Relative error prediction via kernel regression smoothers

Heungsun Park[a], **Key-Il Shin**[a], **M.C. Jones**[b], **S.K. Vines**[b]

[a] *Department of Statistics, Hankuk University of Foreign Studies, YongIn KyungKi, 449-791, Korea*

[b] *Department of Statistics, The Open University, Milton Keynes, MK7 6AA, UK*

Abstract

In this article, we introduce and study local constant and our preferred local linear nonparametric regression estimators when it is appropriate to assess performance in terms of mean squared relative error of prediction. We give asymptotic results for both boundary and non-boundary cases. These are special cases of more general asymptotic results that we provide concerning the estimation of the ratio of conditional expectations of two functions of the response variable. We also provide a good bandwidth selection method for our estimator. Examples of application and discussion of related problems and approaches are also given.

Keywords: Local linear regression; Mean squared relative error; Nadaraya-Watson estimator; Ratio estimation.

Corresponding author: M.C. Jones. *E-mail address*: m.c.jones@open.ac.uk; *Phone*: +44 1908 652209; *Fax*: +44 1908 655515

1. Introduction

Suppose that Y is a response variable and $\tilde{Y}$ is a predictor of Y that is a function of a single predictor variable X. In ordinary predictions, we obtain $\tilde{Y}$ by estimating the conditional mean of a response given predictor value, $E(Y|X)$, because it minimises the expected squared loss, $E\{(Y-\tilde{Y})^2|X\}$, which is Mean Squared Error (MSE). However, when $Y > 0$, it will often be that the ratio of prediction error to the response level, $(Y-\tilde{Y})/Y$, is of prime interest: the expected squared relative loss, $E[\{(Y-\tilde{Y})/Y\}^2|X]$, which is Mean Squared Relative Error (MSRE), is to be minimised. Relative error is considered in many disciplines (Narula and Wellington, 1977, Farnum, 1990, Khoshgoftaar, Bhattacharyya, and Richardson, 1992a, Khoshgoftaar, Munson, Bhattacharyya, and Richardson, 1992b, Park and Shin, 2006), particularly those connected with engineering.

Park and Stefanski (1998) showed that we need to estimate

$$\frac{E\left(Y^{-1}|\,X\right)}{E\left(Y^{-2}|\,X\right)}, \tag{1.1}$$

to minimise MSRE, provided the first two conditional inverse moments of Y given X are finite. They also noted that this Mean Squared Relative Error Predictor (MSREP) is always smaller than the Mean Squared Error Predictor (MSEP), $E(Y|X)$. By way of notation, write $r_\ell(x) = E(Y^{-\ell}|X = x)$ so that if we denote (1.1), when conditioned on $X = x$, by $g(x)$ then $g(x) = r_1(x)/r_2(x)$.

Park and Stefanski went on to consider parametric approaches to the estimation of g which centred on parametric estimators of the mean and variance functions of the inverse response. In this paper, we do not make any such parametric assumptions, but just assume smoothness. Therefore, we

introduce appropriate kernel–based smoothers to estimate g. For background in the case of MSE see, for example, Wand and Jones (1995), Fan and Gijbels (1996) or Simonoff (1996).

The general idea of local polynomial mean squared relative error prediction is described in Section 2, with particular emphasis on its local constant and local linear special cases. Asymptotic MSRE properties of the local constant and local linear MSREPs are given and discussed in Section 3 for both boundary and interior regions of the support of X. In Section 4, we develop a bandwidth selector that seems to work well; it is basically a 'rule-of-thumb' bandwidth selector but we found that quite a sophisticated version of that approach is necessary. Examples from the software quality literature are examined in Section 5. In Section 6, it is noted that the MSEP and MSREP can be obtained as special cases of a slightly more general framework concerning ratios of conditional expectations of functions of the response variable and it is in that framework that outline proofs of the results of Section 3 are provided. A positive alternative to the local linear MSREP is briefly described in Section 7.1 and a related problem mentioned in Section 7.2.

2. Local polynomial MSREP

Suppose we are given observations $\{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \Re$ and $Y_i \in \Re^+$. Introduce a symmetric probability density function K as kernel function which will act in a localising capacity. Associate with K a smoothing parameter, or bandwidth, h, using the notation $K_h(\cdot) = h^{-1}K(h^{-1}\cdot)$. Let $p_m(z) = \sum_{j=0}^m \beta_j z^j$ be a polynomial of degree m. Then, as kernel–localised local polynomial estimators of $r_{-1}(x) = E(Y|x)$ are defined as $\hat{\beta}_0 = \hat{\beta}_0(x)$

where $\hat{\beta}_0, ..., \hat{\beta}_m$ minimise

$$\sum_{i=1}^{n} K_h(X_i - x)\{Y_i - p_m(X_i - x)\}^2,$$

so kernel–localised local polynomial estimators of $g(x)$ are defined in the same way, except with the objective function changed to

$$\sum_{i=1}^{n} K_h(X_i - x)Y_i^{-2}\{Y_i - p_m(X_i - x)\}^2. \tag{2.1}$$

The $m = 0$ and $m = 1$ special cases of this will be of particular interest in this paper, but it is clear that higher–order local polynomials, particularly local quadratics and/or local cubics, could also be of value on occasion (Fan and Gijbels, 1996).

2.1. Local constant MSREP

When $m = 0$, the solution to (2.1) is the local constant estimator

$$\hat{g}_0(x) = \frac{\sum_{i=1}^{n} K_h(X_i - x)Y_i^{-1}}{\sum_{i=1}^{n} K_h(X_i - x)Y_i^{-2}}. \tag{2.2}$$

This is, of course, the direct analogue of the well known Nadaraya–Watson estimator in the MSE case. It is also perhaps the most obvious 'naive' kernel smoothing estimator of $g(x)$.

2.2. Local linear MSREP

Although the Nadaraya–Watson–type estimator is appealing in its simplicity, it will prove better to increase the polynomial order from $m = 0$ to $m = 1$. Define

$$t_\ell(x) = n^{-1}\sum_{i=1}^{n}(X_i - x)^\ell K_h(X_i - x)Y_i^{-1} \tag{2.3}$$

and

$$s_\ell(x) = n^{-1} \sum_{i=1}^{n} (X_i - x)^\ell K_h(X_i - x) Y_i^{-2}, \tag{2.4}$$

$\ell = 0, 1, 2$. The local linear estimator can then be written as

$$\hat{g}_1(x) = \frac{t_0(x)s_2(x) - t_1(x)s_1(x)}{s_0(x)s_2(x) - s_1^2(x)}. \tag{2.5}$$

Note also that $\hat{g}_0(x) = t_0(x)/s_0(x)$. The advantages of $\hat{g}_1(x)$ over $\hat{g}_0(x)$, which are parallel to those in the MSE case, are clarified in the case of MSRE by the asymptotic results of the following section.

3. MSRE of $\hat{g}_0$ and $\hat{g}_1$

3.1. MSRE in general

As essentially in Park and Stefanski (1998), it is easy to show that, for any estimator $\tilde{g}$ of g,

$$\text{MSRE}(\tilde{g}(X)) = E\left[\left\{\frac{Y - \tilde{g}(X)}{Y}\right\}^2 \mid X\right] = \left\{1 - \frac{r_1^2(X)}{r_2(X)}\right\} + r_2(X)E\{\tilde{g}(X) - g(X)\}^2. \tag{3.1}$$

By analogy with MSE prediction, the first term on the right–hand side of (3.1) is due to the extra uncertainty in a future value of Y over and above that accounted for by estimating its location, and the second term is due to uncertainty in estimation of the location; we can affect only the second term. In the MSRE case, this second term turns out to be a weighting function times the usual MSE. It is this second term, called MSRE$_-$ for short, that will be the sole focus of our investigations from here on.

3.2. Asymptotic MSE, MSRE

In Section 6, we note a general formulation that covers asymptotic performance of the local constant and local linear versions of both the MSEP

and MSREP, and we also provide the manipulations that lead to the results. The special cases of the results of Section 6 that pertain to the MSREP are given here.

Suppose that $X_1, ..., X_n$ is a random sample from a density f on [0,1] and that f' is continuous. We will consider biases and variances conditional on $X_1, ..., X_n$. The following assumptions will be made:

(i) K is symmetric about zero with finite support which we take to be $[-1, 1]$; other kernels such as the normal can also readily be dealt with.

(ii) $b_\ell(K) \equiv \int_{-1}^{1} z^\ell K(z)dz$ and $R(K) \equiv \int_{-1}^{1} K^2(z)dz$ are finite.

(iii) $r_\ell(x)$ exists for $\ell = 1, ..., 4$ and $r_1''(x)$, $r_2''(x)$, $r_3(x)$ and $r_4(x)$ are continuous on $x \in [0, 1]$.

(iv) $h = h(n) \to 0$ as $n \to \infty$ in such a way that $nh \to \infty$.

The n–dependent interior and boundary of the support of X are delineated by the points $x = h$ and $x = 1-h$. The first result concerns properties of $\hat{g}_0(x)$ and $\hat{g}_1(x)$ in the interior, under assumptions (i) to (iv).

Result 1. *(a) For* $h \le x \le 1 - h$,

$$E\{\hat{g}_0(x)\} \simeq g(x) + \frac{1}{2}h^2 b_2(K)\left(\frac{r_1''}{r_2} - \frac{r_1 r_2''}{r_2^2} + 2\frac{f'g'}{f}\right)(x),$$

$$V\{\hat{g}_0(x)\} \simeq \frac{R(K)V_g(x)}{f(x)nh}$$

and hence

$$\text{MSRE}_-\{\hat{g}_0(x)\} \simeq \frac{1}{4}h^4 b_2^2(K) r_2(x)\left(\frac{r_1''}{r_2} - \frac{r_1 r_2''}{r_2^2} + 2\frac{f'g'}{f}\right)^2(x) + \frac{r_2(x)R(K)V_g(x)}{f(x)nh}.$$

(b) For $h \le x \le 1 - h$,

$$E\{\hat{g}_1(x)\} \simeq g(x) + \frac{1}{2}h^2 b_2(K) g''(x),$$

$$V\{\hat{g}_1(x)\} \simeq \frac{R(K)V_g(x)}{f(x)nh}$$

and hence

$$\text{MSRE}_-\{\hat{g}_1(x)\} \simeq \frac{1}{4}h^4 b_2^2(K) r_2(x)\{g''(x)\}^2 + \frac{r_2(x)R(K)V_g(x)}{f(x)nh}.$$

Above,

$$V_g(x) = \frac{(r_2^3 - 2r_1r_2r_3 + r_1^2r_4)(x)}{r_2^4(x)}.$$

These results parallel those for the MSE case (Wand and Jones, 1995, p.125). In particular, the leading asymptotic bias of $\hat{g}_0(x)$ is a complicated function of derivatives of r_1 and r_2 and is also a function of the design density f. The equivalent term for $\hat{g}_1(x)$, however, is the much simpler, and perhaps more expected, second derivative of $g(x)$. The asymptotic variances of $\hat{g}_0$ and $\hat{g}_1$ are the same as is the order, h^2, of the bias term. Further parallelling the MSE case, the bandwidth h optimising the asymptotic MSRE is of order $n^{-1/5}$ and the resulting optimised asymptotic MSRE is of order $n^{-4/5}$.

The asymptotic performance of $\hat{g}_0(x)$ and $\hat{g}_1(x)$ near the boundary at 0, under assumptions (i) to (iv), is covered by Result 2; a similar result applies near the boundary at 1. For $0 \leq x < h$, write $x = ch$. Define $a_\ell(K;c) = \int_{-1}^{c} z^\ell K(z)dz$ and $R(K;c) = \int_{-1}^{c} K^2(z)dz$.

Result 2. *(a) For $x = ch$,*

$$E\{\hat{g}_0(x)\} \simeq g(x) - ha_1(K;c)g'(x),$$

$$V\{\hat{g}_0(x)\} \simeq \frac{R(K;c)V_g(x)}{f(x)nh}$$

and hence

$$\text{MSRE}_-\{\hat{g}_0(x)\} = O(h^2 + (nh)^{-1}).$$

(b) For $x = ch$,

$$E\{\hat{g}_1(x)\} \simeq g(x) + \frac{1}{2}h^2 b_2(K_4; c)g''(x),$$

$$V\{\hat{g}_1(x)\} \simeq \frac{R(K_4; c)V_g(x)}{f(x)nh}$$

and hence

$$\text{MSRE}_{-}\{\hat{g}_1(x)\} \simeq \frac{1}{4}h^4 b_2^2(K_4; c)r_2(x)\{g''(x)\}^2 + \frac{r_2(x)R(K_4)V_g(x)}{f(x)nh}.$$

In Result 2(b),

$$K_4(z) = \frac{a_2(K; c) - a_1(K; c)z}{a_2(K; c)a_0(K; c) - a_1^2(K; c)}K(z),$$

the usual fourth order kernel associated with local linear estimation (Wand and Jones, 1995, Section 5.6.1).

Notice that the local constant estimator has an order–of–magnitude increase in bias near the boundary, while the local linear estimator does not. It is this boundary behaviour that is perhaps the most important advantage of local linear over local constant estimation in practice. Note also that although the local linear boundary behaviour retains the interior's asymptotic rates for bias and variance, the variance is inflated somewhat as reflected in the constant term in the variance in Result 2(b).

4. Bandwidth selection

Our bandwidth selector, developed for the specific case of the local linear estimator, is, at once, pragmatic yet non-trivial. It is a form of 'rule-of-thumb' bandwidth selector, but one based on quite a complex pilot estimator. A number of unsuccessful attempts at simpler or 'more standard' approaches

to bandwidth selection based on the asymptotics of Section 3.2 preceded our development of the following.

To obtain a global bandwidth, we consider the asymptotic formula for $\int \mathrm{MSRE}_{-}\{\hat{g}_1(x)\}f(x)dx$ obtained from Result 1(b). This yields the asymptotically optimal bandwidth

$$h_0 = \left[\frac{R(K)\int V_g(x)dx}{b_2^2(K)\int\{g''(x)\}^2 f(x)dx\ n}\right]^{1/5}. \tag{4.1}$$

Now, purely for the purposes of rule-of-thumb bandwidth selection but in the general spirit of the problem at hand, take $Z_i = \log(Y_i)$ and consider the model $Z_i = \mu(X_i) + \epsilon_i$ where the ϵ's are i.i.d. normal errors with variance σ^2 and $\mu(x)$ is fitted as a 'blocked quartic' function, $\hat{\mu}(x)$. After taking logs, this follows precisely the recipe given for their pilot estimator by Ruppert, Sheather and Wand (1995, pp.1261-1262) including the use of Mallows' C_p for selection of the number of blocks and an estimate, $\hat{\sigma}$, of σ; the idea originates from Härdle and Marron (1995). Then, utilise $n^{-1}\sum_{i=1}^{n}\{\widehat{g''}(X_i)\}^2$ as estimator of $\int\{g''(x)\}^2 f(x)dx$ where $\widehat{g''}(x) = [\mu''(x)+\{\mu'(x)^2\}]\exp\{\mu(x)\}$. Also, a simple formula for $r_\ell(x)$, $\ell = 1, ..., 4$, follows immediately from the log–normal nature of the temporary model: $r_\ell(x) = \exp[-(\ell/2)\{2\mu(x) - \ell\sigma^2\}]$. In combination, this affords an estimate of $V_g(x)$ which is averaged over a uniformly spaced grid of values to provide an estimate of $\int V_g(x)dx$.

We have tried this bandwidth selector out with satisfactory consequences in a variety of simulated situations, not shown. However, we do not present a simulation study because we would have to operate under certain specific assumptions for the error structure, and we have eschewed such assumptions throughout this paper. We do, however, present estimates produced for data examples in the next section.

5. Examples

In this section, we consider two datasets on software quality for which relative error regression was considered appropriate by the investigators. The data, from Kitchenham and Pickard (1987), are as analysed by Khoshgoftaar et al. (1992b). The datasets refer to components of a computer operating system, referred to as "Subsystems 1 and 2", each made up of a number of software modules which take the role of experimental units. In each case, the response variable is the number of changes made to a module, both to mend faults and to extend functionality, after the subsystem was placed "under formal configuration control". Note that the response is, therefore, a count. The explanatory variables were measures of software complexity, specifically "Halstead's operator count η_1" for Subsystem 1 and "Halstead's operand count η_2" for Subsystem 2. The sample sizes are 27 and 40, respectively.

* * * Figs 1 and 2 about here * * *

The solid lines in Figs 1 and 2 show the local linear estimates $\hat{g}_1(x)$ with values of h chosen by the method of Section 4 as 10.27 in Fig. 1 and 17.55 in Fig. 2, respectively. They can be compared extremely favourably with the preferred straight line fits of Khosgoftaar et al., produced by a minimum absolute value of relative error procedure (dot-dashed lines). The nonparametric fits suggest a simple increasing but clearly non-linear form for g. The local constant estimates $\hat{g}_0(x)$ (dashed lines), admittedly using the same values of h which are not 'optimised' for them, also compare unfavourably with the local linear fits, with what appears to be domination by adverse boundary effects towards the right-hand sides of the figures. The fourth, dotted, lines on Figs 1 and 2 will be explained in Section 7.1. In each example we

used the normal density as the kernel function, K.

6. Generalisation and outline proof

Consider estimation of a ratio, $\gamma(x)$, of two functions which are each conditional expectations of some function of Y: $\gamma(x) = p(x)/q(x)$, where $p(x) = E\{P(Y)|x\}$ and $q(x) = E\{Q(Y)|x\}$, minimises

$$\int Q(y)\{S(y) - \theta\}^2 f(y|x)dy \tag{6.1}$$

where $S(y) = P(y)/Q(y)$ and $f(y|x)$ is the conditional density of Y given x. The natural data-based estimate of (6.1) is

$$n^{-1}\sum_{i=1}^{n} K_h(X_i - x)Q(Y_i)\{S(Y_i) - \theta\}^2,$$

and for the purposes of local polynomial estimation, we can replace θ by $p_m(X_i - x)$. It is then immediately clear that this formulation covers both the MSEP, for which $Q(y) = 1$, $S(y) = y$ and the MSREP, for which $Q(y) = y^{-2}$, $S(y) = y$.

The asymptotic MSE and/or MSRE$_-$ for local constant and local linear fitting can be obtained for this general case, and look essentially like Results 1 and 2, if we identify p with r_1, q with r_2 and γ with g, and in addition replace $V_g(x)$ by its more general formulation

$$V_\gamma(x) = \frac{\mathrm{Var}\{P(Y)q(x) - Q(Y)p(x)|X = x\}}{q^4(x)}. \tag{6.2}$$

It is also not then difficult to see that in the MSE case, the appropriate formulae reduce to the standard ones (Wand and Jones, 1995, Section 5).

We now give, in outline, the manipulations leading to Results 1 and 2 in this more general formulation. We will work with the boundary case, noting

that the interior results arise from the boundary ones by setting $c = 1$. Generalise (2.3) and (2.4) to

$$t_\ell(x) = n^{-1} \sum_{i=1}^{n} (X_i - x)^\ell K_h(X_i - x) P(Y_i) \tag{6.3}$$

and

$$s_\ell(x) = n^{-1} \sum_{i=1}^{n} (X_i - x)^\ell K_h(X_i - x) Q(Y_i), \tag{6.4}$$

respectively, $\ell = 0, 1, 2$.

6.1. Asymptotic bias

Clearly,

$$E(t_\ell(x)|X_1, ..., X_n) = n^{-1} \sum_{i=1}^{n} (X_i - x)^\ell K_h(X_i - x) p(X_i)$$

and this can be approximated by standard Taylor series expansions to yield

$$E\{t_\ell(x)\} \simeq h^\ell a_\ell(K; c)(pf)(x) - h^{\ell+1} a_{\ell+1}(K; c)(pf)'(x) + \frac{1}{2} h^{\ell+2} a_{\ell+2}(K; c)(pf)''(x), \tag{6.5}$$

and likewise for s_ℓ in terms of q.

For the local constant estimator, since $\hat{g}_0(x) = t_0(x)/s_0(x)$ we will use the standard approximation

$$\hat{g}_0(x) \simeq g(x) + (qf)(x)^{-1}\{t_0(x) - (pf)(x)\} - (q^2 f)(x)^{-1} p(x)\{s_0(x) - (qf(x)\}. \tag{6.6}$$

It follows that

$$\begin{aligned} E\{\hat{g}_0(x)\} &\simeq g(x) - h a_1(K; c) \left\{ \frac{(pf)'(x)}{qf(x)} - \frac{p(x)(qf)'(x)}{q^2 f(x)} \right\} \\ &+ \frac{1}{2} h^2 a_2(K; c) \left\{ \frac{(pf)''(x)}{qf(x)} - \frac{p(x)(qf)''(x)}{q^2 f(x)} \right\}. \end{aligned}$$

Noting that the multiplier of $ha_1(K;c)$ is $(p/q)'(x)$ completes the demonstration of the mean of $\hat{g}_0$ in the boundary case of Result 2(a); and with $a_1(K;1) = b_1(K) = 0$, the term of order h^2 produces the term of that order in the mean of $\hat{g}_0(x)$ in the interior, given in Result 1(a).

The remainder of this subsection concerns the asymptotic bias of the local linear estimator. Now $\hat{g}_1(x)$ is given in terms of the more general t's and s's by (2.5). Write this as $(w_3 - w_4)^{-1}(w_1 - w_2)$ where

$$\begin{aligned}
w_1 &= h^2 a_0(K;c)a_2(K;c)(pqf^2)(x) + h^2 a_2(K;c)(qf)(x)\{t_0(x) - a_0(K;c)(pf)(x))\} \\
&+ a_0(K;c)(pf)(x)\{s_2(x) - h^2 a_2(K;c)(qf)(x)\} \\
&+ \{t_0(x) - a_0(K;c)(pf)(x)\} \times \{s_2(x) - h^2 a_2(K;c)(qf)(x)\}, \\
w_2 &= h^2 a_1^2(K;c)(pqf^2)(x) + ha_1(K;c)(qf)(x)\{t_1(x) - ha_1(K;c)(pf)(x)\} \\
&+ ha_1(K;c)(pf)(x)\{s_1(x) - ha_1(K;c)(qf)(x)\} \\
&+ \{t_1(x) - ha_1(K;c)(pf)(x)\} \times \{s_1(x) - ha_1(K;c)(qf)(x)\},
\end{aligned}$$

and w_3 and w_4 are the same as w_1 and w_2, respectively, with t_i set to s_i and p set to q. Each w_i is thus written as $c_i + d_i$, where c_i is the first term in w_i and $d_i = w_i - c_i$ is the remaining, stochastic, term, so that

$$\hat{g}_1(x) \simeq \frac{(c_1 - c_2)}{(c_3 - c_4)} \left\{1 + \frac{d_1 - d_2}{c_1 - c_2}\right\} \left\{1 - \frac{d_3 - d_4}{c_3 - c_4} + \left(\frac{d_3 - d_4}{c_3 - c_4}\right)^2\right\}. \qquad (6.7)$$

Now, $(c_3 - c_4)^{-1}(c_1 - c_2) = (p/q)(x)$. Also, using (6.5) repeatedly, the expectation of (6.7) has a term of order h which can be shown to be made up of $\{a_0(K;c)a_2(K;c) - a_1^2(K;c)\}^{-1}(qf)^2(x)$ times

$$\begin{aligned}
&- a_1(K;c)a_2(K;c)(qf)(x)(pf)'(x) - a_0(K;c)a_3(K;c)(pf)(x)(qf)'(x) \\
&+ a_1(K;c)a_2(K;c)(qf)(x)(pf)'(x) + a_1(K;c)a_2(K;c)(pf)(x)(qf)'(x) \\
&- (p/q)(x)\{-a_1(K;c)a_2(K;c)(qf)(x)(qf)'(x) - a_0(K;c)a_3(K;c)(qf)(x)(qf)'(x)
\end{aligned}$$

$$+ \quad a_1(K;c)a_2(K;c)(qf)(x)(qf)'(x) + a_1(K;c)a_2(K;c)(qf)(x)(qf)'(x)\}$$

which is zero. Likewise, the order h^2 term is $\{a_0(K;c)a_2(K;c) - a_1^2(K;c)\}^{-1}\frac{1}{2}(qf)^2(x)$ times

$$\begin{aligned}
& a_2^2(K;c)(qf)(x)(pf)''(x) + a_0(K;c)a_4(K;c)(pf)(x)(qf)''(x) \\
+ \quad & 2a_1(K;c)a_3(K;c)(qf)'(x)(pf)'(x) - a_1(K;c)a_3(K;c)(qf)(x)(pf)''(x) \\
- \quad & a_1(K;c)a_3(K;c)(pf)(x)(qf)''(x) - 2a_2^2(K;c)(qf)'(x)(pf)'(x) \\
- \quad & (p/q)(x)\{a_2^2(K;c)(qf)(x)(qf)''(x) + a_0(K;c)a_4(K;c)(qf)(x)(qf)''(x) \\
+ \quad & 2a_1(K;c)a_3(K;c)(qf)'(x)^2 - a_1(K;c)a_3(K;c)(qf)(x)(qf)''(x) \\
- \quad & a_1(K;c)a_3(K;c)(qf)(x)(qf)''(x) - 2a_2^2(K;c)(qf)'(x)^2\}
\end{aligned}$$

plus other terms which come to zero in the manner above. This quantity reduces to $\{a_2^2(K;c) - a_1(K;c)a_3(K;c)\}$ times

$$(qf)(x)(pf)''(x) - 2(qf)'(x)(pf)'(x) - (qf)''(x)(pf)(x) + 2(p/q)(x)(qf)'(x)^2$$

which means the whole $O(h^2)$ bias term is

$$\frac{1}{2}\frac{a_2^2(K;c) - a_1(K;c)a_3(K;c)}{a_0(K;c)a_2(K;c) - a_1^2(K;c)}g''(x)$$

as claimed in Results 1(b) and 2(b).

6.2. Asymptotic variance

Now,

$$V\{t_\ell(x)|X_1,...,X_n\} = n^{-2}\sum_{i=1}^{n}(X_i - x)^{2\ell}K_h^2(X_i - x)V\{P(Y_i)\}$$

and this can be approximated by standard Taylor series expansions by

$$n^{-1}h^{2\ell-1}R(x^\ell K;c)V\{P(Y)|x\}f(x) \tag{6.8}$$

where $x^\ell K$ denotes the replacement of $K(z)$ by $z^\ell K(z)$. A similar expression holds for the variance of s_ℓ in terms of the conditional variance of $Q(Y)$. Likewise,

$$\begin{aligned} C\{t_k(x), s_m(x)|X_1, ..., X_n\} &= n^{-2}\sum_{i=1}^{n}(X_i - x)^{k+m}K_h^2(X_i - x)C\{P(Y_i), Q(Y_i)\} \\ &\simeq n^{-1}h^{k+m-1}R(x^{(k+m)/2}K; c)C\{P(Y), Q(Y)|x\}f(x). \quad (6.9) \end{aligned}$$

Using (6.6),

$$\begin{aligned} V\{\hat{g}_0(x)\} &\simeq (qf)^{-2}(x)V\{t_0(x)\} - 2(q^3f^2)^{-1}(x)p(x)C\{t_0(x), s_0(x)\} \\ &\quad + (q^2f)^{-2}(x)p^2(x)V\{s_0(x)\} \\ &\simeq (nh)^{-1}R(K; c)\{(q^2f)^{-1}(x)V\{P(Y|x)\} - 2(q^3f)^{-1}(x)p(x)C\{P(Y), Q(Y)|x)\} \\ &\quad + (q^4f)^{-1}(x)p^2(x)V\{Q(Y|x)\}\} \end{aligned}$$

which reduces to the quantity in Result 2 and thence to the quantity in Result 1 provided $V_g(x)$ is replaced by $V_\gamma(x)$ given at (6.2).

For the local linear estimator, use is made of (6.7). In particular, we find that

$$V\left(\frac{d_1 - d_2}{c_1 - c_2} - \frac{d_3 - d_4}{c_3 - c_4}\right) \simeq \frac{V_1}{V_2^2}$$

where

$$V_1 = V\left[h^2a_2(K; c)\{(qf/p)(x)t_0(x) - f(x)s_0(x)\} - ha_1(K; c)\{(qf/p)(x)t_1(x) - f(x)s_1(x)\}\right]$$

and

$$V_2 = h^2\{a_0(K; c)a_2(K; c) - a_1^2(K; c)\}qf^2(x).$$

Then (6.8) and (6.9) give the variance of $\hat{g}_1(x)$ as, approximately, $(nh)^{-1}$ times

$$g^2(x)R(K_4; c)\frac{(qf/p)^2(x)V\{P(Y)|x\} - 2(qf^2/p)(x)C\{P(Y), Q(Y)|x\} + f^2(x)V\{Q(Y)|x\}}{(q^2f^3)(x)}$$

which gives the expressions in Results 1(b) and 2(b) when $V_\gamma(x)$ is substituted for $V_g(x)$.

7. Alternatives

7.1. A positive local linear MSREP

The local constant estimator (2.2) is clearly positive when $Y_i > 0$, $i = 1, ..., n$, but the local linear estimator (2.5) does not respect the positivity constraint. Although this will rarely be a problem in practice, it may be thought desirable to provide an always positive alternative. A natural way to do this is to model g locally by an exponentiated polynomial. Specifically, choose a and b to minimise

$$\sum_{i=1}^{n} K_h(X_i - x) Y_i^{-2} [Y_i - \exp\{a + b(X_i - x)\}]^2 \tag{7.1}$$

and take $\hat{g}_P(x) = \exp(\hat{a})$ where $\hat{a}$ is the minimising value of a. Unfortunately, this estimator loses its explicitness and is considerably more difficult to compute. In fact, we have

$$\hat{g}_P(x) = \frac{\sum_{i=1}^n K_h(X_i - x) Y_i^{-1} \exp\{\hat{b}(X_i - x)\}}{\sum_{i=1}^n K_h(X_i - x) Y_i^{-2} \exp\{2\hat{b}(X_i - x)\}} \tag{7.2}$$

where $\hat{b}$ satisfies

$$\frac{\sum_{i=1}^n K_h(X_i - x) Y_i^{-1} \exp\{\hat{b}(X_i - x)\}}{\sum_{i=1}^n K_h(X_i - x) Y_i^{-2} \exp\{2\hat{b}(X_i - x)\}} = \frac{\sum_{i=1}^n (X_i - x) K_h(X_i - x) Y_i^{-1} \exp\{\hat{b}(X_i - x)\}}{\sum_{i=1}^n (X_i - x) K_h(X_i - x) Y_i^{-2} \exp\{2\hat{b}(X_i - x)\}}. \tag{7.3}$$

It is this positive local log-linear estimator that appeared in the form of dotted lines in Figs 1 and 2, with the same bandwidth values as for $\widehat{g}_1$. Its performance is clearly comparable with that of $\widehat{g}_1$ and use of either can be recommended.

7.2. Answering a different question

It is tempting to come up with further apparent alternatives to $\hat{g}_1$ such as taking logs of the Ys and fitting a local log polynomial function to them, prior to exponentiating the result. However, this would be answering a different question: using ordinary MSE for the fitting, the fitted curve would be estimating $\exp\{E(\log Y|x)\}$.

The context for which MSREP is designed is where it is natural to *assess* the quality of a predictor in terms of errors relative to the size of the response. This says nothing about the model for the assumed error structure about a regression function. The case of errors increasing with increasing response levels, but interest remaining centred on the ordinary regression mean function is quite different. It is for that case that a variety of further alternatives come to mind: for example, allowing a non–constant variance function in a normal errors context, fitting using a generalised smooth model framework (Fan, Heckman and Wand, 1995), taking logs (Eagleson and Müller, 1997).

References

Eagleson, G.K., Müller, H.G., 1997. Transformations for smooth regression models with multiplicative errors. J. Roy. Statist. Soc. Ser. B 59, 173–189.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman and Hall, London.

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

Farnum, N.R., 1990. Improving the relative error of estimation. Amer. Statist. 44, 288–289.

Härdle, W., Marron, J.S., 1995. Fast and simple scatterplot smoothing. Comput. Statist. Data Anal. 20, 1–17.

Khoshgoftaar, T.M., Bhattacharyya, B.B., Richardson, G.D., 1992a. Predicting software errors, during development, using nonlinear regression models: a comparative study. IEEE Trans. Reliab. 41, 390–395.

Khoshgoftaar, T.M., Munson, J.C., Bhattacharyya, B.B., Richardson, G.D., 1992b. Predicting modeling techniques of software quality from software measures. IEEE Trans. Soft. Eng. 18, 979–987.

Kitchenham, B., Pickard, L., 1987. Towards a constructive quality model. Part II: statistical techniques for modelling software quality in the ESPRIT REQUEST project. Soft. Eng. J. 2, 114–126.

Narula, S.C., Wellington, J.F., 1977. Prediction, linear regression and the minimum sum of relative errors. Technometrics 19, 185–190.

Park, H., Shin, K.I., 2006. A shrinked forecast in stationary processes favouring percentage error. J. Time Ser. Anal. 27, 129–139.

Park, H., Stefanski, L.A., 1998. Relative-error prediction. Statist. Probab. Lett. 40, 227–236.

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

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer, New York.

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

Fig. 1. Data (circles) on "Subsystem 1" concerning the number of changes to software modules plotted against Halstead's η_1 measure of software complexity. There are four fitted regression lines: $\widehat{g}_1(x)$ (solid); $\widehat{g}_0(x)$ (dashed); $\widehat{g}_P(x)$ (dotted); and a linear fit (dot-dashed) due to Khoshgoftaar et al. (1992b).

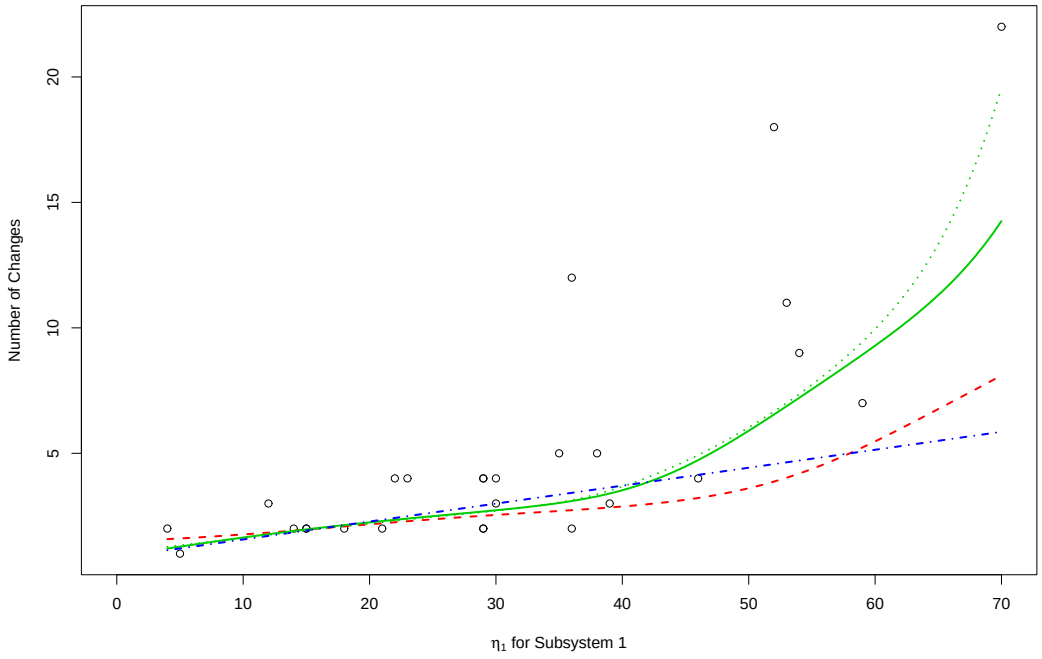

Fig. 2. As Fig. 1 except that the data concern "Subsystem 2" and the software complexity measure is Halstead's η_2.