Sankhyā : The Indian Journal of Statistics
2002, Volume 64, Series A, Pt. 1, pp 95-108

AN OPTIMAL ESTIMATING EQUATION WITH UNSPECIFIED VARIANCES

By ANUP DEWANJI
Indian Statistical Institute, Kolkata
LUE PING ZHAO
Fred Hutchinson Cancer Research Center, Seattle, WA

SUMMARY. In estimating equation technique for estimating the mean regression parameters, one important issue is how to choose the weights to improve the efficiency of the estimation. The key idea of this paper is to replace the weights with the empirically estimated covariances. We discuss regression of an outcome on a vector of covariates, and propose an optimal estimating equation approach which achieves asymptotic efficiency. Asymptotic normality of the regression parameter estimates is also established. A small simulation study indicates improved efficiency of this approach.

1. **Introduction**

We consider a general set-up with Y being the response variable and X as the vector of, say d, explanatory variables or covariates. Our purpose in this paper is to estimate the regression parameter vector θ which specifies the relationship of the mean of the response with the covariates via the following function

$$E[Y|x] = \mu(x, \theta) \ ,$$

where $\mu(x, \theta)$ is a specified function up to a vector of unknown parameters θ, based on n observations (y_i, x_i), for $i = 1, \cdots, n$. Here, (y_i, x_i) denotes the observed values of Y and of covariates X on the ith individual. A special but important case included in the above model is the regression model in which $\mu(x, \theta)$ is a function of $x^T\theta$. Likelihood based estimation of θ is common in such cases when distribution of Y is known, which is covered under the broad heading of generalized linear models (McCullagh and Nelder, 1989).

Paper received March, 2000; revised August, 2001.
AMS (1991) Subject classification.
Key words and phrases. Estimating equations, Generalized linear model, Mean regression, Kernel smoothing

Evolving from the generalized linear model is the quasi-likelihood, which requires, instead of a distributional assumption, an assumption of only variance function, denoted by $V(x, \theta) = Var[Y|X = x]$. The estimate of θ satisfies the quasi-score estimating equation, $\sum_i (\partial\mu_i/\partial\theta)V_i^{-1}(y_i - \mu_i) = 0$, where $V_i = V_i(\theta) = V(x_i, \theta)$ and $\mu_i = \mu_i(\theta) = \mu(x_i, \theta)$. Interestingly, the estimate from such an equation has all of the desirable properties as a likelihood-based estimate, but is more robust (McCullagh and Nelder, 1989, p328).

When neither distribution nor variance has been specified, estimating equation technique has been proposed to estimate θ in the univariate regression (Huber, 1967; White, 1982) and also in the multivariate analysis (Liang and Zeger, 1986; Prentice and Zhao, 1991). Specifically, rather than assuming a particular variance function, one may choose a weighting function, w_i, and then estimate θ as a solution to the estimating equation,

$$\sum_i (\partial\mu_i/\partial\theta)w_i^{-1}(y_i - \mu_i) = 0, \tag{1}$$

for θ. It can be shown easily that the above estimating equation yields consistent estimate of θ, and the estimate has an asymptotic normal distribution with an easily estimated asymptotic variance-covariance matrix (Liang and Zeger, 1986). Hence, this estimate is more robust than either likelihood-based or quasi-likelihood-based estimate. However, the price for this gain of the robustness by the estimating equation technique is possible inefficiency. Godambe and Heyde (1987) (see also McCullagh and Nelder, 1989, ch 9.5) have shown that the optimum choice of this weight function is the variance function itself or its proportionality in the sense that the asymptotic variance of any linear combination of the estimate of the parameter vector θ is minimized. For example, with binary response, the variance function is fully specified by the mean of the response, $V_i = \mu_i(1 - \mu_i)$. Hence, the weighting function should be chosen to equal this variance function. In general, however, assuming a variance function usually requires an untestable and yet nuisance assumption about an aspect of the random response process. For example, with count response, if it is resulted from a sum of independently distributed binary random responses, the count has a binomial distribution, and thus has a specific variance function. However, in the presence of dependence between these binary random responses or of heterogeneity in their means, the count response does not have the binomial distribution, and the variance of the count responses has a so-called over-dispersion, which could be arbitrary depending on the source for the over-dispersion. A usual choice of the over-dispersion parameter, resulted from either beta-binomial or equal within-correlation, represents a strong assumption about the variance function, and, often, is not verifiable from the available data.

How to improve efficiency of estimating θ by the estimating equation with unspecified variance function is of great interest. Our idea is to estimate variance function nonparametrically, and to replace the weighting function w_i in (1) with the estimated variance function. Intuitively, the nonparametrically estimated variance function will approximate the true variance function, and hence the estimating equations with the estimated variance function will achieve the optimality in estimating the regression coefficients in the absence of any knowledge regarding the true variance function. Among many nonparametric regression techniques, we choose the kernel smoothing method. One way of estimating the variance function is to use two applications of kernel smoothing to estimate $E[Y^2|x]$ and $E[Y|x]$, respectively. However, a more efficient way is to exploit the assumed mean structure $\mu(x, \theta)$ and estimate $E[(Y - \mu(x, \theta))^2 \, |x]$ using one application of kernel smoothing. This idea has been considered by Carroll (1982) to deal with linear regression with unknown heteroscedasticity. Carroll (1982) suggested the kernel smoothing technique to estimate the variance function based on the squared residuals which are obtained through an ordinary least square estimates of the parameters. Then, with estimated variances as weights one can obtain the weighted least squares estimates of the regression coefficients in linear regression (see also Carroll and Ruppert, 1988, p110-113). Estimate of θ thus obtained by using nonparametrically estimated variances (weights) has the same asymptotic distribution as if the variances were known. The proof of this result has been provided, with long and tedious algebra, by Carroll (1982) for linear regression with one covariate and also by Robinson (1987) and Müller and Stadtmüller (1987) in some general cases. The purpose of this work, although in principle similar to that of Carroll (1982), is two-fold. First, it considers the estimating equation approach, thus avoiding any distributional assumption, with the most general form of mean function $\mu(x, \theta)$ and suggests an iterative method alternating between estimation of the V_i's and estimation of the parameter of interest θ to achieve both the efficiency and robustness (see Section 2). Secondly, this approach leads to a simple proof of the asymptotic results in the most general case (see section 3) and allows for natural extension to multivariate response data (see section 5). Section 4 presents a small simulation study to investigate the performance of this method based on nonparametric estimation of variance.

2. An Optimal Estimating Equation

We use kernel smoother (Nadaraya, 1964; Watson, 1964) to estimate the variance function $V(x,\theta)$, assumed to be continuous in x, as

$$\hat{V}(x,\theta)=\frac{\sum_{j=1}^{n}K[h^{-1}(x-x_j)][y_j-\mu_j(\theta)]^2}{\sum_{j=1}^{n}K[h^{-1}(x-x_j)]}\ ,$$

where $K(\cdot)$ denotes the kernel function of order r, say, and x denotes a typical covariate vector. In particular, such a kernel estimate for V_i can be written as

$$\hat{V}_i=\hat{V}_i(\theta)=(\sum_{j=1}^{n}D_{ij}S_j^2)/\sum_{j=1}^{n}D_{ij}, \tag{2}$$

where $D_{ij}=K[h^{-1}(x_i-x_j)]$ and $S_j=y_j-\mu_j(\theta)$. The choice of the above kernel smoother does not preclude the use of other smoothing technique, except that this choice offers an advantage for the theoretical proof of the asymptotic properties to be discussed in the next section. As in (1), assuming μ_i to be differentiable in θ, let us write, for a fixed θ, the estimating equations

$$U_n(\theta,V)=n^{-1/2}\sum_{i=1}^{n}\dot{\mu}_iV_i^{-1}S_i\ , \tag{3}$$

where $\dot{\mu}_i=\frac{\partial\mu_i(\theta)}{\partial\theta}$, and

$$U_n(\theta,\hat{V})=n^{-1/2}\sum_{i=1}^{n}\dot{\mu}_i\hat{V}_i^{-1}S_i\ . \tag{4}$$

We suggest estimating θ by solving the optimal estimating equation $U_n(\theta,\hat{V})=0$ for θ. Implementing this is rather straightforward, similar to reweighted least square, via the following algorithm with a fixed bandwidth h for kernel smoothing.

Step 1: Set $V_i=1$ for all i and solve $U_n(\theta,V)=0$ (see (3)) for θ to obtain an initial estimate θ^0.

Step 2: Obtain the nonparametrically estimated variances $\hat{V}_i^0$ with $\theta=\theta^0$ as described in (2), for all i.

Step 3: Solve $U_n(\theta,\hat{V})=0$ for θ, with $\hat{V}_i$'s replaced by $\hat{V}_i^0$'s, to obtain an improved estimate θ^1. This step may be iterative.

Step 4: Go back to *Step 2* with θ^0 replaced by θ^1 and continue the iteration until convergence.

Carroll (1982) suggested one iteration of steps 1-3 above to settle for the estimate θ^1. Clearly, the estimate obtained at the final iteration satisfies $U_n(\theta, \hat{V}) = 0$. This estimate is denoted by $\hat{\theta}(h)$ to emphasize the dependence on the bandwidth h. In principle, one should find an optimal choice of the bandwidth h by minimizing, for example, the asymptotic variance of $\hat{\theta}(h)$ (see the next section). In this paper, we do not consider such optimal choice of h, but estimation of θ for a fixed bandwidth h. Hence, we write $\hat{\theta}(h) = \hat{\theta}$ suppressing the dependence on h. Given the order r, usually the choice of kernel function has little effect on the nonparametric estimates, given here by (2). However, selection of the bandwidth h is crucial. The condition (5) for the asymptotic results in the following section gives a guideline for such a selection.

3. Asymptotic Results

Note that, from (3), $E[U_n(\theta, V)] = 0$ and, since $Var[S_i] = V_i$, we have

$$Var[U_n(\theta, V)] = n^{-1} \sum_{i=1}^{n} \dot{\mu}_i \dot{\mu}_i^T V_i^{-1},$$

where $\dot{\mu}_i^T$ denotes the transpose of $\dot{\mu}_i$. Then, assuming that the limit exists, write $V_U = V_U(\theta) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \dot{\mu}_i \dot{\mu}_i^T V_i^{-1}$ as the asymptotic variance of $U_n(\theta, V)$, for a fixed θ.

Let us first specify the sufficient condition on the bandwidth h which is:

$$h \to 0, \quad nh^{2d} \to \infty \quad \text{and} \quad nh^{2r} \to 0 \quad \text{as} \quad n \to \infty. \tag{5}$$

This condition has been recently worked with by several authors (Carroll and Wand, 1991; Carroll et al., 1995; Wang et al., 1997). This clearly requires $r > d$. Thus, higher order kernels will be required for larger values of d. Note that

$$\begin{aligned} E[\hat{V}_i] &= \frac{\sum_{j=1}^{n} D_{ij} V_j}{\sum_{j=1}^{n} D_{ij}} \\ &= V_i + O\left(h^r + \sqrt{\frac{h^{2-d}}{n}}\right), \end{aligned} \tag{6}$$

for $i = 1, \cdots, n$, which can be easily verified by using the technique similar to Silverman (1986, p38-40). The following result (7) will be extensively used in further derivations:

Assuming $E[S_i^4] < \infty$, we have

$$\hat{V}_i = V_i + O_p(h^r + \frac{1}{\sqrt{nh^d}}) \ , \tag{7}$$

for all $i = 1, \cdots, n$. To give a sketch of the proof for (7), write $W_j = Var[S_j^2]$, which is finite since $E[S_j^4]$ is, for all j. Then, note that

$$Var[\hat{V}_i] = \frac{\sum_{j=1}^n D_{ij}^2 W_j}{(\sum_{j=1}^n D_{ij})^2} = O(\frac{1}{nh^d}) \ , \tag{8}$$

using the same technique (Silverman, 1986, p38-40). Then, using (6), we have the result.

Note that $\hat{\theta}$ satisfies $U_n(\theta, \hat{V}(\theta)) = 0$ (see (4)). By a Taylor series expansion, we have

$$n^{1/2}(\hat{\theta} - \theta) = \left[-n^{-1/2}\frac{\partial}{\partial\theta}U_n(\theta, \hat{V}(\theta))\right]^{-1} U_n(\theta, \hat{V}(\theta)) + o_p(1) \ . \tag{9}$$

Write $\eta_n = \{nh^{2r} + 1/(nh^{2d})\}^{1/2}$. Consider the following linear approximation (see Wang et al., 1997):

$$\begin{aligned}
U_n(\theta, \hat{V}) - U_n(\theta, V) &= n^{-1/2}\sum_{i=1}^n \dot{\mu}_i S_i \left(\frac{1}{\hat{V}_i} - \frac{1}{V_i}\right) \\
&= n^{-1/2}\sum_{i=1}^n \dot{\mu}_i S_i \left[-\frac{\hat{V}_i - V_i}{V_i^2} + O_p(h^{2r} + \frac{1}{nh^d})\right], \\
&\qquad \text{using (7)} \\
&= -n^{-1/2}\sum_{i=1}^n \frac{\dot{\mu}_i S_i}{V_i^2}(\hat{V}_i - V_i) + O_p(\eta_n) \\
&= -n^{-1/2}\sum_{i=1}^n \frac{\dot{\mu}_i S_i}{V_i^2}\left[\sum_{j=1}^n \frac{D_{ij}(S_j^2 - V_i)}{\sum_{k=1}^n D_{ik}}\right] + O_p(\eta_n) \\
&= -n^{-1/2}\sum_{i=1}^n \frac{\dot{\mu}_i S_i}{V_i^2}\frac{D_{ii}(S_i^2 - V_i)}{\sum_{k=1}^n D_{ik}} \\
&\qquad -n^{-1/2}\sum_{i=1}^n \sum_{j \neq i} \frac{\dot{\mu}_i S_i}{V_i^2}\frac{D_{ij}(S_j^2 - V_i)}{\sum_{k=1}^n D_{ik}} + O_p(\eta_n) \\
&= A_n + B_n + O_p(\eta_n), \text{ say.}
\end{aligned} \tag{10}$$

Assuming $E[S_i^6] < \infty$, it is easy to verify that $A_n = O_p(\eta_n)$. Note that $E[B_n] = 0$. With routine (but long and tedious) calculations, one can also verify that $Var[B_n] = O_p(\eta_n^2)$ (see the Appendix for a sketch of the proof). Hence, $B_n = O_p(\eta_n)$. Thus, we have, from (10), $U_n(\theta, \hat{V}) - U_n(\theta, V) = O_p(\eta_n)$; or,

$$U_n(\theta, \hat{V}) = U_n(\theta, V) + O_p(\eta_n) \; . \tag{11}$$

The asymptotic normality of $U_n(\theta, \hat{V})$ is now easily seen from (11) with asymptotic mean and variance same as that of $U_n(\theta, V)$, that is 0 and V_U, respectively. Therefore, $U_n(\theta, \hat{V})$ is asymptotically unbiased and, hence, $\hat{\theta}$ converges in probability to θ (Foutz, 1977), assuming $V_U(\theta)$ to be positive definite. Because of (7), $n^{-1}\sum_{i=1}^n \dot{\mu}_i \dot{\mu}_i^T \hat{V}_i^{-1}$ can be used as a consistent estimate of V_U. The asymptotic normality of $n^{1/2}(\hat{\theta} - \theta)$ now follows from (9).

Note that the term inside [ ] in (9) can be written as

$$-\frac{\partial}{\partial\theta}\left[n^{-1}\sum_{i=1}^n \dot{\mu}_i \hat{V}_i^{-1} S_i\right] = -n^{-1}\sum_{i=1}^n \left[\left(\frac{\partial}{\partial\theta}\dot{\mu}_i \hat{V}_i^{-1}\right) S_i - \dot{\mu}_i \dot{\mu}_i^T \hat{V}_i^{-1}\right] \; .$$

Using (7) and (5), the first term can be shown to be $o_p(1)$, following the same derivation as (11), and the second term is

$$n^{-1}\sum_{i=1}^n \dot{\mu}_i \dot{\mu}_i^T \hat{V}_i^{-1} = n^{-1}\sum_{i=1}^n \dot{\mu}_i \dot{\mu}_i^T V_i^{-1} + o_p(1) \to V_U \; .$$

Hence, the asymptotic variance of $n^{1/2}(\hat{\theta} - \theta)$ is, from (9) and using (11), V_U^{-1}. However, noting that $U_n(\theta, \hat{V}) = \sum_{i=1}^n U_{n,i}(\theta, \hat{V}_i)$ with $U_{n,i}(\theta, \hat{V}_i) = n^{-1/2}\dot{\mu}_i \hat{V}_i^{-1} S_i$, a finite sample approximation for the variance of $U_n(\theta, \hat{V})$ can be taken as

$$\sum_{i=1}^n U_{n,i}(\theta, \hat{V}_i) U_{n,i}(\theta, \hat{V}_i)^T.$$

Then, an alternative estimate for the asymptotic variance of $n^{1/2}(\hat{\theta} - \theta)$ is

$$\left(\frac{1}{n}\sum_{i=1}^n \dot{\mu}_i \dot{\mu}_i^T \hat{V}_i^{-1}\right)^{-1} \left(\sum_{i=1}^n U_{n,i}(\theta, \hat{V}_i) U_{n,i}(\theta, \hat{V}_i)^T\right) \left(\frac{1}{n}\sum_{i=1}^n \dot{\mu}_i \dot{\mu}_i^T \hat{V}_i^{-1}\right)^{-1} \; , \tag{12}$$

evaluated at $\theta = \hat{\theta}$, the so called 'sandwitch estimate'.

4. A Simulation Study

We conduct a simulation study to investigate the finite sample performance of our method based on nonparametric estimate of the variance function. One important objective of this study is also to investigate the extent of efficiency gain by the suggested iterative method over the one-iteration method of Carroll (1982) for linear regression. For this purpose, we simulate regression data from heteroscedastic Normal, Binomial and Poisson distributions and analyze the data assuming only their mean structure. In each case, for a sample of size n, the n values of the regressor X (assumed scalar for our simulation) are generated from $\mathcal{U}[0, 10]$ distribution. The models for simulation, given $X = x$, assume different mean structures for the corresponding response variable Y, given by $\mu(x, \theta)$ as functions of $\alpha + \beta x$, where $\theta = (\alpha, \beta)$.

Firstly, the conditional distribution of Y, given $X = x$, is assumed to be Normal with mean $\mu(x, \theta)$ and variance cx. For the model parameters, we choose different combinations of $\theta = (5.0, 1.0)$, $(20.0, 0.0)$ and $c = 0.5,\ 1.5$. For the Binomial distribution of Y given x, number of trials m is taken as 10 and *logit* of the success probability is assumed to be $\alpha + \beta x$ so that

$$\mu(x, \theta) = 10 \times \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}\ .$$

We choose $\theta = (-4.6, 0.5)$ and $(-2.95, 0.3)$. Thirdly, for the Poisson distribution, we assume $\mu(x, \theta) = \exp(\alpha + \beta x)$ and choose $\theta = (1.61, 0.11)$ and $(1.61, 0.16)$. These parameter values were chosen to reflect different levels of heteroscedasticity.

For a particular model and given the n values of X, we carry out 1000 simulations. In each of them, the n values of Y are generated from the corresponding distribution and θ is estimated by the estimating equation approach with four different weight functions as follows: (I) assuming the variance structure (that is, by solving (3)) which gives asymptotically the most efficient estimate, (II) assuming homoscedasticity (that is, $V_i = V(x_i, \theta) = 1$) which is step 1 of our algorithm in section 2, (III) nonparametrically estimating the variance by kernel smoothing (see (2) and step 2 of our algorithm) with $\theta = \theta^0$ obtained from (II) above and then carrying out step 3 of our algorithm once (similar to Carroll's method), and (IV) same as (III) above but iterating steps 2-4 of our algorithm five times (for computational simplicity, we do not iterate till convergence). For the nonparametric estimates of variance functions, we use the Epanechnikov's kernel function given by $K(x) = 0.75(1 - x^2),\ |x| \leq 1$, and a window length h that is proportional to $n^{-1/3}$.

We first consider estimating the asymptotic relative efficiency (ARE) of the initial estimate of θ, from (II), by assuming equal variance, with respect to that from (I). This is obtained by the ratio of the corresponding variance estimates (from (I) and (II)) based on the estimates of θ from 1000 simulations and denoted by ARE0 in the tables. Next, we estimate ARE for the one-iteration estimate from (III), with respect to that from (I), by using (i) the variance estimate based on $\hat{V}_U$ (that is, $n^{-1}\sum_{i=1}^{n}\dot{\mu}_i\dot{\mu}_i^T\hat{V}_i^{-1}$), and also ($ii$) the sandwitch estimate given by (12). The variance estimates in the numerator (that is, for the estimate from (I)), in both (i) and (ii), are also based on $\hat{V}_U$ and equation (12), respectively, but using the true variance structure. These are denoted by ARE1 and AREs1, respectively. The same is done for our estimates from (IV) using five iterations and are denoted by ARE5 and AREs5, respectively. The two entries in each cell of the tables are the corresponding ARE's for the estimates of the intercept and slope parameters, respectively. The results are presented in Tables 1-3 for Normal, Binomial and Poisson models, respectively. The simulation results indicate small bias in the estimates (not reported here), as expected, which reduces with increasing sample size.

Table 1. Asymptotic relative efficiency for Normal models.

θ	c	n	ARE0	ARE1	ARE5	AREs1	AREs5
(5.0,1.0)	0.5	25	.505	.561	.586	.604	.620
			.639	.849	.861	.868	.885
		50	.307	.481	.510	.656	.676
			.530	.832	.849	.931	.945
		100	.344	.510	.530	.666	.677
			.566	.810	.821	.900	.907
	1.5	25	.401	.542	.586	.767	.801
			.641	.932	.952	.995	.998
		50	.217	.269	.274	.207	.210
			.525	.648	.650	.640	.647
		100	.126	.158	.161	.128	.130
			.417	.531	.535	.580	.585
(20.0,0.0)	0.5	25	.833	.946	.963	.968	.976
			.870	.989	.996	.959	.966
		50	.331	.425	.436	.525	.536
			.571	.768	.772	.840	.849
		100	.281	.425	.441	.541	.554
			.522	.697	.708	.796	.807
	1.5	25	.589	.698	.727	.809	.829
			.711	.912	.927	.959	.975
		50	.250	.348	.356	.424	.433
			.482	.697	.700	.783	.794
		100	.414	.483	.493	.563	.568
			.581	.765	.770	.824	.828

The ARE's of one-iteration and five-iteration estimates (from (III) and (IV), respectively) are evidently larger than those from (II), obtained by assuming homoscedasticity. Therefore, using nonparametric estimate of the variance function is better than not using any weight. Also, the five-iteration estimate is more efficient (although marginally in most cases) than the one-iteration estimate. The ARE of the slope parameter estimate is generally seen to be more than that of the corresponding intercept parameter estimate. So, the uncertainty due to estimation of variance function seems to have more effect on the intercept parameter estimate. The sandwitch estimate of variance, in general, results in higher ARE, which indicates that its use may lead to more efficient interval estimation.

Table 2. Asymptotic relative efficiency for Binomial models.

θ	n	ARE0	ARE1	ARE5	AREs1	AREs5
(-4.6,0.5)	25	.592	.886	.897	.927	.940
		.638	.970	.975	.956	.967
	50	.452	.886	.907	.947	.959
		.508	.951	.964	.973	.984
	100	.573	.836	.841	.957	.961
		.612	.877	.881	.963	.966
(-2.95,0.3)	25	.767	.951	.963	.992	.998
		.802	.980	.987	.990	.995
	50	.812	.948	.955	.997	.999
		.842	.994	.997	.994	.996
	100	.816	.928	.931	.995	.996
		.846	.979	.980	.995	.996

Table 3. Asymptotic relative efficiency for Poisson models.

θ	n	ARE0	ARE1	ARE5	AREs1	AREs5
(1.61,0.11)	25	.881	.964	.988	.988	.993
		.892	.952	.966	.983	.987
	50	.906	.970	.980	.985	.987
		.892	.965	.971	.987	.989
	100	.892	.986	.992	.986	.987
		.888	.981	.983	.971	.972
(1.61,0.16)	25	.775	.957	.982	.973	.985
		.798	.978	.994	.975	.986
	50	.860	.962	.971	.967	.970
		.866	.977	.983	.982	.984
	100	.855	.985	.991	.992	.994
		.847	.980	.982	.977	.979

5. Discussion

The variance of the response variable as a function of the covariates may be misspecified because of the wrong modeling assumption or presence of overdispersion, etc. The assumption of arbitrary variance function, and the use of a nonparametric estimate for it, as in section 2, makes the estimates of the regression parameters robust against such misspecification. Although, in the context of mean regression problem, variance function is a nuisance component, its nonparametric estimation ensures that the estimation of mean parameters is optimal while alleviating the problem of choosing a variance function. For this purpose, one could use any other nonparametric estimate of the V_i's instead of the kernel smoother, for example, using other smoothing techniques or by using replications whenever they are available at the different x_i's (Fuller and Rao, 1978). There could be situations when estimating variance function is of primary interest. For example, in studies of HIV infection, while CD_4 cell counts measure the immune status of patients, their variability over a period describes the stability of infection. In monitoring quality control, the primary objective is to quantify variability of the key outcome. One of the study objectives in genetic analysis of variance is to assess how much variation of phenotype may be explained by genetic factors.

Although the method described in section 2 implicitly assumes that the covariates are continuous randomly selected from a covariate space, it works in principle for fixed covariates and also for discrete or categorized covariates. The asymptotic results are, however, difficult in the latter case. The variance estimates $\hat{V}_i$'s, while plotted against the covariates, gives an idea if only a subset of the covariates affects the variance and hence a lower dimensional kernel smoother may be adopted. If the variance is assumed to be a smooth function of a single argument $x^T\theta$, or the mean μ, as is often the case, one can use a univariate kernel function; it alleviates the rather computational problem of having to work with a multivariate kernel function (Silverman, 1986, Chapter 4). This kind of dimension reduction has also been considered by Carroll et al. (1995) in the context of semiparametric regression with errors in covariates.

In this paper, we worked with only univariate response data. The estimating equation approach described in this paper is easily extended for multivariate response data in which V_i is to be interpreted as the covariance matrix of Y_i, the response vector corresponding to $X = x_i$. The nonparametric estimate of the matrix V_i in this case can be derived as before by obtaining the kernel smoother for each of its elements. For example, the

(j,j')th element of V_i, $V_{i(jj')}$ say, can be estimated as

$$\hat{V}_{i(jj')} = \frac{\sum_l D_{il} \times (y_{lj} - \mu_{lj}(\theta))\,(y_{lj'} - \mu_{lj'}(\theta))}{\sum_l D_{il}} \;,$$

where $Y_{ij} = y_{ij}$ denotes the jth component of $Y_i = y_i$ and $\mu_{ij}(\theta) = E[Y_{ij}|X = x_i]$. The estimate $\hat{V}_i$ thus obtained is used as a 'working covariance matrix' in the usual estimating equation approach.

Acknowledgement. This work has been supported in part by a fellowship grant from UICC, and by NCI grant no. 47658 and 64046. The authors wish to thank Dr. C. Y. Wang for many helpful discussion during this work and also the two anonymous referees for their helpful comments.

Appendix

$$\begin{aligned}
&Var[B_n] \\
&= n^{-1}\left[\sum_i \sum_{j\neq i} Var\left(\frac{\dot{\mu}_i}{V_i^2}\frac{D_{ij}}{\sum_k D_{ik}} S_i(S_j^2 - V_i)\right)\right. \\
&\quad +\sum_{i\neq i'} \sum_{j,j':j\neq i, j'\neq i'} Cov\left(\frac{\dot{\mu}_i S_i}{V_i^2}\frac{(S_j^2 - V_i)D_{ij}}{\sum_k D_{ik}}, \frac{\dot{\mu}_{i'} S_{i'}}{V_{i'}^2}\frac{(S_{j'}^2 - V_{i'})D_{i'j'}}{\sum_k D_{i'k}}\right) \\
&\quad \left. +\sum_i \sum_{j\neq j'(\neq i)} Cov\left(\frac{\dot{\mu}_i S_i}{V_i^2}\frac{(S_j^2 - V_i)D_{ij}}{\sum_k D_{ik}}, \frac{\dot{\mu}_i S_i}{V_i^2}\frac{(S_{j'}^2 - V_i)D_{ij'}}{\sum_k D_{ik}}\right)\right] \\
&= n^{-1}\left[B_{1n} + B_{2n} + B_{3n}\right], \text{ say.} \qquad (13)
\end{aligned}$$

It is easy to see that

$$\begin{aligned}
B_{1n} &= \sum_i \sum_{j\neq i} \frac{\dot{\mu}_i^2}{V_i^4}\frac{D_{ij}^2}{(\sum_k D_{ik})^2} V_i(W_j + V_j^2 - 2V_iV_j + V_i^2) \\
&= \sum_i O(\frac{1}{nh^d}), \quad \text{as in (8).}
\end{aligned}$$

Thus $n^{-1}B_{1n} = O(\frac{1}{nh^d}) = O(\eta_n^2)$. Note that $B_{2n} = 0$ and

$$\begin{aligned}
B_{3n} &= \sum_i \frac{\dot{\mu}_i^2}{V_i^4} \sum_{j\neq j'(\neq i)} \frac{D_{ij}D_{ij'}}{(\sum_k D_{ik})^2} V_i(V_j - V_i)(V_{j'} - V_i) \\
&= \sum_i \frac{\dot{\mu}_i^2}{V_i^3} \sum_{j\neq i} \frac{D_{ij}(V_j - V_i)}{\sum_k D_{ik}} \sum_{j\neq j'(\neq i)} \frac{D_{ij'}(V_{j'} - V_i)}{\sum_k D_{ik}}
\end{aligned}$$

$$= \sum_i \left(O(h^r + \sqrt{\frac{h^{2-d}}{n}}) \right)^2, \quad \text{as in (6).}$$

Thus $n^{-1}B_{3n} = O\left(h^{2r} + \frac{1}{nh^{d-2}}\right) = O(\eta_n^2)$. Hence, from (13), we have $Var[B_n] = O(\eta_n^2)$.

References

Carroll, R.J. (1982). Adapting for heteroscedasticity in linear models. *Ann. Statist.* **10**, 1224-1233.

Carroll R.J. and Ruppert D. (1988). *Transformations and Weighting in Regression.* Chapman and Hall, New York.

Carroll, R.J. and Wand M.P. (1991). Semiparametric estimation in logistic measurement error models. *J. Roy. Stat. Assoc., Series B* **53**, 573-585.

Carroll R.J., Knickerbocker R.K. and Wang, C.Y. (1995). Dimension reduction in a semiparametric regression model with errors in covariates. *Ann. Statist.* **23**, 161-181.

Foutz, R.V. (1977). On the unique consistent solution to the likelihood equations. *J. Amer. Statist. Assoc.* **72**, 147-148.

Fuller, W.A. and Rao. J.N.K. (1978). Estimation for a linear regression model with unknown diagonal covariance matrix. *Ann. Statist.* **6**, 1149-1158.

Godambe V.P. and Heyde, C.C. (1987). Quasi-likelihood and optimal estimation. *Int. Statist. Review* **55**, 231-244.

Huber P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability,* 221-233.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd edition. Chapman and Hall, London.

Müller, H.G. and Stadtmüller U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15**, 610-625.

Nadaraya, E.A. (1964). On estimating regression. *Theor. Probab. Applications* **10**, 186-190.

Prentice, R.L. and Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**, 825-839.

Robinson P.M. (1987). Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica* **55**, 875-892.

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

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.

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

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1-25.

Anup Dewanji
Applied Statistics Unit
Indian Statistical Institute
Kolkata 700 108
E-mail: dewanjia@isical.ac.in

Lue Ping Zhao
Division of Public Health Sciences
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue N
Seattle, WA 98109-1024
E-mail: lzhao@fhcrc.org