Statistica Sinica **16**(2006), 1071-1100

LOCAL QUASI-LIKELIHOOD ESTIMATION WITH DATA MISSING AT RANDOM

Jianwei Chen, Jianqing Fan, Kim-Hung Li and Haibo Zhou

University of Rochester, Princeton University,
The Chinese University of Hong Kong and University of North Carolina

Abstract: Local quasi-likelihood estimation is useful for nonparametric modeling in a widely-used exponential family of distributions, called generalized linear models. Yet, the technique cannot be directly applied to situations where a response variable is missing at random. Three local quasi-likelihood estimation techniques are introduced: the local quasi-likelihood estimator using only complete-data; the locally weighted quasi-likelihood method; the local quasi-likelihood estimator with imputed values. These estimators share basically the same first order asymptotic biases and variances. Our simulation results show that substantial efficiency gains can be obtained by using the local quasi-likelihood estimator with imputed values. We develop the local quasi-likelihood imputation methods for estimating the mean functional of the response variable. It is shown that the proposed mean imputation estimators are asymptotically normal with asymptotic variance that can be easily estimated. Data from an ongoing environmental epidemiologic study is used to illustrate the proposed methods.

Key words and phrases: Bandwidth selection, generalized linear models, local imputation method, nonparametric regression, quasi-likelihood, the mean functional.

1. Introduction

Quasi-likelihood estimation is an important extension of maximum likelihood estimation. This method, proposed by Wedderburn (1974), requires only assumptions on the conditional mean and variance functions rather than on the full likelihood. McCullagh and Nelder (1989) extended the method to the analysis of parametric generalized linear models. The need to reduce possible modeling biases and to validate parametric models leads to the development of nonparametric models. See, for example, Hastie and Tibshrani (1990), Fan and Gijbels (1996) and Eubank (1999). There are a number of papers that study nonparametric function estimation in the context of generalized linear models. Methods for kernel estimation were discussed by Staniswalis (1989), Severini and Staniswalis (1994) and Hunsberger (1994). Fan, Heckman and Wand (1995) developed a local quasi-likelihood estimation via a local polynomial fitting. Carroll, Fan, Gujbels and Wand (1997) considered an extension of such estimators to the multivariate generalized partially linear single-index models. Those techniques allow one to analyze data from many useful families of distributions, including the Bernoulli and the Poisson distributions. Fan, Farmen and Gijbels (1998) proposed a pre-asymptotic substitution method for selecting bandwidths for the local quasi-likelihood estimator. The local estimating equations have been studied in Carroll, Ruppert and Welsh (1998).

The first objective of this paper is to study the local quasi-likelihood estimation when the response Y is missing at random (MAR). Data with missing outcome are common in medical research. For example, in a recent environmental epidemiologic study, the Collaborative Prenatal Projects (CPP) (Niswander and Gordon (1972) and Longnecker, Klebnoff, Zhou and Brock (2002)), the investigators are interested in the nonlinear relationship between the women's PCBs exposure (high or low) and the body mass index at the beginning of the pregnancy. There are about sixty percent of the women's PCBs exposures missing in the study. Our research is motivated by the need to develop a nonparametric estimation method for missing responses. Methods for regression analysis with missing data have been studied by many authors. For an introduction to these, see the books by Little and Rubin (1987) and Gelman, Carlin, Sterm and Rubin (1995), and the papers by Rosenbaum and Rubin (1983), Little (1992), Cheng (1994), Wang, Wang, Zhao and Ou (1997) and Wang and Rao (2002), among others.

In the context of the response variable missing at random, observations for which only covariates have been recorded are not informative about the quantities to be estimated. The methods based on only complete observations provide a valid and suitable analysis. Paik (1997) has shown that the imputation methods which impute missing values by regression fitting can provide a more efficient estimator if the model for the missing data is correctly specified. These techniques are extended to nonparametric models. Three estimation techniques − the local quasi-likelihood estimator using only complete-data, the locally weighted quasi-likelihood method, and the local quasi-likelihood estimator with imputed values − are employed. These techniques are complementary to those in Wang, Wang, Gutierrez and Carroll (1998), which generalized the local linear estimation of Fan, Heckman and Wand (1995) to the case with covariates missing at random. They found that the variances of the locally weighted method are the same when the selection probabilities are either known or unknown, but the biases are different. In the context where the responses are missing at random, we show that the three proposed methods basically share the same asymptotic bias and variance. Yet, finite sample simulations show that the local quasi-likelihood estimator with imputed data substantially outperforms the two other proposals. This finding is consistent with the results in parametric modeling by Paik (1997). Our extensive simulations show that the gains from imputed data are even more substantial than for two other estimators when the bandwidths are automatically selected by data. In fact, with imputed data, there are larger sample sizes and the selected bandwidths are more stable (most bandwidth selection procedures involve some degrees of estimating higher order derivatives, which requires a large amount of data to stabilize the estimate). This leads to more stabilized nonparametric estimators, resulting in smaller variances, than the two completing methods, and hence substantially improves the performance. It should be pointed out that the improvement comes from the fact that the imputation makes a better choice of the optimal bandwidth $-$ it makes more use of the complete observed data in the local fitting.

The second objective of our study is to estimate the mean functional of the response variable when it is missing at random. Tamhane (1978) and Matloff (1981) considered hypothesis testing and estimation of the mean functional with a specified regression function. Cheng (1994) discussed nonparametric estimation of the mean functional by using the kernel regression estimator in this MAR setting. Wang, Linton and Härdle (2004) developed the kernel regression estimator of the mean functional in a semiparametric partially linear regression model. In this paper, local quasi-likelihood estimators are used to impute each missing datum. We develop two quasi-likelihood imputation estimators for the mean functional of the response. The corresponding quasi-likelihood weight imputation estimators are also investigated. The proposed imputation estimators are shown to be asymptotically normal, with an asymptotic variance that can be easily estimated.

The paper is organized as follows. Section 2 introduces the local quasi-likelihood estimator with the complete-case data, the locally weighted quasi-likelihood method and the locally quasi-likelihood imputation method. The asymptotic properties of three methods are studied. The nonparametric quasi-likelihood imputation estimators of the mean functional of the response variable are studied in Section 3. Section 4 shows how to access the bias and variance of the local quasi-likelihood estimators and discusses how to select data-driven bandwidths. In Section 5, we present results from simulation studies comparing the proposed estimators. The methods are demonstrated with a data set from the Collaborative Perinatal Projects in Section 6. Technical proofs are relegated to the Appendix.

2. Model and Methodology

2.1. The models

Let $(X_1, Y_1, \delta_1), \ldots, (X_n, Y_n, \delta_n)$ be a set of independent random variable where, for each i, $\delta_i = 1$ if the Y_i is observed and $\delta_i = 0$ otherwise, and X_i is an observed covariate having density function f. Furthermore, assume that the selection probability is $P(\delta_i = 1|Y_i, X_i) = P(\delta_i = 1|X_i) \equiv \pi(X_i) > 0$. Suppose that the conditional mean and conditional variance of Y given X are

$$E(Y|X = x) = m(x), \qquad \mathrm{Var}\,(Y|X = x) = \sigma^2 V\{m(x)\}, \tag{2.1}$$

for a given function V and an unknown scale parameter σ^2. In (2.1), only the relationship between the conditional mean and the conditional variance is given. Therefore, it is appropriate to apply the quasi-likelihood method. The quasi-likelihood function $Q(\mu, y)$ is defined via

$$\frac{\partial}{\partial\mu} Q(\mu, y) = \frac{y - \mu}{V(\mu)}, \tag{2.2}$$

and the ith data point contributes $Q\{m(X_i), Y_i\}$ to the quasi-likelihood. When all Y's are observed, one can estimate the parameters in the conditional mean via maximizing the quasi-likelihood.

A specific case of (2.1) is the generalized linear model (McCullagh and Nelder (1989)), which assumes that the conditional density of Y given $X = x$ belongs to a canonical exponential family

$$f_{Y|X}(y|x) = \exp\Big(\frac{\Theta(x)y - b\{\Theta(x)\}}{a(\phi)} + c(y, \phi)\Big), \tag{2.3}$$

for some known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$. Here the unknown parameter $\Theta(\cdot)$ is called the canonical parameter and ϕ is called the dispersion parameter. Note that

$$m(x) = E(Y|X = x) = b'\{\Theta(x)\} \;\text{ and }\; \mathrm{Var}\,(Y|X = x) = a(\phi)b''\{\Theta(x)\}. \tag{2.4}$$

Thus (2.3) satisfies (2.1). For the exponential family of models, the quasi-likelihood (2.2) is just the conditional log-likelihood of $(Y_1, \ldots, Y_n)$ given $(X_1, \ldots, X_n)$. Thus, the quasi-likelihood approach is an extension of the likelihood method. If $g = (b')^{-1}$, then g is called the canonical link function. Our interest is to estimate, nonparametrically, the function $\eta(x) = g\{m(x)\}$. A local quasi-likelihood estimator of $\eta(x)$ is given by Fan, Heckman and Wand (1995). The goal of this section is to extend their technique to handle cases with missing data.

2.2. The local quasi-likelihood estimation with the complete-case data

We first study the local quasi-likelihood estimator of $\eta(x)$ basing on the complete-case data $\{(X_i, Y_i) : \delta_i = 1,\ i = 1, \ldots, n\}$. Assume that η possesses $p + 1$ derivatives. For each x, approximate the function locally by

$$\eta(z) \approx \beta_0 + \cdots + \beta_p(z - x)^p, \tag{2.5}$$

for z in a neighborhood of the point x. Following Fan, Heckman and Wand (1995), we construct the local quasi-likelihood for the complete-case data:

$$\ell_C(\beta) \equiv \sum_{i=1}^{n} \delta_i Q[g^{-1}\{\beta_0 + \cdots + \beta_p(X_i - x)^p\}, Y_i]K_{h_1}(X_i - x), \tag{2.6}$$

where $K_{h_1}(\cdot) = K(\cdot/h_1)/h_1$ with K a kernel function and h_1 a bandwidth. Let $\hat{\beta}_C(x) = (\hat{\beta}_{0,C}(x), \ldots, \hat{\beta}_{p,C}(x))^T$ maximize (2.6). Then, the maximum local quasi-likelihood estimator of $\eta^{(\upsilon)}(x)$ with the complete-case data is $\widehat{\eta}_{\upsilon,C}(x) = \upsilon!\widehat{\beta}_{\upsilon,C}(x)$ for $\upsilon = 0, \ldots, p$, with the convention $\widehat{\eta}_C(x) = \widehat{\eta}_{0,C}(x)$.

2.3. The locally weighed quasi-likelihood estimation

An alternative approach to handling missing data is a locally weighted quasi-likelihood estimation. Similar to the discussion in Section 2.2, a locally weighted quasi-likelihood can be defined as

$$\ell_W(\beta) \equiv \sum_{i=1}^{n} \frac{\delta_i}{\pi(X_i)} Q[g^{-1}\{\beta_0 + \cdots + \beta_p(X_i - x)^p\}, Y_i]K_{h_1}(X_i - x), \tag{2.7}$$

where the weight $\pi(x)$ is the selection probability defined in Section 2.1. Let $\hat{\beta}_W(x, \pi) = (\hat{\beta}_{0,W}(x), \ldots, \hat{\beta}_{p,W}(x))^T$ maximize (2.7). Then, the maximum locally weighted quasi-likelihood estimator $\eta^{(\upsilon)}(x)$ can be expressed as $\widehat{\eta}_{\upsilon,W}(x, \pi) = \upsilon!\widehat{\beta}_{\upsilon,W}(x, \pi)$ for $\upsilon = 0, \ldots, p$, with the convention $\widehat{\eta}_W(x, \pi) = \widehat{\eta}_{0,W}(x, \pi)$.

Note that the selection probability in (2.7) is regarded as known. If the selection probability is unknown, it can be estimated by a kernel smoothing method. In that case, an estimated locally weighted quasi-likelihood estimator $\hat{\beta}_W(x, \hat{\pi})$ can be obtained by replacing $\pi(X)$ by its estimator $\hat{\pi}(X)$ in (2.7). It is clear that the locally weighted quasi-likelihood estimators $\hat{\beta}_W(x, \pi)$ and $\hat{\beta}_W(x, \hat{\pi})$ have the same asymptotic properties in our framework.

2.4. The local quasi-likelihood estimation with the imputed values

The local quasi-likelihood estimator with complete-case data and the locally weighted quasi-likelihood estimator do not fully explore the information contained in the data. When there are many missing values, a substantial reduction in estimation efficiency emerges due to discarding incomplete cases in the local fitting. Stability issues also arise when using only complete case data. Since the effective number of local data points can be small, the singularity of the Hessian matrix $[\ell''(\cdot)]$ occurs frequently in the analysis of complete case data, particularly when the bandwidth is small. This is not simply rescued by increasing the size of bandwidth or using the ridge regression technique (see e.g., Seifert and Gasser (1996) and Fan and Chen (1999)), because these introduce extra biases.

We introduce an imputation method to manage the problem of missing data. The procedure consists of two steps. The first step involves imputing missing Y's based on the complete-case data with an initial bandwidth h_0. In the second step, we substitute Y_i by $\hat{Y}_i^* = \delta_i Y_i + (1-\delta_i) g^{-1}(\hat{\eta}(X_i))$, for $i = 1, \ldots, n$. Then the local quasi-likelihood based on imputed values is applied, namely maximizing

$$\ell(\beta) \equiv \sum_{i=1}^{n} Q\left[g^{-1}\{\beta_0 + \cdots + \beta_p (X_i - x)^p\}, \hat{Y}_i^*\right] K_{h_2}(X_i - x) \tag{2.8}$$

with respect to $\beta_j, j = 0, \ldots, p$, where $K_{h_2}(\cdot)$ is a kernel function and h_2 is a bandwidth. Let $\hat{\beta}_I(x) = (\hat{\beta}_{0,I}, \ldots, \hat{\beta}_{p,I})$ maximize (2.8). Then, the local quasi-likelihood estimators based on the imputed values are $\widehat{\eta}_I^{(v)}(x) = v!\widehat{\beta}_{v,I}(x)$ for $v = 0, \ldots, p$, with the convention that $\hat{\eta}_I(x) = \hat{\eta}_I^{(0)}(x)$.

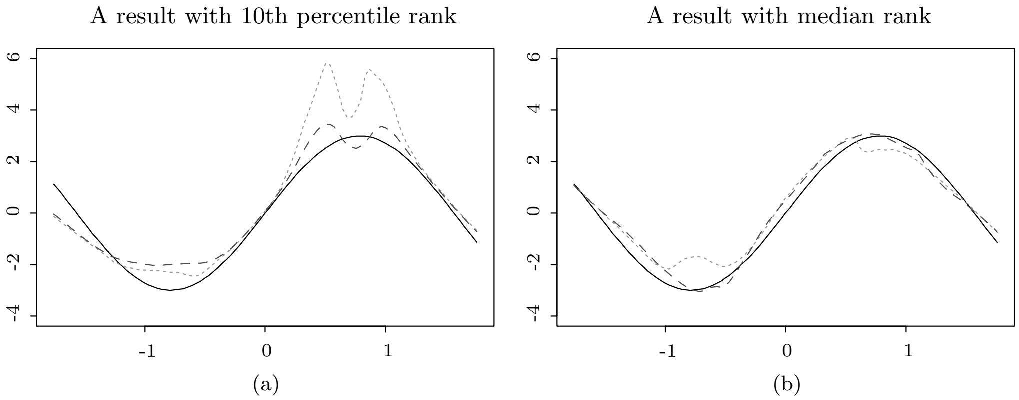

Figure 1. Comparison of the performances of the local quasi-likelihood estimator with complete-case data and the imputed values. (a) The 10th percentile performance among 400 simulations; (b) The median performance among 400 simulations. Solid curve − true function. Dash curves are the local estimators with the complete-case data (short dash) and imputed values (long dash).

The proposed imputation method improves the effectiveness of the local quasi-likelihood estimator with complete-case data. To justify our claim, we use a simulated example to illustrate the proposed method. The selection probability $\pi(x) = 0.4$ and the sample size $n = 500$. Figure 1 plots a typical estimate of the local linear quasi-likelihood estimator with the complete-case data and that of the local imputation method, both using the optimal estimated bandwidth. Details of simulations can be found in Section 5. In Figure 1(a), the local quasi-likelihood estimator with the complete-case data does not perform too well, due partially to the instability caused by missing values. The proposed imputation method improves the quality of the local estimator. The local quasi-likelihood estimator with complete-case data works reasonably well in Figure 1(b). Nevertheless, the local imputation method still improves somewhat on the performance of the local quasi-likelihood estimator with complete-case data.

2.5. Asymptotic properties

We explore the asymptotic distribution of $\hat\beta_C(x)$, $\hat\beta_W(x,\pi)$ and $\hat\beta_I(x)$. For convenience, we use the following notation. Write $\mu_j = \int u^j K(u)du$ and $\nu_j = \int u^j K(u)^2 du$, $j = 0, 1, 2, \ldots$. Let

$$S = (\mu_{i+j-2})_{1\le i,j\le p+1}, \quad S^* = (\nu_{i+j-2})_{1\le i,j\le p+1}, \quad H_1 = \operatorname{diag}(1, h_1, \ldots, h_1^p), \tag{2.9}$$

be $(p+1)\times(p+1)$ matrices. Furthermore, let $\beta(x) = (\eta(x), \eta'(x), \ldots, \eta^{(p)}(x)/p!)^T$ be the true local parameters.

Theorem 1. *Under Condition* 1 *in the Appendix, if* $nh_1 \to \infty$ *and* $h_1 \to 0$*, then* $\hat\beta_C(x)$ *and* $\hat\beta_W(x,\pi)$ *satisfy*

$$\sqrt{nh_1}\left(H_1\{\hat\beta(x) - \beta(x)\} - \frac{\eta^{(p+1)}(x)}{(p+1)!}S^{-1}U_{p+1}h_1^{p+1} + o_p(h_1^{p+1})\right)$$
$$\longrightarrow N\Big(0, \frac{\sigma^2(x)}{\pi(x)f(x)}S^{-1}S^*S^{-1}\Big), \tag{2.10}$$

where $U_{p+1} = (\mu_{p+1}, \ldots, \mu_{2p+1})^T$ *is a* $(p+1)$*-column vector, and* $\sigma^2(x) = [g'\{m(x)\}]^2 \mathrm{Var}\,(Y|X=x)$.

The proof of Theorem 1 is given in the Appendix. Comparing this with Theorem 1 of Fan, Heckman and Wand (1995), one can easily see that the local quasi-likelihood estimator based on complete-case data shares the same asymptotic bias as that with full data, but has larger asymptotic variance. The extra factor $1/\pi(x)$ can have an adverse effect on the efficiency of estimation, especially when there are lots of missing values. The results in Theorem 1 also reveal that the locally weighted quasi-likelihood estimator does not provide any improvement over the local quasi-likelihood estimator with the complete-case data in the local fitting. The asymptotic normality of the local quasi-likelihood estimator with the imputed values $\hat\beta_I(x)$ can be described as follows.

Theorem 2. *Under Conditions* 1 *and* 2 *in the Appendix, as* $n \to \infty, nh_0^4 \to 0$, $nh_0^2/\log(1/h_0) \to \infty$, $h_2 \to 0$ *and* $nh_2 \to \infty$,

$$\sqrt{nh_2}\,\Big(H_2\{\hat\beta_I(x) - \beta(x)\} - \lambda(x)\Big) \longrightarrow N\Big(0, \frac{\sigma^2(x)}{\pi(x)f(x)}S^{-1}S^*S^{-1}\Big), \tag{2.11}$$

where $H_2 = \mathrm{diag}(1, h_2, \ldots, h_2^p)$, $U_k = (\mu_k, \ldots, \mu_{p+k})^T, (k = 0, p+1)$, *and*

$$\lambda(x) = \frac{\eta^{(p+1)}(x)}{(p+1)!} \Big(S^{-1} U_{p+1} h_2^{p+1} + (1 - \pi(x)) e_1^T S^{-1} U_{p+1} S^{-1} U_0 h_0^{p+1} \Big) \\ + o_p(h_2^{p+1} + h_0^{p+1}).$$

In particular, we have the asymptotic expansion

$$\sqrt{nh_2}\,(\hat{\eta}_I(x) - \eta(x) - \lambda_0(x)) \longrightarrow N\Big(0, \frac{\sigma^2(x)}{\pi(x) f(x)} S^{-1} S^* S^{-1}\Big), \tag{2.12}$$

with $\lambda_0(x) = \eta^{(p+1)}(x) e_1^T S^{-1} U_{p+1} \Big(h_2^{p+1} + (1 - \pi(x)) h_0^{p+1} \Big) / (p+1)! + o_p(h_2^{p+1} + h_0^{p+1})$.

The proof of Theorem 2 is in the Appendix. Note that the imputation method has a rather simple asymptotic expression. If $h_0 = o(h_2)$ in the imputation, the bias of $\hat{\beta}_I$ is the same as that of $\hat{\beta}_C$. We can select h_0 and h_2 such that the bias of $\hat{\beta}_I$ less than $\hat{\beta}_C$. For example, if $h_2 = h_0 < (2 - \pi(x))^{-1/(p+1)} h_1$, then the bias of $\hat{\beta}_I$ is smaller than that of $\hat{\beta}_C$. Another gain is that the imputation method has more local data points and is more stable in implementation. Thus, it improves the finite sample properties. The simulation results in Section 5 reinforce this statement about finite sample behavior.

2.6. Optimal bandwidth

The bandwidth parameter is important in nonparametric curve estimation. From Theorems 1 and 2, the asymptotic optimal bandwidths $h_{1,opt}$ and $h_{2,opt}$ can be chosen by minimizing their asymptotic weighted mean integrated squared errors, respectively. Denote $S^{-1} = (S^{ij})_{0 \le i,j \le p}$ and let

$$K_v^*(t) = e_{v+1}^T S^{-1} (1, t, \ldots, t^p)^T K(t) = \Big(\sum_{j=0}^{p} S^{vj} t^j \Big) K(t)$$

be the equivalent kernel. Then, the asymptotically optimal bandwidth for estimating $\hat{\eta}_C(\cdot)$ based on the complete-case data is given by

$$h_{1,opt} = C_{0,p}(K) \left[\frac{\int \frac{\sigma^2(x) w(x)}{\pi(x) f(x)} dx}{\int \{\eta^{(p+1)}(x)\}^2 w(x) dx} \right]^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}}, \tag{2.13}$$

and the asymptotically optimal bandwidth $h_{2,opt}$ for the estimator $\hat{\eta}_I(x)$ can be expressed by solving the following equation:

$$h_{2,opt}^{2p+3} + \frac{\int \{\eta^{(n+1)}(x)\}^2 (1-\pi(x)) w(x) dx}{2(p+1) \int \{\eta^{(p+1)}(x)\}^2 w(x) dx} h_0^{p+1} h_{2,opt}^{p+2}$$
$$= C_{0,p}(K)^{2p+3} \frac{\int \frac{\sigma^2(x) w(x)}{\pi(x) f(x)} dx}{n \int \{\eta^{(p+1)}(x)\}^2 w(x) dx}, \qquad (2.14)$$

where w is a weight function, h_0 is an initial bandwidth and

$$C_{0,p}(K) = \left[\frac{(p+1)!^2 \int K_0^{*^2}(t) dt}{2(p+1)\{\int t^{p+1} K_0^*(t) dt\}^2} \right]^{\frac{1}{2p+3}}.$$

This theoretical optimal bandwidth depends on unknown quantities. In Section 4, we provide a data-driven bandwidth selection based on the generalized pre-asymptotic methods of Fan and Gijbels (1995) and Fan, Farmen and Gijbels (1998).

3. Quasi-likelihood Imputation Estimation of The Mean Functionals

We turn to the investigation of the estimation of the mean functionals $\theta = E(Y)$ using imputed values. Cheng (1994) studied a kernel regression imputation estimator of the mean functional. Each missing datum Y is imputed by kernel regression imputation. The estimate of the mean parameter θ can be taken as

$$\hat{\theta}_K = \frac{1}{n} \sum_{i=1}^{n} [\delta_i Y_i + (1-\delta_i) \hat{m}(X)], \qquad (3.1)$$

where $\hat{m}(X)$ is the Nadaraya-Watson kernel estimator based on the complete-case data $\{(X_i, Y_i) : \delta_i = 1, i = 1, \ldots, n\}$. Wang, Linton and Härdle (2004) developed the kernel regression estimators of the mean functionals in a semiparametric partially linear regression model.

In the section, we develop the nonparametric local quasi-likelihood estimators of the mean functionals. The local quasi-likelihood estimators developed in Section 2 are used to impute each missing datum. Four quasi-likelihood imputation estimators for the mean functional of the response are introduced. The quasi-likelihood imputation estimator and the weighted quasi-likelihood imputation estimator can be expressed by, respectively,

$$\hat{\theta}_Q = \frac{1}{n} \sum_{i=1}^{n} [\delta_i Y_i + (1-\delta_i) g^{-1}\{\hat{\eta}_C(X_i)\}], \qquad (3.2)$$

$$\hat{\theta}_{GW} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\delta_i}{\hat{\pi}(X_i)} Y_i + (1 - \frac{\delta_i}{\hat{\pi}(X_i)}) g^{-1}\{\hat{\eta}_C(X_i)\} \right], \qquad (3.3)$$

where $\hat{\pi}(X_i) = \sum_{j=1}^n \delta_j K_a(X_j - X_i)/\sum_{j=1}^n K_a(X_j - X_i)$, with $K_a(\cdot)$ a kernel function and a a bandwidth.

By using the local quasi-likelihood estimator with the imputed values, a two-step quasi-likelihood imputation estimator of the mean functional and the corresponding two-step weighted quasi-likelihood imputation estimator can be expressed as

$$\hat{\theta}_T = \frac{1}{n}\sum_{i=1}^n [\delta_i Y_i + (1-\delta_i) g^{-1}\{\hat{\eta}_I(X_i)\}], \tag{3.4}$$

$$\hat{\theta}_{TW} = \frac{1}{n}\sum_{i=1}^n \left[\frac{\delta_i}{\hat{\pi}(X_i)} Y_i + (1 - \frac{\delta_i}{\hat{\pi}(X_i)}) g^{-1}\{\hat{\eta}_I(X_i)\}\right]. \tag{3.5}$$

The following Theorem 3 establishes the asymptotic normality of the proposed quasi-likelihood imputation estimators of mean functionals.

Theorem 3. *Under Conditions* 1 *and* 2 *in the Appendix, as* $n \to \infty, nh_0^4 \to 0$, *and* $nh_0^2/\log(1/h_0) \to \infty$, *the quasi-likelihood imputation estimator* $\sqrt{n}(\hat{\theta} - \theta - \lambda)$ *has an asymptotic normal distribution* $N(0, \Lambda)$, *where* $\lambda = O_p(h_0^{p+1})$ *and* $\Lambda = \mathrm{Var}\{m(X)\} + E\{\mathrm{Var}(Y|X)/\pi(X)\}$.

The proof is given in the Appendix. The result shows that the proposed local quasi-likelihood imputation estimators have the $(p+1)$st order asymptotic bias $O_p(h_0^{p+1})$. Comparing our results with those in Theorem 1 in Cheng (1994), the proposed imputation estimators have a higher order asymptotic bias than that of the kernel imputed method although both estimators share the same asymptotic variance. When the data are full, i.e., the selection probability $\pi(x) = 1$, Theorem 3 reduces to the classical result: the asymptotic variance of $\hat{\theta}$ is $(1/n)[\mathrm{Var}\{m(X)\} + E\{\mathrm{Var}(Y|X)\}] = (1/n)\mathrm{Var}(Y)$. A consistent estimator for the asymptotic variance Λ is

$$\hat{\Lambda} = \frac{1}{n}\sum_{i=1}^n \left\{ \left(g^{-1}(\hat{\eta}(X_i)) - \frac{1}{n}\sum_{j=1}^n g^{-1}(\hat{\eta}(X_j))\right)^2 + \frac{\hat{\mathrm{Var}}(Y|X_i)}{\hat{\pi}(X_i)}\right\}, \tag{3.6}$$

where $\hat{\mathrm{Var}}(Y|x) = \sum_{i=1}^n \delta_i Y_i^2 K_a(X_i - x)/\sum_{i=1}^n \delta_i K_a(X_i - x) - g^{-1}(\hat{\eta}(X_i))^2$. For the logistic regression and Poisson regression with the canonical links, we can specifically estimate the variance by $\hat{\mathrm{Var}}(Y|x) = \exp\{\hat{\eta}(x)\}/(1 + \exp\{\hat{\eta}(x)\})^2$ and $\hat{\mathrm{Var}}(Y|x) = \exp\{\hat{\eta}(x)\}$, respectively.

4. Estimation of Bias and Variance and Bandwidth Selection

In this section, we assess the bias and variance of the local quasi-likelihood estimator with complete-case data. Combining the generalized pre-asymptotic methods of Fan and Gijbels (1995) and Fan, Farmen and Gijbels (1998), we propose a new data-driven bandwidth selector. The method is also applicable to the locally weighted estimator and the local imputation estimator.

4.1. Assessing the bias and variance of the estimator

Suppose that $\eta(x)$ has a $(p+1+b)$th derivative at the point x for a positive integer a. Let

$$r(X_i) = \eta(X_i) - \sum_{j=0}^{p} \eta^{(j)}(x) \frac{(X_i - x)^j}{j!}, \quad \text{for} \quad i = 1, \ldots, n.$$

Since the bias of the estimator $\hat{\beta}_C(x)$ comes from the approximation error in the Taylor expansion, $r(X_i)$ can be approximated by

$$\beta_{p+1}(X_i - x)^{p+1} + \cdots + \beta_{p+b}(X_i - x)^{p+b} \equiv r_i, \tag{4.1}$$

where b denotes the order of the approximation. In general, we put $b = 2$ as recommended by Fan and Gijbels (1995). Therefore, for given quantities r_i, a more precise local quasi-likelihood is given by

$$\ell^*(\beta) \equiv \sum_{i=1}^{n} \delta_i Q[g^{-1}\{\beta_0 + \cdots + \beta_p(X_i - x)^p + r_i\}, Y_i] K_h(X_i - x). \tag{4.2}$$

Let $\hat{\beta}^*_C(x) = (\hat{\beta}^*_{0,C}(x), \ldots, \hat{\beta}^*_{p,C}(x))$ maximize (4.2). Then the bias of $\hat{\beta}_C(x)$ can be estimated by $\hat{\beta}_C(x) - \hat{\beta}^*_C(x)$. Following Fan, Farmen and Gijbels (1998), the estimated bias and variance of $\hat{\beta}_C(x)$ can be written, respectively, as

$$\begin{aligned} \hat{B}(\hat{\beta}_C(x), h) = & \left(\sum_{i=1}^{n} \delta_i q_2 \left(\mathbf{X}_i^T \hat{\beta}_C(x) + r_i, Y_i \right) \mathbf{X}_i \mathbf{X}_i^T K_h(X_i - x) \right)^{-1} \\ & \times \sum_{i=1}^{n} \delta_i q_1 \left(\mathbf{X}_i^T \hat{\beta}_C(x) + r_i, Y_i \right) \mathbf{X}_i K_h(X_i - x), \end{aligned} \tag{4.3}$$

$$\begin{aligned} \hat{V}(\hat{\beta}_C(x), h) \doteq [\sigma^2(x)]^{-1} & \left(\sum_{i=1}^{n} \delta_i q_2 \left(\mathbf{X}_i^T \hat{\beta}_C(x), Y_i \right) \mathbf{X}_i \mathbf{X}_i^T K_h(X_i - x) \right)^{-1} \\ & \times \left(\sum_{i=1}^{n} \delta_i \mathbf{X}_i \mathbf{X}_i^T K_h^2(X_i - x) \right) \\ & \times \left(\sum_{i=1}^{n} \delta_i q_2 \left(\mathbf{X}_i^T \hat{\beta}_C(x), Y_i \right) \mathbf{X}_i \mathbf{X}_i^T K_h(X_i - x) \right)^{-1}, \end{aligned} \tag{4.4}$$

where $\sigma^2(x) = [g'\{m(x)\}]^2 \mathrm{Var}\,(Y|x)$. For binary and Poisson regression with canonical links, $\hat{\sigma}^2(x) = (1+\exp\{\mathbf{X}_i^T\hat{\beta}_C(x)\})^2/\exp\{\mathbf{X}_i^T\hat{\beta}_C(x)\}$ and $\hat{\sigma}^2(x) = 1/\exp\{\mathbf{X}_i^T\hat{\beta}_C(x)\}$, respectively. In general, $\sigma^2(x)$ can be estimated by the local residual variance defined by

$$\hat{\sigma}^2(x) = \frac{\sum_{i=1}^{n} \delta_i \left(Y_i - g^{-1}(\mathbf{X}_i^{*T}\hat{\beta}^{(p+b)})\right)^2 [g'\{\hat{m}^*(X_i)\}]^2 K_{h^*}(X_i - x)}{\sum_{i=0}^{n} \delta_i K_{h^*}(X_i - x)}, \qquad (4.5)$$

where $\hat{\beta}^{(p+b)} = (\hat{\beta}_0, \ldots, \hat{\beta}_{p+b})^T$ is the result of the $(p+b)$th-order local polynomial fit (2.6) using the pilot bandwidth h^* and $\mathbf{X}_i^* = (1, X_i - x, \ldots, (X_i - x)^{p+b})^T$, and $\hat{m}^*(X_i) = g^{-1}(\mathbf{X}_i^*\hat{\beta}^{(p+b)})$. In the same way, $r_1, \ldots, r_n$ in the estimated bias (4.3) can be estimated by (4.1), using the estimator $\hat{\beta}^{(p+b)}$.

4.2. Bandwidth selection

In order to derive an estimator of the theoretical optimal bandwidth, a pilot bandwidth is needed in order to assess the bias and variance. This in turn requires the selection of a pilot bandwidth. We define a Generalized Residual Squares Criterion (GRSC) as follows:

$$\mathrm{GRSC}(x;h) = \hat{\sigma}^2(x)\{1 + (p+1)N\}, \qquad (4.6)$$

where $\hat{\sigma}^2(x)$ is given by (4.5) and N is the first diagonal element of the matrix

$$\left(\sum_{i=1}^{n}\delta_i \mathbf{X}_i\mathbf{X}_i^T K_h(X_i-x)\right)^{-1}\left(\sum_{i=1}^{n}\delta_i \mathbf{X}_i\mathbf{X}_i^T K_h^2(X_i-x)\right)\left(\sum_{i=1}^{n}\delta_i \mathbf{X}_i\mathbf{X}_i^T K_h(X_i-x)\right)^{-1}.$$

The intuition behind the GRSC criterion is the same as for the least squares case proposed in Fan and Gijbels (1995). When the bandwidth h is too large, the polynomial function may not fit well − the bias is large and so is $\hat{\sigma}^2(x)$. On the other hand, if the bandwidth h is too small, the variance of the fit will be large and hence N will be large as well. The GRSC quantity protects against both extreme choices. A justification for the GRSC can be found in the following theorem.

Theorem 4. *Under Condition* 1 *in the Appendix, as* $h_n \to 0$ *and* $nh_n \to \infty$, *then*

$$E\{\mathrm{GRSC}(x;h_n)|X_1,\ldots,X_n\}$$
$$= \sigma^2(x) + C_p \frac{h_n^{2p+2}}{(p+1)!^2}[\eta^{(p+1)}(x)]^2 + a_0 \frac{(p+1)\sigma^2(x)}{\pi(x)f(x)nh_n} + o_p\left(h_n^{2p+2} + \frac{1}{nh_n}\right), \quad (4.7)$$

where $C_p = (\mu_{2p+2} - U_{p+1}^T S^{-1} U_{p+1})/\mu_0$, $a_0 = e_1^T S^{-1} S^* S^{-1} e_1$ *and* $e_1 = (1, 0, \ldots, 0)^T$, *and other symbols are the same as those in Theorem* 1.

Theorem 4 can be proved by using arguments similar to those in Fan and Gijbels (1995). The minimizer of the weighted integration of (4.7) leads to the optimal pilot bandwidth as follows:

$$\begin{aligned} h_p^* &= \arg\min_h \left\{ \int \mathrm{GRSC}(x;h) w(x) dx \right\} \\ &= \left[\frac{(p+1)!^2 \int K_0^{*2}(t)dt \int \sigma^2(x)w(x)/(\pi(x)f(x))dx}{2nC_p \int \{\eta^{(p+1)}\}^2 w(x)dx} \right]^{\frac{1}{2p+3}}, \qquad (4.8) \end{aligned}$$

for some given weight function w. The relationship between h_p^* and $h_{1,opt}$ in (2.13) is given by

$$\begin{aligned} h_{1,opt} &= \left[\frac{C_p \int K_0^{*2}(t)dt}{(p+1)\left\{\int t^{p+1} K_0^*(t)dt\right\}^2 \int K_0^{*2}(t)dt} \right]^{\frac{1}{2p+3}} h_p^* \\ &\equiv \mathrm{adj}_{0,opt}(K) h_p^*. \qquad (4.9) \end{aligned}$$

This shows that the minimizer of (4.6) is only a constant factor away from the target optimal bandwidth. The adjusting constant $\mathrm{adj}_{0,opt}(K)$ is a known constant that depends only on the kernel function K.

After selecting a pilot bandwidth h_{p+b}^* from (4.6) and (4.8), an estimator $\hat\beta^{(p+b)} = (\hat\beta_0, \ldots, \hat\beta_{p+b})^T$ can be derived by fitting a polynomial of degree $p+b$ locally based on (2.6). Then the estimated bias $\hat B(\hat\beta_C(x), h)$ and variance $\hat V(\hat\beta_C(x), h)$ of $\hat\beta_C(x)$ are obtained. Therefore, the MSE of the first element of $\hat\beta_C(x)$ can be estimated by

$$\hat{\mathrm{MSE}}_{p,0}(x,h) = \hat B_{p,0}(\hat\beta_C(x), h)^2 + \hat V_{p,0}(\hat\beta_C(x), h). \qquad (4.10)$$

Furthermore, a data-driven optimal bandwidth selector is given by

$$\hat h_{1,opt} = \arg\min_h \left\{ \int \hat{\mathrm{MSE}}_{p,0}(x,h) w(x) dx \right\}, \qquad (4.11)$$

where $w(\cdot)$ is a given weight function.

When the curve $\eta(\cdot)$ admits various degrees of smoothness at different locations, a variable bandwidth selector can be used to enhance the performance of the proposed local quasi-likelihood estimator.

5. Simulation Studies

Monte Carlo simulations are designed to evaluate the finite sample properties of the three proposed estimators. The incomplete random sample is $\{(X_i, Y_i, \delta_i),$

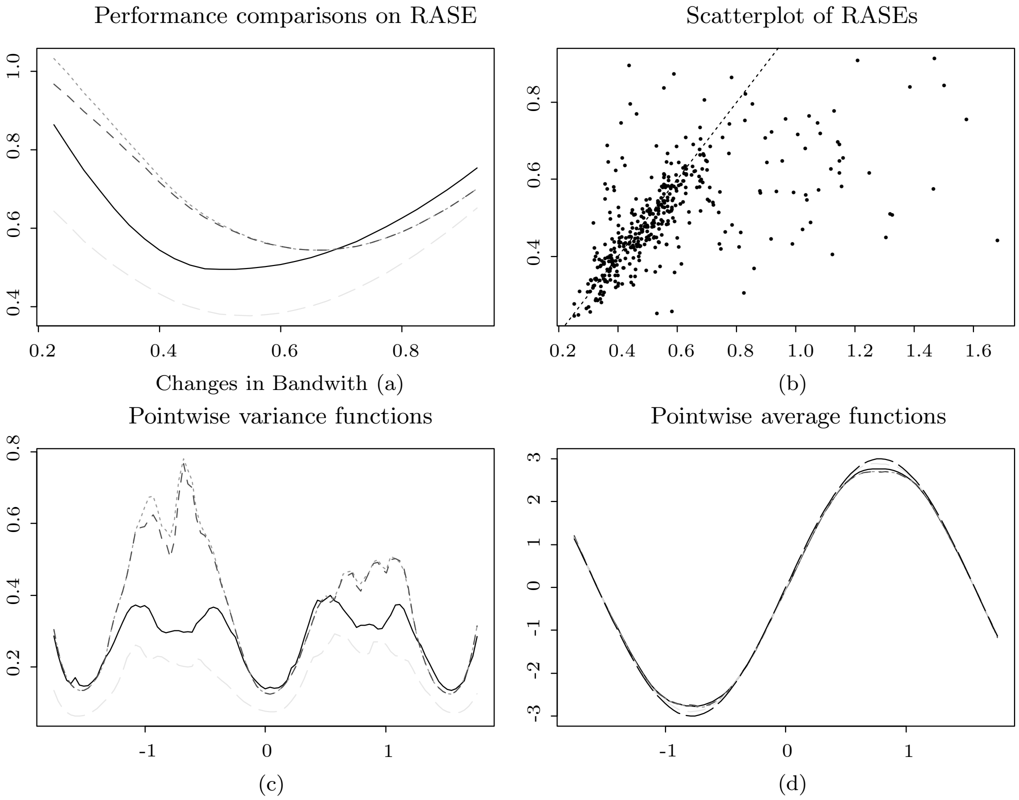

Figure 2. Simulation results for Example 1 with selection probability $\pi_1(x)$ and $n = 500$. (a) The average RASE as a function of bandwidth for the three estimation methods; (b) The scatter plot for the RASE of the local quasi-likelihood estimator with the imputed values versus that of the complete-case data, using data driven bandwidth (dashed line indicates that both methods have the same performance); (c) Pointwise variance of 400 estimated functions of the three estimation procedures; (d) Pointwise averages of 400 estimated functions of the three estimation procedures. Solid curve is the local quasi-likelihood estimator with the imputed-data $\hat{\eta}_I(x)$. Dash curves (from shortest to longest dash) are the local quasi-likelihood estimator with the complete-case data $\hat{\eta}_C(x)$, the local weighted quasi-likelihood estimator $\hat{\eta}_W(x, \pi)$ and the local quasi-likelihood estimator with full data, respectively. Note that the longest dashed curve in (d) is the true function.

$i = 1, \ldots, n\}$, where the (X_i, Y_i)'s come from a population (X, Y) whose conditional distribution is a Bernoulli with $P(Y = 1|X = x) = \exp\{\eta(x)\}/(1 + \exp\{\eta(x)\})$, and δ is a 0-1 random variable with the selection probability $\pi(\delta = 1|Y, X) = \pi(X)$. Two simulation models from Fan, Farmen and Gijbels (1998) are employed here as testing examples. Their logistic regression functions are given by

Example 1. $\eta(x) = 3\sin(2x)$

Example 2. $\eta(x) = 7[\exp\{-(x+1)^2\} + \exp\{-(x-1)^2\}] - 5.5,$

where the design density is the uniform distribution on $[-2, 2]$. The selection probability given X is taken in turn as $\pi_1(x) = \exp\{x\}/(1+\exp\{x\})$, $\pi_2(x) = 0.4$, and $\pi_3(x) = 0.9 - 0.2|x|$ if $|x| \leq 0.6$, and $= 0.3 + 0.1|x|$ if $0.6 < |x| < 2$. For each of the above examples with three selection probabilities, 400 simulations are conducted with sample sizes $n = 250, 500$ and $1,000$.

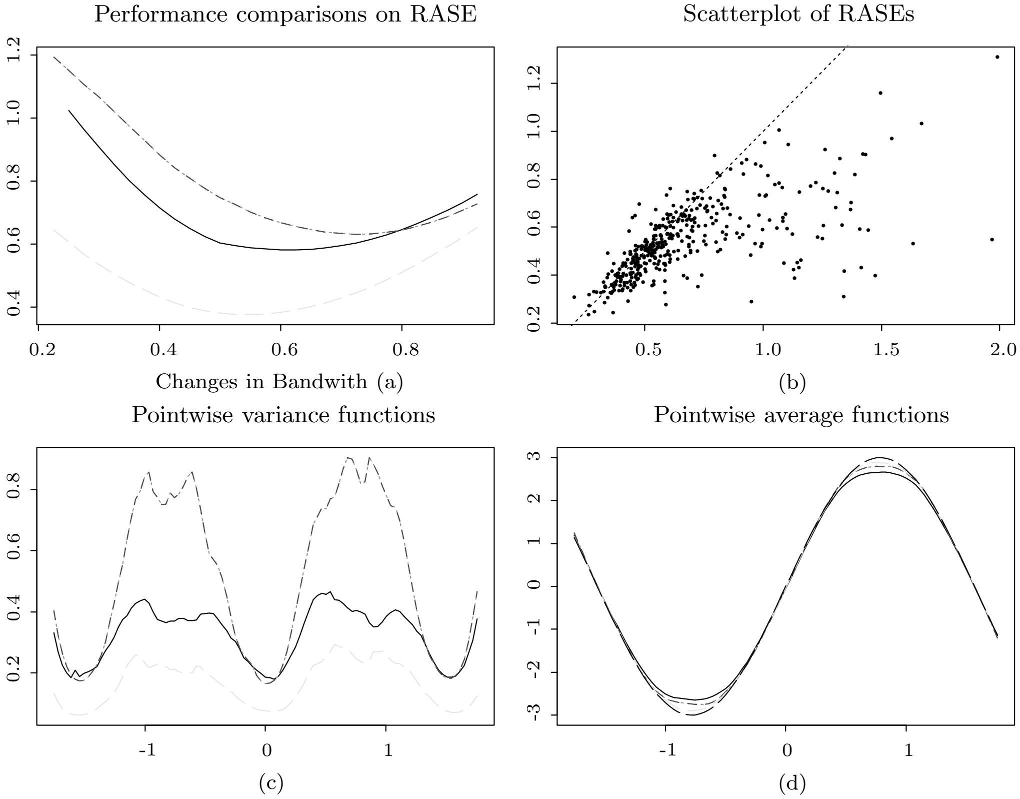

Figure 3. Simulation results for Example 1 with the selection probability $\pi_2(x)$ and n=500. Similar captions to Figure 2 are used. Note that $\hat{\eta}_C(x)$ and $\hat{\eta}_W(x, \pi)$ are identical under $\pi_2(x)$ and hence there are only two procedures.

In the implementation, we employ the local linear fit with the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$. The performance of each given estimator $\hat{\eta}(x)$ is assessed via the square-root of the Average Square Errors (RASE):

$$\text{RASE} = \left(n_{grid}^{-1} \sum_{j=1}^{n_{grid}} \{\hat{\eta}(x_j) - \eta(x_j)\}^2 \right)^{\frac{1}{2}}, \tag{5.1}$$

where $\{x_j, j = 1, \ldots, n_{grid}\}$ are the grid points at which the function $\eta(\cdot)$ is estimated.

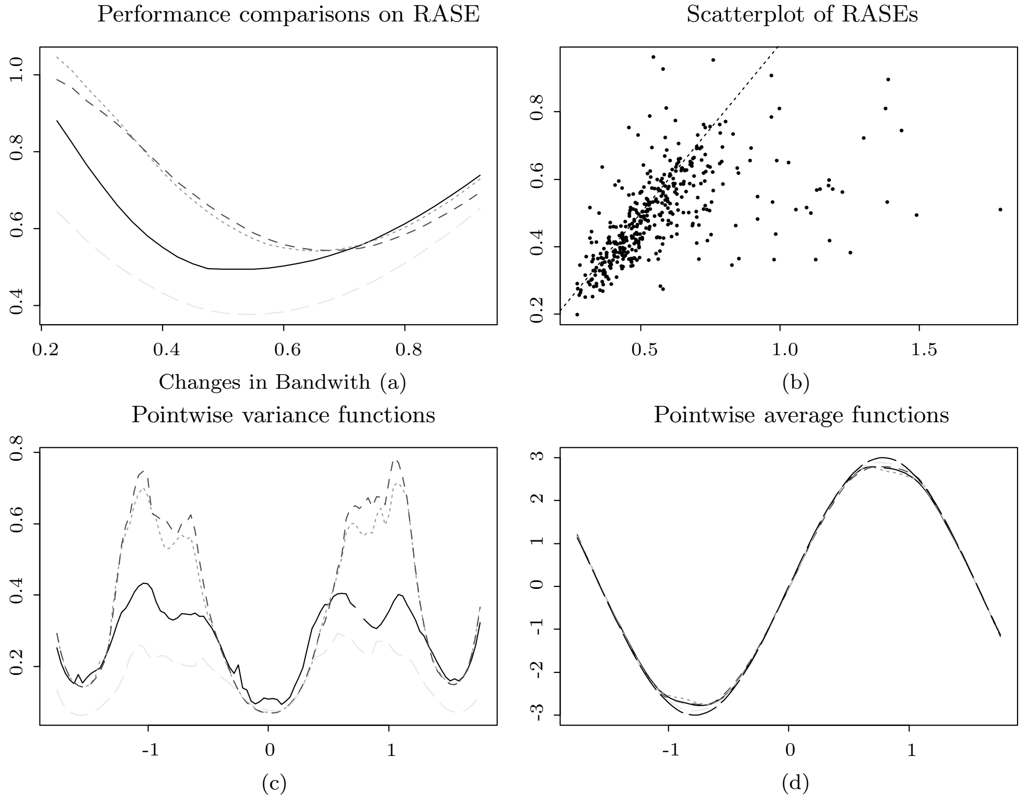

Figure 4. Simulation results for Example 1 with the selection probability $\pi_3(x)$. Captions similar to Figure 2 are used.

Consider Example 1 with the three selection probabilities $\pi_1(x), \pi_2(x)$ and $\pi_3(x)$. Figures 2−4 summarize the performance of the local quasi-likelihood estimator with the complete-case data $\hat{\eta}_C(x)$, the local weighted quasi-likelihood estimator $\hat{\eta}_W(x, \pi)$, and the locally quasi-likelihood estimator with the imputed values $\hat{\eta}_I(x)$, based on 400 simulations with the sample size $n = 500$. To examine the efficacy of the imputed-values method in the local fitting, we also give the performance of the local quasi-likelihood estimator with the full data (use both the complete-case data and missing data) $\hat{\eta}(x)$.

Part (a) in Figures 2−4 depicts the average of RASEs, resulting from 400 simulations, as a function of the bandwidth. The average RASE curve for $\hat{\eta}_I(x)$ is smaller than those for $\hat{\eta}_C(x)$ and $\hat{\eta}_W(x,\pi)$ for a reasonable range of the bandwidth, while the average RASE curve for $\hat{\eta}_C(x)$ is almost the same as that for $\hat{\eta}_W(x,\pi)$. This implies that $\hat{\eta}_I(x)$ outperforms $\hat{\eta}_C(x)$ and $\hat{\eta}_W(x,\pi)$ when bandwidths are properly chosen and that $\hat{\eta}_C(x)$ and $\hat{\eta}_W(x,\pi)$ perform nearly the same. Of course, the average RASE curve for $\hat{\eta}(x)$ is the smallest for a reasonable range of the bandwidth since it uses all data including the complete-case data and missing data. In part (b) of Figures 2−4, we compare, sample-by-sample, the performance of $\hat{\eta}_I(x)$ with $\hat{\eta}_C(x)$, both using the optimal estimated bandwidth in Section 4. For each simulated sample, the RASE of $\hat{\eta}_I(x)$ is plotted against the RASE of $\hat{\eta}_C(x)$. The slope dashed line is a diagonal which marks the position where $\hat{\eta}_I(x)$ and $\hat{\eta}_C(x)$ have the same performance. Figure (2b) indicates that $\hat{\eta}_I(x)$ outperforms $\hat{\eta}_C(x)$ more frequently among 400 simulation trials. To better understand the performance of the three procedures, we plot in part (c) the pointwise variance of 400 estimated curves resulting from 400 simulations. It is evident that $\hat{\eta}_I(x)$ has smaller variance and hence gives more stable performance. The variances of $\hat{\eta}_I(x)$ and $\hat{\eta}(x)$ are closer for all local points. This indicates that $\hat{\eta}_I(x)$ can improve substantially on $\hat{\eta}_C(x)$ and $\hat{\eta}_W(x,\pi)$. Part (d) of the figures shows the average of 400 estimated curves. It indicates the bias of three competing procedures − they have about the same amount of bias.

We now consider Example 2. The performance is summarized in Figure 5. Similar to that in Example 1, substantial gains are obtained by the local quasi-likelihood estimator with the imputed values. The proposed imputation method outperforms those that use only the complete-case data. This is somewhat expected since the imputation method makes better use of the observed data in the local fitting. Again, $\hat{\eta}_I(x)$ is a more stable estimator.

Table 1 presents the results for the imputation estimators $\hat{\theta}_K$ $\hat{\theta}_G$, $\hat{\theta}_{GW}$, $\hat{\theta}_T$ and $\hat{\theta}_{TW}$ for both Examples 1 and 2 with the different missing selection probabilities. The sample sizes studied are $n = 250$, and 400 simulations are conducted. We make the following observations. (i) All the imputation estimators of the mean functionals are valid. (ii) The proposed estimated variances $\hat{SE}$ in (3.6) provide good estimates for the variances of θ. (iii) The proposed quasi-likelihood imputation estimators $\hat{\theta}_G$, $\hat{\theta}_{GW}$, $\hat{\theta}_T$ and $\hat{\theta}_{TW}$ have smaller variances than that of the kernel imputation $\hat{\theta}_K$, while the two-step imputation estimator $\hat{\theta}_T$ has the smallest variance.

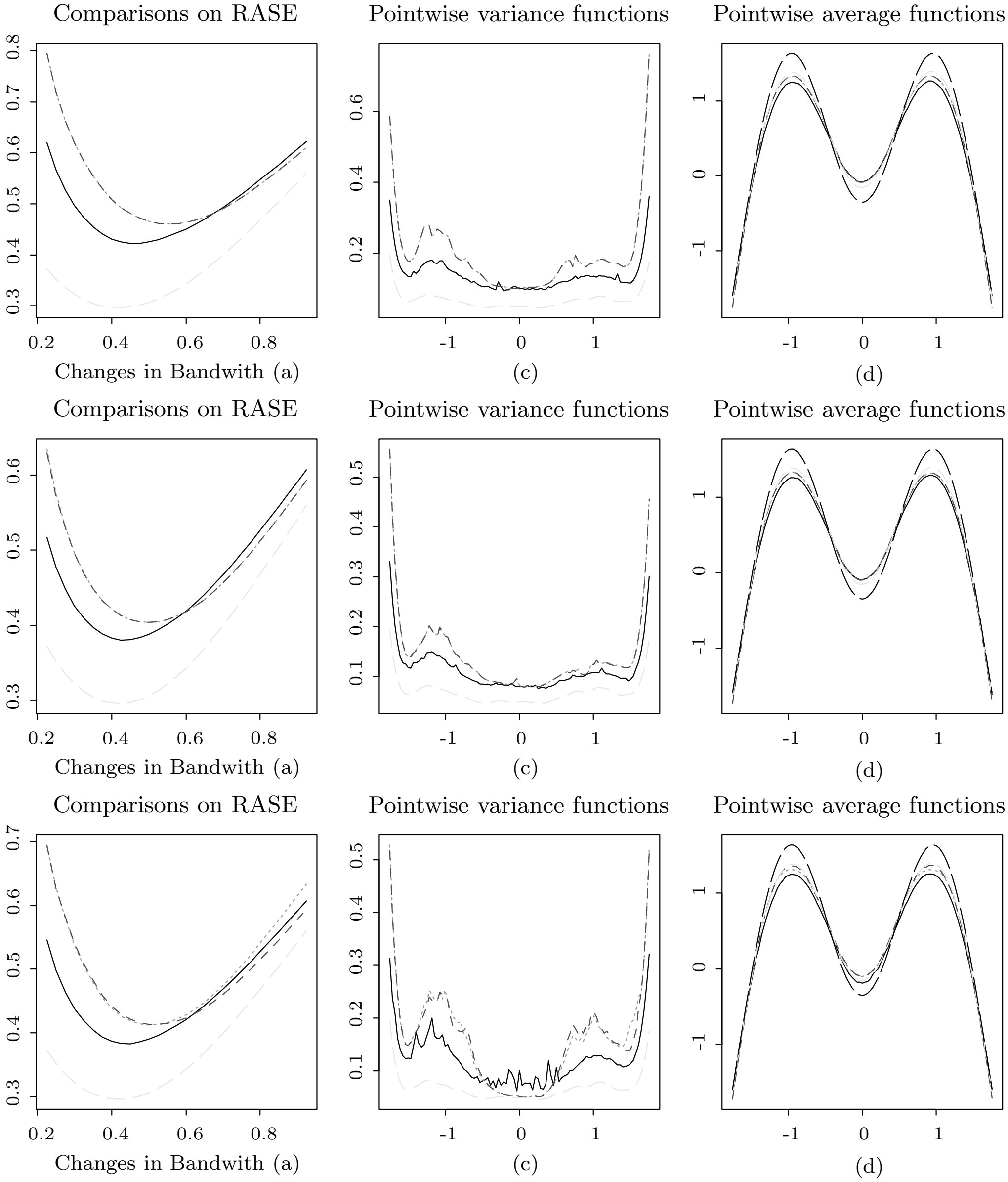

Figure 5. Simulation results for Example 2 with the three selection probabilities. Captions similar to Figure 2−4 are used.

6. Application to Data

We illustrate the proposed methods by analyzing a data set from the Collaborative Perinatal Project (CPP). CPP was a prospective study designed to

Table 1. Simulation study for the estimators of the mean functionals. The results are based on different missing functions $\pi(x)$, and the sample size $n = 250$.

Examples	$\pi(x)$	θ	Estimator	$\hat{\theta}_K$	$\hat{\theta}_G$	$\hat{\theta}_{GW}$	$\hat{\theta}_T$	$\hat{\theta}_{TW}$
1	$\pi_1(x)$	0.5000	Mean	0.4981	0.4981	0.4982	0.4994	0.4983
			SE	0.0409	0.0409	0.0401	0.0397	0.0401
			$\widehat{\mathrm{SE}}$	0.0400	0.0401	0.0401	0.0411	0.0411
	$\pi_2(x)$	0.5000	Mean	0.4960	0.4951	0.4956	0.4956	0.4958
			SE	0.0434	0.0446	0.0429	0.0426	0.0428
			$\widehat{\mathrm{SE}}$	0.0436	0.0439	0.0439	0.0458	0.0458
	$\pi_3(x)$	0.5000	Mean	0.4964	0.4960	0.4961	0.4965	0.4962
			SE	0.0389	0.0387	0.0383	0.0374	0.0383
			$\widehat{\mathrm{SE}}$	0.0392	0.0395	0.0395	0.0405	0.0405
2	$\pi_1(x)$	0.5571	Mean	0.5574	0.5598	0.5560	0.5560	0.5562
			SE	0.0455	0.0452	0.0450	0.0444	0.0450
			$\widehat{\mathrm{SE}}$	0.0423	0.0425	0.0425	0.0429	0.0429
	$\pi_2(x)$	0.5571	Mean	0.5564	0.5573	0.5545	0.5515	0.5546
			SE	0.0496	0.0499	0.0491	0.0488	0.0488
			$\widehat{\mathrm{SE}}$	0.0468	0.0471	0.0471	0.0478	0.0478
	$\pi_3(x)$	0.5571	Mean	0.5502	0.5523	0.5530	0.5557	0.5526
			SE	0.0431	0.0429	0.0423	0.0422	0.0423
			$\widehat{\mathrm{SE}}$	0.0413	0.0415	0.0415	0.0420	0.0420

identify determinants of nuerodevelopmental deficits in children (Niswander and Gordon (1972)) About 56,000 pregnant women were recruited between 1959 and 1966 in the U.S. and blood serum samples were collected during pregnancy. In a recent study (Longnecker, Klebnoff, Zhou and Brock (2002)). the investigators thawed a subsample of 344 of the frozen blood serum samples and measured the extent of exposure to Polychlorinated biphenyls (PCBs), an ubiquitous environmental contaminant, for these women. In this analysis, we looked at the relationship between the women's PCBs exposure (high or low) and their body mass index at the beginning of the pregnancy. In addition to the 344 women with the outcome variable (level of PCBs), we also have an additional subsample of 696 where we observe only the covariate (the body mass index). The total sample size in this example is 1,040.

We apply the Bernoulli likelihood and logit link function to these data. The proposed local quasi-likelihood estimation with complete-case data and the local quasi-likelihood estimation with imputed values are used to estimate $\mathrm{logit}P(Y = 1|X = x)$. The Epanechnikov kernel is used. We apply the data-driven bandwidth developed in Section 4 to select the optimal bandwidths. The corresponding bandwidths are $h_1 = 2.79$ and $h_2 = 2.32$. Figure 6(a) shows the estimated logit functions for both local estimators. The solid line is the local quasi-likelihood estimator with the imputed values and the dashed line is the local quasi-likelihood estimator with the complete-case data. Figure 6(b) plots the estimated conditional probability by the imputation method. The dashed lines are the estimated function plus (or minus) twice the estimated standard errors at each given point. They give us rough ideas about the estimation errors. For comparison purposes, we have also employed local estimation with the complete-case data. Figure 6(c) summarizes the results. Comparing the figures, one can see that the imputation estimator is more efficient than the one that uses only complete cases. The reduction in the variance is evident.

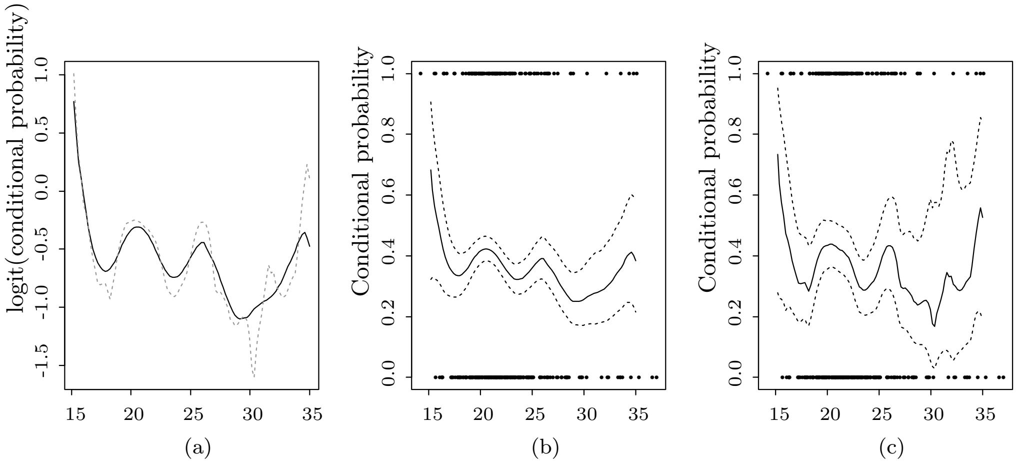

Figure 6. Analysis of CPP data where Y is level of PCBs and X is the body mass index. (a) estimated logit function. Solid curve is the local quasi-likelihood estimator with the imputed data, and dashed curve is the local quasi-likelihood estimator with the complete-case data; (b) estimated conditional probability function by the local quasi-likelihood estimator with the imputed values and the observed data; (c) estimated conditional probability function of the local estimator with the complete-case data and the observed data.

7. Concluding Remarks

Local maximum quasi-likelihood estimation is a useful technique for nonparametric fitting of generalized linear models. We extend this technique to handle data with the response variable missing at random. Three local quasi-likelihood estimators have been proposed. We have shown that the locally weighted quasi-likelihood estimator does not provide any improvement over the local quasi-likelihood estimator with the complete-case data. The proposed imputation method, on the other hand, outperforms the other two methods. The local imputation method provides a more stable estimate. Specifically, the local imputation method is more efficient than the one that uses only complete cases for the finite small sample. The improvement does not merely come from incomplete observations, but from the fact that imputed data allow for more stable choices of the optimal bandwidth.

We have developed a class of quasi-likelihood imputation estimations of the mean functional with the use of the imputed values. It is shown that the proposed mean estimators are asymptotically normal with asymptotic variance that can be easily estimated. It has a higher order asymptotic bias than the kernel imputed method. Simulation studies have shown that the proposed imputation estimators perform well. In the implementation of the local linear regression smoothers, a data-driven bandwidth selection with the missing data has been established. Numerical examples in Section 5 show that the choices of bandwidth parameters are useful.

Acknowledgement

The authors would like to thank the Co-Editor and two referees for helpful comments and suggestions. This research was supported in part by a grant from the National Institutes of Health (CA 79949).

Appendix. Proofs of Theorems 1, 2 and 3

We now impose some regularity conditions for the following theorems.

Let $q_l(x, y) = (\partial^l/\partial x^l)Q\{g^{-1}(x), y\}$, $l = 1, 2, 3$. Then

$$q_1(x, y) = \{y - g^{-1}(x)\}\rho_1(x) \quad \text{and} \quad q_2(x, y) = \{y - g^{-1}(x)\}\rho_1'(x) - \rho_2(x), \quad \text{(A.1)}$$

where $\rho_l(x) = \left[g'(m(x))^l V(m(x))\right]^{-1}$, $l = 1, 2$.

For each given point x, the following conditions are needed.

Condition 1.

(1) The function $q_2(x, y) < 0$ for $x \in \Re$ and y in the range of the response variable.

(2) The functions $f(\cdot), \pi(\cdot), \eta^{(p+1)}(\cdot), V'(\cdot)$ and $g'''(\cdot)$ exist and are continuous at the point x.

(3) Assume that $\rho_2(x) \neq 0$, $Var(Y|X = x) \neq 0$, $g'\{m(x)\} \neq 0$, $f(x) \neq 0$ and $\pi(x) \neq 0$.

(4) $E(Y^4|X = .)$ is bounded in a neighborhood of the point x.

(5) K has bounded support.

Condition 2.

(1) The density function of X has a continuous second derivative.

(2) The function $\eta^{(p+1)}(\cdot)$ is continuous on its support D, which is assumed to be bounded.

(3) The function $V''(\cdot)$ is continuous.

(4) The function $\rho_2(x)$ is twice differentiable in $x \in D$.

In order to prove Theorems 2 and 3, the following lemmas on uniform convergence are needed.

Lemma 1. *Let* $(\mathbf{X}_1, Y_n), \ldots, (\mathbf{X}_n, Y_n)$ *be i.i.d. random vectors, where the* $Y_i's$ *are scalar random variables. Assume that* $E|Y|^s < \infty$ *and* $\sup_x \int |y|^s f(x,y)dy < \infty$, *where* f *denotes the joint density of* $(\mathbf{X}, Y)$. *Let* K *be a bounded positive function with a bounded support, satisfying a Lipschitz condition. Then,*

$$\sup_{\mathbf{x}\in D}\left| n^{-1}\sum_{i=1}^{n}\{K_{h_0}(\mathbf{X_i}-\mathbf{x})Y_i - E[K_{h_0}(\mathbf{X}_i-\mathbf{x})Y_i]\}\right| = O_p\left\{\left[\frac{nh_0}{\log(\frac{1}{h_0})}\right]^{-\frac{1}{2}}\right\},$$

provided that $n^{2\varepsilon-1}h_0 \to \infty$ *for* $\varepsilon < 1 - s^{-1}$.

Proof. It follows immediately from the result obtained by Mack and Silverman (1982).

Lemma 2. *Under the assumptions in Theorem* 3, *we have*

$$\sup_{x\in D}\left|\hat{\eta}(x) - \eta(x) - \frac{1}{nf(x)\pi(x)\rho_2(x)}\sum_{i=1}^{n} W_i K_{h_0}(X_i - x)\right|$$
$$= O_p\left\{h_0^{p+1}c_n + c_n^2 \log^{\frac{1}{2}}(\frac{1}{h_0})\right\},$$

where W_i *is the first element of the vector* $\delta_i q_1(\bar{\eta}_i(x), Y_i)S^{-1}\mathbf{Z}_i$, *and* $c_n = (nh_0)^{-1/2}$.

Proof. The proof can be completed by using Lemma 1 and Lemma A.1 in Carroll, Fan, Gijbels and Wand (1997).

Proof of Theorem 1. The proof of this theorem can be completed by methods similar to those in Fan, Heckman and Wand (1995).

Proof of Theorem 2. First of all, we consider the normalized estimator

$$\hat{\beta}_I^* = a_n^{-1}H_2(\hat{\beta}_I(x) - \beta(x)) = a_n^{-1}(\hat{\beta}_0 - \eta(x), \ldots, h_2^p\{\hat{\beta}_p - \frac{\eta^{(p)}(x)}{p!}\})^T,$$

where $a_n = (\sqrt{nh_2})^{-1}$. Let $\mathbf{Z}_i^* = (1, (X_i - x)/h_2, \ldots, (X_i - x)^p/h_2^p)^T$, $\hat{Y}_i^* = \delta_i Y_i + (1-\delta_i)g^{-1}(\hat{\eta}(X_i))$, $Y_i^* = \delta_i Y_i + (1-\delta_i)g^{-1}(\eta(X_i))$. Then, $\hat{\beta}_I^*$ maximizes the normalized function

$$\ell_n(\beta_I^*) \equiv h_2 \sum_{i=1}^n \left\{ Q\left[g^{-1}\{\bar{\eta}_i(x) + a_n \beta_I^{*^T} \mathbf{Z}_i^*\}, \hat{Y}_i^*\right] - Q\left[g^{-1}\{\bar{\eta}_i(x), \hat{Y}_i^*\right]\right\} \\ \times K_{h_2}(X_i - x). \qquad (A.2)$$

By Taylor expansion of $Q[g^{-1}\{\cdot\}, \hat{Y}_i^*]$, it follows that

$$\ell_n(\beta_I^*) = \mathbf{V}_n^T \beta_I^* + \frac{1}{2}\beta_I^* \mathbf{B}_n \beta_I^* \{1 + o_p(1)\}, \qquad (A.3)$$

where $\mathbf{V}_n = h_2 a_n \sum_{i=1}^n q_1\left(\bar{\eta}_i(x), \hat{Y}_i^*\right)\mathbf{Z}_i^* K_{h_2}(X_i - x)$, and $\mathbf{B}_n = h_2 a_n^2 \sum_{i=1}^n q_2\left(\bar{\eta}_i(x), \hat{Y}_i^*\right)\mathbf{Z}_i^* \mathbf{Z}_i^{*^T} K_{h_2}(X_i - x)$. Now define $\| \hat{\eta} - \eta \|_\infty = \sup_{x \in D} |\hat{\eta}(x) - \eta(x)|$. By Lemma 2, $\| \hat{\eta} - \eta \|_\infty = O_p(h_0^{p+1})$. It can be shown that

$$\mathbf{B}_n = -\rho_2(x) f(x) (\mu_{i+j-2})_{1 \le i,j \le p+1} + o_p(1) \equiv -\mathbf{B} + o_p(1). \qquad (A.4)$$

In fact,

$$\left| \mathbf{B}_n - h_2 a_n^2 \sum_{i=1}^n q_2\left(\bar{\eta}_i(x), Y_i^*\right)\mathbf{Z}_i^* \mathbf{Z}_i^{*^T} K_{h_2}(X_i - x) \right| \\ \le O_P\left(\| \hat{\eta} - \eta \|_\infty\right) \left| \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i^* \mathbf{Z}_i^{*^T} K_{h_2}(X_i - x)\right| = o_p(1).$$

By $(B_n^*)_{ij} = (EB_n^*)_{ij} + O_p(\{\text{Var}\,(B_n^*)_{ij}\}^{1/2})$, we have

$$h_2 a_n^2 \sum_{i=1}^n q_2\left(\bar{\eta}_i(x), \hat{Y}_i^*\right)\mathbf{Z}_i^* \mathbf{Z}_i^{*^T} K_{h_2}(X_i - x) \\ = E\left[q_2\left(\bar{\eta}_1(x), \hat{Y}_1^*\right)\mathbf{Z}_1^* \mathbf{Z}_1^{*^T} K_{h_2}(X_1 - x)\right] + o_p(1) \\ = E\left[q_2\left(\bar{\eta}_1(x), m(X_1)\right)\mathbf{Z}_1^* \mathbf{Z}_1^{*^T} K_{h_2}(X_1 - x)\right] + o_p(1) \equiv -\mathbf{B} + o_p(1).$$

This establishes the result in (A.4). On the other hand, we have

$$\mathbf{V}_n = h_2 a_n \sum_{i=1}^n q_1\left(\bar{\eta}_i(x), Y_i^*\right)\mathbf{Z}_i^* K_{h_2}(X_i - x) \\ + h_2 a_n \sum_{i=1}^n \left[q_1\left(\bar{\eta}_i(x), \hat{Y}_i^*\right) - q_1\left(\bar{\eta}_i(x), Y_i^*\right)\right]\mathbf{Z}_i^* K_{h_2}(X_i - x) \\ \equiv V_{n1} + V_{n2}.$$

By Lemma 2, the second term in the above expression can be expressed as

$$\begin{aligned}
V_{n2} &= h_2 a_n \sum_{i=1}^{n} (1-\delta_i)[g^{-1}(\eta(X_i))]'(\hat{\eta}(X_i)-\eta(X_i))\rho_1(\bar{\eta}_i(x))\mathbf{Z}_i^* K_{h_2}(X_i-x) \\
&\quad + O_P\left\{(nh_2)^{1/2} \parallel \hat{\eta}-\eta \parallel_\infty^2\right\} \\
&= h_2 a_n \sum_{i=1}^{n} \frac{(1-\delta_i)\rho_1(\bar{\eta}_i(x))}{nf(X_i)\pi(X_i)\rho_1(X_i)} \\
&\quad \times \sum_{j=1}^{n} e_1^T \delta_j q_1(\bar{\eta}_j(X_i),Y_j) S^{-1}\mathbf{Z}_j\mathbf{Z}_i^* K_{h_2}(X_i-x)K_{h_0}(X_j-X_i) \\
&\quad + O_p\left\{(nh_2)^{1/2} h_0^{p+1} c_n + (nh_2)^{\frac{1}{2}} c_n^2 \log^{\frac{1}{2}}(\frac{1}{h_0})\right\} + O_P\left\{(nh_2)^{\frac{1}{2}} h_0^{2(p+1)}\right\} \\
&\equiv T_n + o_p(1).
\end{aligned}$$

By Taylor's expansion, $\bar{\eta}_j(X_i) - \eta(X_j) = -\eta^{(p+1)}(X_i)(X_j-X_i)^{p+1}/(p+1)! + o_p\left\{(X_j-X_i)^{p+1}\right\}$, and therefore

$$\begin{aligned}
T_n &= h_2 a_n \sum_{j=1}^{n} \delta_j q_1(\eta(X_j),Y_j) \\
&\qquad \times \left[\sum_{i=1}^{n} \frac{(1-\delta_i)\rho_1(\bar{\eta}_i(x))}{nf(X_i)\pi(X_i)\rho_1(X_i)} e_1^T S^{-1}\mathbf{Z}_j\mathbf{Z}_i^* K_{h_2}(X_i-x)K_{h_0}(X_j-X_i)\right] \\
&\quad - \frac{h_2 a_n}{(p+1)!} \sum_{j=1}^{n} \delta_j q_2(\eta(X_j),Y_j) \\
&\qquad \times \left[\sum_{i=1}^{n} \frac{(1-\delta_i)\rho_1(\bar{\eta}_i(x))}{nf(X_i)\pi(X_i)\rho_1(X_i)} \eta^{(p+1)}(X_i)(X_j-X_i)^{p+1}\right. \\
&\qquad\qquad \left. \times e_1^T S^{-1}\mathbf{Z}_j\mathbf{Z}_i^* K_{h_2}(X_i-x)K_{h_0}(X_j-X_i)\right] \\
&\quad + o_p\left\{(nh_2)^{\frac{1}{2}} h_0^{p+1}\right\} \\
&\equiv T_{n1} + T_{n2} + o_p\left\{(nh_2)^{\frac{1}{2}} h_0^{p+1}\right\}.
\end{aligned}$$

It can be shown via calculating the second moment that

$$T_{n1} - T'_{n1} \xrightarrow{p} 0 \text{ and } T_{n2} - T'_{n2} \xrightarrow{p} 0, \tag{A.5}$$

where

$$T'_{n1} = h_2 a_n \sum_{j=1}^{n} \delta_j q_1(\eta(X_j), Y_j) \frac{(1-\pi(X_j))\rho_1(\bar{\eta}_j(x))}{\pi(X_j)\rho_1(X_j)} e_1^T S^{-1} U_0 \mathbf{Z}_j^* K_{h_2}(X_j - x),$$

$$T'_{n2} = -\frac{h_0^{p+1} h_2 a_n}{(p+1)!} \sum_{j=1}^{n} \delta_j q_2(\eta(X_j), Y_j) \times \frac{(1-\pi(X_j))\rho_1(\bar{\eta}_j(x))}{\pi(X_j)\rho_1(X_j)} \eta^{(p+1)}(X_j) e_1^T S^{-1} U_{p+1} \mathbf{Z}_j^* K_{h_2}(X_j - x),$$

with $U_0 = (1, \mu_1, \ldots, \mu_p)^T$ and $e_1^T S^{-1} U_0 = 1$. Combining (A.3)-(A.5), we have $\ell_n(\beta_I^*) = (V_{n1} + T'_{n1} + T'_{n2})^T \beta_I^* - \beta_I^* \mathbf{B} \beta_I^*/2 + o_p(1)$. Then, applying the same citation as before, we get

$$\hat{\beta}_I^* = \mathbf{B}^{-1}(V_{n1} + T'_{n1} + T'_{n2}) + o_p(1). \tag{A.6}$$

Since $V_{n1} + T'_{n1} + T'_{n2}$ is a sum of i.i.d. random vectors, we can establish the asymptotic normality of $\hat{\beta}_I^*$ by calculating the first two moments. Similar to the proof of Theorem 1, it is easy to show that

$$\begin{aligned} &E(V_{n1} + T'_{n1} + T'_{n2}) \\ &= a_n^{-1} E\left\{q_1\left(\bar{\eta}_1(x), \delta_1 Y_1 + (1-\delta_1) g^{-1}(\eta(X_1))\right) \mathbf{Z}_1^* K_{h_2}(X_1 - x)\right\} \\ &\quad - \frac{h_0^{p+1} a_n^{-1}}{(p+1)!} E\Bigg\{\delta_1 q_2(\eta(X_1), Y_1) \frac{(1-\pi(X_1))\rho_1(\bar{\eta}_1(x))}{\pi(X_1)\rho_1(X_1)} \eta^{(p+1)}(X_1) e_1^T S^{-1} U_{p+1} \\ &\qquad \times \mathbf{Z}_1^* K_{h_2}(X_1 - x)\Bigg\} \\ &= \sqrt{nh_2} \frac{\eta^{(p+1)}(x) h_2^{p+1}}{(p+1)!} \rho_2(x) f(x) U_{p+1} \{1 + o_p(1)\} \\ &\quad + \sqrt{nh_2} \frac{\eta^{(p+1)}(x) h_0^{p+1}}{(p+1)!} (1-\pi(x)) \rho_2(x) f(x) e_1^T S^{-1} U_{p+1} U_0 \{1 + o_p(1)\}, \end{aligned}$$

$$\begin{aligned} &\mathrm{Var}\,(V_{n1} + T'_{n1} + T'_{n2}) \\ &= h_2 \mathrm{Var}\Bigg\{q_1\left(\bar{\eta}_1(x), \delta_1 Y_1 + (1-\delta_1) g^{-1}(\eta(X_1))\right) \mathbf{Z}_1^* K_{h_2}(X_1 - x) \\ &\quad + \delta_1 q_1(\eta(X_1), Y_1) \frac{(1-\pi(X_1))\rho_1(\bar{\eta}_1(x))}{\pi(X_1)\rho_1(X_1)} \mathbf{Z}_1^* K_{h_2}(X_1 - x) + O_p(h_0^{p+1}) + O_p(h_2^{p+1})\Bigg\} \end{aligned}$$

$$
\begin{aligned}
&= \Bigg[\pi(x)\mathrm{Var}\,(Y|x)\rho_1(x)^2 f(x)S^* + 2(1-\pi(x))\mathrm{Var}\,(Y|x)\rho_1(x)^2 f(x)S^* \\
&\quad +\pi(x)\mathrm{Var}\,(Y|x)\rho_1(x)^2\frac{(1-\pi(x))^2}{\pi(x)^2}f(x)S^*\Bigg]\left\{1+o_p(1)\right\} \\
&= \frac{\rho_2(x)f(x)}{\pi(x)}S^*\left\{1+o_p(1)\right\} \\
&\equiv \mathbf{W}\left\{1+o_p(1)\right\}.
\end{aligned}
$$

It can be shown that Liapounov's condition is satisfied and hence

$$
\Big(\hat\beta_I^* - \mathbf{B}^{-1}E(V_{n1}+T'_{n1}+T'_{n2})\Big) \to N\Big(0,\mathbf{B}^{-1}\mathbf{W}\mathbf{B}^{-1}\Big).
$$

We have established Theorem 4.

Proof of Theorem 3. (i) We prove Theorem 3 for $\hat\theta_G$. It follows from (3.2) that

$$
\begin{aligned}
\sqrt{n}\left(\hat\theta_G-\theta\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\left\{g^{-1}(\eta(X_i))-\theta\right\} + \frac{1}{\sqrt{n}}\sum_{i=1}^n \delta_i\left\{Y_i - g^{-1}(\eta(X_i))\right\} \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^n(1-\delta_i)\left\{g^{-1}(\hat\eta(X_i)) - g^{-1}(\eta(X_i))\right\} \\
&\equiv I_1 + I_2 + I_3. \qquad\qquad \text{(A.7)}
\end{aligned}
$$

By Lemma 2, the third term in the above expression can be written as

$$
\begin{aligned}
I_3 &= \frac{1}{\sqrt{n}}\sum_{i=1}^n(1-\delta_i)[g^{-1}(\eta(X_i))]'\left\{\hat\eta(X_i)-\eta(X_i)\right\} + O_P\left(n^{\frac12}\parallel \hat\eta-\eta\parallel_\infty^2\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n(1-\delta_i)\frac{[g^{-1}(\eta(X_i))]'}{nf(X_i)\pi(X_i)\rho_2(X_i)}\sum_{j=1}^n e_1^T\delta_j q_1(\bar\eta_j(X_i),Y_j)S^{-1}\mathbf{Z}_jK_{h_0}(X_j-X_i) \\
&\quad + O_p\left\{n^{\frac12}h_0^{p+1}c_n + n^{\frac12}c_n^2\log^{\frac12}\left(\frac{1}{h_0}\right)\right\} + O_P\left(n^{\frac12}h_0^{2(p+1)}\right) \\
&\equiv I_{31} + o_p(1).
\end{aligned}
$$

Since $\bar\eta_j(X_i) - \eta(X_j) = O_p\left\{(X_j-X_i)^{p+1}\right\}$, we obtain

$$
\begin{aligned}
I_{31} &= \frac{1}{\sqrt{n}}\sum_{j=1}^n e_1^T\delta_j q_1(\eta(X_j),Y_j)S^{-1} \\
&\quad \times\left[\sum_{i=1}^n(1-\delta_i)\frac{g'\{m(X_i))\}\mathrm{Var}\,(Y_i|X_i)}{nf(X_i)\pi(X_i)}\mathbf{Z}_jK_{h_0}(X_j-X_i)\right] + O_P\left(n^{\frac12}h_0^{p+1}\right) \\
&\equiv I_{32} + O_p\left(n^{\frac12}h_0^{p+1}\right).
\end{aligned}
$$

By calculating the second moment, it can be shown that

$$I_{32} - I'_{32} \xrightarrow{p} 0, \tag{A.8}$$

where $I'_{32} = n^{-1/2}\sum_{j=1}^{n} \delta_j q_1(\eta(X_j), Y_j) e_1^T S^{-1} U_0 (1-\pi(X_j))/[\pi(X_j)\rho_1(X_j)]$. Combining (A.7), (A.8) and the result $e_1^T S^{-1} U_0 = 1$ leads to

$$\begin{aligned}
&\sqrt{n}\left(\hat{\theta}_G - \theta\right)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\left[g^{-1}(\eta(X_i)) - \theta\right] + \left[\delta_i\left(Y_i - g^{-1}(\eta(X_i)\right)\right]\right.\\
&\quad \left. + \left[\delta_i\left(\frac{1}{\pi(X_i)} - 1\right)\left(Y_i - g^{-1}(\eta(X_i))\right) e_1^T S^{-1} U_0\right]\right\} + O_p\left(n^{\frac{1}{2}} h_0^{p+1}\right).\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\left[g^{-1}(\eta(X_i)) - \theta\right] + \left[\frac{\delta_i}{\pi(X_i)}\left(Y_i - g^{-1}(\eta(X_i)\right)\right]\right\} + O_p\left(n^{\frac{1}{2}} h_0^{p+1}\right). \quad \text{(A.9)}
\end{aligned}$$

By the Central Limit Theorem, $\hat{\theta}_G$ is asymptotically normal.

(ii) We prove Theorem 3 for $\hat{\theta}_{GW}$. Based on (3.3), we have

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_{GW} - \theta\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{g^{-1}(\eta(X_i)) - \theta\right\} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{\pi(X_i)}\left\{Y_i - g^{-1}(\eta(X_i))\right\}\\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1 - \frac{\delta_i}{\pi(X_i)}\left\{g^{-1}(\hat{\eta}(X_i)) - g^{-1}(\eta(X_i))\right\}\\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i(\hat{\pi}(X_i) - \pi(X_i))}{\pi(X_i)}\left\{Y_i - g^{-1}(\eta(X_i))\right\} + o_p(1)\\
&\equiv J_1 + J_2 + J_3 + J_4 + o_p(1). \qquad \text{(A.10)}
\end{aligned}$$

Similar to the proof of (i), the third term can be reduced to

$$\begin{aligned}
J_3 &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n} e_1^T \delta_j q_1(\eta(X_j), Y_j) S^{-1}\\
&\quad \times \left[\sum_{i=1}^{n}\left(1 - \frac{\delta_i}{\pi(X_i)}\right)\frac{g'\{m(X_i))\}\mathrm{Var}\,(Y|X_i)}{nf(X_i)\pi(X_i)}\mathbf{Z}_j K_{h_0}(X_j - X_i)\right] + o_p(1)\\
&= O_p\left(n^{\frac{1}{2}} h_0^{p+1}\right). \qquad \text{(A.11)}
\end{aligned}$$

Next, by using the kernel estimator $\hat{\pi}(X_i) = \sum_{j=1}^{n}\delta_j K_a(X_j - Xi)/\sum_{j=1}^{n} K_a(X_j -$

Xi), we have

$$\begin{aligned}J_4 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i(\hat{\pi}(X_i)-\pi(X_i))}{\pi^2(X_i)}\left\{Y_i - g^{-1}(\eta(X_i))\right\}\\ &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}[\delta_j-\pi(X_j)]\frac{1}{na}\sum_{i=1}^{n}\frac{\delta_j}{\pi^2(X_i)f(X_i)}\left\{Y_i - g^{-1}(\eta(X_i))\right\}K_a(X_i-X_j)\\ &\quad +o_p(1)\\ &= o_p(1). \qquad (A.12)\end{aligned}$$

Therefore, we have

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_{GW}-\theta\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{g^{-1}(\eta(X_i))-\theta\right\} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{\pi(X_i)}\left\{Y_i - g^{-1}(\eta(X_i))\right\}\\ &\quad +O_P\left(n^{\frac{1}{2}}h_0^{p+1}\right). \qquad (A.13)\end{aligned}$$

This together with a limit theorem, establishes Theorem 3 for $\hat{\theta}_{GW}$.

(iii) With the same proofs as in (i) and (ii), we have proved the asymptotic normality of $\hat{\theta}_T$ and $\hat{\theta}_{TW}$. This completes the proof of Theorem 3.

References

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477-489.

Carroll, R. J., Ruppert, D. and Welsh, A. H. (1998). Nonparametric estimation via local estimating equations. *J. Amer. Statist. Assoc.* **93**, 214-227.

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

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing.* 2nd edition. Marcel Dekker, New York.

Fan, J. and Chen, J. (1999). One-step local local quasi-likelihood estimation. *J. Roy. Statist. Soc. Ser. B* **61**, 927-943.

Fan, J., Farmen, M. and Gijbels, I. (1998). A blueprint of local maximum likelihood estimation. *J. Roy. Statist. Soc. Ser. B* **60**, 591-608.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B* **57**, 371-394.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications.* Chapman and Hall, London.

Fan, J., Heckman, N. E. and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90**, 477-489.

Gelman, A., Carlin, J. B., Sterm, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis.* Chapman and Hall, London.

Hastie, T. J. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hunsberger, S. (1994). Semiparametric regression in likelihood-based models. *J. Amer. Statist. Assoc.* **89**, 1354-1365.

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

Little, R. J. A. (1992). Regression with missing X's: A Review. *J. Amer. Statist. Assoc.* **87**, 1227-1237.

Longnecker, M. P., Klebnoff, M. A., Zhou, H. and Brock J. W. (2002). Maternal serum level of the DDT metabolite DDE is associated with preterm and small-for-gestational-age birth. *Amer. J. Epidemiology* **155**, 313-322.

Mack, Y. P. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression and density estimation. *Z. Wahrsch. Verw. Gebiete* **61**, 405-415.

Matloff, N. S. (1981). Use of regression functions for improved estimation of means. *Biometrika* **68**, 685-689.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Niswander, K. R. and Gordon, M. (1972). *The Women and Their Pregnancies*. USDHEW Publication No. (NIH) 73-379. USGPO, Washington, DC.

Paik, M. C. (1997). The generalized estimating equation approach when data are not completely missing at random. *J. Amer. Statist. Assoc.* **92**, 1320-1329.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41-55.

Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.* **89**, 501-511.

Seifert, B. and Gasser, T. (1996). Finite-sample variance of local polynomials: Analysis and solutions. *J. Amer. Statist. Assoc.* **91**, 267-275.

Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *J. Amer. Statist. Assoc.* **84**, 276-283.

Tamhane, A. C. (1978). Inference based on regression estimator in double sampling. *Biometrika* **64**, 419-428.

Wang, C. Y., Wang, S., Gutierrez, R. G. and Carroll, R. J. (1998). Local linear regression for generalized linear models with missing data. *Ann. Statist.* **26**, 1028-1050.

Wang, C. Y., Wang, S., Zhao, L. P. and Ou, S. T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate. *J. Amer. Statist. Assoc.* **92**, 512-525.

Wang, Q. H. and Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation with missing response. *Ann. Statist.* **30**, 896-924.

Wang, Q. H., Linton, O. and Härdle, W. (2004). Semiparametric regression analysis with missing response at random. *J. Amer. Statist. Assoc.* **99**, 334-345.

Wedberburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and Gauss-Newton method. *Biometrika* **61**, 439-447.

Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Avenue, Box 630, Rochester, New York 14642, U.S.A.

E-mail: jchen@bst.rochester.edu

Department of Operation Research and Financial Engineering, Princeton University, Princeton, NJ 08544, U.S.A.

E-mail: jqfan@princeton.edu

Department of Statistics, The Chinese University of Hong Kong, Hong Kong.

E-mail: khli@cuhk.edu.hk

Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420, U.S.A.

E-mail: zhou@bios.unc.edu

(Received July 2004; accepted April 2005)