For regression methods that are based on residual sums of squares (such as F-tests), use of df_{res} is more appropriate than df_{fit}. Details are given in Section 6.6.2.

3.15 Other Approaches to Scatterplot Smoothing

Spline-based smoothers form just one class of the large collection of scatterplot smoothers developed over the years. In this section we briefly describe some of the other main classes.

3.15.1 *Local Polynomial Fitting*

One of the most popular methods for smoothing a scatterplot is local polynomial fitting. One of its advantages compared with spline-based smoothers is simpler theoretical analysis. This has allowed greater insight into the smoothing process. Summaries of this theory are given in Wand and Jones (1995), Fan and Gijbels (1996), and Loader (1999).

Figure 3.23 provides an illustration of the basic idea. The smooth at $x = u$ is obtained by fitting a weighted least-squares line where the weights correspond to the height of the *kernel function,* which is shown at the base of the plot. The estimate at $x = v$ is obtained similarly and also illustrated in Figure 3.23. If this procedure is applied over a grid of x-values then the solid curve results.

In Figure 3.23, local lines are being fitted. However, polynomials of any degree could be used. Let p be the degree of the polynomial being fit. At a point x, the smooth is obtained by fitting the pth-degree polynomial model

$$\mathsf{E}(y_i) = \beta_0 + \beta_1(x_i - x) + \cdots + \beta_p(x_i - x)^p$$

using weighted least squares with *kernel weights* $K\{b^{-1}(x_i - x)\}$. The kernel function K is usually taken to be a symmetric positive function with $K(x)$ decreasing as $|x|$ increases. For example, Figure 3.23 uses the standard normal density function. The parameter $b > 0$ is the smoothing parameter for local polynomial smoothers and is usually referred to as the *bandwidth*. The value of the curve estimate is the height of the fit $\hat{\beta}_0$, where $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \ldots, \hat{\beta}_p]^{\mathsf{T}}$ minimizes

$$\sum_{i=1}^{n} \{y_i - \beta_0 - \cdots - \beta_p(x_i - x)^p\}^2 K\left(\frac{x_i - x}{b}\right).$$

Assuming the invertibility of $\mathbf{X}_x^{\mathsf{T}}\mathbf{W}_x\mathbf{X}_x$, standard weighted least-squares theory leads to the solution

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_x^{\mathsf{T}}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^{\mathsf{T}}\mathbf{W}_x\mathbf{y},$$

where

$$\mathbf{X}_x = \begin{bmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^p \end{bmatrix}$$

is an $n \times (p + 1)$ design matrix and

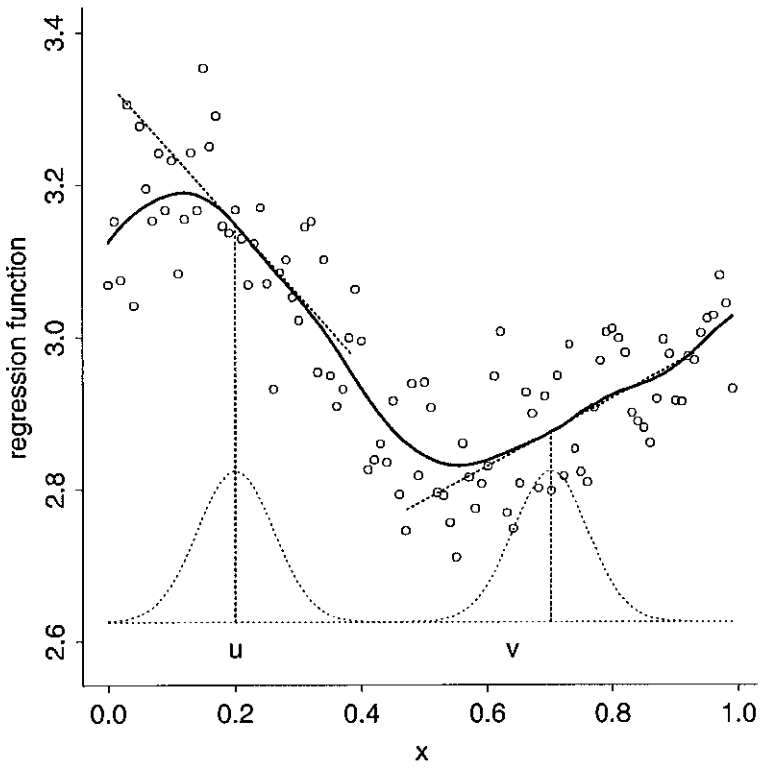

Figure 3.23 Local linear scatterplot smooth (solid curve) based on 100 simulated observations (represented by circles). The dotted curves are the kernel weights and cubic fits at the points u and v.

$$\mathbf{W}_x = \text{diag}\left\{K\left(\frac{x_1 - x}{b}\right), \ldots, K\left(\frac{x_n - x}{b}\right)\right\}$$

is an $n \times n$ diagonal matrix of weights. Since the estimator of $f(x) = \mathsf{E}(y|x)$ is the intercept coefficient, we obtain

$$\hat{f}(x; p, b) = \mathbf{e}_1^{\mathsf{T}}(\mathbf{X}_x^{\mathsf{T}}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^{\mathsf{T}}\mathbf{W}_x\mathbf{y},$$

where $\mathbf{e}_1$ is the $(p+1) \times 1$ vector having 1 in the first entry and 0 elsewhere.

The case $p = 0$ results in the *Nadaraya–Watson* (Nadaraya 1964; Watson 1964) estimator:

$$\hat{m}(x; 0, b) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x}{b}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{b}\right)}.$$

The data analyst must choose p and b. Our experience is that $p = 1$ works well if f appears to be monotonically increasing; otherwise, $p = 2$ is satisfactory. The bandwidth b can be chosen by trial and error with visual inspection, but it can also be chosen from the data using one of the automatic smoothing parameter selection approaches discussed in Chapter 5.

The Nadaraya–Watson (or "local constant") estimator has long been studied by theoreticians, but the local linear ($p = 1$) estimator seems to have been more

widely used in practice after the seminal paper of Cleveland (1979). The reasons for the superior practical performance of local linear over local constant estimation became clearer with the papers of Fan (1992, 1993). Near the boundaries of the data – and also in the interior, if the x are unequally spaced – local linear estimation is less biased than local constant estimation. Fan (1992, 1993) showed that, as $n \to \infty$ and $b \to 0$, the bias of $\hat{f}(x; b, p)$ is $O(b^2)$ for all x but the bias of $\hat{m}(x; b, p)$ is $O(b)$ at the boundaries and $O(b^2)$ at the interior. Ruppert and Wand (1994) showed that this effect of greater asymptotic bias near the boundary than in the interior holds for all even values of p. However, experience with data and simulation studies is required when interpreting this asymptotic result. The effect of this "boundary bias" is most severe for $p = 0$. In practice, $p = 2$ is an excellent choice for the degree of the local polynomials and is much less variable near the boundaries as compared to $p = 3$. In simulation studies, $p = 2$ often outperforms $p = 1$ and $p = 3$.

A sequence a_n is $O(c_n)$ if there exists a constant M such that $|a_n| \leq M|c_n|$ for n. In other words, a_n is bounded by a multiple of c_n. Thus, saying that the bias is $O(b^2)$ means that there is a constant M such that the bias when using bandwidth b is bounded in absolute value by Mb^2 for any b.

There are several variations on the basic local polynomial fitting idea depicted in Figure 3.23. Mostly they involve changing the value of the bandwidth across the estimation region. For example, the method of Cleveland (1979) sets the bandwidth so that the number of points used to estimate $f(x)$ is fixed, regardless of the estimation location x. The resulting scatterplot smooth is named LOESS (short for "local regression").

Relative to penalized splines, local polynomial regression is slow to compute if programmed directly. However, there are several strategies for speeding up the calculations (see e.g. Cleveland and Grosse 1991; Härdle and Scott 1992; Fan and Marron 1994; Seifert et al. 1994).

3.15.2 *Series-Based Smoothers*

Without loss of generality, assume that the regression function f is defined on the unit interval $[0, 1]$. Under certain regularity conditions, f admits the *Fourier series* representation

$$f(x) = \beta_0 + \sum_{j=1}^{\infty} \{\beta_j^s \sin(j\pi x) + \beta_j^c \cos(j\pi x)\}.$$

For higher values of j, the functions $\sin(j\pi x)$ and $\cos(j\pi x)$ become more oscillatory, as shown in Figure 3.24. The more oscillatory functions account for the finer structure in f. For smoother f, the corresponding coefficients will be small. This suggests the model

$$\hat{f}(x) = \hat{\beta}_0 + \sum_{j=1}^{J} \{\hat{\beta}_j^s \sin(j\pi x) + \hat{\beta}_j^c \cos(j\pi x)\},$$

where $\hat{\beta}_j^s$, $\hat{\beta}_j^c$ $(1 \leq j \leq J)$ and $\hat{\beta}_0$ are all estimated by least squares. The cut-off value J is the smoothing parameter in this case.

Other basis functions that are ordered by amount of oscillation may be used instead of the trigonometric basis functions. An example is the *Demmler–Reinsch*