routinely follows from Condition (1) that $|E[\widehat{g}(x)]-g(x)|\leq Cb$ *wp* 1, for $x\in\mathcal{T}$, the region interior to the support of X; but this fails for $x\in\mathcal{B}$, the boundary region, which is within distance b from the endpoints of the support of X. However, for $x\in\mathcal{B}$, it can be checked that $E[\widehat{g}(x)]-g(x)\geq -||g'||b\,K_0^+-g(x)/2$, where $||g'||=\sup_x|g'(x)|$ and $K_0^+=|\int_{-1}^1 uK(u)du|$. Therefore, by Condition (1) and the fact that $\widehat{g}(x)=\{\widehat{g}(x)-E[\widehat{g}(x)]\}+\{E[\widehat{g}(x)]-g(x)\}+g(x)$, we have

$$\widehat{g}(X)\geq -a(\log N/(Nb))^{1/2}-||g'||bK_0^+ +g(X)/2\geq c+o(1) \tag{2.9}$$

wp 1, since $g(X)\geq 2c$. Thus, $S_N(y)$ and $T_N(y)$ are almost surely evaluated over the set $\cap_i\{\widehat{g}(X_i)\geq c+o(1)\}$ eventually. This preliminary fact will be used to establish the following two lemmas concerning the strong uniform consistency of S_N and T_N.

Lemma 2.2. *Assume Conditions* (1) *to* (4); *then each summand of* (2.7) *satisfies* (2.4), *and so does* S_N.

Proof. Consider the ith summand of (2.7). Given $X_i=x$, it is clear that

$$\sum_{j=1}^N W_j^2(x)\leq \max_j W_j(x)I[\widehat{g}(x)\geq c]\leq ||K||/(cNb) \tag{2.10}$$

wp 1, since $\sum_{j=1}^N W_j(x)=1$ over the set $[\widehat{g}(x)\geq c]$ by (2.9). On the other hand, there is a set of positive measure where K is greater than some positive constant by Condition (2). Thus, we may assume, without loss of generality, that K is unimodal and that $K(u)\geq K(l)>0$ for all u, $0<|u|\leq l$, for some $0<l<1$. Then, an opposite inequality

$$\sum_{j=1}^N W_j^2(x)\geq (N/b)\Big\{\frac{1}{Nb}\sum_{j=1}^N K^2\Big(\frac{X_j-x}{b}\Big)\delta_j I[|X_j-x|\leq l]\Big\}/(N^2||\widehat{g}||^2)\geq C/(Nb) \tag{2.11}$$

holds wp 1 as N large. This is due to the classical strong law that the brace in the middle of (2.11) converges to $\int_{|u|\leq l}K^2(u)g(x+bu)du\geq 2c\int_{|u|\leq l}K^2(u)du$ wp

1, and that $||\widehat{g}||^2 \le ||g||^2 + o(1) < \infty$ wp 1, eventually. As a consequence of (2.10) and (2.11), we find, by Singh's (1975) theorem, that

$$\begin{aligned} &P\Big\{\sup_{y\in\mathbb{R}} |\sum_j W_j(X_i)[I(Y_j \le y) - G(y|X_j)]| > t|X_i = x\Big\} \\ &\le C(Nb\log N)^{1/2}\exp\{-2c(Nb)t^2/||K||\}, \end{aligned}$$

where $C \le 4e^2(||K||/c)^{1/2}$, $t \simeq (||K||\log N/(cNb))^{1/2}$. Obviously, the r.h.s. above is also an upper bound to the r.h.s. of the following inequality:

$$\begin{aligned} &P\Big\{\sup_{y\in\mathbb{R}} |\sum_j W_j(X_i)[I(Y_j \le y) - G(y|X_j)]| > t\Big\} \\ &\le E_x\Big(P\Big\{\sup_{y\in\mathbb{R}} |\sum_j W_j(X_i)[I(Y_j \le y) - G(y|X_j)]| > t|X_i = x\Big\}\Big), \end{aligned}$$

which is summable in N by Condition (3). This proves Lemma 2.2.

By analogous arguments, we shall prove that each term in the braces of (2.8) satisfies (2.4).

Lemma 2.3. *Assume Conditions* (1) *to* (4). *Then* $\limsup_{N\to\infty}(1/b^2)\sup_y |T_N(y)| \le C/c$ *wp* 1, *where the constant* C *is defined by* (2.4).

Proof. It suffices to show that wp 1

$$\limsup_{N\to\infty}(1/b^2)\sup_y\Big|\sum_j W_j(X_i)[G(y|X_j) - G(y|X_i)]\Big| \le C/c. \tag{2.12}$$

By (2.9), it suffices to establish the bound C for the numerator of the l.h.s. of (2.12). By Condition (4), this numerator is, ignoring the remainder term in a Taylor's expansion, bounded by

$$\sup_y |G'(y|x)|\ |N^{-1}\sum_{j\ne i} K_b(X_j - X_i)\delta_j(X_j - X_i)|,$$

where the first factor is bounded due to Condition (4). Next, conditioned on the X_i, the second factor is an average of i.i.d. bounded r.v.'s, which by the classical strong law converges almost surely to $E[K_b(X_j - X_i)\delta_j(X_j - X_i)]$. This mean is bounded by $b^2||g'||\int_{-1}^1 u^2K(u)du$ in view of Conditions (1) and (2). By the bandwidth choice of Condition (3), (2.12) is verified, and so is Lemma 2.3. The proof of Theorem 2.1 is now complete.

Remark 2.1. When X is $\mathbb{R}^d$-valued, $d > 1$, it is routinely defined in (2.2) that $K_b(X_k - X_i) = b^{-d}K((X_k - X_i)/b)$. In this case, it can also be checked that Lemmas 2.2 and 2.3 are valid, if Conditions (1) to (4) are modified to be Conditions (1′) to (4′), respectively.

(1′) g is bounded away from 0, and has bounded partial derivatives (within the compact support of X) up to order $k-1$, where k is even and $2 \le k \le d$;
(2′) K is a kernel of order k, e.g. $K = \prod_{i=1}^{d} K_0$, and the one-dimensional symmetric kernel K_0 satisfies $\int_{-1}^{1} u^l K_0(u)du = 1$, if $l = 0$; $= 0$, if $1 \le l \le k-1$; $\neq 0$ for $l = k$;
(3′) $Nb^{d+2k} \simeq \log N$ as $N \to \infty$, where $d < 2k$;
(4′) $G(y|x)$ has bounded partial derivatives with respect to x (within the support of X) up to order $k-1$.

Remark 2.2. Suppose we wish to estimate the quantiles of the distribution of the incomplete data Y. Let ξ_q be the unique solution y satisfying $G(y-) \le q \le G(y)$, $0 < q < 1$. Assume all the conditions of Theorem 2.1. Then it follows from (2.4) that for some positive constant C

$$\limsup_{N\to\infty} (Nb/\log N)^{1/2} |\widehat{G}^{-1}(q) - \xi_q| \le C \text{ wp } 1, \tag{2.13}$$

where $\widehat{G}^{-1}$ is the left-continuous sample quantile function derived from $\widehat{G}$. For $\mathbb{R}^d$-valued X, the conditions of Remark 2.1 can be effectively employed. The proof is fairly routine and consequently omitted.

3. Mean Square Consistency and Asymptotic Normality

This section aims at establishing two asymptotic properties for the estimate $\widehat{G}(y)$ in a more general setting. We noted in Section 2 that the stringent part of Condition (1) (the function g is strictly greater than a positive constant over the compact support of X) was imposed only for the strong uniform convergence property. Indeed, this condition excludes possible radical fluctuations of the weights $W_j(X_i)$ by trimming off the events $[\widehat{g}(X_i) < c]$, for any small $c > 0$, in order to apply the inequality (2.9). A similar trimming method, using the events $[\widehat{f}(X) < c]$ with compactly-supported X, has been discussed by Härdle and Stoker (1989), and Härdle, Hart, Marron and Tsybakov (1992) in the study of a semiparametric model. In contrast to their trimming technique in the complete data case, it is remarkable that Condition (1) can be relaxed when the goal is to obtain mean square consistency and the asymptotic distribution for the incomplete data (2.1). Allowing noncompact support of X, we relax Condition (1) but simply retain its smoothness requirement;
Condition (1*): g has bounded partial derivatives up to the order specified by (1′).

For ease of exposition, we consider only scalar-valued X. Let $c = c(n) \sim b^t$, $0 < t \le 1/4$, be a decreasing sequence of constants such that the corresponding sets $\{x : g(x) < c\}$ have small probability contents that tend to zero as n increases. Define the sets $A_c = \{g(X) \ge c\}$, the trimming events $\widehat{A}_{ic} = \{\widehat{g}(X_i) \ge c\}$,

and the associated events $A_{ic} = \{g(X_i) \geq c\}$. To implement the trimming, tentatively multiply each summand of S_N and T_N (in (2.7) and (2.8)) by $I(\widehat{A}_{ic})$, and set the remaining ones with factors $\{1 - I(\widehat{A}_{ic})\}$ equal to zero. For notational simplicity, let us redefine S_N of (2.7) to be the trimmed version $S_N(y) = N^{-1}\sum_{i=1}^N \varphi_i(y)/\widehat{g}(X_i)$ just specified, and let $S_N^*(y) = N^{-1}\sum_{i=1}^N \varphi_i(y)/g(X_i)$ be the untrimmed counterpart, where

$$\varphi_i(y) = (1-\delta_i)\Big\{\sum_j \frac{1}{(N-1)}\delta_j K_b(X_j - X_i)[I(Y_j \leq y) - G(y|X_j)]\Big\}. \tag{3.1}$$

Analogous terms T_N and T_N^* are likewise defined as in (2.8). Clearly, it suffices to justify our modified trimming method by establishing the asymptotic mean square equivalence between S_N and S_N^*. The analogous proof for T_N and T_N^* will be omitted. We begin with the basic Lemma 3.1, from which the mean square consistency of $\widehat{G}$ follows. The proof of Lemma 3.1 will be sketched in the Appendix. Here, we need to impose a natural condition that is void in the complete data case;

Condition (5): $E[\sigma^2(y|X)/p(X)] < \infty$, and $E[\{1-p(X)\}\{G'(y|X)/g(X)\}^2] < \infty$.

Lemma 3.1. *Assume Conditions* (1*) *and* (2) *to* (5). *Then* $E(S_N^*)^2 = 4\zeta_1/N + O(1/N^2 b)$ *and* $E(T_N^*)^2 = O(b^4)$, *where* $\zeta_1 = E[\{1-p(X)\}^2\sigma^2(y|X)/p(X)]/4 + O(b)$.

We now establish the asymptotic mean square equivalence between the trimmed S_N and S_N^* (suppressing a similar argument for T_N and T_N^*) so as to justify the trimming method. In the derivation of $E(S_N^*)^2$ (Lemma 3.1, see also Appendix), we find, by inserting the factor $[I(A_c) - 1]$ into the expectation of ζ_1, that

$$E\Big\{\frac{1}{N}\sum_{i=1}^N (1-\delta_i)\Big[\sum_{j\neq i}^N \Big(\frac{K_b(X_j - X_i)\delta_j\varepsilon_j}{(N-1)g(X_i)}\Big)\Big]\cdot[I(A_{ic}) - 1]\Big\}^2 \to 0, \tag{3.2}$$

where $\varepsilon_j \equiv I(Y_j \leq y) - G(y|X_j)$. Referring to the proof of (2.9), it follows that for some $a > 1$

$$\sum_{N=1}^{\infty}(N^2/b^t)P[\sup_{x\in A_c}|\widehat{g}(x) - E[\widehat{g}(x)]| > a(\log N/Nb^{1+t})^{1/2}] < \infty. \tag{3.3}$$

Then, by the uniform integrability of S_N^* (due to Lemma 3.1), (3.3), $\sup_x |E[\widehat{g}(x)] - g(x)| \leq Cb$ wp 1, and $[g(X)I(\widehat{A}_c)/\widehat{g}(X) - I(A_c)] \to 0$ wp 1 (indeed, $I(\widehat{A}_c) - I(A_c) \to 0$ wp 1), it is seen that

$$E\Big\{\frac{1}{N}\sum_{i=1}^N (1-\delta_i)[\sum_{j\neq i}^N \Big(\frac{K_b(X_j - X_i)\delta_j\varepsilon_j}{(N-1)g(X_i)}\Big)]\cdot[\frac{g(X_i)}{\widehat{g}(X_i)}I(\widehat{A}_{ic}) - I(A_{ic})]\Big\}^2 \to 0, \tag{3.4}$$

where we have used the condition $0 < t \leq 1/4$ together with a condition yet unspecified;

Condition (3*): $\log N = O(Nb^2)$.

It follows from (3.2) and (3.4) that the contribution to the asymptotic mean square (AMS) by the part with factors $\{1 - I(\widehat{A}_{ic})\}$ (being trimmed off) actually converges to 0, and the complementary part with factors $I(\widehat{A}_{ic})$ (being kept) eventually contributes the total AMS. The asymptotic mean square equivalence between the trimmed S_N and S_N^*, and hence the modified trimming, is now justified.

As a result of the mean square consistency, we have the following asymptotic normality.

Theorem 3.1. *Assume* (2.1), *conditions* (1*), (2), (3*), (4), *and* (5). *Assume also* $Nb^4 \to 0$ *as* $N \to \infty$. *Suppose that the weights of* $\widehat{G}$ *are defined by a trimming sequence* $c(n) \simeq b^t$, $0 < t \leq 1/4$. *Then, for each scalar* y, $N^{1/2}[\widehat{G}(y) - G(y)]$ *converges to a normal distribution with mean* 0 *and variance*

$$\sigma^2(y) = E[\sigma^2(y|X)/p(X)] + E[G^2(y|X)] - G^2(y). \tag{3.5}$$

Proof. By (2.5) to (2.8), Lemma 3.1 and the foregoing mean square equivalence, it suffices to consider $\widehat{G}(y) - G(y) \simeq Q_N + R_N + S_N^* + T_N^*$, suppressing the arguments y. First, it is straightforward that $N^{1/2}Q_N$ and $N^{1/2}R_N$ converge to the normal distributions with means 0, and variances $G(y) - G^2(y)$ and $E[\{1 - p(X)\}\sigma^2(y|X)]$, respectively. Clearly, $2\,\mathrm{Cov}(N^{1/2}Q_N, N^{1/2}R_N) = -2E[\{1 - p(X)\}\sigma^2(y|X)]$. It remains to consider the sum of

$$S_N^* = N^{-1}\sum_{i=1}^{N}(1 - \delta_i)\Big\{\sum_{j\neq i}\delta_j K_b(X_j - X_i)[I(Y_j \leq y) - G(y|X_j)]\Big\}\Big/g(X_i),$$

and

$$T_N^* = N^{-1}\sum_{i=1}^{N}(1 - \delta_i)\Big\{\sum_{j\neq i}\delta_j K_b(X_j - X_i)[G(y|X_j) - G(y|X_i)]\Big\}\Big/g(X_i).$$

By analogy with (5.1) in the appendix, it follows from Lemma 3.1 that $N^{1/2}S_N^*$ converges to a normal distribution with mean 0 and variance $E[\{1 - p(X)\}^2\sigma^2(y|X)/p(X)]$. Likewise, for T_N^*, we find by standard U-statistics arguments that $E[N^{1/2}(T_N^* - \widehat{T}_N)]^2 = O(1/N)$, where the projection $\widehat{T}_N$ satisfies that $\widehat{T}_N = O(b^2) = o(1/N^{1/2})$ wp 1 by Condition (3*). Consequently, $N^{1/2}T_N^* \to 0$ in probability. To compute the covariances between S_N^* and $Q_N + R_N$, it suffices, by the proof of Lemma 3.1, to consider the projection $\widehat{U}_N = 2\sum_{i=1}^{N} H_1(Z_i)/N$, noting

that $H_1(Z_i) = V_i\{1-p(X_i)\}/(2p(X_i)) + O(b)$ with $V_i = \delta_i[I(Y_i \le y) - G(y|X_i)]$. Therefore, $2\,\mathrm{Cov}(N^{1/2}Q_N, N^{1/2}S_N^*) = 2E[\{1-p(X)\}\sigma^2(y|X)]$, and $\mathrm{Cov}(N^{1/2}R_N, N^{1/2}S_N^*) = 0$ because $\delta_i(1-\delta_i) = 0$ and $E[V_i|X_i] = 0$. It is then easy to see that $\mathrm{Var}\{N^{1/2}(Q_N + R_N + S_N^*)\} = \sigma^2(y)$ given by (3.5). The proof of Theorem 3.1 is now complete.

Remark 3.1. Since $E\{N^{1/2}[\widehat{G}(y) - G(y)]\}^2$ is bounded, it follows from *condition* (3^*) that the optimal choice of the bandwidth b can only be determinded by the second-order term (of magnitude $o(1/N)$) of the mean squared error $E[\widehat{G}(y) - G(y)]^2$. Let $K_1 = \int_{-1}^{1} K^2(u)du$ and recall $K_2 = \int_{-1}^{1} u^2K(u)du$ for notational convenience. It follows by Lemma 3.1 that

$$\begin{aligned} &E[\widehat{G}(y) - G(y)]^2 \\ &= \frac{1}{N}\Big\{\mathrm{Var}[G(y|X)] + E[\sigma^2(y|X)/p(X)]\Big\} + (\theta_N)^2 + 2\zeta_2/N^2 + O(1/N^2), \end{aligned} \tag{3.6}$$

where the first summand on the r.h.s. is equal to $\sigma^2(y)/N$ by (3.5); further, $2\zeta_2/N^2 = K_1e_1(N^2b)^{-1}$, $(\theta_N)^2 = [E\widehat{G}(y) - G(y)]^2 = (K_2e_2b^2)^2$, where $e_1 = E[\{1-p(X)\}\sigma^2(y|X)/g(X)]$, and $e_2 = E\{[p'(X) + (1-p(X))(g'/g)(X)]G'(y|X)\}$. Thus, the asymptotic minimum mean squared error is found to be

$$\sigma^2(y)/N + (K_1e_1)^{4/5}(K_2e_2)^{2/5}\Big/(8^{1/5}N^{8/5}) + O(1/N^2), \tag{3.7}$$

which obtains when the theoretically optimal bandwidth $b_{opt} = (K_1e_1)^{1/5} \times (2K_2e_2)^{-2/5}N^{-2/5}$ is implemented. It is remarkable, *perhaps surprising*, that the local linear regression smoother discussed by Fan (1992) would lead to the same formulae (3.6) and (3.7); that is, the advantage of the local linear smoother in reducing the bias θ_N^* (in fact, e_2) directly to a quadratic term, does *not* apply to the incomplete data case when the indicators δ_i are involved. On the other hand, neither could the well-known Gasser-Müller (1979) regression estimator yield the same formulae (in the random design case) due to creating a larger asymptotic variance compared to the Nadaraya-Watson (1964) estimator (see the discussions by Mack and Müller (1988), and Chu and Marron (1991)). Adapting b_{opt} will involve estimating e_1 and e_2, that requires estimating p', f' and G'. This extra task can also be done by using some U-statistics defined through some formulae comparable to (2.2). For brevity, adaptation of b_{opt} is not discussed here because, it does not yield any particular advantage, as evidenced by the flexible choices of b in the simulation study of Section 4.

We conclude this section by complementing Remark 2.2 with the asymptotic normality of the sample quantile $\widehat{G}^{-1}(q)$.

Theorem 3.2. *Assume in addition to the conditions of Theorem* 3.1 *that G has a density $G^{(1)}$ which is positive and continuous at ξ_q. Then* $N^{1/2}(\widehat{G}^{-1}(q) - \xi_q)$

converges to a normal distribution with mean 0 *and variance* $(\sigma(\xi_q)/G^{(1)}(\xi_q))^2$, *where* $\sigma^2(\cdot)$ *is defined by* (3.5).

Proof. Fix a real t. Let $\Phi(t)$ be the standard normal cdf, and $A > 0$ be a normalizing constant to be specified later. Letting $\xi_{q,N} = \xi_q + tAN^{-1/2}$ and $\widehat{\xi}_q = \widehat{G}^{-1}(q)$, we have

$$P[\,N^{1/2}(\widehat{\xi}_q - \xi_q)A^{-1} \le t\,] = P(\widehat{\xi}_q \le \xi_{q,N}) = P[q \le \widehat{G}(\xi_{q,N})] \tag{3.8}$$
$$= P[\,N^{1/2}(q - G(\xi_{q,N}))\sigma(\xi_{q,N})^{-1} \le N^{1/2}(\widehat{G}(\xi_{q,N}) - G(\xi_{q,N}))\sigma(\xi_{q,N})^{-1}\,].$$

By (3.5) and the Lebesgue dominated convergence theorem, $\sigma(\xi_{q,N}) \to \sigma(\xi_q)$. Thus $N^{1/2}\{G(\xi_{q,N}) - q\}/\sigma(\xi_{q,N})$ converges to $tAG^{(1)}(\xi_q)/\sigma(\xi_q)$ which equals t if $A = \sigma(\xi_q)/G^{(1)}(\xi_q)$. Therefore, the limit in N of (3.8) equals $1 - \Phi(-t)$ by Theorem 3.1. The proof is complete.

4. A Simulation Study

In this section we carry out a study comparing the performance of three distribution function estimators $\widehat{G}$, G_s and G_p given the missing data pattern (1.1). Here, $\widehat{G}$ is the estimator of (2.3), G_s is the basic naive pairwise-deletion estimator defined by

$$G_s(y) = \sum_{i=1}^{N} \delta_i I(Y_i \le y) / \sum_{i=1}^{N} \delta_i,$$

and G_p is a standard linear-regression estimator defined by

$$G_p(y) = N^{-1} \sum_{i=1}^{N} [\delta_i I(Y_i \le y) + (1 - \delta_i) I(Y_i^* \le y)],$$

where $Y_i^* = \overline{Y} + (S_{xy}/S_{xx})(X_i - \overline{X})$, and $\overline{X}$, $\overline{Y}$, S_{xx} and S_{xy} denote the sample averages, variance and covariance, respectively, of the sub-sample of the complete pairs (cf. Little and Rubin (1987) Chapter 6). It is worth noting that $G_s(y)$ is a consistent estimate of $E[p(X)G(y|X)]/p$ by assumption (2.1), which is not equal to $G(y)$ in general unless $p(x)$ identically equals the constant p. For the kernel estimator $\widehat{G}$, we have tested some symmetric pdf kernels in computations, including the biweight kernel and the Epanechnikov (1969) kernel $K(u) = \frac{3}{4}(1 - u^2)$. Results from the latter kernel, showing no noticeable difference from the former, will be reported below (see Silverman (1986) and Müller (1988) for discussions on the choices of multivariate kernels in case of multivariate X). It is worth noting that we did not use the Gasser-Müller (1979) estimator as a competitor to the Nadaraya-Watson (1964) estimator, because the former has larger asymptotic variance and a comparable bias property (see Remark 3.1, and Chu and Marron, 1991).

The report below presents a typical simulation study where a parametric linear regression model holds. Two hundred data sets, each of sample size $N = 50$, were generated from the bivariate normal distribution $((0,0), (1,1), \rho = 0.8)$. The ignorable missing mechanism was generated by $p(x) = I(x \leq 0) + 0.3 \cdot I(x > 0)$ with $p = Ep(X) = 0.65$, which is specially selected to violate Condition (1^*) of Section 3. Our unreported extensive simulations indicate that a moderate jump discontinuity of $p(x)$ in the interior of the support of X usually has little effect on the performance of $\widehat{G}$. In the sparse data regions, either in the tails of the joint distribution (of X and Y) or due to heavy missing, the performance of $\widehat{G}$ may be unsatisfactory. A modified version of $\widehat{G}$ (see (4.1) below) employing the trimming of Section 3 is suggested for improvement.

The results of the present simulation study are exhibited in Figures 1 and 2, and Table 1. In Figures 1 and 2, plots of the bias and the variance of four estimators are depicted at values $y = -4.0$ (0.2) 4.0, respectively. Dotted curves are for G_s, starred ones for G_p, solid ones for $\widehat{G}$ using $c = 0$, and dashed ones for a modified $\widehat{G}$ using $c = 0.02$. The bandwidths tested for $\widehat{G}$ include $b = 0.1$ (0.1) 1.2. The case $b = 0.4$ is reported, since similar performances are obtained within a wide range: $0.2 \leq b \leq 0.9$. Likewise, for the modified $\widehat{G}$, values 0.01 (0.01) 0.05 for c also provided similar behavior in the current normal linear regression example. Table 1 presents three measurements of sampling variation for the four estimators in contrast to Figures 1 and 2. They are the means and the variances of the Kolmogorov-Smirnov distances, and the mean integrated square errors (MISE) defined by, say, $\mathrm{MISE}(\widehat{G}) = E[\int_{-\infty}^{\infty} \{\widehat{G}(y) - G(y)\}^2 dy]$.

To reduce the effect of sparse data, the modified $\widehat{G}$ above-mentioned goes one step beyond the trimming of Section 3. Consider the following weighting design that modifies (2.2) by

$$\widehat{W}_j(X_i) = W_j(X_i) I[\hat{g}(X_i) \geq c] + W_j^*(X_i) I[\hat{g}(X_i) < c], \tag{4.1}$$

where equal weights $W_j^*(X_i) = 1/[c^* N b]$, for some $c^* > 0$, are assigned to each member of the set $\{Y_j : \delta_j = 1$, and the concomitant X_j is one of the $[c^* N b]$ nearest neighbors to $X_i\}$. Intuitively speaking, (4.1) is designed to help reduce the bias in the sparse data area $\{g(X) < c\}$. In this simulation example, proper values for c^* that are relatively larger than c also form a small flexible range $c^* = 0.06$ (0.01) 0.15 where stable bias performance is obtained. Thus, we merely report the value $c^* = 0.1$. However, more study concerning proper combination choices of the values b, c and c^* need to be carried out to make (4.1) or any analog a more practical method.

In summary, we find from the means, the standard deviations, and the MISE of the Kolmogorov-Smirnov distances in Table 1 that G_p has the least sampling variation and MISE, G_s is the worst by these measurements, and both $\widehat{G}$ perform fairly well. On the other hand, Figures 1 and 2 indicate that both $\widehat{G}$ still have

unsatisfactory sample variances in some regions of the y values, whereas the modified $\widehat{G}$ has the best bias performance among all. Although it is difficult, if not impossible, to cover a wide range of simulations combining general regression models and missing data patterns, the modified $\widehat{G}$ appears to be a reasonably good nonparametric alternative to G_p without assuming a parametric model. Further study needs to be done for the general cases, especially for nonrandom missing data.

Table 1. The sample mean and standard deviation (SD) of the Kolmogorov-Smirnov distance and mean integrated square error (MISE) of $G_s(y)$, of $G_p(y)$, of $\hat{G}(y)$ with $b = 0.4$ and $c = 0$, and of $\hat{G}(y)$ with $b = 0.4$, $c = 0.02$ and $c^* = 0.1$, over the 200 simulated data sets of sample size $N = 50$. The MISE of $\hat{G}$ is defined by $\text{MISE}(\hat{G}) = E[\int(\hat{G}(y) - G(y))^2 dy]$, and those of G_s and G_p are defined similarly.

	Mean	SD	MISE
$\hat{G}(y), b = 0.4, c = 0$	0.1652	0.0515	0.0034
$\hat{G}(y), b = 0.4, c = 0.02, c^* = 0.1$	0.1710	0.0558	0.0029
$G_s(y)$	0.2276	0.0744	0.0067
$G_p(y)$	0.1502	0.0481	0.0024

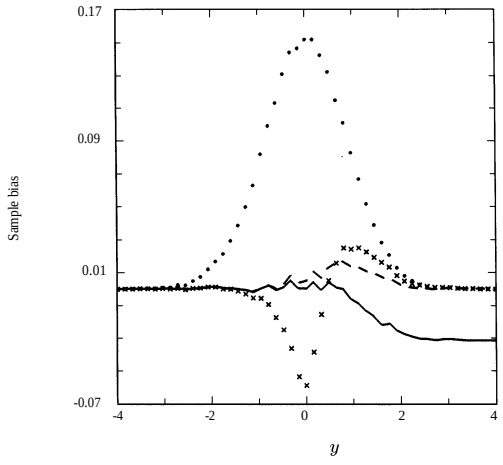

Figure 1. Plot of the sample biases of $\widehat{G}(y)$ with $b = 0.4$ and $c = 0$ (solid curve), $\widehat{G}(y)$ with $b = 0.4$, $c = 0.02$ and $c^* = 0.1$ (dashed curve), $G_s(y)$ (dotted curve), and $G_p(y)$ (starred curve) over the 200 simulated data sets of sample size $N = 50$.

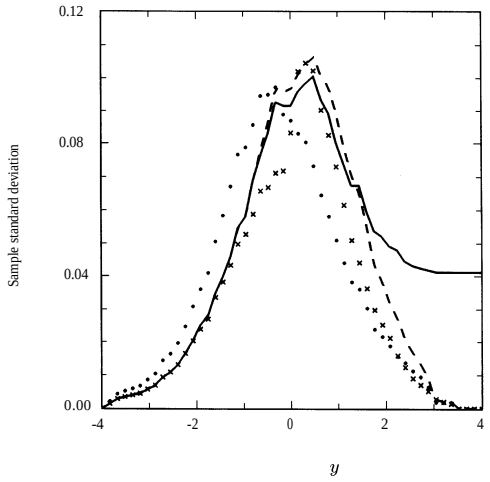

Fig. 2. Plot of the sample standard deviations of $\widehat{G}(y)$ with $b = 0.4$ and $c = 0$ (solid curve), $\widehat{G}(y)$ with $b = 0.4$, $c = 0.02$ and $c^* = 0.1$ (dashed curve), $G_s(y)$ (dotted curve), and $G_p(y)$ (starred curve) over the 200 simulated data sets of sample size $N = 50$.

Acknowledgments

The authors would like to thank the associate editor and the referees for insightful comments that have led to improvement in the presentation.

Appendix : Proof of Lemma 3.1.

For $S_N^*(y)$ defined by (3.1), let $V_i = \delta_i[I(Y_i \le y) - G(y|X_i)]$ for all i. Set $Z_i = (X_i, Y_i, \delta_i)$ and define for all pairs (i, j), $H(Z_i, Z_j) = K_b(X_j - X_i)\{(1 - \delta_j)V_i/g(X_j) + (1 - \delta_i)V_j/g(X_i)\}/2$, which is a symmetric function by Condition (2). Thus $S_N^*(y) = N^{-1}\sum_{i=1}^N \varphi_i(y)/g(X_i) = \sum_{i=1}^N \sum_{j\neq i} H(Z_i, Z_j)/N^2 = NU_N(y)/(N-1)$ (since $H(Z_i, Z_i) = 0$), where U_N is a standard U-statistic with a symmetric kernel of order 2.

Now, consider $U_N(y)$ for each y. We note that $E[H(Z_i, Z_j)] = 0 = E(V_i)$, and that for each i the conditional expectation $H_1(Z_i) \equiv E[H(Z_i, Z_j)|Z_i] = V_i\{1 - p(X_i)\}/(2p(X_i)) + O(b)V_i$ wp 1. A few useful moments for U_N are then computed:

$$\zeta_1 \equiv \mathrm{Var}\, H_1(Z_i) = E[\{1 - p(X)\}^2\sigma^2(y|X)/(4p(X))] + O(b),$$

and

$$\zeta_2 \equiv \mathrm{Var}\, H(Z_i, Z_j) = E[\{1 - p(X)\}\sigma^2(y|X)/g(X)](\int_{-1}^{1} K^2(u)du)/(2b) + O(1).$$

The U-statistic projection of U_N is $\widehat{U}_N = 2\sum_{i=1}^N H_1(Z_i)/N$, $E(\widehat{U}_N^2) = 4\zeta_1/N$, and $E(U_N^2) = \mathrm{Var}(U_N) = 4(N-2)\zeta_1/(N(N-1)) + 2\zeta_2/(N(N-1))$. Therefore, $E(U_N - \widehat{U}_N)^2 = 2\zeta_2/(N(N-1)) + O(1/N^2)$. It follows by a standard argument (Serfing, (1980) p. 189-192) that

$$N^{1/2}U_N \to \text{Normal } (0, 4\zeta_1), \tag{5.1}$$

which is a useful preliminary result for Theorem 3.1.

Similar analyses for $T_N^*(y)$ via U-statistic are now sketched. Like (3.1), (omitting the arguments y) we have the expression $T_N^* = \sum_{i=1}^N \sum_{j\neq i} L(Z_i, Z_j)/N^2$, where $Z_i = (X_i, \delta_i)$ and

$$L(Z_i, Z_j) = \Big\{ g(X_i)^{-1}(1-\delta_i)\delta_j K_b(X_j - X_i)[G(y|X_j) - G(y|X_i)]/2 \\ + g(X_j)^{-1}(1-\delta_j)\delta_i K_b(X_i - X_j)[G(y|X_i) - G(y|X_j)]/2 \Big\}.$$

Thus T_N^* is a U-statistic for estimating a bias quantity $\theta_N \equiv E[L(Z_i, Z_j)]$. Define $\widehat{T}_N = 2\sum_{i=1}^N E[L(Z_i, Z_j)|Z_i]/N - \theta_N = 2\sum_{i=1}^N L_1(Z_i)/N - \theta_N$ to be the projection of T_N^*, such that $E(\widehat{T}_N) = \theta_N$. By Conditions (1), (2) and (4), direct computation using the SLLN like lemma 2.3 leads to the fact that $L_1(Z_i) \leq Cb^2$ wp 1, for N large. A similar proof shows that $\theta_N = E[L_1(Z_i)] \leq Cb^2$, and $|\widehat{T}_N| \leq Cb^2$ wp 1, for N large.

Further, the bounds for the second moments are easily verified to be $\rho_1 \equiv E[L_1^2(Z_i)] \leq Cb^4$ and $\rho_2 \equiv E[L^2(Z_i, Z_j)] \leq C$. It follows that $\mathrm{Var}(\widehat{T}_N) = 4(\rho_1 - \theta_N^2)/N = O(b^4/N) = o(1/N^2)$ by Condition (3^*) specified in Theorem 3.1, and $\mathrm{Var}(T_N^* - \widehat{T}_N) = E[T_N^* - \widehat{T}_N]^2 = 2\rho_2/(N(N-1))$. The proof of Lemma 3.1 is now complete.

References

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

Cheng, P. E. and Cheng, K. F. (1990). Asymptotic normality for robust R-estimators of regression function. J. Statist. Plann. Inference **24**, 137-149.

Cheng, P. E. and Wei, L. J. (1986). Nonparametric inferences under ignorable missing data process and treatment assignment. 1986 International Statistical Symposium, Taipei Vol. 1, 97-111. Institute of Statistical Science, Academia Sinica, Taipei.

Chu, C.-K. and Marron, J. S. (1991). Choosing a kernel regression estimator. Statist. Sci. **6**, 404-419.

Cochran, W. G. (1963). Sampling Techniques, 2nd edition. John Wiley, New York.

Efron, B. (1994). Missing data, imputating, and the bootstrap. J. Amer. Statist. Assoc. **89**, 463-475.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. Theory Probab. Appl. **14**, 153-158.

Fan, J. (1992). Design-adaptive nonparametric regression. J. Amer. Statist. Assoc. **87**, 998-1004.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In Smoothing Techniques for Curve Estimation. Lecture Notes in Math. **757**, 23-68, Springer-Verlag, New York.

Härdle, W. (1990). Applied Nonparametric Regression. Econometric Society Monographs No. **19**, Cambridge University Press.

Härdle, W., Janssen, P. and Serfling, R. J. (1988). Strong uniform consistency rates for estimators of conditional functionals. Ann. Statist. **16**, 1428-1449.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. J. Amer. Statist. Assoc. **84**, 986-995.

Härdle, W., Hart, J., Marron, J. S. and Tsybakov, A. B. (1992). Bandwidth choice for average deriative estimation. J. Amer. Statist. Assoc. **87**, 218-226.

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

Kiefer, J. (1961). On large deviations of the empirical D. F. of vector chance variables and a law of iterated logarithm. Pacific J. Math. **11**, 649-660.

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

Mack, Y. P. and Müller, H.-G. (1988). Convolution type estimators for nonparametric regression. Statist. Probab. Lett. **7**, 229-239.

Müller, H.-G. (1988). Nonparametric Regression Analysis of Longitudinal Data. Springer-Verlag, Berlin.

Nadaraya, E. A. (1964). On estimating regression. Theory Probab. Appl. **9**, 141-142.

Neyman, J. (1938). Contribution to the theory of sampling human populations. J. Amer. Statist. Assoc. **33**, 101-116.

Rubin, D. B. (1976). Inference and missing data. Biometrika **63**, 581-590.

Rubin D. B. (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley, New York.

Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics. John Wiley, New York.

Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Singh, R. S. (1975). On the Glivenko-Cantelli theorem for weighted empiricals based on independent random variables. Ann. Probab. **3**, 371-374.

Titterington, D. M. and Mill, G. M. (1983). Kernel-based density estimates from incomplete data. J. Roy. Statist. Soc. Ser.B **45**, 258-266.

Watson, G. S. (1964). Smooth regression analysis. Sankhyā A **26**, 359-372.

Institute of Statistical Science, Academia Sinica, Taipei, 115, Taiwan.

Institute of Statistics, National Tsing Hua University, Hsinchu, Taiwan.

(Received February 1993; accepted June 1995)