The Annals of Statistics
1998, Vol. 26, No. 3, 1028–1050

LOCAL LINEAR REGRESSION FOR GENERALIZED LINEAR MODELS WITH MISSING DATA

By C. Y. Wang,[1] Suojin Wang,[2] Roberto G. Gutierrez and R. J. Carroll[3]

Fred Hutchinson Cancer Research Center, Texas A&M University, Southern Methodist University and Texas A&M University

Fan, Heckman and Wand proposed locally weighted kernel polynomial regression methods for generalized linear models and quasilikelihood functions. When the covariate variables are missing at random, we propose a weighted estimator based on the inverse selection probability weights. Distribution theory is derived when the selection probabilities are estimated nonparametrically. We show that the asymptotic variance of the resulting nonparametric estimator of the mean function in the main regression model is the same as that when the selection probabilities are known, while the biases are generally different. This is different from results in parametric problems, where it is known that estimating weights actually decreases asymptotic variance. To reconcile the difference between the parametric and nonparametric problems, we obtain a second-order variance result for the nonparametric case. We generalize this result to local estimating equations. Finite-sample performance is examined via simulation studies. The proposed method is demonstrated via an analysis of data from a case-control study.

1. Introduction. This paper is concerned with nonparametric function estimation via quasilikelihood when the predictor variable may be missing, and the missingness depends upon the response. We use local polynomials with kernel weights, generalizing the work of Staniswalis (1989), Severini and Staniswalis (1994) and Fan, Heckman and Wand (1995) to the missing data problem.

In practice, covariates may be missing due to reasons such as loss to follow up. For example, in a study of acute graft versus host disease of bone marrow transplants of 97 female subjects conducted at the Fred Hutchinson Cancer Research Center, the outcome is the acute graft host disease and one covariate of interest is the donor's previous pregnancy status, which was missing for 31 patients because of the incompleteness of the donors' medical history. In

Received January 1997; revised January 1998.

[1] Supported by National Cancer Institute Grants CA-53996 and AG 15026.

[2] Supported by NSF Grant DMS-95-04589, NSA Grant MDA904-96-1-0029, National Cancer Institute Grant CA-57030 and Texas A&M University's Scholarly and Creative Activities Program (95-59).

[3] Supported by National Cancer Institute Grant CA-57030 and partially completed while visiting the Institut für Statistik und Ökonometrie, Sonderforschungsbereich 373, Humboldt Universität zu Berlin, with partial support from a senior Alexander von Humboldt Foundation research award.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62G20.

Key words and phrases. Generalized linear models, kernel regression, local linear smoother, measurement error, missing at random, quasilikelihood functions.

this paper, we consider the missing covariate data problem in nonparametric generalized linear models. We assume that covariates are missing at random (MAR). In this case, the missingness is ignorable [Rubin (1976)] so that the missingness mechanism depends only on the observed data but not the missing data.

In parametric problems, two approaches are common. Likelihood methods assume a joint parametric distribution for covariates and response, and under our assumptions ignore the missing data mechanism [Little and Rubin (1987)]. Complete-case analysis assumes nothing about the distribution of covariates, and is in this sense semiparametric. Estimation is based on the "complete-cases," that is, those with no missing data, with weighting inversely proportional to the probability that the covariate is observed given the response [Horvitz and Thompson (1952)]. We call these *selection probabilities*. We use the second approach. Our methods apply as well to other semiparametric schemes, for example, that of Robins, Rotnitzky and Zhao (1994). In this paper, the missing data probabilities are also modeled via a generalized linear model. We estimate the missing data probabilities by nonparametric regression.

In parametric problems, the Horvitz–Thompson weighting scheme has a curious and important property. Consider two estimators: (a) the one with known selection probabilities and weights; and (b) one where the selection probabilities are estimated by a properly specified parametric model. The two methods yield consistent estimates, but that with estimated weights generally has a *smaller* asymptotic variance [Robins, Rotnitzky and Zhao (1994)]. A heuristic argument of this phenomenon was given in Robins, Rotnitzky and Zhao [(1994), Section 6.1]. However, it is a somewhat counterintuitive finding. In the parametric case, neither estimator has an asymptotic bias problem.

One might expect the same sort of result to hold in the nonparametric regression case with nonparametrically estimated selection probabilities, especially in view of the work of Wang, Wang, Zhao and Ou (1997). However, this is not the case, and we show (Theorem 1) that whether weights are estimated or not has no effect on asymptotic variance, while it does have an effect on the bias in general.

In simulations, however, we observed repeatedly that estimating weights was beneficial in the sense that the resulting estimator of the mean function in the main regression model is more efficient than that using true weights for small to moderate samples. To understand whether this numerical evidence was at all general, we developed a second-order variance result (Theorem 2) showing that the estimator with estimated weights can be expected to have smaller finite-sample variance than if the weights are known. This second-order variance result provides a reconciliation between the different first-order results in the parametric and nonparametric cases. This phenomenon could also be because the local regression estimation is effectively finite dimensional.

The statistical models are described in Section 2. In Section 3, we propose the methodology and the asymptotic result for the weighted method with both known and estimated selection probabilities. The method is demonstrated in

Section 4 by analyzing the data from a case-control study of bladder cancer. In Section 5, we investigate the finite-sample performance by conducting a simulation study. We note that estimating the selection probabilities has a finite-sample effect on the estimation of the mean function of our primary interest. We explain the possible finite-sample efficiency gain by a second-order variance approximation in Section 6.

The major result of Section 3 can be described as follows:

1. An unknown function $\pi(\cdot)$ is estimated nonparametrically, by $\widehat{\pi}(\cdot)$.
2. If $\pi(\cdot)$ were known, one would use it to estimate nonparametrically a second function $\mu(\cdot)$, by $\widehat{\mu}(\cdot, \pi)$.
3. The estimates $\widehat{\mu}(\cdot, \pi)$ and $\widehat{\mu}(\cdot, \widehat{\pi})$ have the same asymptotic variance.

In Section 7, we sketch a result showing that this phenomenon is quite general, and not restricted to our particular context. All detailed proofs are given in the Appendix.

2. The models.

2.1. *Full data models.* We let $(Y_1, X_1), \dots, (Y_n, X_n)$ be a set of independent random variables, where Y_i is a scalar response variable, and X_i is a scalar covariate variable. In a classical generalized linear model [(Nelder and Wedderburn (1972); McCullagh and Nelder (1989)], the conditional density (or probability mass function) of Y given X belongs to a canonical exponential family $f_{Y|X}(y|x) = \mathscr{C}(y)\exp[y\theta(x) - \mathscr{B}\{\theta(x)\}]$ for known functions $\mathscr{B}$ and $\mathscr{C}$, where the function θ is called the canonical or natural parameter. The unknown function $\mu(x) = E(Y|X = x)$ is modeled in X by a link function g by $g\{\mu(x)\} = \eta(x)$. In a parametric generalized linear model, $\eta(x) = c_0 + c_1 x$ for some unknown parameter c_0, c_1. The link function g is assumed to be known. For example, in logistic regression $g(u) = \log\{u/(1-u)\}$, and in linear regression $g(u) = u$. In our nonparametric setting, there is no model assumption about $\eta(x)$.

Fan, Heckman and Wand (1995) considered quasilikelihood models, where only the relationship between the mean and the variance is specified. If the conditional variance is modeled as $\mathrm{var}(Y|X = x) = V\{\mu(x)\}$, for some known positive function V, then the corresponding quasilikelihood function $Q(w, y)$ satisfies $(\partial/\partial w)Q(w, y) = (y - w)/V(w)$ [Wedderburn (1974)]. The primary interest is to estimate $\mu(x)$, or equivalently $\eta(x)$, nonparametrically.

2.2. *Missing data models.* In a missing covariate data problem, some covariates may be missing and we let $\delta_i = 1$ if X_i is observed, $\delta_i = 0$ otherwise. Furthermore, let

$$\pi_i = \mathrm{pr}(\delta_i = 1|Y_i, X_i) = \mathrm{pr}(\delta_i = 1|Y_i) = \pi(Y_i) \tag{1}$$

be the selection probability which does not depend on X_i, that is, X_i is MAR. In a two-stage design [White (1982)], often the selection probabilities are known. In many missing data problems, however, the selection probabilities are unknown and need to be estimated. To model the selection probabilities, we assume that, given Y, there is a known link function g^* such that $g^*\{\pi(y)\} = \eta^*(y)$, where $\eta^*(y)$ is a smooth function. Let the conditional variance be modeled by $\mathrm{var}(\delta|Y = y) = V^*\{\pi(y)\}$ for some known positive function V^*. The corresponding quasilikelihood function $Q^*(w, \delta)$ satisfies $(\partial/\partial w)Q^*(w, \delta) = (\delta - w)/V^*(w)$. We say that *two-stage data models* occur when the selection probabilities are known, and *missing data models* occur when the selection probabilities are unknown. In the missing data models, $\pi(y)$, or $\eta^*(y)$, is a nuisance component which needs to be estimated.

3. Methodology.

3.1. *The weighted method.* When (Y_i, X_i) are fully observable, Fan, Heckman and Wand (1995) proposed the local linear kernel estimator of $\eta(x)$ as $\widehat{\eta}(x, h) = \widehat{\beta}_0$, where h is the bandwidth of a kernel function K and $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)^t$ maximizes

$$\sum_{i=1}^{n} Q[g^{-1}\{\beta_0 + \beta_1(X_i - x)\}, Y_i]K_h(X_i - x), \tag{2}$$

where $K_h(\cdot) = K(\cdot/h)$. We assume that the maximizer exists, and this can be verified for standard choices of Q. The mean function $\mu(x)$ is estimated by $\widehat{\mu}(x) = g^{-1}(\widehat{\beta}_0)$. When data are missing, a naive method is to apply (2) by using the complete-case (CC) analysis, that is, solving (2) by restricting to pairs in which both Y and X are observed. However, complete-case analysis may cause considerable bias when the missingness probabilities (1) depend on the response [Little and Rubin (1987)].

To accommodate the missingness in the observed data, we propose a Horvitz–Thompson inverse-selection weighted method, so that the estimator of β maximizes

$$\sum_{i=1}^{n} Q[g^{-1}\{\beta_0 + \beta_1(X_i - x)\}, Y_i]\frac{\delta_i}{\pi(Y_i)}K_h(X_i - x). \tag{3}$$

Note that here $\pi(Y_i)$ is assumed to be known and strictly positive in the support of Y. For notational purposes, we denote the solution to (3) by $\widehat{\beta}(\pi)$.

We now define some notation for the presentation of the asymptotic properties of $\widehat{\beta}_0 = \widehat{\eta}(x, \pi)$. Suppose that K is supported on $[-1, 1]$. For any set $\mathscr{A} \subset R$, and $i = 0, 1, 2, 3$, let $\gamma_i(\mathscr{A}) = \int_{\mathscr{A}} z^i K(z)\,dz$, $\tau_i(\mathscr{A}) = \int_{\mathscr{A}} z^i K^2(z)\,dz$. Define

$$N_x^h = \{z\colon x - hz \in \mathrm{supp}(f_X)\} \cap [-1, 1],$$

$$b_x = \frac{1}{2}\eta^{(2)}(x)[g'\{\mu(x)\}]^{-1}\frac{\gamma_2^2(N_x^h) - \gamma_1(N_x^h)\gamma_3(N_x^h)}{\gamma_0(N_x^h)\gamma_2(N_x^h) - \gamma_1^2(N_x^h)},$$

$$\sigma_x^2 = f_X^{-1}(x)\mathscr{L}(x)\frac{\gamma_2^2(N_x^h)\tau_0(N_x^h) - 2\gamma_1(N_x^h)\gamma_2(N_x^h)\tau_1(N_x^h) + \gamma_1^2(N_x^h)\tau_2(N_x^h)}{\{\gamma_0(N_x^h)\gamma_2(N_x^h) - \gamma_1^2(N_x^h)\}^2},$$

where $f_X(x)$ is the density of X and

$$\mathscr{I}(x) = E\left[\frac{\{Y_1 - \mu(x)\}^2}{\pi(Y_1)}\Big| X_1 = x\right]. \tag{4}$$

As we will see later, σ_x^2 is the asymptotic variance of $\widehat{\mu}(x, \pi)$. For a bandwidth h, x is an interior point of $\operatorname{supp}(f_X)$ if and only if $N_x^h = [-1, 1]$. To estimate $\mu(x) = g^{-1}\{\eta(x)\}$, we let $\widehat{\mu}(x, \pi) = g^{-1}\{\widehat{\eta}(x, \pi)\} = g^{-1}(\widehat{\beta}_0)$. The limit distribution of $\widehat{\mu}(x, \pi)$ presented in Theorem 1 below can be obtained by calculations similar to those in Fan, Heckman and Wand (1995). Theorem 1 also indicates that the bias b_x is affected by estimating the selection probabilities, while the variance is not.

3.2. *Main theorem.* We now investigate the case with unknown selection probabilities. To estimate the selection probabilities, we again apply the local linear smoother of Fan, Heckman and Wand (1995). For a fixed point y, we estimate $\pi(y)$ by

$$\widehat{\pi}(y) = g^{*^{-1}}(\widehat{\alpha}_0), \tag{5}$$

where $\widehat{\alpha} = (\widehat{\alpha}_0, \widehat{\alpha}_1)$ maximizes $\sum_{i=1}^n Q^*[g^{*^{-1}}\{\alpha_0 + \alpha_1(Y_i - y), \delta_i\}]K_\lambda(Y_i - y)$, where we use λ as the smoothing parameter to distinguish it from the other smoothing parameter h used in estimating β for estimating the primary mean function μ. Note that if the outcome Y is categorical such as the situation in Section 4, then, as $\lambda \to 0$, the estimate of π is equal to the empirical averages.

Let $\widehat{\beta}(\widehat{\pi})$ maximize

$$\sum_{i=1}^n Q[g^{-1}\{\beta_0 + \beta_1(X_i - x)\}, Y_i]\frac{\delta_i}{\widehat{\pi}(Y_i)}K_h(X_i - x), \tag{6}$$

where $\widehat{\pi}(y)$ is given in (5). Similar to the definition of $\widehat{\mu}(x, \pi)$, we define $\widehat{\mu}(x, \widehat{\pi}) = g^{-1}\{\widehat{\eta}(x, \widehat{\pi})\}$ where $\widehat{\eta}(x, \widehat{\pi}) = \widehat{\beta}_0(\widehat{\pi})$. We now present our main result.

THEOREM 1. *Suppose that Conditions* (A1)–(A7) *in the Appendix are satisfied. Then, if $h = h_n \to 0$, $nh^3 \to \infty$ and $\lambda = \lambda_n = c^*h$ for a constant $c^* > 0$, we have that, for any $x \in \operatorname{supp}(f_X)$, there exist $b_{nj}(x) = b_x\{1 + o(1)\}$, $j = 1, 2$, such that $(nh)^{1/2}\{\widehat{\mu}(x, \pi) - \mu(x) - h^2 b_{n1}(x)\}$ converges in distribution to a normal random variable with mean* 0 *and variance σ_x^2. Further, if Conditions* (B1)–(B6) *in the Appendix are also satisfied, then $(nh)^{1/2}\{\widehat{\mu}(x, \widehat{\pi}) - \mu(x) - h^2 b_{n2}(x) - \lambda^2 f_X(x) S_3(x)\}$ converges in distribution to a normal random variable with mean* 0 *and variance σ_x^2, where $S_3(x)$ is given in* (22) *in the Appendix, and $S_3(x) = 0$ if either Y is a lattice random variable or π is a constant.*

One important implication of this result is that the effect on the asymptotic variance due to estimating selection probabilities, which is nonnegligible in the parametric or semiparametric models [Robins, Rotnitzky and Zhao (1994);

Wang et al. (1997)], disappears in the corresponding fully nonparametric problems. The difference appears in the bias term, but it vanishes if either Y is a lattice random variable or π is a constant. The proof of Theorem 1 is in the Appendix.

3.3. *Bandwidth selection.*

BANDWIDTH FOR THE SELECTION PROBABILITY ESTIMATION. Fan, Heckman and Wand (1995) suggested a bandwidth selector based on "plugging-in" estimates of unknown quantities. For the rest of the paper, the notation $\phi^{(k)}(\cdot)$ denotes the kth derivative of a function $\phi(\cdot)$. Because we consider the local linear smoother for π, an approximate asymptotic mean integrated square error for $\widehat{\eta}^*(\lambda)$ is

$$\begin{aligned}\text{AMISE}\{\widehat{\eta}^*(\lambda)\} = {} & \frac{\lambda^4\gamma_2^2}{4}\int\{\eta^{*(2)}(y)\}^2 f_Y(y)\,dy \\ & + (n\lambda)^{-1}\tau_0\int V^*\{\pi(y)\}[g^{*\prime}\{\pi(y)\}]^2\,dy,\end{aligned}$$

where $\gamma_2 = \gamma_2([-1,1])$ and $\tau_0 = \tau_0([-1,1])$ are given in Section 3.1 and $f_Y(y)$ denotes the density of Y. This approximation excludes the boundary regions. With respect to this criterion, by taking the derivative of AMISE with respect to λ, we have that the optimal bandwidth for the estimate of π is then

$$\lambda_{AMISE} = \left[\frac{\tau_0\int V\{\pi(y)\}[g^{*\prime}\{\pi(y)\}]^2\,dy}{n\gamma_2^2\int\{\eta^{*(2)}(y)\}^2 f_Y(y)\,dy}\right]^{1/5}.$$

Note that $\pi(y)$ and $f_Y(y)$ are unknown. An "ad hoc" plug-in bandwidth selection is to estimate $\eta^*(y)$ by a third- (or higher-) degree polynomial parametric fit to the selection probabilities and to estimate $f_Y(y)$ by a usual kernel estimate. We also note that this criterion is an approximation which does not consider the γ_0 and τ_0 as a function of λ on the boundary points. In practice, this selector seems to perform reasonably well for a wide range of functions.

BANDWIDTH FOR THE PRIMARY ESTIMATION. Now we study the bandwidth selection for our primary estimation. As in Fan, Heckman and Wand (1995), by excluding the boundary regions, an approximate asymptotic mean integrated square error for $\widehat{\eta}$ is

$$\text{AMISE}\{\widehat{\eta}(h)\} = \frac{h^4\gamma_2^2}{4}\int\{\eta^{(2)}(x)\}^2 f_X(x)\,dy + (nh)^{-1}\tau_0\int \mathscr{L}(x)[g'\{\mu(x)\}]^2\,dx.$$

With respect to this criterion, by taking the derivative of AMISE with respect to h, we have that the optimal bandwidth for the estimate of μ is then

$$h_{AMISE} = \left[\frac{\tau_0\int \mathscr{L}(x)[g'\{\mu(x)\}]^2\,dx}{n\gamma_2^2\int\{\eta^{(2)}(x)\}^2 f_X(x)\,dx}\right]^{1/5}.$$

Similar to the argument of the selection of λ, we may estimate $\eta(x)$ by a third- (or higher-) degree polynomial. In addition, we may estimate $\mathscr{L}(x) =$

$E[\{Y_1 - \mu(X_1)\}^2/\pi(Y_1)|X_1 = x]$ and $f_X(x)$ by nonparametric estimation based on validation data with inverse selection weights. This gives a global bandwidth selection. Alternatively, Schucany (1995) proposed an adaptive local bandwidth estimator for the Nadaraya–Watson estimator, and found that it has improvements over a global bandwidth estimator. It may be a worthwhile future project to study the local bandwidth selector in the problem of generalized linear missing data models.

4. Data analysis. In this section, we consider an example of a case-control study of bladder cancer conducted at the Fred Hutchinson Cancer Research Center. Eligible subjects were residents of three counties of western Washington state who were diagnosed between January 1987 and June 1990 with invasive or noninvasive bladder cancer. This population-based case-control study was designed to address the association between bladder cancer and some nutrients. We use the data here for illustrative purposes. Some detailed results can be found in Bruemmer, White, Vaughan and Cheney (1996).

In our demonstration, the response variable is the bladder cancer history and the covariate X is the smoking package year. The smoking package year of a participant is defined as the average number of cigarette packages smoked per day multiplied by the years one has been smoking. There are a total of 262 cases and 405 controls. However, the smoking package year information of one case and 215 controls was missing. In addition, we treated past smokers as in the nonvalidation set since we are primarily interested in the smoking effect of current smokers. One case with $X = 200$ has high leverage (X has mean 26 and standard deviation 30) and was not included in the validation set. As a result, there were 167 cases and 179 controls in the validation set.

To analyze the data, one may consider the complete-case logistic regression of Y on X, with and without adjustment by estimated inverse selection weights. The estimates of the slope (s.e.) are .0276 (.0047) and .0268 (.0046), respectively. The resulting estimates of $E(Y|X)$, called global estimates, are given in Figure 1. We note that a parametric estimator is based on global estimation. Based on this logistic regression analysis, one would argue that the risk of developing bladder cancer increases monotonically as a function of the average smoking year.

Alternatively, we may employ the weighted local estimation method. We used the Epanechnikov kernel function that $K(u) = .75(1 - u^2)$ on $[-1, 1]$. The unweighted estimates of $E(Y|X)$, denoted by $\widehat{\mu}_{CC}(\cdot)$, and the weighted estimator, $\widehat{\mu}(\cdot, \widehat{\pi})$, are given in Figure 1. Because Y is binary, $\pi(Y)$ was estimated by the empirical average at the corresponding Y value. Based on the bandwidth selection criteria given in Section 3.3, we used 24.2 as the bandwidth h for the weighted local smoother and 19.6 for the unweighted one. We notice that the CC analysis has basically captured the effect of the average package year, as it is somewhat parallel to $\widehat{\mu}(\cdot, \widehat{\pi})$. Because the missing data are mainly from controls, the unweighted estimator thus overestimates $\mathrm{pr}(Y = 1|X)$ of the case-control data (assuming no missingness). Hence, the unweighted estimator is always above the weighted one. Based on this non-

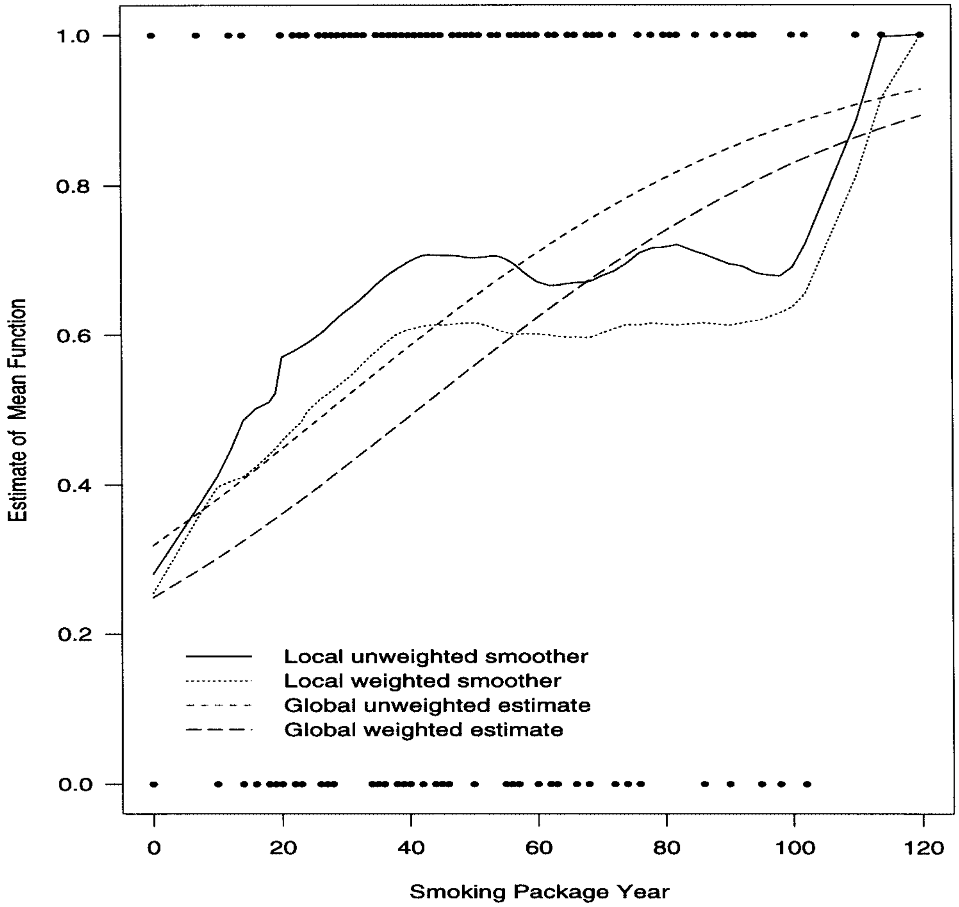

FIG. 1. *Bladder cancer case-control data analysis.*

parametric analysis, the argument is somewhat different from the previous parametric one. For example, the curves between $X = 40$ and $X = 95$ do not increase as much as the other two segments ($X < 40$ or $X > 95$). Although it is true that the average package year has a significant effect on bladder cancer, our analysis suggests that piecewise logistic regression is more proper if parametric inference is to be made.

One small point concerns the interpretation of Figure 1. Prentice and Pyke (1979) showed that in a case-control study with an ordinary parametric logistic regression model, the logits of the observed case-control data differ from that of the population only in the intercept term. The same is true in our problem. This means that the basic monotonicities and flatness observed in Figure 1 are not affected by the case-control sampling, although the levels of estimated disease probability of course would differ. The result of Prentice and Pyke (1979) ignoring missingness is equivalent to that of the global unweighted estimate. Our estimate is the local weighted estimate. They are approximately parallel when the smoking package year (X) is smaller than 40, but not so elsewhere. Because in this example the selection probabilities depend mainly on the disease status, parallelism of weighted and unweighted estimates is expected. The reason that our estimate has a different flatness from that of Prentice and Pyke for $X \geq 40$ is due to the nonparametric estimation, which is not restricted by the linear logit model.

5. Simulation studies. We conducted simulations to better understand the finite-sample performance of the weighted estimator and the finite-sample effect due to estimating the selection probabilities. Recall that $\widehat{\mu}_{CC}$ is the unweighted method which applies the local linear smoother of Fan, Heckman and Wand (1995) directly to the validation set only. We compare the biases and variances of $\widehat{\mu}_{CC}$, $\widehat{\mu}(\cdot, \pi)$ and $\widehat{\mu}(\cdot, \widehat{\pi})$.

We first consider the case of continuous response Y. We generated $n = 200$ X's from a uniform $[-1, 1]$ distribution, and the response variable Y's follow the linear link such that $Y_i = \mu(X_i) + .3\varepsilon_i$, where $\mu(x_i) = x_i^2$, ε_i ($i = 1, \dots, n$) is a random sample from normal $(0, 1)$ distribution and is independent of X_i. The selection probability given Y is from the logistic model with intercept 0.0 and slope 1.0. Approximately 42% of the data are missing under the above selection probabilities. We ran 1,000 independent replicates in this simulation experiment, and we applied the linear link and logit link to estimate $\mu(\cdot)$ and $\pi(\cdot)$, respectively. In each replicate, $\widehat{\mu}_{CC}(\cdot)$, $\widehat{\mu}(\cdot, \pi)$ and $\widehat{\mu}(\cdot, \widehat{\pi})$ were obtained using the Epanechnikov kernel function $K(u) = .75(1 - u^2)$ on $[-1, 1]$ and the bandwidth selection criteria as described in Section 3.3.

The empirical biases of the estimators are shown in Figure 2 for $x \in (-1, 1)$. The curves are the averages of the bias estimates over 1,000 runs. Note that the CC analysis has considerable bias and that $\widehat{\mu}(\cdot, \pi)$ and $\widehat{\mu}(\cdot, \widehat{\pi})$ are very close in most points. Figure 3 shows the sample variances of $\widehat{\mu}(x, \pi)$ and $\widehat{\mu}(x, \widehat{\pi})$. It appears that the weighted estimator using estimated selection probabilities is at least as efficient as the one using the true $\pi(\cdot)$. There is considerable gain using estimated π for a range of X values, especially when X is around zero. The relative efficiency of $\widehat{\mu}(x, \widehat{\pi})$ to $\widehat{\mu}(x, \pi)$ at $x = 0$ is 1.29 when $n = 200$. If we increase the sample size to $n = 2{,}000$, then the corresponding relative efficiency is 1.22. In Section 6, we explain the finite-sample efficiency gain from estimating the selection probabilities by a second-order variance approximation.

We have also investigated the case when the response is binary, and the findings are similar to those for continuous response. We omit the details here.

6. Second-order variance approximation. The simulations in the previous section show that there is finite-sample gain from estimating the selection probabilities. Recall that the first-order asymptotic result of Theorem 1 shows no asymptotic efficiency gain from estimating the selection probabilities. To explain this, we now present the second-order variance approximation. The proof is given in the Appendix.

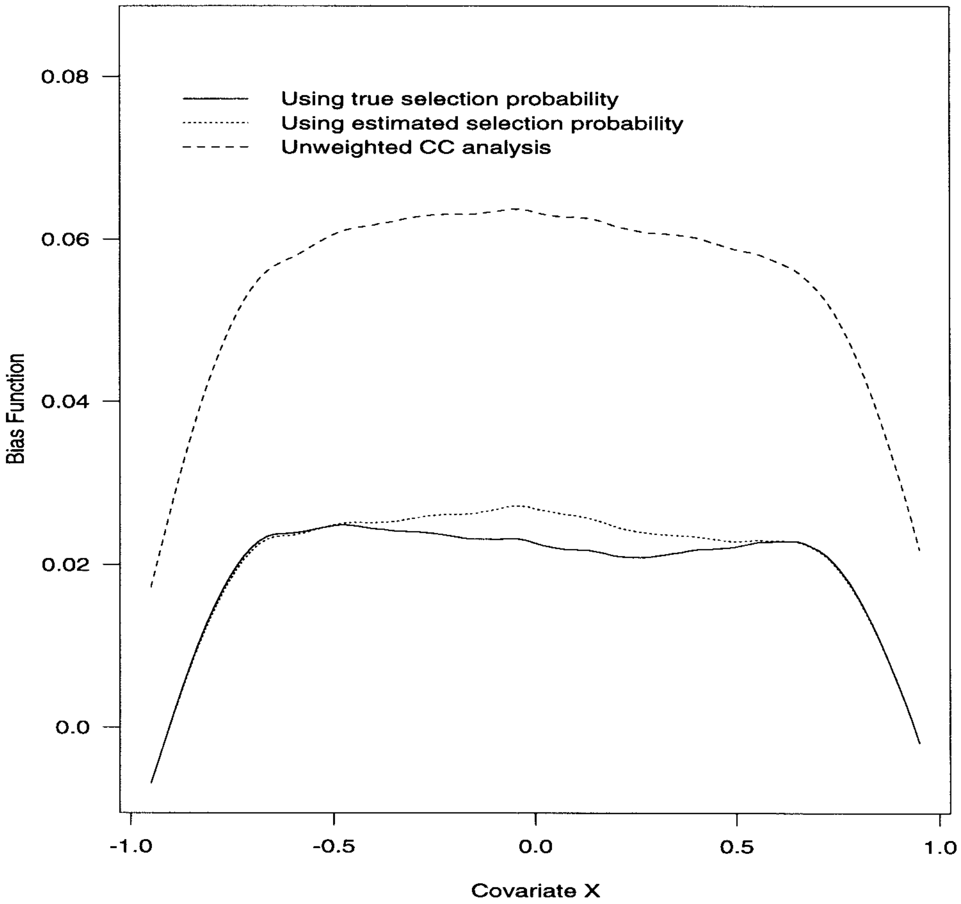

FIG. 2. *Simulation study for biases from estimating μ for continuous response.*

THEOREM 2. *Under the same conditions as in Theorem* 1 *and for any* $x \in \operatorname{supp}(f_X)$ *with* $\operatorname{var}\{\widehat{\mu}(x, \pi)\} < \infty$, *there exists* $\widehat{\mu}_*(x) = \widehat{\mu}(x, \widehat{\pi}) + o_p(h^{1/2}n^{-1/2})$, *such that*

$$\operatorname{var}\{\widehat{\mu}_*(x)\} = \operatorname{var}\{\widehat{\mu}(x, \pi)\} - n^{-1}v(x)\{1 + o(1)\},$$

for some $v(x) > 0$.

Theorem 2 shows that using the estimated selection probabilities improves the efficiency at the rate of n^{-1}. Note that the second-order efficiency gain is valid even when Y is a lattice random variable. For a fixed point x, let the relative efficiency gain by using the estimated selection probabilities be defined by $[\operatorname{var}\{\widehat{\mu}(x, \pi)\} - \operatorname{var}\{\widehat{\mu}_*(x)\}]/\operatorname{var}\{\widehat{\mu}_*(x)\}$. It is easy to see from Theorem 2 that the relative efficiency gain is of order $O(h)$, which goes to zero slowly. This supports the results of our simulations.

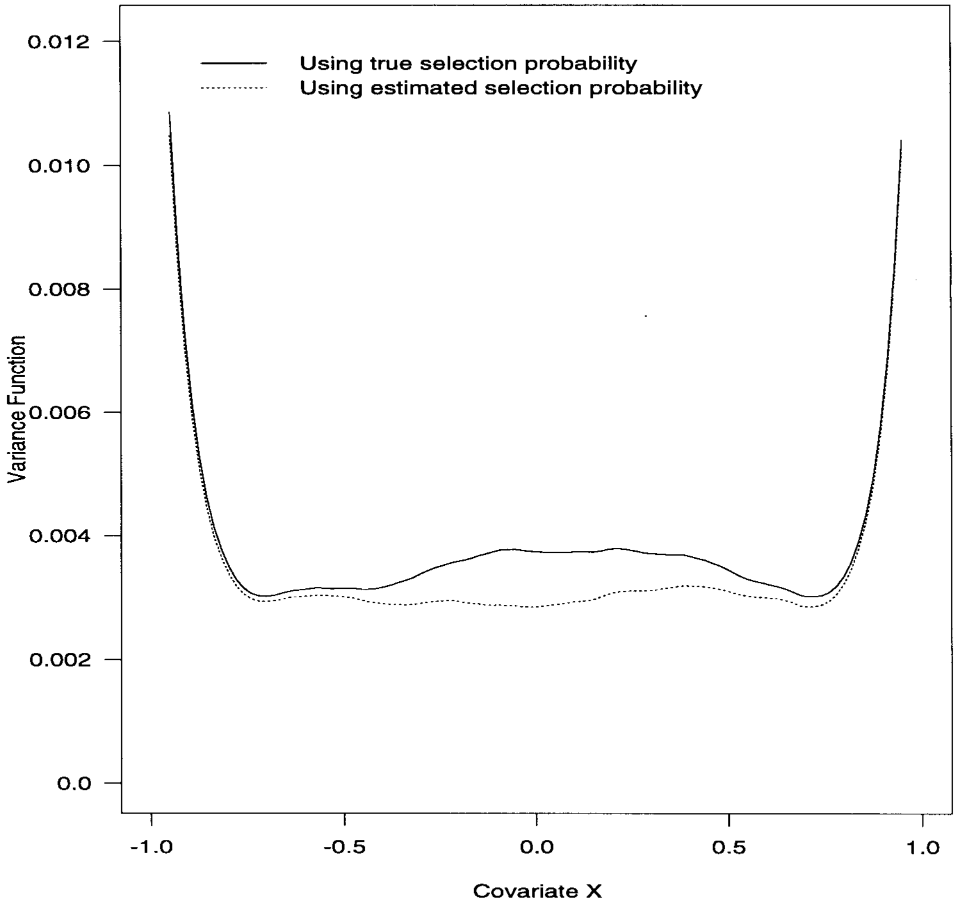

FIG. 3. *Simulation study for variances from estimating μ for continuous response.*

7. Generalizations. Theorem 1 is a special case of a general phenomenon, which we outline here. Suppose that one has interest in a function $\eta(\cdot)$. If a nuisance function $\pi(\cdot)$ were known, one would estimate $\eta(\cdot)$ at x by solving a local estimating equation of the form

$$0 = n^{-1} \sum_{i=1}^{n} K_h(X_i - x)\Psi\{\tilde{Y}_i, \pi(Z_i), \beta_0 + \beta_1(X_i - x)\}\{1, (X_i - x)\}^t, \tag{7}$$

where Ψ is an estimating function, Z is the covariate variable for $\pi(\cdot)$ and $\tilde{Y}$ represents a vector which may or may not include Z. We note that in (7), we are primarily interested in the estimation of β_0 and $\widehat{\beta}_0 = \widehat{\eta}(x)$. Equation (7) includes the weighted estimating equation obtained from the derivation of (3). In our problem, both $\tilde{Y}$ and Z equal the response Y.

Now suppose that $\pi(z)$ is also estimated by a local estimating equation but with bandwidth λ, so that

$$0 = n^{-1}\sum_{i=1}^{n} K_\lambda(Z_i - z)\Phi\{\tilde{Y}_i, X_i, \alpha_0 + \alpha_1(Z_i - z)\}\{1, (Z_i - z)\}^t,$$

for some function Φ such that the resulting estimate $\widehat{\alpha}_0 = \widehat{\pi}(z)$. The estimating functions Ψ and Φ are assumed to satisfy

$$0 = E\big[\Psi\{\tilde{Y}, \pi(Z), \eta(X)\}\big], \qquad 0 = E\big[\Phi\{\tilde{Y}, X, \pi(Z)\}|Z\big].$$

Under this setup, in Appendix C we sketch a result showing that

1. the bias of $\widehat{\eta}(x)$ is of order h^2, is independent of the design densities of (Z, X), but is generally affected by the estimation of $\pi(\cdot)$.
2. the variance of $\widehat{\eta}(x)$ is asymptotically the same as if $\pi(\cdot)$ were known.

Both these conclusions are reflected in our Theorem 1.

The extension to multivariate covariates may be made by applying the multivariate kernel function as in Fan, Heckman and Wand (1995, Section 3.2). However, the curse of dimensionality may occur. A more appealing approach is to consider a regression model by the generalized partial linear single-index model [Carroll, Fan, Gijbels and Wand (1997)]. Asymptotic distribution theory in this setting requires further investigation.

APPENDIX: TECHNICAL PROOFS

A. Proof of Theorem 1. First, we present a brief proof of the limit distribution of $\widehat{\mu}(\cdot, \pi)$. The readers are referred to Fan, Heckman and Wand (1995) for some related calculations. Recall that we use known π now. Define $\rho(x) = ([g'\{\mu(x)\}]^2 V\{\mu(x)\})^{-1}$, and let $q_i(x, y) = (\partial^i/\partial x^i)Q\{g^{-1}(x), y\}$. Fan, Heckman and Wand (1995) noted that q_i is linear in y for a fixed x and that $q_1\{\eta(x), \mu(x)\} = 0$ and $q_2\{\eta(x), \mu(x)\} = -\rho(x)$.

CONDITIONS. (A1) The function $q_2(x, y) < 0$ for $x \in R$ and y in the range of the response variable.

(A2) The functions f'_X, $\eta^{(3)}$, $\mathrm{var}(Y|X = \cdot)$, $V^{(2)}$ and $g^{(3)}$ are continuous.

(A3) For each $x \in \mathrm{supp}(f_X)$, $\rho(x)$, $\mathrm{var}(Y|X = x)$ and $g'\{\mu(x)\}$ are nonzero.

(A4) The kernel function K is a symmetric probability density with support $[-1, 1]$.

(A5) For each point x_0 on the boundary of $\mathrm{supp}(f_X)$, there exists a nontrivial interval $\mathscr{C}$ containing x_0 such that $\inf_{x\in\mathscr{C}} f_X(x) > 0$.

(A6) The selection probability $\pi(y) > 0$ for all $y \in \mathrm{supp}(f_Y)$.

(A7) $E[q_1\{\eta(X_1), Y_1\}(\delta_1/\pi_1)]^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$.

PROOF OF THE ASYMPTOTIC DISTRIBUTION OF $\widehat{\mu}(\cdot, \pi)$. We study the asymptotic properties of $\widehat{\beta}^* = (nh)^{1/2}[\widehat{\beta}_0 - \eta(x), h\{\widehat{\beta}_1 - \eta'(x)\}]^t$. Let $\overline{\eta}(x, u) = \eta(x) + \eta'(x)(u - x)$, $X_i^* = \{1, (X_i - x)/h\}^t$ and $\beta^* = (nh)^{1/2}[\beta_0 - \eta(x), h\{\beta_1 - \eta'(x)\}]^t$. Since $\beta_0 + \beta_1(X_i - x) = \overline{\eta}(x, X_i) + (nh)^{-1/2}\beta^{*t}X_i^*$, if $(\widehat{\beta}_0, \widehat{\beta}_1)$ maximizes (3), then $\widehat{\beta}^*$ maximizes

$$\sum_{i=1}^n Q[g^{-1}\{\overline{\eta}(x, X_i) + (nh)^{-1/2}\beta^{*t}X_i^*\}, Y_i]\frac{\delta_i}{\pi_i}K_h(X_i - x), \tag{8}$$

as a function of β^*, where $\pi_i = \pi(Y_i)$. We consider the normalized function

$$\begin{aligned} &l_n(\beta^*, \pi) \\ &\quad = \sum_{i=1}^n \Big(Q[g^{-1}\{\overline{\eta}(x, X_i) + (nh)^{-1/2}\beta^{*t}X_i^*\}, Y_i] \\ &\qquad\qquad - Q[g^{-1}\{\overline{\eta}(x, X_i)\}, Y_i]\Big)\frac{\delta_i}{\pi_i}K_h(X_i - x). \end{aligned} \tag{9}$$

Then $\widehat{\beta}^* = \widehat{\beta}^*(\pi)$ maximizes $l_n(\cdot, \pi)$. Let

$$W_n(\pi) = (nh)^{-1/2}\sum_{i=1}^n q_1\{\overline{\eta}(x, X_i), Y_i\}\frac{\delta_i}{\pi_i}K_h(X_i - x)X_i^*, \tag{10}$$

$$A_n(\pi) = (nh)^{-1}\sum_{i=1}^n q_2\{\overline{\eta}(x, X_i), Y_i\}\frac{\delta_i}{\pi_i}K_h(X_i - x)X_i^*X_i^{*t}. \tag{11}$$

Similarly to Fan, Heckman and Wand (1995), we have that

$$\begin{aligned} l_n(\beta^*, \pi) &= W_n^t(\pi)\beta^* + \tfrac{1}{2}\beta^{*t}A_n(\pi)\beta^* + O_p\{(nh)^{-1/2}\} \\ &= W_n^t(\pi)\beta^* - \tfrac{1}{2}\beta^{*t}(\Sigma_x + h\Lambda_x)\beta^* + O_p\{(nh)^{-1/2}\} + o_p(h), \end{aligned}$$

where

$$\begin{aligned} \Sigma_x &= \rho(x)f_X(x)\begin{pmatrix} \gamma_0(N_x^h) & \gamma_1(N_x^h) \\ \gamma_1(N_x^h) & \gamma_2(N_x^h) \end{pmatrix}, \\ \Lambda_x &= (\rho f_X)'(x)\begin{pmatrix} \gamma_1(N_x^h) & \gamma_2(N_x^h) \\ \gamma_2(N_x^h) & \gamma_3(N_x^h) \end{pmatrix}. \end{aligned} \tag{12}$$

By the Quadratic Approximation Lemma of Fan, Heckman and Wand (1995) and under the bandwidth condition that $nh^3 \to \infty$, we have that

$$\widehat{\beta}^* = \Sigma_x^{-1}W_n(\pi) - h\Sigma_x^{-1}\Lambda_x\Sigma_x^{-1}W_n(\pi) + o_p(h). \tag{13}$$

Similarly to Fan, Heckman and Wand (1995), we can also show that

$$\begin{aligned} E\{W_n(\pi)\} &= \frac{1}{2}(nh^5)^{1/2}\eta^{(2)}(x)\rho(x)f_X(x)\begin{pmatrix}\gamma_2(N_x^h)\\ \gamma_3(N_x^h)\end{pmatrix} + O\{(nh^7)^{1/2}\}\\ &\equiv n^{1/2}h^{5/2}B_x + O\{(nh^7)^{1/2}\},\\ \operatorname{var}\{W_n(\pi)\} &= \frac{f_X(x)\mathscr{L}(x)}{[V\{\mu(x)\}g'\{\mu(x)\}]^2}\begin{pmatrix}\tau_0(N_x^h) & \tau_1(N_x^h)\\ \tau_1(N_x^h) & \tau_2(N_x^h)\end{pmatrix} + o(h)\\ &\equiv \Gamma_x + o(h), \end{aligned} \tag{14}$$

where $\mathscr{L}(x)$ is given in (4).

It can be shown by checking Lyapounov's condition and using the Cramér–Wold device that $\widehat{\beta}^*$ is asymptotically normally distributed. From (13), we get the approximations

$$\begin{aligned} E(\widehat{\beta}^*) &= \Sigma_x^{-1}n^{1/2}h^{5/2}B_x + O\{(nh^7)^{1/2}\} + o(h);\\ \operatorname{var}(\widehat{\beta}^*) &= \Sigma_x^{-1}\Gamma_x\Sigma_x^{-1} + o(h). \end{aligned}$$

The proof of the first part of Theorem 1 thus follows since we are only concerned with the first component of $\widehat{\beta}^*$, and $\mu(x) = g^{-1}\{\eta(x)\}$.

We now present some additional conditions for dealing with the asymptotic distribution of $\widehat{\mu}(\cdot, \widehat{\pi})$. Define $\rho^*(y) = [\{g^{*(1)}(\pi(y))\}^2 V^*\{\pi(y)\}]^{-1}$, and let $q_i^*(y, z) = (\partial^i/\partial y^i)Q^*(g^{*-1}(y), z)$. Again, we have that

$$q_1^*\{\eta^*(y), \pi(y)\} = 0, \qquad q_2^*\{\eta^*(y), \pi(y)\} = -\rho^*(y). \tag{15}$$

In addition to Conditions (A1)–(A7), we need the following conditions.

CONDITIONS. (B1) The function $q_2^*(y, \delta) < 0$ for $y \in R$ and $\delta = 0, 1$.

(B2) The functions f_Y' (when Y is continuous), $\eta^{*(3)}$, $\operatorname{var}(\delta|Y = \cdot)$, $V^{*(2)}$, $g^{*(3)}$ and $\pi^{(2)}$ are continuous.

(B3) For each $y \in \operatorname{supp}(f_Y)$, $\rho^*(x)$, $V^*(y)$ and $g'\{\pi(y)\}$ are nonzero.

(B4) For each point y_0 on the boundary of $\operatorname{supp} f_Y$, there exists a nontrivial interval $\mathscr{C}$ containing y_0 such that $\inf_{y\in\mathscr{C}} f_Y(y) > 0$.

(B5) $\inf\{\pi(y)\colon y \in \operatorname{supp}(f_Y)\} > 0$.

(B6) The conditional density of X given Y is bounded a.e.

Before proving the main part of Theorem 1, we present some lemmas which will be used in the proof. Recall that $\widehat{\pi}$ was defined in (5).

LEMMA 1. *Under the same conditions as those of Theorem* 1, $G_n = o_p(h)$, *where*

$$G_n = (nh)^{-1}\sum_{i=1}^{n} q_2\{\overline{\eta}(x, X_i), Y_i\}K_h(X_i - x)X_i^* X_i^{*t}\frac{\delta_i}{\pi_i^2}(\widehat{\pi}_i - \pi_i).$$

LEMMA 2. *Under the same conditions as those in Theorem* 1, $C_n = o_p(h^{1/2})$, *where*

$$C_n = (nh)^{-1/2}\sum_{i=1}^{n}\left[\frac{\delta_i - \pi_i}{\pi_i}\frac{\widehat{\pi}_i - \pi_i}{\pi_i}q_1\{\overline{\eta}(x, X_i), Y_i\}K_h(X_i - x)X_i^*\right].$$

LEMMA 3. *Under the same conditions as those of Theorem* 1, *let*

$$D_n = (nh)^{-1/2}\sum_{i=1}^{n}\left[\frac{\widehat{\pi}_i - \pi_i}{\pi_i}q_1\{\overline{\eta}(x, X_i), Y_i\}K_h(X_i - x)X_i^*\right].$$

Then there exists an $S_3(x)$ *and* D_n^* *with* $E(D_n^*) = o(n^{1/2}h^{5/2})$ *and* $\mathrm{var}(D_n^*) = o(h^2)$, *such that*

$$D_n - n^{1/2}h^{5/2}(c^*)^2 f_X(x)S_3(x) = (nh)^{-1/2}\sum_{i=1}^{n}\{(\delta_i - \pi_i)/\pi_i\}\mathscr{M}_h(Y_i) + D_n^*,$$

where

$$\mathscr{M}_h(Y_i) = E\big[q_1\{\overline{\eta}(x, X_i), Y_i\}X_i^* K_h(X_i - x)|Y_i\big]. \tag{16}$$

The proofs of Lemmas 1–3 will be postponed until after the proof of the limit distribution of $\widehat{\mu}(\cdot, \widehat{\pi})$.

PROOF OF THE ASYMPTOTIC DISTRIBUTION OF $\widehat{\mu}(\cdot, \widehat{\pi})$. Recall that $\widehat{\beta} = \widehat{\beta}(\widehat{\pi})$ maximizes (6), and we defined $l_n(\beta^*, \pi)$ in (9). The main step here is to derive the asymptotic expression of $l_n(\beta^*, \widehat{\pi})$. Note that $A_n(\pi)$ was defined in (11). We have

$$\begin{aligned} A_n(\widehat{\pi}) - A_n(\pi) = \Bigg[&(nh)^{-1}\sum_{i=1}^{n} q_2\{\overline{\eta}(x, X_i), Y_i\} \\ &\times K_h(X_i - x)X_i^* X_i^{*t}\frac{\delta_i}{\pi_i^2}(\widehat{\pi}_i - \pi_i)\Bigg]\{1 + o_p(1)\} \\ \equiv{}& G_n\{1 + o_p(1)\}. \end{aligned}$$

By Lemma 1, we have $A_n(\widehat{\pi}) = A_n(\pi) + o_p(h)$. Using calculations similar to those in the proof of the distribution of $\widehat{\mu}(\cdot, \pi)$,

$$l_n(\beta^*, \widehat{\pi}) = W_n^t(\widehat{\pi})\beta^* - \tfrac{1}{2}\beta^{*t}(\Sigma_x + h\Lambda_x)\beta^* + O_p\{(nh)^{-1/2}\} + o_p(h).$$

For simplicity in this proof, we continue to use $\widehat{\beta}^* = \widehat{\beta}^*(\widehat{\pi})$ as the maximizer of (8) with estimated $\widehat{\pi}$. By the quadratic approximation lemma of Fan, Heckman and Wand (1995), we have that

$$\widehat{\beta}^* = \Sigma_x^{-1}W_n(\widehat{\pi}) - h\Sigma_x^{-1}\Lambda_x\Sigma_x^{-1}W_n(\pi) + o_p(h). \tag{17}$$

We now find the limit distribution of $W_n(\widehat{\pi})$, where $W_n(\cdot)$ was defined in (10). By a linearization technique as in Wang et al. (1997), we have

$$\begin{aligned} W_n(\widehat{\pi}) &= (nh)^{-1/2}\sum_{i=1}^n\left[q_1\{\overline{\eta}(x, X_i), Y_i\}\frac{\delta_i}{\pi_i}K_h(X_i-x)X_i^*\left(1-\frac{\widehat{\pi}_i-\pi_i}{\pi_i}\right)\right] \\ &\quad+\left((nh)^{-1/2}\sum_{i=1}^n\left[q_1\{\overline{\eta}(x, X_i), Y_i\}\frac{\delta_i}{\pi_i}K_h(X_i-x)X_i^*\frac{(\widehat{\pi}_i-\pi_i)^2}{\pi_i}\right]\right) \\ &\qquad\times\{1+o_p(1)\}. \end{aligned}$$

Denote the second term of the above equation by R_n. Then it can be shown to have mean $O(n^{1/2}h^{9/2})$ and variance $o(h)$. Therefore, we have that

$$\begin{aligned} W_n(\widehat{\pi}) &= (nh)^{-1/2}\sum_{i=1}^n\left(\frac{\delta_i}{\pi_i}-\frac{\widehat{\pi}_i-\pi_i}{\pi_i}\right)q_1\{\overline{\eta}(x, X_i), Y_i\}K_h(X_i-x)X_i^* \\ &\quad-(nh)^{-1/2}\sum_{i=1}^n\frac{\delta_i-\pi_i}{\pi_i}\frac{\widehat{\pi}_i-\pi_i}{\pi_i} \\ &\qquad\times q_1\{\overline{\eta}(x, X_i), Y_i\}K_h(X_i-x)X_i^*+R_n \\ &\equiv W_n(\pi)-D_n-C_n+R_n. \end{aligned}\tag{18}$$

By Lemmas 2 and 3, we have that

$$\begin{aligned} W_n(\widehat{\pi}) &= W_n(\pi)-n^{1/2}h^{5/2}(c^*)^2f_X(x)S_3(x) \\ &\quad-(nh)^{-1/2}\sum_{i=1}^n\frac{\delta_i-\pi_i}{\pi_i}\mathscr{M}_h(Y_i)+R_n^* \end{aligned}\tag{19}$$

for some R_n^* that has mean $o(n^{1/2}h^{5/2})$ and variance $o(1)$. Let $f_{X|Y}$ denote the conditional density of X given Y. By direct calculations,

$$\begin{aligned} \mathscr{M}_h(Y_j) &= h[q_1\{\eta(x), Y_j\}f_{X|Y}(x)\{\gamma_0(N_x^h), \gamma_1(N_x^h)\}^t]\{1+o_p(1)\} \\ &\equiv h\mathscr{D}(Y_j)\{1+o_p(1)\}. \end{aligned}$$

By (14), (17) and (19), the asymptotic distribution of $\widehat{\mu}(\cdot, \widehat{\pi})$ follows because

$$E\left\{\frac{\delta_i-\pi_i}{\pi_i}\mathscr{M}_h(Y_i)\right\} = E\left[E\left\{\frac{\delta_i-\pi_i}{\pi_i}\mathscr{M}_h(Y_i)|Y_i\right\}\right] = 0;$$

$$\begin{aligned} &\operatorname{var}\left\{(nh)^{-1/2}\sum_{i=1}^n\frac{\delta_i-\pi_i}{\pi_i}\mathscr{M}_h(Y_i)\right\} \\ &\quad= h^{-1}E\left\{\mathscr{M}_h(Y_1)\mathscr{M}_h^t(Y_1)\operatorname{var}\left(\frac{\delta_1-\pi_1}{\pi_1}|Y_1\right)\right\} \\ &\quad= h^{-1}E\left\{\frac{1-\pi_1}{\pi_1}\mathscr{M}_h(Y_1)\mathscr{M}_h^t(Y_1)\right\} \\ &\quad= hE\left\{\frac{1-\pi_1}{\pi_1}\mathscr{D}(Y_1)\mathscr{D}^t(Y_1)\right\}\{1+o(1)\} = O(h). \end{aligned}$$

The last equation holds by Conditions (B5) and (B6). □

PROOF OF LEMMA 1. First, we note that $E\{(\widehat{\pi}_1 - \pi_1)|Y_1\} = \lambda^2 c_1(Y_1)\{1 + o_p(1)\}$ and $\operatorname{var}\{(\widehat{\pi}_1 - \pi_1)|Y_1\} = (n\lambda)^{-1} c_2(Y_1)\{1 + o_p(1)\}$, for some functions c_1 and c_2. Let $\widehat{\pi}_{i(j)}$ denote $\widehat{\pi}_i$ without using subject j. Then

$$\begin{aligned}
E(G_n) &= E\biggl[q_2\{\overline{\eta}(x, X_1), Y_1\}\frac{1}{h}K_h(X_1 - x)X_1^* X_1^{*t}\frac{\delta_1}{\pi_1^2}(\widehat{\pi}_1 - \pi_1)\biggr]\\
&= E\biggl(E\biggl[q_2\{\overline{\eta}(x, X_1), Y_1\}\frac{1}{h}K_h(X_1 - x)X_1^* X_1^{*t}\\
&\qquad\qquad \times \frac{\delta_1}{\pi_1^2}(\widehat{\pi}_{1(1)} - \pi_1)\big|X_1, Y_1\biggr]\biggr)\\
&\quad + O\biggl(\frac{1}{n\lambda}\biggr)\\
&= E\biggl[q_2\{\overline{\eta}(x, X_1), Y_1\}\frac{1}{h}K_h(X_1 - x)X_1^* X_1^{*t}\frac{1}{\pi_1}E\{(\widehat{\pi}_1 - \pi_1)\big|Y_1\}\biggr]\\
&\quad + O\biggl(\frac{1}{n\lambda}\biggr)\\
&= \lambda^2 E\biggl[q_2\{\overline{\eta}(x, X_1), Y_1\}\frac{1}{h}K_h(X_1 - x)X_1^* X_1^{*t}\frac{1}{\pi_1}c_1(Y_1)\{1 + o_p(1)\}\biggr]\\
&\quad + O\biggl(\frac{1}{n\lambda}\biggr)\\
&= o(h).
\end{aligned}$$

The last equation holds since $\lambda = c^* h$ for some $c^* > 0$. Note that if we let $S_i = q_2\{\overline{\eta}(x, X_i), Y_i\}h^{-1}\ K_h(X_i - x)X_i^* X_i^{*t}(\delta_i/\pi_i^2)(\widehat{\pi}_i - \pi_i)$, then, following calculations similar to those above, we obtain that $\operatorname{cov}(S_i, S_j) = o(h^2)$ when $i \neq j$. Therefore, the variance of the left-upper element of G_n is

$$\begin{aligned}
\operatorname{var}[\{G_n\}_{11}] &= n^{-1}\operatorname{var}\biggl[q_2\{\overline{\eta}(x, X_1), Y_1\}\frac{1}{h}K_h(X_1 - x)\frac{\delta_1}{\pi_1^2}(\widehat{\pi}_1 - \pi_1)\biggr] + o(h^2)\\
&= n^{-1}E\biggl[q_2^2\{\overline{\eta}(x, X_1), Y_1\}\frac{1}{h^2}K_h^2(X_1 - x)\\
&\qquad\qquad \times \frac{1}{\pi_1^4}\operatorname{var}\{(\widehat{\pi}_{1(1)} - \pi_1)\big|Y_1\}\biggr]\\
&\quad + n^{-1}\operatorname{var}\biggl[q_2\{\overline{\eta}(x, X_1), Y_1\}\frac{1}{h}K_h(X_1 - x)\\
&\qquad\qquad \times \frac{\delta_1}{\pi_1^2}E\{(\widehat{\pi}_{1(1)} - \pi_1)\big|Y_1\}\biggr] + o(h^2)\\
&= (nh)^{-1}E\biggl[q_2^2\{\overline{\eta}(x, X_1), Y_1\}\frac{1}{h}K_h^2(X_1 - x)\\
&\qquad\qquad \times \frac{1}{\pi_1^3}(n\lambda)^{-1}c_2(Y_1)\{1 + o_p(1)\}\biggr]
\end{aligned}$$

$$\begin{aligned}&+(nh)^{-1}\operatorname{var}\biggl[q_2\{\overline{\eta}(x, X_1), Y_1\}\frac{1}{\sqrt{h}}K_h(X_1-x)\\&\qquad\qquad\times\frac{\delta_1}{\pi_1^2}\lambda^2 c_1(Y_1)\{1+o_p(1)\}\biggr]+o(h^2)\\&=O\{(nh)^{-1}(n\lambda)^{-1}+(nh)^{-1}\lambda^4\}+o(h^2)\\&=o(h^2).\end{aligned}$$

The last equation holds since $\lambda = c^*h$ and $nh^3 \to \infty$. Similar calculations lead to $\operatorname{var}[\{G_n\}_{12}] = o(h^2)$ and $\operatorname{var}[\{G_n\}_{22}] = o(h^2)$, completing the proof of Lemma 1. □

PROOF OF LEMMA 2. We assume that Y is a continuous random variable; a similar approach can be easily applied to discrete Y. Using calculations similar to those of Fan, Heckman and Wand (1995) and under Conditions (B1)–(B4), we can show that, for each y, as $n\lambda \to \infty$,

$$\begin{aligned}\widehat{\pi}(y)-\pi(y)&=(n\lambda)^{-1/2}[g^{*(1)}\{\pi(y)\}]^{-1}\{\Sigma_y^{*-1}W_n^*\}_1\\&\qquad+(n\lambda)^{-1/2}\mathscr{R}(y,\tilde{\delta},\tilde{Y}),\end{aligned}\tag{20}$$

where $\{\cdot\}_1$ denotes the first component of a vector, $\tilde{\delta} = (\delta_1, \dots, \delta_n)$, $\tilde{Y} = (Y_1, \dots, Y_n)$,

$$W_n^* = (n\lambda)^{-1/2}\sum_{i=1}^n q_1^*\{\overline{\eta}^*(y, Y_i), \delta_i\}K_\lambda(Y_i-y)Y_i^*,$$

$$\Sigma_y^* = \rho^*(y)f_Y(y)\begin{pmatrix}\gamma_0(N_y^\lambda) & \gamma_1(N_y^\lambda)\\ \gamma_1(N_y^\lambda) & \gamma_2(N_y^\lambda)\end{pmatrix},$$

$\overline{\eta}^*(y,u) = \eta^*(y) + \eta^{*(1)}(u)(u-y)$, $Y_i^* = (1, (Y_i - y)/\lambda)^t$, $N_y^\lambda = \{z\colon y - \lambda z \in \operatorname{supp}(f_Y) \cap [-1, 1]\}$, $\mathscr{R}(y, \tilde{\delta}, \tilde{Y}) = -\lambda[g^{*(1)}\{\pi(y)\}]^{-1}\{\Sigma_y^{*(-1)}\Lambda_y^*\Sigma_y^{*(-1)}W_n^*\}_1\{1 + o_p(1)\}$, and Λ_y^* is the same as Λ_x defined in (12) except replacing ρf_X by $\rho^* f_Y$, $f_Y(\cdot)$ being the density of Y. From (20), we have that there is a function $g_\pi^*(y)$ and a function $\mathscr{X}(\delta, y)$ such that

$$\begin{aligned}\widehat{\pi}(y)-\pi(y)&=\lambda^2 g_\pi^*(y)+(n\lambda)^{-1}\sum_{i=1}^n K_\lambda(Y_i-y)\mathscr{X}(\delta_i,Y_i)\\&\qquad+o_p(\lambda^2)+o_p(n^{-1/2}),\end{aligned}\tag{21}$$

where $E\{\mathscr{X}(\delta_i, Y_i)|Y_i\} = 0$ and the $o_p(\lambda^2)$ term does not depend on the δ's. Note that as in Fan, Heckman and Wand (1995), it can be shown that the term associated with the bias $g_\pi^*(y) = 0$ if $\pi'(y) = 0$. Therefore,

$$(nh)^{-1/2}\sum_{i=1}^n\left[\frac{\delta_i-\pi_i}{\pi_i}\frac{\lambda^2 g_\pi^*(Y_i)}{\pi_i}q_1\{\overline{\eta}(x,X_i),Y_i\}K_h(X_i-x)X_i^*\right]=O_p(\lambda^2)$$

by calculating the mean and the variance. Also,

$$(nh)^{-1/2}\sum_{i=1}^{n}\biggl[\frac{\delta_i-\pi_i}{\pi_i^2}\{(n\lambda)^{-1}\sum_{j=1}^{n}K_\lambda(Y_j-Y_i)\mathscr{X}(\delta_j,Y_j)\} \\ \times q_1\{\overline{\eta}(x,X_i),Y_i\}K_h(X_i-x)X_i^*\biggr]$$

can be shown to be $o_p(h)$ because each summand has a factor $(\delta_i-\pi_i)\mathscr{X}(\delta_j, Y_j)$ which has mean 0 if $i\neq j$. Further,

$$(nh)^{-1/2}\sum_{i=1}^{n}\biggl[\frac{\delta_i-\pi_i}{\pi_i^2}q_1\{\overline{\eta}(x,X_i),Y_i\}K_h(X_i-x)X_i^*\{o_p(\lambda^2)+o_p(n^{-1/2})\}\biggr] \\ = o_p(h^{1/2}),$$

because the $o_p(\lambda^2)$ term above does not depend on the δ's. Therefore, $C_n = o_p(h^{1/2})$. □

PROOF OF LEMMA 3. Apply (21) to D_n and write $D_n = D_{1n}+D_{2n}+D_{3n}$. First,

$$D_{1n}=(nh)^{-1/2}\sum_{i=1}^{n}\biggl[\frac{\lambda^2 g_\pi^*(Y_i)}{\pi_i}q_1\{\overline{\eta}(x,X_i),Y_i\}K_h(X_i-x)X_i^*\biggr].$$

Then note that

$$E(D_{1n}) = n^{1/2}h^{-1/2}\lambda^2\int\frac{g_\pi^*(y_1)}{\pi_1}q_1\{\overline{\eta}(x,x_1),y_1\} \\ \times K_h(x_1-x)x_1^*f_{Y,X}(y_1,x_1)\,dy_1\,dx_1 \\ = n^{1/2}h^{-1/2}\lambda^2\int\frac{g_\pi^*(y_1)}{\pi_1}\frac{y_1-\mu(x)-[g^{(1)}\{\mu(x)\}]^{-1}\eta^{(1)}(x)(x_1-x)}{g^{(1)}\{\mu(x)\}V\{\mu(x)\}} \\ \times K_h(x_1-x)x_1^*f_{Y,X}(y_1,x_1)dy_1dx_1.$$

Let

$$S_1(x) = E\bigl(g_\pi^*(Y_1)Y_1[\pi(Y_1)g^{(1)}\{\mu(x)\}V\{\mu(x)\}]^{-1}\bigr),$$
$$S_2(x) = E\bigl(g_\pi^*(Y_1)[\pi(Y_1)g^{(1)}\{\mu(x)\}V\{\mu(x)\}]^{-1}\bigr)$$

and

$$S_3(x) = S_1(x)-\mu(x)S_2(x). \tag{22}$$

Note that $S_3(x)=0$ if either Y is a lattice random variable or $\pi'(Y)=0$ a.e., because under these circumstances $g_\pi^*(Y)=0$ a.e. Then it is easily seen that $E\{(D_{1n})_1\} = n^{1/2}h^{5/2}(c^*)^2f_X(x)S_3(x)+o_p(n^{1/2}h^{5/2})$ and it can also be shown that $\operatorname{var}\{(D_{1n})_1\} = O(\lambda^4)$.

Now consider

$$D_{2n} = (nh)^{-1/2} \sum_{j=1}^{n} \mathscr{X}(\delta_j, Y_j)\Bigg((n\lambda)^{-1} \sum_{i=1}^{n} \bigg[\frac{q_1\{\overline{\eta}(x, X_i), Y_i\}X_i^*}{\pi_i} K_h(X_i - x)\bigg] \times K_\lambda(Y_j - Y_i)\Bigg).$$

To estimate D_{2n}, we note that, by some further calculations, the $\mathscr{X}(\delta_i, Y_i)$ in (21) can be written as $\mathscr{X}(\delta_i, Y_i) = \{f_Y(y)\psi(y)\}^{-1}\phi(Y_i, y)\{\delta_i - \pi(Y_i)\} + o_p(1)$, where $\phi(Y_i, y) = \gamma_2(N_y^\lambda) - \gamma_1(N_y^\lambda)\{(Y_i - y)/\lambda\}$ and $\phi(y) = \gamma_0(N_y^\lambda)\gamma_2(N_y^\lambda) - \gamma_1^2(N_y^\lambda)$. Therefore, by some standard calculations, we have that

$$D_{2n} = (nh)^{-1/2} \sum_{j=1}^{n} (\delta_j - \pi_j)\mathscr{M}_h(Y_j)/\pi_j + o_p(n^{-1/2}).$$

Finally,

$$D_{3n} = (nh)^{-1/2} \sum_{i=1}^{n} \bigg[\frac{q_1\{\overline{\eta}(x, X_i), Y_i\}K_h(X_i - x)X_i^*}{\pi_i}\{o_p(h^2) + o_p(n^{-1/2})\}\bigg].$$

By some standard calculations, we have that $E(D_{3n}) = o(n^{1/2}h^{5/2})$, $\operatorname{var}(D_{3n}) = o(h^2)$. Combining the calculations of D_{1n}, D_{2n} and D_{3n}, we have that there is a D_n^* such that $E(D_n^*) = o(n^{1/2}h^{5/2})$, $\operatorname{var}(D_n^*) = o(h^2)$ and

$$D_n - n^{1/2}h^{5/2}(c^*)^2 f_X(x)S_3(x) = (nh)^{-1/2} \sum_{j=1}^{n} \frac{\delta_j - \pi_j}{\pi_j}\mathscr{M}_h(Y_j) + D_n^*. \qquad \square$$

B. Proof of Theorem 2. By (18), $W_n(\widehat{\pi}) = W_n(\pi) - D_n - C_n + R_n$. As in the proof of Lemma 3, $D_n = (nh)^{-1/2}\sum_{i=1}^{n}\{(\delta_i - \pi_i)/\pi_i\}\mathscr{M}_h(Y_i) + n^{1/2}h^{5/2}(c^*)^2 f_X(x)S_3(x) + D_n^*$. By some standard calculations, we have

$$\begin{aligned}
&\operatorname{cov}\{W_n(\pi), D_n\} \\
&\quad = (nh)^{-1} \sum_{i=1}^{n} \operatorname{cov}\bigg[q_1\{\overline{\eta}(x, X_i), Y_i\}\frac{\delta_i}{\pi_i}K_h(X_i - x)X_i^*, \frac{\delta_i - \pi_i}{\pi_i}\mathscr{M}_h(Y_i)\bigg] \\
&\qquad + o(h) \\
&\quad = h^{-1}E\bigg[\frac{1 - \pi_1}{\pi_1}q_1\{\overline{\eta}(x, X_1), Y_1\}X_1^*K_h(X_1 - x)\mathscr{M}_h^t(Y_1)\bigg] + o(h) \\
&\quad = h^{-1}E\bigg\{\frac{1 - \pi_1}{\pi_1}\mathscr{M}_h(Y_1)\mathscr{M}_h^t(Y_1)\bigg\} + o(h) \\
&\quad = hE\bigg\{\frac{1 - \pi_1}{\pi_1}\mathscr{D}(Y_1)\mathscr{D}^t(Y_1)\bigg\} + o(h).
\end{aligned}$$

The $\mathscr{D}(Y)$ in the above calculations was defined in the proof of Theorem 1. The covariances due to $\operatorname{cov}\{W_n(\pi), C_n\}$ and $\operatorname{cov}\{W_n(\pi), R_n\}$ can be shown to be smaller than the rate of $\operatorname{cov}\{W_n(\pi), D_n\}$ since $\operatorname{cov}(C_n) = o(h), \operatorname{cov}(R_n) =$

$o(h)$. We may, in fact, apply (21) to get more precise rates, for example, $\mathrm{cov}\{W_n(\pi), C_n\} = o(h^2)$. Therefore, letting

$$\begin{aligned}\widehat{B}^* &= \Sigma_x^{-1}\{W_n(\pi) - D_n\} - h\Sigma_x^{-1}\Lambda_x\Sigma_x^{-1}W_n(\pi)\\ &= \widehat{\beta}^*(\pi) - \Sigma_x^{-1}D_n + o_p(h) = \widehat{\beta}^*(\widehat{\pi}) + o_p(h)\end{aligned}$$

by (13) and (17), Theorem 2 follows since $\mathrm{var}\{\widehat{B}^*\} = \mathrm{var}\{\widehat{\beta}^*(\pi)\} - h\Sigma_x^{-1}\, E[\{(1-\pi_1)/\pi_1\}\mathscr{D}(Y_1)\mathscr{D}^t(Y_1)]\Sigma_x^{-1} + o(h)$. □

C. Sketch of the Proof of Generalizations in Section 7. Here we sketch the arguments of Section 7. We renormalize so that $\gamma_0 = \gamma_2 = 1$, $\gamma_1 = \gamma_3 = 0$. Carroll, Ruppert and Welsh (1998) showed that there is a function $g_\pi(z)$ which does not depend on the density $f_Z(\cdot)$ of Z, and a function $\Omega(\cdot)$ such that

$$\begin{aligned}\widehat{\pi}(z) - \pi(z) = (h^2/2)g_\pi(z) + (n\lambda)^{-1}\sum_{i=1}^{n} K_\lambda(Z_i - z)\Omega(\tilde{Y}_i, X_i, Z_i)\\ + o_p(h^2) + o_p(n^{-1/2}),\end{aligned}\tag{23}$$

$$E\{\Omega(\tilde{Y}, X, Z)|Z\} = 0.\tag{24}$$

By a Taylor series expansion of (7), it is easily seen that

$$-f_X(x)\Sigma_1(x)\{\widehat{\eta}(x) - \eta(x)\} \approx B_{n1} - B_{n2} + B_{n3},$$

where, if $\Psi_\pi = (\partial/\partial v)\Psi(\tilde{Y}, v, \eta)$, $\Psi_\eta = (\partial/\partial v)\Psi(\tilde{Y}, \pi, v)$. Then

$$\begin{aligned}B_{n1} &= (nh)^{-1}\sum_{i=1}^{n} K_h(X_i - x)\Psi\{\tilde{Y}_i, \pi(Z_i), \eta(X_i)\},\\ B_{n2} &= (nh)^{-1}\sum_{i=1}^{n} K_h(X_i - x)\big[\Psi\{\tilde{Y}_i, \pi(Z_i), \eta(X_i)\}\\ &\qquad\qquad - \Psi\{\tilde{Y}_i, \pi(Z_i), \eta(x) + \eta'(x)(X_i - x)\}\big],\\ B_{n3} &= (nh)^{-1}\sum_{i=1}^{n} K_h(X_i - x)\Psi_\pi\{\tilde{Y}_i, \pi(Z_i), \eta(X_i)\}\{\widehat{\pi}(Z_i) - \pi(Z_i)\},\\ \Sigma_1(x) &= E[\Psi_\eta\{\tilde{Y}, \pi(Z), \eta(x)\}|X = x].\end{aligned}$$

It is easily seen that $B_{n2} = (h^2/2)f_X(x)\Sigma_1(x)\eta^{(2)}(x)\{1 + o_p(1)\}$. Writing $\Psi_{\pi_i} = \Psi_\pi\{\tilde{Y}_i, \pi(Z_i), \eta(X_i)\}$, and writing Ω_i similarly, we note that $B_{n3} \approx B_{n31} + B_{n32}$, where, using (23),

$$\begin{aligned}B_{n31} &= (h^2/2)(nh)^{-1}\sum_{i=1}^{n} K_h(X_i - x)\Psi_{\pi_i} g_\pi(Z_i)\\ &= (h^2/2)f_X(x)E\{\Psi_\pi(\cdot)g_\pi(\cdot)|X = x\}\{1 + o_p(1)\},\\ B_{n32} &= n^{-2}(h\lambda)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} K_h(X_i - x)\Psi_{\pi_i}K_\lambda(Z_j - Z_i)\Omega_j.\end{aligned}$$

Thus, we have shown that

$$\widehat{\eta}(x) - \eta(x) \approx (h^2/2)\big[\eta^{(2)}(x) - E\{\Psi_\pi g_\pi(\cdot)|X = x\}/\Sigma_1(x)\big] \\ - \{f_X(x)\Sigma_1(x)\}^{-1}(B_{n1} + B_{n32}). \tag{25}$$

The first term in (25) is the bias, which is independent of the design density but affected by estimation of $\pi(\cdot)$, as claimed. To complete the argument, we merely need to show that $B_{n32} = o_p\{(nh)^{-1/2}\}$. Recalling that $K(\cdot)$ is symmetric, rewrite

$$B_{n32} = n^{-1}\sum_{i=1}^{n}\Omega_i\left\{n^{-1}\sum_{j=1}^{n}h^{-1}K_h(X_j - x)\lambda^{-1}K_\lambda(Z_j - Z_i)\Psi_{\pi_j}\right\}. \tag{26}$$

Using Chebychev's inequality, detailed algebra gives the designed result. In the interests of space, we forego the calculations, but note that the term in braces in (26) is a *bivariate* kernel regression of $\Psi_\pi\{\tilde{Y}_i, \pi(Z_i), \eta(X_i)\}$ on (X, Z) evaluated at $X = x$, $Z = Z_i$, and hence converges to $r(x, Z_i)$, where $r(x, Z_i) = E[\Psi_\pi\{\tilde{Y}, \pi(Z), \eta(X)\}|X = x, Z = Z_i]$. Therefore, $B_{n32} \approx n^{-1}\sum_{i=1}^{n}\Omega\{\tilde{Y}_i, \pi(Z_i), \eta(X_i)\}\, r(x, Z_i)$, which is $O_p(n^{-1/2})$ from (24). □

Acknowledgments. We are grateful to Barbara Bruemmer and Emily White for the case-control data.

REFERENCES

BRUEMMER, B., WHITE, E., VAUGHAN, T. and CHENEY, C. (1996). Nutrient intake in relationship to bladder cancer among middle aged men and women. *Amer. J. Epidemiology* **144** 485–495.

CARROLL, R. J., RUPPERT, D. and WELSH, A. H. (1998). Local estimating equations. *J. Amer. Statist. Assoc.*, **93** 214–227.

CARROLL, R. J., FAN, J., GIJBELS, I. and WAND, M. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477–489.

FAN, J., HECKMAN, N. E. and WAND, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasilikelihood functions. *J. Amer. Statist. Assoc.* **90** 141–150.

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.

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

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972), Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

PRENTICE, R. L. and PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66** 403–411.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592.

SCHUCANY, W. R. (1995). Adaptive bandwidth choice for kernel regression. *J. Amer. Statist. Assoc.* **90** 535–540.

SEVERINI, T. A. and STANISWALIS, J. G. (1994). Quasilikelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.* **89** 501–511.

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

WANG, C. Y., WANG, S., ZHAO, L. P. and OU, S. T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *J. Amer. Statist. Assoc.* **92** 512–525.

WEDDERBURN, R. W. M. (1974). Quasilikelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61** 439–447.

WHITE, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *Amer. J. Epidemiology* **115** 119–128.

C. Y. WANG
DIVISION OF PUBLIC HEALTH SCIENCES
FRED HUTCHINSON CANCER RESEARCH CENTER
1124 COLUMBIA STREET, MP 1002
SEATTLE, WASHINGTON 98104
E-MAIL: cywang@fhcrc.org

SUOJIN WANG
DEPARTMENT OF STATISTICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143

ROBERTO G. GUTIERREZ
DEPARTMENT OF STATISTICAL SCIENCE
SOUTHERN METHODIST UNIVERSITY
DALLAS, TEXAS 75275-0332

R. J. CARROLL
DEPARTMENT OF STATISTICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143