Loader, Catherine

Working Paper

Smoothing: Local Regression Techniques

Papers, No. 2004,12

Provided in Cooperation with:
CASE - Center for Applied Statistics and Economics, Humboldt University Berlin

Suggested Citation: Loader, Catherine (2004) : Smoothing: Local Regression Techniques, Papers, No. 2004,12, Humboldt-Universität zu Berlin, Center for Applied Statistics and Economics (CASE), Berlin

This Version is available at:
https://hdl.handle.net/10419/22186

Smoothing: Local Regression Techniques

Catherine Loader[1]

Department of Statistics, Case Western Reserve University, Cleveland, OH 44106, USA. c@herine.net

Smoothing methods attempt to find functional relationships between different measurements. As in the standard regression setting, the data is assumed to consist of measurements of a response variable, and one or more predictor variables. Standard regression techniques (Chapter **??**) specify a functional form (such as a straight line) to describe the relation between the predictor and response variables. Smoothing methods take a more flexible approach, allowing the data points themselves to determine the form of the fitted curve.

This article begins by describing several different approaches to smoothing, including kernel methods, local regression, spline methods and orthogonal series. A general theory of linear smoothing is presented, which allows us to develop methods for statistical inference, model diagnostics and choice of smoothing parameters.

The theory is then extended to more general settings, including multivariate smoothing and likelihood models.

1 Smoothing

Given a dataset consisting of several variables and multiple observations, the goal of smoothing is to construct a functional relationship among the variables.

The most common situation for smoothing is that of a classical regression setting, where one assumes that observations occur in (predictor, response) pairs. That is, the available data has the form

$$\{(x_i, Y_i); i = 1, \ldots, n\},$$

where x_i is a measurement of the predictor (or independent) variable, and Y_i is the corresponding response. A functional model relating the variables takes the form

$$Y_i = \mu(x_i) + \epsilon_i, \tag{1}$$

where $\mu(x_i)$ is the mean function, and ϵ_i is a random error term. In classical regression analysis, one assumes a parametric form for the mean function; for example, $\mu(x) = a_0 + a_1 x$. The problem of estimating the mean function then reduces to estimating the coefficients a_0 and a_1.

The idea of smoothing methods is not to specify a parametric model for the mean function, but to allow the data to determine an appropriate functional form. Loosely stated, one assumes only that the mean function is smooth. Formal mathematical analysis may state the smoothness condition as a bound on derivatives of μ; for example, $|\mu''(x)| \leq M$ for all x and a specified constant M.

Section 2 describes some of the most important smoothing methods. These all fall into a class of linear smoothers, and Section 3 develops important properties, including bias and variance. These results are applied to derive statistical procedures, including bandwidth selection, model diagnostics and goodness-of-fit testing in Section 4. Multivariate smoothing, when there are multiple predictor variables, is discussed in Section 5. Finally, Section 5.2 discusses extensions to likelihood smoothing.

2 Linear Smoothing

In this section, some of the most common smoothing methods are introduced and discussed.

2.1 Kernel Smoothers

The simplest of smoothing methods is a kernel smoother. A point x is fixed in the domain of the mean function $\mu(\,\cdot\,)$, and a smoothing window is defined around that point. Most often, the smoothing window is simply an interval $(x-h, x+h)$, where h is a fixed parameter known as the *bandwidth*.

The kernel estimate is a weighted average of the observations within the smoothing window:

$$\hat{\mu}(x) = \frac{\sum_{i=1}^{n} W\left(\frac{x_i - x}{h}\right) Y_i}{\sum_{j=1}^{n} W\left(\frac{x_j - x}{h}\right)}, \tag{2}$$

where $W(\,\cdot\,)$ is a weight function. The weight function is chosen so that most weight is given to those observations close to the fitting point x. One common choice is the bisquare function,

$$W(x) = \begin{cases} (1-x^2)^2 & -1 \leq x \leq 1 \\ 0 & x > 1 \text{ or } x < -1 \end{cases}.$$

The kernel smoother can be represented as

$$\hat{\mu}(x) = \sum_{i=1}^{n} l_i(x) Y_i, \tag{3}$$

where the coefficients $l_i(x)$ are given by

$$l_i(x) = \frac{W\left(\frac{x_i - x}{h}\right)}{\sum_{j=1}^{n} W\left(\frac{x_j - x}{h}\right)}.$$

A *linear smoother* is a smoother that can be represented in the form (3) for appropriately defined weights $l_i(x)$. This linear representation leads to many nice statistical and computational properties, which will be discussed later.

The kernel estimate (2) is sometimes called the Nadaraya-Watson estimate (Nadaraya, 1964; Watson, 1964). Its simplicity makes it easy to understand and implement, and it is available in many statistical software packages. But its simplicity leads to a number of weaknesses, the most obvious of which is boundary bias. This can be illustrated through an example.

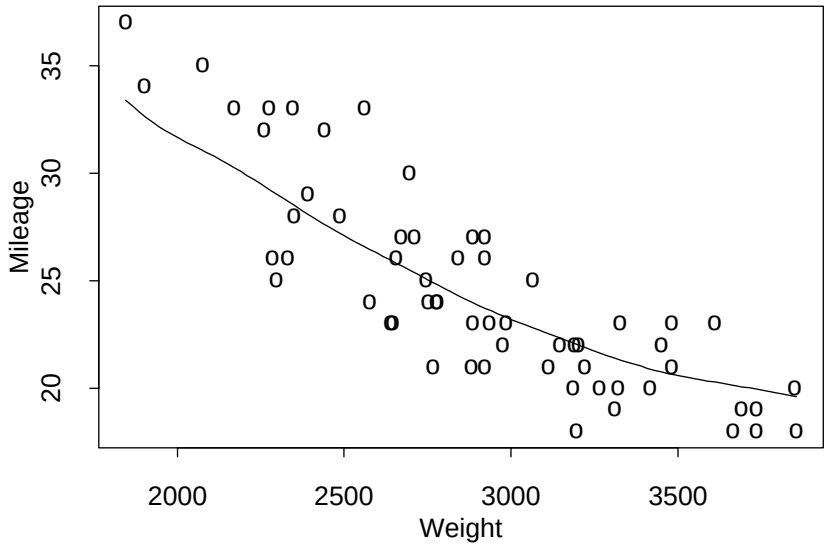

Fig. 1. Kernel smooth of the fuel economy dataset. The bisquare kernel is used, with bandwidth $h = 600$ pounds.

The fuel economy dataset consists of measurements of fuel usage (in miles per gallon) for sixty different vehicles. The predictor variable is the weight (in pounds) of the vehicle. Fig. 1 shows a scatterplot of the sixty data points, together with a kernel smooth. The smooth is constructed using the bisquare kernel and bandwidth $h = 600$ pounds.

Over much of the domain of Fig. 1, the smooth fit captures the main trend of the data, as required. But consider the left boundary region; in particular, vehicles weighing less than 2200 pounds. All these data points lie *above* the

fitted curve; the fitted curve will underestimate the economy of vehicles in this weight range. When the kernel estimate is applied at the left boundary (say, at Weight = 1800), all the data points used to form the average have Weight > 1800, and correspondingly slope of the true relation induces boundary bias into the estimate.

More discussion of this and other weaknesses of the kernel smoother can be found in Hastie and Loader (1993). Many modified kernel estimates have been proposed, but one obtains more parsimonious solutions by considering alternative estimation procedures.

2.2 Local Regression

Local regression estimation was independently introduced in several different fields in the late nineteenth and early twentieth century (Henderson, 1916; Schiaparelli, 1866). In the statistical literature, the method was independently introduced from different viewpoints in the late 1970's (Cleveland, 1979; Katkovnik, 1979; Stone, 1977). Books on the topic include Fan and Gijbels (1996) and Loader (1999b).

The underlying principle is that a smooth function can be well approximated by a low degree polynomial in the neighborhood of any point x. For example, a local linear approximation is

$$\mu(x_i) \approx a_0 + a_1(x_i - x) \tag{4}$$

for $x - h \le x_i \le x + h$. A local quadratic approximation is

$$\mu(x_i) \approx a_0 + a_1(x_i - x) + \frac{a_2}{2}(x_i - x)^2.$$

The local approximation can be fitted by locally weighted least squares. A weight function and bandwidth are defined as for kernel regression. In the case of local linear regression, coefficient estimates $\hat{a}_0, \hat{a}_1$ are chosen to minimize

$$\sum_{i=1}^{n} W\left(\frac{x_i - x}{h}\right)(Y_i - (a_0 + a_1(x_i - x)))^2. \tag{5}$$

The local linear regression estimate is defined as

$$\hat{\mu}(x) = \hat{a}_0. \tag{6}$$

Each local least squares problem defines $\hat{\mu}(x)$ at one point x; if x is changed, the smoothing weights $W\left(\frac{x_i - x}{h}\right)$ change, and so the estimates $\hat{a}_0$ and $\hat{a}_1$ change.

Since (5) is a weighted least squares problem, one can obtain the coefficient estimates by solving the normal equations

$$\mathbf{X}^\top \mathbf{W}\left(Y - \mathbf{X}\begin{pmatrix}\hat{a}_0 \\ \hat{a}_1\end{pmatrix}\right) = 0, \tag{7}$$

where $\mathbf{X}$ is the *design matrix*:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 - x \\ \vdots & \vdots \\ 1 & x_n - x \end{pmatrix}$$

for local linear regression, $\mathbf{W}$ is a diagonal matrix with entries $W\left(\frac{x_i - x}{h}\right)$ and $Y = (\, Y_1 \quad \dots \quad Y_n \,)^\top$.

When $\mathbf{X}^\top \mathbf{W} \mathbf{X}$ is invertible, one has the explicit representation

$$\begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} Y. \tag{8}$$

This shows that the local regression estimate is a linear estimate, as defined by (3). Explicitly, the coefficients $l_i(x)$ are given by

$$l(x)^\top = (\, l_1(x) \quad \dots \quad l_n(x) \,) = e_1^\top (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}, \tag{9}$$

where $e_1^\top$ is the unit vector $(\, 1 \quad 0 \,)$.

For local quadratic regression and higher order fits, one simply adds additional columns to the design matrix $\mathbf{X}$ and vector $e_1^\top$.

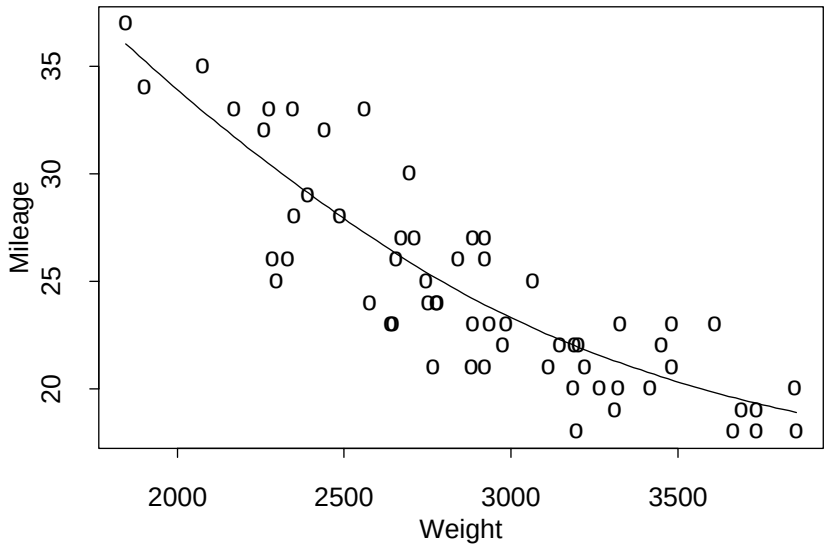

Fig. 2. Local Linear Regression fitted to the fuel economy dataset. A bandwidth $h = 1000$ pounds is used.

Fig. 2 shows a local linear regression fit to the fuel economy dataset. This has clearly fixed the boundary bias problem observed in Fig. 1. With the

reduction in boundary bias, it is also possible to substantially increase the bandwidth, from $h = 600$ pounds to $h = 1000$ bounds. As a result, the local linear fit is using much more data, meaning the estimate has less noise.

2.3 Penalized Least Squares (Smoothing Splines)

An entirely different approach to smoothing is through optimization of a penalized least squares criterion, such as

$$\sum_{i=1}^{n}(Y_i - \mu(x_i))^2 + \lambda \int \mu''(x)^2 dx, \tag{10}$$

where λ is specified constant. This criterion trades off fidelity to the data (measured by the residual sum-of-squares) versus roughness of the mean function (measured by the penalty term). The penalized least squares method chooses $\hat{\mu}$ from the class of twice differentiable functions to minimize the penalized least squares criterion.

The solution to this optimization problem is a piecewise polynomial, or spline function, and so penalized least squares methods are also known as smoothing splines. The idea was first considered in the early twentieth century (Whitaker, 1923). Modern statistical literature on smoothing splines began with work including Wahba and Wold (Wahba and Wold, 1975) and Silverman (Silverman, 1985). Books devoted to spline smoothing include Green and Silverman (1994) and Wahba (1990).

Suppose the data are ordered; $x_i \leq x_{i+1}$ for all i. Let $\hat{a}_i = \hat{\mu}(x_i)$, and $\hat{b}_i = \hat{\mu}'(x_i)$, for $i = 1, \ldots, n$. Given these values, it is easy to show that between successive data points, $\hat{\mu}(x)$ must be the unique cubic polynomial interpolating these values:

$$\hat{\mu}(x) = a_i\phi_0(u) + b_i\Delta_i\psi_0(u) + a_{i+1}\phi_1(u) + b_{i+1}\Delta_i\psi_1(u),$$

where $\Delta_i = x_{i+1} - x_i$; $u = (x - x_i)/\Delta_i$ and

$$\begin{aligned}
\phi_0(u) &= 1 - u^2(3 - 2u) \\
\psi_0(u) &= u(1 - u(2 - u)) \\
\phi_1(u) &= u^2(3 - 2u) \\
\psi_1(u) &= u^2(u - 1).
\end{aligned}$$

Letting $\alpha^\top = (\, a_1 \quad b_1 \quad \ldots \quad a_n \quad b_n \,)$, the penalty term $\int \mu''(x)^2 dx$ is a quadratic function of the parameters, and so (10) can be written as

$$\|Y - \mathbf{X}\alpha\|^2 + \lambda\alpha^\top \mathbf{M}\alpha,$$

for appropriate matrices $\mathbf{M}$ and $\mathbf{X}$. The parameter estimates are given by

$$\hat{\alpha} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{M})^{-1}\mathbf{X}^\top Y.$$

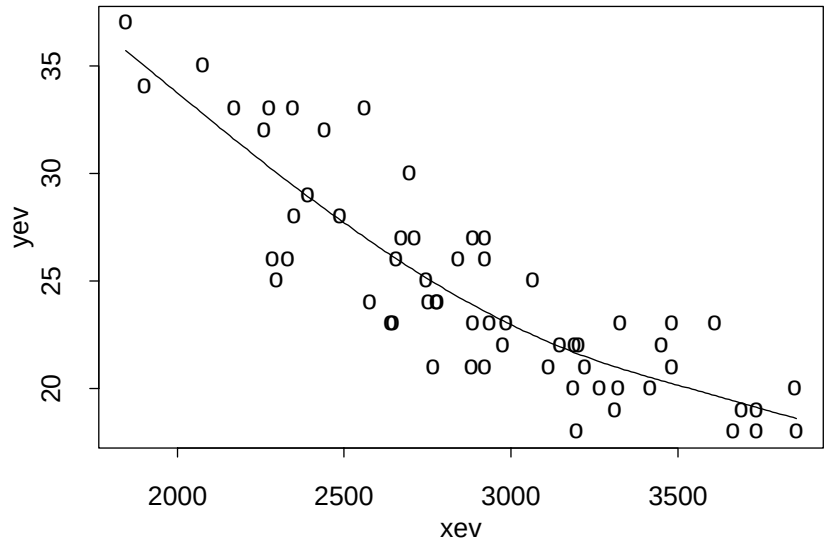

Fig. 3. Smoothing Spline fitted to the fuel economy dataset. The penalty is $\lambda = 1.5 \times 10^8$ pounds^3.

Fig. 3 shows a smoothing spline fitted to the fuel economy dataset. Clearly, the fit is very similar to the local regression fit in Fig. 2. This situation is common for smoothing problems with a single predictor variable; with comparably chosen smoothing parameters, local regression and smoothing spline methods produce similar results. On the other hand, kernel methods can struggle to produce acceptable results, even on relatively simple datasets.

2.4 Regression Splines

Regression splines begin by choosing a set of knots (typically, much smaller than the number of data points), and a set of basis functions spanning a set of piecewise polynomials satisfying continuity and smoothness constraints.

Let the knots be $v_1 < \ldots < v_k$ with $v_1 = \min(x_i)$ and $v_k = \max(x_i)$. A linear spline basis is

$$f_j(x) = \begin{cases} \frac{x - v_{j-1}}{v_j - v_{j-1}} & v_{j-1} \leq x \leq v_j \\ \frac{v_{j+1} - x}{v_{j+1} - v_j} & v_j < x \leq v_{j+1} \\ 0 & \text{otherwise} \end{cases} ;$$

note that these functions span the space of piecewise linear functions with knots at $v_1, \ldots, v_k$. The piecewise linear spline function is constructed by regressing the data onto these basis functions.

The linear spline basis functions have discontinuous derivatives, and so the resulting fit may have a jagged appearance. It is more common to use piecewise

cubic splines, with the basis functions having two continuous derivatives. See Chapter 3 of Ruppert et al. (2003) for a more detailed discussion of regression splines and basis functions.

2.5 Orthogonal Series

Orthogonal series methods represent the data with respect to a series of orthogonal basis functions, such as sines and cosines. Only the low frequency terms are retained. The book Efromovich (1999) provides a detailed discussion of this approach to smoothing.

Suppose the x_i are equally spaced; $x_i = i/n$. Consider the basis functions

$$f_\omega(x) = a_\omega \cos(2\pi\omega x);\ \ \omega = 0, 1, \ldots, \lfloor n/2 \rfloor$$
$$g_\omega(x) = b_\omega \sin(2\pi\omega x);\ \ \omega = 1, \ldots, \lfloor (n-1)/2 \rfloor,$$

where the constants a_ω, b_ω are chosen so that $\sum_{i=1}^n f_\omega(x_i)^2 = \sum_{i=1}^n g_\omega(x_i)^2 = 1$. Then the regression coefficients are

$$c_\omega = \sum_{i=1}^n f_\omega(x_i) Y_i$$
$$s_\omega = \sum_{i=1}^n g_\omega(x_i) Y_i$$

and the corresponding smooth estimate is

$$\hat{\mu}(x) = \sum_\omega h(\omega) \left(c_\omega f_\omega(x) + s_\omega g_\omega(x) \right).$$

Here, $h(\omega)$ is chosen to 'damp' high frequencies in the observations; for example,

$$h(\omega) = \begin{cases} 1 & \omega \le \omega_0 \\ 0 & \omega > \omega_0 \end{cases}$$

is a low-pass filter, passing all frequencies less than or equal to ω_0.

Orthogonal series are widely used to model time series, where the coefficients c_ω and s_ω may have a physical interpretation: non-zero coefficients indicate the presence of cycles in the data. A limitation of orthogonal series approaches is that they are more difficult to apply when the x_i are not equally spaced.

3 Statistical Properties of Linear Smoothers

Each of the smoothing methods discussed in the previous section has one or more 'smoothing parameters' that control the amount of smoothing being performed. For example, the bandwidth h in the kernel smoother or local

regression methods, and the parameter λ in the penalized likelihood criterion. In implementing the smoothers, the first question to be asked is how should the smoothing parameters be chosen? More generally, how can the performance of a smoother with given smoothing parameters be assessed? A deeper question is in comparing fits from different smoothers. For example, we have seen for the fuel economy dataset that a local linear fit with $h = 1000$ (Fig. 2) produces a fit similar to a smoothing spline with $\lambda = 1.5 \times 10^8$ (Fig. 3). Somehow, we want to be able to say these two smoothing parameters are equivalent.

As a prelude to studying methods for bandwidth selection and other statistical inference procedures, we must first study some of the properties of linear smoothers. We can consider measures of goodness-of-fit, such as the mean squared error,

$$\mathrm{MSE}(x) = E\left((\hat{\mu}(x) - \mu(x))^2\right) = \mathrm{var}(\hat{\mu}(x)) + \mathrm{bias}(\hat{\mu}(x))^2,$$

where $\mathrm{bias}(\hat{\mu}(x)) = E(\hat{\mu}(x)) - \mu(x)$.

Intuitively, as the bandwidth h increases, more data is used to construct the estimate $\hat{\mu}(x)$, and so the variance $\mathrm{var}(\hat{\mu}(x))$ decreases. On the other hand, the local polynomial approximation is best over small intervals, so we expect the bias to increase as the bandwidth increases. Choosing h is a tradeoff between small bias and small variance, but we need more precise characterizations to derive and study selection procedures.

3.1 Bias

The bias of a linear smoother is given by

$$E(\hat{\mu}(x)) - \mu(x) = \sum_{i=1}^{n} l_i(x) E(Y_i) - \mu(x) = \sum_{i=1}^{n} l_i(x)\mu(x_i) - \mu(x). \qquad (11)$$

As this depends on the unknown mean function $\mu(x)$, it is not very useful by itself, although it may be possible to estimate the bias by substituting an estimate for $\mu(x)$. To gain more insight, approximations to the bias are derived. The basic tools are

1. A low order Taylor series expansion of $\mu(\,\cdot\,)$ around the fitting point x.
2. Approximation of the sums by integrals.

For illustration, consider the bias of the local linear regression estimate defined by (6). A three-term Taylor series gives

$$\mu(x_i) = \mu(x) + (x_i - x)\mu'(x) + \frac{(x_i - x)^2}{2}\mu''(x) + o(h^2)$$

for $|x_i - x| \le h$. Substituting this into (11) gives

$$E(\hat{\mu}(x)) - \mu(x) = \mu(x)\sum_{i=1}^{n} l_i(x) + \mu'(x)\sum_{i=1}^{n}(x_i - x)l_i(x) + \frac{\mu''(x)}{2}\sum_{i=1}^{n}(x_i - x)^2 l_i(x) - \mu(x) + o(h^2).$$

For local linear regression, it can be shown that

$$\sum_{i=1}^{n} l_i(x) = 1$$
$$\sum_{i=1}^{n}(x_i - x)l_i(x) = 0.$$

This is a mathematical statement of the heuristically obvious property of the local linear regression: if data Y_i fall on a straight line, the local linear regression will reproduce that line. See Loader (1999b), p37, for a formal proof. With this simplification, the bias reduces to

$$E(\hat{\mu}(x)) - \mu(x) = \frac{\mu''(x)}{2}\sum_{i=1}^{n}(x_i - x)^2 l_i(x) + o(h^2). \tag{12}$$

This expression characterizes the dependence of the bias on the mean function: the dominant term of the bias is proportional to the second derivative of the mean function.

The next step is to approximate summations by integrals, both in (12) and in the matrix equation (9) defining $l_i(x)$. This leads to

$$E(\hat{\mu}(x)) - \mu(x) \approx \mu''(x)h^2\frac{\int v^2 W(v)dv}{2\int W(v)dv}. \tag{13}$$

In addition to the dependence on $\mu''(x)$, we now see the dependence on h: as the bandwidth h increases, the bias increases quadratically with the bandwidth.

Bias expansions like (13) are derived much more generally by Ruppert and Wand (1994); their results cover arbitrary degree local polynomials and multidimensional fits also. Their results imply that when p, the degree of the local polynomial, is odd, the dominant term of the bias is proportional to $h^{p+1}\mu^{(p+1)}(x)$. When p is even, the first-order term can disappear, leading to bias of order h^{p+2}.

3.2 Variance

To derive the variance of a linear smoother, we need to make assumptions about the random errors ϵ_i in (1). The most common assumption is that the

errors are independent and identically distributed, with variance $\mathrm{var}(\epsilon_i) = \sigma^2$. The variance of a linear smoother (3) is

$$\mathrm{var}(\hat{\mu}(x)) = \sum_{i=1}^{n} l_i(x)^2 \mathrm{var}(Y_i) = \sigma^2 \|l(x)\|^2. \tag{14}$$

As with bias, informative approximations to the variance can be derived by replacing sums by integrals. For local linear regression, this leads to

$$\mathrm{var}(\hat{\mu}(x)) \approx \frac{\sigma^2}{nhf(x)} \frac{\int W(v)^2 dv}{(\int W(v) dv)^2}, \tag{15}$$

where $f(x)$ is the density of the design points x_i. The dependence on the sample size, bandwidth and design density through $1/(nhf(x))$ is universal, holding for any degree of local polynomial. The term depending on the weight function varies according to the degree of local polynomial, but generally increases as the degree of the polynomials increases. See Ruppert and Wand (1994) for details.

3.3 Degrees of Freedom

Under the model (1) the observation Y_i has variance σ^2, while the estimate $\hat{\mu}(x_i)$ has variance $\sigma^2 \|l(x_i)\|^2$. The quantity $\|l(x_i)\|^2$ measures the variance reduction of the smoother at a data point x_i. At one extreme, if the 'smoother' interpolates the data, then $\hat{\mu}(x_i) = Y_i$ and $\|l(x_i)\|^2 = 1$. At the other extreme, if $\hat{\mu}(x_i) = \bar{Y}$, $\|l(x_i)\|^2 = 1/n$. Under mild conditions on the weight function, a local polynomial smoother satisfies

$$\frac{1}{n} \leq \|l(x_i)\|^2 \leq 1,$$

and $\|l(x_i)\|^2$ is usually a decreasing function of the bandwidth h.

A global measure of the amount of smoothing is provided by

$$\nu_2 = \sum_{i=1}^{n} \|l(x_i)\|^2.$$

This is one definition of the 'degrees of freedom' or 'effective number of parameters' of the smoother. It satisfies the inequalities

$$1 \leq \nu_2 \leq n.$$

An alternative representation of ν_2 is as follows. Let $\mathbf{H}$ be the 'hat matrix', which maps the data to fitted values:

$$\begin{pmatrix} \hat{\mu}(x_1) \\ \vdots \\ \hat{\mu}(x_n) \end{pmatrix} = \mathbf{H}Y.$$

For a linear smoother, $\mathbf{H}$ has rows $l(x_i)^\top$, and $\nu_2 = \text{trace}(\mathbf{H}^\top\mathbf{H})$.

The diagonal elements of $\mathbf{H}$, $l_i(x_i)$ provide another measure of the amount of smoothing at x_i. If the smooth interpolates the data, then $l(x_i)$ is the corresponding unit vector with $l_i(x_i) = 1$. If the smooth is simply the global average, $l_i(x_i) = 1/n$. The corresponding definition of degrees of freedom is

$$\nu_1 = \sum_{i=1}^{n} l_i(x_i) = \text{trace}(\mathbf{H}).$$

For a least-squares fit, the hat matrix is a perpendicular projection operator, which is symmetric and idempotent. In this case, $\mathbf{H} = \mathbf{H}^\top\mathbf{H}$, and $\nu_1 = \nu_2$. For linear smoothers, the two definitions of degrees-of-freedom are usually not equal, but they are often of similar magnitude.

For the local linear regression in Fig. 2, the degrees of freedom are $\nu_1 = 3.54$ and $\nu_2 = 3.09$. For the smoothing spline smoother in Fig. 3, $\nu_1 = 3.66$ and $\nu_2 = 2.98$. By either measure the degrees of freedom are similar for the two fits. The degrees of freedom provides a mechanism by which different smoothers, with different smoothing parameters, can be compared: we simply choose smoothing parameters producing the same number of degrees of freedom. More extensive discussion of the degrees of freedom of a smoother can be found in Cleveland and Devlin (1988) and Hastie and Tibshirani (1990).

Variance Estimation

The final component needed for many statistical procedures is an estimate of the error variance σ^2. One such estimate is

$$\hat{\sigma}^2 = \frac{1}{n - 2\nu_1 + \nu_2} \sum_{i=1}^{n} (Y_i - \hat{\mu}(x_i))^2. \tag{16}$$

The normalizing constant is chosen so that if the bias of $\hat{\mu}(x_i)$ is neglected, $\hat{\sigma}^2$ is unbiased. See Cleveland and Devlin (1988).

4 Statistics for Linear Smoothers: Bandwidth Selection and Inference

We also want to perform statistical inference based on the smoothers. As for parametric regression, we want to construct confidence bands and prediction intervals based on the smooth curve. Given a new car that weighs 2800 pounds, what is its fuel economy? Tests of hypotheses can also be posed: for example,

is the curvature observed in Fig. 2 significant, or would a linear regression be adequate? Given different classifications of car (compact, sporty, minivan e.t.c.) is there differences among the categories that cannot be explained by weight alone?

4.1 Choosing Smoothing parameters

All smoothing methods have one or more smoothing parameters: parameters that control the 'amount' of smoothing being performed. For example, the bandwidth h in the kernel and local regression estimates. Typically, bandwidth selection methods are based on an estimate of some goodness-of-fit criterion. Bandwidth selection is a special case of model selection, discussed more deeply in Chapter **??**.

How should smoothing parameters be used? At one extreme, there is full automation: optimization of the goodness-of-fit criterion produces a single 'best' bandwidth. At the other extreme is purely exploratory and graphical methods, using goodness-of-fit as a guide to help choose the best method.

Automation has the advantage that it requires much less work; a computer can be programmed to perform the optimization. But the price is a lack of reliability: fits with very different bandwidths can produce similar values of the goodness-of-fit criterion. The result is either high variability (producing fits that look undersmoothed) or high bias (producing fits that miss obvious features in the data).

Cross Validation

Cross validation (CV) focuses on the prediction problem: if the fitted regression curve is used to predict new observations, how good will the prediction be? If a new observation is made at $x = x_0$, and the response Y_0 is predicted by $\hat{Y}_0 = \hat{\mu}(x_0)$, what is the prediction error? One measure is

$$E((Y_0 - \hat{Y}_0)^2).$$

The method of CV can be used to estimate this quantity. In turn, each observation (x_i, Y_i) is omitted from the dataset, and is 'predicted' by smoothing the remaining $n - 1$ observations. This leads to the CV score

$$\mathrm{CV}(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_{-i}(x_i))^2, \tag{17}$$

where $\hat{\mu}_{-i}(\,\cdot\,)$ denotes the smoothed estimate when the single data point (x_i, Y_i) are omitted from the dataset; only the remaining $n - 1$ data points are used to compute the estimate.

Formally computing each of the leave-one-out regression estimates $\hat{\mu}_{-i}(\,\cdot\,)$ would be highly computational, and so at a first glance computation of the

CV score (17) looks prohibitively expensive. But there is a remarkable simplification, valid for nearly all common linear smoothers (and all those discussed in Section 2):

$$\hat{\mu}_{-i}(x_i) = \frac{\hat{\mu}(x_i) - l_i(x_i)Y_i}{1 - l_i(x_i)}.$$

With this simplification, the CV criterion becomes

$$\mathrm{CV}(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}(x_i))^2}{(1 - l_i(x_i))^2}.$$

Generalized cross validation (GCV) replaces each of the influence values $l_i(x_i)$ by the average, ν_1/n. This leads to

$$\mathrm{GCV}(\hat{\mu}) = n\frac{\sum_{i=1}^{n}(Y_i - \hat{\mu}(x_i))^2}{(n - \nu_1)^2}.$$

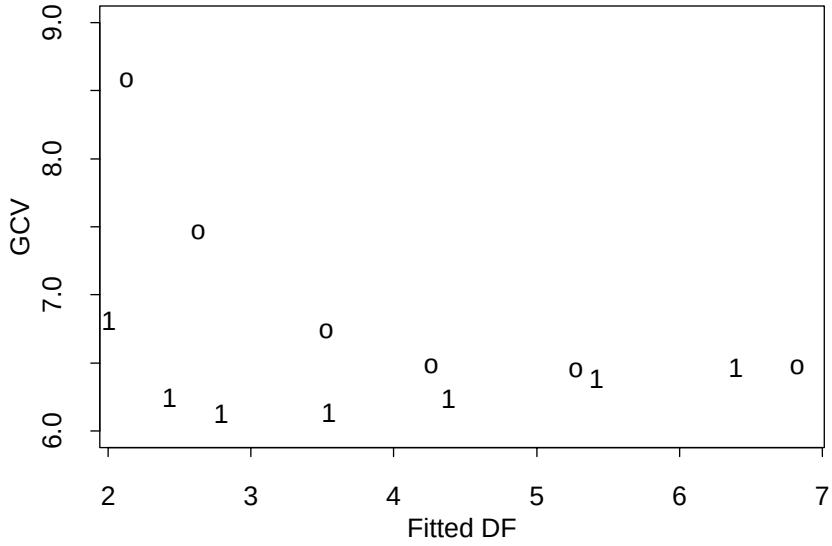

Fig. 4. GCV scores for the fuel economy dataset. Points marked 0 are for kernel smoothers with a range of bandwidths h, and points markes 1 are for a local linear smoother.

Fig. 4 shows the GCV scores for the fuel economy dataset, and using kernel and local linear smoothers with a range of bandwidths. Note the construction of the plot: the fitted degrees of freedom ν_1 are used as the x axis. This allows us to meaningfully superimpose and compare the GCV curves arising from different smoothing methods. From right to left, the points marked '0' represent a

kernel smoother with $h = 300, 400, 500, 600, 800$ and 1000, and points marked '1' represent a local linear smoother with $h = 400, 500, 700, 1000, 1500, 2000$ and ∞.

The interpretation of Fig. 4 is that for any fixed degrees of freedom, the local linear fit outperforms the kernel fit. The best fits obtained are the local linear, with 3 to 3.5 degrees of freedom, or h between 1000 and 1500.

Unbiased Risk Estimation

A risk function measures the distance between the true regression function and the estimate; for example,

$$R(\mu, \hat{\mu}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} E\left((\hat{\mu}(x_i) - \mu(x_i))^2\right). \tag{18}$$

Ideally, a good estimate would be one with low risk. But since μ is unknown, $R(\mu, \hat{\mu})$ cannot be evaluated directly.

Instead, the risk must be estimated. An unbiased estimate is

$$\hat{R}(\mu, \hat{\mu}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \hat{\mu}(x_i))^2 - n + 2\nu_1$$

(Mallows, 1973; Cleveland and Devlin, 1988). The unbiased risk estimate is equivalent to Akaike's Information Criterion (Akaike, 1972, 1974). To implement the unbiased risk estimate one needs to substitute an estimate for σ^2; Cleveland and Devlin recommend using (16) with a small bandwidth.

The unbiased risk estimate can be used similarly to GDV. One computes $\hat{R}(\mu, \hat{\mu})$ for a range of different fits $\hat{\mu}$, and plots the resulting risk estimates versus the degrees of freedom. Fits producing a small risk estimate are considered best.

Bias Estimation and Plug-in Methods

An entirely different class of bandwidth selection methods, often termed plug-in methods, attempt to directly estimate a risk measure by estimating the bias and variance. The method has been developed mostly in the context of kernel density estimation, but adaptations to kernel regression and local polynomial regression can be found in Fan and Gijbels (1995) and Ruppert et al. (1995).

Again focusing on the squared-error risk, we have the bias-variance decomposition

$$\begin{aligned} \sigma^2 R(\mu, \hat{\mu}) &= \sum_{i=1}^{n} \text{bias}(\hat{\mu}(x_i))^2 + \sum_{i=1}^{n} \text{var}(\hat{\mu}(x_i)) \\ &= \sum_{i=1}^{n} \left(\sum_{j=1}^{n} l_j(x_i)\mu(x_j) - \mu(x_i) \right)^2 + \sigma^2 \sum_{i=1}^{n} \|l(x_i)\|^2. \end{aligned} \tag{19}$$

A plug-in estimate begins by constructing a preliminary *pilot estimate* of the mean function $\mu(\,\cdot\,)$. This is then substituted into the risk estimate (19), which can then be minimized over the bandwidth h.

There are many variants of the plug-in idea in the statistics literature. Most simplify the risk function using asymptotic approximations such as (13) and (15) for the bias and variance; making these substitutions in (19) gives

$$\sigma^2 R(\mu,\hat{\mu}) \approx h^4 \left(\frac{\int v^2 W(v)dv}{2\int W(v)dv}\right)^2 \sum_{i=1}^{n} \mu''(x_i)^2 + \frac{\sigma^2}{nh}\frac{\int W(v)^2 dv}{(\int W(v)dv)^2}\sum_{i=1}^{n}\frac{1}{f(x_i)}.$$

If the design points are uniformly distributed on an interval $[a,b]$ say, then approximating the sums by integrals gives

$$\sigma^2 R(\mu,\hat{\mu}) \approx nh^4 \left(\frac{\int v^2 W(v)dv}{2\int W(v)dv}\right)^2 \frac{1}{b-a}\int_a^b \mu''(x)^2 dx + \frac{(b-a)\sigma^2}{h}\frac{\int W(v)^2 dv}{(\int W(v)dv)^2}.$$

Minimizing this expression over h yields an asymptotically optimal bandwidth:

$$h_{\text{opt}}^5 = \frac{\sigma^2 (b-a)^2 \int W(v)^2 dv}{n(\int v^2 W(v)dv)^2 \int_a^b \mu''(x)^2 dx}.$$

Evaluation of h_{opt} requires substitution of estimates for $\int_a^b \mu''(x)^2 dx$ and of σ^2. The estimate (16) can be used to estimate σ^2, but estimating $\int_a^b \mu''(x)^2 dx$ is more problematic. One technique is to estimate the second derivative using a 'pilot' estimate of the smooth, and then use the estimate

$$\int_a^b \hat{\mu}''(x)^2 dx.$$

If a local quadratic estimate is used at the pilot stage, the curvature coefficient $\hat{a}_2$ can be used as an estimate of $\mu''(x)$.

But the use of a pilot estimate to estimate the second derivative is problematic. The pilot estimate itself has a bandwidth that has to be selected, and the estimated optimal bandwidth $\hat{h}_{\text{opt}}$ is highly sensitive to the choice of pilot bandwidth. Roughly, if the pilot estimate smooths out important features of μ, so will the estimate $\hat{\mu}$ with bandwidth $\hat{h}_{\text{opt}}$. More discussion of this point may be found in Loader (1999a).

4.2 Normal-based inference

Inferential procedures for smoothers include the construction of confidence bands for the true mean function, and procedures to test the adequacy of simpler models. In this section, some of the main ideas are briefly introduced; more extensive discussion can be found in the books Azzalini and Bowman (1997), Härdle (1990), Hart (1997) and Loader (1999b).

Confidence intervals

If the errors ϵ_i are normally distributed, then confidence intervals for the true mean can be constructed as

$$\hat{\mu}(x) \pm c\hat{\sigma}\|l(x)\|.$$

The constant c can be chosen from the Student's t distribution with degrees of freedom equal to $n - 2\nu_1 + \nu_2$ (alternative choices are discussed below in the context of testing). These confidence intervals are pointwise intervals for $E(\hat{\mu}(x))$:

$$P(|\hat{\mu}(x) - E(\hat{\mu}(x))| < c\hat{\sigma}\|l(x)\|) = 1 - \alpha.$$

To construct confidence intervals for $\mu(x)$, one must either choose the bandwidth sufficiently small so that the bias can be ignored, or explicitly estimate the bias. The latter approach suffers from the same weaknesses observed in plug-in bandwidth selection.

Tests of Hypothesis

Consider the problem of testing for the adequacy of a linear model. For example, in the fuel economy dataset of Figs. 1 and 2, one may be interested in knowing whether a linear regression, $\mu(x) = a + bx$ is adequate, or alternatively whether the departure from linearity indicated by the smooth is significant. This hypothesis testing problem can be stated as

$$\begin{aligned} H_0 &: \mu(x) = a + bx \text{ for some } a, b \\ H_1 &: \text{otherwise.} \end{aligned}$$

In analogy with the theory of linear models, an F ratio can be formed by fitting both the null and alternative models, and considering the difference between the fits. Under the null model, parametric least squares is used; the corresponding fitted values are $\mathbf{M}Y$ where $\mathbf{M}$ is the hat matrix for the least squares fit. Under the alternative model, the fitted values are $\mathbf{H}Y$, where $\mathbf{H}$ is the hat matrix for a local linear regression. An F ratio can then be formed as

$$F = \frac{\|\mathbf{H}Y - \mathbf{M}Y\|^2/\nu}{\hat{\sigma}^2},$$

where $\nu = \text{trace}((\mathbf{H} - \mathbf{M})^\top(\mathbf{H} - \mathbf{M}))$.

What is the distribution of F when H_0 is true? Since $\mathbf{H}$ is not a perpendicular projection operator, the numerator does not have a χ^2 distribution, and F does not have an exact F distribution. None-the-less, we can use an approximating F distribution. Based on a one-moment approximation, the degrees of freedom are ν and $n - 2\nu_1 + \nu_2$.

Better approximations are obtained using the two-moment Satterwaite approximation, as described in Cleveland and Devlin (1988). This method

matches both the mean and variance of chi-square approximations to the numerator and denominator. Letting $\Lambda = (\mathbf{H} - \mathbf{M})^\top (\mathbf{H} - \mathbf{M})$, the numerator degrees of freedom for the F distribution are given by $\text{trace}(\Lambda)^2/\text{trace}(\Lambda^2)$. A similar adjustment is made to the denominator degrees of freedom. Simulations reported in Cleveland and Devlin (1988) suggest the two-moment approximation is adequate for setting critical values.

For the fuel economy dataset, we obtain $F = 7.247$, $\nu = 1.0866$ and $n - 2\nu_1 + \nu_2 = 55.997$. Using the one-moment approximation, the p-value is 0.0079. The two-moment approximation gives a p-value of 0.0019. Both methods indicate that the nonlinearity is significant, although there is some discrepancy between the P-values.

4.3 Bootstrapping

The F-tests in the previous section are approximate, even when the errors ϵ_i are normally distributed. Additionally, the degrees-of-freedom computations (particularly for the two-moment approximation) require $O(n^3)$ computations, which is prohibitively expensive for n more than a few hundred.

An alternative to the F approximations is to simulate the null distribution of the F ratio. A bootstrap method (Chapter **??**) performs the simulations using the empirical residuals to approximate the true error distribution:

- Let $r_i = Y_i - \hat{\mu}(x_i)$.
- Resample: $Y_i^* = \hat{\mu}(x_i) + \epsilon_i^*, i = 1, \ldots, n$, where ϵ_i^* is drawn from $r_1, \ldots, r_n$.
- Compute the F statistic based on the resampled data:

$$F^* = \frac{\|\mathbf{H}Y - \mathbf{M}Y\|^2/\nu}{(\hat{\sigma}^*)^2}.$$

This procedure is repeated a large number of times (say $B = 1000$) and tabulation of the resulting F^* values provides an estimate of the true distribution of the F ratio.

Remark. Since the degrees of freedom do not change with the replication, there is no need to actually compute the normalizing constant. Instead, one can simply work with the modified F ratio,

$$F_B = \frac{\|\mathbf{H}Y - \mathbf{M}Y\|^2}{\|(I - \mathbf{H})Y\|^2}.$$

Fig. 5 compares the bootstrap distribution of the F ratio and the 1 and 2 moment F approximations for the fuel economy dataset. The bootstrap method uses 10000 bootstrap replications, and the density is estimated using the Local Likelihood method (Section 5.2 below). Except at the left end-point, there is generally good agreement between the bootstrap density and the two-moment density. The upper 5% quantiles are 3.21 based on the two-moment approximation, and 3.30 based on the bootstrap sample. The one-moment

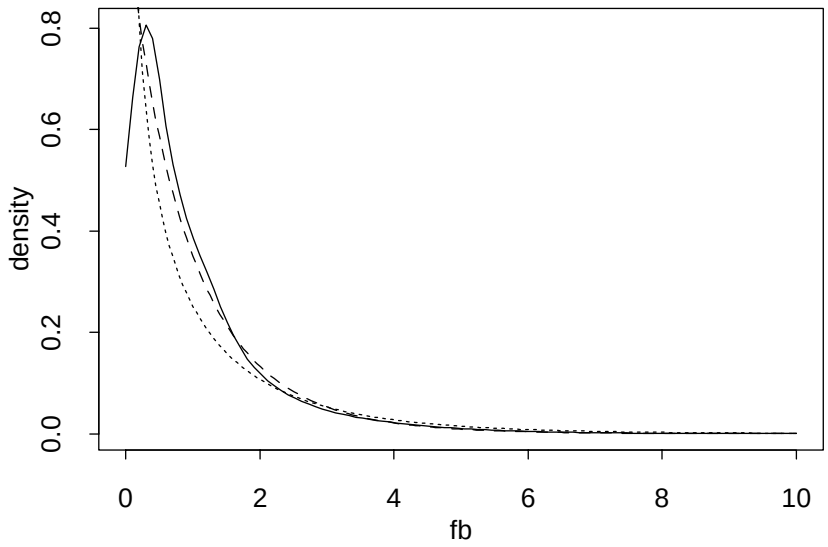

Fig. 5. Estimated density of the F ratio, based on the bootstrap method (solid line); 1-moment F approximation (short dashed line) and 2-moment F approximation (long dashed line).

approximation has a critical value of 3.90. Based on the observed $F = 7.248$, the bootstrap p-value is 0.0023, again in close agreement with the two-moment method.

5 Multivariate Smoothers

When there are multiple predictor variables, the smoothing problem becomes multivariate: $\mu(x)$ is now a surface. The definition of kernel and local regression smoothers can be extended to estimate a regression surface with any number of predictor variables, although the methods become less useful for more than 2 or 3 variables. There are several reasons for this:

- Data sparsity - the curse of dimensionality.
- Visualization issues - how does one view and interpret a high dimensional smooth regression surface?
- Computation is often much more expensive in high dimensions.

For these reasons, use of local polynomials and other smoothers to model high dimensional surfaces is rarely recommended, and the presentation here is restricted to the two-dimensional case. In higher dimensions, smoothers can be used in conjunction with dimension reduction procedures (Chapter **??**), which

attempt to model the high-dimensional surface through low-dimensional components. Examples of this type of procedure include Projection Pursuit (Friedman and Stuetzle, 1981), Additive Models (Hastie and Tibshirani, 1990), Semiparametric Models (Ruppert et al. (2003) and Chapter **??**) and recursive partitioning (Chapter **??**).

5.1 Two predictor variables

Suppose the dataset consists of n vectors (u_i, v_i, Y_i), where u_i and v_i are considered predictor variables, and Y_i is the response. For simplicity, we'll use $x_i = \begin{pmatrix} u_i & v_i \end{pmatrix}^\top$ to denote a vector of the predictor variables. The data are modeled as

$$Y_i = \mu(u_i, v_i) + \epsilon_i = \mu(x_i) + \epsilon_i.$$

Bivariate smoothers attempt to estimate the surface $\mu(u_i, v_i)$. Kernel and local regression methods can be extended to the bivariate case, simply by defining smoothing weights on a plane rather than on a line. Formally, a bivariate local regression estimate at a point $x = (u, v)^\top$ can be constructed as follows:

1. Define a distance measure $\rho(x, x_i)$ between the data points and fitting point. A common choice is Euclidean distance,
$$\rho(x, x_i) = \sqrt{(u_i - u)^2 + (v_i - v)^2}.$$
2. Define the smoothing weights using a kernel function and bandwidth:
$$w_i(x) = W\left(\frac{\rho(x, x_i)}{h}\right).$$
3. Define a local polynomial approximation, such as a local linear approximation
$$\mu(u_i, v_i) \approx a_0 + a_1(u_i - u) + a_2(v_i - v)$$
when (u_i, v_i) is close to (u, v). More generally, a local polynomial approximation can be written
$$\mu(x_i) \approx \langle a, A(x_i - x)\rangle,$$
where a is a vector of coefficients, and $A(\,\cdot\,)$ is a vector of basis polynomials.
4. Estimate the coefficient vector by local least squares. That is, choose $\hat{a}$ to minimize
$$\sum_{i=1}^{n} w_i(x)\left(Y_i - \langle a, A(x_i - x)\rangle\right)^2.$$
5. The local polynomial estimate is then
$$\hat{\mu}(x) = \hat{a}_0.$$

5.2 Likelihood Smoothing

A likelihood smoother replaces the model (1) with a distributional assumption

$$Y_i \sim f(y, \mu_i),$$

where $f(y, \mu)$ is a specified family of densities, parameterized so that $E(Y_i) = \mu_i$. The family may be chosen depending on the response variable. If Y_i is a count, then the Poisson family is a natural choice:

$$f(y, \mu) = \frac{\mu^y e^{-\mu}}{y!}; y = 0, 1, 2, \ldots.$$

If Y_i is a 0/1 (or no/yes) response, then the Bernoulli family is appropriate:

$$f(y, \mu) = \mu^y (1 - \mu)^{1-y}; y = 0, 1.$$

Given the data, the log-likelihood is

$$\mathcal{L}(\mu_1, \ldots, \mu_n) = \sum_{i=1}^{n} \log f(Y_i, \mu_i).$$

The goal is to estimate the mean function, $\mu_i = \mu(x_i)$ for an observed set of covariates x_i. A generalized linear model (Chapter **??**) uses a parametric model for the mean function. Likelihood smoothers assume only that the mean is a smooth function of the covariates.

The earliest work on likelihood smoothing is Henderson (1924), who used a penalized binomial likelihood to estimate mortality rates. The local likelihood method described below can be viewed as an extension of local polynomial regression, and was introduced by Tibshirani and Hastie (1987).

Local Likelihood Estimation

Local likelihood estimation is based on a locally weighted version of the log-likelihood:

$$\mathcal{L}_x(\mu_1, \ldots, \mu_n) = \sum_{i=1}^{n} w_i(x) \log f(Y_i, \mu_i).$$

A local polynomial approximation is then used for a transformation of the mean function. For example, a local quadratic approximation is

$$\begin{aligned} \theta(x_i) &= g(\mu(x_i)) \\ &\approx a_0 + a_1(x_i - x) + \frac{a_2}{2}(x_i - x)^2. \end{aligned}$$

The function $g(\mu)$ is the link function. Its primary goal is to remove constraints on the mean by mapping the parameter space to $(-\infty, \infty)$. For example, in

the Poisson case, the parameter space is $0 < \mu < \infty$. If the log transformation $\theta = \log(\mu)$ is used, then the parameter space becomes $-\infty < \theta < \infty$.

Let $l(y, \theta) = \log f(y, \mu)$ where $\theta = g(\mu)$, so that the locally weighted log-likelihood becomes

$$\mathcal{L}_x = \sum_{i=1}^{n} w_i(x) l(Y_i, \theta(x_i)).$$

The maximizer satisfies the likelihood equations,

$$\sum_{i=1}^{n} w_i(x) \begin{pmatrix} 1 \\ x_i - x \\ \frac{1}{2}(x_i - x)^2 \end{pmatrix} \dot{l}(Y_i, \theta(x_i)) = 0, \tag{20}$$

where

$$\dot{l} = \frac{\partial}{\partial \theta} l(y, \theta).$$

In matrix notation, this system of equations can be written in a form similar to (7):

$$\mathbf{X}^\top \mathbf{W} \dot{l}(Y, \mathbf{X}a) = 0. \tag{21}$$

This system of equations is solved to find parameter estimates $\hat{a}_0, \hat{a}_1$ and $\hat{a}_2$. The local likelihood estimate is defined as

$$\hat{\mu}(x) = g^{-1}(\hat{a}_0).$$

Solving the Local Likelihood Equations

The local likelihood equations (20) are usually non-linear, and so the solution must be obtained through iterative methods. The Newton-Raphson updating formula is

$$\hat{a}^{(j+1)} = \hat{a}^{(j)} + (\mathbf{X}^\top \mathbf{W} \mathbf{V} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \dot{l}(Y, \mathbf{X}\hat{a}^{(j)}), \tag{22}$$

where $\mathbf{V}$ is a diagonal matrix with entries

$$-\frac{\partial^2}{\partial \theta^2} l(y, \theta).$$

For many common likelihoods $l(Y, \theta)$ is concave. Under mild conditions on the design points, this implies that the local likelihood is also concave, and has a unique global maximizer. If the Newton-Raphson algorithm converges, it must converge to this global maximizer.

The Newton-Raphson algorithm (22) cannot be guaranteed to converge from arbitrary starting values. But for concave likelihoods, $\hat{a}^{(j+1)} - \hat{a}^{(j)}$ is guaranteed to be an ascent direction, and convergence can be ensured by controlling the step size.

Statistics for the Local Likelihood Estimate

Since the local likelihood estimate does not have an explicit representation, statistical properties cannot be derived as easily as in the local regression case. But a Taylor series expansion of the local likelihood gives an approximate linearization of the estimate, leading to theory parallel to that developed in Sections 3 and 4 for local regression. See Chapter 4 of Loader (1999b).

5.3 Extensions of Local Likelihood

The local likelihood method has been formulated for regression models. But variants of the method have been derived for numerous other settings, including robust regression, survival models, censored data, proportional hazards models, and density estimation. References include Tibshirani and Hastie (1987), Hjort and Jones (1996), Loader (1996) anb Loader (1999b).

Robust Smoothing

Robust smoothing combines the ideas of robust estimation (Chapter **??**) with smoothing. One method is local M-estimation: choose $\hat{a}$ to minimize

$$\sum_{i=1}^{n} w_i(x)\rho(Y_i - \langle a, A(x_i - x)\rangle),$$

and estimate $\hat{\mu}(x) = \hat{a}_0$. If $\rho(u) = u^2$, this corresponds to local least squares estimation. If $\rho(u)$ is a symmetric function that increases more slowly than u^2, then the resulting estimate is more robust to outliers in the data. One popular choice of $\rho(u)$ is the Huber function:

$$\rho(u) = \begin{cases} u^2 & |u| \le c \\ c(2|u| - c) & |u| > c \end{cases}.$$

References include Härdle (1990) and Loader (1999b). Another variant of M-estimation for local regression is the iterative procedure of Cleveland (1979).

Density Estimation

Suppose $X_1, \ldots, X_n$ are an independent sample from a density $f(x)$. The goal is to estimate $f(x)$. The local likelihood for this problem is

$$\mathcal{L}_x(a) = \sum_{i=1}^{n} w_i(x)\,\langle a, A(x_i - x)\rangle - n\int_{\mathcal{X}} W\left(\frac{u-x}{h}\right) e^{\langle a, A(u-x)\rangle} du.$$

Letting $\hat{a}$ be the maximizer of the local log-likelihood, the local likelihood estimate is $\hat{f}(x) = \exp(\hat{a}_0)$. See Hjort and Jones (1996) and Loader (1996).

The density estimation problem is discussed in detail, together with graphical techniques for visualizing densities, in Chapter **??**.

Acknowledgments

This work was supported by National Science Foundation Grant DMS 0306202.

References

Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csàki, F., editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akademia Kiadó.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Azzalini, A. and Bowman, A. W. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, Oxford.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610.

Efromovich, S. (1999). *Nonparametric Curve Estimation*. Springer, New York.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Series B*, 57:371–394.

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

Friedman, J. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823.

Green, P. J. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach*. Chapman and Hall, London.

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

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer, New York.

Hastie, T. J. and Loader, C. R. (1993). Local regression: Automatic kernel carpentry (with discussion). *Statistical Science*, 8:120–143.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Henderson, R. (1916). Note on graduation by adjusted average. *Transactions of the Actuarial Society of America*, 17:43–48.

Henderson, R. (1924). A new method of graduation. *Transactions of the Actuarial Society of America*, 25:29–40.

Hjort, N. L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation. *The Annals of Statistics*, 24:1619–1647.

Katkovnik, V. Y. (1979). Linear and nonlinear methods of nonparametric regression analysis. Автоматика *(Soviet Automatic Control)*, 5:35–46 (25–34).

Loader, C. (1996). Local likelihood density estimation. *The Annals of Statistics*, 24:1602–1618.

Loader, C. (1999a). Bandwidth selection: Classical or plug-in? *The Annals of Statistics*, 27:415–438.

Loader, C. (1999b). *Local Regression and Likelihood*. Springer, New York.

Mallows, C. L. (1973). Some comments on c_p. *Technometrics*, 15:661–675.

Nadaraya, E. A. (1964). On estimating regression. Теория Вероятностей и ее Применения *(Theory of Probability and its Applications)*, 9:157–159 (141–142).

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90:1257–1270.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22:1346–1370.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

Schiaparelli, G. V. (1866). Sul modo di ricavare la vera espressione delle leggi delta natura dalle curve empiricae. *Effemeridi Astronomiche di Milano per l'Arno*, 857:3–56.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society, Series B*, 47:1–52.

Stone, C. J. (1977). Consistent nonparametric regression (with discussion). *The Annals of Statistics*, 5:595–645.

Tibshirani, R. J. and Hastie, T. J. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82:559–567.

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

Wahba, G. and Wold, S. (1975). A completely automatic French curve: Fitting spline functions by cross-validation. *Communications in Statistics*, 4:1–17.

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

Whitaker, E. T. (1923). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41:62–75.

Index