Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting

WILLIAM S. CLEVELAND and SUSAN J. DEVLIN*

Locally weighted regression, or *loess*, is a way of estimating a regression surface through a multivariate smoothing procedure, fitting a function of the independent variables locally and in a moving fashion analogous to how a moving average is computed for a time series. With local fitting we can estimate a much wider class of regression surfaces than with the usual classes of parametric functions, such as polynomials. The goal of this article is to show, through applications, how loess can be used for three purposes: data exploration, diagnostic checking of parametric models, and providing a nonparametric regression surface. Along the way, the following methodology is introduced: (a) a multivariate smoothing procedure that is an extension of univariate locally weighted regression; (b) statistical procedures that are analogous to those used in the least-squares fitting of parametric functions; (c) several graphical methods that are useful tools for understanding loess estimates and checking the assumptions on which the estimation procedure is based; and (d) the M plot, an adaptation of Mallows's C_p procedure, which provides a graphical portrayal of the trade-off between variance and bias, and which can be used to choose the amount of smoothing.

1. INTRODUCTION

Locally weighted regression, or *loess*, is a procedure for fitting a regression surface to data through multivariate smoothing: The dependent variable is smoothed as a function of the independent variables in a moving fashion analogous to how a moving average is computed for a time series. The basic framework is this. Let y_i ($i = 1, \ldots, n$) be measurements of the dependent variable, and let $x_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, be n measurements of p independent variables. Suppose that the data are generated by $y_i = g(x_i) + \varepsilon_i$. As in the most commonly used framework for regression, we suppose that the ε_i are independent normal variables with mean 0 and variance σ^2. In the usual framework, we would also suppose that g is a member of a parametric class of functions, such as polynomials, but here we will suppose only that g is a smooth function of the independent variables. With local fitting we can estimate a wide class of smooth functions, much wider, in fact, than what we could reasonably expect from any specific parametric class of functions.

Smoothing by local fitting is actually an old idea that is deeply buried in the methodology of time series, where data measured at equally spaced points in time were smoothed by local fitting of polynomials (Macaulay 1931). Watson (1964), Stone (1977), and Cleveland (1979) introduced local-fitting methods into the more general case of regression analysis. Hastie and Tibshirani (1986) took local fitting one step further; in any situation where a dependent variable depends on independent variables, we can carry out a local likelihood procedure. Cleveland (1979) introduced the specific local-fitting methodology that is the subject of this article, locally weighted regression, and Devlin (1986) expanded the methodology and addressed mathematical properties; in this article we further expand the methodology. The original methodology also included a robust version in which M estimation is incorporated so that the assumption of normality can be relaxed, but we do not address robustness here.

The applications in this article illustrate three major uses of the local-fitting methodology. The first is simply to provide an exploratory graphical tool; graphing smooth surfaces that are fitted to the data can give us insight into the behavior of the data and help us choose parametric models. The second is to provide additional regression diagnostics to check the adequacy of parametric models fitted to the data. The third is to use the loess estimate as the estimated regression surface, without resorting to a parametric class of functions. While presenting these three uses we introduce new methods and review and apply some old ones.

In Section 2 we introduce the multivariate smoother: It is a straightforward extension of the univariate loess smoother discussed by Cleveland (1979). Section 3 has an application to velocity measurements of galaxy NGC 7531. Locally weighted regression is used to fit a velocity surface as a function of position on the celestial sphere. In Section 4 we discuss the statistical properties of loess. Fortunately, analogs of the statistical procedures used in parametric function fitting—for example, analysis of variance (ANOVA) and t intervals—involve statistics whose distributions are well approximated by familiar distributions. Section 5 has an application to measurements of ozone concentration and three meteorological variables. Locally weighted regression is used to provide a regression surface and to carry out prediction. In Section 6 we introduce the M plot, using Mallows's C_p idea (Mallows 1966, 1973) with appropriate modifications for the new context, and graphing an estimate of mean squared error against degrees of freedom of the fit. The principal use of the M plot is to choose the amount of smoothing, that is, the neighborhood size of the multivariate smoother. Section 7 has an application to data from an industrial experiment mea-

* William S. Cleveland is in Statistics Research, AT&T Bell Laboratories, Murray Hill, NJ 07974. Susan J. Devlin is in Measurements Research, Bell Communications Research, Piscataway, NJ 08854. This article benefited greatly from discussions with Trevor Hastie, who shared his substantial experience with the backfitting algorithm. The authors are grateful to John Chambers, Trevor Hastie, Jon Kettenring, and Colin Mallows for helpful suggestions about the methods. They also thank two editors and three referees whose comments led to a substantial improvement of the exposition.

Journal of the American Statistical Association
September 1988, Vol. 83, No. 403, Applications & Case Studies

suring the abrasion loss of rubber specimens. A locally weighted regression analysis suggests that there is no interaction between the two independent variables, so the regression surface is estimated by additive fitting (Hastie and Tibshirani 1986). Section 8 has an application to measurements of NO_x in engine exhaust. The history of these data includes an estimation of the regression surface by alternating conditional expectations (ACE) (Breiman and Friedman 1985), a procedure that transforms the dependent variable and fits an additive surface to the data. An analysis by locally weighted regression shows that the regression surface of these data is such that no nontrivial transformation of the data could lead to additivity. Section 9 describes simulations that investigate the distributional approximations of Section 4. Section 10 discusses qualifications to the methodology and discusses other methods.

We also introduce graphical methodology in addition to the M plot. Because it is easier to discuss these methods with graphs at hand, however, we introduce this methodology in the applications sections. Sections 5 and 7 set forth conditioning plots, Section 7 presents component-residual plots, and Section 5 discusses diagnostic plots for checking the assumptions made about ε_i.

The shortened name *loess* has some semantic substance. A loess (pronounced "lō´ is") is a deposit of fine clay or silt along river valleys; in a vertical cross-section of earth, a loess would appear as a narrow, curve-like stratum running through the section.

2. MULTIVARIATE SMOOTHING

Locally weighted regression provides an estimate $\hat{g}(x)$ of the regression surface at any value x in the p-dimensional space of the independent variables. Let q be an integer, where $1 \leq q \leq n$. The estimate of g at x uses the q observations whose x_i values are closest to x. That is, we define a neighborhood in the space of the independent variables. Each point in the neighborhood is weighted according to its distance from x; points close to x have large weight, and points far from x have small weight. A linear or a quadratic function of the independent variables is fitted to the dependent variable using weighted least squares with these weights; $\hat{g}(x)$ is taken to be the value of this fitted function at x. Of course, we must do this computation for each value of x for which we want $\hat{g}(x)$, and thus loess is a computer-intensive method, but algorithms exist for doing the computations efficiently (Cleveland, Devlin, and Grosse 1988).

To carry out locally weighted regression we must have a distance function ρ in the space of the independent variables. For one independent variable we let ρ be Euclidean distance. For the multiple-regression case it is sensible to take ρ to be Euclidean distance in applications where the independent variables are measurements of position in physical space; for example, the independent variables might be geographical location and the dependent variable temperature. If the independent variables are measured on different scales, then it is typically sensible to divide each variable by an estimate of scale before applying a standard distance function. For the applications of Sections 5 and 7, we divide each independent variable by its standard deviation and then use Euclidean distance. (In applications where one or more of the univariate sample distributions of the independent variables has outliers, it is sensible to standardize with a resistant measure of scale such as the interquartile range.) For the application of Section 3 we use Euclidean distance without adjusting the scale.

Locally weighted regression also requires a weight function and a specification of neighborhood size. The weight function used in all of our examples is the tricube function: $W(u) = (1 - u^3)^3$ for $0 \leq u < 1$, and 0 otherwise. We now show how the weight function is used. Let $d(x)$ be the distance of the qth-nearest x_i to x. Then the weight for the observation (y_i, x_i) is

$$w_i(x) = W(\rho(x, x_i)/d(x)).$$

Thus $w_i(x)$ as a function of i is a maximum for x_i close to x, decreases as the x_i increase in distance from x, and becomes 0 for the qth-nearest x_i to x. Instead of thinking in terms of q, the number of points in the neighborhood, we think in terms of $f = q/n$, the fraction of points in the neighborhood. As f increases, $\hat{g}(x)$ becomes smoother. The M plot, which is discussed in Section 6, is an aid to choosing f in applications.

If locally linear fitting is used, the fitting variables are just the independent variables. If locally quadratic fitting is used, the fitting variables are the independent variables, their squares, and their cross-products. Locally quadratic fitting tends to perform better in situations where the regression surface has substantial curvature, such as local maxima and minima (e.g., see the application in Sec. 3).

3. NGC 7531 VELOCITY DATA: AN APPLICATION ILLUSTRATING THE BEHAVIOR OF THE MULTIVARIATE SMOOTHER

NGC 7531 is a spiral galaxy in the Southern Hemisphere with a very bright inner ring. Buta (1987) made measurements of the velocities of this galaxy at a collection of points in the celestial sphere that covered about 200 arc seconds in the north–south direction and about 135 arc seconds in the east–west direction. The measurements were derived from nine spectrograms taken at Cerro Tololo Inter-American Observatory in July and October 1981. Each spectrogram was made along a narrow slit, and the velocity measurements were made at points along the slit by observing the redshift. The locations of these velocity measurements are shown in Figure 1. As can be seen from the figure, there are seven unique positions of the nine slits, since two positions were used twice; the seven unique slit lines intersect at a point in the middle of the observation region. The maximum velocity measurement is 1,785 kilometers per second and the minimum is 1,409 km/sec. The data are scattered because of measurement noise and do not form a smooth velocity field.

The velocity surface was estimated by locally quadratic fitting with $f = .4$. Figure 2 is a contour plot. The fitted surface does a good job of following the underlying pattern in the data. For example, the surface follows the peaks

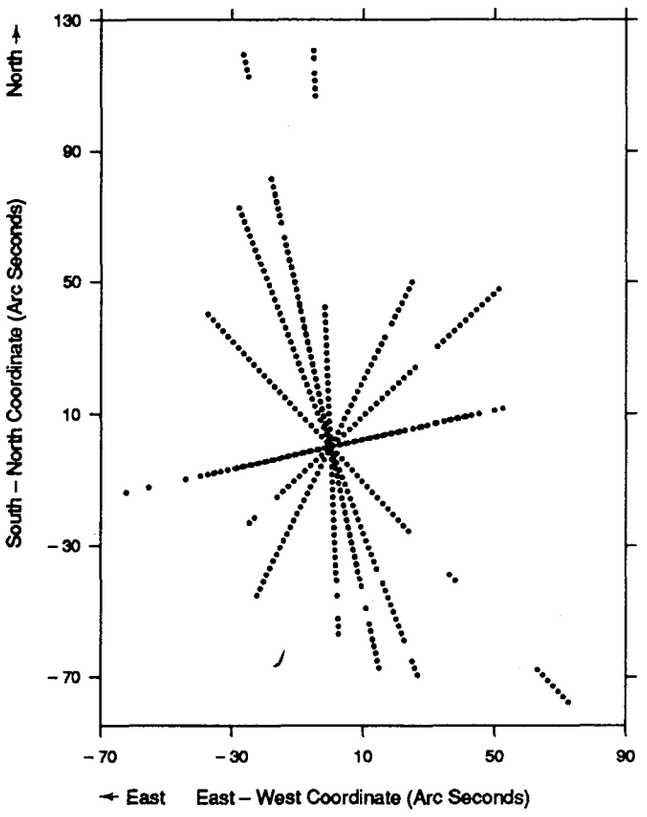

Figure 1. NGC 7531 Velocity Data. The plot shows the locations in the celestial sphere at which the NGC 7531 velocity measurements were made.

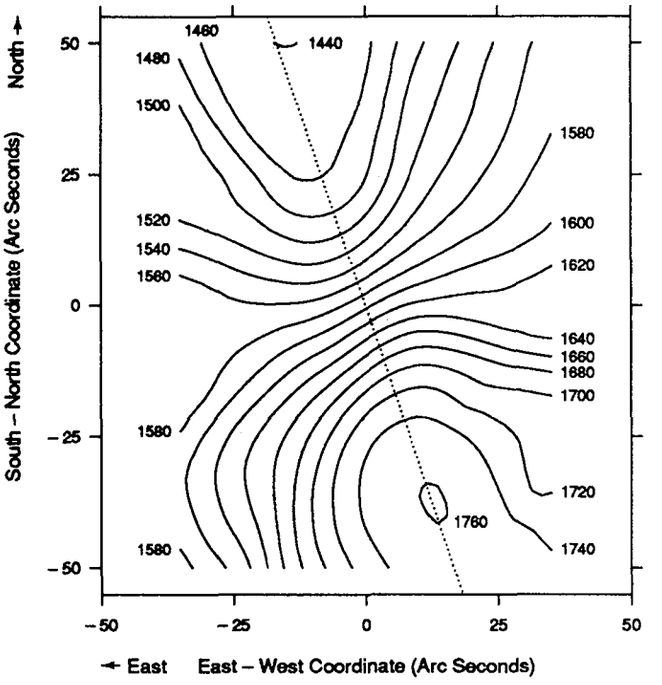

Figure 2. NGC 7531 Velocity Data. The velocity surface was estimated by locally quadratic fitting with f = .4. The figure shows surface contours. The dotted line has a slope of 108°; the surface is roughly symmetric about this line.

and troughs in the data: The maximum value of the estimates at the positions where the measurements were made is 1,757 km/sec, and the minimum value is 1,440 km/sec. When locally linear fitting is used, the fit is poorer and cannot track the substantial curvature unless f is taken to be very small, about .1, in which case the estimated surface is very noisy.

The velocity pattern revealed by the contours is interesting. There appears to be an axis of symmetry of about 108° (the axis is shown by the dotted line in Fig. 2). As we move from north to south along this axis, the velocity increases by about 320 km/sec. Suppose that the only motions of the galaxy (relative to the earth) were a rotation about an axis through its center and a recession due to the expansion of the universe. Then the velocity surface would be linear, the contours would be straight lines parallel to the projection of the axis of rotation on the viewing plane from the earth, and the velocity along this projection would be equal to the recession velocity. Figure 2 does not follow such a pattern. The velocity is not linear along the 108° axis: As we move from the center outward along the axis, the rate of change of the velocity decreases rather than staying constant. Furthermore, the contours are curved, bending one way below the 1,580 km/sec contour and the other way above this contour. Nevertheless, the contours suggest that the predominant motion of the galaxy (aside from the recession) is circular. The motion superimposed on this rotation, which results in the bending of the contours, is not yet known (Buta 1987).

4. STATISTICAL PROPERTIES

The loess estimate, $\hat{g}(x)$, is a linear combination of the y_i,

$$\hat{g}(x) = \sum_{i=1}^{n} l_i(x)\, y_i,$$

where the $l_i(x)$ depend on x_k for $k = 1, \ldots, n$, W, ρ, and f, but not on the y_i. Let $\hat{y}_i = \hat{g}(x_i)$ be the fitted values, let $\hat{\varepsilon}_i = y_i - \hat{y}_i$ be the residuals, and let $y = (y_1, \ldots, y_n)'$, $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)'$, and $\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)'$. Since each $\hat{y}_i$ is a linear combination of the elements of y, we have that $\hat{y} = Ly$, where L (locally weighted regression) is an $n \times n$ matrix and $\hat{\varepsilon} = (I - L)y$, where I is the $n \times n$ identity matrix. This is analogous to parametric least squares: For least squares, the fitted values are Gy, where G (Gauss) is the projection operator onto the space spanned by the fitting variables. If we apply both G and L to the values of one of the fitting variables, we get the same values back. One way to write this is $GG = G$ and $LG = G$. But unlike G, L is neither symmetric nor idempotent (Devlin 1986).

There are three key ingredients for discussing the sampling variability of the loess estimate: (a) that $\hat{g}(x)$ is a linear combination of the y_i; (b) the assumption that y has a normal distribution; and (c) the assumption that $\hat{g}(x)$ estimates g with no bias. For locally linear fitting the assumption of no bias can only be exactly true when g is linear, and for locally quadratic fitting it can only be exactly true when g is quadratic. Nevertheless, the goal of part of the diagnostic checking (discussed in Sec. 5) and the M plot (discussed in Sec. 6) is to find estimates with

negligible bias. Note that lack of bias also underlies the distributional results of parametric regression.

The major conclusion of this section is that several statistics defined analogously with those used in fitting parametric functions by least squares have distributions that are well approximated by those used in parametric regression. This is good news, because familiar techniques can thus be used in making inferences based on loess. In the remainder of this section we present the distributional approximations, and in Section 9 we describe simulations that studied the quality of the approximations.

4.1 Distributions of Residuals, Fitted Values, and Residual Sum of Squares

Because of the linearity and normality, $\hat{y}$ and $\hat{\varepsilon}$ have normal distributions with covariance matrices $\sigma^2 LL'$ and $\sigma^2(I - L)(I - L)'$, respectively. Now

$$\hat{\varepsilon}'\hat{\varepsilon} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \text{residual sum of squares.}$$

Because of the unbiasedness, $E(\hat{\varepsilon}'\hat{\varepsilon}) = \sigma^2 \operatorname{tr}(I - L)'(I - L)$, and we can estimate σ^2 by

$$\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}/\operatorname{tr}(I - L)(I - L)'.$$

Thus, since the variance of $\hat{g}(x)$ is

$$\sigma^2(x) = \sigma^2 \sum_{i=1}^{n} l_i^2(x),$$

we can estimate it by

$$\hat{\sigma}^2(x) = \hat{\sigma}^2 \sum_{i=1}^{n} l_i^2(x).$$

We can approximate the distribution of a quadratic form in normal variables such as $\hat{\varepsilon}'\hat{\varepsilon}$ by the distribution of a constant multiplied by a χ^2 variable; the degrees of freedom and the constant are chosen so that the first two moments of the approximating distribution match those of the distribution of the quadratic form (Kendall and Stuart 1977). Let $\delta_1 = \operatorname{tr}(I - L)(I - L)'$ and let $\delta_2 = \operatorname{tr}[(I - L)(I - L)']^2$. Using this method of approximation, the distribution of $(\delta_1^2\hat{\sigma}^2)/(\delta_2\sigma^2)$ is approximated by a χ^2 distribution with δ_1^2/δ_2 df, and the distribution of $(\hat{g}(x) - g(x))/\hat{\sigma}(x)$ is approximated by a t distribution with δ_1^2/δ_2 df. We can use this result to get approximate confidence intervals for $g(x)$ based on $\hat{g}(x)$.

4.2 Analysis of Variance

Suppose that Ny and Ay are two vectors of fitted values for two regression procedures. We think of N as yielding a fit for a null hypothesis and A as yielding a fit for an alternative hypothesis. For example, N might be linear least squares so that $N = G$ and A might be loess so that $A = L$, or A might be loess with a small value of f, say .3, and N might be loess with a larger value of f, say .9, so that $N = L_{.9}$ and $A = L_{.3}$. Let $y'R_N y = y'(I - N)(I - N)'y$ and $y'R_A y = y'(I - A)(I - A)'y$ be the residual sum of squares of the two fits. If we want to test N against A, the likelihood ratio test leads us to $(y'R_N y)/(y'R_A y) > c$. Thus we will use in analogy with ANOVA a test based on $(y'R_N y - y'R_A y)/y'R_A y$. In this test the reduction due to A in the residual sum of squares is compared with the residual sum of squares of A. [Devlin (1986) discussed a somewhat different approach to testing for the special case where $N = G$.] Let $\nu_1 = \operatorname{tr}(R_N - R_A)$, $\nu_2 = \operatorname{tr}(R_N - R_A)^2$, $\delta_1 = \operatorname{tr} R_A$, and $\delta_2 = \operatorname{tr} R_A^2$. The idea is to use the two-moment χ^2 approximation for the numerator of the aforementioned statistic and the denominator, and approximate the test statistic by an F distribution. That is,

$$\hat{F} = \frac{(y'R_N y - y'R_A y)/\nu_1}{(y'R_A y)/\delta_1}$$

is the test statistic and its distribution is approximated by an F distribution with ν_1^2/ν_2 and δ_1^2/δ_2 df. We refer to ν_1 as the *numerator divisor* of the F test and to ν_1^2/ν_2 as the *numerator degrees of freedom*. Similar terminology holds for δ_1 and δ_1^2/δ_2.

5. OZONE AND METEOROLOGICAL DATA: AN APPLICATION ILLUSTRATING THE USE OF THE STATISTICAL PROPERTIES, DIAGNOSTIC CHECKING, AND CONDITIONING PLOTS

The data in this application are 111 measurements of four variables—ozone (an air pollutant), solar radiation, temperature, and wind speed—on 111 days between May 1 and September 30, 1973, at sites in the New York City metropolitan region (Bruntz, Cleveland, Kleiner, and Warner 1974). We analyzed these data to describe the dependence of ozone on the meteorological variables so that ozone concentrations can be predicted from forecasts of the meteorology. Figure 3 is a scatterplot matrix of the data. The first step in the analysis of these data was to smooth ozone as a function of the meteorological variables by a locally linear fitting with $f = .4$.

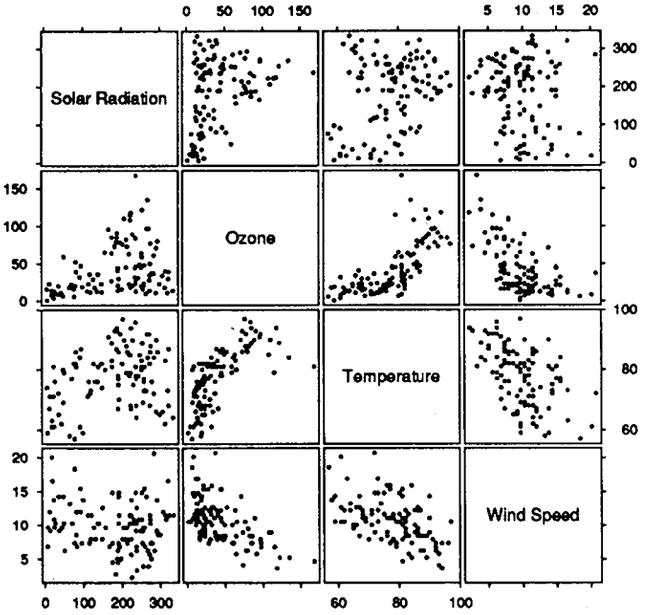

Figure 3. Ozone and Meteorological Data. The figure is a scatterplot matrix of 111 measurements of ozone, wind speed, temperature, and solar radiation. The goal is to predict the ozone concentrations from the meteorological variables.

The loessial methodology discussed in this article widens the domain of applicability compared with the much-practiced parametric-function fitting; nevertheless, the methodology is still based on certain critical assumptions. One is that the errors, ε_i, are independently and normally distributed with constant variance. Another is that the fitted function follows the pattern of the data, that is, provides a nearly unbiased estimate. Such assumptions must be checked. When assumptions are violated we can often take corrective actions similar to those used in parametric regression. There already exists a wealth of diagnostic procedures for regression models (Belsley, Kuh, and Welsch 1980; Chambers, Cleveland, Kleiner, and Tukey 1983; Cook and Weisberg 1982; Daniel and Wood 1971). Much of it is applicable to locally weighted regression; for example, one can make a normal probability plot of $\hat{\varepsilon}_i$ to check the normality assumption, make a plot of $|\varepsilon_i|$ against $\hat{y}_i$ to check the assumption of a constant variance, and graph $\hat{\varepsilon}_i$ against the independent variables to check for bias.

Figures 4 and 5 are diagnostic plots for the locally linear fit to the ozone data. The top panel of Figure 4 is a normal probability plot of the $\hat{\varepsilon}_i$. The curvature suggests that the ε_i have a distribution that is skewed to the right. The bottom panel of Figure 4 is a plot of $|\hat{\varepsilon}_i|$ versus $\hat{y}_i$. The smooth curve is a locally linear fit to the points of the plot with $f = \frac{2}{3}$. The plot suggests that the variance of ε_i depends on the level of g. Figure 5 shows plots of $\hat{\varepsilon}_i$ against the independent variables. The curves on the graphs are locally linear fits with $f = \frac{2}{3}$. No distortion appears in the top panel, but a small effect appears in the middle panel and a more serious one appears in the bottom panel, which suggests that the estimated surface is not following the pattern in the data. Of course, it is possible that the distortion is also causing the inadequacies in Figure 4.

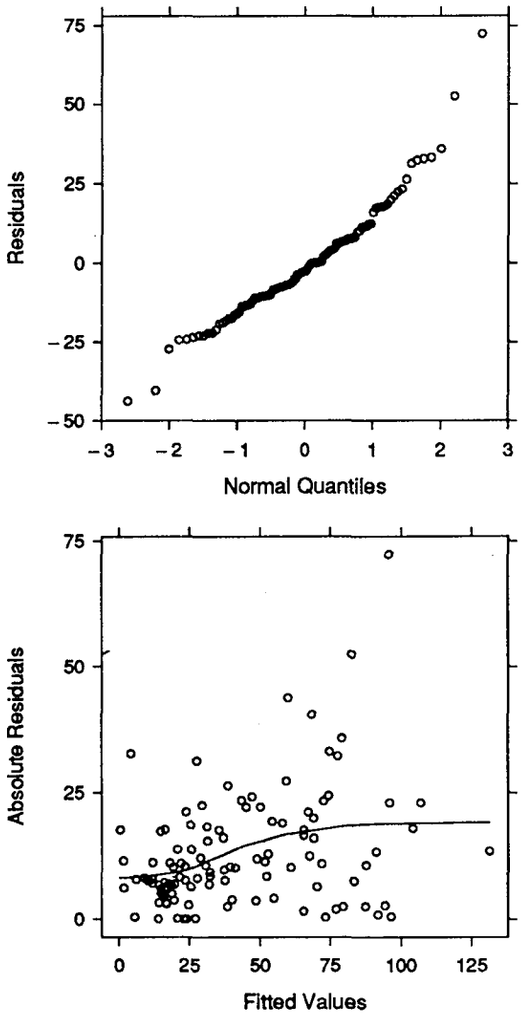

Figure 4. Ozone and Meteorological Data. Ozone was regressed on the meteorological variables using locally linear fitting and f = .4. The top panel is a normal probability plot of the residuals. The bottom panel is a graph of the absolute residuals against the fitted values; the smooth curve is a loess fit to the data of the plot, with f = 2/3. The plots show nonnormality and a dependence of variance on the level of the dependent variable.

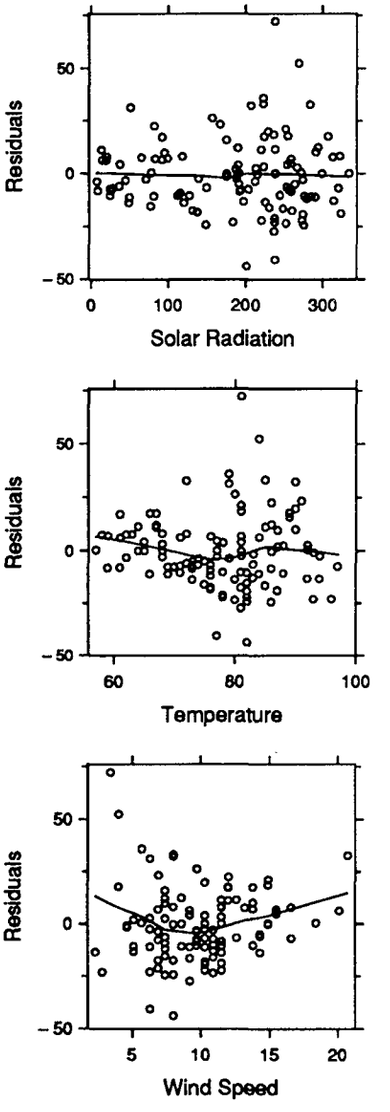

Figure 5. Ozone and Meteorological Data. The residuals for the ozone data are graphed against the independent variables; the smooth curves are loess fits to the data of the plots, with f = 2/3. The plots indicate that the estimated regression surface does not fit the data.

We could reduce the distortion by decreasing the value of f, which is .4. But since f is already fairly small, and since the fitted surface has substantial curvature, we decided to combat the distortion by switching to locally quadratic fitting with $f = .8$. The distortion disappeared, but the inadequacies of Figure 4 remained. Thus we took the cube roots of the ozone concentrations and again computed a locally quadratic fit with $f = .8$. This estimate passed the diagnostic checks.

Figures 6–8 are *three-variable conditioning plots* for the locally quadratic fit. In each panel of Figure 6, $\hat{g}$ is graphed against temperature for fixed values of solar radiation and wind speed, and confidence intervals (computed as described in Sec. 4.1) are shown at five values of temperature. For example, in the panels of the bottom row, solar radiation is 50 langleys; in the panels of the leftmost column, wind speed is 5 miles per hour. Figures 7 and 8 graph $\hat{g}$ against solar radiation and wind speed, respectively, for fixed values of the other variables. The conditioning plots show clearly the nonlinearity of the regression surface and the interaction among the independent variables.

One major reason for fitting a regression surface to ozone data is prediction, either retrospective or prospective. We want to predict the severity of ozone pollution from actual or predicted values of the meteorological variables. For example, during the period of measurement, May 1–September 30, 1973, there were many days with missing ozone measurements because of malfunctioning equipment. Two of these days, August 10 and 11, followed three days of relatively high concentrations, 122, 89, and 110 parts per billion (ppb), all of which were above the federal standard of 80 ppb. Did the pollution episode continue on these two days, or was it reduced? We can use the loess surface to estimate the missing ozone concentrations from the meteorological measurements. The right and left endpoints of approximate 95% confidence intervals, all on the ppb scale, are the following: August 10—68 and 97; August 11—34 and 57. Thus ozone might have been somewhat elevated on the 10th, but with high probability it dropped on the 11th.

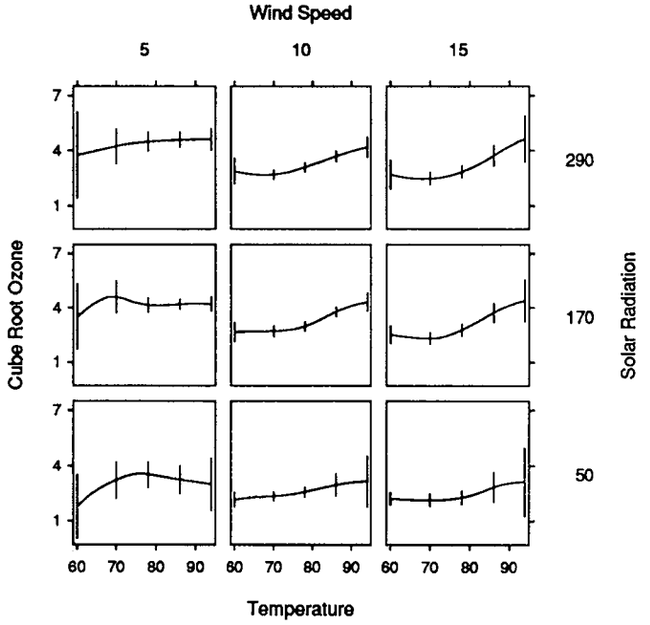

Figure 6. Ozone and Meteorological Data. Because of the problems indicated by the diagnostic plots in Figures 4 and 5, ozone was transformed by cube roots and locally quadratic fitting with f = .8 was used. This figure shows a conditioning plot. Each panel shows a slice of the regression surface as a function of temperature for fixed values of solar radiation and wind speed; the vertical lines are 95% confidence intervals.

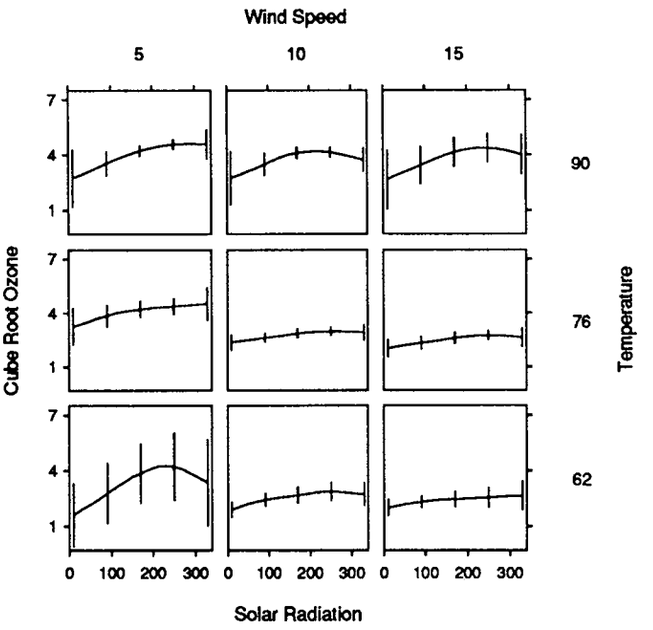

Figure 7. Ozone and Meteorological Data. Each panel of this conditioning plot shows a slice of the regression surface as a function of solar radiation for fixed values of temperature and wind speed.

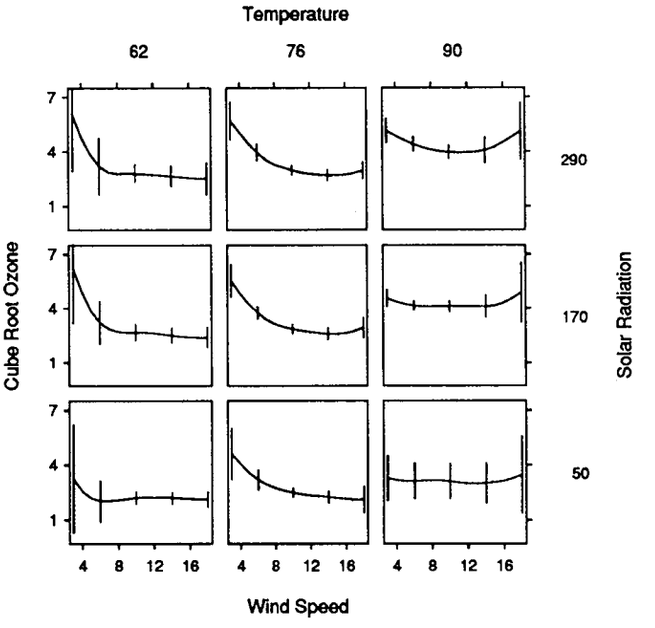

Figure 8. Ozone and Meteorological Data. Each panel of this conditioning plot shows a slice of the regression surface as a function of wind speed for fixed values of temperature and solar radiation.

6. THE *M* PLOT

Mallows (1966) invented a procedure called C_p for choosing a subset of the independent variables based on estimates of the mean squared error for each subset. Later, Mallows (1973) extended this to a more general class of estimates and applied it to choosing the parameter in ridge regression. We can also extend it to locally weighted regression to help choose the value of f. The expected mean squared error summed over the x_i in the sample and divided by σ^2 is

$$M_f = \left[E \sum_{i=1}^{n} (\hat{g}_f(x_i) - g(x_i))^2 \right] \Big/ \sigma^2,$$

where the notation for the fitted values, $\hat{g}_f(x_i)$, now has a subscript to show the dependence on f. Suppose that $\hat{\sigma}_s^2$ is an estimate of σ^2 from a smoothing where s, the value of f, is small, usually in the range from .2 to .4. The idea is to choose a small s so that the bias of $\hat{g}_s(x_i)$ will be negligible, which results in a nearly unbiased estimate of σ^2. Now, let

$$\hat{B}_f = \hat{\varepsilon}_f' \hat{\varepsilon}_f / \hat{\sigma}_s^2 - \operatorname{tr}(I - L_f)'(I - L_f)$$

and

$$V_f = \operatorname{tr} L_f' L_f.$$

A simple derivation shows that we can estimate M_f by $\hat{M}_f = \hat{B}_f + V_f$. $\hat{B}_f$ is the contribution of bias to the estimated mean squared error, and V_f is the contribution of variance. If, for a particular f, $\hat{g}_f$ is a nearly unbiased estimate, then using a standard δ-method argument (Kendall and Stuart 1977) the expected value of $\hat{B}_f$ is nearly 0, so the expected value of $\hat{M}_f$ is nearly V_f. If as f increases bias is introduced, $\hat{B}_f$ has a positive expected value, so the expected value of $\hat{M}_f$ exceeds V_f.

Here V_f is the equivalent number of parameters of the fit, a measure of the amount of smoothing done by the local-fitting procedure. We use this name because if we had done ordinary linear least squares, then the operator matrix L_f would be replaced by G and tr $G'G$ = tr G, the number of parameters used in the fit. In the forthcoming applications, V_f decreases as f increases, so more smoothing results in a smaller equivalent number of parameters.

The M plot is a graph of $\hat{M}_f$ against V_f for a selection of f values between s and 1; this lets us see the trade-off between the contributions of variance and bias to the mean squared error as f changes. It is also helpful, for judging variation on the plot, to show information about the distribution of $\hat{M}_f$ when there is no bias. We can proceed exactly as in Section 4.2. Let R_N be the matrix for the residual sum of squares when the smoothing parameter is f, that is, $\hat{\varepsilon}_f' \hat{\varepsilon}_f = y' R_N y$, and let R_A be the matrix when the parameter is s. Then

$$\hat{M}_f = \nu_1 \frac{(y' R_N y - y' R_A y)/\nu_1}{(y' R_A y)/\delta_1} + \delta_1 - n + 2 \operatorname{tr} L_f$$

$$= \nu_1 \hat{F} + \delta_1 - n + 2 \operatorname{tr} L_f.$$

As before, we approximate the distribution of $\hat{F}$ by an F with ν_1^2/ν_2 and δ_1^2/δ_2 df, and thereby approximate the distribution of $\hat{M}_f$.

It is important to emphasize that the M plot is not intended to produce hard-and-fast rules for the choice of f. Rather, by showing the trade-off between variance and bias as f changes and some information about sampling variability, it assists in our judgment of an appropriate f. Sometimes we want to minimize the mean squared error; this might be the case when we want to use $\hat{g}(x)$ for prediction. In other applications we may decide that low variance is important and thus choose an f that inflates the bias somewhat; this might be the case when the sample size is small or we are searching for a simple description of the data structure that captures the salient features. In still other applications we might decide that low bias is critical; this is often the case when the loess estimate is used for graphical exploration, since our eyes can tolerate some noise but cannot recover a missed effect. Routinely choosing f by minimizing $\hat{M}_f$ is a poor procedure because it ignores variance and bias, which are important to consider in most applications. [Mallows (1973) made the same point about the use of C_p.] Furthermore, at the minimum, $\hat{M}_f$ is often flat compared with its sampling variability, so a range of values of f with different variance and bias properties gives the same mean squared error.

The M plot can be used for more general purposes than comparing loess smoothings with different values of f. For example, we can add $\hat{M}$ from any parametric fit or $\hat{M}$ from other local-fitting procedures such as additive fitting (discussed in Sec. 8). We do this by computing a value of $\hat{M}$ in a manner analogous to the computation of $\hat{M}_f$, and with σ^2 still estimated by $\hat{\sigma}_s^2$.

7. ABRASION–LOSS DATA: AN APPLICATION IN WHICH THE *M* PLOT IS USED TO CHOOSE *f* AND AN ADDITIVE SURFACE FITS THE DATA

An industrial experiment was run measuring three variables for each of 30 rubber specimens (Davies 1957). Each specimen was rubbed with an abrasive material, and the abrasion loss was measured; the experiment was to relate this loss to measurements of the hardness and tensile strength of the specimens. Figure 9 is a scatterplot matrix of the data, which we analyzed by fitting a linear regression model. We intend to evaluate this model. In an initial pass over the data an outlier was found and removed; we analyze the remaining 29 observations. (Since the outlier did not result in extreme values in any of the univariate sample distributions, the independent variables were standardized based on sample standard deviations computed from all 30 observations.)

Figure 10 is an M plot with s = .3. The circles show $\hat{M}_f$ versus V_f for f ranging from f = 1 (the leftmost circle) to f = .3 (the rightmost circle) in steps of .05. The line $\hat{M}_f = V_f$ has been drawn; note that $\hat{M}_s$ must lie on this line. The vertical line segments and their tick marks portray the sampling distribution of $\hat{M}_f$, under the hypothesis of no bias and using the distributional approximation described in the previous section: The top of each line is the

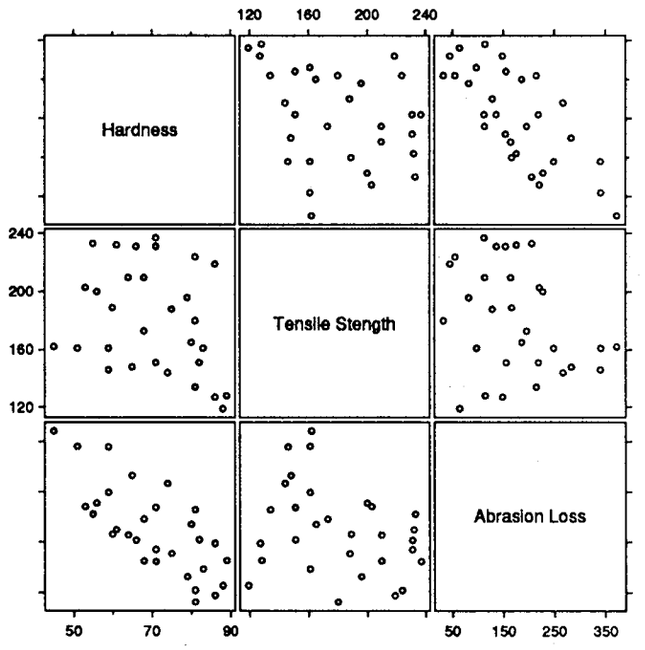

Figure 9. Abrasion-Loss Data. The figure is a scatterplot matrix of data from an industrial experiment in which abrasion loss was studied as a function of hardness and tensile strength.

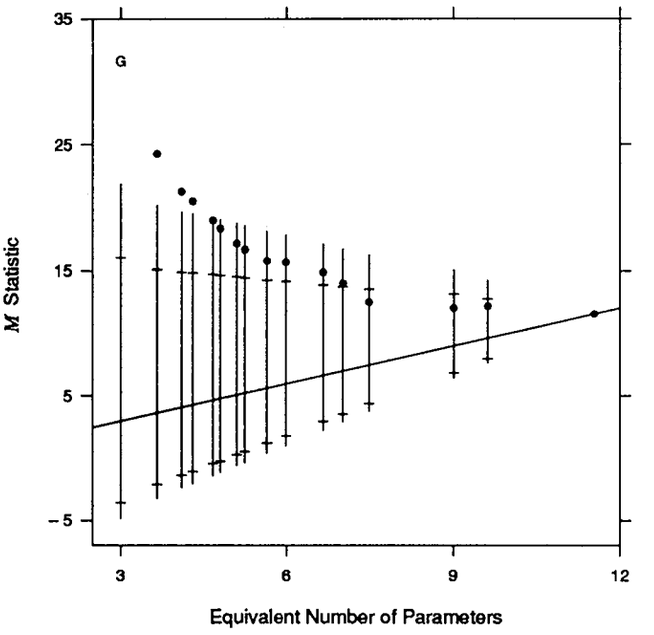

Figure 10. Abrasion-Loss Data. The M plot is a graphical method for choosing the smoothing parameter, f, in locally weighted regression. The filled circles show M statistics, estimates of the mean squared error, for f ranging from .3 (rightmost circle) to 1.0 (leftmost circle). The G shows the M statistic for a linear least-squares fit. The M statistics are graphed against their expected values under an assumption of no bias. The slanted line on the plot is y = x, so the vertical distance of an M statistic to the line is the contribution of bias to the estimate of the mean squared error. The ends of the vertical lines show 90% intervals, and the tick marks show 80% intervals of the distributions of the M statistics under an assumption of no bias. On the basis of this plot, f was chosen to be .5.

95% point, the upper tick mark is the 90% point, the lower tick mark is the 10% point, and the bottom of the line is the 5% point. The G on the plot is the value of $\hat{M}$ for linear least squares. Note that the equivalent number of parameters for the least-squares fit is less than that for any local-regression smoothing, because least squares does more data smoothing than local regression. In Figure 10 there is no clearly defined point where the $\hat{M}_f$ begin a precipitous rise, and $\hat{M}_f$ is flat compared with its sampling variability, from $f = .3$ to $f = .5$; we chose f to be .5, preferring an estimate that had as low a variance as possible, in view of the small sample size, without introducing undue bias. Note that the $\hat{M}$ value for least squares shows that the linear-model fit in the original analysis is inappropriate.

Figure 11 plots the fit with $f = .5$ in the following way: Consider the top curve in the bottom panel. The value of hardness has been set to 60. The curve is a graph of the fitted surface against tensile strength for this fixed value of hardness. For the other curves on the panel, hardness has been set to other values. The graph in the bottom panel is similar, but the conditioning is on tensile strength. This graphical tool is a *two-variable conditioning plot* that can be used generally to explore loess fits with two independent variables. Of course, it is analogous to the three-variable conditioning plots of Figures 6 to 8. Figure 11 reveals several important properties of the estimated surface. On each panel the four curves have roughly the same shape, varying mostly in level, suggesting that there is little interaction between tensile strength and hardness. Furthermore, the plots suggest that abrasion loss is a linear function of hardness and a nonlinear function of tensile strength.

Figure 11 suggests that we incorporate lack of interaction and linearity of hardness into the smoothing. We can do this by following the additive-estimation approach of Hastie and Tibshirani (1986). An *additive estimate* consists of a sum of smooth functions of the independent variables, $\hat{g}_1(x_{i1}) + \cdots + \hat{g}_p(x_{ip})$. The $\hat{g}_k$ are the *component functions*. The salient feature of the estimate is that although the regression surface is nonlinear, there is no interaction among the independent variables.

Additive estimation can be carried out by using the backfitting algorithm from projection selection (Breiman and Friedman 1985; Friedman and Stuetzle 1981; Hastie and Tibshirani 1986). Backfitting is an iterative procedure. In each iteration a component function, say the kth, is updated by smoothing y_i minus the sum of the other component functions as a function of x_{ik}. In our implementation the smoothing is carried out by loess. The final fit is a linear operator applied to y. For this reason the distributional results of Section 4 apply to backfitting as well, but with L replaced by the backfitting operator.

Additive fitting was used for the abrasion-loss data, with the component function for hardness estimated by linear least squares and the component function for tensile strength estimated by loess with varying values of f. Figure 12 shows $\hat{M}$ for these fits. The values of f used in the loess smoothing range from $f = 1$ (leftmost circle) to $f = .3$ (rightmost circle) in steps of .05, just as for the multivariate

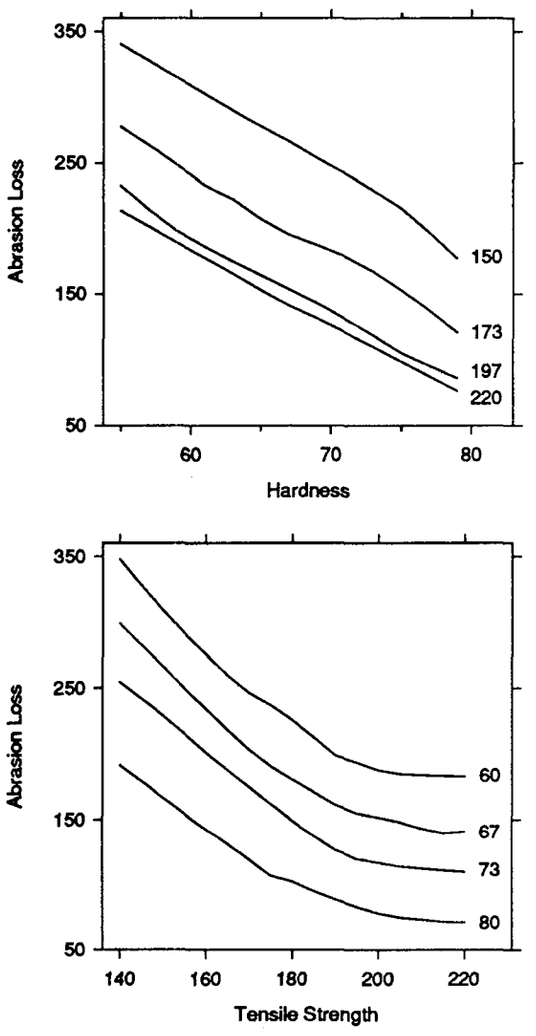

Figure 11. Abrasion-Loss Data. Conditioning plots show a loess fit to the abrasion-loss data, with f = .5. The graphs suggest the dependence on hardness is linear and that there is no interaction between tensile strength and hardness.

loess in Figure 10. Also, the estimate of σ^2 is the same as that in Figure 10. The plot shows that an additive smoothing can provide an acceptable fit to the data; we chose f to be .75, preferring a low-variance estimate without unduly inflating the mean squared error, again in view of the small sample size.

Additive fits can be graphed by *component-residual plots*. As before, let $\hat{g}_r(x_{ir})$ be the estimated component functions, and let $\hat{\varepsilon}_i$ be the residuals. To study the properties of the fit we can make one plot for each component function: $\hat{g}_r(x_{ir})$ is graphed against x_{ir} for $i = 1$ to n by connecting successive points by line segments, and $\hat{g}_r(x_{ir}) + \hat{\varepsilon}_i$ is graphed against x_{ir} by circles. These plots allow us to see the form of the estimated surface and to see whether any signal has leaked into the residuals. The plotting method follows that used in partial residual plots (Landwehr 1983; Larsen and McCleary 1972), where the component functions have a different form.

Figure 13 shows component-residual plots for the additive fit to the abrasion-loss data with $f = .75$. The top panel shows clearly the form that the nonlinearity takes;

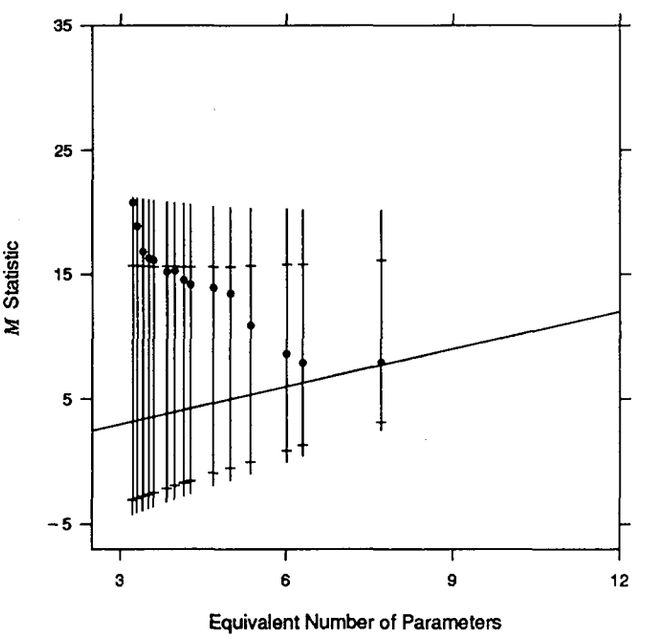

Figure 12. Abrasion-Loss Data. Figure 11 suggests that an additive nonparametric smoothing with no interaction will fit the data. This figure is an M plot for additive fits with hardness estimated by linear least squares and abrasion loss estimated by loess, with f ranging from .3 to 1 in steps of .05. On the basis of this plot, f was chosen to be .75.

there is a hockey-stick dependence. A logical next step in the analysis of these data would be to fit a parametric model in which the dependence of tensile strength is continuous and piecewise linear.

8. NO_x DATA: AN APPLICATION IN WHICH THE *M* PLOT IS USED TO CHOOSE *f* AND AN ADDITIVE SURFACE DOES NOT FIT THE DATA

The data in this application are from an experiment in which a single-cylinder engine was run with ethanol or indolene (Brinkman 1981). There are 110 measurements of compression ratio (C), equivalence ratio (E), and NO_x in the exhaust. The purpose of the analysis was to see how NO_x depends on E and C. There were 88 runs with ethanol; for these runs, E varied from .535 to 1.232, C took one of five values ranging from 7.5 to 18, and the values of E and C were nearly uncorrelated. There were 22 runs with indolene; for these runs, C took just one value, 7.5, and E ranged from .665 to 1.224.

Rodriguez (1985) analyzed these data using ACE (Breiman and Friedman 1985) and MORALS (Young, De-Leeuw, and Takane 1976), with type of fuel as a categorical variable and C and E as continuous variables. In ACE analysis the resulting surface is an additive fit to a transformation of the dependent variable. Thus an ACE fit to the NO_x concentrations results in a surface with no interaction.

Our goal was to explore the data to see if an additive fit was reasonable. To allow for general interactions, we treated C and type of fuel as a single categorical variable with six levels, since C was equal to 7.5 for all indolene runs and to five values ranging from 7.5 to 18 for the

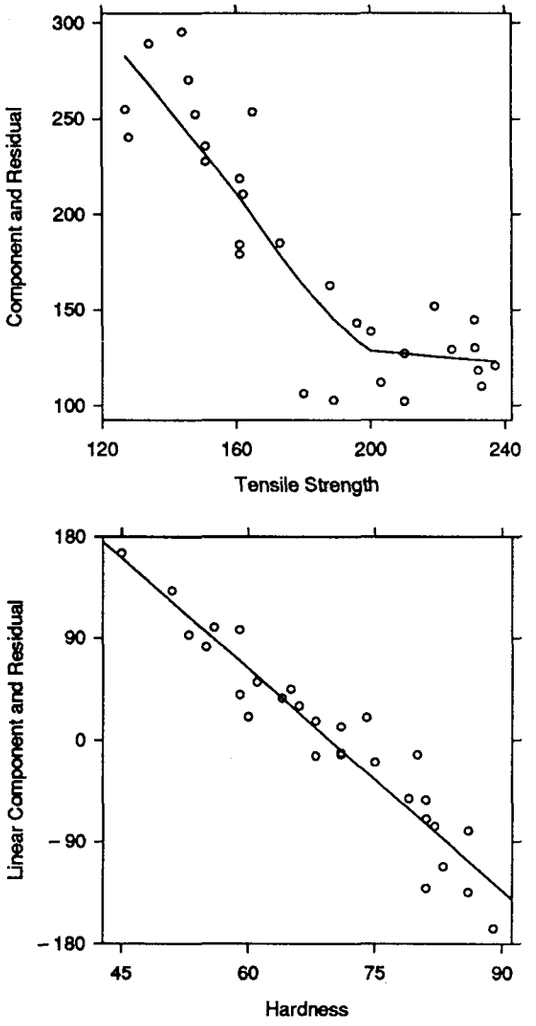

Figure 13. Abrasion-Loss Data. Component-residual plots show the additive fit to the abrasion-loss data, with f = .75. The curve on each plot is the estimated component function for one independent variable, and the plotting symbols show the component function value plus the residual for each observation.

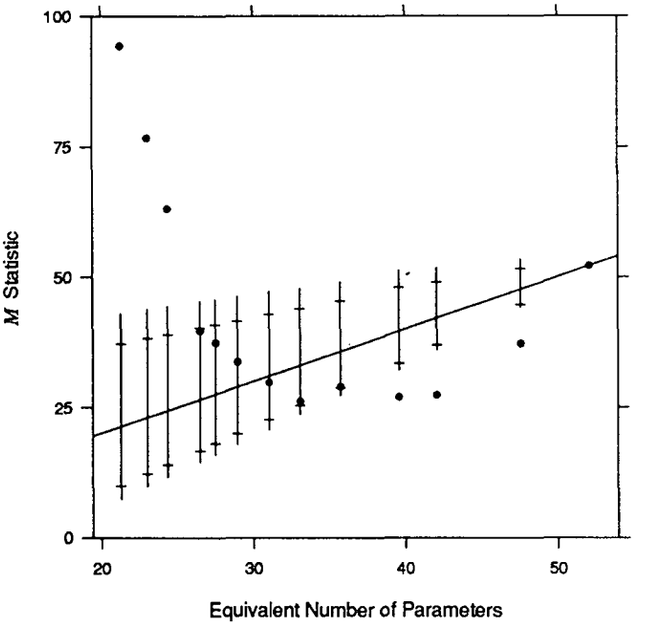

Figure 14. NO_x Data. The data are from an experiment studying the dependence of NO_x exhaust emissions on equivalence ratio, compression ratio, and type of fuel. The figure is an M plot for locally quadratic fitting, with f ranging from .4 to 1 in steps of .05. The type of fuel and the level of compression ratio, which took on one of five values, were both entered as categorical variables. On the basis of the plot, f was chosen to be .85.

ethanol runs. Thus there are two independent variables, E and this categorical variable. Furthermore, the NO_x concentrations were transformed by cube roots. Thus the loess analysis consists of six separate smoothings of cube root NO_x as a function of E, one for each level of the categorical variable. The smoother in this case was locally quadratic fitting because, as we shall see, the functional dependence of cube root NO_x on the equivalence ratio has a local maximum and substantial curvature.

Figure 14 is an M plot for the locally quadratic smoother; the value of s is .4, and in moving from left to right f goes from 1 to .4 in steps of .05. On the basis of this plot f was chosen to be .85; $\hat{M}_f$ jumps considerably for larger values of f.

The top panel of Figure 15 shows the six local-regression estimates, $\hat{h}_k(x)$ for $k = 1$–6, for the six levels of the categorical variable. Each estimate was computed at 50 equally spaced values of E from .6 to 1.15; let the 50 values be denoted by x_j^* for $j = 1$ to 50. Each estimate is graphed in the top panel by connecting successive values by line segments. The bottom panel of Figure 15 is an *interaction plot*. Each curve is a graph of

$$\hat{h}_k(x_j^*) - \frac{1}{6}\sum_{l=1}^{6} \hat{h}_l(x_j^*)$$

against x_j^*.

Figure 15 shows something important: For the ethanol runs, there is a substantial interaction between C and E. As C increases $NO_x^{1/3}$ generally increases, but the effect is reduced as E increases and eventually becomes nearly 0 when E is at its maximum value. Indolene adds to this interaction, because its behavior as a function of E is different from that of ethanol with C equal to 7.5. Thus an additive fit is completely inappropriate for these data. (The M plot for the additive fits, as one would expect, shows very large biases.) Furthermore, Figure 15 shows that the form of the interaction is such that a nontrivial transformation of NO_x cannot possibly remove the interaction, which means that ACE cannot lead to a satisfactory model for these data.

9. LABORATORY AND FIELD SIMULATIONS

Monte Carlo simulations with normal ε_i were run to investigate the distributional approximations discussed in Section 4. We constructed a wide collection of design configurations (i.e., sets of values of the independent variables) for up to five independent variables. Three items were studied in the simulations: (a) distribution of $\hat{\sigma}^2/\sigma^2$, (b) confidence intervals for $g(x)$, and (c) ANOVA for N = linear least squares and A = locally linear fitting. The

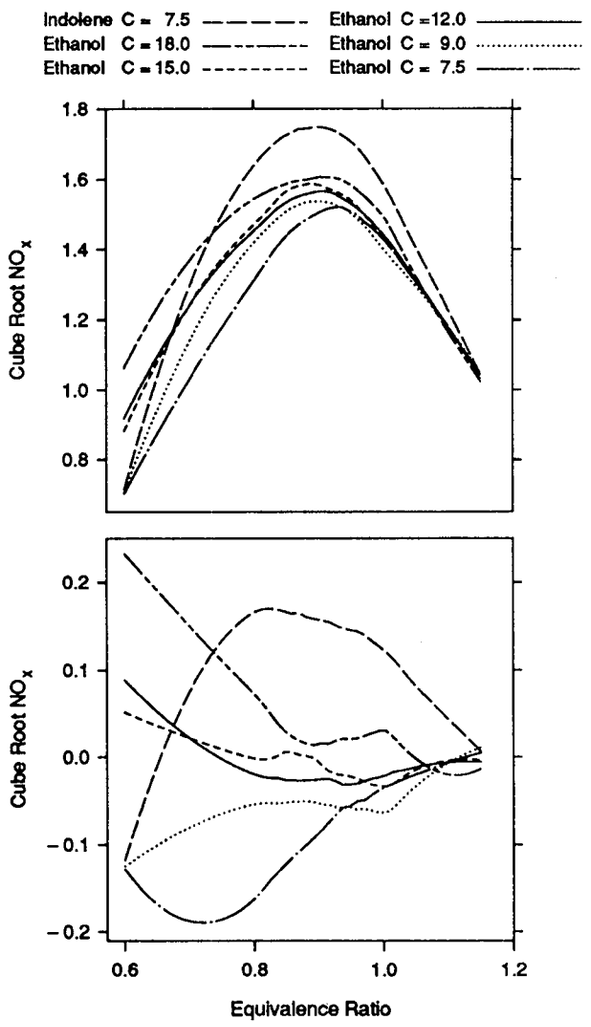

Figure 15. NO_x Data. The top panel shows the six separate smoothings of NO_x as a function of equivalence ratio. The bottom panel shows the curves with a mean curve subtracted. The graphs show a strong interaction among the independent variables that cannot be removed by a nontrivial transformation of the dependent variable. Thus an additive fit is not possible for these data.

distributional approximations of Section 4 were exceedingly close to the true distributions for (a) and (b). For (c) they were close, except when the degrees of freedom of the fit were a large fraction of n; however, this situation is not relevant in practice. We refer to these simulations as *laboratory* simulations, because they employ artificially constructed design configurations. In Section 9.1, (c) is investigated; in Section 9.2, (a) and (b) are investigated; and in Section 9.3, (c) is investigated for a modification of the loess procedure.

For normal ε_i, the true distributions of the statistics involved in (a)–(c) depend on the value of f and the design configuration. A data analyst can check the distributional approximation for any particular application through a simulation using the design configuration of the data and the value of f used in the smoothing. We call these *field* simulations. If the diagnostic checking of the residuals shows that the sample distribution of the residuals is well approximated by a normal distribution, then the field simulation can use samples from the normal. If significant nonnormality appears in the residuals, then sampling can be from the sample distribution of the residuals. Two field simulations are discussed in Section 9.4.

9.1 Laboratory Simulations: Analysis of Variance

In this section we discuss laboratory simulations for testing N = linear least squares against A = locally linear fitting. Figure 16 shows some of the results for one collection of 60 simulations; each simulation employed 16,000 replications, which gave high accuracy even at the .01 significance level. The 60 simulations employed 18 design configurations and 4 values of f; not all values of f were used with each configuration, since we limited our investigations to practical situations.

There were nine design configurations for $p = 1$. For each of three values of n, 100, 50, and 25, there were three sets of values of the independent variable. Each set was of the form $F^{-1}[(i - .5)/n]$ for $i = 1, \ldots, n$, where F was either the uniform, normal, or Cauchy distribution. Simulations with $f = .3$, .5, and .7 were run for each configuration, resulting in 27 simulations.

There were six design configurations for $p = 2$. For each of two values of n, 50 and 100, there were three sets of values of the independent variables. Each set was derived in the following manner: One independent variable was initially set equal to one of the sets of values used for $p = 1$; the second variable was initially set equal to a random permutation of these values; and then the two variables were rotated and scaled to have correlation 0 and variance 1. Simulations with $f = .3$, .5, .7, and .9

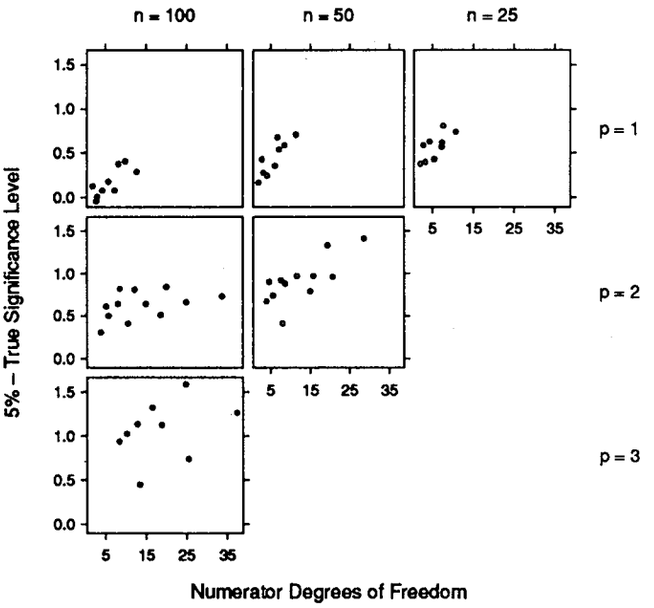

Figure 16. Simulations. The figure shows the results of laboratory simulations investigating the ANOVA test for global linearity. On each panel the vertical scale is 5% (the nominal significance level) minus the true significance level, and the horizontal scale is the degrees of freedom of the numerator. The panels are arranged by p, the number of independent variables, and n, the number of observations. The figure shows that the distributional approximations work exceedingly well.

were run for each configuration, resulting in 24 simulations.

There were three design configurations for $p = 3$. Only the value of $n = 100$ was used, and the configurations were generated in a manner analogous to that for the case with $p = 2$ and $n = 100$. Simulations with $f = .5, .7$, and .9 were run for each configuration, resulting in nine simulations.

Figure 16 shows information about the test at the 5% level of significance. The values plotted on the vertical scales are 5% minus the actual significance, and the horizontal scales are the degrees of freedom of the numerator, that is, v_1^2/v_2. The panels are arranged by p and n. Most important, Figure 16 shows that the approximating 5% significance level is close to the true levels in each of the 60 simulations. The largest absolute deviation is 1.59%. In fact, the situation is even better than that, because the largest departures occur for the largest degrees of freedom, and these values are somewhat larger than those typically used in practice. For the cases with less than 10 df, the largest absolute deviation is .94%. Similar results hold for the deviations at the 10% and 2.5% levels of significance. For the former, the largest absolute deviation is 2.18%; for the latter, the largest is 1.05%. Figure 16 also shows that the deviation of the true level from the nominal level increases as p increases, as n decreases, or as the degrees of freedom increase.

The good performance of the approximations for ANOVA occurs even though the numerator of the test statistic is not independent of the denominator. The approximation works partly because the dependence is not strong and partly because unless n or f is very small the numerator is contributing the most to the variability of the statistic.

9.2 Laboratory Simulations: Confidence Intervals for σ^2 and $g(x)$

The 60 simulations described in Section 9.1 were also used to investigate confidence intervals for σ^2. For the 90% confidence level, the maximum absolute deviation of the actual level from the nominal level was .50%; for the 95% level the maximum was .48%. Clearly, the approximating distributions performed excellently in these cases.

The 27 simulations for $p = 1$ that were described in Section 9.1 were also used to investigate confidence intervals for $g(x)$ at two values of x: the mean of the x_i and the largest of the x_i. For the 90% confidence interval, the largest absolute deviation was .44% for the mean and .65% for the extreme. For the 95% interval, the largest absolute deviation was .45% for the mean and .65% for the extreme. Again, the approximations performed excellently.

9.3 Other Laboratory Simulations

In distributional approximations for ANOVA, the divisors for the sums of squares, v_1 for the numerator and δ_2 for the denominator, are not generally the same as the degrees of freedom for the approximating F distribution, v_1^2/v_2 for the numerator and δ_1^2/δ_2 for the denominator. Nevertheless, one might hope that v_1 is close to v_2 and that δ_1 is close to δ_2, and then take the degrees of freedom to be v_1 and δ_1. The 60 simulations described in Section 9.1 were also used to investigate this one-moment approximation. For the 10%, 5%, and 2.5% levels of significance, the maximum absolute deviations are 3.84%, 2.68%, and 1.62%, respectively. The corresponding values for the two-moment approximation (given in Sec. 9.1) are 2.18%, 1.59%, and 1.05%. The degradation in the approximation for the one-moment case is just large enough that we have continued with the somewhat more complicated two-moment approximation.

9.4 Field Simulations

As we stated earlier, a data analyst can check the performance of the approximating distribution in any application by a field simulation. If the approximating distribution performed poorly, the simulation distribution could be used to make inferences. But we have not yet encountered an application in which the residuals have a sample distribution that is well approximated by the normal and the approximating distribution performed poorly. We will illustrate the use of two field simulations for two of the applications in this article.

For the estimation of the ozone surface in Section 5, it is sensible to ask whether the observed curvature in the fitted surface is significant, because the estimate of the standard error of the residuals is $\hat{\sigma} = .43$, which is not small compared with the sample standard deviation of the cube root ozone concentrations, which is .89. To address whether data with this much noise can support other than a global fit, we carried out ANOVA (described in Sec. 4.2), testing the locally weighted regression fit against a quadratic least-squares fit. The $\hat{F}$ statistic is 2.10 and the approximating distribution is F, with 19.2 and 89.0 df. The significance level is .011, so the curvature is highly significant. We also ran a field simulation with 1,200 replications: The simulated significance level was .010, which is quite close to the approximating level.

The result of the abrasion-loss application in Section 7 was a nonlinear additive fit. Since the number of observations (29) is small, we might reasonably ask whether the data really support a nonlinear regression surface. Thus we tested the additive model against a linear least-squares fit: The significance level was .00256, making the nonlinearity highly significant. (Of course, the test needs to be viewed with some caution, because the model arose after several passes of the fitting process and because f was selected from the M plot.) A field simulation was also run: The simulated significance level was .00211, which is quite close to the approximating level.

10. DISCUSSION

10.1 Locally Weighted Regression for Applications

The methodology introduced here can be an integral part of the analysis in many regression studies. In fact, it represents a new approach, compared with what is most

often practiced today. This methodology can potentially penetrate a regression study most deeply when the dependent variable is a nonlinear function of the independent variables. Today, the two most common approaches to fitting nonlinear surfaces in applications are searching for transformations of the variables that linearize the surface and fitting polynomials of the independent variables. These methods, however, do not lead to a nearly rich enough class of surfaces to model adequately the wide variety of surfaces encountered in practice. But even when the final result of a regression study is a parametric surface, the methodology can help substantiate the validity of the fit.

10.2 Current Restrictions to the Methodology

One current restriction of the applicability of our methodology is the assumption of normality and constant variance of the errors. Nevertheless, future work might relax this restriction. A method for estimating $g(x)$ when the ε_i are assumed only to be symmetric already exists: robust locally weighted regression (Cleveland 1979). What is needed for this robust procedure, however, is distributional results similar to those in Sections 4 and 6. Smoothing techniques without distributional results often leave the analyst with too little methodology to make informed inferences.

Another current restriction is to studies in which the relevance of each independent variable in explaining the dependent variable has already been ascertained. To remove this restriction, work is needed to determine how to incorporate into loess methodology procedures for selecting a subset of the independent variables.

10.3 The Curse of Dimensionality

As the number of independent variables, p, increases, a fixed number of points, n, rapidly becomes sparse. This is referred to as the *curse of dimensionality*. Some have mistakenly supposed that the curse makes multivariate smoothing—that is, smoothing with $p > 1$—a method to avoid. What must be avoided is allowing f to remain fixed as p increases, because for fixed f the equivalent number of parameters of the fit increases as p increases. Of course, we must maintain control of the equivalent number of parameters; this is done by increasing f. As long as we maintain control and do not allow the equivalent number of parameters to become a large fraction of n, we can expect multivariate smoothing to behave reliably. In this article we have successfully carried out multivariate smoothing for data sets with two and three independent variables. Fowlkes (1986) demonstrated that smoothing with more than three independent variables is reasonable in certain circumstances, even for moderately sized data sets. Of course, as p and f increase for fixed n there will be a decrease in the amount of curvature that can be estimated without serious bias. This is not a defect in the method but a statement that the more complicated a regression surface becomes, the larger n must be to get good estimates of it. Exactly the same considerations obtain whatever the method of estimation.

10.4 Weight Functions and the Poor Performance of the Uniform

The general form of the tricube weight function, particularly the smooth contact with 0 at 1, enhances the performance of locally weighted regression. Any reasonable function with smooth contact can also be expected to perform well. Nevertheless, the uniform weight function, with the discontinuity at 1, performs poorly.

A problem with the uniform is that its discontinuity results in local roughness in $\hat{g}(x)$ that is almost always noise and not signal. This is a well-known phenomenon in digital filtering and spectrum analysis, that boxcar windows have Fourier transforms with side lobes that fall off slowly as a function of frequency and thus pass unacceptably large amounts of high-frequency noise (Bloomfield 1976). A second problem with the uniform weight function is that it leads to less satisfactory distributional approximations, because for the uniform, the eigenvalues of L—which, again, are related to the Fourier transform of the weight function—do not lend themselves as well to the approximations as to a continuous weight function such as tricube (Devlin 1986).

We mention the weight-function issue, in part, because asymptotic results for nonparametric regression show that the overall form of the weight function does not have an appreciable effect with respect to mean squared error (e.g., Priestley and Chao 1972). This, however, should not be interpreted to mean that the form of the weight function does not matter in all respects.

10.5 Other Methods

Another approach to smoothing a dependent variable as a function of two or more independent variables is *projection pursuit*, an iterative procedure (Friedman and Stuetzle 1981). At each stage of the iterations, y_i is smoothed as a function of a linear combination of the independent variables. The linear combination is chosen to give a maximum reduction in the residual sum of squares. The smoother is similar to univariate locally weighted regression, but with modifications to decrease the computation time and with a method for choosing the amount of smoothing. The multivariate smoothing introduced here is attractive because of its simplicity: For a particular f, $\hat{g}(x)$ has a straightforward definition and is simply a linear combination of the y_i, so the statistical properties are easy to fathom. This simplicity leads to much of the additional methodology in this article. The full projection-pursuit algorithm results in a considerably more complicated function of the y_i, because the linear combinations of the independent variables are chosen to minimize the residual sum of squares. Consequently, almost nothing is known about its distributional properties (Huber 1985). In addition, full projection pursuit also has its restricted domain of applicability; not all regression surfaces can be well approximated by a moderate number of smooth functions of linear combinations of the independent variables (Huber 1985).

Locally weighted regression falls into a class of regression procedures that some call *nonparametric regression*.

Stone (1977), Collomb (1981), Wegman and Wright (1983), and Titterington (1985) reviewed other procedures. Many studies of nonparametric regression focused on asymptotic properties such as consistency, normality, and rates of convergence (e.g., Benedetti 1977; Devlin 1986; Härdle and Gasser 1984; Stone 1977, 1982; Wahba 1979). For example, Stone (1977), using elegant arguments, showed the asymptotic consistency of a wide class of nonparametric estimates.

One well-known nonparametric regression procedure is smoothing splines (Henderson 1924; Reinsch 1967; Silverman 1985; Wahba 1978; Whittaker 1923). Splines have an attractive property: They are the solution to an intuitively appealing mathematical criterion. Another attractive property is that they have a Bayesian interpretation (Wahba 1978; Whittaker 1923). [Weerahandi and Zidek (1985) provided a Bayesian interpretation for univariate locally weighted regression with a particular weight function.] But splines also have some unattractive properties. First, they optimize a global criterion and are not generally local. [Although, as Silverman (1985) pointed out, when n is large and the amount of smoothing is neither large nor small, spline methods behave, to a good approximation, as smoothing by local fitting with a weight function with exponential decay; thus splines are nearly local in this case.] A second unattractive property is that because splines arise as the result of an optimization, it can be difficult to determine how they operate on the data. In contrast, the operational characteristics of local-fitting procedures are easier to fathom because they are defined directly. For example, because of its definition, one knows that the locally weighted regression estimate, $\hat{g}(x)$, is determined by $100f\%$ of the data at each x, for any n and for any configuration of the x_i (except when ties in the x_i leave more than $100f\%$ of the data at a particular point). It is considerably more difficult to determine the effective bandwidth of a spline estimate at x (Silverman 1984). In many cases this is only possible by numerically working out the coefficients of the linear combination of y_i that make up the estimate.

The most serious problem with splines is computational. Although fast $O(n)$ algorithms exist for one independent variable (Silverman 1985; Whittaker 1923), fitting "thin plate" splines to two or more independent variables is an $O(n^3)$ computation (Wahba 1984). The expected computation time of a loess estimate at a single value of x is $O(n)$. For a fixed value of f (i.e., a fixed number of degrees of freedom of the fit), the number of points at which one needs to compute $\hat{g}$ to characterize it for practical applications is fixed: By using blending functions and $k - d$ trees, local-fitting computations in practice can be kept to $O(n)$ time (Cleveland et al. 1988) and are thus feasible even in computing environments that do not have fast, powerful processors. Note that this strategy is not available in spline smoothing, because one cannot get $\hat{g}$ at a single value of x without carrying out the full optimization. Thus another strategy that has been employed for splines is to solve an altered optimization that requires less computation and that yields a solution close to the original one when n is large (Wahba 1984). But the computing is still substantial and complex, and many questions remain (Silverman 1985).

Two popular methods for choosing the smoothing parameter in spline-fitting are cross-validation (Stone 1974) and generalized cross-validation (Craven and Wahba 1979). Unfortunately, users of these methods generally focus exclusively on the mean squared error, which in Section 6 we criticized as too limiting. One exception, however, is the work by Clark (1980). Of course, one could use cross-validation or generalized cross-validation in place of the M statistic to choose the amount of smoothing for locally weighted regression, or one could use the M statistic for splines. That is, these methods for choosing the amount of smoothing are not dependent on the method of smoothing.

[*Received September 1986. Revised December 1987.*]

REFERENCES

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics*, New York: John Wiley.

Benedetti, J. K. (1977), "On the Nonparametric Estimation of Regression Functions," *Journal of the Royal Statistical Society*, Ser. B, 39, 248–253.

Bloomfield, P. (1976), *Fourier Analysis of Time Series: An Introduction*, New York: John Wiley.

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Correlation and Regression," *Journal of the American Statistical Association*, 80, 580–598.

Brinkman, N. D. (1981), "Ethanol Fuel—A Single-Cylinder Engine Study of Efficiency and Exhaust Emissions," *SAE Transactions*, 90, No. 810345, 1410–1424.

Bruntz, S. M., Cleveland, W. S., Kleiner, B., and Warner, J. L. (1974), "The Dependence of Ambient Ozone on Solar Radiation, Wind, Temperature, and Mixing Height," in *Symposium on Atmospheric Diffusion and Air Pollution*, Boston: American Meteorological Society, pp. 125–128.

Buta, R. (1987), "The Structure and Dynamics of Ringed Galaxies, III: Surface Photometry and Kinematics of the Ringed Nonbarred Spiral NGC 7531," *The Astrophysical Journal Supplement Series*, 64, 1–37.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Monterey, CA: Wadsworth.

Clark, R. M. (1980), "Calibration, Cross-Validation, and Carbon-14, II," *Journal of the Royal Statistical Society*, Ser. A, 143, 177–194.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Cleveland, W. S., Devlin, S. J., and Grosse, E. (1988), "Regression by Local Fitting: Methods, Properties, and Computational Algorithms," *Journal of Econometrics*, 37, 87–114.

Collomb, G. (1981), "Estimation Non-Paramétrique de la Regression: Revue Bibliographique," *International Statistical Review*, 49, 75–93.

Cook, R. D., and Weisberg, S. (1982), *Influence and Residuals in Regression*, New York: Chapman & Hall.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerische Mathematik*, 31, 377–403.

Daniel, C., and Wood, F. (1971), *Fitting Equations to Data*, New York: John Wiley.

Davies, O. L. (ed.) (1957), *Statistical Methods in Research and Production* (3rd ed.), New York: Hafner Press.

Devlin, S. J. (1986), "Locally-Weighted Multiple Regression: Statistical Properties and Its Use to Test for Linearity," technical memorandum, Bell Communications Research, Piscataway, NJ.

Fowlkes, E. B. (1986), "Some Diagnostics for Binary Logistic Regression Via Smoothing" (with discussion), in *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 54–64.

Friedman, J. H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823.

Härdle, W., and Gasser, T. (1984), "Robust Non-Parametric Function Fitting," *Journal of the Royal Statistical Society*, Ser. B, 46, 42–51.

Hastie, T. J., and Tibshirani, R. J. (1986), "Generalized Additive Models," *Statistical Science*, 1, 297–318.

Henderson, R. (1924), "A New Method of Graduation," *Transactions of the Actuarial Society of America*, 25, 29–40.
Huber, P. (1985), "Projection Pursuit" (with discussion), *The Annals of Statistics*, 13, 435–525.
Kendall, M., and Stuart, A. S. (1977), *The Advanced Theory of Statistics* (Vol. 1, 4th ed.), New York: Macmillan.
Landwehr, J. M. (1983), "Using Partial Residual Plots to Detect Nonlinearity," technical memorandum, AT&T Bell Laboratories, Murray Hill, NJ.
Larsen, W. A., and McCleary, S. J. (1972), "The Use of Partial Residual Plots in Regression Analysis," *Technometrics*, 14, 781–790.
Macaulay, F. R. (1981), *The Smoothing of Time Series*, New York: National Bureau of Economic Research.
Mallows, C. L. (1966), "Choosing a Subset Regression," unpublished paper presented at the annual meeting of the American Statistical Association, Los Angeles.
——— (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675.
Priestley, M. B., and Chao, M. T. (1972), "Non-parametric Function Fitting," *Journal of the Royal Statistical Society*, Ser. B, 34, 385–392.
Reinsch, C. (1967), "Smoothing by Spline Functions," *Numerische Mathematik*, 10, 177–183.
Rodriguez, R. N. (1985), "A Comparison of the ACE and MORALS Algorithms in an Application to Engine Exhaust Emissions Modeling," in *Computer Science and Statistics: Proceedings of the Sixteenth Symposium on the Interface*, ed. L. Billard, New York: North-Holland, pp. 159–167.
Silverman, B. W. (1984), "Spline Smoothing: the Equivalent Variable Kernel Method," *The Annals of Statistics*, 12, 898–916.
——— (1985), "Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 47, 1–52.
Stone, C. J. (1977), "Consistent Nonparametric Regression," *The Annals of Statistics*, 5, 595–620.
——— (1982), "Optimal Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053.
Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.
Titterington, D. M. (1985), "Common Structure of Smoothing Techniques in Statistics," *International Statistical Review*, 53, 141–170.
Wahba, G. (1978), "Improper Priors, Spline Smoothing, and the Problem of Guarding Against Model Errors in Regression," *Journal of the Royal Statistical Society*, Ser. B, 40, 364–372.
——— (1979), "Convergence Rates of 'Thin Plate' Smoothing Splines When the Data Are Noisy," in *Smoothing Techniques for Curve Estimation*, eds. T. Gasser and M. Rosenblatt, Berlin: Springer-Verlag, pp. 233–245.
——— (1984), "Cross-Validated Spline Methods for the Estimation of Multivariate Functions From Data on Functionals," in *Statistics: An Appraisal*, eds. H. A. David and H. T. David, Ames: Iowa State University Press, pp. 205–235.
Watson, G. S. (1964), "Smooth Regression Analysis," *Sankhya*, Ser. A, 26, 359–372.
Weerahandi, S., and Zidek, J. V. (1985), "Smoothing Locally Smooth Processes by Bayesian Nonparametric Methods," Technical Report 26, University of British Columbia, Dept. of Statistics.
Wegman, E. J., and Wright, I. W. (1983), "Splines in Statistics," *Journal of the American Statistical Association*, 78, 351–365.
Whittaker, E. T. (1923), "On a New Method of Graduation," in *Proceedings of the Edinburgh Mathematical Society* (Vol. 41), pp. 63–75.
Young, F. W., DeLeeuw, J., and Takane, Y. (1976), "Regression With Qualitative and Quantitative Variables: An Alternating Least-Squares Method With Optimal Scaling Features," *Psychometrika*, 41, 505–529.