Biometrika (2003), **90,** 1, *pp.* 43–52
© 2003 Biometrika Trust
Printed in Great Britain

Marginal nonparametric kernel regression accounting for within-subject correlation

By NAISYIN WANG

Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.

nwang@stat.tamu.edu

Summary

There has been substantial recent interest in non- and semiparametric methods for longitudinal or clustered data with dependence within clusters. It has been shown rather inexplicably that, when standard kernel smoothing methods are used in a natural way, higher efficiency is obtained by assuming independence than by using the true correlation structure. It is shown here that this result is a natural consequence of how standard kernel methods incorporate the within-subject correlation in the asymptotic setting considered, where the cluster sizes are fixed and the cluster number increases. In this paper, an alternative kernel smoothing method is proposed. Unlike the standard methods, the smallest variance of the new estimator is achieved when the true correlation is assumed. Asymptotically, the variance of the proposed method is uniformly smaller than that of the most efficient working independence approach. A small simulation study shows that significant improvement is obtained for finite samples.

Some key words: Asymptotic relative efficiency: Asymptotics; Bandwidth: Generalised estimating equation; Local linear estimator: Working covariance matrix; Working independence estimator.

1. Introduction

There has been substantial recent interest in nonparametric marginal estimation for longitudinal/clustered data; see Lin & Carroll (2000, 2001) for a brief summary of the literature. Various authors such as Hoover et al. (1998), Lin & Carroll (2000, 2001), Lin & Yin (2001), Ruckstuhl et al. (2000) and Zeger & Diggle (1994) recommend an approach which effectively ignores the within-subject correlation completely and treats the data as if they are independent. Lin & Carroll (2000) refer to this approach as a 'working independence' method. In the case where the cluster size is finite, they also provide theoretical evidence indicating that the most efficient standard kernel estimator of the nonparametric function is obtained by completely ignoring the within-subject correlation; that is, correct specification of the true correlation structure results in an asymptotically less efficient estimator than the 'working independence' approach.

On the other hand, the semiparametric efficient score derived by Lin & Carroll (2001) indicates that, in order to obtain a semiparametric efficient estimator, the true correlation has to be incorporated into both parametric and nonparametric estimation. Furthermore, the standard nonparametric estimator will not be a suitable choice even when it uses the true correlation structure. In fact, the estimated parametric parameters could be

$\sqrt{n}$-inconsistent unless either a 'working independence' assumption or an under-smoothing step is adopted.

In view of these interesting but somewhat inexplicable results, there is a need to develop an alternative nonparametric kernel method that can make proper use of the correlation among observations from the same subject. The purpose of this paper is to construct such an estimator. We do so by first showing in § 2·3 that Lin & Carroll's results are a natural consequence of the standard estimating equation. In particular, we show that for bounded cluster sizes, as bandwidth tends to zero, the chance that two or more observations from the same cluster have non-negligible kernel weights tends to zero. We then modify the original estimating equation to one that can account properly for the correlation. The resulting estimator can be viewed as a natural extension of the parametric generalised estimating equations of Liang & Zeger (1986) and Zeger & Liang (1986) in the sense that this approach maintains various properties possessed by the original parametric approach. In particular, the optimal solution is achieved when the true correlation is taken into consideration. The resulting new estimator has a uniformly smaller variance than the working independence estimator.

2. Kernel estimating equations

2·1. *The model*

We consider the same model structure as in Lin & Carroll (2000). Suppose that the data consist of n clusters with the ith cluster having m_i observations. As in Lin & Carroll (2000), we assume that m_i is finite. Let Y_{ij} and X_{ij} respectively be the response variable and the time-varying covariates of the jth observation in the ith cluster. Let $Y^i=(Y_{i1}, \ldots, Y_{im_i})^{\mathrm{T}}$ and analogously define X^i. The observations (X^i, Y^i) are random samples from the distribution of (X, Y). Given X^i, $E(Y_{ij}|X^i)=E(Y_{ij}|X_{ij})=\mu_{ij}$ and $\mathrm{var}(Y_{ij}|X^i)=\phi w_{ij}^{-1}v(\mu_{ij})$, where ϕ is a scale parameter, w_{ij} is a known weight and $v(.)$ is a variance function. The marginal mean μ_{ij} depends on X_{ij} through a known and differentiable link function $\mu(.)$: $\mu_{ij}=\mu\{\theta(X_{ij})\}$, where $\theta(.)$ is an unknown smooth function. In matrix notation,

$$E(Y^i|X^i)=\mu^i=\mu(\theta^i), \quad \mathrm{var}(Y^i|X^i)=\Sigma^i=\Sigma(\mu^i), \tag{1}$$

where θ^i denotes $\theta(X^i)=\{\theta(X_{i1}), \ldots, \theta(X_{im_i})\}^{\mathrm{T}}$. To focus the presentation on the main concepts and keep the paper compact, we will let X_{ij} be a scalar, as in Lin & Carroll (2000). We will also concentrate on the local linear regression setting. Generalisation to a multivariate X using local polynomial regression can be easily achieved following Ruppert & Wand (1994). Throughout the remainder of the paper we let $m_i=m$.

2·2. *Use of standard kernel methods*

Let $K_h(s)=h^{-1}K(s/h)$, where K is a symmetric zero-mean density function. Lin & Carroll (2000) estimate $\theta(x)$ by $\tilde{\theta}(x)=\tilde{\alpha}_0(x)$. With $p=1$, $\tilde{\alpha}(x)=(\tilde{\alpha}_0(x), \tilde{\alpha}_1(x))^{\mathrm{T}}$ solves the following estimating equation:

$$0=n^{-1}\sum_{i=1}^{n}\{G^i(x)\}^{\mathrm{T}}\Delta^i(x)[\{K_h^i(x)\}^{\frac{1}{2}}(V^i)^{-1}\{K_h^i(x)\}^{\frac{1}{2}}][Y^i-\mu\{G^i(x)\}^{\mathrm{T}}\tilde{\alpha}(x)\}], \tag{2}$$

where

$$K_h^i(x)=\mathrm{diag}\{K_h(X_{ij}-x)\},$$

$G(s)=(1,s)^{\mathrm{T}}$, the jth column of $G^i(x)$ contains $G\{(X_{ij}-x)/h\}$,

$$\Delta^i=\operatorname{diag}(\mu^{(1)}[G\{(X_{ij}-x)/h\}^{\mathrm{T}}\tilde{\alpha}(x)]),$$

with $d^{(r)}(.)$ denoting the rth derivative of a function $d(.)$, and $V^i=V(\mu^i,\delta)$ is an invertible working covariance matrix which may depend on an unknown parameter, δ. For example, Liang & Zeger (1986) consider $V^i=(S^i)^{\frac{1}{2}}R^i(\delta)(S^i)^{\frac{1}{2}}$, where $S^i=\operatorname{diag}\{\operatorname{var}(Y_{ij}|X_{ij})\}$ and $R^i(\delta)$ is an invertible working correlation matrix. Equation (2) is a modified extension of equation (6) for independent observations in Severini & Staniswalis (1994).

The most counter-intuitive result of Lin & Carroll is their claim that the most efficient $\tilde{\theta}(x)$ is obtained by entirely ignoring the within-cluster correlation, and that correctly specifying the within-cluster correlation actually results in an asymptotically less efficient estimator of $\theta(x)$. This asymptotic behaviour is quite different from that of the parametric generalised estimating equation. To be precise, for finite m, Lin & Carroll's results indicate that the variance of $\tilde{\theta}(x)$ is minimised when working independence is assumed and it becomes

$$(nh)^{-1}\gamma_0(K)\left[\sum_{j=1}^{m}E\{(\mu_j^{(1)})^2\sigma_{jj}^{-1}|X_j=x\}f_j(x)\right]^{-1}, \tag{3}$$

where $\mu_j^{(1)}$ denotes the jth element of $\mu^{(1)}\{\theta(X)\}$, $f_j(.)$ is the marginal density of X_{ij}, $\gamma_r(K)=\int z^rK^2(z)\,dz$, and σ_{jk} is the (j,k)th element of Σ defined in (1).

2·3. *How standard kernel methods use correlation*

Further understanding of the working independence estimator has been helpful to our construction of the new estimator. Suppose that m is finite and observe that $(V^i)^{-1}$ in (2) is pre- and post-multiplied by a diagonal matrix, $(K_h^i)^{1/2}$, with elements $K_h^{1/2}(X_{ij}-x)$. We note that, if the marginal density f_j of the time-varying X_{ij} is bounded away from 0 and K has compact support, $[-1,1]$, as $h\to 0$ when $n\to\infty$, K_h^i will eventually have only one nonzero element. This observation has two implications: (i) asymptotically, there is effectively one data point per cluster that contributes to the estimate of $\theta(x_0)$ under (2); (ii) with $(V^i)^{-1}$ pre- and post-multiplied by $(K_h^i)^{1/2}$ and if the contributing data point in the ith cluster is point j say, it is indicated below that the matrix inside the square bracket prceding $[Y^i-\mu\{G^i(x)^{\mathrm{T}}\tilde{\alpha}(x)\}]$ in (2) also only has one nonzero element, which is $K_h(X_{ij}-x_0)(v^i)^{jj}$, where $(v^i)^{jj}$ denotes the (j,j) entry of $(V^i)^{-1}$. Thus, asymptotically, the only contributing data point in the ith cluster will be weighted with $K_h(X_{ij}-x_0)(v^i)^{jj}$. Since data from different clusters are independent, the estimator with the smallest variance under the given kernel weighted structure can be obtained by letting $(v^i)^{jj}$ be the reciprocal of the variance of Y_{ij} given X_{ij}. Taking V^i to be diagonal, with σ^{ll}'s as the diagonal elements, would thus provide an estimator with the smallest variance under (2). Choices of V^i other than this 'working independence' covariance matrix may not lead to $(v^i)^{jj}=1/\sigma^{jj}$ and could give a less efficient estimator.

Rather than proving that, under certain regularity conditions, K_h^i will have only one nonzero element for any fixed m when $h\to 0$ *as* $n\to\infty$, we simply illustrate the concept with the following simple example. Suppose X_j, for $j=1,\ldots,m$, are independent and uniformly distributed over $[0,1]$. Also, let $X_{[k]}$ denote the kth order statistic among the X_j, and let $D_k=X_{[k]}-X_{[k-1]}$, for $k=2,\ldots,m$, the distance between two consecutive order statistics. For two or more observations from the same cluster to be used in estimating $\theta(x_0)$, we must have at least one $D_k\leqslant 2h$. However, it is easy to show that D_k follows a beta distribution with parameters $(1, m)$, for all k. Given k, $\mathrm{pr}(D_k \leqslant 2h) \rightarrow 0$, as $h \rightarrow 0$ at any rate. Thus, for finite m, the probability of having at least one $D_k \leqslant 2h$ also goes to 0. The concept embedded in this simple example can be generalised to cases of correlated X_j's with non-uniform marginal distributions f_j. Asymptotically, only one data point per cluster could contribute to the estimate of $\theta(x_0)$ under (2).

2·4. *Marginal kernel method incorporating correlation*

The explanation in § 2·3 shows that not all correlated elements in Y^i are used, mainly because one needs to use kernel weights to control biases, and the way (2) eliminates biases also eliminates input from correlated elements in Y^i. It seems difficult to control the biases and to reduce the variation simultaneously. We thus propose a two-step procedure for accomplishing both tasks. The basic idea is as follows: once a data point, point j say, within a cluster has its X-value within $\pm h$ of x_0 and is used to estimate $\theta(x_0)$, all points in that cluster are used. To avoid bias, the contributions of all points but point j within that cluster to the local estimate of $\theta(x_0)$ are through their residuals, which are calculated by subtracting from the responses Y their estimated means at step one.

Let $G^i_{*j} = [e_j | e_j(x - X_{ij})/h]$, be an $m \times 2$ matrix, where e_j denotes the indicator vector with jth entry equal to 1, and 0 elsewhere. Also, let $\check{\theta}(x)$ be the consistent estimator of $\theta(x)$ obtained at the first step; for example, we might take $\check{\theta}(x) = \tilde{\theta}(x)$, the working independence estimator. In the second step, we determine $\hat{\theta}(x) = \hat{\alpha}_0$, where $\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1)^{\mathrm{T}}$ solves the kernel-weighted estimating equation

$$0 = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(x - X_{ij})[\mu^{(1)}\{\hat{\alpha}_0 + \hat{\alpha}_1(x - X_{ij})\}(G^i_{*j})^{\mathrm{T}}] \times (V^i)^{-1}[Y^i - \mu_{*j}\{X^i, \hat{\alpha}, \check{\theta}(X^i)\}], \tag{4}$$

where the lth element of $\mu_{*j}\{X^i, \hat{\alpha}, \check{\theta}(X^i)\}$ is $\mu\{\hat{\alpha}_0 + \hat{\alpha}_1(x - X_{il})/h\}$, when $l = j$, and is $\mu\{\check{\theta}(X_{il})\}$, when $l \neq j$.

The structure of μ_{*j} indicates that, for a Y_{il} whose X-value is not within h of x, the residual $Y_{il} - \check{\theta}(X_{il})$, rather than Y_{il}, is taken into account in the weighted average that gives $\hat{\theta}(x) = \hat{\alpha}_0(x)$. Since the mean of this residual goes to 0 as n goes to ∞, at worse, the asymptotic bias of the proposed estimator is still 0.

For illustration, we consider the linear case where $Y_{ij} = \theta(X_{ij}) + \varepsilon_{ij}$, and obtain the following asymptotic expression for $\hat{\theta}$ when $\check{\theta} = \tilde{\theta}$:

$$\hat{\theta}(x) \simeq \frac{\sum_i \sum_j K_h(X_{ij} - x)\{(v^i)^{jj} Y_{ij} + \sum_{l \neq j} (v^i)^{jl}(Y_{il} - \tilde{\theta}_{il})\}}{\sum_i \sum_j K_h(X_{ij} - x)(v^i)^{jj}} \tag{5}$$

$$\simeq \theta(x) + \frac{\sum_i \sum_j K_h(X_{ij} - x)\{\sum_l (v^i)^{jl} \varepsilon_{il}\}}{\sum_i \sum_j K_h(X_{ij} - x)(v^i)^{jj}} + \text{Bias}, \tag{6}$$

where $(v^i)^{jl}$ denotes the (j, l) entry of $(V^i)^{-1}$. If we compare (5) to the first of the two following expressions of the working independence estimator $\tilde{\theta}$, namely

$$\tilde{\theta}(x) \simeq \frac{\sum_i \sum_j K_h(X_{ij} - x)(v^i)^{jj} Y_{ij}}{\sum_i \sum_j K_h(X_{ij} - x)(v^i)^{jj}}$$

$$\simeq \theta(x) + \frac{\sum_i \sum_j K_h(X_{ij} - x)(v^i)^{jj} \varepsilon_{ij}}{\sum_i \sum_j K_h(X_{ij} - x)(v^i)^{jj}} + \text{Bias}, \tag{7}$$

we see that the updating step is simply adding weighted 'residuals' obtained from the vector Y^i when X_{ij} is within h of x but X_{il} is not. This can be implemented by placing the lth residual in place of Y_{il} in Y^i, pre-multiplying the new vector by $(V^i)^{-1}$, and then using the resulting vector in the original local polynomial regression algorithm. The extra computation is therefore minimal. The advantage of the new estimator is a reduction in variance. This can be easily seen heuristically by comparing (6) and (7). In (6), we can account properly for the correlation structure in $(V^i)^{-1}$ and use all random elements in Y^i through ε^i without inducing nonzero asymptotic biases. The formal variance comparison is given in § 3 and is proved in the Appendix.

When $\check{\theta}=\tilde{\theta}$, $\hat{\theta}$ is a one-step update of the working independence estimator. An alternative approach is to use that $\hat{\theta}$ as a new $\check{\theta}$ and iterate to solve (4). The theoretical investigation in § 3 indicates that, to first order, the two estimators have very similar asymptotic properties.

3. Asymptotic properties

Let $\mathscr{C}$ be the union of supports of f_j and assume that there exists a j such that $\inf_{x\in\mathscr{C}} f_j(x)$ is bounded away from 0. Unless otherwise stated, x is assumed to be an interior point of $\mathscr{C}$. Standard regularity conditions, such as that μ, f_j and θ are twice continuously differentiable, are also assumed. Note that the unknown parameter δ in V^i can be estimated at a parametric rate, which is faster than the rate for nonparametric estimation. For purposes such as presenting asymptotic theory, it can be assumed without loss of generality that δ is known. Standard calculations result in Theorem 1, which provides the asymptotic expression for $\hat{\theta}(x)=\hat{\alpha}_0$ defined in (4).

Theorem 1. *Assume that* $h_s^3=o(n^{-\frac{1}{2}})$, *that* $(nh_s)^{-1}=o(n^{-\frac{1}{2}})$ *and that* $\check{\theta}(x)$ *based on the starting bandwidth,* h_s, *has the following asymptotic expression uniformly for* $x\in\mathscr{C}$:

$$\check{\theta}(x)-\theta(x)=W_2^{-1}(x)\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{ij}^{(1)}(v^i)^{jj}K_{h_s}(X_{ij}-x)D_{ij} + c_2(K)b(x)h_s^2+o_p\{(nh_s/\log n)^{-\frac{1}{2}}+h_s^2+n^{-\frac{1}{2}}\}, \qquad (8)$$

where $b(.)$ *is a bounded fixed function,* $c_q(K)=\int t^qK(t)\,dt$, D_{ij} *is a zero-mean bounded variance random variable which can be written as a linear combination of elements in* $Y^i-\mu^i$ *and*

$$W_2(x)=\sum_{j=1}^{m}E\{(\mu_j^{(1)})^2v^{jj}\,|\,X_j=x\}f_j(x). \qquad (9)$$

The regularity conditions which allow this type of uniform property are addressed in the appendix of Carroll et al. (1997). *Assume that* $h\to 0$ *and* $(nhh_s)^{-1}\to 0$. *Then we obtain*

$$\hat{\theta}(x)-\theta(x)=W_2^{-1}(x)\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{ij}^{(1)}K_h(X_{ij}-x)\left\{\sum_{l=1}^{m}(v^i)^{jl}(Y_{il}-\mu_{il})\right\} + c_2(K)\mathscr{B}(x)+o_p\{(nh)^{-\frac{1}{2}}+h^2+n^{-\frac{1}{2}}\}, \qquad (10)$$

where

$$\mathscr{B}(x)=h^2\theta^{(2)}(x)/2-h_s^2W_2^{-1}(x)\sum_{j=1}^{m}E\left\{\sum_{l\neq j}\mu_j^{(1)}v^{jl}\mu_l^{(1)}b(X_l)\,\middle|\,X_j=x\right\}f_j(x). \qquad (11)$$

The proof of (10) is sketched in the Appendix. The following corollary is an immediate result of Theorem 1.

COROLLARY 1. (i) *To first order, the variance of* $\hat{\theta}(x)$ *is*

$$(nh)^{-1}\gamma_0(K)W_2^{-2}(x)\left[\sum_{j=1}^{m} E\{(\mu_j^{(1)})^2\eta_{jj}|X_j=x\}f_j(x)\right], \tag{12}$$

where η_{jk} *is the* (j, k)*th element of* $V^{-1}\Sigma V^{-1}$.

(ii) *The variance in* (12) *is minimised when* $V=\Sigma$ *is correctly specified. It can then be simplified to*

$$(nh)^{-1}\gamma_0(K)\left[\sum_{j=1}^{m} E\{(\mu_j^{(1)})^2\sigma^{jj}|X_j=x\}f_j(x)\right]^{-1}, \tag{13}$$

where σ^{jj} *is the* (j, j)*th element of* Σ^{-1}.

(iii) *For any covariance structure* Σ, *the variance in* (13) *is always less than or equal to that in* (3); *that is, if we use the asymptotic variance as the criterion, this estimator uniformly outperforms the working independence estimator.*

Result (i) above can be easily obtained by standard arguments. Result (ii) is a direct consequence of an extended Cauchy–Schwartz inequality given in Johnson & Wichern (1982, § 2.7). A sketch proof for (iii) is given in the Appendix.

Remark 1. The asymptotic bias for the working independence estimator is $c_2(K)b(x)h^2$, where $b(x)=\theta^{(2)}(x)/2$. When $h_s=h$, equation (11) indicates that this asymptotic bias could differ from that of the proposed estimator. The difference between the two asymptotic biases is a function of $\mu^{(1)}(.)$, $\theta^{(2)}(.)$, the joint distribution of (X_{ij}, X_{il}) and the assumed covariance structure through v^{jl}. The structure of this difference is thus complex. Numerically, its value could be positive or negative; that is, when we are considering asymptotic biases with $h_s=h$, neither estimator outperforms the other uniformly. In general, a simple way of comparing the two theoretical biases is to let $h_s=o(h)$. In this situation, the two theoretical asymptotic biases are the same up to first order.

Remark 2. The first term in (10) determines the asymptotic variation of $\hat{\theta}$. Note that this term does not depend on the D_{ij} of the asymptotic expression of $\check{\theta}$ in (8); that is, the one-step update version of $\hat{\theta}$, besides its computational convenience, achieves the same variation as the fully iterated version.

The following corollary provides the first-order asymptotic results for the fully iterated estimator. In the linear case, the fully iterated estimator has a closed-form expression, which can be used to show the convergence of the iterative procedure.

COROLLARY 2. *Upon convergence, the limit obtained from fully iterated algorithm has, up to first order, the asymptotic variance given in* (12). *Therefore, properties* (ii) *and* (iii) *in Corollary* 1 *still hold. If we use the same bandwidth h throughout, the asymptotic bias of* $\hat{\theta}$ *with a full iteration is* $c_2(K)\mathscr{B}^*(x)$, *where* $\mathscr{B}^*(x)$ *satisfies the integral equation*

$$\mathscr{B}^*(x)=\frac{h^2}{2}\theta^{(2)}(x)-W_2^{-1}(x)\sum_{j=1}^{m} E\left\{\sum_{l\neq j}\mu_j^{(1)}v^{jl}\mu_l^{(1)}\mathscr{B}^*(X_l)\middle|X_j=x\right\}f_j(x). \tag{14}$$

Equation (14) was obtained by direct derivation after replacing $\check{\theta}$ in (4) by $\hat{\theta}$.

4. Numerical results

In this section, we use a simple simulation study to illustrate the numerical performance of the proposed estimator. Comparisons with the working independence estimator are also investigated. The scenario we consider is as follows. The continuous covariate, X_{ij}, was generated for each simulation from a uniform distribution on the interval $[-2, 2]$. The response, Y_{ij} has a conditional mean, $E(Y_{ij}|X_{ij}) = \sin(2X_{ij})$, a common variance, 1, and an exchangeable correlation structure with correlation ρ. We let $n=100$ and $m=3$. Two values of ρ, 0·4 and 0·6, were considered. They represent a mild and a somewhat stronger correlation, respectively.

Three estimators were evaluated, the working independence estimator, the one step update version of the proposed estimator and the estimator from the fully iterated algorithm. Five hundred datasets were generated. Since the X vary from dataset to dataset, estimates were obtained at 300 fixed equally-spaced grid points within the range of X throughout the simulation. Numerical properties of these estimates were then investigated.

We chose K to be the Epanechnikov kernel (Fan & Gijbels, 1996, § 2.2). A simple bandwidth selector was adopted. To be specific, all calculation rules of the empirical bias bandwidth selector of Ruppert (1997) were followed except that the bias was still calculated using the asymptotic formula with $\theta^{(2)}$ estimated by the corresponding derivative of a local fourth-order polynomial. A pilot bandwidth equal to half of the range of X was used here to mimic Fan & Gijbels' suggestion of using a global fit to obtain $\theta^{(2)}$; see Fan & Gijbels (1996, § 4.2). This slight modification allowed the use of the same estimation program to estimate $\theta^{(2)}$. Even though the estimator loses the major advantage of Ruppert's method, namely that it does not require an asymptotic bias structure, since the asymptotic bias formula is readily available the current approach is much faster than the empirical bias bandwidth selection. Little difference was found when using a pilot bandwidth varying from half the range of X to the full range, as is to be expected.

In the simulation study, the bias, variance and mean squared error for each estimator were obtained at each of the 300 fixed points. Bias, variance and mean squared error ratios were calculated, with numerators obtained from the working independence estimator and denominators obtained from the proposed estimator. A ratio above 1 indicates that the proposed estimator outperforms the working independence estimator. Table 1 summarises these ratios. For each type of ratio, the sample mean and sample quartiles are provided. It is evident that the improvement achieved by the proposed estimator increases as the within-subject correlation increases. It is also clear that the one-step update estimator behaves almost as well as the fully iterated version.

As a comparison, the ratio of (3) over (13) gives the value 1·296 for $\rho = 0{\cdot}4$ and 1·818 for $\rho = 0{\cdot}6$. Even though the theoretical result indicates that the ratio of (3) over (13) is the same for all locations, in reality we still observe different Monte Carlo variance ratios at different locations. For $\rho = 0{\cdot}6$, the numerical values of the variance ratios are not as high as indicated by the asymptotic result. Both proposed estimators exhibit much smaller variation compared to the working independence estimator.

Most simulation studies compare the performances of two estimators through averages over all simulated datasets. Here, we calculate an index R for comparing the proposed and traditional estimators for each dataset. The sums of squared deviations between the estimated and the true f values at the 300 fixed points are obtained for all estimators. We then define R to be the ratio of the two sums of squared deviations, with the numerator again calculated using the working independence estimator, and the denominator using

Table 1. *Summary results of the Monte Carlo bias, variance and mean squared error ratios at* 300 *fixed points. Each ratio was calculated with the numerator given by the working independence estimator and the denominator given by one of the two proposed estimators as indicated in the first column. The biases, variances and mean squared errors at the* 300 *locations were based on* 500 *simulated datasets. Entries reading from left to right are for the sample mean and sample quartiles obtained from the ratios.*

		$\rho=0{\cdot}4$				$\rho=0{\cdot}6$			
Estimator		Mean	Q1	Median	Q3	Mean	Q1	Median	Q3
One-step	Bias	1·116	1·099	1·118	1·140	1·230	1·203	1·231	1·257
	Variance	1·245	1·180	1·228	1·292	1·549	1·449	1·510	1·633
	MSE	1·235	1·174	1·221	1·280	1·513	1·414	1·497	1·588
Full iteration	Bias	1·117	1·101	1·120	1·139	1·232	1·204	1·231	1·261
	Variance	1·248	1·186	1·241	1·301	1·556	1·453	1·521	1·645
	MSE	1·237	1·178	1·231	1·288	1·518	1·419	1·507	1·609

MSE, mean squared error; Q1, lower quartile; Q3, upper quartile.

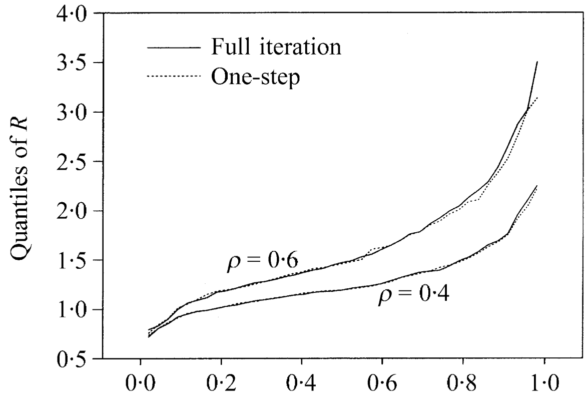

Fig. 1. Quantile plot of R for the one-step and the fully iterated estimators versus the working independence among the 500 simulated datasets.

the new estimator. The quantile plots of R for $\rho=0{\cdot}4$ and $\rho=0{\cdot}6$ are given in Fig. 1. The solid and dashed lines correspond to the quantile curves with the new estimator being the fully iterated and the one-step update estimators respectively. As in Table 1, an R larger than 1 indicates that the proposed estimator has a smaller sum of squared errors than the working independence estimator. Figure 1 clearly shows that the proposed estimators are superior to the working independence estimator for almost all 500 datasets. Furthermore, by the closeness of the solid and dashed curves, we observe that the one-step update estimator behaves approximately as well as the fully iterated version.

5. Discussion

We have concentrated on the scenario in which the numbers of within-subject observations, m_i, are finite. When the m_i increase with n, additional assumptions are required before the proposed estimator is bounded. Properties of the proposed estimator and perhaps also of the working independence estimator are not well understood in that scenario.

We have used the asymptotic variance as the criterion of comparison. We have not used asymptotic mean squared error mainly because of the complex format of the asymptotic biases. In view of the structure of the semiparametric efficient bound (Lin & Carroll, 2001), we expect the bound to be reached by an extension of the proposed method. This will be reported elsewhere.

Acknowledgement

This research was supported by a grant from the US National Cancer Institute and a Texas Advanced Research Program. The author thanks the editor and the associate editor, whose helpful comments greatly improved the paper. She also thanks Ray Carroll, Xihong Lin, Chuck McCulloch, Jamie Robins and Jane-Ling Wang for helpful discussions and Jim Calvin and Kathy Prewitt for suggestions about the presentation.

Appendix

Proofs

Derivation of equation (10). Define $\alpha_l = h^l \theta^{(l)}(t)/l!$ and suppose that $h \to 0$ and $nh \to \infty$. Standard derivations for local polynomial regression lead to

$$\hat{\theta}(x) - \theta(x) = \{W_2(x)\}^{-1}(A_{1n} + A_{2n} + A_{3n})\{1 + o_p(1)\},$$

where $W_2(x)$ is defined in (9) and

$$A_{1n} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(x - X_{ij}) \mu_{ij}^{(1)} v_i^{jj} [Y_{ij} - \mu\{\alpha_0 + \alpha_1(x - X_{ij})/h\}],$$

$$A_{2n} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(x - X_{ij}) \mu_{ij}^{(1)} \left\{ \sum_{l \neq j} v_i^{jl}(Y_{il} - \mu_{il}) \right\},$$

$$A_{3n} = -n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(x - X_{ij}) \mu_{ij}^{(1)} \left[\sum_{l \neq j} v_i^{jl} \mu_{il}^{(1)} \{\check{\theta}(X_{il}) - \theta(X_{il})\} \right]. \tag{A1}$$

Let $f_{jl}(t, s)$ be the joint density of (X_j, X_l) evaluated at (t, s) and define

$$Q(t, s) = \sum_{l} \sum_{j \neq l} f_{jl}(t, s) E[\mu_j^{(1)} v^{jl} \mu_l^{(1)} \{W_2(X_l)\}^{-1} | X_j = t, X_l = s].$$

By (8), we can rewrite $A_{3n} = (A_{31n} + A_{32n})\{1 + o_p(1)\}$, where

$$A_{31n} = -n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(x - X_{ij}) \mu_{ij}^{(1)} \left\{ \sum_{l \neq j} v_i^{jl} \mu_{il}^{(1)} W_2^{-1}(X_{il}) \sum_{i'=1}^{n} \sum_{l'=1}^{m} \mu_{i'l'}^{(1)} v_{i'}^{l'l'} K_{h_s}(X_{i'l'} - X_{il}) D_{i'l'} \right\}, \tag{A2}$$

$$A_{32n} = -(h^2/2) n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(x - X_{ij}) \mu_{ij}^{(1)} \left\{ \sum_{l \neq j} v_i^{jl} \mu_{il}^{(1)} b(T_{il}) \right\}.$$

It can be shown that the right-hand side of (A2) equals

$$-n^{-1} \sum_{i'=1}^{n} \sum_{l'=1}^{m} \mu_{i'l'}^{(1)} v_{i'}^{l'l'} \left[n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{j \neq l} \mu_{ij}^{(1)} v_i^{jl} \mu_{il}^{(1)} \{W_2(X_{il})\}^{-1} K_h(x - X_{ij}) K_{h_s}(X_{i'l'} - X_{il}) \right] D_{i'l'}$$

$$= -\left\{ n^{-1} \sum_{i'=1}^{n} \sum_{l'=1}^{m} Q(t, T_{i'l'}) D_{i'l'} \right\} \{1 + o_p(1)\}.$$

With $h \to 0$, $h_s \to 0$ and $nhh_s \to \infty$, $Q(t, s)$ is of order 1, and thus $A_{31n} = O_p(n^{-1/2}) = o_p\{(nh)^{-1/2}\}$. It is then straightforward to obtain (10), where $c_2(K)\mathscr{B}(x)$, with $\mathscr{B}(x)$ defined in (11), contains biases resulting from A_{1n} and A_{31n}.

Proof that formula (13) *is less than or equal to formula* (3). Since $\gamma_0(K)$, $f_j(x)$ and $\{\mu_j^{(1)}\}^2$ are nonnegative, we only need to show that $\sigma^{jj} \geqslant 1/\sigma_{jj}$. Recall that σ^{jj} is the (j, j)th element of Σ^{-1}, σ_{jj} is the (j, j)th element of Σ, and Σ is positive definite. Without loss of generality, we only need to consider $j = 1$. If we write

$$\Sigma = \begin{bmatrix} \sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^{\mathrm{T}} & \Sigma_{22} \end{bmatrix},$$

then

$$\sigma^{11} = (\sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^{\mathrm{T}})^{-1}.$$

The result holds because Σ_{22} is positive definite.

References

Carroll, R. J., Fan, J., Gijbels, I. & Wand, M. P. (1997). Generalized partially linear single-index models. *J. Am. Statist. Assoc.* **92**, 477–89.

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

Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, Y. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–22.

Johnson, R. A. & Wichern, D. W. (1982). *Applied Multivariate Statistical Analysis.* New York: Prentice and Hall.

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

Lin, D. Y. & Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data (with Discussion). *J. Am. Statist. Assoc.* **96**, 103–26.

Lin, X. & Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Am. Statist. Assoc.* **95**, 520–34.

Lin, X. & Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *J. Am. Statist. Assoc.* **96**, 1045–56.

Ruckstuhl, A. F., Welsh, A. H. & Carroll, R. J. (2000). Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statist. Sinica* **10**, 51–71.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Am. Statist. Assoc.* **92**, 1049–62.

Ruppert, D. & Wand, M. P. (1994). Multivariate weighted least squares regression. *Ann. Statist.* **22**, 1346–70.

Severini, T. A. & Staniswalis, J. G. (1994). Quasilikelihood estimation in semiparametric models. *J. Am. Statist. Assoc.* **89**, 501–11.

Zeger, S. L. & Diggle, P. J. (1994). Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689–99.

Zeger, S. L. & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–30.

[*Received August* 2001. *Revised May* 2002]