Robust Locally Weighted Regression and Smoothing Scatterplots

WILLIAM S. CLEVELAND*

The visual information on a scatterplot can be greatly enhanced, with little additional cost, by computing and plotting smoothed points. Robust locally weighted regression is a method for smoothing a scatterplot, (x_i, y_i), $i = 1, \ldots, n$, in which the fitted value at x_k is the value of a polynomial fit to the data using weighted least squares, where the weight for (x_i, y_i) is large if x_i is close to x_k and small if it is not. A robust fitting procedure is used that guards against deviant points distorting the smoothed points. Visual, computational, and statistical issues of robust locally weighted regression are discussed. Several examples, including data on lead intoxication, are used to illustrate the methodology.

KEY WORDS: Graphics; Scatterplots; Nonparametric regression; Smoothing; Robust estimation.

1. INTRODUCTION

Figure A shows a scatterplot of points (x_i, y_i), for $i = 1, \ldots, n$, where $n = 50$. In Figure B the same scatterplot is summarized by another set of points $(x_i, \hat{y}_i)$, for $i = 1, \ldots, n$, which are plotted by joining successive values by straight lines. The point $(x_i, \hat{y}_i)$ portrays the location of the distribution of the variable on the vertical axis, Y, given the value of the variable on the horizontal axis, $X = x_i$. The formation of the new points will be referred to as smoothing the scatterplot. The point $(x_i, \hat{y}_i)$ is called the smoothed point at x_i and $\hat{y}_i$ is called the fitted value at x_i. The example in Figure A was generated by taking $x_i = i$, and

$$y_i = .02x_i + \epsilon_i \ ,$$

where ϵ_i is a random sample from a normal distribution with mean 0 and variance 1. The linear effect is not easily perceived from the scatterplot alone, but is revealed when the smoothed points are superimposed.

In this article we shall discuss a method for smoothing scatterplots called robust locally weighted regression. Local fitting of polynomials has been used for many decades to smooth time series plots in which the x_i are equally spaced (Macauley 1931). Locally weighted regression is an extension of this technique to more general configurations of the x_i. In addition, a robust fitting procedure is used that guards against deviant points distorting the smoothed points. The procedure is an adaptation of iterated weighted least squares, a recent technique of robust estimation (Beaton and Tukey 1974; Andrews 1974). Thus, robust locally weighted regression is a combination of old ideas for smoothing and new ideas for robust estimation.

An early example of smoothing scatterplots is given by Ezekiel (1941, p. 51). The points are grouped according to x_i, and for each group the mean of the y_i is plotted against the mean of the x_i. More recently, Stone (1977) proves the consistency of a wide class of nonparametric regression estimates under very general conditions and presents a discussion and bibliography of methods that have appeared in the literature. Another method, which appeared after Stone's review, is that of Clark (1977), who proposes a technique for smoothing scatterplots in which the plot is interpolated by joining successive points with straight lines and is then smoothed by convolution with a weight function.

In the remainder of this article we shall first describe the details of robust locally weighted regression. Then, we shall use examples to show how the methodology can be put to use in practice and give guidelines for choosing certain parameters that are needed for carrying out the procedure. An algorithm is given that allows efficient computation of smoothed points. Various statistical topics, including the sampling distributions of fitted values, an estimate of the error variance, and the equivalent number of parameters, are presented. Finally, the interplay between bias and variance is discussed and conditions are given that ensure that increasing a parameter that controls the amount of smoothing will decrease the variance of the fitted values.

2. LOCALLY WEIGHTED REGRESSION AND ROBUST LOCALLY WEIGHTED REGRESSION

We shall first attempt to give the rough idea of the smoothing procedure before giving the precise details. Let W be a weight function with the following properties:

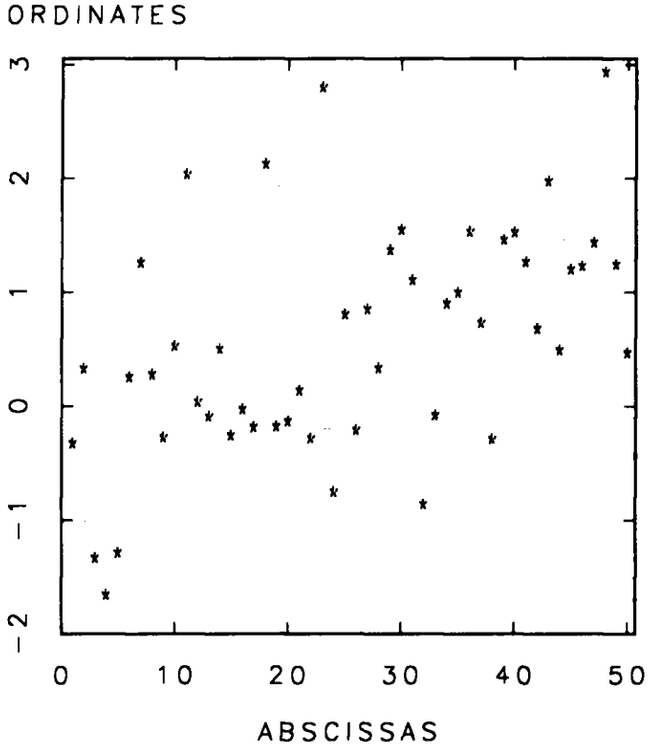
1. $W(x) > 0$ for $|x| < 1$;
2. $W(-x) = W(x)$; (2.1)
3. $W(x)$ is a nonincreasing function for $x \geq 0$;
4. $W(x) = 0$ for $|x| \geq 1$.

Let $0 < f \leq 1$ and let r be fn rounded to the nearest integer. Roughly, the procedure is the following. For each x_i, weights, $w_k(x_i)$, are defined for all x_k, $k = 1, \ldots, n$, using the weight function W. This is done by centering W at x_i and scaling it so that the point at which W first becomes zero is at the rth nearest neighbor of x_i. The

* William S. Cleveland is Member, Technical Staff, Bell Telephone Laboratories, Murray Hill, NJ 07974. The author wishes to thank Richard A. Becker, Roberta Guarino, Colin L. Mallows, and Christine Waternaux for many helpful suggestions.

© Journal of the American Statistical Association
December 1979, Volume 74, Number 368
Theory and Methods Section

A. Scatterplot of Artificially Generated Data



initial fitted value, $\hat{y}_i$, at each x_i is the fitted value of a dth degree polynomial fit to the data using weighted least squares with weights $w_k(x_i)$. This procedure for computing the initial fitted values is referred to as locally weighted regression. A different set of weights, δ_i, is now defined for each (x_i, y_i) based on the size of the residual $y_i - \hat{y}_i$. Large residuals result in small weights and small residuals result in large weights. New fitted values are now computed as before but with $w_k(x_i)$ replaced by $\delta_i w_k(x_i)$. The computation of new weights and new fitted values is now repeated several times. The entire procedure, including the initial computation and the iterations, is referred to as robust locally weighted regression.

The smoothing procedure has been designed to accommodate data for which

$$y_i = g(x_i) + \epsilon_i \tag{2.2}$$

where g is a smooth function and the ϵ_i are random variables with mean 0 and constant scale. Within such a framework, $\hat{y}_i$ is an estimate of $g(x_i)$. The assumption of smoothness allows points in a neighborhood of (x_i, y_i) to be used in forming $\hat{y}_i$. For a weight function, $W(x)$, which decreases for increasing nonnegative x, the weights $w_k(x_i)$ decrease as the distance of x_k from x_i increases. Thus points whose abscissas are close to x_i play a large role in the determination of $\hat{y}_i$, while points far away play a lesser role. Increasing f increases the neighborhood of influential points and therefore tends to increase the smoothness of the smoothed points.

We shall now give the details of the procedures. For each i let h_i be the distance from x_i to the rth nearest neighbor of x_i. That is, h_i is the rth smallest number among $|x_i - x_j|$, for $j = 1, \ldots, n$. For $k = 1, \ldots, n$, let

$$w_k(x_i) = W(h_i^{-1}(x_k - x_i)) \ .$$

Locally weighted regression and robust locally weighted regression are defined by the following sequence of operations:

1. For each i compute the estimates, $\hat{\beta}_j(x_i)$, $j = 0, \ldots, d$, of the parameters in a polynomial regression of degree d of y_k on x_k, which is fit by weighted least squares with weight $w_k(x_i)$ for (x_k, y_k). Thus the $\hat{\beta}_j(x_i)$ are the values of β_j that minimize

$$\sum_{k=1}^{n} w_k(x_i)(y_k - \beta_0 - \beta_1 x_k - \ldots - \beta_d x_k^d)^2 \ .$$

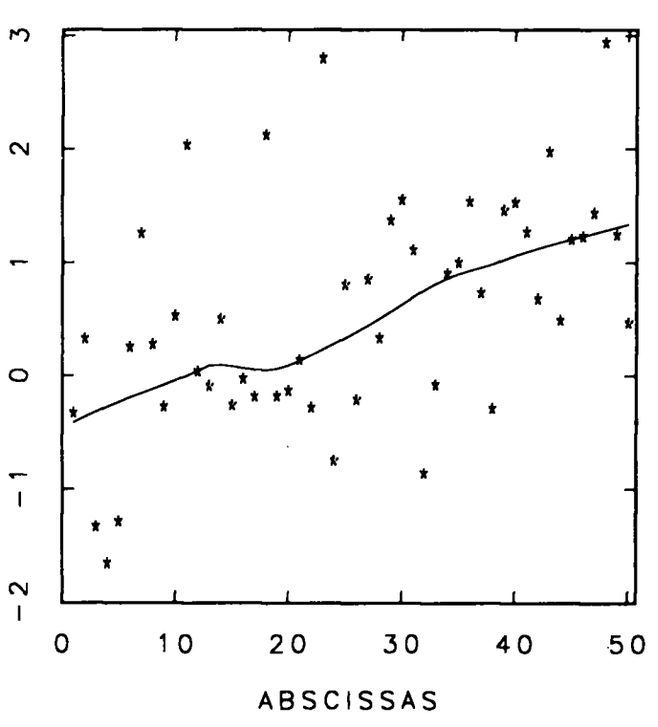
The smoothed point at x_i using locally weighted regression of degree d is $(x_i, \hat{y}_i)$, where $\hat{y}_i$ is the fitted value of the regression at x_i. Thus

$$\hat{y}_i = \sum_{j=0}^{d} \hat{\beta}_j(x_i) x_i^j = \sum_{k=1}^{n} r_k(x_i) y_k \ ,$$

where $r_k(x_i)$ does not depend on y_j, $j = 1, \ldots, n$. We have used the notation "$r_k(x_i)$" to remind us that these are the coefficients for the y_k that arise from the regression.

2. Let B be the bisquare weight function that is de-

B. Scatterplot of Artificially Generated Data and Robust Smoothed Values With $f = .5$



fined by

$$B(x) = (1 - x^2)^2 , \quad \text{for} \quad |x| < 1$$
$$= 0 , \quad \text{for} \quad |x| \geq 1 .$$

Let

$$e_i = y_i - \hat{y}_i$$

be the residuals from the current fitted values. Let s be the median of the $|e_i|$. Define robustness weights by

$$\delta_k = B(e_k/6s) .$$

3. Compute new $\hat{y}_i$ for each i by fitting a dth degree polynomial using weighted least squares with weight $\delta_k w_k(x_i)$ at (x_k, y_k).

4. Repeatedly carry out steps 2 and 3 a total of t times. The final $\hat{y}_i$ are robust locally weighted regression fitted values.

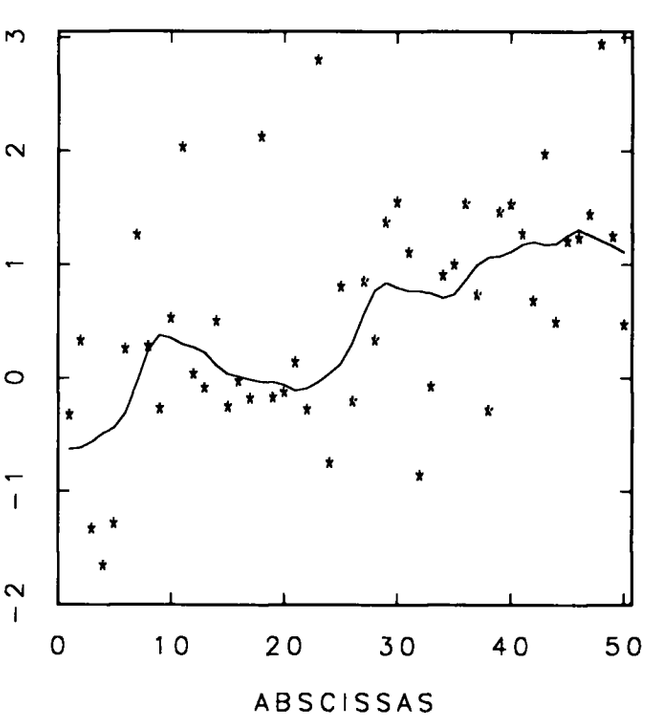
For the smoothed points in Figure B, $f = .5$, $d = 1$, $t = 2$, and the weight function is "tricube,"

$$W(x) = (1 - |x|^3)^3 , \quad \text{for} \quad |x| < 1$$
$$= 0 , \quad \text{for} \quad |x| \geq 1 .$$

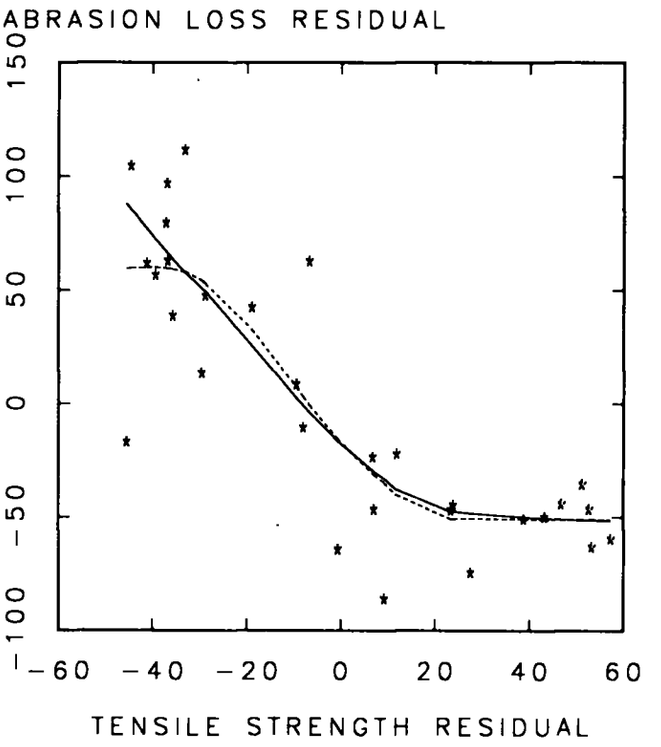
In Figure C, f has been decreased to .2 with the result that the smoothed points are "rougher" than those in Figure B. Section 4 contains guidelines and methods for choosing f, d, t, and W in practice.

The iterative fitting in steps 2 to 4 is carried out to achieve robust smoothed points in which a small fraction of outliers does not distort the results. The outliers, which can be thought of as arising when ϵ_i has a long-tailed distribution, tend to have small robustness weights, δ_k, and therefore do not play a large role in the determination of the smoothed points. The bisquare function is used because other investigations have shown it to perform well for robust estimation of location (Gross 1976) and for robust regression (Gross 1977).

C. Scatterplot of Artificially Generated Data and Robust Smoothed Values With f = .2



D. Scatterplot of Abrasion Loss Regression Residuals, Nonrobust Smoothed Values (Connected by Dotted Lines), and Robust Smoothed Values (Connected by Solid Lines)



Once the robustness weights δ_k have been determined, the fitted value at x (not necessarily equal to some x_i) can be computed by fitting a polynomial using the weights $\delta_k\ w_k(x)$. Thus the fitted values could, for example, be computed and plotted at an equally spaced set of points on the horizontal axis.

The smoothed points can be plotted by joining successive points by straight lines as in Figure B or by symbols at the points $(x_i, \hat{y}_i)$. When the smoothed points are superimposed on the scatterplot, the first method provides greater visual discrimination with the points of the scatterplot. But using lines raises the danger of an inappropriate interpolation. One possible approach is to use symbols initially when the data are being analyzed; then if a particular plot is needed for further use, such as presentation to others, the lines can be used if the initial plot indicates that linear interpolation would not lead to a distortion of the results. Another method is to plot the smoothed points separately with the same scales as the original scatterplot. This is particularly attractive for low-resolution plots such as printer plots.

The method of summarizing the scatterplot described here is appropriate when Y is the response or dependent

variable and X is the explanatory variable. In cases in which neither variable can be designated as the response, the scatterplot can be summarized by plotting the smoothed points of Y given X and the smoothed points of X given Y.

The smoothed points $(x_i, \hat{y}_i)$ portray the location of the distribution of Y given $X = x_i$. It is often useful to have, in addition, a summary of the scale. This can be done by plotting $|y_i - \hat{y}_i|$ versus x_i and computing and plotting smoothed points for this scatterplot.

3. EXAMPLES

3.1 Abrasion Loss Data

The importance of the robust procedure is illustrated in Figure D. The data are from a linear regression analysis (Box et al. 1957, p. 210) that related the abrasion losses of 30 rubber specimens to their hardnesses and tensile strengths. In Figure D the residuals from regressing abrasion loss on hardness are plotted against the residuals from regressing tensile strength on hardness. Superimposed on the plot are the smoothed points using locally weighted regression and robust locally weighted regression with $t = 2$. In both cases, $f = .5$, $d = 1$, and the weight function is tricube. The outlier in the lower left of the plot has substantially distorted the nonrobust smoothed points, while the robust smoothed points appear quite adequate. The smoothed points in this example show a substantial nonlinear effect; thus a regression model that is linear in the explanatory variables is not appropriate.

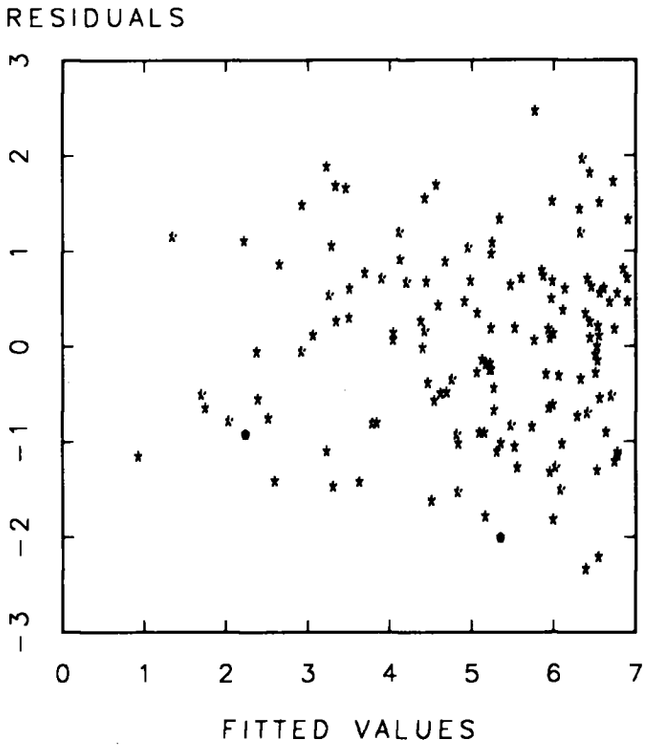
3.2 Residuals vs. Fitted Values

It has long been argued that plotting residuals against fitted values from a regression analysis is useful for, among other things, detecting a dependence of the scale of the errors on the level of the fitted values (Daniel and Wood 1971; Draper and Smith 1966). Such a plot has been made in Figure E for artifically generated data. The informal visual test is to look at the scale of the ordinates of the plot and determine if it is changing (e.g., increasing) with changing (e.g., increasing) values of the abscissa. The reader is invited to do this for Figure E.

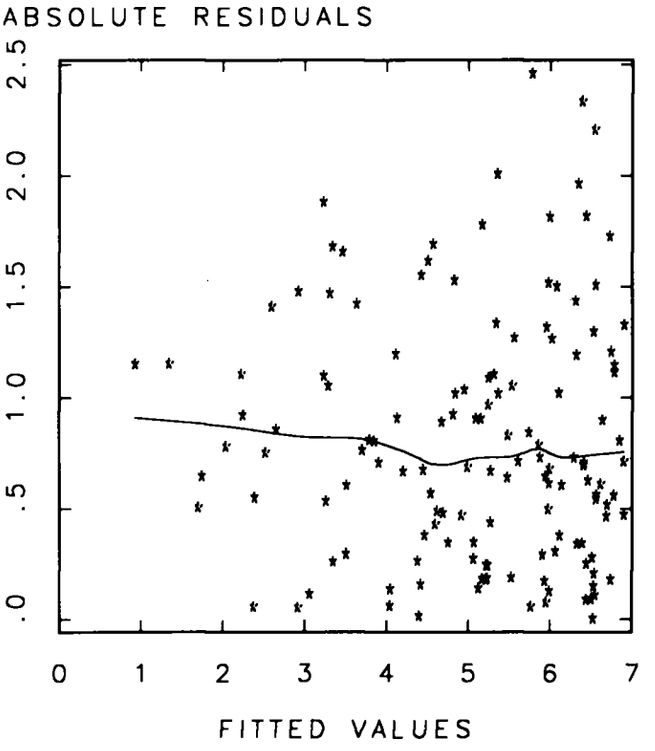
In fact, such an informal procedure is often confusing and too frequently misleading. For example, we might conclude from Figure E that the scale increases with increasing fitted values. In fact, the scale is constant. The misleading effect arises because the density of the points increases in going from left to right on the plot so that the ranges of the residuals tend to increase. Our visual assessment of scale is heavily dominated by our perception of the range, which of course does not properly measure scale because of the changing density.

A far better procedure for assessing the scale is to plot the absolute values of the residuals against the fitted values, superimpose smoothed points, and look for a consistent change. This has been done in Figure F for the same data plotted in Figure E. The plot correctly shows a constant scale since there is little change in the smoothed points.

E. Scatterplot of Residuals Against Fitted Values



F. Scatterplot of Absolute Values of Residuals Against Fitted Values and Robust Smoothed Values



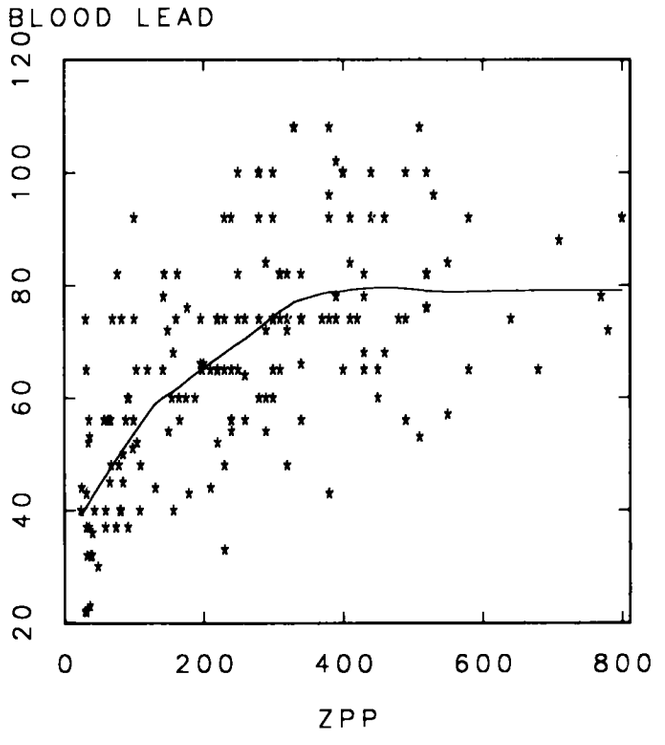
3.3 Lead Intoxication

Robust locally weighted regression has been used (Moody and Tukey 1979) in the investigation of the lead exposure of 158 workers in lead-smelting plants. The data involve two different screening methods for determining lead intoxication. The first is the traditional method in which lead levels in a blood sample are measured by atomic absorption spectrophotometry. The second, which is both newer and considerably simpler, is a hematofluorameter measurement of zinc protoporphyrin (ZPP), an enzyme released into the blood stream as a result of lead intoxication.

Figure G is a scatterplot of the blood lead versus ZPP level for the 158 workers. Superimposed on the plot are robust locally weighted regression smoothed values with $d = 1$, $f = .49$, the tricube weight function, and $t = 2$. The value of f was selected by using the cross-validation procedure described in Section 4.4. The purpose of computing the fitted values, $\hat{y}_i$, is to provide a typical blood lead value given the value of a ZPP measurement. The curve has a quadraticlike behavior for ZPP in the range 0 to 400 μg/dl and is constant for ZPP above 400 μg/dl.

For these data we are not in a situation in which there is a theoretical model to explain the dependence of blood lead on ZPP. Such a model would require a consideration of many physiological variables and a level of knowledge that does not now exist. Thus a summary of blood lead given ZPP must be determined empirically. It is clear that a single low-order polynomial would not adequately describe the entire curve in Figure G. We could attempt, of course, to find some other parametric family of curves to fit the data, but this would seem to require more effort than the relatively simple robust locally weighted regression.

G. Scatterplot of Blood Lead Against ZPP and Robust Smoothed Values (Units on both axes are μg/dl.)



4. CHOOSING d, W, t, AND f

There are four items that the user must select in order to carry out robust locally weighted regression: d, the order of the polynomial that is locally fit to each point on the scatterplot; W, the function used to determine the weights; t, the number of iterations of the robust fitting procedure; and f, the parameter used to determine the amount of smoothing. For the first three of these items certain preselected choices should serve almost all situations. Only f needs to be chosen on the basis of the properties of the data on the scatterplot.

4.1 Choosing d

Choosing d to be 1 appears to strike a good balance between computational ease and the need for flexibility to reproduce patterns in the data. The case $d = 0$ is the simplest, computationally, but in the practical situation an assumption of local linearity seems to serve far better than an assumption of local constancy because the tendency is to plot variables that are related to one another. For $d = 2$, however, computational considerations begin to override the need for having flexibility. Taking $d = 1$ should almost always provide adequate smoothed points and computational ease.

4.2 Choosing W

In (2.1) four requirements for W were described for the following reasons: (a) is necessary, of course, since negative weights do not make sense; (b) is required since there is no reason to treat points to the left of x_i differently from those to the right; (c) is required for it seems unreasonable to allow a particular point to have less weight than one that is further from x_i; (d) is required for computational reasons that are described in Section 5.

In addition it seems desirable that $W(x)$ decrease smoothly to 0 as x goes from 0 to 1. Such a weight function produces smoothed points that have a smooth appearance. That is, using time series terminology, the smoothed points have relatively small power at high frequencies. Among the weight functions that decrease to 0, tricube has been chosen since, as will be discussed in Section 6, it enhances a chi-squared distributional approximation of an estimate of the error variance. Tricube should provide an adequate smooth in almost all situations.

4.3 Choosing t

One procedure for carrying out the robust iterations would be to define a convergence criterion and iterate

until the criterion is satisfied. This seems needlessly complicated. Experimentation with a large number of real and artificial data sets indicates that two iterations should be adequate for almost all situations.

4.4 Choosing f

As stated earlier, increasing f tends to increase the smoothness of the smoothed points $(x_i, \hat{y}_i)$. The goal in the choice of f is to pick a value as large as possible to minimize the variability in the smoothed points without distorting the pattern in the data. In situations such as Figures B, C, D, and F where the sole purpose of the smooth is just to enhance the visual perception of patterns in the plot, the choice of f is not so critical since the eyes can partially correct for a less than optimal choice of f. For example, in Figure C the noisy smooth with $f = .2$ still provides a clear description of the increasing overall trend. In such situations choosing f in the range .2 to .8 should serve most purposes; in situations in which there is no clear idea of what is needed, taking $f = .5$ is a reasonable starting value.

In situations such as Figure G, where the smoothed values $(x_i, \hat{y}_i)$ are to be used as a regression function of y_i on x_i and might be communicated without the plot, more care in choosing f seems warranted. In such cases the PRESS procedure of Allen (1974), used ordinarily for choosing a subset of the independent variables in a regression, can be tailored to robust locally weighted regression to choose f. As in Section 2, the procedure begins with locally weighted regression (without the robust fitting) and iterates. Let $\hat{y}_i(f)$ be the locally weighted regression–fitted value of x_i for a given value of f with y_i not included in the computation. Then an initial value, f_0, of f is chosen by minimizing

$$\sum_{k=1}^{n} (y_k - \hat{y}_k(f))^2 .$$

Now let δ_k be the robustness weights for the residuals from the locally weighted regression fit with $f = f_0$ (as computed in step 2 in Section 2). Let $\hat{y}_i(f)$ be the fitted value at x_i for a given value of f with y_i not included in the computation and using the robustness weights δ_k (as in step 3 in Section 2). The next value of f is chosen by minimizing

$$\sum_{k=1}^{n} \delta_k (y_k - \hat{y}_k(f))^2 .$$

The procedure can then be repeated several times to produce a final value of f. For the blood-lead example described in Section 3.3 the successive values of f were $f_0 = .48$, $f_1 = .49$, $f_2 = .49$.

5. COMPUTATIONS

5.1 Reducing the Computations

Suppose the x_i are ordered from smallest to largest and let $x_{a(i)}, \ldots, x_{b(i)}$ be the ordered r nearest neighbors of x_i. The values of $a(i + 1)$ and $b(i + 1)$ can be found from $a(i)$ and $b(i)$ by using the following scheme:

1. Let $A = a(i)$ and $B = b(i)$.
2. Let $d_A = x_{i+1} - x_A$ and $d_B = x_{B+1} - x_{i+1}$.
3. If $d_A \leqslant d_B$, then $a(i + 1) = A$ and $b(i + 1) = B$. If $d_A > d_B$ replace A by $A + 1$ and B by $B + 1$ and return to step 2.
4. h_{i+1} is the maximum of $x_{i+1} - x_A$ and $x_B - x_{i+1}$.

Thus this scheme can be used to save computations by computing the fitted values at x_1, then x_2, and so on. Only $x_{a(i)}, \ldots, x_{b(i)}$ need be considered in the weighted least squares computation of $\hat{y}_i$ since $W(x) = 0$ for $|x| > 1$. This saving would not be achieved by using a weight function that becomes small but not zero for large x, such as the full normal probability density.

Portable FORTRAN programs that incorporate these savings are available from the author on request.

5.2 Grouping

The computations for the nearest-neighbor algorithm are approximately of the order fn^2. For scatterplots with fewer than 100 points, the computations present no problems. For plots with more points, computations can be saved simply by grouping the x_i. The saving results from the fact that if $x_{i+1} = x_i$ then $\hat{y}_{i+1} = \hat{y}_i$.

6. ESTIMATION AND SAMPLING DISTRIBUTIONS FOR LOCALLY WEIGHTED REGRESSION

In this section we shall suppose, as is generally done in ordinary least squares regression, that the ϵ_i are independent and identically distributed.

6.1 Estimation of the Error Variance and the Standard Errors of Fitted Values for Normal ϵ_i

Let us further suppose that the ϵ_i are normally distributed with variance σ^2. For such an error structure we would be content to smooth by locally weighted regression and not employ the robust fitting algorithm. Thus we shall suppose the fitted values $\hat{y}_i$ are the result of step 1 in Section 2.

Let R be the matrix whose (i, k)th element is $r_k(x_i)$. Let $\hat{\epsilon}_i = y_i - \hat{y}_i$ be the residuals. The fitted values and residuals have multivariate normal distributions with covariance matrices $\sigma^2 RR'$ and $\sigma^2 C$, respectively, where I is the identity matrix and $C = (I - R)(I - R)'$. Let $t_s = trC^s$. If we suppose the bias in the fitted values is negligible, then $E\hat{y}_i = g(x_i)$ and

$$\hat{\sigma}^2 = t_1^{-1} \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

is an unbiased estimate of σ^2. Thus the standard error of $\hat{y}_i$ may be estimated by

$$\hat{\sigma} \left(\sum_{k=1}^{n} r_k^2(x_i) \right)^{\frac{1}{2}} .$$

$\hat{\sigma}^2$ is a quadratic form in normal variables. A standard procedure for approximating the distribution of such a quadratic form (Box 1953) is to use a constant times a

chi-squared distribution whose first two moments match those of the quadratic form. Thus

$$t_1^2 t_2^{-1} \sigma^{-2} \hat{\sigma}^2$$

may be approximated by a chi-squared distribution with degrees of freedom equal to $t_1^2 t_2^{-1}$ rounded to the nearest integer. The chi-squared approximation will be enhanced if, in addition, we can make the third cumulants of the actual and the approximating distributions as close as possible by the proper choice of the weight function W. Straightforward calculations (Cleveland 1977) show that the tricube weight function provides such a third-moment match.

The quantity

$$\begin{aligned} \lambda &= n - \sigma^{-2} E \sum_{i=1}^{n} \hat{\epsilon}_i^2 \\ &= n - t_1 \\ &= 2 \sum_{i=1}^{n} r_i(x_i) - \sum_{i,k=1}^{n} r_k^2(x_i) \end{aligned}$$

can be used to assist in judging the relative amounts of smoothing for different values of f. If the $\hat{\epsilon}_i$ were the residuals from a linear least squares fit with q parameters, then λ would be equal to q. Thus, for locally weighted regression, λ can be interpreted as an equivalent number of parameters.

λ is not necessarily an integer, as in ordinary regression, but it is always nonnegative. To see this note that since $r_k(x_i)$, for $k = 1, \ldots, n$, result from a weighted least squares regression we have

$$r_k(x_i) = b_{ik} \left(\frac{w_k(x_i)}{w_i(x_i)} \right)^{.5} ,$$

where, for fixed i, $[b_{jk}]$ is an idempotent matrix with n rows and n columns. Since W has its maximum at 0, $w_i(x_i) \geqslant w_k(x_i)$. Thus

$$\begin{aligned} \sum_{k=1}^{n} r_k^2(x_i) &= \sum_{k=1}^{n} b_{ik}^2 w_k(x_i) w_i^{-1}(x_i) \\ &\leq \sum_{k=1}^{n} b_{ik}^2 \\ &= b_{ii} \\ &= r_i(x_i) . \end{aligned}$$

Thus

$$2 r_i(x_i) \geqslant \sum_{k=1}^{n} r_k^2(x_i)$$

and $\lambda \geqslant 0$.

Straightforward approximations (Cleveland 1977) show that for $d = 1$ and for the tricube weight function the quantity $2(1 + f^{-1})$ provides a good approximation of λ.

6.2 Estimating the Standard Error of the Fitted Values for More Generally Distributed ϵ_i

If we do not assume normality as in Section 6.1, then generally it will be wise to use the robust fitting procedure described in Section 2. Let $u_k = (y_k - \hat{y}_k)/6s$ and let $\theta_k = 1$ if $|\delta_k| > 0$ and let θ_k be 0 otherwise. Following Huber's (1973) suggestion for estimating standard errors in robust regression we might try estimating the standard error of $\hat{y}_i$ by

$$\hat{\sigma} \left(\sum_{k=1}^{n} r_k^2(x_i) \right)^{\frac{1}{2}}$$

where

$$\hat{\sigma}^2 = \frac{n^2}{n - \lambda} \left[\sum_{k=1}^{n} \delta_k^2 (y_k - \hat{y}_k)^2 \right] \cdot \left[\sum_{k=1}^{n} \theta_k (1 - u_k^2)(1 - 5u_k^2) \right]^{-2} .$$

More experimentation (e.g., Monte Carlo) with this estimate is needed in order to understand its properties.

7. VARIANCE, BIAS, AND MEAN SQUARED ERROR FOR LOCALLY WEIGHTED REGRESSION OF DEGREE ZERO

Suppose the y_i satisfy the model in (2.2) but with the additional assumption that the ϵ_i are independent with common finite variance σ^2. Let $\hat{y}$ be the fitted value at x (not necessarily equal to an x_i). The variance and bias of $\hat{y}$ are related to the mean squared error by

$$E(\hat{y} - g(x))^2 = (E\hat{y} - g(x))^2 + \operatorname{var} \hat{y} .$$

Let h be the distance of x to its rth nearest neighbor. Increasing the value of h tends to decrease the contribution of the variance term to the mean squared error, but runs the risk of increasing the bias. For locally weighted regression the variance of $\hat{y}$,

$$v(h) = \sigma^2 \sum_{k=1}^{n} r_k^2(x) ,$$

is generally (but not always) a nonincreasing function of h, since increasing h generally pools more information from the data. To illustrate this the behavior of $v(h)$ for the special case $d = 0$ will be investigated.

We shall begin with a lemma whose proof is from Colin L. Mallows. (In the lemma and the theorem to follow all summations run from 1 to n.)

Lemma: Let a_k and b_k for $k = 1, \ldots, n$ be two sequences of numbers with the following properties:

1. $a_k > 0$ and $b_k \geqslant 0$,
2. a_k, b_k, and b_k/a_k are nonincreasing sequences,
3. $\sum a_k = \sum b_k = 1$.

Then

$$\sum a_k^2 \leqslant \sum b_k^2 .$$

Equality occurs only if $a_k = b_k$ for all k.

Proof:

$$\begin{aligned} c &= \sum a_k b_k - \sum a_k^2 \\ &= \sum (a_k + a)((b_k)/(a_k) - 1) a_k , \end{aligned}$$

where a is any real number. Since $a_k + a$ and $b_k/a_k - 1$

are nonincreasing we may choose a so that the signs of these two sequences match. Thus $c \geqslant 0$. This inequality together with the Cauchy-Schwarz inequality for a_k and b_k proves the lemma.

The following theorem gives a necessary and sufficient condition that $\nu(h)$ be a nonincreasing function of h for locally weighted regression of degree 0.

Theorem: Let

$$\nu_k(h) = \frac{W(h^{-1}(x - x_k))}{\sum_j W(n^{-1}(x - x_j))},$$

where W is a weight function as defined in Section 2. (Note that for locally weighted regression with $d = 0$, we have $r_k(x) = \nu_k(h)$.) Let

$$\nu(h) = \sigma^2 \sum \nu_k^2(h)$$

and let

$$C(z) = \log W(e^z)$$

be defined for all real z such that $W(e^z) > 0$. Then $\nu(h)$ is a nonincreasing function of h for any set of x_i and any x if and only if C is a concave function.

Proof: Suppose $C(z)$ is concave. Let $\beta > \alpha > 0$, $a_k = \nu_k(\alpha^{-1})$, and $b_k = \nu_k(\beta^{-1})$. For simplicity of notation let us suppose $|x - x_k| = t_k$ is nondecreasing in k so that, since W is nonincreasing, we have a_k and b_k are nonincreasing. Furthermore, $a_k = 0$ implies $b_k = 0$, so that with no loss of generality we may suppose $a_k > 0$.

We shall now show that the sequence $b_k/a_k = c_k$ is nonincreasing. Suppose $b_k = 0$ for $k = s + 1, \ldots, n$, but $b_s > 0$. Then clearly c_k is nonincreasing for $k = s, \ldots, n$. Now suppose $t_k = 0$, for $k = 1, \ldots, r$, but $t_{r+1} > 0$. Then

$$\frac{c_{r+1}}{c_r} = \frac{b_{r+1}}{a_{r+1}} \frac{a_r}{b_r} = \frac{W(\beta t_{r+1})}{W(\alpha t_{r+1})}.$$

Since $\beta > \alpha$ and since W is nonincreasing we have $c_{r+1}/c_r \leqslant 1$. Thus c_k is nonincreasing for $k = 1, \ldots, r+1$. It remains to show c_k is nonincreasing for $k = r+1, \ldots, s$. For $k = r + 1, \ldots, s - 1$

$$\log \frac{c_{k+1}}{c_k} = \log \left[\frac{W(\beta t_{k+1})}{W(\beta t_k)} \frac{W(\alpha t_k)}{W(\alpha t_{k+1})} \right]$$
$$= [C(z_4) - C(z_3)] - [C(z_2) - C(z_1)],$$

where $z_4 = \log(\beta t_{k+1})$, $z_3 = \log(\beta t_k)$, $z_2 = \log(\alpha t_{k+1})$, and $z_1 = \log(\alpha t_k)$. Since $z_4 \geqslant z_3$, $z_2 \geqslant z_1$, $z_2 < z_4$, and $z_4 - z_3 = z_2 - z_1$ and since C is concave we have $\log c_{k+1}/c_k \leqslant 0$. Thus b_k/a_k is nonincreasing and from the lemma,

$$\sum a_k^2 \leqslant \sum b_k^2 .$$

Thus completes the proof of sufficiency.

To prove necessity suppose C is not concave. Then there exists $z_1 < z_2 < z_3 < z_4$ such that $z_2 - z_1 = z_4 - z_3$ and

$$C(z_2) - C(z_1) < C(z_4) - C(z_3) . \qquad (7.1)$$

Let $n = 2$, $x = 0$, $x_1 = e^{z_1}$, $x_2 = e^{z_2}$, and $\alpha = e^{z_3 - z_1}$. Furthermore let

$$a_k = W(\alpha x_k)(\sum W(\alpha x_j))^{-1}$$

and

$$b_k = W(x_k)(\sum W(x_j))^{-1} .$$

For the smoothed value at x,

$$\nu(\alpha^{-1}) = \sum a_j^2$$

and

$$\nu(1) = \sum b_j^2 .$$

Since $\log b_2 - \log b_1 = C(z_2) - C(z_1)$ and $\log a_2 - \log a_1 = C(z_4) - C(z_3)$ we have, from (7.1), $b_1/a_1 > b_2/a_2$. Thus, from the lemma,

$$\sum a_k^2 < \sum b_k^2 .$$

Since $\alpha^{-1} < 1$ we have proved necessity.

For the tricube weight function

$$C(z) = 3 \log (1 - e^{3z})$$

for $-\infty < z < 0$, and

$$C''(z) = \frac{-27e^{3z}}{(1 - e^{3z})^2},$$

which is negative. Thus C is concave and $\nu(h)$ is a nonincreasing function of h for tricube.

[*Received March 1978. Revised April 1979.*]

REFERENCES

Allen, David M. (1974), "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction," *Technometrics*, 16, 125–127.

Andrews, David F. (1974), "A Robust Method for Multiple Linear Regression," *Technometrics*, 16, 523–531.

Beaton, Albert E., and Tukey, John W. (1974), "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data," *Technometrics*, 16, 147–185.

Box, George E.P. (1953), "Normality and Tests on Variances," *Biometrika*, 40, 318–335.

———, Cousins, W.R., Davies, O.L., Hinsworth, F.R., Henney, H., Milbourn, M., Spendley, W., Stevens, W.L. (1957), *Statistical Methods in Research and Production* (3rd ed.), London: Oliver and Boyd.

Clark, R.M. (1977), "Non-parametric Estimation of a Smooth Regression Function," *Journal of the Royal Statistical Society*, Ser. B, 39, 107–113.

Cleveland, William S. (1977), "Locally Weighted Regression and Smoothing Scatterplots," Bell Laboratories memorandum.

Daniel, Cuthbert, and Wood, Fred S. (1971), *Fitting Equations to Data*, New York: John Wiley & Sons.

Draper, N.R., and Smith, H. (1966), *Applied Regression Analysis*, New York: John Wiley & Sons.

Ezekiel, M. (1941), *Methods of Correlation Analysis* (2nd ed.), New York: John Wiley & Sons.

Gross, Alan M. (1976), "Confidence Interval Robustness With Long-Tailed Symmetric Distributions," *Journal of the American Statistical Association*, 71, 409–416.

——— (1977), "Confidence Intervals for Bisquare Regression Estimates," *Journal of the American Statistical Association*, 72, 341–354.

Huber, Peter J. (1973), "Robust Regression: Asymptotics, Conjectures, and Monte Carlo," *Annals of Statistics*, 1, 799–821.

Macauley, Frederick R. (1931), *The Smoothing of Time Series*, New York: National Bureau of Economic Research.

Moody, Ivy, and Tukey, Paul A. (1979), "An Exploratory Analysis of Data on Lead Intoxication," Bell Laboratories memorandum.

Stone, Charles J. (1977), "Consistent Nonparametric Regression," *Annals of Statistics*, 5, 595–620.

Tukey, John W. (1977), *Exploratory Data Analysis*, Reading, Mass.: Addison-Wesley.