An Introduction to Kernel and Nearest Neighbor Nonparametric Regression

by

N.S. Altman

BU-1065MA

June 1991

An Introduction to Kernel and Nearest Neighbor Nonparametric Regression

N. S. Altman*

Biometrics Unit

Cornell University

Ithaca, NY14853

ABSTRACT

Nonparametric regression is a set of techniques for estimating a regression curve without making strong assumptions about the shape of the true regression function. These techniques are therefore useful for building and checking parametric models, as well as for data description. Kernel and nearest neighbor regression estimators are local versions of univariate location estimators, and so they can readily be introduced to beginning students, and consulting clients who are familiar with such summaries as the sample mean and median.

Key Words: Confidence intervals; Local linear regression; Model building; Model checking; Smoothing.

*N. S. Altman is Assistant Professor, Biometrics Unit, Cornell University, Ithaca, NY14853. Preparation of this article was partially funded by Hatch Grant 151410 NYF. The author thanks C. McCulloch for comments that substantially improved this article and L. Molinari and H. Henderson, for providing the data used in Examples B and C. Two anonymous referees and an associate editor provided numerous comments that substantially improved the presentation of the material.

1. INTRODUCTION

Nonparametric regression is a collection of techniques for fitting a curve when there is little a priori knowledge about its shape. The estimators discussed in this article provide estimates that are smooth functions, and the estimation procedure is called smoothing. Running averages, a very simple type of smoother, have been used since at least the late 1800's for determining trends in time series (for example, Wolfenden, 1942, attributes the method to De Forest in the 1870's). Since the 1970's there has been renewed interest in this area. A number of new smoothing techniques have been suggested, and their properties are becoming well-understood. This article introduces local location estimators such as kernel (Nadaraya 1964; Priestley and Chao 1972; Watson 1964) and nearest neighbor regression estimators (Benedetti 1977; Stone 1977; Tukey 1977) as simple extensions of ordinary univariate location estimators. These nonparametric regression estimators are powerful data-analytic tools, both as stand-alone techniques and as supplements to parametric analyses.

Estimators of location, such as the sample mean and median, are generally taught in elementary statistics courses, along with estimates of their precision. Summarizing a bivariate relationship using local location estimators is readily introduced to beginning students as an extension of these techniques.

Scatterplots are generally used to introduce bivariate relationships. Students have little difficulty with the idea of summarizing the trend in a scatterplot with a curve, fit by eye, particularly if the initial examples are not too scattered about the regression line. In my experience, students readily accept the idea that a more accurate summary may be obtained by dividing the scatterplot into vertical strips, and computing a location estimator in each strip. Error bars can be computed in each strip using univariate confidence intervals.

Practical application of this method generally leads the students to question how the strips should be located on the plot, and how the number of strips (or the width of the strips) should be chosen. These questions lead naturally to the idea of "moving" strips (windows) and selection of window size (bandwidth or span) that are central to kernel and nearest neighbor regression.

Introducing local means or medians as a summary of a bivariate relationship emphasizes to students that regression estimators attempt to represent the population location at fixed values of the predictor variables. Parametric fits can then be introduced as a means of summarizing the observed relationship with an equation, and the parametric and nonparametric fits can be compared as the first step in assessing goodness-of-fit of the parametric model. For example, the elementary text by Freedman, Pisani and Purves (1978, Chap. 10) makes good use of this method before introducing linear regression. When residual plots are introduced as diagnostic tools, it is natural to think of smoothing them as well, to detect trends not described by the parametric model.

There is an unfortunate lack of off-the-shelf software for nearest neighbor and kernel smoothing. However, related smoothers are now available in a number of software packages, including JMP (SAS Institute 1989), Minitab (Ryan, Joiner and Ryan 1985), S (Becker, Chambers and Wilks 1988) and Systat (Wilkinson 1988).

2. LOCAL LOCATION ESTIMATORS

The simplest nonparametric regression estimators are local versions of location estimators. For a random variable (t, y),the regression curve, $\mu(t) = E(y|t)$, shows how the mean of the dependent variable, y, varies with the independent variable, t.

If we are interested in estimation only at a single value of the independent variable, say, t^*, (and if t is under experimental control, so that we can sample the dependent variable at t^*), we would do best to sample only at this value. Then we could use a location estimator, such as the sample mean, trimmed mean, or median. Confidence intervals for this estimator would be formed using the usual intervals for the location estimator.

If we have several design points, $(t_1 \cdots t_n)$, with several replicates of the dependent variable at each design point, we can estimate $\mu(t_i)$ by a location estimator of the observations taken at t_i. Confidence intervals can also be computed at each point, or, if the variance is assumed to be constant, can be computed using a pooled estimate of the population variance.

This method is shown in Figure 1, using the sample mean as the location estimator. The data is mortality rate (y) as a function of average July temperature (t) in a set of American cities (Velleman 1988). As temperature was rounded to the nearest degree, there are replicates at many temperatures. In Figure 1, the data and $\bar{y}_{i.}$, the average mortality at temperature t_i, are plotted. (When there are no replicates, the sample average is just the observed data point.) Between data points, the average mortality rates are estimated by linear interpolation.

Two sets of normal theory confidence intervals are illustrated. If the variance of mortality is assumed to vary with temperature, confidence intervals for the mean should be based on local estimates of the variance,

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

where n_i is the number of data points at t_i and y_{ij} is the j^{th} data value taken at t_i. These intervals are shown by vertical bars in Figure 1. Notice that temperatures with no replicates do not have confidence intervals because no local estimate of standard error can be computed. The $1 - \alpha$ confidence interval at t_i is computed as

$$\bar{y}_{i.} \pm t(n_i - 1, \alpha)\hat{\sigma}_i / \sqrt{n_i},$$

where $t(h, \alpha)$ denotes the α^{th} quantile of the Student's t distribution on h degrees of freedom. Whenever the number of replicates is small, the interval is very wide, due to the small number of degrees of freedom for the t-statistic, and the large standard error of the sample mean.

If the variance of mortality is assumed to be constant, a pooled variance estimate can be used:

$$\hat{\sigma}^2 = \frac{1}{N - n} \sum_{i=1}^{n} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

where n is the number of design points, and $N = \sum_{i=1}^{n} n_i$ is the total number of data points. The ends of these intervals are shown on Figure 1 by circles. Confidence intervals can be computed at every data point, and the length of the confidence interval is inversely related to the square root of the number of replicates, since the degrees of freedom do not vary. The confidence interval at t_i is computed as

$$\bar{y}_{i.} \pm t(N-n, \alpha)\hat{\sigma}/\sqrt{n_i}.$$

The intervals are, in general, much shorter than those based on local estimates of variance.

Usually there is only one observation at each design point. However, if we know that $\mu(t)$ is smooth, points that are close together should have approximately the same mean. So, if we want to estimate $\mu(t^*)$, we could pick some neighborhood of t^* (a vertical strip on the scatterplot, as in Figure 2a), and proceed as if the data values falling in the neighborhood are actually a sample taken at t^*. The estimator is then

$$\begin{aligned} \hat{\mu}(t^*) &= \frac{1}{n^*} \sum_{t_i \in N(t^*)} y_i \\ &= \mu(t^*) + \frac{1}{n^*} \sum_{t_i \in N(t^*)} [\mu(t_i) - \mu(t^*)] + \frac{1}{n^*} \sum_{t_i \in N(t^*)} \varepsilon_i \end{aligned} \tag{1}$$

where $N(t^*)$ is the neighborhood, n^* is the number of data points in the neighborhood, y_i is the datum at t_i, and ε_i is the deviation of y_i from $\mu(t_i)$. Since the y-values in the neighborhood have mean close to, but not equal to, $\mu(t^*)$, this estimator has bias

$$Bias[\hat{\mu}(t^*)] = \frac{1}{n^*} \sum_{t_i \in N(t^*)} [\mu(t_i) - \mu(t^*)].$$

On the other hand, the estimate based on this subset of the data will have smaller variance than the estimate based on a single observation. If the goodness of the estimator is assessed by the Euclidean distance (squared error) between the estimator and the true regression function, the estimator based on the neighborhoods is an improvement if the variance decreases more than the squared bias increases.

If we have an unbiased estimator of the population variance, confidence intervals for the predicted values can be estimated using the usual normal theory approximations. The confidence intervals will, however, be centered around the biased estimate of the mean. Under repeated sampling using the same design points, the bias depends only on the unknown regression function and the design points. The confidence intervals will have the correct coverage properties for the expectation of the estimator, $\mu(t^*) + Bias[\hat{\mu}(t^*)]$, but not for the true regression curve. Adjusting confidence intervals so that they have the correct coverage for the true regression function is a topic of current research.

Two methods are commonly used to determine the size of the neighborhoods. Kernel estimators use strips of constant width (bandwidth). This is illustrated in Figure 2. In Figure 2a, the bandwidth is 0.1. The two strips have 6 and 3 points respectively. In Figure 2b, the bandwidth is 0.25. The strips have 12 and 10 points respectively. For a constant value of the bandwidth, the number of data points, and thus the variance of the estimator, varies from strip to strip. As the bandwidth increases, the number of points in the neighborhood is nondecreasing and so is the maximum distance between a point in the neighborhood and the point of estimation. As a result, the variance of the estimator decreases, but the bias, in general, increases.

Nearest neighbor estimators use strips of constant sample size (span). Usually the neighborhood is chosen so that an equal number of design points is taken from either side of the point of estimation. This is illustrated in Figure 3. In Figure 3a, the span is 5. The width of the first strip is 0.1, and the width of the second strip is 0.15. In Figure 3b, the span is 13. The width of the first strip is 0.28 and the width of the second strip is 0.35. Although width of the neighborhood varies from strip to strip, if the population variance is constant and there are no replicates, the variability of the estimator will be the same in every neighborhood (that is, the confidence intervals will all have the same width). For nearest neighbor estimators, it is not always clear how to handle replicates. In this paper we will base the span on the number of design points covered. The true sample size will be taken into account when forming confidence intervals.

Placement of the boundaries of the neighborhood can have a very strong effect on the regression estimator. This is avoided in practice by using moving strips, as in Figures

2 and 3. Instead of cutting the t-axis into fixed strips, the strip is moved along the axis, and centered at each estimation point in turn. Generally estimation is done only at observed design points, by centering the strip at the design point, and extended by interpolation between design points. For kernel estimators, this avoids the problem of empty intervals. However, in principle, for kernel estimators the strip can be moved continuously for estimation at each point on the t-axis. For nearest neighbor estimators, moving the strip continuously gives a step function estimator, as the nearest neighbors are constant between design points.

Figure 4 displays 4 regression estimates of the mortality data, with accompanying pointwise confidence intervals. (The pooled within variance has been used to estimate σ^2). In each plot, the heavier central line is the regression estimate. Figure 4a is the estimate based on the sample mean at each design point. The estimate is unbiased, but quite wiggly. The main features of the plot are the dips at about $68°F$ and $73°F$, and the peaks at about $71°F$ and $78°F$. However, there are a number of other local peaks.

Figure 4b displays an estimate based on fitting mortality with ordinary polynomial regression, using a polynomial of degree 3. The estimate is very smooth. The only feature of the plot is the peak near $78°F$ and a possible, shallow dip near $67°F$. Although polynomial regression can also be viewed as a nonparametric regression technique, it is somewhat more limited for exploratory analysis than kernel and nearest neighbor regression, due to the severe shape restrictions of low order polynomials. Unless the relationship is truly cubic, the polynomial regression estimator is also biased. Use of a higher degree polynomial is similar to using smaller span or bandwidth - the fit is less biased but more variable.

Figures 4c and 4d display, respectively, a kernel regression estimate based on a bandwidth of $6°F$, and a nearest neighbor estimate based on a span of 5. (The span was chosen to produce bandwidths close to $6°F$.) The plots are both quite similar to the cubic fit, although they are somewhat less smooth. The main difference between the polynomial and nonparametric curves is the shape of the bump, which appears to be skewed right in the polynomial fit, but is quite symmetric in the nonparametric fits.

3. ESTIMATING VARIANCE

To compute confidence intervals, an estimate of variance is needed. If the variance is assumed constant, some type of pooled estimator can be used. When the data contain replicates at most design points, and the variance is assumed constant, it is natural to use the pooled within variance, $\hat{\sigma}^2$, to estimate the population variance. If the data are normally distributed, $\hat{\sigma}^2/\sigma^2$ is distributed as a chi-squared on $N-n$ degrees of freedom (χ^2_{N-n}).

Usually, however, there are few replicates in the data. The residual sum of squares is a natural candidate for estimating variance. However, Equation (1) shows that the residuals are inflated by bias. When the choice of bandwidth or span is based on minimizing the squared error distance between the estimated and true regression functions, the bias and random error are of the same order of magnitude, so that the residual mean square is much larger than the true variance.

An idea that works well is detrending the data locally, and using the sample variance of the detrended data. When the design points are not too clustered, a simple, effective way to detrend for variance estimation is to use the pseudo-residuals $r_i = y_i - (y_{i+1} + y_{i-1})/2$ (Altman and Paulson 1990; Rice 1984). The variance estimator is then

$$\tilde{\sigma}^2 = \frac{2}{3(n-2)} \sum_{i=2}^{n-1} r_i^2 \qquad (2)$$

The distribution of $\tilde{\sigma}^2/\sigma^2$ can be approximated by χ^2_h where $h = (n-2)/2$ (Box 1954). The usual normal theory pointwise confidence intervals, using a Student's t ordinate on h degrees of freedom should then be adequate and will have the form:

$$\bar{y}_{i.} \pm t(h,\alpha)\tilde{\sigma}/\sqrt{n_i},$$

where $\bar{y}_{i.}$ is the average of the data in the neighborhood of t_i and n_i is the number of points in the neighborhood.

When the design points are highly clustered, more sophisticated detrending may be needed. The method of Gasser, Sroka and Jennen-Steinmetz (1986) seems to be effective for this situation. Altman and Paulson (1990) gives a simplification of some of the computations for this variance estimator.

4. REDUCING THE BIAS USING WEIGHTED AVERAGES

Equation (1) shows that kernel and nearest neighbor estimators are biased. Since we assume that points which are close together have means which are more similar than points which are far apart, it makes sense to use a weighted average, with smaller weight for points farther from the center of the strip. This decreases the bias of the estimator without much increase in its variance. Standard errors can be computed in the usual way, using the formula for the variance of a weighted average.

Since it is often useful to compare the smooth (regression estimate) for several spans or bandwidths, it is helpful if the weights can be defined so that they are readily adjusted to the size of the neighborhood. One way to do this is to define the weights using a function, $K(t)$, called a kernel weight function, that is large near zero and dies away to 0 as it reaches 1/2. For example, the quadratic kernel is the function:

$$K(t) = 6(1/4 - t^2) \qquad for\ |t| \leq 1/2.$$

Then, for a kernel estimator with bandwidth λ, the estimate at t^* is the weighted average

$$\hat{\mu}(t^*) = \sum_{i=1}^{n} w_i y_i / \sum_{i=1}^{n} w_i$$

where the weights are defined by $w_i = K[(t^* - t_i)/\lambda]$. The nearest neighbor estimator with span λ has a similar form, with weights defined by $w_i = K[(r^* - r_i)/(\lambda - 1)]$, where r_i is the rank of the ordered design points and r^* is the rank of t^* among the design points. The effect of changing the neighborhood size on the kernel weights is illustrated in Figure 5. As the neighborhood size is increased, more points fall in the neighborhood, but each point receives proportionately less weight.

The sum of the weights in the denominator is a normalizing factor, which is sometimes replaced by other expressions. The kernel function, $K(t)$, is generally chosen to be symmetric. However, it need not be unimodal or even positive. The quadratic kernel given above is optimal in a certain sense among positive kernels. See Gasser and Müller (1979) for a discussion of the choice of kernel function.

Another way of computing weights is to use local linear (or polynomial) regressions. Instead of a weighted mean, a linear regression estimate is computed in every neighborhood (Cleveland 1979; Cleveland and Devlin 1988; Friedman 1984). This is illustrated in Figure 6. Essentially, instead of estimating the mean at every point, the curve is approximated by estimating a tangent at every point. Local regressions are popular because if the points lie on a line (or polynomial), the line (or polynomial) will be reproduced. The computations can be done more rapidly than the computations for weighted averages based on kernel functions, and normal theory confidence intervals can still readily be produced.

Kernel, nearest neighbor and local linear estimators are all linear functions of the data - that is $\hat{\mu}(t) = \sum_{i=1}^{n} \gamma_i(t) y_i$ where $\gamma_i(t)$ are the kernel weights (for kernel and nearest neighbor regression) or the elements of the hat matrix (for local linear regression). As a result, pointwise normal theory confidence intervals can be computed in the usual way as

$$\hat{\mu}(t) \pm t(h,\alpha)\tilde{\sigma} \left(\sum_{i=1}^{n} \gamma_i^2(t) \right)^{1/2}$$

where $\tilde{\sigma}$ and h are defined by Equation (2). Once again, it should be noted that these intervals are centered about a biased estimator of the true regression function.

5. CHOOSING THE SIZE OF THE NEIGHBORHOOD

The problem of selection of the smoothing parameter (that is, the size of the nieghborhood) is closely related to the problems of selecting degree for a polynomial regression, or selecting variables in multiple regression. The need to avoid overfitting, and to "trade" bias for variance to obtain a better fit is very evident in nonparametric regression. When students understand these ideas, it is easy to introduce parametric model selection problems.

Figures 2 and 3 show how the regression estimate changes with neighborhood size.

Choice of the appropriate neighborhood size is critical to the performance of most nonparametric regression estimators. When the bandwidth or span is very small, the estimate will be very close to the original data, and so will be very wiggly. Due to overfitting, the estimate will be almost unbiased, but will have large variance under repeated sampling. At the other extreme, the estimate will be very smooth, lying close to the mean of all the data or to a simple parametric curve, such as a line or low order polynomial, depending on the form of estimator used. The estimate will have small variance, but will be highly biased.

One way to select the smoothing parameter is simply to look at plots of the smooth for several spans or bandwidths. If the overall trend is the feature of most interest to the investigator, a very smooth estimate may be desirable. If the investigator is interested in local extrema, a less smooth estimate may be preferred. Subjective choice of smoothing parameter offers a great deal of flexibility, as well as a comprehensive look at the data, and is readily introduced to beginning students. However, objective methods may be preferred in order to produce an automatic smoothing technique, or for consistency of results among investigators.

Model selection techniques based on measures of prediction error are often used for choosing the smoothing parameter. We could, for example, proceed by minimizing the least squares criterion, $\sum_{i=1}^{n} r_i^2$, where $r_i = y_i - \hat{\mu}(t_i)$ are the regression residuals. However, just as in variable selection, this criterion leads to fitting the largest available model. For nonparametric regression, this is the model with bandwidth 0 (span 1), $\hat{\mu}(t_i) = y_i$. The result is analogous to polynomial regression, where choosing the degree by minimizing the residual sum of squares also leads to $\hat{\mu}(t_i) = y_i$.

This overfitting occurs because, when the model is not known, the criterion is biased down for squared prediction error, and the bias increases as the bandwidth or span decreases. A number of less biased estimators of squared prediction error have been developed in the context of variable selection and other model building situations, and are applicable to bandwidth and span selection.

A popular heuristic is leave-one-out cross-validation, or PRESS (prediction sum of squares) (Allen 1974; Geisser 1975; Stone 1974). Case i is deleted from the data for predicting $\mu(t_i)$ giving the estimate

$$\hat{\mu}_{-i}(t_i) = \sum_{i \neq j} w_j y_j / \sum_{i \neq j} w_j$$

where the weights are defined as in Section 4. The deleted residuals

$$r_{-i} = y_i - \hat{\mu}_{-i}(t_i)$$

are then computed. The method of cross-validation then chooses the bandwidth or span to minimize the sum of squared deleted residuals, $\sum_{i=1}^{n} r_{-i}^2$. This method provides consistent estimates of the regression function (Härdle and Marron 1985). Although the rate of convergence of the smoothing parameter to its optimum is known to be slow (Härdle, Hall and Marron, 1986), the selected parameter value often works well even for moderate sample sizes. The computational burden is small, as simple algebra shows that $r_{-i} = r_i/[1-\gamma_i(t_i)]$, where $\gamma_i(t_i) = w_i/\sum_{j=1}^{n} w_j$, so that the regression estimate need only be computed once for each value of the smoothing parameter.

6. LARGE SAMPLE PROPERTIES

It can be seen intuitively that increasing the bandwidth of a smoother increases the bias, while increasing the span reduces the variance. These ideas can be made more precise by investigating the large sample properties of the regression estimators under some simplifying assumptions.

The assumptions that will be made are:

I) The average distance between design points is about $1/n$ for large sample sizes n. That is $|t_i - t_{i-1} - 1/n| = o(1/n)$.

This is required to ensure that there are no gaps in the data, since we cannot get a good estimate in or near a gap.

II) The regression function $\mu(t)$ has $p \geq 2$ square integrable derivatives. (Actually, only continuity is needed, but the algebra is more difficult.)

III) The errors are uncorrelated with mean 0 and variance σ^2.

We will also choose our kernel function, $K(t)$, so that

A) $K(t)$ is symmetric on $[-1/2, 1/2]$ and 0 off the interval.

B) $\int t^k K(t)dt = 0$ for $k < p$ and $\int t^p K(t)dt \neq 0$.

With these assumptions, the asymptotic bias of a kernel estimator can readily be computed using a Taylor series expansion around the true value $\mu(t)$. As the sample size n goes to infinity, and if the bandwidth λ is chosen so that λ goes to zero and $n\lambda$ goes to infinity, then:

$$Bias[\hat{\mu}(t)] = (-1)^p \lambda^p \mu^{(p)}(t) \int x^p K(x)dx/p! + o(\lambda^p),$$

where $\mu^{(p)}(t)$ is the p^{th} derivative of $\mu(t)$. The variance of the kernel estimator is:

$$Var[\hat{\mu}(t)] = \sigma^2 \int K^2(x)dx/n\lambda + o(1/n\lambda)$$

The results for nearest neighbor estimators are the same if the span is allowed to be $n\lambda$.

For a positive kernel, p must be 2. If the regression function is known to have more than 2 derivatives, the asymptotic bias of the estimator can be reduced by using a kernel that attains negative values. Also, notice that the bias is greatest where the function has large p^{th} derivative. The estimate is biased down in the neighborhoods of local maxima, and up in the neighborhoods of local minima. Kernel and nearest neighbor estimators erode hills and fill in valleys.

These results show explicitly the bias versus variance trade-off. The bias disappears when λ goes to zero. The variance disappears when $n\lambda$ goes to infinity.

The asymptotic mean squared error is:

$$\begin{aligned} E[\hat{\mu}(t) - \mu(t)]^2 &= Bias^2[\hat{\mu}(t)] + Var[\hat{\mu}(t)] \\ &= \lambda^{2p} \left(\mu^{(p)}(t) \int x^p K(x)dx/p!\right)^2 + \sigma^2 \int K^2(x)dx/n\lambda + o(\lambda^{2p}) + o(1/n\lambda) \end{aligned}$$

Ignoring the higher order terms and setting the derivative of this expression equal to zero shows that, asymptotically, the distance between the estimate and true value is minimized when $\lambda = Cn^{-1/(2p+1)}$, where C is a constant depending on the p^{th} derivative of $\mu(t)$, and on the kernel.

For the optimal value of λ, the bias and standard error of the estimate are the same order of magnitude. This explains why the residual mean square of the smooth is not a good estimate of σ^2 when the smoothing parameter is chosen to minimize mean squared error. As well, it shows that the centering of confidence intervals is an important problem. Intervals are centered around a biased estimate of the regression function, and the widths of the intervals are too small to compensate for the incorrect centering.

7. SOFTWARE

There is a shortage of off-the-shelf software for smoothing. However, for moderate sample sizes, weighted averages can readily be computed on a pocket calculator. Running medians of 3 or 5 provide rougher estimates, but can readily be computed by eye on a scatterplot.

The IMSL subroutines (IMSL 1984) include routines ICSSCU and ICSSCV which compute smoothing splines (Wahba 1990). (Spline smoothing is a more sophisticated smoothing technique, which produces results similar to kernel estimators.) Spline smoothing is also available in JMP (SAS Institute 1989). Smoothers based on running medians (Tukey 1977) are available in Minitab (Ryan, Joiner, and Ryan 1985), S, (Becker, Chambers and Wilks 1988) and Systat (Wilkinson 1988). Lowess (Cleveland 1979), a method based on local linear regressions, is available in S and Systat. Systat also offers unweighted averages. However, these packaged routines do not include estimates of the pointwise confidence intervals.

8. EXAMPLES

The 3 examples below demonstrate a number of uses of nonparametric regression estimation. Example A shows the use of nonparametric regression to summarize a complex regression relationship not readily captured by a parametric model. Example B, taken from Gasser, Müller, Köhler, Molinari and Prader (1984), shows how nonparametric regression can be used to supplement parametric modelling. Example C shows the use of nonparametric regression in model building and model checking for a discrete regression problem.

Smoothing in Example A was done using kernel regression with unweighted means. Error bars were computed using the variance estimator, $\tilde{\sigma}^2$, described in Section 3. Smoothing in Example B was done using the implementation of spline smoothing in JMP. Smoothing in Example C was done using kernel regression with quadratic weights. Error bars were computed using a local variance estimator.

Example A) Summarizing a nonlinear relationship

Figure 7 is a plot of the insurance market activity in ZIP code areas of Chicago as a function of theft rate (Andrews and Herzberg 1985). Market activity increases sharply with theft rate at low levels of theft, and then decreases. Kernel regression with bandwidth 12, chosen subjectively, has been used to smooth the data. The complicated shape of the curve could not readily be approximated by a parametric function, although polynomial regression provides a comparable fit.

The somewhat jagged appearance of the curve is due to the use of unweighted means. There is little data for theft rates beyond 50/1000. For the 3 highest theft rates, the neighborhoods contain only one data point, and the estimator simply interpolates the data. The wide error bands in this region reflect the sparcity of information.

Example B) Supplementing a parametric model

Parametric models for predicting human height have been under development since the 1930's (for example, Jenss and Bayley 1937). Recently developed models, (for example, Preece and Baines 1978), have very good predictive value. A parametric fit and residual plot for a child in a longitudinal study by the University of Zurich, (Gasser et al 1984) are displayed in Figures 8a and 8b. The fit was done using SAS PROC NONLIN (SAS Institute, 1988) and Preece and Baines Model 1. (The model was developed to fit growth after age 48 months.) The curvature following the initial peak evident in the residual plot was also observed in fits done by Preece and Baines, and attributed by those authors to autocorrelation in the data. However, most of the children show positive departures from the fitted curve at similar ages, indicating that this a systematic, not random, departure from the model.

Gasser et al analyzed the data using nonparametric regression. A fit using smoothing splines, with smoothing parameter chosen to give the same residual sum of squares as the parametric fit, is displayed with its residuals in Figures 8c and 8d. No systematic deviation appears in the residuals, except for the first few months when the spline curve cannot pick up the very rapid initial growth.

The source of the curvature in the residuals from the parametric model appears to be a mid-growth spurt, which occurs in most children around age 7. The form of the Preece and Baines model allows only a single growth spurt occurring in the adolescent years. A nonparametric estimate of growth rate shows the mid-growth spurt. This mid-growth spurt had been discussed in the early literature on human growth, but had disappeared from the literature following the development of parametric models which did not allow for it. Nonparametric regression, which has very weak assumptions on the shape of the regression curve, was able to pick up the extra peak.

Example C) Model building and model checking in generalized linear models

Nonparametric regression can provide great assistance in model building, particularly when the data is very noisy, or has other features which make patterns difficult to see. Binary response data is one example in which nonparametric regression can be useful, since scatterplots of the raw data and of regression residuals are often difficult to interpret. Figure 9a is a plot of survival of periparturient recumbent cows as a function of serum urea (Clark, Henderson, Hoggard, Ellison, and Young 1987). In cattle, increased serum urea may be due to a number of causes such as shock, increased protein catabolism and/or kidney damage. The asterisks are the observed proportions surviving. (Since there are few replicates, most of the proportions are 0 or 1.) The smooth indicates that survival increases and then falls, so a linear logistic curve is not appropriate.

The nonparametric fit was done using a kernel estimator with quadratic weights and bandwidth 0.2, chosen subjectively. The fitted curve (dark line) is smoother than the regression estimate in Example A due to the use of the quadratic weights. When uniform weights were used with the same bandwidth, the estimated regression function was quite jagged although it had the same general shape and features as shown in Figure 9a.

Estimated pointwise error bars were computed using local variance estimator $\hat{\sigma}_i^2 = \hat{\mu}(t_i)[1.0 - \hat{\mu}(t_i)]$, motivated by a Binomial model for the response. The estimated error bars show that the peaks at 1.9 and 3.0 are likely to be real features of the data. The peak at 1.4 may be spurious, due to sparser data in this region.

Figure 9b shows the same data fit with a quadratic logistic regression (dark line). Following Azzalini, Bowman and Härdle (1989), the goodness-of-fit of the parametric model is assessed by determining if the nonparametric fit falls within the parametric error bars. The peak of the parametric fit is located very near the primary peak of the smooth, but the smooth lies outside the error bars, indicating a sharper increase in survival than allowed by the quadratic model. Also, the quadratic fit does not allow the extra peak at 3.0. The quadratic model does not appear to be a good fit to this data.

Figure 9c is a plot of survival as a function of serum aspartate amino transferase (AST), a blood fraction which indicates muscle damage. The dark line is the logistic fit to the data. Except for small regions at extreme values of AST, probably caused by sparse data, the smooth lies entirely within the error bars, indicating that the logistic curve may be a reasonable model for the data.

Another way to use smoothing to check the model is to smooth the residual plots. This is a sensitive means of detecting nonlinearities in the data. In my experience, beginning students, in particular, find it easier to interpret residual plots if the plots are augmented by a smooth. Formal tests of goodness-of-fit of a parametric model versus smooth alternatives now exist for a variety of situations. Azzalini, Bowman and Härdle (1989) suggest formal tests for generalized linear models. Cox and Koh (1989), Cox, Koh, Wahba and Yandell (1988), and Eubank and Spiegelman (1990) suggest tests for linear and polynomial regression.

9. CONCLUDING REMARKS

Many other nonparametric regression techniques are available. They have not been discussed here, due to lack of space. Smoothing splines have many optimal properties, and are readily extended to complicated situations. Techniques using sequential knot selection, such as regression trees (Breiman, Friedman, Olshen and Stone 1984) and regression splines (Eubank 1988), are computationally and heuristically more complex, but are especially useful in multiple regression problems. Proper selection of smoothing parameters, such as span or bandwidth, seems to be critical to the success of all techniques.

Polynomial regression, with degree determined from the data, is the most popular nonparametric regression technique and is often taught in courses on multiple regression. Because the curve can be summarized by the regression coefficients, it is a useful technique for comparing curves, and for checking for nonlinearity. However, low degree polynomials do not offer the flexibility in shape of kernel and nearest neighbor estimators, limiting the usefulness of polynomial regression for data exploration and summary. The need to use polynomials of successively higher degree as the sample size increases (Eubank 1988) is seldom emphasized.

In this article, normal theory confidence intervals have been discussed. Confidence bands based on resampling techniques, such as the bootstrap, can also be used (Efron and Tibshirani 1986) and preserve the nonparametric flavor of the analysis.

Nonparametric regression techniques are flexible, powerful methods for estimating an unknown regression function. These techniques are useful in their own right, for data exploration, and estimation of the mean function, its derivatives, and features such as maxima and zeroes. They can also be used for model building and model checking in parametric regression. A number of texts giving fuller details of these methods have recently become available. These include Eubank (1988), Györfi, Härdle, Sarda and Vieu (1989), Härdle (1990), Müller (1990), and Wahba (1990).

Because the theory supporting nonparametric regression is more complicated than that of least squares linear regression, most treatments of the topic are in advanced texts such as those cited above. (A notable exception to this is Tukey 1977.) However, the heuristic motivation behind local location techniques can be easily understood. Computationally, local location techniques are no more difficult than the location estimators on which they are based. For these reasons, the powerful tools of nonparametric regression can readily be made accessible even to beginning statistics students. This paper has attempted to show, as well, that there are good pedagogical reasons for introducing nonparametric regression techniques prior to, or in parallel with, parametric techniques.

REFERENCES

Allen, D. M. (1974) "The relationship between variable selection and data augmentation and a method for prediction," *Technometrics*, **16**, 1307-1325.

Altman, N. S., and Paulson, C. P. (1990) "Some Remarks about the Gasser - Sroka - Jennen-Steinmetz Variance Estimator," Biometrics Unit Memo BU-1088-M, Cornell University.

Andrews, D. F., and Herzberg, A. M. (1985) *Data : a collection of problems from many fields for the student and research worker*, Springer Series in Statistics, Springer-Verlag, New York.

Azzalini, A., Bowman, A. W., and Härdle, W. (1989) "On the use of nonparametric regression for model checking," *Biometrika*, **76**, 1-11.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*, Wadsworth & Brooks/Cole Computer Science Series, Pacific Grove.

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

Box, G. E. P. (1954) "Some theorems on quadratic forms applied in the study of analysis of variance problems, I: effect of inequality of variance in the one-way classification," *Annals of Mathematical Statistics*, **25**, 290-302.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*, Wadsworth International Group, Belmont.

Clark, R. G., Henderson, H. V., Hoggard, G. K., Ellison, R. S. and Young, B. J. (1987) "The ability of biochemical and haematological tests to predict recovery in periparturient recumbent cows," *New Zealand Veterinary Journal*, **35**, 126-133.

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

Cleveland, W. S., and Devlin, S. J. (1988) "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, **83**, 596-610.

Cox, D. D., and Koh E. (1989), "A smoothing spline based test of model adequacy in polynomial regression," *Annals of the Institute of Statistical Mathematics*, **41**, 383-400.

Cox, D. D., Koh, E., Wahba, G. and Yandell, B. (1988) "Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models," *Annals of Statistics*, **16**, 113-119.

Efron, B. and Tibshirani, R. (1986) "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science*, **1**, 54-77.

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

Eubank, R. L. and Spiegelman, C. H. (1990) "Testing the Goodness of Fit of a Linear Model Via Nonparametric Regression Techniques," *Journal of the American Statistical Association*, **85**, 387-392.

Freedman, D., Pisani, R. and Purves, R. (1978) *Statistics*, W. W. Norton and Co. Inc., New York.

Friedman, J. H. (1984)"A Variable Span Smoother," LCS Technical Report 5 , Stanford, Department of Statistics.

Gasser, T., Müller, H. G. (1979) "Kernel estimation of regression functions," in ***Smoothing Techniques for Curve Estimation*** ed. Gasser T. and Rosenblatt, M., 23-67, Lecture Notes in Mathematics 757, Springer-Verlag, Heidelberg.

Gasser, T., Müller, H. G., Köhler, W., Molinari, L., Prader, A. (1984) "Nonparametric Regression Analysis of Growth Curves," ***Annals of Statistics***, **12**, 210-229.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C., (1986) "Residual variance and residual pattern in nonlinear regression," ***Biometrika***, **73**, 625-633.

Geisser, S. (1975) "The Predictive Sample Reuse Method with Applications," ***Journal of the American Statistical Association***, **70**, 320-328.

Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1989) ***Nonparametric Curve Estimation from Time Series***, Lecture Notes in Statistics 60, Springer-Verlag, Berlin.

Härdle, W., (1990) ***Applied Nonparametric Regression***, Cambridge University Press, Econometric Society Monograph Series, New York.

Härdle, W., Hall, P., and Marron, J. S. (1986) "How Far are Automatically Chosen Regression Smoothing Parameters From Their Optimum?" with discussion, ***Journal of the American Statistical Association***, **83**, 86-95.

Härdle, W., and Marron, J. S. (1985) "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," ***Annals of Statistics***, **13**, 1465-1481.

IMSL Incorporated (1984) ***IMSL Library Reference Library***, Houston.

Jenss, R. M. and Bayley, N. (1937) "A mathematical method for studying growth in children," ***Human Biology***, **9**, 556-563.

Müller, H. G. (1988) ***Nonparametric Regression Analysis of Longitudinal Data***, Lecture Notes in Statistics 46, Springer-Verlag, Berlin.

Nadaraya, E. A. (1964) "On estimating regression," ***Theory of Probability and its Applications***, **9**, 141-142.

Preece, M. A. and Baines, M. J. (1978) "A new family of mathematical models describing the human growth curve," ***Annals of Human Biology***, **5**, 1-24.

Priestley, M. B. and Chao, M. T. (1972) "Non-parametric function fitting," *Journal of the Royal Statistical Society, Series B*, **34**, 385-392.

Rice, J. (1984) "Bandwidth choice for nonparametric regression," *Annals of Statistics*, **12**, 1215-1230.

Ryan, B. F., Joiner, B. L. and Ryan, T. A. (1985) *Minitab Handbook Second Edition*, PWS-Kent Publishing Co., Boston.

SAS Institute Inc., (1985) *SAS User's Guide: Statistics Version 5 Edition*, Cary.

SAS Institute Inc., (1989) *JMP User's Guide*, Cary.

Stone, C.J. (1977) "Consistent Nonparametric Regression," *Annals of Statistics*, **5**, 595-645.

Stone, M. (1974) "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society Series B*, **39**, 44-47.

Tukey, J. (1977) *Exploratory Data Analysis*, Addison Wesley, Reading.

Velleman, P.F. (1988) *DataDesk Handbook Volume I*, Odesta Corporation, Northbrook.

Wahba, G. (1990) *Spline models for observational data*, Society for Industrial and Applied Mathematics, CBMS-NSF regional conference series in applied mathematics 59.

Watson, G.S. (1964) "Smooth Regression Analysis," *Sankhya, Series A*, **26**, 359-372.

Wilkinson, L. (1988) *SYSTAT: The System For Statistics*, SYSTAT, Inc., Evanston.

Wolfenden, H. H. (1942) *The Fundamental Principles of Mathematical Statistics*, The MacMillan Company of Canada, Toronto.

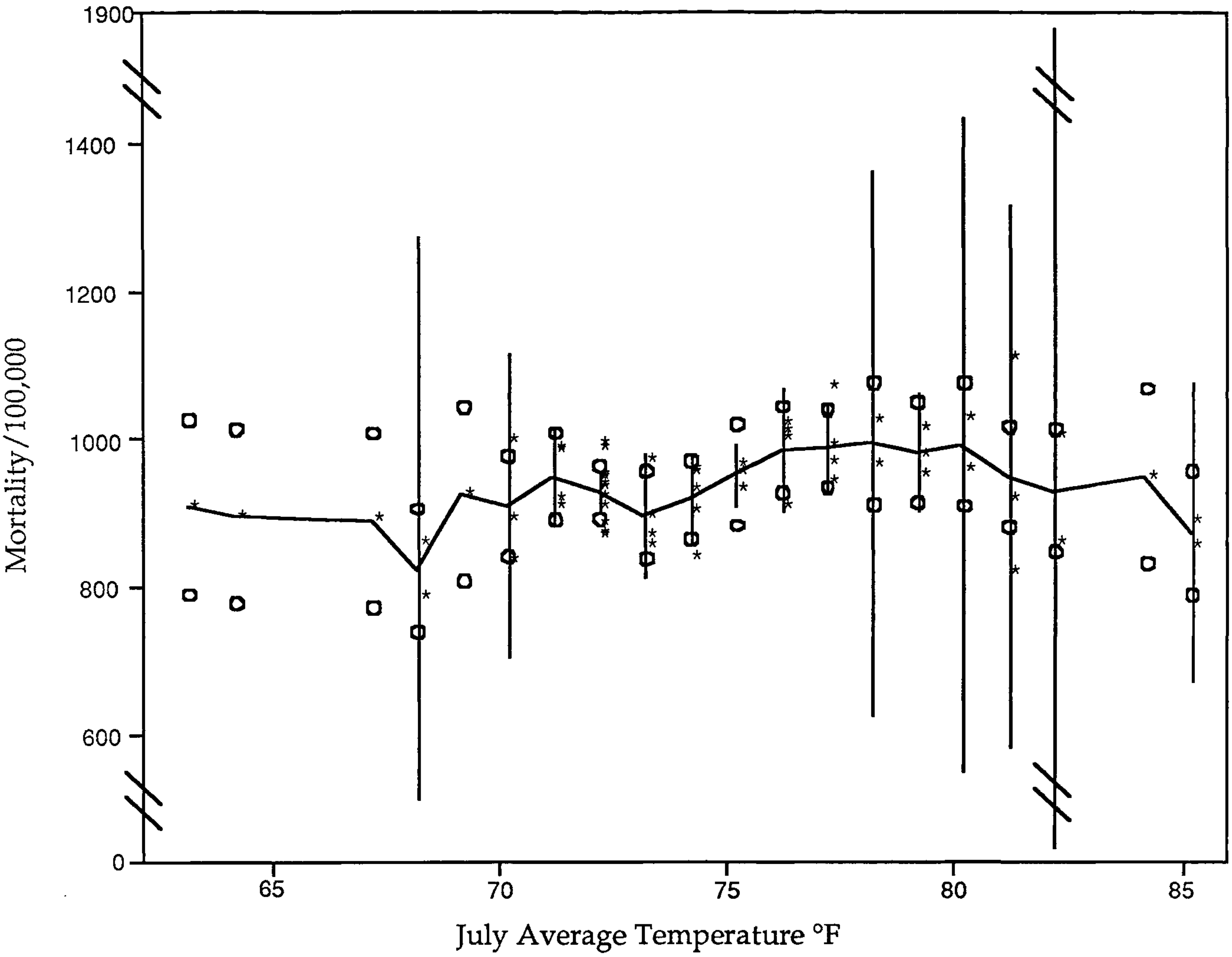
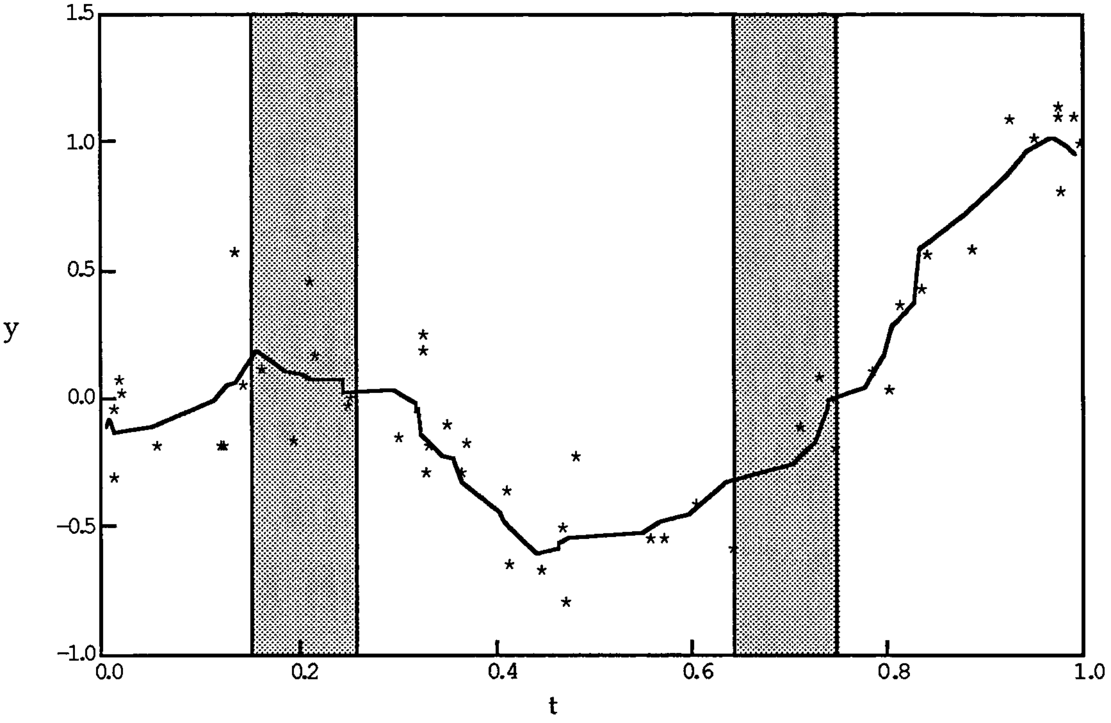
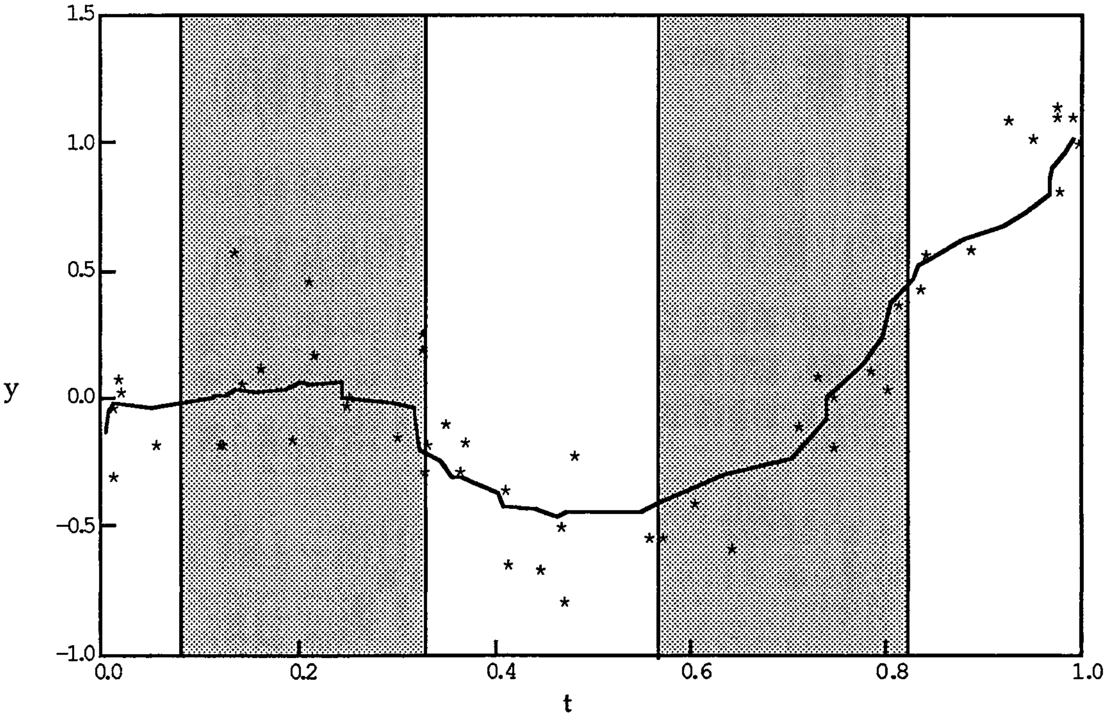


Figure 1: Mortality as a function of July average temperature in a sample of American cities. The curve joins the mean mortality at each temperature. The error bars are the 95% confidence intervals for the mean when the variance is estimated separately at each temperature. (There are no error bars for the 5 temperatures with no replicates.) The error bar at 82°F extends from 12.5 to 1855.5. The circles mark the endpoints of the 95% confidence intervals for the mean when the pooled variance estimate is used.

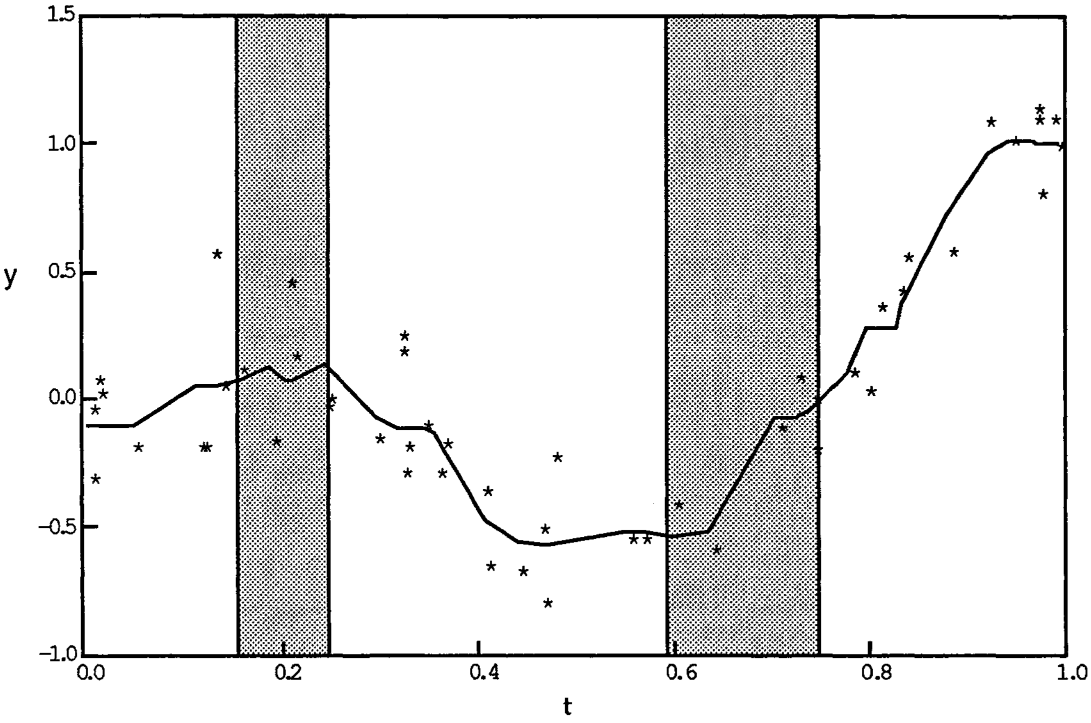


a. Bandwidth=.1

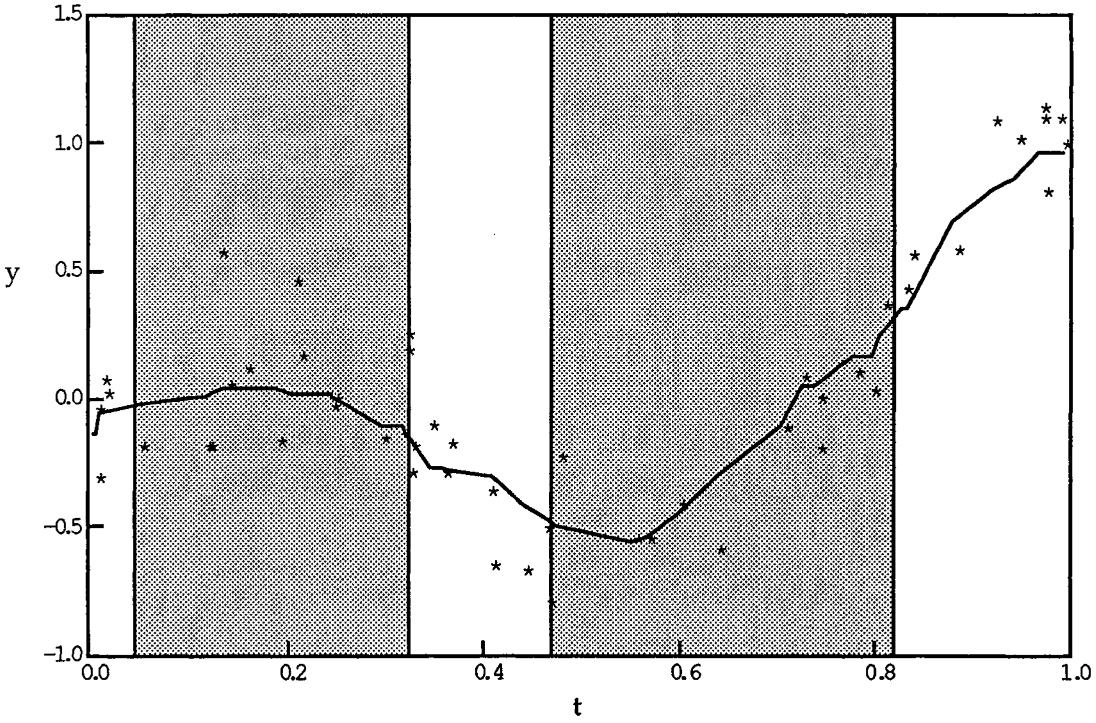


b. Bandwith=.25

Figure 2: Kernel estimate of the curve y=t sin(2.5πt)+ε at various bandwidths. The design points were generated from a Uniform(0,1). The errors were generated from a Normal(0,.01). The neighborhoods of t*=.2 and .7 are shown by the shaded strips on each plot.

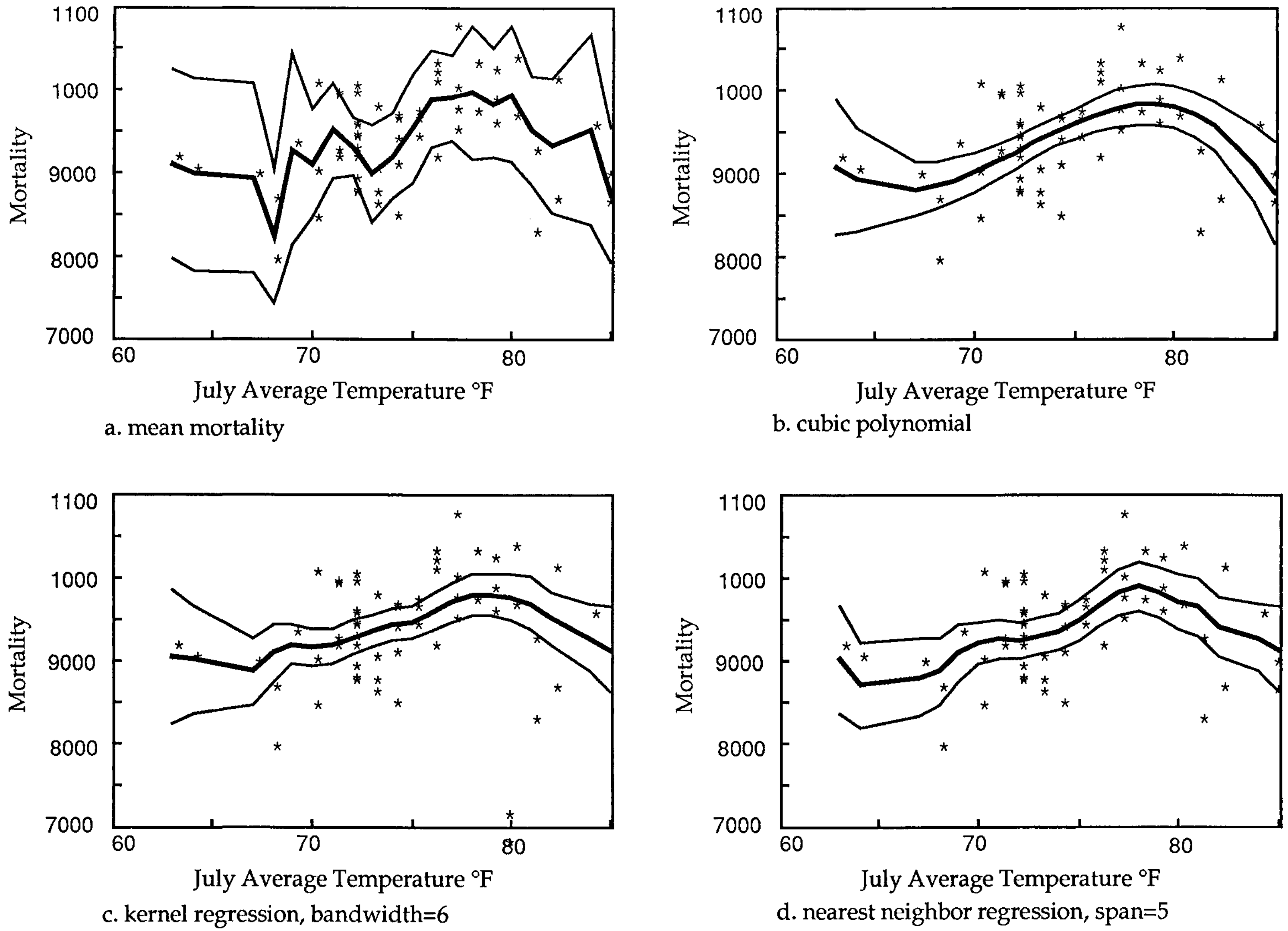


a. Span= 5



b. Span=13

Figure 3: Nearest neighbor estimate of the curve y=t sin(2.5πt)+ε at various spans. The design points were generated from a Uniform(0,1). The errors were generated from a Normal(0,.01). The neighborhoods of t*=.2 and .7 are shown by the shaded strips on each plot.



a. mean mortality

b. cubic polynomial

c. kernel regression, bandwidth=6

d. nearest neighbor regression, span=5

Figure 4: Mortality as a function of July average temperature in a sample of American cities. Various fits to the data including the pointwise mean (a) cubic polynomial (b) kernel regression (c) and nearest neighbor regression (d). In each case the pointwise 95% confidence intervals are given using a pooled variance estimate.

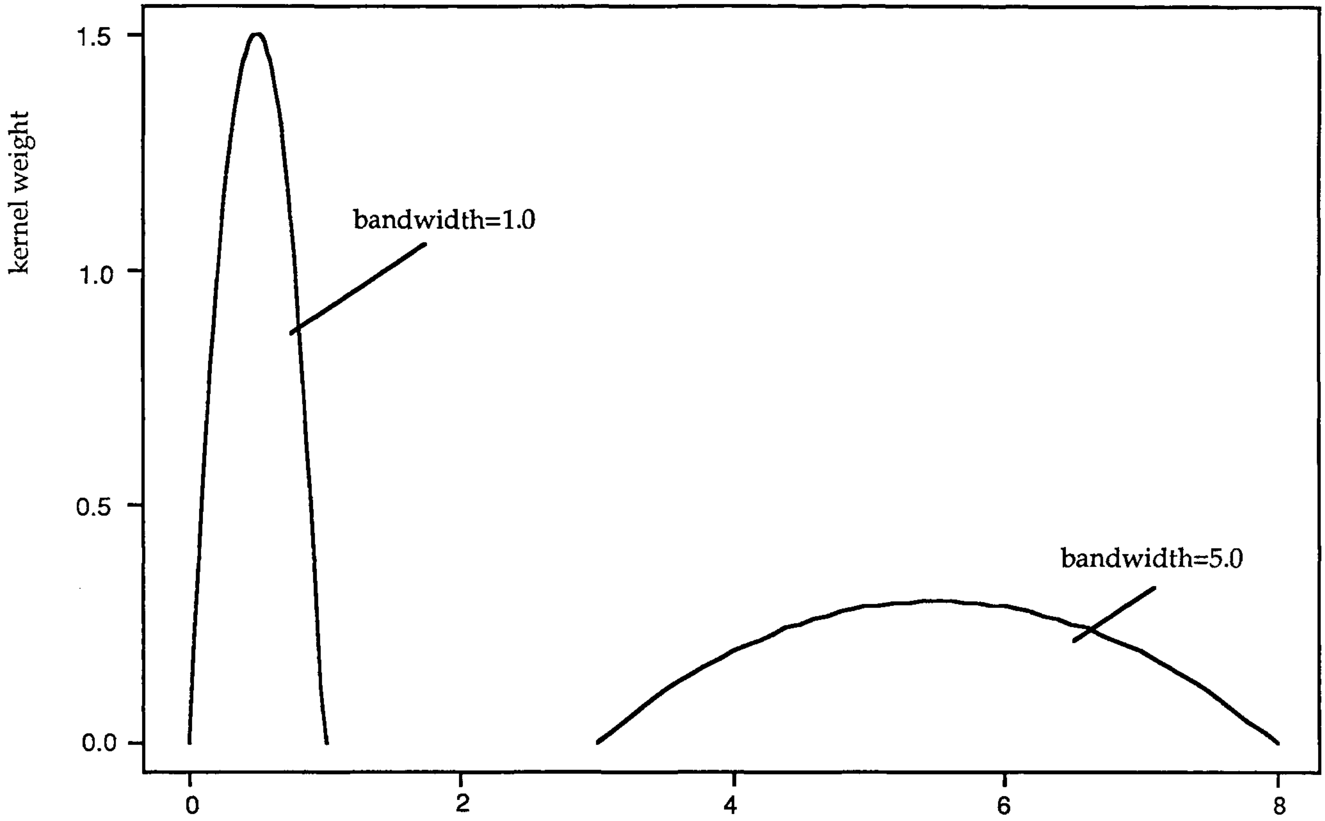


Figure 5: The effect of bandwidth on kernel weights
At larger bandwidths, more points get non-zero weight, but the weight of each is smaller.

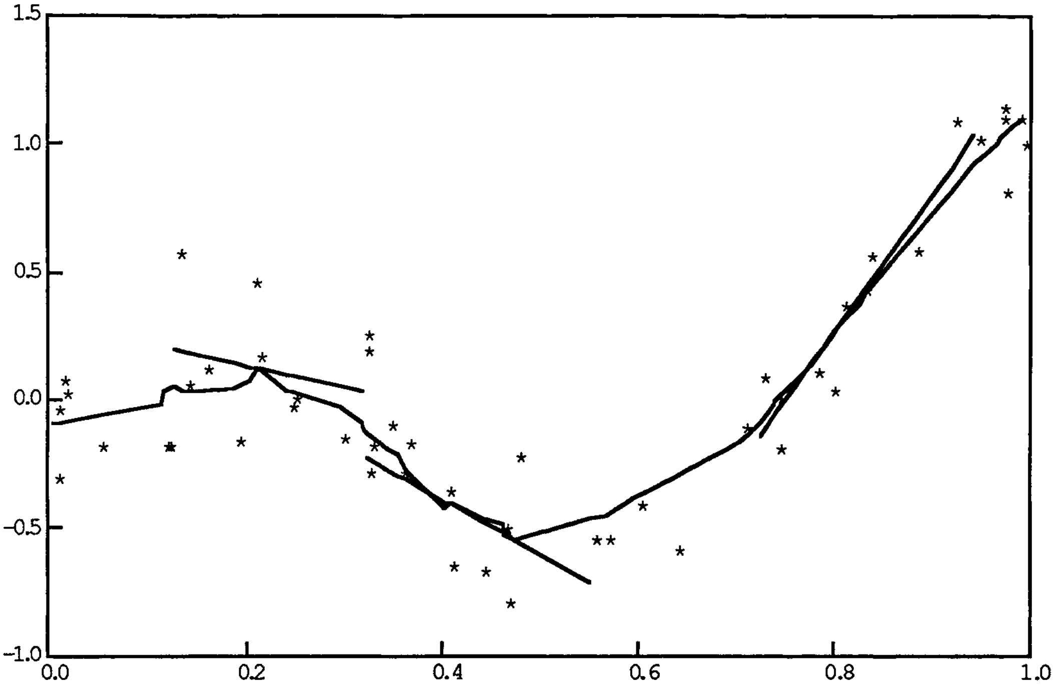


Figure 6: Running linear regression estimate of the curve y=t sin(2πt)+ε at span=.1. The design points were generated from a Uniform(0,1). The errors were generated from a Normal(0,.01). The local linear regressions are shown at 3 points along the curve.

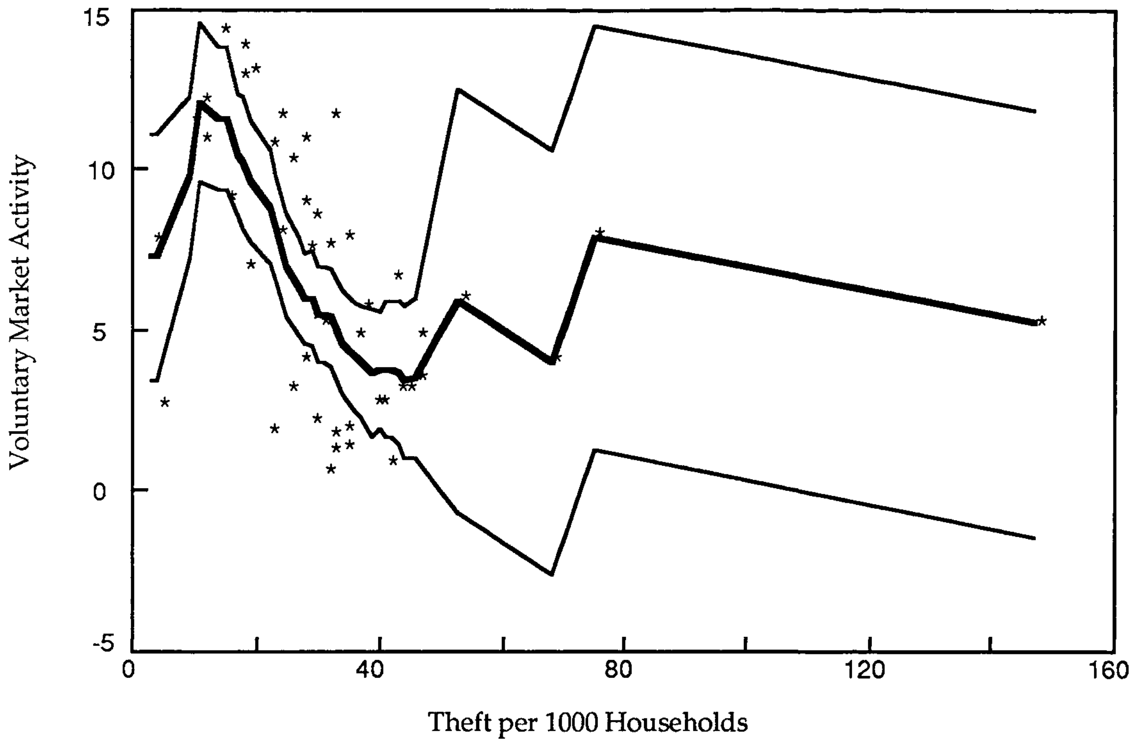


Figure 7: Voluntary market activity as a function of theft rate in Chicago neighborhoods summarized by ZIP code area.

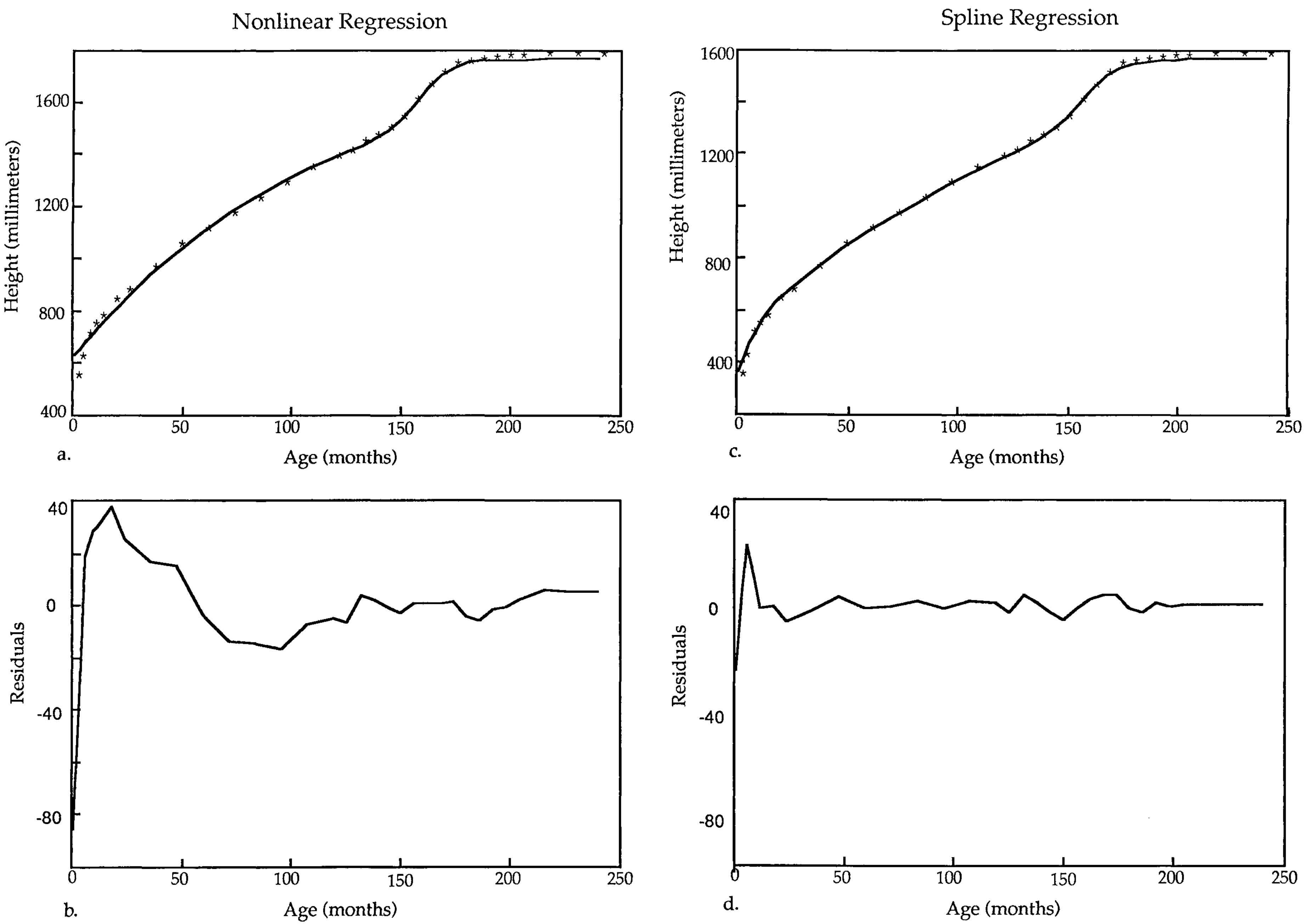


Figure 8: Parametric fit (a) and residual (b) and smoothing spline fit (c) and residual (d) to the height of a boy as a function of age.

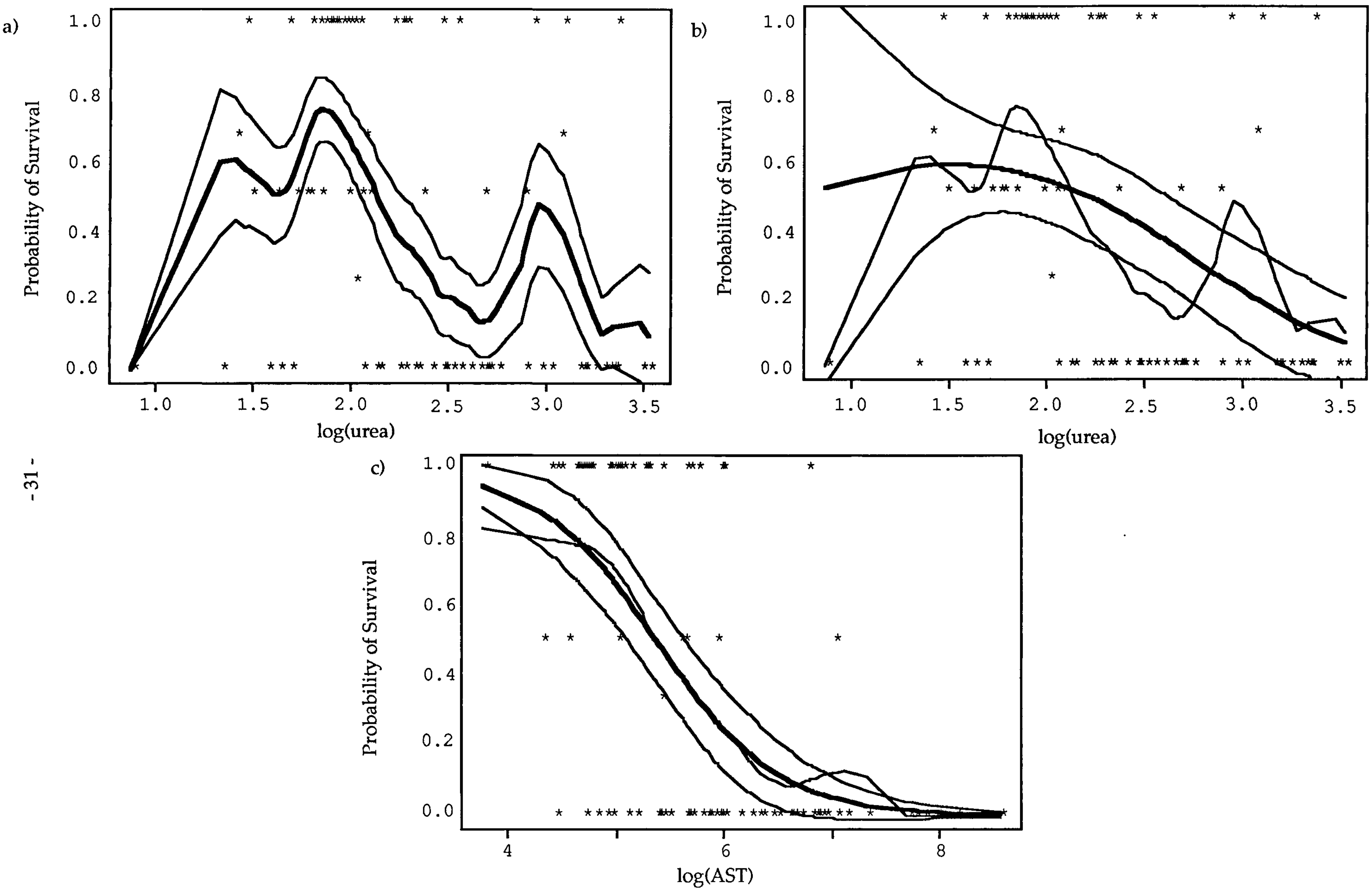


Figure 9: An example of the use of smoothing to supplement parametric modelling. The data is the survival of periparturient recumbent cows as a function of various blood serum measurements.