Statistica Sinica **7**(1997), 1171-1180

A VARIATION ON LOCAL LINEAR REGRESSION

M. C. Jones

Open University

Abstract: There has been much justifiable recent interest in local polynomial regression, and in particular in its local linear special case. Local linear regression has advantages in terms of desirable theoretical properties both in the interior and near the boundaries of the region of interest. For implementation, binning is useful. In this paper, we describe a variation on local linear regression which can be considered an alternative binning thereof. We show that existing and novel methods are almost indistinguishable. The point of the paper is not to extol the virtues of the new version over the old, but rather (i) to show that the good properties of local linear regression can be achieved in more than one way, and (ii) to elucidate close links between local linear regression and other kernel smoothing methods. The latter include, most closely, a boundary corrected 'naive' kernel estimator and a recent proposal of Wu and Chu (1992), as well as binned Nadaraya–Watson estimators and methods for binomial regression.

Key words and phrases: Binning, boundary correction, kernel smoothing, nonparametric regression.

1. Introduction

Local linear regression (see e.g. Hastie and Loader (1993), Wand and Jones (1995), Chapter 5) has various advantages over other kernel–based methods for nonparametric regression. Assume that the regression mean m has two continuous derivatives and that a symmetric probability density function is used as kernel K. In the interior of the design space, local linear regression achieves a desirable asymptotic bias which is proportional to m''. It does so while retaining the variance expected of most kernel-type estimators (but not attained by all) (Fan (1992, 1993)). That the combination of desirable bias and variance is not easy to achieve by direct kernel means is clear from Jones, Davies and Park (1994). Furthermore, local linear regression behaves very well near the boundaries of the design space in the sense that it retains $O(h^2)$ bias everywhere, as its single smoothing parameter, or bandwidth, $h \to 0$ (Fan (1992), Ruppert and Wand (1994)).

For fast implementation, binning algorithms have been developed; these are most thoroughly investigated by Fan and Marron (1994). The binning necessitates specification of a second smoothing parameter, the binwidth, or equivalently

of the number of bins. But it is clear from experience that the precise value taken for the binwidth is not crucial. The binned local linear method retains the good properties of the unbinned version.

This paper contributes to our understanding of the processes involved in achieving desirable performance by showing how a variation on binned local linear regression has the same desirable properties. The novel method is developed in Section 2 and its properties described in Section 3. This new method has close links to "naive" kernel regression (Section 2), and further links between these and other existing methods are given in Section 4. A complementary paper is mentioned briefly in Section 5.

In simulations, existing and novel methods can be seen to be virtually indistinguishable in practice; we save space by not illustrating this here.

2. The Modifications

We follow the notation of Fan and Marron (1994). The data are $\{(X_i, Y_i) : i = 1, \ldots, n\}$. We specify an equally spaced grid $x_1, \ldots, x_g$, defining the binwidth $\Delta = x_j - x_{j-1}$ for any $2 \leq j \leq g$. Let $I_j \equiv \{i : \ X_i \to x_j\}$ where the arrow denotes replacing X_i by it *nearest* gridpoint. (Variations on this are possible, too, but will not be considered here.) The binned data are $\{(x_j, Y_j^{\Sigma}, c_j) : \ j = 1, \ldots, g\}$ where the "bin counts" are $c_j \equiv \sum I(i \in I_j)$, and the "bin totals" $Y_j^{\Sigma} \equiv \sum Y_i I(i \in I_j)$ are the sums of the responses in the bins.

For any $l = 0, 1, \ldots,$ define

$$\bar{T}_l(x) \equiv \sum_{j=1}^{g} K_h(x - x_j)(x - x_j)^l Y_j^{\Sigma} \text{ and } \bar{S}_l(x) \equiv \sum_{j=1}^{g} K_h(x - x_j)(x - x_j)^l c_j.$$

Here, $K_h(u) \equiv h^{-1}K(h^{-1}u)$. The quantities $\bar{T}_l(x)$ and $\bar{S}_l(x)$ are binned approximations to

$$T_l(x) \equiv \sum_{i=1}^{n} K_h(x - X_i)(x - X_i)^l Y_i \text{ and } S_l(x) \equiv \sum_{i=1}^{n} K_h(x - X_i)(x - X_i)^l,$$

respectively. The usual binned local linear regression estimator (Fan and Marron (1994)) is

$$\hat{m}_{\ell}^{B}(x) \equiv \frac{\bar{S}_2(x)\bar{T}_0(x) - \bar{S}_1(x)\bar{T}_1(x)}{\bar{S}_2(x)\bar{S}_0(x) - \bar{S}_1^2(x)}, \tag{2.1}$$

and this approximates the actual local linear regression estimator (Fan (1992)):

$$\hat{m}_{\ell}(x) \equiv \frac{S_2(x)T_0(x) - S_1(x)T_1(x)}{S_2(x)S_0(x) - S_1^2(x)}.$$

Our alternative binning can be introduced just as simply as follows. Whenever $c_j \neq 0$, define the "bin averages" $\bar{Y}_j \equiv c_j^{-1} Y_j^{\Sigma}$. Now replace smooths of Y_j^{Σ} and c_j in $\hat{m}_\ell^B$ by smooths of $\bar{Y}_j I(c_j > 0)$ and $I(c_j > 0)$, respectively. That is, set

$$\tau_l(x) = \sum_{j=1}^{g} K_h(x - x_j)(x - x_j)^l \bar{Y}_j I(c_j > 0)$$

and

$$\sigma_l(x) \equiv \sum_{j=1}^{g} K_h(x - x_j)(x - x_j)^l I(c_j > 0).$$

Then define

$$\tilde{m}_\ell^B(x) \equiv \frac{\sigma_2(x)\tau_0(x) - \sigma_1(x)\tau_1(x)}{\sigma_2(x)\sigma_0(x) - \sigma_1^2(x)}. \tag{2.2}$$

The estimator $\tilde{m}_\ell^B$ is our proposal to compete with $\hat{m}_\ell^B$, and we shall see in Sections 3 and 4 that the two barely differ at all, in theory or in practice.

The following alternative development provides further motivation for $\tilde{m}_\ell^B$ and makes a first link to alternative kernel regression approaches. Each bin average $\bar{Y}_j$ estimates $m(x)$ for x in bin j, of course. Because of the uniform nature of the grid, we introduce no complications (in the form of a non–uniform "design density" f) if we simply kernel smooth the set $\{(x_j, \bar{Y}_j I(c_j > 0), I(c_j > 0)) : j = 1, \ldots, g\}$ i.e. consider using $\check{m}^B(x) = \Delta\ \tau_0(x)$. The estimator $\check{m}^B$ is the natural binned version of the basic "naive" kernel regression estimator $\check{m}(x) = n^{-1} T_0(x)$ (e.g. Priestley and Chao (1972); something very closely related to $\check{m}$ has recently been proposed by Chu (1993)). We will see in Section 4.3 that $\check{m}^B$ is closely related to a proposal of Wu and Chu (1992). We repeat that consideration of $\check{m}^B(x)$ as an estimator of $m(x)$ is justified by the uniformity of the gridpoints which avoids inadequacies of $\check{m}(x)$ due to nonuniform f (Jones, Davies and Park (1994)).

It seems, therefore, as though binning might allow us to revert to naive smoothing, and this is true in the interior of the design space; but there remains the other inadequacy of this approach which is its failure to cope with boundaries. We need to introduce some form of boundary kernel (e.g. Müller (1991), Jones (1993)). Suppose that there is a boundary at, without loss of generality, 0. One appropriate choice of boundary kernel is to replace $K(u)$, at $x = ph$ say, by the linear multiple

$$K_L(u) = \{a_0(p)a_2(p) - a_1^2(p)\}^{-1}\{a_2(p) - a_1(p)u\}K(u),$$

where $a_l(p) = \int_{-\infty}^{p} z^l K(z)dz$. This boundary kernel can variously be motivated as being asymptotically equivalent to that implicitly used by $\hat{m}_\ell$ (Ruppert and

Wand (1994)), as a sensible *ad hoc* device (e.g. Hart and Wehrly (1992)) and via generalised jackknifing (Jones (1993)). It seems to be as good a choice as other boundary kernels (Jones (1993)). Note that as we move away from the boundary, K_L tends to K.

Now, unless there is no data near x, $\sigma_l(x)$ is a Monte Carlo or quadrature approximation (depending on the genesis of $X_1, \ldots, X_n$) to $h^l a_l(p)$. (The $\sigma_l(x)$ formula in fact works near any boundary — and we might well be working on [0,1] — since it targets $\int_{C_f} K_h(x-u)(x-u)^l du$ where C_f is the support of f.) That is, $K_L(u)$ can be approximated by

$$\{\sigma_0(x)\sigma_2(x) - \sigma_1^2(x)\}^{-1}\{\sigma_2(x) - \sigma_1(x)hu\}K(u)$$

and replacement of K in $\check{m}^B(x)$ by this yields precisely $\tilde{m}_\ell^B(x)$ as in (2.2).

Here, we have built, via binning, on the well–known approximate equivalence (e.g. Ruppert and Wand (1994)) between local linear regression and the use of kernel regression (in fact, naive smoothing when the design is uniform) with boundary kernel K_L. Many readers will prefer the local linear motivation because it does such boundary correction automatically, without the user imposing it. (Now that the conceptual role of K_L has been clarified, K_L will not occur again in the remainder of the paper.) But the formulae used are essentially the same and of the same complexity in either approach, and the similarity of (2.1) and (2.2) reflects this.

3. Theoretical Properties

Asymptotic mean squared errors (AMSEs) of $\hat{m}_\ell^B(x)$ and $\tilde{m}_\ell^B(x)$, conditional on $X_1, \ldots, X_n$, turn out to be identical, so the two are dealt with in the following theorem. For proof of the theorem, see the Appendix.

Theorem. *Suppose that the regression model is* $Y_i = m(X_i) + \epsilon_i$ *where the* ϵ*s are independent of each other and of the* X*s, having mean zero and variance* $v^2(x_i)$. $X_1, \ldots, X_n$ *have design density* f *in either sense of being chosen randomly from* f *or of being placed* (*close to or*) *at the quantiles of* f. *Suppose* $f > 0$ *on* C_f. *Assume that* K *has jump discontinuities in its* $(2t-1)$*st derivative for some* $t \geq 1$ *and that* m *has* $\max\{2, 2t-1\}$ *continuous derivatives. To cope with a boundary at, without loss of generality,* 0, *let* $x = ph$. *Assume also, for simplicity, that* ph *is a gridpoint* (*otherwise, one should think in terms of interpolating between values at gridpoints in such a way as not to introduce further approximation error; this is unimportant in practice*). *Let* $n \to \infty$, $h = h(n) \to 0$ *such that* $nh \to \infty$, *and* $\Delta = \Delta(n) \to 0$ *such that* $\Delta/h \to 0$. *Then*

$$\mathrm{AMSE}\,(M_h(x)) \simeq \Big(\frac{1}{2}\frac{\{a_2^2(p) - a_1(p)a_3(p)\}}{\{a_0(p)a_2(p) - a_1^2(p)\}}m''(x)h^2 + O(\Delta^2 + (\Delta/h)^{2t})\Big)^2$$

$$+\frac{1}{nh}\Big[\frac{v^2(x)}{f(x)}\int_{-\infty}^{p}\frac{\{a_2(p)-ua_1(p)\}^2}{\{a_2(p)a_0(p)-a_1^2(p)\}^2}K^2(u)du+O(\Delta^2+(\Delta/h)^{2t})\Big],$$

where M_h denotes either $\hat{m}_\ell^B$ or $\tilde{m}_\ell^B$. If K has support $[-1,1]$, then for an interior point $x \geq h$, this reduces to

$$\begin{aligned}\mathrm{AMSE}(M_h(x)) \simeq &\Big(\frac{1}{2}s_2 m''(x)h^2+O(\Delta^2+(\Delta/h)^{2t})\Big)^2\\ &+\frac{1}{nh}\Big\{\frac{v^2(x)}{f(x)}\int_{-1}^{1}K^2(u)du+O(\Delta^2+(\Delta/h)^{2t})\Big\},\end{aligned}$$

where $s_2=\int_{-1}^{1}u^2K(u)du$.

The theorem shows that if Δ is small enough relative to h then the effect of binning on the AMSEs can be ignored. In this case, the theorem reduces to results of Fan (1992). The theorem is clearly novel for $\tilde{m}_\ell^B$ but does not seem to have been written down elsewhere for $\hat{m}_\ell^B$ either.

4. Relationships With Other Methods

4.1. Links with the Nadaraya-Watson estimator

It is worth commenting on various other kernel-based nonparametric regression estimators in the light of this work. Possibly the most popular method, at least until recently, was the Nadaraya-Watson estimator (e.g. Härdle (1990)) which in our notation is simply $\hat{m}_c(x)=T_0(x)/S_0(x)$. Its binned version, considered by Härdle and Scott (1992) and Fan and Marron (1994), is $\hat{m}_c^B(x)=\bar{T}_0(x)/\bar{S}_0(x)$. But, of course, we can now introduce an alternative binned implementation, namely $\tilde{m}_c^B(x)=\tau_0(x)/\sigma_0(x)$.

Theoretical properties follow straightforwardly from the proof of the theorem. Under the same conditions as in the theorem, it is not difficult to see that

$$\begin{aligned}&\mathrm{AMSE}(\hat{m}_c^B(x))\\ &\simeq\Big(-\frac{a_1(p)}{a_0(p)}m'(x)h+\Big[\frac{1}{2}\frac{a_2(p)}{a_0(p)}m''(x)+\frac{\{a_2(p)a_0(p)-a_1^2(p)\}}{a_0^2(p)}\frac{m'(x)f'(x)}{f(x)}\Big]h^2\\ &\quad+O(\Delta^2+(\Delta/h)^{2t})\Big)^2+\frac{1}{nh}\Big\{\frac{v^2(x)}{f(x)}\frac{\int_{-\infty}^{p}K^2(u)du}{a_0^2(p)}+O(\Delta^2+(\Delta/h)^{2t})\Big\}\end{aligned}$$

which reduces to

$$\begin{aligned}\mathrm{AMSE}(\hat{m}_c^B(x)) \simeq &\Big(\frac{1}{2}s_2\Big\{m''(x)+2\frac{m'(x)f'(x)}{f(x)}\Big\}h^2+O(\Delta^2+(\Delta/h)^{2t})\Big)^2\\ &+\frac{1}{nh}\Big\{\frac{v^2(x)}{f(x)}\int_{-1}^{1}K^2(u)du+O(\Delta^2+(\Delta/h)^{2t})\Big\}\end{aligned}$$

in the interior. (See also Härdle and Scott (1992), who cite a Diploma thesis of K. Breuer.) Aside from the binning terms, the latter is the usual expression for the AMSE of the Nadaraya-Watson estimator in the interior (e.g. Härdle (1990)), and the former is its analogue near the zero boundary.

For $\tilde{m}_c^B$, however, the asymptotics work out differently. We get

$$\begin{aligned}\mathrm{AMSE}(\tilde{m}_c^B(x)) \simeq \Big(-\frac{a_1(p)}{a_0(p)} m'(x)h + \frac{1}{2}\frac{a_2(p)}{a_0(p)} m''(x)h^2 + O(\Delta^2 + (\Delta/h)^{2t})\Big)^2 \\ + \frac{1}{nh}\Big\{ \frac{v^2(x)}{f(x)} \frac{\int_{-\infty}^{p} K^2(u)du}{a_0^2(p)} + O(\Delta^2 + (\Delta/h)^{2t})\Big\}\end{aligned}$$

giving the same interior formula as for the local linear estimators $\hat{m}_\ell^B$ and $\tilde{m}_\ell^B$ in the theorem, but leading boundary behaviour like that of $\hat{m}_c^B$.

The following behaviour is discernible in simulations but we shall not explicitly illustrate it. For small samples, $\tilde{m}_c^B$ stays close to its roots in $\hat{m}_c^B$. However, for quite large sample sizes, in the interior, three estimators, namely $\tilde{m}_\ell^B$, $\hat{m}_\ell^B$ and $\tilde{m}_c^B$, remain very close together and $\hat{m}_c^B$ is indeed somewhat apart. However, at the boundaries, $\tilde{m}_c^B$ reverts to the less desirable behaviour of $\hat{m}_c^B$. This well reflects the asymptotic results above.

An explanation of this surprising phenomenon is at hand. For n very small relative to g, Y_j^Σ is either 0 or $\bar{Y}_j$, and $\bar{T}_0$ and τ_0 follow T_0 in estimating $m \times f$. Also, $\bar{S}_0$ and $\sigma_0(x)$ follow S_0 as estimators of f. Thus $\tilde{m}_c^B$ behaves like $\hat{m}_c^B$, and also like $\hat{m}_c$. On the other hand, if n is very large relative to (large) g, τ_0 estimates m directly, and σ_0 becomes an estimate of unity. Estimators which target m directly (or, more exactly, through estimating $n^{-1}\sum_{i=1}^n K_h(x - X_i)Y_i/f(X_i)$), which include the local linear approaches as well as $\tilde{m}_c^B$, have quite different properties from those taking the "$\widehat{mf}/\hat{f}$" route (Jones, Davies and Park (1994)), and these are reflected above. The estimators $\hat{m}_c$ and $\hat{m}_c^B$ continue to work in the latter way when n is large also.

A general note is that one sometimes has to work quite hard to arrange that f is non–uniform enough in a steep enough region of m for the differences above to be considerable, and indeed for the differences between $\hat{m}_c$ and $\hat{m}_\ell$ to be of major qualitative importance (except at a boundary). The above also implies that boundary correction of $\tilde{m}_c^B$ is called for to align it with the better estimators. But this is exactly how $\tilde{m}_\ell^B$ comes about in any case.

4.2. Links with binomial regression

Suppose that *on an equispaced grid* $\{Z_j, j = 1, \ldots, g\}$, we have independent binomial data $B_j \sim \mathrm{Bin}(n_j, p_j)$. Kernel weighted local likelihood, in its local constant form (Staniswalis (1989)), estimates the assumed smooth function p at

x by maximising

$$\sum_{j=1}^{n} K_h(x - Z_j)\{B_j \log a + (n_j - B_j)\log(1-a)\}$$

over a. The result is precisely $\hat{m}_c^B$ with Z_j, B_j and n_j replacing x_j, Y_j^Σ and c_j, respectively. That is, for equispaced binomial regression — which may be of practical interest for example in bioassay problems on a log dose scale — local likelihood essentially suggests $\hat{m}_c^B$.

On the other hand, $\tilde{m}_c^B$ is essentially a kernel smooth of the actual maximum likelihood estimates B_j/n_j, the analogue of $\bar{Y}_j$. Our arguments earlier, therefore, indicate a preference for kernel smoothing a maximum likelihood estimate rather than kernel smoothing the log-likelihood function itself for this particular ratio estimation problem. As an associate editor says, an explanation may be that $\tilde{m}_c^B$ is more faithful to the observed likelihood than is $\hat{m}_c^B$. This may have analogues in other indirect curve estimation problems. Gavin, Haberman and Verrall (1994) discuss these two smooth binomial estimators in an actuarial setting, and also have a general preference for the latter.

For non-equispaced binomial regression data, one ought, again, to fit local lines (Tibshirani and Hastie (1987), Fan, Heckman and Wand (1995)). When doing so for normal errors, we observe that it is discretising the local likelihood (for fast computation purposes) that leads to $\hat{m}_\ell^B$, while we have to start with modified data to derive $\tilde{m}_\ell^B$.

4.3. Links with another kernel–type regression estimator

The method of Wu and Chu (1992) can be thought of as being very much in the spirit of the new proposal of this paper but in perhaps a slightly less desirable form. Wu and Chu's "double smoothing type estimator" (DSTE) (i) obtains binned data, but using a Nadaraya Watson weighted average (with bandwidth half the binwidth) of Ys in the bin, then (ii) uses essentially $\Delta\tau_0(x)$ with the Nadaraya-Watson smooths replacing the simple bin averages. The DSTE is clearly close to $\tilde{m}_\ell^B$ but it does not employ the boundary correction which fully lines $\tilde{m}_\ell^B$ up with local linear regression. Also, Wu and Chu (1992) state that the binning "is not related to the asymptotic performance" of DSTE but do not fully exploit this for practice, while the initial Nadaraya-Watson estimation seems unnecessarily complicated. The current paper delves more deeply into appropriately modified DSTE.

5. A Complementary Paper

On completion of the current paper, the author's attention was drawn to Kneip and Engel (1996), a paper written independently and at much the same

time. Kneip and Engel (1996) also have as main focus an estimator closely related to $\check{m}^B$: see their (3.3) (Kneip and Engel use Gasser-Müller weights (e.g. Müller (1988)); because of the binning, these weights will differ little from ours). But Kneip and Engel's work has a different, and complementary, emphasis; they are particularly concerned with using a coarser binning with the aim of alleviating the design sparsity difficulty investigated by Seifert and Gasser (1996).

Acknowledgements

Thanks are due to the many authors who provided me with prepublication copies of recently appeared or yet to be published papers in the reference list, and to the referees for prompting clarification in the presentation of this material.

Appendix: Proof of Theorem

Conditional on $X_1, \ldots, X_n$,

$$E(Y_j^{\Sigma}) = \sum_{i=1}^{n} m(X_i) I(i \in I_j) \simeq n \int_{I_j} m(x) f(x) dx = n\{\Delta m(x_j) f(x_j) + O(\Delta^3)\}$$

and

$$E(c_j) = c_j \simeq n \int_{I_j} f(x) dx = n\{\Delta f(x_j) + O(\Delta^3)\}.$$

It follows that $E(\bar{Y}_j) \simeq m(x_j) + O(\Delta^2)$. Let $*$ denote convolution. It then follows that

$$E(\bar{T}_l(x)) \simeq n\{h^l (x^l K)_h * (mf + O(\Delta^2)) + O((\Delta/h)^{2t})\},$$
$$\bar{S}_l(x) \simeq n\{h^l (x^l K)_h * (f + O(\Delta^2)) + O((\Delta/h)^{2t})\},$$

and, provided we now take n large enough that we can assume that $I(c_j > 0) = 1$,

$$E(\tau_l(x)) \simeq \Delta^{-1}\{h^l (x^l K)_h * (m + O(\Delta^2)) + O((\Delta/h)^{2t})\}$$

and

$$\sigma_l(x) \simeq \Delta^{-1}\{h^l (x^l K)_h * (1) + O((\Delta/h)^{2t})\}.$$

(See Hall and Wand (1996) for the genesis of the $O((\Delta/h)^{2t})$ terms.) Note that $(x^l K)_h * r$, say, can be expanded as $a_l(p) r(x) - h a_{l+1}(p) r'(x) + \frac{1}{2} h^2 a_{l+2}(p) r''(x)$. The biases of the theorem then follow by employing these expressions in the formulae for $\hat{m}_\ell^B$ and $\tilde{m}_\ell^B$.

The variance terms will also include remainders which result in the $O(\Delta^2) + O((\Delta/h)^{2t})$ terms in the theorem, but these will not be explicitly written out from now on for convenience. Let $\bar{A}_l(x)$ be defined like $\bar{S}_l(x)$ except for K being replaced by K^2. Then, conditional on $X_1, \ldots, X_n$,

$$\mathrm{Cov}(\bar{T}_l(x), \bar{T}_k(x)) \simeq \sum_{j=1}^{g} K_h^2 (x - x_j)(x - x_j)^{l+k} c_j v^2(x_j) \simeq v^2(x) \bar{A}_{l+k}(x).$$

Thus,

$$\begin{aligned}&\mathrm{Var}(\hat{m}_\ell^B(x))\\ &\simeq v^2(x)\frac{\bar{S}_2^2(x)\bar{A}_0(x)-2\bar{S}_2(x)\bar{S}_1(x)\bar{A}_1(x)+\bar{S}_1^2(x)\bar{A}_2(x)}{\{\bar{S}_2(x)\bar{S}_0(x)-\bar{S}_1^2(x)\}^2}\\ &\simeq (nh)^{-1}v^2(x)\int_{-\infty}^{p}K^2(u)\frac{a_2(p)^2f^3(x)-2a_2(p)a_1(p)f^3(x)u+u^2a_1^2(p)f^3(x)}{\{a_2(p)a_0(p)f^2(x)-a_1^2(p)f^2(x)\}^2}du\end{aligned}$$

by virtue of the above approximation to $\bar{S}_l(x)$, and one small final step gives the variance in the theorem. Likewise, conditional on $X_1,\ldots,X_n$,

$$\mathrm{Cov}\,(\tau_l(x),\tau_k(x))\simeq\sum_{j=1}^{g}K_h^2(x-x_j)(x-x_j)^{l+k}c_j^{-1}v^2(x_j)\simeq v^2(x)\bar{\alpha}_{l+k}(x).$$

Thus,

$$\mathrm{Var}(\tilde{m}_\ell^B(x))\simeq v^2(x)\frac{\sigma_2^2(x)\alpha_0(x)-2\sigma_2(x)\sigma_1(x)\alpha_1(x)+\sigma_1^2(x)\alpha_2(x)}{\{\sigma_2(x)\sigma_0(x)-\sigma_1^2(x)\}^2},$$

where

$$\alpha_l=\sum_{j=1}^{g}K_h^2(x-x_j)(x-x_j)^lc_j^{-1}\simeq h^l\int_{-\infty}^{p}u^lK^2(u)du/f(x),$$

and the result follows.

References

Chu, C. K. (1993). A new version of the Gasser-Mueller estimator. *J. Nonparametric Statist.* **3**, 187-193.

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

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196-216.

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

Fan, J., Heckman, N. and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90**, 141-150.

Fan, J. and Marron, J. S. (1994). Fast implementations of nonparametric curve estimators. *J. Comput. Graphical Statist.* **3**, 35-56.

Gavin, J., Haberman, S. and Verrall, R. (1994). On the choice of bandwidth for kernel graduation. *J. Inst. Actuar.* **121**, 119-34.

Hall, P. and Wand, M. P. (1996). On the accuracy of binned kernel density estimators. *J. Multivariate Anal.* **56**, 165-184.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.

Härdle, W. and Scott, D. W. (1992). Smoothing by weighted averaging of rounded points. *Comput. Statist.* **7**, 97-128.

Hart, J. D. and Wehrly, T. E. (1992). Kernel regression estimation when the boundary region is large, with an application to testing the adequacy of polynomial models. *J. Amer. Statist. Assoc.* **87**, 1018-1024.

Hastie, T. and Loader, C. (1993). Local regression: automatic kernel carpentry (with comments). *Statist. Sci.* **8**, 120-143.

Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statist. Comput.* **3**, 135-146.

Jones, M. C., Davies, S. J. and Park, B.-U. (1994). Versions of kernel-type regression estimators. *J. Amer. Statist. Assoc.* **89**, 825-832.

Kneip, A. and Engel, J. (1996). A remedy for kernel estimation under random design. *Statistics* **28**, 201-225.

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

Müller, H.-G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika* **78**, 521-530.

Priestley, M. B. and Chao, M. T. (1972). Non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B* **34**, 385-392.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346-1370.

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

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

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82**, 559-567.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* Chapman and Hall, London.

Wu, J. S. and Chu, C. K. (1992). Double smoothing for kernel estimators in nonparametric regression. *J. Nonparametric Statist.* **1**, 375-386.

Department of Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, U.K.
E-mail: m.c.jones@open.ac.uk

(Received April 1995; accepted February 1997)