Statistica Sinica **6**(1996), 347-366

ROBUST ESTIMATION OF SMOOTH REGRESSION AND SPREAD FUNCTIONS AND THEIR DERIVATIVES

A. H. Welsh

The Australian National University

Abstract. We consider the application of kernel weighted local polynomial regression methods to estimate regression and spread functions and their derivatives. In particular, we consider both an extension of the regression quantile methodology introduced by Koenker and Bassett (1978) and an approach based on M-estimation for heteroscedastic regression models. The present work is partly motivated by the paper of Ruppert and Wand (1994) who show that, by analysing local polynomial fitting directly as a weighted regression method rather than as an approximate kernel smooth, asymptotic results for estimating the regression function can be obtained for complex problems including vector covariates, general polynomials, derivative estimation and boundary problems. We extend their results to allow for robust fitting, for modelling general heteroscedasticity and for derivative estimation in the multivariate case. Our results confirm that local polynomial fitting procedures produce robust estimators of the regression and spread functions and their derivatives. Moreover, it is shown that we can reduce the bias of the estimators by increasing the order of the polynomials being fitted. The excellent edge-effect behaviour of local polynomial methods extends to derivative estimation and the multivariate case. We apply the methodology to two data sets to illustrate its practical utility.

Key words and phrases: Boundary effects, kernel function, local regression, M-estimation, nonparametric regression, regression quantiles, weighted regression.

1. Introduction

A common problem in data analysis is to describe the conditional distribution of Y given X on the basis of n independent observations $(Y_1, X_1), \ldots, (Y_n, X_n)$ on (Y, X). The interesting structure in this distribution, at least at a gross level, is often captured by the location and spread of the conditional distribution. In such cases, it is natural to adopt the model

$$Y_i = m(X_i) + s(X_i)e_i, \quad 1 \leq i \leq n, \tag{1.1}$$

where m and $s > 0$ are smooth location (or regression) and spread functions respectively and $\{e_i\}$ is a sequence of independent random variables which are

independent of $\{X_i\}$. This model entails

$$\begin{aligned} \text{location}(Y_i|X_i) &= m(X_i) + s(X_i)\text{location}(e), \\ \text{spread}(Y_i|X_i) &= s(X_i)\text{spread}(e), \quad 1 \leq i \leq n, \end{aligned}$$

and so is identifiable when we specify a location and scale functional on the e_i for which $\text{location}(e_i) = 0$ and $\text{spread}(e_i) = 1$. If we are prepared to identify s only up to a multiplicative constant, we do not need to specify a spread functional. This is of course part of the motivation for the common practice of treating $\log\{s(x)\}$ rather than $s(x)$ as the parameter of interest.

Exploratory analysis based on the model (1.1) requires at least estimation of the regression and spread functions m and s. However, in particular problems, we may also be interested in the derivatives of these functions. For example, in fitting growth curves, Müller (1988) has argued that the first two derivatives of the regression function m are of interest and for fitting generalised linear models (see for example McCullagh and Nelder (1990)) without assuming that the link or variance function is known as in Weisberg and Welsh (1995), m, m' and $v = s^2$ are required for an algorithm based on the method of scoring while m'' and v' are additionally required for a Newton-Raphson algorithm. Derivatives of m and s are also nuisance parameters in the substitution or "plug-in" method of optimal bandwidth estimation. Thus, we need to be able to estimate the regression and spread functions and their derivatives. It is important that we produce good estimates on the boundary of the support of X and that X can be allowed to be a vector of general dimension d. Finally, it is important, particularly in exploratory work, that all of these requirements be achieved robustly in the sense that the estimators are insensitive to at least large $\{e_i\}$.

There is an extensive literature on a variety of different approaches to most of these problems though they are typically treated as distinct. See for example the books by Eubank (1988), Müller (1988), Härdle (1990) and Wahba (1990). Our purpose in this paper is to present a unified approach to achieving all the above requirements simultaneously by using local polynomial fitting. Recent papers on local polynomial fitting include Stone (1977, 1982), Cleveland (1979), Tsybakov (1986), Müller (1987), Cleveland and Devlin (1988), Fan (1990, 1992, 1993), Chaudhuri (1991), Ruppert and Wand (1994), Fan, Hu and Truong (1992) and Hastie and Loader (1993). Hastie and Loader (1993) gave an excellent review of local polynomial smoothing and discuss its advantages over traditional kernel smoothing methodology. Local polynomial fitting involves the fitting of approximating polynomial models by weighted regression methods, where suitable weights are defined through a kernel function K and a bandwidth $h > 0$. The fitted intercept of the approximating polynomial is the estimator of the func-

tion and the fitted slope coefficients contain information about the derivatives of the function.

The present work is motivated by the paper of Ruppert and Wand (1994) who show that by analysing local polynomial fitting directly as a weighted regression method rather than as an approximate kernel smooth, asymptotic results for estimating m can be obtained for complex problems including vector X, general polynomials, derivative estimation and boundary problems. We extend their results to allow for robust fitting, for modelling general heteroscedasticity and for derivative estimation in the multivariate case. Tsybakov (1986) and Fan, Hu and Truong (1992) have considered local robust fitting of linear polynomials when X is scalar but do not treat higher polynomial fitting or derivative estimation. Moreover, Tsybakov treats the errors as homoscedastic and both articles ignore the scale. Even if the errors are homoscedastic (i.e. $s(x)$ is constant, which probably should not be assumed in an initial exploratory analysis) this means that their estimators are not scale equivariant unless the scale is known or can be factored out of the estimation problem; this greatly restricts the applicability of their results.

In this paper, we consider two approaches to fitting the model (1.1). The details of the approaches are given in Section 2. Both methods are based on making methods of fitting polynomial models local by incorporating kernel weights into the estimating equations. Specifically, we consider local regression quantile fitting and local M-estimation by incorporating kernel weights into the regression quantile procedure of Koenker and Bassett (1978) and the procedure for M-estimation of heteroscedastic regression models.

Running regression quantile estimators were considered by Chaudhuri (1991) who obtained similar results to those in Theorem 3.1 below. However, Chaudhuri treated $s(x)$ as constant so the errors are homoscedastic and did not consider the spread problem. Fan, Hu and Truong (1992) considered only linear polynomials and ignored both the spread and derivative problems. In contrast, our approach enables us to estimate the spread and the derivatives of both the running regression quantile functions and the spread function.

The quantile approach requires the distribution function F to be smooth so it is not generally applicable in the generalised linear model framework. We, therefore, consider the general class of local M-estimators applied to an approximating heteroscedastic regression model which pertain whether F is smooth or not and allow the possibility of increased efficiency in estimation.

We introduce some further notation, the required assumptions and explicit definitions of the procedures proposed in Section 2. We then derive the conditional asymptotic biases and variances of the running regression quantile and heteroscedastic M-estimators in Section 3. These results confirm that our local

polynomial fitting procedures produce estimators of the regression and spread functions and their derivatives and that they can be chosen to achieve robustness. Moreover, provided m and s are smooth enough, we can reduce the bias of the estimators by increasing the order of the polynomials being fitted. We also discuss the implications of our results, window estimation, the performance of the estimators on the boundary of the support of X and the extension of the results to the case of vector X. One obvious advantage of the local polynomial approach is that when $d = 1$ a single window can be used for all the estimators though this does not simultaneously optimise the mean squared error of all the estimators. As shown by Fan (1990, 1992), Ruppert and Wand (1994), and Fan, Hu and Truong (1992), local polynomial methods have excellent edge-effect behaviour and as pointed out by Ruppert and Wand (1994), this extends to the multivariate case and is also true for derivative estimation. Finally, in Section 4, we present three applications which illustrate the practical utility of the methodology.

2. Robust Local Polynomial Procedures

To simplify the presentation, we adopt the model (1.1) throughout and make the additional assumption that the $\{e_i\}$ are identically distributed with common distribution function F. The extension to the important case of non-identical distributions is straightforward but complicates matters without adding anything conceptually. In any case, if procedures do not work in simple models, they are unlikely to work in more complicated ones; so model (1.1) can be regarded as a baseline model in which to evaluate our procedures.

We assume throughout that model (1.1) holds and in addition that, for some nonnegative integers p and q,

(i) $\{X_i\}$ are independent and identically distributed with density function g, $g(x) > 0$ and g' is continuous in a neighbourhood of x;
(ii) $m^{(p+2)}(x)$ and $s^{(q+2)}(x)$ are continuous in a neighbourhood of x;
(iii) K is bounded, symmetric, has compact support and satisfies $\int K(u)du = 1$; and
(iv) N_p and N_q are non-singular, where N_p denotes the $(p+1) \times (p+1)$ matrix with (i,j)th element μ_{i+j-2} and $\mu_j = \int u^j K(u)du$.

Condition (iii) on K is adopted to simplify the proofs of the theoretical results; it can be relaxed to allow kernels with noncompact support.

The statements of our results involve the $(p+1) \times (p+1)$ matrix T_p with (i,j)th element $\int u^{i+j-2}K(u)^2du$. Note that this matrix can be obtained explicitly as soon as K is specified. The assumption that K is symmetric gives N_p a checker-board pattern which is preserved when N_p is inverted but the squaring of

K in the elements of T_p means that it is not similarly patterned. Also, to identify particular components in our vector and matrix expressions, let k_r denote the $(p+1)$-vector with a one in the rth position and zeros elsewhere, $1 \le r \le p+1$.

The first approach we consider involves the use of a local regression quantile fit.

- **Running regression quantiles**

Let $\hat{\beta}(\alpha) \in R^{p+1}$ denote the minimum of

$$\sum_{i=1}^{n} \rho_\alpha\{Y_i - z_i(x)^T\beta\}\ K\{(X_i - x)/h\}, \quad h > 0,$$

where $z_i(x) = (1, X_i - x, (X_i - x)^2, \ldots, (X_i - x)^p)^T$, $1 \le i \le n$, $p \ge 0$, and $\rho_\alpha(x) = x\{\alpha - I(x < 0)\}$. Then set

$$\hat{m}^{(r)}(x) = r!\hat{\beta}_{r+1}(1/2), \quad 0 \le r \le p,$$

and

$$\mathrm{IQR}^{(r)}(x) = r!\{\hat{\beta}_{r+1}(3/4) - \hat{\beta}_{r+1}(1/4)\}, \quad 0 \le r \le p.$$

The running regression quantile lines $\hat{\beta}_1(\alpha)$, $0 < \alpha < 1$, which are estimating the conditional quantiles $m(x)+s(x)F^{-1}(\alpha)$, are useful in their own right for studying the conditional distribution of Y given X and can be used to explore whether the reduction to consideration of regression and spread in (1.1) is sensible in any given problem. Here $\mathrm{IQR}^{(r)}(x)$ is estimating $\{F^{-1}(3/4) - F^{-1}(1/4)\}s^{(r)}(x)$, the interquartile range of F times the rth derivative of of s; so we may choose (a) to ignore the constant $F^{-1}(3/4) - F^{-1}(1/4)$, (b) estimate the interquartile range of F and renormalise or (c) work on the log scale.

Our results for running regression quantiles depend on a vector $\gamma(\alpha)$ which is defined to be a vector for which $\hat{\beta}(\alpha)$ is asymptotically unbiased so that the asymptotic variance of $\hat{\beta}(\alpha)$ about $\gamma(\alpha)$ can be obtained by standard arguments. The leading bias term is obtained from an asymptotic expansion of $\gamma(\alpha)$ about the vector of derivatives of m and s. This part of the argument requires some attention to detail because the pattern in the N_p and Q_p matrices results in very different expressions for the bias of the $(r+1)$th component depending on whether $p - r$ (or $q - r$) is odd or even.

The function ρ_α in the definition of the running regression quantiles is not sufficently smooth for direct expansion arguments to apply. This means that the burden of smoothness must be carried by F, the distribution of the errors $\{e_i\}$. We require

(v) the density $f = F'$ is continuous and positive on its support.

Part of our motivation for exploring the heteroscedastic M-estimation approach is to remove the burden of smoothness from F which is not under our control. We, therefore, consider

• **Heteroscedastic M-estimation**

Let $(\hat{\beta}^T, \hat{\theta}^T)^T \in R^{p+q+2}$ denote a solution of the system of equations

$$A_n(\beta, \theta) = 0,$$

where

$$A_n(\beta,\theta) = (nh)^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} z_i(x)\exp(-w_i(x)^T\theta)\psi\{\exp(-w_i(x)^T\theta)(Y_i - z_i(x)^T\beta)\} \\ w_i(x)\chi\{\exp(-w_i(x)^T\theta)(Y_i - z_i(x)^T\beta)\}\end{pmatrix} K\Big\{\frac{X_i - x}{h}\Big\},$$

$h > 0$, ψ and χ are real functions satisfying $\int \psi(e)dF(e) = \int \chi(e)dF(e) = 0$,

$z_i(x) = (1, X_i - x, (X_i - x)^2, \ldots, (X_i - x)^p)^T, \quad 1 \le i \le n,\ p \ge 0,$ and
$w_i(x) = (1, X_i - x, (X_i - x)^2, \ldots, (X_i - x)^q)^T, \quad 1 \le i \le n,\ q \ge 0.$ Then set

$$\hat{m}^{(r)}(x) = r!\hat{\beta}_{r+1}$$

and

$$\hat{l}^{(r)}(x) = r!\hat{\theta}_{r+1}.$$

Here $\hat{l}^{(r)}(x)$ is estimating $l^{(r)}(x)$, the rth derivative of $\log\{s(x)\}$. We can either adopt this as the natural scale for treating spread or we can convert back to the raw scale using $s(x) = \exp\{l(x)\}$, $s^{(1)}(x) = l^{(1)}(x)s(x)$ etc.

Note that we obtain the heteroscedastic M-estimation procedure by approximating (1.1) by the heteroscedastic regression model

$$Y_i = z_i(x)^T\beta + \exp(w_i(x)^T\theta)e_i, \quad 1 \le i \le n,$$

writing down estimating equations for M-estimation of the parameters β and θ, and then making the fit local by including kernel weights. If the errror distribution F is known and has density f, we can obtain the "maximum likelihood" estimator by taking $\psi(x) = -f'(x)/f(x)$ and $\chi(x) = x\psi(x) - 1$. The usual choice when F is Gaussian is obtained by taking $\psi(x) = x$. However, ψ and χ can be chosen to be bounded functions to ensure robustness (Huber (1981, p.135ff)). Heteroscedastic M-estimation can be regarded both as local maximum likelihood and local pseudo-likelihood estimation since these are identical in this case. A study of these methods for fitting parametric heteroscedastic regression models under general conditions is given by Welsh, Carroll and Ruppert (1994).

For the heteroscedastic M-estimation method, the order of the bias is determined by $\min(q, p)$ unless the estimators of regression and scale are orthogonal in the sense that

$$\int e\psi'(e)dF(e) = \int \chi'(e)dF(e) = 0.$$

This condition is satisfied in the important case when $\psi(x) = x$ and $\chi(x) = x\psi(x) - 1$ but is not generally satisfied by robust estimators unless F is symmetric. It is difficult to assess the validity of an assumption of symmetry in the presence of heteroscedasticiy so we avoid this assumption. Thus, unless we are confident that F is symmetric, general scale equivariant robust estimation of the regression function requires estimation of the spread to the same order. This is easily achieved by setting $p = q$.

Our results for the heteroscedastic M-estimation approach require

(vi) As $a, b \to 0$,

$$E\psi\{\exp\{a\}[e + b]\} = Ee\psi'(e)a + E\psi'(e)b + o(|a| + |b|)$$

and

$$E\chi\{\exp\{a\}[e + b]\} = Ee\chi'(e)a + E\chi'(e)b + o(|a| + |b|).$$

Moreover, the matrices

$$K = \begin{pmatrix} s(x)^{-2}E\psi(e)^2 & s(x)^{-1}E\psi(e)\chi(e) \\ s(x)^{-1}E\psi(e)\chi(e) & E\chi(e)^2 \end{pmatrix} \text{ and } M = \begin{pmatrix} s(x)^{-2}E\psi'(e) & s(x)^{-1}Ee\psi'(e) \\ s(x)^{-1}E\chi'(e) & Ee\chi'(e) \end{pmatrix}$$

are both finite and M is nonsingular.

(vii) For $a, b \to 0$ and $|a_1|, |b_1| \le C < \infty$,

$$E\sup\{|\psi[\exp\{b_1{+}b\}\{e{+}a_1{+}a\}] - \psi[\exp\{b_2{+}b\}\{e{+}a_2{+}a\}]| : |a_1{-}a_2|, |b_1{-}b_2| \le \delta\} \le C\delta$$

and

$$E\sup\{|\chi[\exp\{b_1{+}b\}\{e{+}a_1{+}a\}] - \chi[\exp\{b_2{+}b\}\{e{+}a_2{+}a\}]| : |a_1{-}a_2|, |b_1{-}b_2| \le \delta\} \le C\delta$$

for $\delta > 0$ sufficiently small, and, as $\delta \to 0$

$$E\sup\{|\psi[\exp\{b_1{+}b\}\{e{+}a_1{+}a\}] - \psi[\exp\{b_2{+}b\}\{e{+}a_2{+}a\}]|^2 : |a_1{-}a_2|, |b_1{-}b_2| \le \delta\} = o(1)$$

and

$$E\sup\{|\chi[\exp\{b_1{+}b\}\{e{+}a_1{+}a\}] - \chi[\exp\{b_2{+}b\}\{e{+}a_2{+}a\}]|^2 : |a_1{-}a_2|, |b_1{-}b_2| \le \delta\} = o(1).$$

These are essentially the conditions used by Carroll and Ruppert (1982) in their study of robust estimation for parametric heteroscedastic regression models; (see also Härdle and Luckhaus (1984) for further motivation). Although condition (vi) is written for the case that ψ and χ are smooth, when χ and ψ are monotone but not smooth and F is smooth, the results of Section 3 hold with only notational changes. The suprema inside the expectation in (vii) can be removed when ψ is of bounded variation and χ is of bounded variation on $[0, \infty)$.

3. Theoretical Results

3.1. Theorems

We first prove that the $(r+1)$st component of the running regression quantile estimator $\hat{\beta}_{r+1}(\alpha)$ estimates $\{m^{(r)}(x) + s^{(r)}(x)F^{-1}(\alpha)\}/r!$. It follows immediately that if we impose $F^{-1}(1/2) = 0$, the $(r+1)$st component of the running median estimator $\hat{\beta}_{r+1}(1/2)$ estimates $m^{(r)}(x)/r!$. The result gives the order of the asymptotic bias and the variance of $\hat{\beta}_{r+1}(\alpha)$.

Theorem 3.1. *Suppose that Conditions* (i)-(v) *hold, x is an element of the interior of the support of g and that $nh^{2r+1} \to \infty$ as $n \to \infty$ and $h \to 0$. Then for $0 < \alpha < 1$ the running regression quantile estimator $\hat{\beta}_{r+1}(\alpha)$ estimates $\{m^{(r)}(x) + s^{(r)}(x)F^{-1}(\alpha)\}/r!$ with asymptotic bias of $O(h^{p-r+1})$ for $p-r$ odd and $O(h^{p-r+2})$ for $p-r$ even. In either case, the asymptotic variance is*

$$n^{-1}h^{-2r-1}f\{F^{-1}(\alpha)\}^{-2}\alpha(1-\alpha)g(x)^{-1}s(x)^2k_{r+1}^T N_p^{-1}T_pN_p^{-1}k_{r+1}.$$

The exact expression for the asymptotic bias and full details of the proof are available from the author. The algebra in the proof is complicated but the broad approach is conceptually simple. Let

$$A_n(\beta) = (nh)^{-1/2}\sum_{i=1}^n z_i(x)\psi\{Y_i - z_i(x)^T\beta\}\ K\{(X_i - x)/h\},$$

where $\psi(x) = \alpha - I(x < 0)$ denotes the estimating equations for $\hat{\beta}_{r+1}(\alpha)$, and define $\gamma(\alpha) \equiv \gamma_n(\alpha)$ by $E[A_n\{\gamma(\alpha)\}|X] = 0$. The bias is studied by making a local Taylor series expansion of $E[A_n\{\gamma(\alpha)\}|X]$ and then solving for $\gamma(\alpha)$. The asymptotic variance is established by a modification to the proof of Lemma 4.1 of Bickel (1975), to prove that

$$\sup\{|A_n(\beta) - A_n(\gamma) - E\{A_n(\beta)|X\}| : |\beta - \gamma| \le C(nh)^{-1/2}\} = o_p(1).$$

Arguing as in the proof of Lemma A.2 of Ruppert and Carroll (1980) we can then prove that $A_n(\hat{\beta}) = o_p(1)$ and then; using the method of Jurečková (1977), that $\hat{\beta} - \gamma = O_p\{(nh)^{-1/2}\}$. Then, since $EA_n(\gamma) = 0$, we have

$$(nh)^{1/2}(\beta - \gamma) = f\{F^{-1}(\alpha)\}^{-1}D_n^{-1}A_n(\gamma) + o_p(1)$$

from which the result follows.

It is now straightforward to establish the behaviour of the running interquantile range estimator. We see that $\hat{\beta}_{r+1}(1-\alpha)-\hat{\beta}_{r+1}(\alpha)$ estimates $s^{(r)}(x)\{F^{-1}(1-\alpha)-F^{-1}(\alpha)\}/r!$.

Corollary 3.1. *Suppose that Conditions* (i)-(v) *hold, x is an element of the interior of the support of g and that $nh^{2r+1}\to\infty$ as $n\to\infty$ and $h\to 0$. Then for $0<\alpha<0.5$, the running interquartile range estimates $s^{(r)}(x)\{F^{-1}(1-\alpha)-F^{-1}(\alpha)\}/r!$ with asymptotic bias of $O(h^{p-r+1})$ for $p-r$ odd and $O(h^{p-r+2})$ for $p-r$ even. In either case, the asymptotic variance is*

$$n^{-1}h^{-2r-1}g(x)^{-1}s(x)^2\alpha\Big[(1-\alpha)f\{F^{-1}(1-\alpha)\}^{-2}+(1-\alpha)f\{F^{-1}(\alpha)\}^{-2} - 2\alpha f\{F^{-1}(1-\alpha)\}^{-1}f\{F^{-1}(\alpha)\}^{-1}\Big]k_{r+1}^T N_p^{-1}T_pN_p^{-1}k_{r+1}.$$

The result follows directly from the preceding theorem.

We next establish that the $(r+1)$st component of the heteroscedastic M-estimators $\hat{\beta}_{r+1}$ and $\hat{\theta}_{r+1}$ estimate $m^{(r)}(x)/r!$ and $l^{(r)}(x)/r!$ respectively, when the location and scale functionals on F which identify the model (1.1) are implicitly defined by the requirement that $\int\psi(e)dF(e)=\int\chi(e)dF(e)=0$. We first give the result for the case $p=q$ and then for general p and q but assuming that $\int e\psi'(e)dF(e)=\int\chi'(e)dF(e)=0$. The proofs are similar to that of Theorem 3.1 and so are omitted. However, exact expressions for the asymptotic bias and the full details of the proofs are available from the author.

Theorem 3.2. *Suppose that Conditions* (i)-(iv) *and* (vi)-(vii) *hold, x is an element of the interior of the support of g and that $nh^{2r+1}\to\infty$ as $n\to\infty$ and $h\to 0$. Then $\hat{\beta}_{r+1}$ and $\hat{\theta}_{r+1}$ estimate $m^{(r)}(x)/r!$ and $l^{(r)}(x)/r!$, respectively, with biases of $O(h^{p-r+2})$ if $p-r$ is even and $O(h^{p-r+1})$ if $p-r$ is odd. In either case, the asymptotic variances are*

$$\mathrm{Var}[\hat{\beta}_{r+1}|X]\sim n^{-1}h^{-2r-1}(M^{-1}KM^{-1})_{11}g(x)^{-1}k_{r+1}^TN_p^{-1}T_pN_p^{-1}k_{r+1}$$

and

$$\mathrm{Var}[\hat{\theta}_{r+1}|X]\sim n^{-1}h^{-2r-1}(M^{-1}KM^{-1})_{22}g(x)^{-1}k_{r+1}^TN_p^{-1}T_pN_p^{-1}k_{r+1}.$$

Theorem 3.3. *Suppose that Conditions* (i)-(iv) *and* (vi)-(vii) *hold, x is an element of the interior of the support of g and that $nh^{2r+1}\to\infty$ as $n\to\infty$ and $h\to 0$. Then $\hat{\beta}_{r+1}$ estimates $m^{(r)}(x)/r!$ with bias of $O(h^{p-r+2})$ if $p-r$ is even and of $O(h^{p-r+1})$ if $p-r$ is odd, and $\hat{\theta}_{r+1}$ estimates $l^{(r)}(x)/r!$ with biases of $O(h^{q-r+2})$ if $q-r$ is even and of $O(h^{q-r+1})$ if $q-r$ is odd. In either case, the asymptotic variances are*

$$E[\{\hat{\beta}_{r+1}-\gamma_{r+1}\}^2|X]\sim n^{-1}h^{-2r-1}g(x)^{-1}s(x)^2\{E\psi'(e)\}^{-2}E\psi(e)^2k_{r+1}^TN_p^{-1}T_pN_p^{-1}k_{r+1}$$

and

$$E[\{\hat{\theta}_{r+1} - \tau_{r+1}\}^2|X] \sim n^{-1}h^{-2r-1}g(x)^{-1}\{Ee\chi'(e)\}^{-2}E\chi(e)^2 k_{r+1}^T N_p^{-1} T_p N_p^{-1} k_{r+1}.$$

Moreover, $\hat{\beta}$ and $\hat{\theta}$ are asymptotically conditionally independent given X.

3.2. Remarks

Remark 1. Although our proofs deal naturally with vectors of estimators, we have stated all our results for a single component of those vectors. This is partly because it is simpler and partly because of the need to at least consider using different bandwidths to estimate the different functions of interest. However, matrix expressions are easily obtained. Moreover, we have chosen to state the results in terms of asymptotic conditional biases and variances rather than local Bahadur representations or asymptotic conditional normality as has been done by Fan, Hu and Truong (1992). However, since our proofs produce asymptotic linearity results for the estimators, conversion to these formats is straightforward. Thus, in the special case $s(x) = 1$, $r = p = 1$, $\chi \equiv 0$ and $w_i(x) \equiv 0$, Theorems 3.1 and 3.3 have essentially the same content Theorem 1 of Fan, Hu and Truong (1992). It is also possible (but not very useful) to restate our results in terms of equivalent kernels as has been done by Ruppert and Wand (1994). If we take $\psi(x) = x$, $\chi \equiv 0$ and $w_i(x) \equiv 0$ Theorem 3.3 reduces to Theorems 4.1 and 4.2 of Ruppert and Wand (1994).

Remark 2. In all our results, if $p - r$ (or $q - r$) is odd, the leading term of the conditional bias of the estimator $m^{(r)}(x)$ or $s^{(r)}(x)$ does not involve the density g of X. This property was called design adaptivity by Fan (1992, 1993). However, as noted by Ruppert and Wand (1994), it does not hold when $p - r$ is even.

Remark 3. We often choose $p = r + 1$ or $r + 2$ in applications. Thus if we want to estimate the regression and spread functions (for which $r = 0$), we can take $p = 1$ to incur a bias of $O(h^2)$. The slope terms estimate the derivatives ($r = 1$) at no extra cost with a bias of $O(h^2)$ too. If we increase p so $p = 2$, the bias for the estimators of the regression and spread functions is $O(h^4)$ while that of the derivatives remains $O(h^2)$. For $p = 3$, in estimating the regression and spread functions, we incur a bias of $O(h^4)$ and for estimating their derivatives, a bias of $O(h^4)$. These results are summarised in Table 1 below. Of course, in particular applications, we may require the bias to be smaller than some specified order and in this case we may need to choose p to satisfy these requirements.

Remark 4. Our results are easily manipulated to produce asymptotic conditional mean squared errors. For example, for the running regression quantile estimators, the asymptotic conditional mean squared error of $r!\hat{\beta}_{r+1}(\alpha)$ is the

sum of two terms, the first of which is of $O(n^{-1}h^{-2r-1})$ and the second of which is of $O(h^{2(p-r+1)})$ for $p-r$ odd and $O(h^{2(p-r+2)})$ for $p-r$ even. The asymptotic conditional mean integrated squared error can be obtained by integrating these expressions with respect to a weight function over the support of g. These criteria can then be minimised with respect to h to obtain bandwidths which produce optimal estimators with respect to asymptotic conditional mean or mean integrated squared error. The optimal window width h_0 for estimating the rth derivative of the regression function satisifies $h_0 = O(n^{-1/(2p+3)})$ for $p-r$ odd and $h_0 = O(n^{-1/(2p+5)})$ for $p-r$ even. We can use a single h to estimate m and s and their derivatives simultaneously, but the optimal window width for estimating $m^{(r)}$ or $s^{(r)}$ changes as r changes so simultaneous optimal estimation is not possible. This is often not important when these quantities are merely nuisance parameters because then very different optimality criteria should be used. Alternatively, we can use different weights (corresponding to different bandwidths) to estimate each derivative. To use the optimal window widths in practice, we need to construct preliminary estimates of the unknown quantities in the appropriate asymptotic conditional mean or mean integrated squared error. This problem is has been investigated by Ruppert, Sheather and Wand (1995) for the least squares estimator. Since it should be straightforward to extend their results to the present problem, we will not pursue the matter in detail here.

Remark 5. The results obtained above all require x to be interior to the support of g but we can also obtain results which apply when x is close to the boundary of the support of g. Let $\text{supp}(g) = [a,b]$ and $x = \partial x + hc$, where $\partial x = a$ or b, and c is chosen so $\partial x + hc \in [a,b]$. Finally, let $S = \{-c + (a - \partial x)/h < u < -c + (b - \partial x)/h\} \cap \text{supp}(K)$. Now for the running regression quantiles, the asymptotic bias for the $(r+1)$th entry is readily shown to be of $O(h^{p-r+1})$. In the asymptotic variance, T_p is replaced by $T_p(S)$, the matrix with entries $\int_S u^{i+j-2}K(u)^2du$. However, T_p was not patterned so the form of the expression is unchanged and we have

$$\text{Var}\{(nh)^{1/2}(\hat{\beta}-\gamma)\} \sim g(x)^{-1}f\{F^{-1}(\alpha)\}^{-2}\alpha(1-\alpha)H^{-1}N_p(S)^{-1}T_p(S)N_p(S)^{-1}H^{-1}.$$

The general point is that the effect of the boundary is to restrict the domain of integration of all integrals to S and thereby destroy the all important patterns in N_p. This simplifies the expression for the asymptotic bias and means that the bias of the $(r+1)$th component is $O(h^{p-r+1})$ whether $p-r$ is odd or even. In particular, for $p-r$ odd, this means that the asymptotic bias on the boundary is of the same order as that in the interior, a point noted in the work of Fan (1990, 1992), Ruppert and Wand (1994), and Fan, Hu and Truong (1992).

Table 1. The order of the bias of estimators of the regression and spread functions ($r = 0$) and their first derivatives ($r = 1$) for polynomials of order p when x is an interior or a boundary point.

	x in Interior		x on Boundary	
	$r = 0$	$r = 1$	$r = 0$	$r = 1$
$p = 0$	h^2	–	h	–
$p = 1$	$h^{2\,*}$	h^2	$h^{2\,*}$	h
$p = 2$	h^4	$h^{2\,*}$	h^3	$h^{2\,*}$
$p = 3$	$h^{4\,*}$	h^4	$h^{4\,*}$	$h3$

* Cases where $p - r$ is odd.

This, together with the design adaptivity property, is the primary motivation for using $p = r + 1$ as a default choice. It is straightforward to obtain similar results for boundary points for the other estimators.

Remark 6. Problems with vector X of dimension $d \geq 1$ are also of interest. The case $d = 2$ is of particular interest because it arises naturally with spatial geographic data and the results can still be examined graphically. Asymptotics in a general multidimensional setting have been addressed by Stone (1980, 1982). The extension of our results to this case is notationally complex but otherwise straightforward. The details have been omitted to save journal space but are available from the author.

4. Examples

In this section we illustrate the practical utility of the methodolgy we have proposed by applying it to two data sets. All the analysis was carried out in Splus.

We used local linear fits ($p = 1$) for simplicity throughout. Thus, we estimated the regression and spread functions simultaneously without changing the order of the local polynomial or the window width to improve derivative estimation. The choice of $p = 1$ ensures that the bias of the regression and spread estimators is of $O(h^2)$ both in the interior and on the boundary. On the other hand, the bias of the estimators of the derivatives of these functions is of $O(h^2)$ in the interior but only $O(h)$ on the boundary and so are not expected to perform as well. Nonetheless, as is shown below, the derivative estimators perform rather well. This performance is remarkable in view of the notorious difficulty of using conventional smoothing methods to estimate derivatives.

We used the Gaussian density function as the kernel function in all the fits.

The window width h was chosen by inspection based on trial and error. An initial plausible value was chosen as

$$h = 2\min\{SD(X),\ MAD(X)\},$$

where $SD(X)$ denotes the usual sample standard deviation of the covariates and $MAD(X)$ denotes 1.48 times the median absolute deviation from the median of the covariates. This is a slightly modified version of the suggestion of Silverman (1986, p 45ff). The window was then adjusted until a reasonably smooth fit with little apparent bias was obtained.

The running regression quantile fits were obtained using the function written by Koenker and D'Orey (1987). There is no direct provision for passing weights to the function but the fit at x is easily effected by applying the function to $Y_iK\{(Xi - x)/h\}$ and $(X_i - x)K\{(X_i - x)/h\}$. For the heteroscedastic M-estimators we solved the defining estimating equations at x by means of a modification of the robust pseudo-likelihood algorithm described by Carroll and Ruppert (1988, p196).

For our robust fits, we need to specify ψ and χ. We used a variant of Huber's proposal 2 for the heteroscedastic M-estimators. That is, we used the Huber ψ function for which $\psi(x) = \max(-c, \min(c, x))$ with $c = 1.35$ and set $\chi(x) = \psi(x)^2 - \int_{-\infty}^{\infty} \psi(t)^2 d\Phi(t)$ with $c = 2.00$, where Φ is the standard Gaussian distribution function. This choice is simple and familiar but there are other possibilities; an interesting possibility would be to replace the Huber ψ function by the bisquare function for which $\psi(x) = x\{1-(x/c)^2\}^2 I(|x| \leq c)$ and $c = 4.685$. As noted in Section 2, a local maximum likehood fit based on the Gaussian distribution is obtained by taking $\psi(x) = x$ and $\chi(x) = x^2 - 1$. The resulting fit is not robust but we use it in the second example below to show the benefits of the robust procedure.

• The raptor data

The first example involves data presented by Olsen, Cunningham and Donnelly (1994) as part of a study of the evolutionary ecology of reproduction of raptors. For $n = 267$ species of raptor, the response variable Y is the logarithm of the average egg volume and the explanatory variable X is the logarithm of the average female weight. The data exhibit some curvature and clear heteroscedasticity on the raw scale. Standard applied practice in such situations is often to transform the data; the logarithmic transformation applied to both variables produces approximate linearity and removes the heteroscedasticity. We will apply local polynomial smoothing to assess the effectiveness of the transformation strategy.

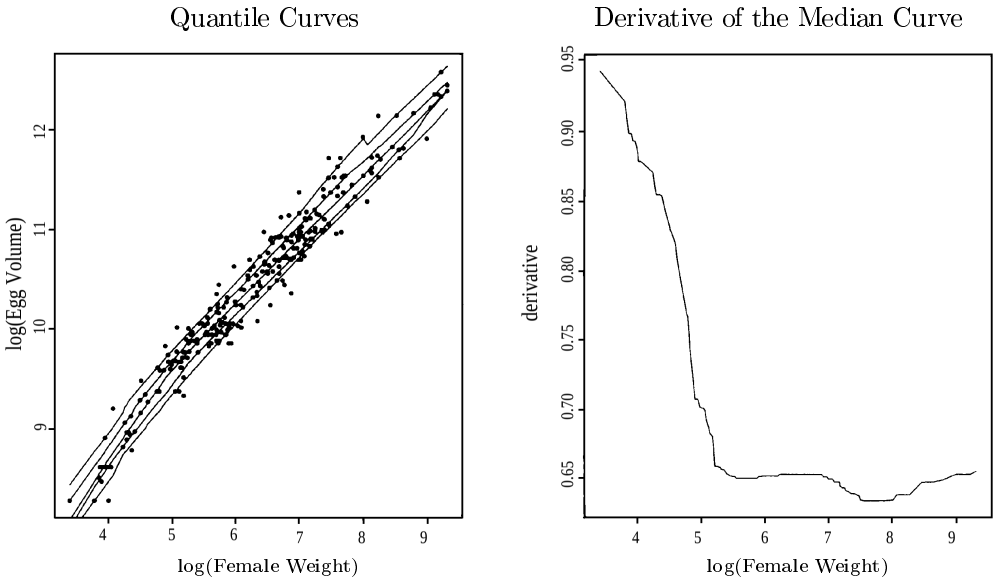

Figure 4.1. Modelling the response in the raptor data

Figure 4.1 is a scatterplot of the transformed data with the 0.10, 0.25, 0.50, 0.75 and 0.90 running regression quantile curves superimposed. (These are too close together to label but obviously α increases from bottom to top. Thus the top curve is the 0.90 curve, the middle the 0.50 curve and the bottom the 0.10 curve.) For simplicity, the window width for each curve is $h = 0.75$. The fact that the curves are all roughly parallel shows that the transformation has indeed stabilised the variance. However, there is evidence of curvature in the lower left corner of the plot. This is highlighted by the adjacent plot of the derivative of the median curve (also obtained with $h = 0.75$) which decreases before becoming roughly constant. This shows clearly that there is curvature in the lower left corner but that otherwise the relationship is linear.

The outer quantile curves appear to be less smooth than the inner ones. Indeed, there is a clear jump in the 0.90 quantile. This suggests that when the outer quantiles are of interest they should be estimated using a larger window width than we would use for the less extreme quantiles. This is intuitively reasonable as the estimators of the outer quantiles are more variable than those of the less extreme quantiles. Theorem 3.1 shows that the asymptotic variance of the running quantile curves depends on α only through $f\{F^{-1}(\alpha)\}^{-2}\alpha(1-\alpha)$ which, for nice log-concave densities, increases as α tends to zero or one. The exact expression for the asymptotic bias shows that it depends on α only through $m^{(2)}(x) + s^{(2)}(x)F^{-1}(\alpha)$, the second derivative of the quantile function, so the squared bias also increases in α if $s^{(2)}(x) \neq 0$. Another alternative is to simply

smooth the quantile curves a second time or to use a method like twicing. We have not used any of these methods in the present examples.

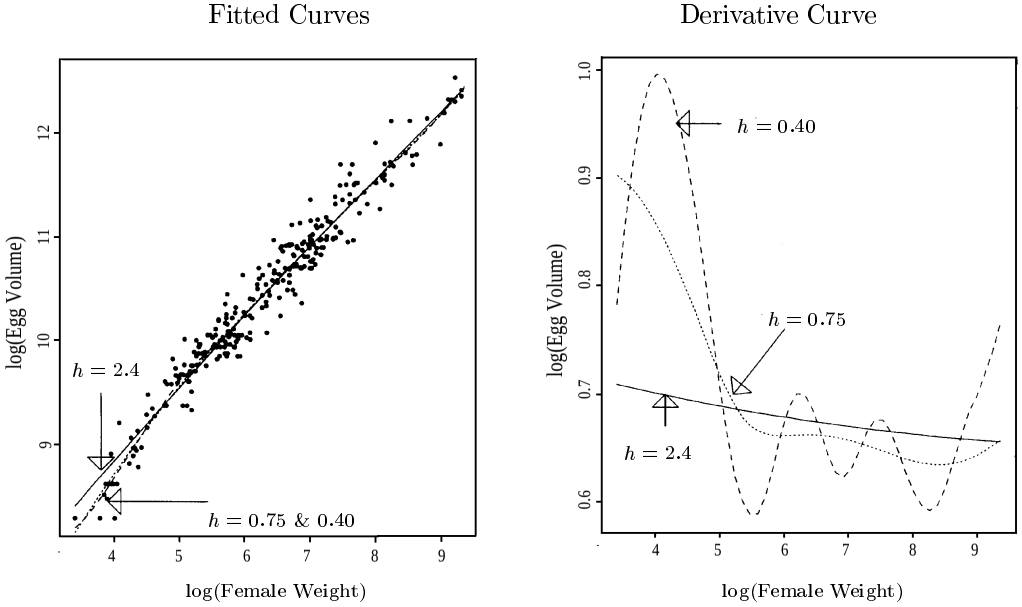

Figure 4.2. Window width and the running Huber curve for the raptor data

Figure 4.2 shows the effect of using a running Huber M-estimator and changing the window width on the regression function and its first derivative. With larger window widths, the relationship appears much more linear and we require relatively small window widths to pick up the curvature. The regression functions with $h = 0.75$ and 0.40 are essentially identical though the corresponding derivatives are rather different. As the window width decreases, the derivative becomes noticeably less smooth, supporting the view that it is often sensible to use larger window widths to estimate a derivative than to estimate the function itself. Nonetheless, the results for both the regression function and its derivative are excellent for $h = 0.75$ and confirm the finding from the regression quantiles.

The phenomenon of curvature remaining in the lower left corner of a scatterplot after transformation is not uncommon. However, what we decide to do about it depends strongly on the context and objectives of the analysis. In the present example, a linear model fits the bulk of the data extremely well and produces a familiar, interpretable allometric relationship. It is reasonable to use this model to describe the large raptors and make rough generalisations about all raptors. However, if we are specifically interested in the small raptors, then our analysis shows that the simple model may be inadequate.

It is interesting to note that the curve and derivative obtained using the

smooth Huber M-estimator are smoother than those obtained from the running median fit. This can be attributed to the fact that the Huber M-estimator has a smooth influence function while the median has an influence function with a jump discontinuity at the origin. This property of producing smoother estimates may be an advantage of the M-estimation procedure over the regression quantile procedure in some applications.

Finally, note that the quantile curves can cross. This slightly disturbing feature is a consequence of the local exact fit property which forces each weighted regression quantile fit to pass through at least two points in the neighbourhood of points with non-zero weight. For the fit at x, the quantile curves are correctly ordered only at the kernel weighted mean of the covariates. For interior points, this will often be close to the centre of the neighbourhood, namely x, but for boundary points, this will differ from x. Combined with the local exact fit property, it is clear that crossing can occur and that if it does occur, it is more likely to do so near the boundary.

• The moderate sized beef farm data

The second example involves the set of $n = 376$ farms with between 50 and 2500 beef cattle which participated in the 1988 Australian Agricultural and Grazing Industries Survey carried out by the Australian Bureau of Agricultural and Resource Economics. The response Y here is the farm income and the covariate X is the number of beef cattle.

This set of moderately sized beef farms was used as a population of interest by Welsh and Ronchetti (1994) to illustrate the use of robust methods in the analysis of sample survey data. The analysis was carried out using the model

$$Y_i = \beta X_i + g(X_i)e_i, \qquad 1 \leq i \leq n,$$

where $g(x) = x$. Models of this type are widely used in the super population approach to the analysis of sample survey data but their utility depends on their validity. We will use robust local polynomial smoothing with the default window-width $h = 400$ to assess the validity of this model.

Figure 4.3 shows the running quantile curves and the local IQR curve for the moderately sized beef farm data. There appear to be a few very extreme outliers present in the data. The plots show a roughly linear relationship with severe heteroscedasticity and some asymmetry.

Figure 4.4 shows a scatterplot of the data with both a Gaussian and a Huber local polynomial fit. The Gaussian fit is increased by the asymmetry and the outliers so the Gaussian curve lies above the Huber curve in the plot. Both fits are roughly linear until the right hand edge of the plot although the Gaussian curve is perturbed slightly by the extreme outliers and so is slightly less smooth

than the Huber curve. This effect is shown even more clearly in the plot of the derivatives where the Gaussian derivative curve is more variable than the robust derivative curve. The nonlinearity at the edge of the plot is due to only two points which are surprisingly small relative to their neighbours and should not be overemphasised. Indeed, increasing the window width tends to remove the nonlinearity. Thus we conclude that the linear regression model is reasonable.

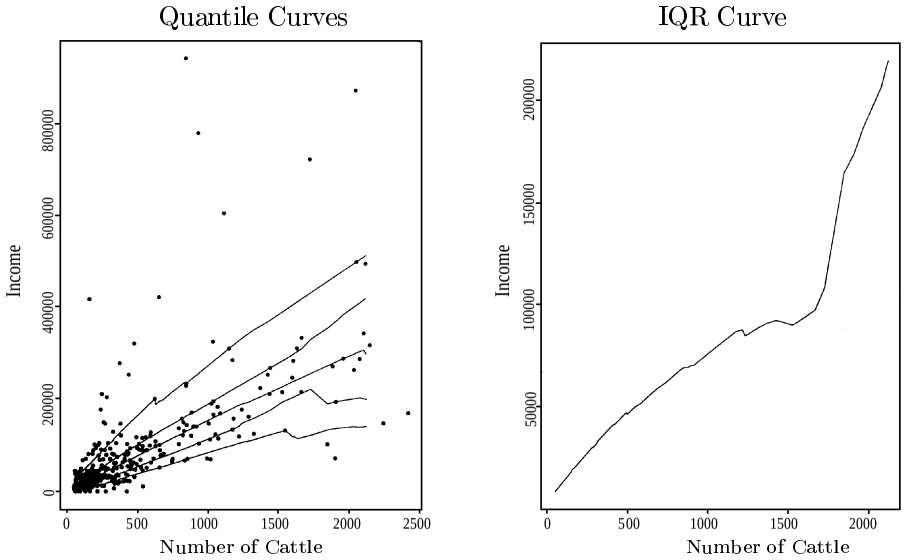

Figure 4.3. Modelling the beef farm data

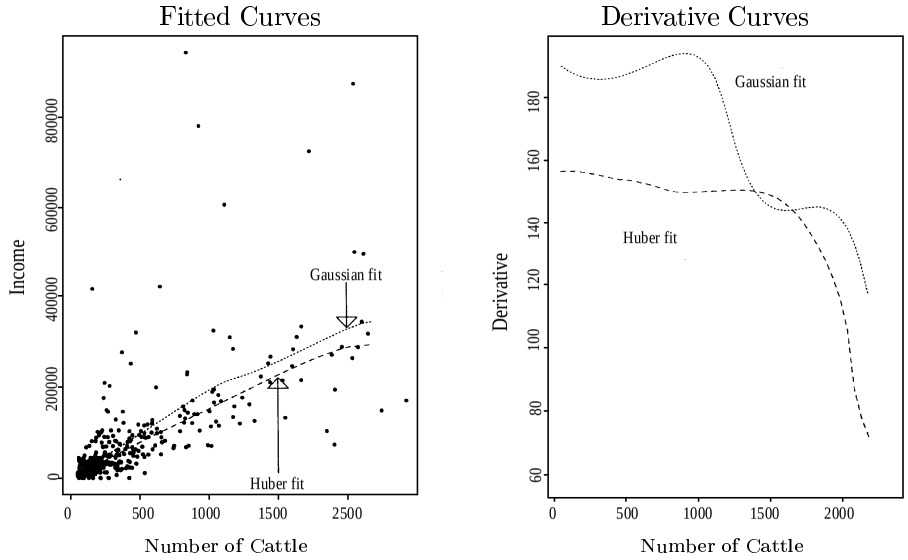

Figure 4.4. Robust and nonrobust fits to the response for the beef farm data

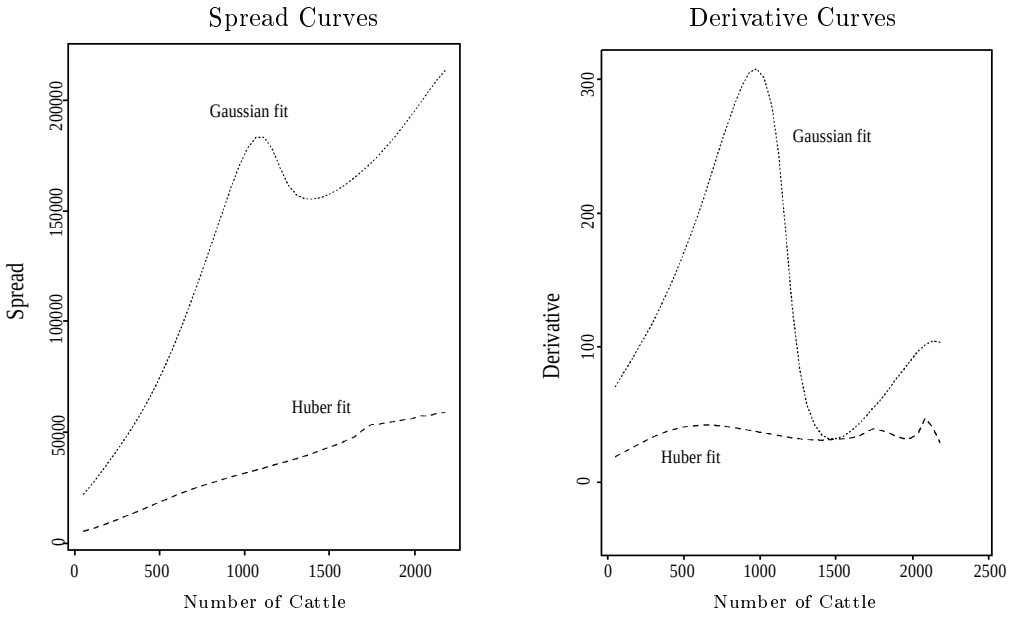

Figure 4.5. Robust and nonrobust fits to the variability for the beef farm data

The effect of the outliers on nonrobust estimators of regression functions is to increase their variability and, when the outliers are also asymmetric, to increase their bias relative to robust estimators. Since the easiest visual comparison of curves is of their relative biases, examples with asymmetric outliers are visually powerful. Nonetheless, at least relative to the scale of the data (which is not necessarily the best scale for the comparison), the magnitudes of such biases can seem misleadingly small. This is not of course the case in spread estimation where outliers typically have a substantial effect. This is shown clearly in Figure 4.5 where the Gaussian estimate of the spread function is far too large and visibly tracks the extreme outliers in the data. In contrast, the Huber estimate is much smaller and is robust against the extreme outliers. These differences also occur in the estimates of the derivatives of the spread functions.

Finally, the linearity of the robust spread function estimate and the constancy of its derivative estimate support the assumption of Welsh and Ronchetti (1994) that $g(x) = x$, the spread is linear in the number of cattle.

Acknowledgements

I am grateful to David Ruppert and J. Fan for sending me preprints of their papers, to David Ruppert and Matthew Wand for helpful conversations, to Penny Olsen, Ross Cunningham and Christine Donnelly for letting me use the raptor data, and to the Editor and an Associate Editor for their helpful suggestions.

References

Bickel, P. J. (1975). One-step Huber estimates in the linear model. J. Amer. Statist. Assoc. **70**, 428-434.

Carroll, R. J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. Ann. Statist. **10**, 429-441.

Carroll, R. J. and Ruppert, D. (1988). Transformation and Weighting in Regression. Chapman and Hall, London.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local bahadur representation. Ann. Statist. **19**, 760-777.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. J. Amer. Statist. Assoc. **74**, 829-836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally-weighted regression and smoothing scatterplots. J. Amer. Statist. Assoc. **83**, 597-610.

Eubank, R. J. (1988). Spline Smoothing and Nonparametric Regression. Marcel Dekker, New York.

Fan, J. (1990). Local linear regression smoothers and their minimax efficiency. Preprint.

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., Hu, T.-C. and Truong, Y. K. (1992). Robust nonparametric function estimation. Preprint.

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

Härdle, W. and Luckhaus, S. (1984). Robust nonparametric regression with simultaneous scale curve estimation. Ann. Statist. **16**, 120-135.

Hastie, T. and Loader, C. (1993). Local regression: Automatic kernel carpentry (with discussion). Statist. Sci. **8**, 139-143.

Huber, P. J. (1981). Robust Statistics. John Wiley, New York .

Jurečková, J. (1977). Asymptotic relations of M-estimates and R-estimates in linear regression model. Ann. Statist. **5**, 464-472.

Koenker, R. and Bassett, G. (1978). Regression quantiles. Econometrica **46**, 33- 50.

Koenker, R. and D'Orey, V. (1987). Computing regression quantiles. Appl. Statist. **36**, 383-393.

McCullagh, P. and Nelder, J. A. (1990). Generalised Linear Models (Second edition). Chapman and Hall, London.

Müller, H. G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. J. Amer. Statist. Assoc. **82**, 231-238.

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

Olsen, P. D., Cunningham, R. B. and Donnelly, C. F. (1994). Avian egg morphometrics: Allometric models of egg volume, clutch volume and shape. Austral. J. Zoo. To appear.

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. J. Amer. Statist. Assoc. **75**, 828-838.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. J. Amer. Statist. Assoc. **90**, 1257-1270.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. Ann. Statist. **21**, 1346-1370.

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

Stone, C. J. (1977). Consistent nonparametric regression. Ann. Statist. **5**, 595-620.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. Ann. Statist. **8**, 1348-1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. Ann. Statist. **10**, 1040-1053

Tsybakov, A. B. (1986). Robust reconstruction of functions by local-approximation method. Problems Inform. Transmission **22**, 133-146.

Wahba, G. (1990). Spline Models for Observational Data. SIAM, Philadelphia.

Weisberg, S. and Welsh, A. H. (1995). Adapting for a missing link. Ann. Statist. **22**, 1674-1700.

Welsh, A. H., Carroll, R. J. and Ruppert, D. (1994). Fitting heteroscedastic regression models. J. Amer. Statist. Assoc. **89**, 100-116.

Welsh, A. H. and Ronchetti, E. (1994). Bias calibrated estimation of totals and quantiles from sample surveys containing outliers. Preprint.

Department of Statistics, The Faculties, The Australian National University, Canberra ACT 0200, Australia.

(Received November 1993; accepted April 1995)