Week 7

7.1 Nonparametric regression

Consider random variables Y and X, where Y is viewed as the "response variable" and X is viewed as a "predictor variable" and suppose that we are interested in the relationship between Y and X.

If the relationship between Y and X is believed to be a linear one, then we may consider a regression model of the form

$$Y = \alpha + \beta X + \epsilon$$

where α and β are unknown parameters and ϵ is an unobserved random variable such that $\mathrm{E}(\epsilon|X) = 0$. Then we may describe this model by the expression

$$\mathrm{E}(Y|X) = \alpha + \beta X$$

so that the linear function $\alpha + \beta x$ gives the mean value of Y corresponding to $X = x$.

Consider a sample of i.i.d. pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ such that each pair (X_j, Y_j) has the same distribution as (X, Y). Then, for some α, β,

$$Y_j = \alpha + \beta X_j + \epsilon_j, \quad j = 1, 2, \ldots, n,$$

where $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are i.i.d. random variables satisfying

$$\mathrm{E}(\epsilon_j|X_j) = 0, \quad j = 1, 2, \ldots, n;$$

it follows that

$$\mathrm{E}(Y_j|X_j) = \alpha + \beta X_j.$$

Under this model, we may estimate the parameters α, β using a method such as least-squares.

Recall that, in this type of regression model, the variables X_j might be non-stochastic, in the sense that they are chosen by the experimenter. In this case, we will continue to use the same notation as in the random-predictor case, writing, for example,

$$\mathrm{E}(Y_j|X_j) = \alpha + \beta X_j$$

to express the idea that the expected value of Y_j is a linear function of X_j. For simplicity, here we will use the terminology corresponding to random X_j; however, the methodology applies to both cases and it is a relatively simple matter to describe the results in terms of fixed predictors.

Although models based on linear relationships are commonly used, the relationship between Y and X might be nonlinear. For instance, it could be the case that

$$\mathrm{E}(Y|X) = \alpha + \beta_1 X + \beta_2 X^2$$

or

$$\mathrm{E}(Y|X) = \frac{\alpha}{\beta + X}.$$

In the first of these, we can estimate the parameters α, β_1, β_2 using standard least-squares methods; in the second, α and β can be estimated using a method such as nonlinear least-squares. An important feature of both of these models is that the relationship between Y and X is a *parametric* one, in the sense that the relationship can be described by a fixed (finite) number of parameters: α, β_1, β_2 in the first model and α, β in the second one.

However, in some cases, it may be reasonable to assume that Y and X are related, but we do not have enough information to be able to describe the relationship by a simple parametric model. In such cases, we may write

$$\mathrm{E}(Y|X) = m(X)$$

or

$$Y = m(X) + \epsilon \quad \text{where} \quad \mathrm{E}(\epsilon|X) = 0,$$

for an unknown "smooth" function $m(\cdot)$; that is, $m(\cdot)$ is an unknown, continuous function, for which one or more derivatives exists.

This is a *nonparametric* regression model and in nonparametric regression we are concerned with estimation of the function $m(\cdot)$ based on i.i.d. pairs of random variables (X_j, Y_j), $j = 1, 2, \ldots, n$, each with the distribution of (X, Y).

Consider estimating $m(x) = \mathrm{E}(Y|X = x)$ for some specific value x in the range of X. If several X_j are equal to x, we can estimate $m(x)$ by the average of the corresponding Y values:

$$\widehat{m}(x) = \frac{\sum_{j:X_j=x} Y_j}{\sum_{j:X_j=x} 1}.$$

Clearly, this is an unbiased estimator of $m(x)$.

However, such an approach will not be useful for estimating the entire function $m(\cdot)$. One reason is that, in many cases, most (if not all) X_j will be unique so that for $x = X_j$ we will estimate $m(x)$ by a single Y value, Y_j. Moreover, for nearly all values of x, there will be no X_j equal to x.

Hence, to deal with both of these issues, we can use *local averaging*, in which we take the average, or a weighted average, of the Y_j for which the corresponding X_j are "close to" x.

The general form of such an estimator is

$$\sum_{j=1}^{n} W_{nj}(x) Y_j$$

where $W_{nj}(x)$ denotes the weight given to Y_j when estimating $m(x)$; this weight generally depends on some measure of the distance between X_j and x.

Typically, the weights sum to 1:

$$\sum_{j=1}^{n} W_{nj}(x) = 1.$$

That is the case for the estimators we will consider in this course. Such an estimator is called a *linear smoother* because it is a linear function of the random variables $Y_1, Y_2, \ldots, Y_n$ and the result may be viewed as a "smooth curve" describing the relationship between Y_j and X_j.

7.2 Kernel regression

In *kernel regression*, the weight function $W_{nj}(\cdot)$ is based on a kernel function, as we used in density estimation; the kernel converts the distance between X_j and x to a weight applied to Y_j when estimating $m(x)$. Specifically,

$$W_{nj}(x) = \frac{K(\frac{x-X_j}{h})}{\sum_{j=1}^{n} K(\frac{x-X_j}{h})}$$

where $K(\cdot)$ is a kernel function and h is a smoothing parameter. As in density estimation, $K(\cdot)$ is generally taken to be a symmetric density function, such as the Gaussian kernel

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right), \quad -\infty < u < \infty.$$

Therefore, a kernel regression estimator has the form

$$\widehat{m}(x) = \frac{\sum_{j=1}^{n} K(\frac{x-X_j}{h}) Y_j}{\sum_{j=1}^{n} K(\frac{x-X_j}{h})}, \quad -\infty < x < \infty.$$

As in density estimation, the choice of the kernel has a relatively small effect on the estimator and, hence, we will always use the Gaussian kernel. As is also the case in density estimation, the the value of h controls the degree of smoothing.

When h is large, $(x - X_j)/h$ is close to 0 unless $|x - X_j|$ is very large and, hence, $K((x-X_j)/h)$ is relatively constant as j changes; that is, in the weighted sum defining $\widehat{m}(x)$, most of the Y_j receive a relatively large weight. The result is a smooth estimated regression function. When h is small, $(x - X_j)/h$ is large unless $|x - X_j|$ is very small. Because $K(\cdot)$ is the normal density function, if $(x - X_j)/h$ is large, then $K((x - X_j)/h)$ is close to 0. Hence, in the weighted sum defining $\widehat{m}(x)$, only those Y_j for which the corresponding X_j is close x receive large weights. The result is a less-smooth regression estimate. These properties are illustrated in the following example.

Example 7.1 A study was conducted on the relationship between the ratio of strontium isotopes in a fossil and the fossil's age. The R variable `sratio` contains the strontium ratios of 106 fossils and the variable `age` contains the corresponding ages of the fossils. For convenience, the values in `sratio` have been standardized by subtracting 0.707 and then multiplying by 10^4; these data are available in the dataset "fossil". Figure 7.1 contains a scatterplot of the data; note that the relationship between strontium ratio and age is clearly nonlinear and it does not appear to follow a low-order polynomial, such as a quadratic function.

Figure 7.2 contains scatterplots of strontium ratio vs. age with kernel estimates of the nonparametric regression function superimposed; four choices for the smoothing parameter were used, $h = 0.5, 1, 2, 4$. Note that for larger values of h the estimated regression function is smoother and it captures less of the local variation in the relationship.

The kernel estimates in Figure 7.2 were calculated using the function `sm.regression` in the package "sm". For instance, the following command was used to calculate and display the kernel estimate corresponding to $h = 1$:

```
> library(sm)
> sm.regression(x=age, y=sratio, poly.index=0, h=1, ngrid=1000)
```

Here the argument `poly.index` is set to 0 to specify that the kernel estimator described in this section is used; an alternative choice for this argument will be discussed later in this chapter. The argument `ngrid` specifies the number of points to use when plotting the kernel estimate; the default value (50) is rather low, and specifying a larger value results in a smoother plotted function.

The function `sm.regression` automatically produces a plot of the data together with the estimate of the regression function. You can also save the output of the function in a variable

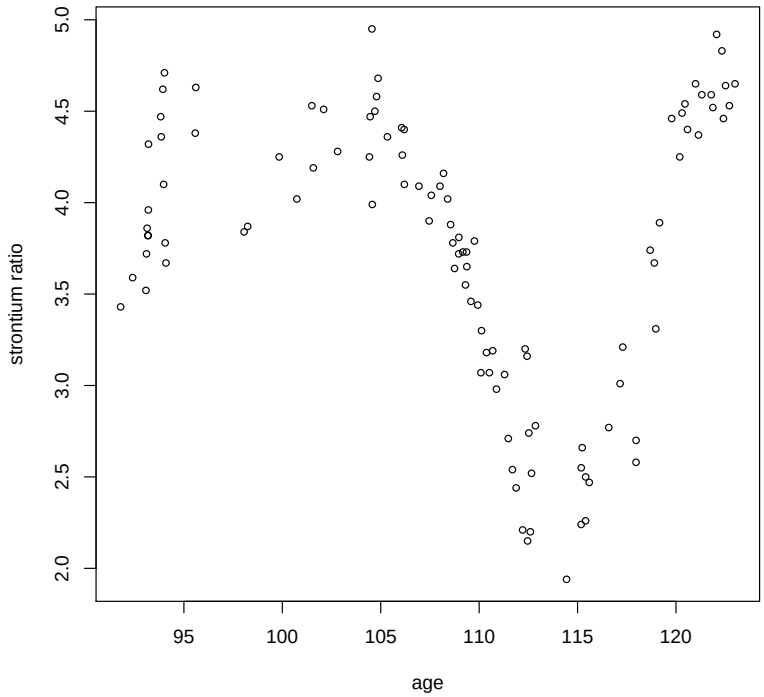

Figure 7.1: Plot of Strontium Ratio vs. Age for the Fossil Data

and access certain results, as components of the output. The most important of these are `eval.points`, the points at which the regression function estimate is evaluated as part of the estimation procedure, and `estimate`, which contains the corresponding estimates.

For instance, consider

```
> out<-sm.regression(x=age, y=sratio, poly.index=0, h=1, ngrid=1000)
```

so that the output from the estimation procedure is stored in `out`. Then

```
> head(out$eval.points)
[1] 91.79 91.82 91.85 91.88 91.91 91.94
> head(out$estimate)
[1] 3.784 3.789 3.795 3.800 3.805 3.810
```

Note that the function `head` returns the first part of a vector or matrix.

Hence,

$$\widehat{m}(91.79) = 3.784,\ \widehat{m}(91.82) = 3.789,$$

and so on. Because we specified `ngrid=1000`, the length of `out$eval.points` is 1000 and, hence, `out$estimate` also has length 1000.

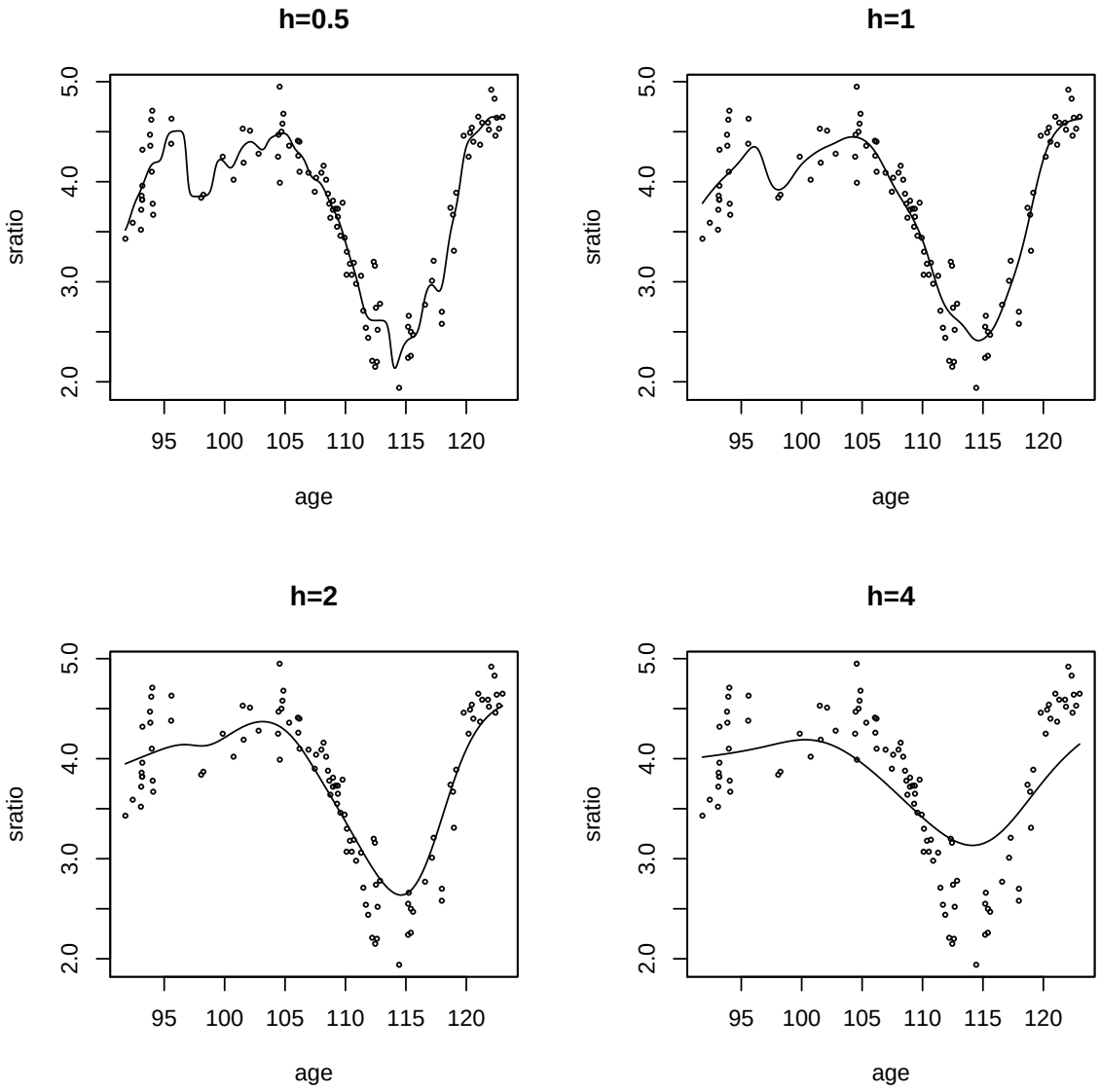

Figure 7.2: Kernel Estimates of the Regression Function for the Fossil Data Using Different Values of h

To find $\widehat{m}(x)$ at a point x not included the vector of evaluation points, we can use the function `approx`, which performs linear interpolation. For instance, to find $\widehat{m}(91.81)$, we can use

```
> approx(out$eval.points, out$estimate, xout=91.81)
$x
[1] 91.81

$y
[1] 3.788
```

Hence,

$$\widehat{m}(91.81) = 3.788.$$

□

7.3 Accuracy of kernel estimators

Let (X, Y) denote a pair of random variables and let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ denote i.i.d. random vectors such that for each $j = 1, 2, \ldots, n$, (X_j, Y_j) has the same distribution as (X, Y). For each x in the range of X, let

$$m(x) = \mathrm{E}(Y|X = x)$$

where $m(\cdot)$ is an unknown differentiable function. Note that, under these conditions, we may write

$$Y_j = m(X_j) + \epsilon_j \quad j = 1, 2, \ldots, n,$$

where $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are i.i.d. random variables satisfying

$$\mathrm{E}(\epsilon_j|X_j) = 0, \quad j = 1, 2, \ldots, n.$$

Under the conditions of the model, it may be shown that this can be extended to

$$\mathrm{E}(\epsilon_j|X_1, X_2, \ldots, X_n) = 0;$$

thus, $\mathrm{E}(Y_j|X_1, \ldots, X_n) = m(X_j)$.

We will consider the accuracy of the kernel estimator $\widehat{m}(\cdot)$ given by

$$\widehat{m}(x) = \frac{\sum_{j=1}^{n} K(\frac{x-X_j}{h})Y_j}{\sum_{j=1}^{n} K(\frac{x-X_j}{h})}, \quad -\infty < x < \infty$$

as an estimator of $m(\cdot)$. As with other estimators, we will consider two aspects of the properties of $\widehat{m}(x)$ as an estimator of $m(x)$: its bias and its variance. These can be combined to form the mean squared error; however, it also useful to consider them separately.

Recall that, in a linear regression analysis, the properties of the regression coefficient $\hat{\beta}$ are considered conditionally on the predictor values $X_1, X_2, \ldots, X_n$; for example, in a standard regression model with one predictor, the variance of $\hat{\beta}$ is reported as $\sigma_\epsilon^2 / \sum_{j=1}^{n}(X_j - \bar{X})^2$, where σ_ϵ^2 is the error variance. That is, when considering the hypothetical "repeated sampling" on which this variance is based, we consider samples using the same X-values as those actually observed. Hence, we use the same approach here, considering the conditional bias and conditional variance of $\widehat{m}(x)$.

In many respects, the analysis is similar to that used for kernel density estimators; however, because of the more complicated setting – e.g., we now have pairs of observations (X_j, Y_j), $j = 1, 2, \ldots, n$: the analysis is more difficult and, hence, we will focus on the main ideas rather than on the technical details (of which there are still many). Therefore, when reading this section, you should focus on the main results; the general approach used to obtain the results is described but not all details (which involve Taylor's series expansions along with algebraic manipulations) are given.

First note that we may write

$$\widehat{m}(x) = \frac{\frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{h})Y_j}{\widehat{p}(x)}$$

where

$$\widehat{p}(x) = \frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{h})$$

denotes the kernel density estimator of $p_X(\cdot)$, the marginal density of X.

An important property of $\widehat{p}(x)$ is that it depends only on $X_1, X_2, \ldots, X_n$, that is, it does not depend on $Y_1, Y_2, \ldots, Y_n$, so that when conditioning on $X_1, X_2, \ldots, X_n$, $\widehat{p}(x)$ is, in a sense, non-random. Hence, we focus on the properties of the numerator of the expression for $\widehat{m}(x)$,

$$\frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{h})Y_j.$$

Bias

Because $\mathrm{E}(Y_j|X_1, \ldots, X_n) = m(X_j)$,

$$\mathrm{E}\left(\frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{h})Y_j \;\middle|\; X_1, X_2, \ldots, X_n\right) = \frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{j})m(X_j)$$

and, hence,

$$\mathrm{E}\left(\widehat{m}(x)|X_1, X_2, \ldots, X_n\right) = \frac{\frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{j})m(X_j)}{\widehat{p}(x)}$$

It follows that the conditional bias of $\widehat{m}(x)$ is given by

$$\frac{\frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{j})m(X_j)}{\widehat{p}(x)} - m(x). \tag{7.1}$$

Note that

$$\frac{\frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{j})m(X_j)}{\widehat{p}(x)}.$$

is simply the kernel regression estimator applied to $m(X_1), m(X_2), \ldots, m(X_n)$ (instead of to $Y_1, Y_2, \ldots, Y_n$). This fact is convenient for analyzing the bias of $\widehat{m}(x)$ for different choices of $m(\cdot)$ and we will make use of it later in this chapter.

The expression (7.1) is useful for evaluating the bias of $\widehat{m}(x)$ for a specific set of data (for which the values of $X_1, X_2, \ldots, X_n$ are known) and for a specific choice of $m(\cdot)$. However, because it requires the values of $X_1, X_2, \ldots, X_n$, it is not useful for describing the general properties of $\widehat{m}(\cdot)$.

Hence, we consider the average conditional bias, given by the expected value of (7.1):

$$\mathrm{E}\left(\frac{\frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{j})m(X_j)}{\widehat{p}(x)}\right) - m(x). \tag{7.2}$$

A simple expression for this expected value is difficult (if not impossible) to obtain, because of the presence of $\widehat{p}(x)$ in the denominator. However, we know that $\widehat{p}(x)$ approaches $p_X(x)$ as $n \to \infty$ and $h \to 0$, where $p_X(\cdot)$ denotes the density function of X, so we can expand the ratio in terms of $\widehat{p}(x) - p_X(x)$:

$$\frac{\frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{j})m(X_j)}{\widehat{p}(x)} = \frac{\frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{h})Y_j}{p_X(x)} - \frac{\frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{h})Y_j}{p_X(x)}\frac{\widehat{p}(x)-p_X(x)}{p_X(x)} + \cdots. \tag{7.3}$$

Therefore, an expansion for the average conditional bias (7.2) can be obtained by computing the expected value of the expansion in (7.3). For instance, the first term has expected value

$$\begin{aligned}\frac{1}{p_X(x)}\mathrm{E}\left(\frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{j})m(X_j)\right) &= \frac{1}{p_X(x)}\mathrm{E}\left(\frac{1}{h}K(\frac{x-X}{h})m(X)\right)\\ &= \frac{1}{p_X(x)}\int_{-\infty}^{\infty}\frac{1}{h}K(\frac{x-t}{h})m(t)p_X(t)dt. \end{aligned}\tag{7.4}$$

We may now approximate this integral using the same technique used in analyzing kernel density estimators: use the change-of-variable $u = (x-t)/h$ and then use a Taylor's series expansion around $h = 0$, leading to

$$\frac{1}{p_X(x)}\int_{-\infty}^{\infty}\frac{1}{h}K(\frac{x-t}{h})m(t)p_X(t)dt = \frac{(m''(x)p_X(x) + 2m'(x)p'_X(x) + m(x)p''_X(x))}{p_X(x)}h^2 + O(h^4).$$

Note that the numerator on the right-hand side of this expression is simply the second derivative of $m(x)\,p_X(x)$, expanded by using the chain rule.

The same approach may be used for each term in the expansion in (7.3). However, note that $\widehat{p}(x) - p_X(x)$ converges to 0; for instance, recall that MSE of $\widehat{p}(x)$ as an estimator of $p_X(x)$ converges to 0 as the rate

$$O(h^4) + O(\frac{1}{nh});$$

see Section 5.1. It follows that the additional terms in the expansion are of successively smaller order and, hence, we only need to consider the first two terms on the right-hand side of (7.3).

The result is the following expansion for the average conditional bias of $\widehat{m}(x)$:

$$\frac{1}{2}\left(m''(x) + 2\frac{p'_X(x)}{p_X(x)}m'(x)\right)h^2 + O(h^4). \tag{7.5}$$

Note that, because for random variables Z, V, $\mathrm{E}(Z) = \mathrm{E}\left(E(Z|V)\right)$, the average conditional bias is the same as the unconditional bias.

Variance

Now consider the conditional variance of $\widehat{m}(x)$. Note that

$$\mathrm{Var}\left(\frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{h})Y_j \;\Big|\; X_1, X_2, \ldots, X_n\right) = \frac{1}{(nh)^2}\sum_{j=1}^{n} K(\frac{x-X_j}{h})^2\sigma^2(X_j)$$

where $\sigma^2(X) = \mathrm{Var}(Y|X)$. It follows that the conditional variance of $\widehat{m}(x)$ is given by

$$\mathrm{Var}\left(\widehat{m}(x) \;\Big|\; X_1, X_2, \ldots, X_n\right) = \frac{1}{\widehat{p}(x)^2}\frac{1}{(nh)^2}\sum_{j=1}^{n} K(\frac{x-X_j}{h})^2\sigma^2(X_j). \tag{7.6}$$

We now consider the average condtional variance by taking the expected value of the result in (7.6). As with the bias, to deal with $\widehat{p}(x)$ in the denominator, we first expand the right-hand side of (7.6) in terms of $\widehat{p}(x) - p_X(x)$. For instance, the first term in such an expansion is given by

$$\frac{1}{p_X(x)^2}\frac{1}{(nh)^2}\sum_{j=1}^{n} K(\frac{x-X_j}{h})^2\sigma^2(X_j).$$

Note that

$$\mathrm{E}\left(K(\frac{x-X}{h})^2\sigma^2(X)\right) = \int_{-\infty}^{\infty} K(\frac{x-t}{h})^2\sigma^2(t)p_X(t)dt.$$

The integral appearing in this expression can be approximated in the usual way: use the change-of-variable $u = (x-t)/h$ and a Taylor's series expansion that is valid for small h. t

Using this approach, it may be shown that the average conditional variance of $\widehat{m}(x)$ has the expansion

$$K_2\frac{\sigma^2(x)}{p_X(x)}\frac{1}{nh} + O(\frac{h}{n}) \tag{7.7}$$

where

$$K_2 = \int_{-\infty}^{\infty} K(u)^2 du.$$

Recall that, for random variables Z, V, the variance of Z may be written

$$\text{Var}(Z) = \text{E}\left(\text{Var}(Z|V)\right) + \text{Var}\left(\text{E}(Z|V)\right).$$

The average conditional variance corresponds to $\text{E}\left(\text{Var}(Z|V)\right)$; hence, in contrast to the bias, the average conditional variance is not the same as the unconditional variance.

Here the only bias and variance of $\widehat{m}(\cdot)$ that we will consider will be the the average conditional bias and average conditional variance; hence, when referring to these quantities, we will use the simpler terms "bias" and "variance", respectively.

Summarizing the accuracy of kernel estimators

According to the above results, $\widehat{m}(x)$ has bias

$$\frac{1}{2}\left(m''(x) + 2\frac{p'_X(x)}{p_X(x)}m'(x)\right)h^2 + O(h^4)$$

and variance

$$K_2\frac{\sigma^2(x)}{p_X(x)}\frac{1}{nh} + O(\frac{h}{n})$$

and these expressions have important implications for the properties of kernel estimators of $m(x)$.

- Like with kernel density estimators, as the smoothing parameter h increases, the variance of the kernel regression estimator decreases but the bias increases. The reason for this is that, for large h, $\widehat{m}(x)$ gives non-negligible weights to a larger number of Y-values, leading to a smaller variance. However, those Y-values will tend to correspond to X-values farther away from the "target" value of x, leading to larger bias.

- A larger value of $p_X(x)$ indicates that there tends to be many X values near x so that more Y values will receive non-negligible weight in $m(x)$. Therefore, for a given value of h, we expect the variance of $\widehat{m}(x)$ to be smaller when $p_X(x)$ is large. This explains the presence of $p_X(x)$ in the denominator of the expression for the variance of $\widehat{m}(x)$.

- If $p'_X(x)/p_X(x)$ is large, the density $p_X(\cdot)$ exhibits a relatively large proportional change in a neighborhood of x; hence, it is likely that more X-values are observed on one side of x than the other, leading to a larger bias of $\widehat{m}(x)$ if $m(\cdot)$ also changes considerably near x, that is, if $m'(x)$ is also large. Hence, the term $m'(x)p'_X(x)/p_X(x)$ appearing in the bias of $\widehat{m}(x)$ is not surprising.

- A large value of $m''(x)$ indicates that the function $m(\cdot)$ exhibits a high degree of curvature near x, leading to larger bias.

- A higher value of $\sigma^2(x)$ indicates higher variability of the Y-values corresponding to X values near x. These are the Y values that have the largest contribution to $\widehat{m}(x)$, leading to a larger variance of $\widehat{m}(x)$.

7.4 Choosing the value of the smoothing parameter

The usefulness of a kernel estimator of $m(\cdot)$ depends crucially on the numerical value of the smoothing parameter h used in constructing the estimator. Thus, in this section, we consider this choice.

The expressions for the bias and variance of $\widehat{m}(x)$, given by

$$\frac{1}{2}\left(m''(x) + 2\frac{p'_X(x)}{p_X(x)}m'(x)\right)h^2 + O(h^4)$$

and

$$K_2\frac{\sigma^2(x)}{p_X(x)}\frac{1}{nh} + O(\frac{h}{n}),$$

respectively, can be combined in the usual way to obtain the following expression for the MSE of $\widehat{m}(x)$:

$$\frac{1}{4}\left(m''(x) + 2\frac{p'_X(x)}{p_X(x)}m'(x)\right)^2 h^4 + K_2\frac{\sigma^2(x)}{p_X(x)}\frac{1}{nh} + O(h^6) + O(\frac{h}{n}). \tag{7.8}$$

Thus, the asymptotic MSE of $\widehat{m}(x)$ is given by

$$\frac{1}{4}\left(m''(x) + 2\frac{p'_X(x)}{p_X(x)}m'(x)\right)^2 h^4 + K_2\frac{\sigma^2(x)}{p_X(x)}\frac{1}{nh}. \tag{7.9}$$

Choosing h to minimize this expression yields

$$h = \frac{[K_2\sigma^2(x)/p_X(x)]^{\frac{1}{5}}}{[m''(x) + 2m'(x)p'_X(x)/p_X(x)]^{\frac{2}{5}}}\frac{1}{n^{\frac{1}{5}}}. \tag{7.10}$$

Thus, as expected, for larger n, a smaller value of h should be used, with the optimal value of h being of order $O(1/n^{\frac{1}{5}})$. Using this result in the expression for the asymptotic MSE shows that the optimal rate at which the MSE of $\widehat{m}(x)$ converges to 0 is $O(n^{-\frac{4}{5}})$.

The result (7.10) gives some rough guidelines for choosing h:

- If there is more variability around the true regression curve $m(\cdot)$, so that $\sigma^2(\cdot)$ is larger, then a larger value of h should be used, leading to more averaging, that is, more smoothing.

- If $m(\cdot)$ exhibits a high degree of local variation, as measured by $m'(\cdot)$ and $m''(\cdot)$, then a smaller value of h should be used, so that the estimator of $m(x)$ uses primarily Y-values corresponding to X-values near x. On the other hand, if $m(\cdot)$ is relatively constant, then a large value of h should be used; in the extreme case, in which $m(\cdot)$ is contant, h should be taken to be extremely large, so that $\widehat{m}(x)$ is essentially equal to the sample mean of $Y_1, Y_2, \ldots, Y_n$.

- The optimal value of h for estimating $m(x)$ depends on the value of x. Thus, when using a "global" smoothing parameter, i.e., one that applies for all x, we can expect to have too much smoothing for some x and too little for others, particularly if $m'(x)$, $m''(x)$, and $\sigma^2(x)$ vary considerably with x.

Although such information is useful for understanding the factors that affect the choice of h, the expression given in (7.10) is not particularly useful for determining the value of h to use because it depends on a number of unknown quantities that would be need to be estimated: $\sigma^2(x)$, $m''(x)$, $m'(x)$, $p'_X(x)$, and $p_X(x)$. In addition, this expression depends on the value of x under consideration.

One approach to choosing h is to use a subjective approach, in which a number of different values are considered and the selection is made so that the resulting estimate is useful for the goals of the analysis. Although that approach often works well in specific examples, it is also useful to have a more objective method available, even if it is only used as a starting point for a more subjective choice.

Because our goal is to choose a single value of h that applies to the entire estimator $\widehat{m}(\cdot)$, our choice of h should be based on "global" criteria that considers the properties of $\widehat{m}(x)$ for all x. One approach is to integrate the expression for the asymptotic MSE of $\widehat{m}(x)$, given by (7.9), times the density $p_X(x)$, in order to obtain a type of integrated MSE that can be used as the basis for our selection of h.

A simpler, but closely related, approach is to use an empirical average in place of integration, leading to the criterion

$$\rho(h) = \mathrm{E}\left(\frac{1}{n} \sum_{j=1}^{n} (\widehat{m}_h(X_j) - m(X_j))^2 \;\middle|\; X_1, X_2, \ldots, X_n \right), \quad h > 0;$$

here the kernel estimator is denoted by $\widehat{m}_h(\cdot)$ in order to emphasize the role of h in the estimator. A small value of $\rho(h)$ indicates a more accurate kernel estimator and, hence, the optimal choice of h minimizes $\rho(\cdot)$.

Of course, calculation of $\rho(\cdot)$ requires knowledge of the true function $m(\cdot)$, as well as the probability distribution needed to calculate the expected value used in the definition of $\rho(\cdot)$.

Thus, our goal is to estimate the function $\rho(\cdot)$ and then use that estimate to guide the choice of h.

Define

$$\hat{\rho}(h) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \widehat{m}_h(X_j))^2, \quad h > 0,$$

which uses Y_j as an estimtor of $m(X_j)$; recall that $\mathrm{E}(Y_j|X_j) = m(X_j)$. This is analogous to the "error sum of squares" as used in regression, divided by n, and $\hat{\rho}(h)$ may be viewed as an estimator of $\rho(h)$. Alternatively, we can simply think of $\hat{\rho}(h)$ as a measure of how well the function $\widehat{m}_h(\cdot)$ fits the data.

Unfortunately, the function $\hat{\rho}(\cdot)$ suffers from a crucial drawback: it is an increasing function of h so that smaller values of h always lead to smaller values of $\hat{\rho}(h)$.

Example 7.2 Consider the analysis of the fossil data in Example 7.1. Recall that Figure 7.2 contains plots of the data along with the kernel estimate of the regression function for $h = 0.5, 1, 2, 4$. For each value of h, we can compute $\hat{\rho}(h) = \sum_{j=1}^{n}(Y_j - \widehat{m}_h(X_j))^2/n$:

$$\hat{\rho}(0.5) = 0.047, \quad \hat{\rho}(1) = 0.061, \quad \hat{\rho}(2) = 0.099, \quad \text{and} \quad \hat{\rho}(4) = 0.248.$$

Looking at Figure 7.2, it is easy to see why smaller values of h lead to smaller values of $\hat{\rho}(h)$: when there is less smoothing (i.e., when h is closer to 0), there is less local averaging so that $\widehat{m}(\cdot)$ tends to track the observed data (X_j, Y_j), $j = 1, 2, \ldots, n$ more closely. In particular, very small values of h yields values of $\hat{\rho}(h)$ that are very close to 0:

$$\hat{\rho}(0.2) = 0.036, \quad \hat{\rho}(0.1) = 0.027, \quad \text{and} \quad \hat{\rho}(0.05) = 0.018.$$

□

Note that

$$\lim_{h\to 0} K(\frac{x - X_j}{h}) = \begin{cases} K(0) & \text{if } x = X_j \\ 0 & \text{if } x \neq X_j \end{cases};$$

hence, it is straightforward to show that

$$\lim_{h\to 0} \widehat{m}_h(X_j) = Y_j.$$

Thus, the property illustrated in Example 7.2 holds in general and the value of h that minimizes $\hat{\rho}(h)$ over $h > 0$ is effectively 0. One explanation for this behavior is that, because all of $Y_1, Y_2, \ldots, Y_n$ are used when predicting Y_j, we choose $h = 0$ so that the predictor of Y_j is simply Y_j.

Another explanation is based on the properties of $\hat{\rho}(h)$ as an estimator of $\rho(h)$: $\hat{\rho}(h)$ is a biased estimator of $\rho(h)$. As in the density estimation case, this bias arises from the fact

that the random variables $Y_1, Y_2, \ldots, Y_n$ are used in two different contexts: one as the points to which $\widehat{m}_h(\cdot)$ is compared and the other as the data points used to calculate $\widehat{m}_h(\cdot)$.

Clearly, the solution is to not use Y_j when forming $\widehat{m}_h(X_j)$. Thus, let $\widehat{m}_{h,j}(\cdot)$ denote the kernel estimator of $m(\cdot)$ based on the smoothing parameter h, using all the observations **except** (X_j, Y_j):

$$\widehat{m}_{h,j}(x) = \frac{\sum_{i \neq j} K(\frac{x - X_i}{h}) Y_i}{\sum_{i \neq j} K(\frac{x - X_i}{h})}.$$

Then

$$\lim_{h \to 0} \widehat{m}_{h,j}(X_j) \neq Y_j.$$

The functions $\widehat{m}_{h,j}(\cdot)$, $j = 1, 2, \ldots, n$, may be used to form the *cross-validation criterion*, given by

$$\hat{\rho}_{cv}(h) = \frac{1}{n} \sum_{j=1}^{n} (Y_j - \widehat{m}_{h,j}(X_j))^2, \quad h > 0;$$

it may be shown that

$$\mathrm{E}\left(\hat{\rho}_{cv}(h) | X_1, X_2, \ldots, X_n\right) = \frac{1}{n} \sum_{j=1}^{n} \sigma^2(X_j) + \rho(h).$$

Because the first term in this expression does not depend on h, for the purpose of choosing the smoothing parameter, $\hat{\rho}_{cv}(h)$ is an unbiased estimator of $\rho(h)$.

It is worth noting that, with a little algebra, we may write

$$\hat{\rho}_{cv}(h) = \frac{1}{n} \sum_{j=1}^{n} \alpha_j(h)^2 (Y_j - \widehat{m}_h(X_j))^2$$

where

$$\alpha_j(h) = \frac{1}{1 - K(0)/(nh\widehat{p}(X_j))}.$$

There are two interesting implications of this result. One is that, when calculating $\hat{\rho}_{cv}(h)$, it is not necessary to actually compute n kernel estimators, leaving out one observation each time. The other is that the cross-validation criterion may be viewed as a type of weighted sum of the squared prediction errors.

Also, using the expression for kernel density estimators,

$$(nh)\widehat{p}(X_j) = \sum_{i=1}^{n} K(\frac{X_j - X_i}{h})$$

so that as $h \to 0$, $nh\widehat{p}(X_j) \to K(0)$ and, hence,

$$\alpha_j(X_j) \to \infty.$$

This fact keeps $\hat{\rho}_{cv}(h)$ from approaching 0 as $h \to 0$.

The cross-validation choice of h, h_{cv}, is then taken to be that value of h that minimizes the function $\hat{\rho}_{cv}(\cdot)$.

Example 7.3 Consider the fossil data analyzed in Example 7.1. To calculate the kernel estimate based on $h = h_{cv}$ we may use the command

```
> sm.regression(age, sratio, poly.index=0, method="cv", ngrid=1000)
```

The expression `method="cv"` specifies that the smoothing parameter h is to be chosen using cross-validation; hence, when it is included, an expression such as `h=1` specifying the value of h is not needed.

A plot of the kernel regression estimate calculated by the above command is given in Figure 7.3. The value of h_{cv} is given as the component `$h` of the output of `sm.regression`:

```
> out<-sm.regression(age, sratio, poly.index=0, method="cv", ngrid=1000)
> out$h
[1] 0.569
```

Hence, for these data, $h_{cv} = 0.569$.

A useful feature of `sm.regression` is the argument `hmult`. Specifying `hmult = 2` (for example) in `sm.regression` specifies the estimate uses $2h_{cv}$ as the smoothing parameter:

```
> out<-sm.regression(age, sratio, poly.index=0,
+      method="cv", ngrid=1000, hmult=2)
> out$h
[1] 1.138
```

Thus, it is easy to modify an estimate based on the cross-validation choice of h to use a little more or little less smoothing. □

7.5 A closer look at the bias of a kernel estimator

Although the approximation to the bias of $\widehat{m}(x)$ given in Section 7.3 is useful for understanding general properties of a kernel estimator, such approximations are not needed to calculate numerically the bias of $\widehat{m}(x)$. As we have seen, for a given choice for the true underlying function $m(\cdot)$, a simple expression for the exact conditional bias is readily available:

$$\mathrm{E}\left(\widehat{m}(x)|X_1, X_2, \ldots, X_n\right) - m(x) = \frac{\frac{1}{nh}\sum_{j=1}^{n} K(\frac{x-X_j}{h})m(X_j)}{\widehat{p}(x)} - m(x). \tag{7.11}$$

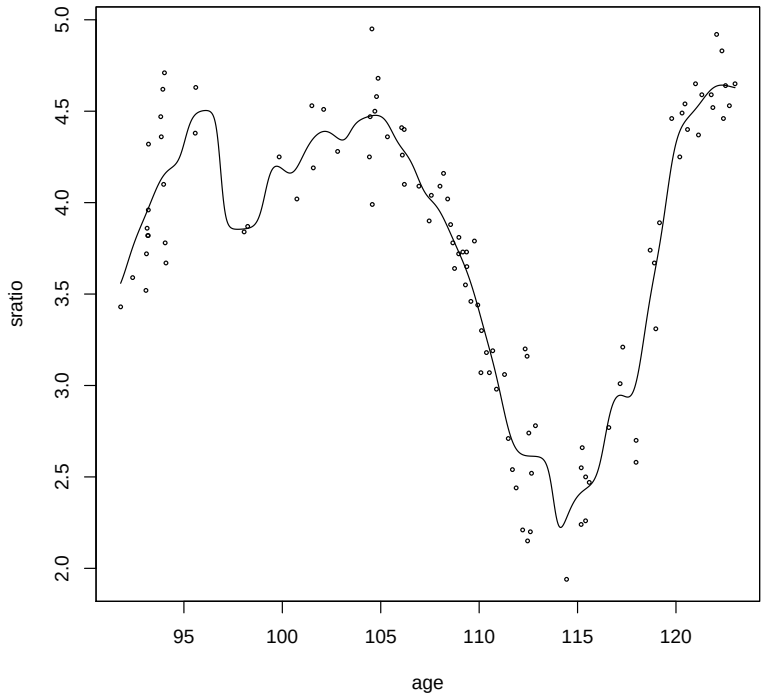

Figure 7.3: Kernel Estimate of the Regression Function for the Fossil Data using Cross-Validation

That is, given $m(\cdot)$ and the X-values $X_1, X_2, \ldots, X_n$, we compute the bias of $\widehat{m}(x)$ by calculating the kernel estimator using $m(X_1), m(X_2), \ldots, m(X_n)$ in place of $Y_1, Y_2, \ldots, Y_n$.

For instance, if the variable `x` contains the vector $(X_1, \ldots, X_n)$, then estimates of the bias of $\widehat{m}(\cdot)$ when $m(x) = x^2$ and $h = 1$ can be calculated using

```
> out<-sm.regression(x, x^2, poly.index=0, method="cv", ngrid=1000)
> bias<-out$estimate - out$eval.points^2
```

Example 7.4 Consider a sample of size 100 from the model

$$Y = m(X) + \epsilon$$

where ϵ is normally distributed with $\mathrm{E}(\epsilon|X) = 0$ and standard deviation σ.

Assume that X takes values in the interval $[0, 1]$ and consider four choices for $m(\cdot)$:

$$m(x) = x^2, \quad m(x) = 10(x - 0.5)^4, \quad m(x) = 1 - \exp(-3x), \quad \text{or} \quad m(x) = \sin(2\pi(x + 1/8)).$$

Figure 7.4 contains plots of these four functions.

Here we consider the exact conditional bias of the kernel estimator when estimating the function $m(\cdot)$, using the expression (7.11), when $m(\cdot)$ is equal to one of these four functions.

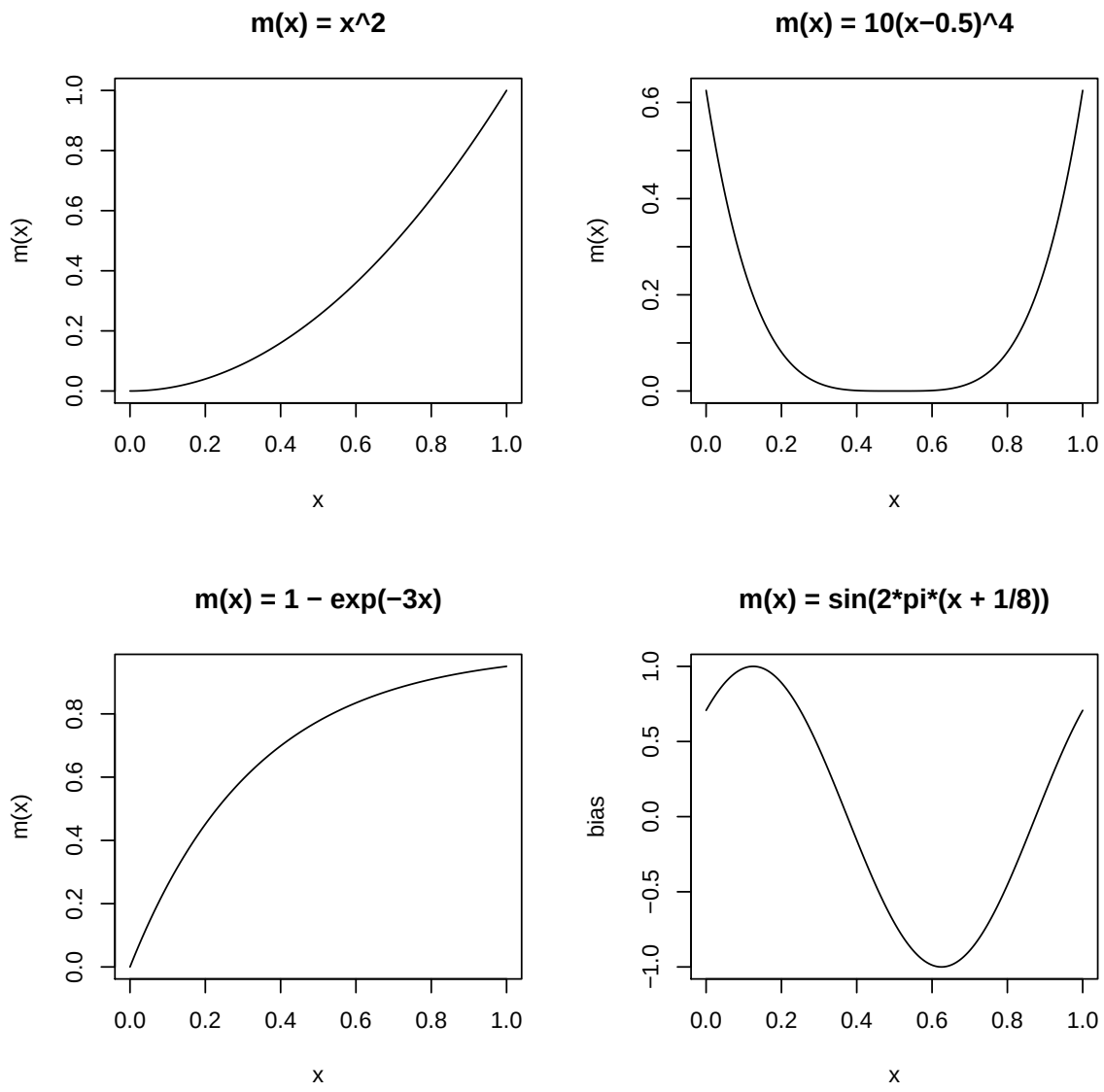

Figure 7.4: Four Functions Analyzed in Example 7.4

In each case, the values of $X_1, X_2, \ldots, X_{100}$ where chosen as a random sample from the distribution with density $p_X(x) = (4/3)x^{\frac{1}{3}}$, $0 < x < 1$. The value of the error standard deviation σ was chosen so that, for each choice of $m(\cdot)$, the proportion of the variation in Y "not explained by X" is approximately 0.10 and the value of h used was chosen by averaging h_{cv} for a large number of randomly-generated datasets; hence, the value of h used may be viewed as an estimate of the value that minimizes the MSE.

Figure 7.5 contains the bias of the kernel estimate $\widehat{m}(\cdot)$ as a function of x; the dots along the x-axis denote the X-values used in the estimation procedure.

We may make the following observations regarding the results in Figure 7.5.

(1) For $m(x) = x^2$, the bias tends to increase as x approaches 1; this may be due to the

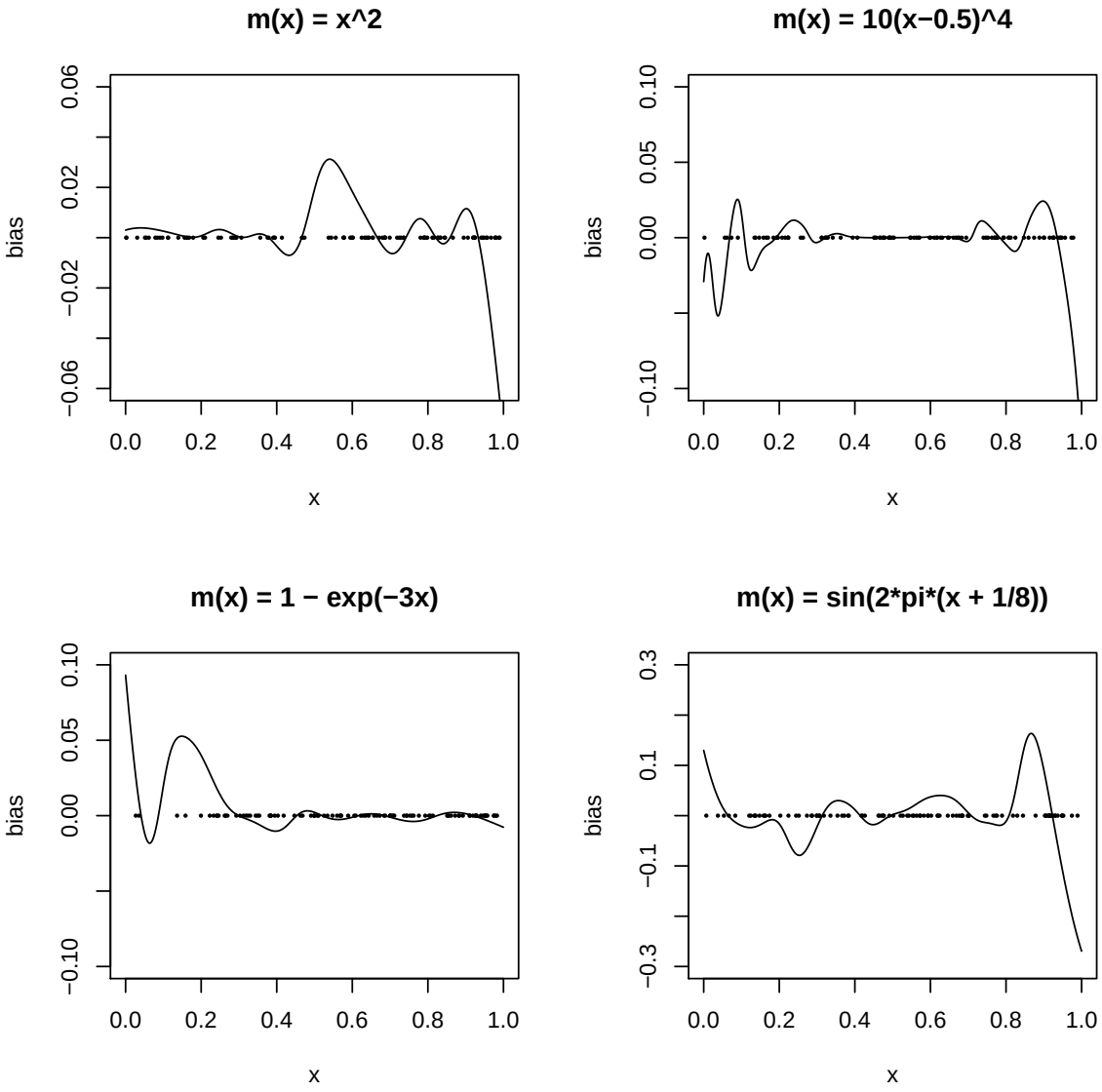

Figure 7.5: Bias of the Kernel Estimator When $m(\cdot)$ is Equal to One of the Four Functions Analyzed in Example 7.4

fact $m'(x)$ is small in magnitude for x near 0 and increases as x increases. The bias is relatively large in the region where there are few observations and also near the boundary $x = 1$; the bias is not large near the boundary $x = 0$ but there $m'(x) \doteq 0$ so that the function is approximately constant.

(2) For $m(x) = 10(x-0.5)^4$, the bias is large near both boundaries $x = 0$ and $x = 1$, where $m'(x)$ and $m''(x)$ are relatively large; however, the bias is essentially 0 for x near 0.5, a region over which the function is approximately constant.

(3) For $m(x) = 1 - \exp(-3x)$ the bias is large for $x < 0.2$, particularly where there are

few observed X-values; the bias is very large near the boundary $x = 0$. On the other hand, the bias is very small for $x > 0.5$. Note that $m'(x) = 3\exp(-3x)$ and $m''(x) = -9\exp(-3x)$ so that both of these are relatively small for even moderate values of x and are fairly small for x near 1; for instance, $m'(0.5) \doteq 0.67$, $m''(0.5) \doteq -2.0$, $m'(1) \doteq 0.15$ and $m''(1) \doteq -0.45$ while $m'(0) = 3$ and $m''(0) = -9$.

(4) For $m(x) = \sin(2\pi(x + 1/8))$, both $m'(x)$ and $m''(x)$ vary considerably over the region $0 < x < 1$ and, similarly, the bias varies considerably over this region. The bias is particularly large near the boundaries $x = 0$ and $x = 1$ and in regions where there are relatively few X-values.

□

These examples show that the bias of the kernel estimator tends to be large when

- there are few X_j near x
- $m'(x)$ and $m''(x)$ are large
- x is near the boundary of the range of X.

The first two of these situations are not surprising given the form of the average conditional bias of $\widehat{m}(x)$:

$$\frac{1}{2}\left(m''(x) + 2\frac{p'_X(x)}{p_X(x)}m'(x)\right)h^2 + O(h^4).$$

Thus, for those values of x for which $p_X(x)$ is small, or $m'(x)$ and $m''(x)$ are large, we expect the bias to be large, generally speaking.

The third situation listed, regarding the bias of the kernel estimator near the boundary of the range of X, also holds in general. This behavior is a consequence of the local averaging on which the kernel estimator is based.

Note that, for x values near a given point x_0, the function $m(\cdot)$ is approximately linear:

$$m(x) \doteq m(x_0) + m'(x_0)(x - x_0).$$

The expected values of the Y_j corresponding to $X_j \doteq x_0$ are close to, but different than $m(x_0)$.

However, when x_0 is an interior point of the range of X, those Y_j with expected values greater than $m(x_0)$ are balanced by Y_j with expected values less than $m(x_0)$, leading to an estimator that is approximately unbiased. However, when x_0 is near the boundary of the range of X, there are more X_j on one side of x_0 than on the other. Because the value of $m(x)$ is greater than $m(x_0)$ for x on one side of x_0 and smaller than $m(x_0)$ for x on the other

side of x_0, this lack of balance leads to an increase in the bias of $\widehat{m}(x_0)$. This is why the magnitude of $m'(x_0)$ plays an important role in the bias of $\widehat{m}(x_0)$.

It is natural to ask why this boundary behavior does not appear in the expression for the average conditional bias,

$$\frac{1}{2}\left(m''(x) + 2\frac{p'_X(x)}{p_X(x)}m'(x)\right)h^2 + O(h^4),$$

which holds for any value of x.

The reason is that this derivation applies to any value of x in the range of X, which was implictly assumed to be an open interval, that is, an interval of the form (a, b), where $-\infty \leq a < b \leq \infty$.

However, suppose that the range of X is a bounded interval such as $[0, 1]$. We will say that x is "near the boundary" 0 when the kernel estimator uses if it is within h of the endpoint of the range of X. For instance, if the range of X is the interval $[0, 1]$, we will consider x to be near the boundary if either $x \leq h$ or $x \geq 1 - h$. Note that, because $h \to 0$ as $n \to \infty$, the property of being "near the boundary" depends on the sample size.

The fact that "nearness to the boundary" depends on the sample size complicates the approximation of the bias of $\widehat{m}(\cdot)$, because those approximations are derived so that they are valid as $n \to \infty$. Thus, to approximate the bias of $\widehat{m}(x)$ for x near the boundary, we assume that x is approaching the boundary as $n \to \infty$. Specifically, if the range of X is $[0, 1]$, we assume that either x is of the form $x = ch$ or $x = 1 - ch$ where $0 < c < 1$; it is easy to adapt this condition to the case in which the range of X is any given interval.

For such values of x, it may be shown that the bias of the kernel estimator is of order $O(h)$ as $h \to 0$. Recall that for an arbitrary value x, the bias is of order $O(h^2)$. Because h is small a term that is $O(h^2)$ tends to be smaller than a term that is $O(h)$, at least when n is large (so that h is "small"). That is, the bias of the kernel estimator is, in general, larger near the boundary of the range of X.

The remainder of this section sketches the argument for this result. However, that material, which is rather technical, is optional and you can skip to the start of the following section if desired.

Bias of a kernel estimator near the boundary

Recall the basic form of the argument used to evaluate the average conditional bias of $\widehat{m}(x)$:

$$\mathrm{E}\left(\widehat{m}(x)|X_1, X_2, \ldots, X_n\right) = \frac{1}{\widehat{p}(x)}\frac{1}{nh}\sum_{j=1}^{n} K(\frac{x - X_j}{h})m(X_j)$$

and

$$\mathrm{E}\left(\frac{1}{nh}\sum_{j=1}^{n}K(\frac{x-X_j}{h})m(X_j)\right)=\int_{-\infty}^{\infty}\frac{1}{h}K(\frac{x-t}{h})m(t)p_X(t)dt.$$

Using the change-of-variable $u=(x-t)/h$,

$$\int_{-\infty}^{\infty}\frac{1}{h}K(\frac{x-t}{h})m(t)p_X(t)dt=\int_{-\infty}^{\infty}K(u)m(x-uh)p_X(x-uh)du$$

and, using a Taylor's series expansion around $uh=0$, along with the facts that

$$\int_{-\infty}^{\infty}K(u)du=1 \quad \text{and} \quad \int_{-\infty}^{\infty}uK(u)du=0,$$

it follows that

$$\begin{aligned}&\int_{-\infty}^{\infty}K(u)m(x-uh)p_X(x-uh)du\\&\quad=m(x)p_X(x)\int_{-\infty}^{\infty}K(u)du+\frac{d}{dx}(m(x)p_X(x))\int_{-\infty}^{\infty}uK(u)du\;h+O(h^2)\\&\quad=m(x)p_X(x)+O(h^2)\end{aligned}\tag{7.12}$$

leading to the result that the bias of $\widehat{m}(x)$ is of order $O(h^2)$.

A key part of this argument is that the range of the variable u in the integrals

$$\int_{-\infty}^{\infty}K(u)du \quad \text{and} \quad \int_{-\infty}^{\infty}uK(u)du$$

is $(-\infty,\infty)$, so that the integral of $K(u)$ over the range of u is 1 and the integral of $uK(u)$ over the range of u is 0.

Now suppose that the range of X is a bounded interval; for concreteness, take the range to be $[0,1]$. Then the range of the variable $u=(x-t)/h$ is $[(x-1)/h, x/h]$ (recall that x is fixed and t is the variable of integration) so that the integral

$$\int_{-\infty}^{\infty}K(u)m(x-uh)p_X(x-uh)du$$

becomes

$$\int_{(x-1)/h}^{x/h}K(u)m(x-uh)p_X(x-uh)du.$$

The Taylor's series expansion is still valid, but the lead term in the expansion (7.12) is now

$$m(x)p_X(x)\int_{(x-1)/h}^{x/h}K(u)du$$

instead of

$$m(x)p_X(x)\int_{-\infty}^{\infty}K(u)du.$$

For a fixed value of x, $0 < x < 1$,

$$\frac{x-1}{h} \to -\infty \quad \text{and} \quad \frac{x}{h} \to \infty$$

as $h \to 0$ and, hence,

$$\int_{(x-1)/h}^{x/h} K(u)du \to 1$$

so that the result that the bias of $\widehat{m}(x)$ is $O(h^2)$ continues to hold.

However, suppose that $x = ch$ for some $c > 0$. Then, as $h \to 0$,

$$\frac{x-1}{h} = c - \frac{1}{h} \to -\infty$$

but

$$\frac{x}{h} = c.$$

Thus,

$$\int_{(x-1)/h}^{x/h} K(u)du \to \int_{-\infty}^{c} K(u)du \neq 1.$$

Similarly, the integral in the second term in the expansion (7.12)

$$\int_{-\infty}^{\infty} uK(u)du = 0$$

becomes

$$\int_{(x-1)/h}^{x/h} uK(u)du \to \int_{-\infty}^{c} uK(u)du \neq 0.$$

It follows that

$$\int_{-\infty}^{\infty} K(u)m(x-uh)p_X(x-uh)du$$

can be expanded as

$$m(x)p_X(x)\int_{-\infty}^{c} K(u)du + \left(\frac{d}{dx}m(x)p_X(x)\right)\int_{-\infty}^{x} uK(u)du\,h + O(h^2).$$

Recall that the kernel density estimator $\widehat{p}(x)$ is in the denominator of the $\widehat{m}(x)$. The same issue affects $\widehat{p}(\cdot)$ so that

$$\widehat{p}(x) = p_X(x)\int_{-\infty}^{c} K(u)du - p'_X(x)\int_{-\infty}^{x} uK(u)du\,h + O(h^2).$$

When dividing these expansions, the term

$$\int_{-\infty}^{c} K(u)du$$

in the numerator and denominator cancels, leaving an expansion of the form

$$m(x) + O(h).$$

That is, the bias of $\widehat{m}(x)$ for $x = ch$ is of order $O(h)$, instead of the usual $O(h^2)$ that applies to the "interior" points x. For instance, for $h = O(n^{-\frac{1}{5}})$ (corresponding to the optimal choice for h), the bias near the boundary is $O(n^{-\frac{1}{5}})$ instead of the usual $O(n^{-\frac{2}{5}})$. Note that the same result holds if x is near the upper boundary of the interval $[0, 1]$, i.e., $x = 1 - ch$ for some $c > 0$.

Thus, the bias of the kernel estimator is, in general, greater near the boundary than it is in the interior.

7.6 Local linear regression

Many of the properties of the kernel estimator $\widehat{m}(\cdot)$, such as those described in the previous section, follow from the fact that the kernel estimator may be described as a "local constant" estimator. Such a description is based on the following considerations.

Suppose we find the value of a that minimizes

$$\sum_{j=1}^{n}(Y_j - a)^2;$$

it is straightforward to show that this minimizing value is $\bar{Y}$, the sample mean of $Y_1, Y_2, \ldots, Y_n$.

In nonparametric regression, we use a "local" version of this approach when estimating $m(x)$. That is, consider the value of a that minimizes

$$\sum_{j=1}^{n}(Y_j - a)^2 K(\frac{x - X_j}{h}); \tag{7.13}$$

note that, in this expression, the jth term receives more weight when X_j is near x and it receives less weight when X_j is far from x.

Differentiating (7.13) with respect to a, setting the result equal to 0 and solving for a yields the result

$$a = \frac{\sum_{j=1}^{n} K(\frac{x-X_j}{h})Y_j}{\sum_{j=1}^{n} K(\frac{x-X_j}{h})},$$

which is the kernel estimator of $m(x)$. Because the "local model" being fit is just a constant a, the kernel estimator we have been using is sometimes called a *local constant estimator*.

Using the same general idea, we may construct a "local linear estimator" based on a "local" regression model. This local regression model is based on the fact that, because

$$\mathrm{E}(Y_j|X_j) = m(X_j),$$

for X_j near x,

$$\mathrm{E}(Y_j|X_j) \doteq m(x) + m'(x)(X_j - x). \tag{7.14}$$

Therefore, to estimate $m(x)$, we might consider fitting a regression model

$$Y_j = a_x + b_x(X_j - x) + \epsilon_j \tag{7.15}$$

that applies for (X_j, Y_j) for which $X_j \doteq x$. Comparing the equation for the regression model (7.15) to the equation (7.14), we see that a_x corresponds to $m(x)$ and b_x corresponds to $m'(x)$; hence, an estimator of a_x yields an estimator of $m(x)$.

To estimate a_x and b_x we can use weighted least squares, giving more weight to observations (X_j, Y_j) with $X_j \doteq x$. That is, we estimate a_x, b_x by minimizing

$$\sum_{j=1}^{n}(Y_j - a - b(X_j - x))^2 K(\frac{x - X_j}{h}). \tag{7.16}$$

It is straightforward to show that the local linear kernel estimator of $m(\cdot)$, which we denote by $\widehat{m}_L(\cdot)$, has the form

$$\widehat{m}_L(x) = \widehat{a}_x = \frac{\sum_{j=1}^n K(\frac{x-X_j}{h})Y_j}{\sum_{j=1}^n K(\frac{x-X_j}{h})} - \widehat{b}_x \left(\frac{\sum_{j=1}^n K(\frac{x-X_j}{h})X_j}{\sum_{j=1}^n K(\frac{x-X_j}{h})} - x \right).$$

Note that

$$\frac{\sum_{j=1}^n K(\frac{x-X_j}{h})Y_j}{\sum_{j=1}^n K(\frac{x-X_j}{h})}$$

is the "local constant" kernel estimator $\widehat{m}(x)$ and, as discussed previously, $\widehat{b}_x$ is an estimator of $m'(x)$.

The quantity

$$\widehat{g}(x) = \frac{\sum_{j=1}^n K(\frac{x-X_j}{h})X_j}{\sum_{j=1}^n K(\frac{x-X_j}{h})} \tag{7.17}$$

is essentially a local constant kernel estimator but with X_j in place of Y_j; hence, it may be viewed as an estimator of $\mathrm{E}(X|X = x)$ which, of course, is just x.

Using properties of the local constant kernel estimator, we know that

$$\widehat{m}(x) \approx m(x) + \frac{1}{2}\left(m''(x) + 2\frac{p'_X(x)}{p_X(x)}m'(x) \right) h^2,$$

where the second term in this expression is the leading term in the expansion of the bias of $\widehat{m}(x)$. A similar result holds for $\widehat{g}$, which may be viewed as an estimator of $g(x) = x$; because $g'(x) = 1$ and $g''(x) = 0$,

$$\widehat{g}(x) \approx x + \frac{p'_X(x)}{p_X(x)}h^2.$$

It follows that

$$\begin{aligned}\widehat{m}_L(x) &= \widehat{m}(x) - b_x\left(\widehat{g}(x) - x\right) \\ &\approx m(x) + \frac{1}{2}\left(m''(x) + 2\frac{p'_X(x)}{p_X(x)}m'(x)\right)h^2 - b_x\frac{p'_X(x)}{p_X(x)}h^2.\end{aligned} \tag{7.18}$$

Finally, recall that b_x is an estimator of $m'(x)$ and, because in (7.18) b_x is multiplied by a term of order h^2, we only need to use the fact that $b_x \approx m'(x)$. It follows that

$$\begin{aligned}\widehat{m}_L(x) &\approx m(x) + \frac{1}{2}\left(m''(x) + 2\frac{p'_X(x)}{p_X(x)}m'(x)\right)h^2 - b_x\frac{p'_X(x)}{p_X(x)}h^2 \\ &\approx m(x) + \frac{1}{2}\left(m''(x) + 2\frac{p'_X(x)}{p_X(x)}m'(x)\right)h^2 - m'(x)\frac{p'_X(x)}{p_X(x)}h^2 \\ &= m(x) + \frac{1}{2}m''(x)h^2.\end{aligned}$$

Hence, the local linear estimator may be viewed as version of the local constant estimator that includes an adjustment for $m'(x)$. That is, the local linear estimator effectively adjusts for the locally linear behavior of $m(\cdot)$.

Example 7.5 Consider the data on the relationship between the ratio of strontium isotopes in a fossil and the fossil's age, analyzed in Example 7.1.

A local linear kernel estimate may calculated using the function `sm.regression`. The argument `poly.index` specifies the order of the "local polynomial" for the estimate, with 0 specifying a local constant estimate (i.e., the "standard" kernel estimator) and 1 specifying a local linear estimate. However, because 1 is the default, for a local linear estimate the argument `poly.index` can be omitted.

Thus, the following command produces the plot of the local linear kernel estimate given in Figure 7.6.

```
> sm.regression(age, sratio, method="cv", ngrid=1000)
```

Note that the local linear estimate is generally similar to the local constant estimate, displayed in Figure 7.3, although there are some differences. Figure 7.7 gives a plot of the local linear estimate with the local constant estimate included as a dashed curve; the smoothing parameter was chosen by cross-validation in both cases. □

The local linear estimator has some important advantages over the kernel estimator, primarily in terms of bias. Note that, as we did when analyzing $\widehat{m}(x)$, we can compute the exact bias of $\widehat{m}_L(x)$ by by calculating the local linear estimator using $m(X_1), m(X_2), \ldots, m(X_n)$ in place of $Y_1, Y_2, \ldots, Y_n$.

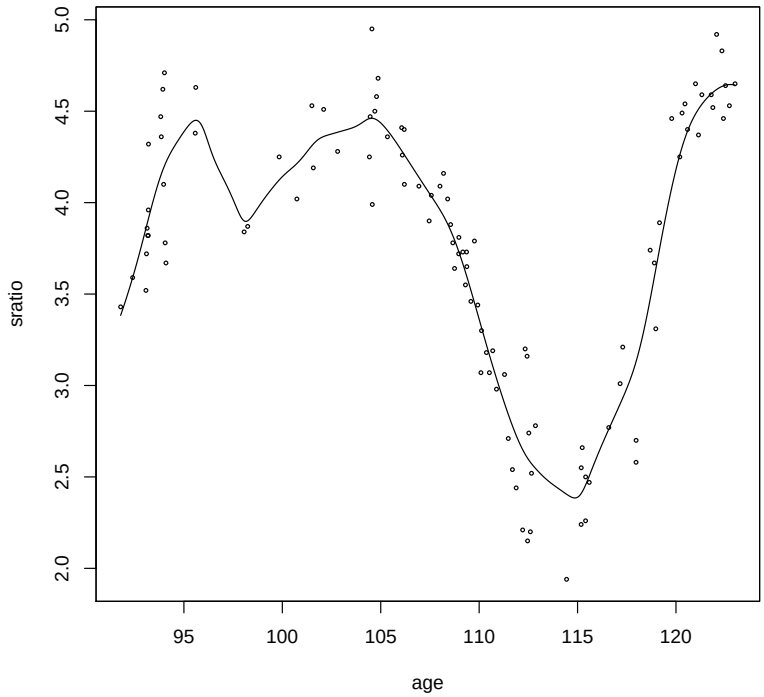

Figure 7.6: Local Linear Kernel Estimate for the Fossil Data using Cross-Validation

Example 7.6 Here we redo the analysis described in Example 7.4, using the local linear kernel estimator in place of the local constant kernel estimator. Plots of the four functions considered for the true regression function $m(\cdot)$ are given in Figure 7.4. The exact conditional bias of the local linear kernel estimator based on a sample of size $n = 100$ was calculated for each choice of $m(\cdot)$, using the same procedure used in Example 7.4.

The results are given in Figure 7.8; for comparison, each plot also contains a plot of the exact conditional bias of the local constant kernel estimate, given as a dotted curve. □

Note that the local linear estimator appears to improve on many of the poor properties of the kernel estimator. In particular, the bias is less dependent on the density of the X values and the bias does not increase appreciably near the boundary of the range of X. However, in the interior of the range of X, often the local constant estimator has smaller bias than does the local linear estimator.

The following results show that all of the these properties hold in general, at least to some extent.

Consider x in the interior of the range of X. As suggested by the analysis given previously in this section, the bias of $\widehat{m}_L(x)$ has an expansion of the form

$$\mathrm{E}\left(\widehat{m}_L(x)\right) - m(x) = \frac{1}{2}m''(x)h^2 + O(h^4);$$

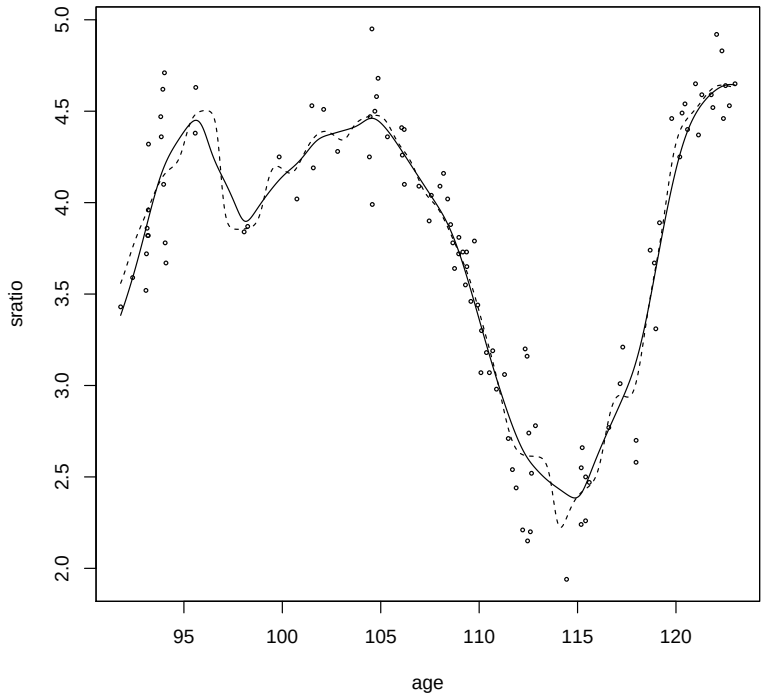

Figure 7.7: Local Linear and Local Constant Kernel Estimates for the Fossil Data

for comparison, recall that for the local constant estimator

$$\mathrm{E}\left(\widehat{m}(x)\right) - m(x) = \frac{1}{2}\left(m''(x) + 2\frac{p'_X(x)}{p_X(x)}m'(x)\right)h^2 + O(h^4).$$

Because $p_X(x)$ only appeared in conjuction with $m'(x)$, adjusting for $m'(x)$ also removed the dependence of the bias on the density $p_X(x)$, at least based on the approximation we are using. That is, a region in which $p_X(\cdot)$ is relatively small, so that there are relatively few X values in that region, does not inflate the bias in that region; note that this is not true for the kernel estimator, the bias of which depends on $p_X(\cdot)$ through the term $p'_X(x)/p_X(x)$. Thus, $\widehat{m}_L(\cdot)$ is sometime described as being "design adaptive".

For x near the boundary, there exists a function $b(\cdot)$ such that

$$\mathrm{E}\left(\widehat{m}_L(x)\right) - m(x) = \frac{b(x)}{2}m''(x)h^2 + O(h^4);$$

for comparison, recall that for x near the boundary,

$$\mathrm{E}\left(\widehat{m}(x)\right) - m(x) = O(h).$$

Of course, it does not necessarily follow that a term of order $O(h^2)$ will be smaller than a term of order $O(h)$, although that will tend to be the case for small h, that is, for large n.

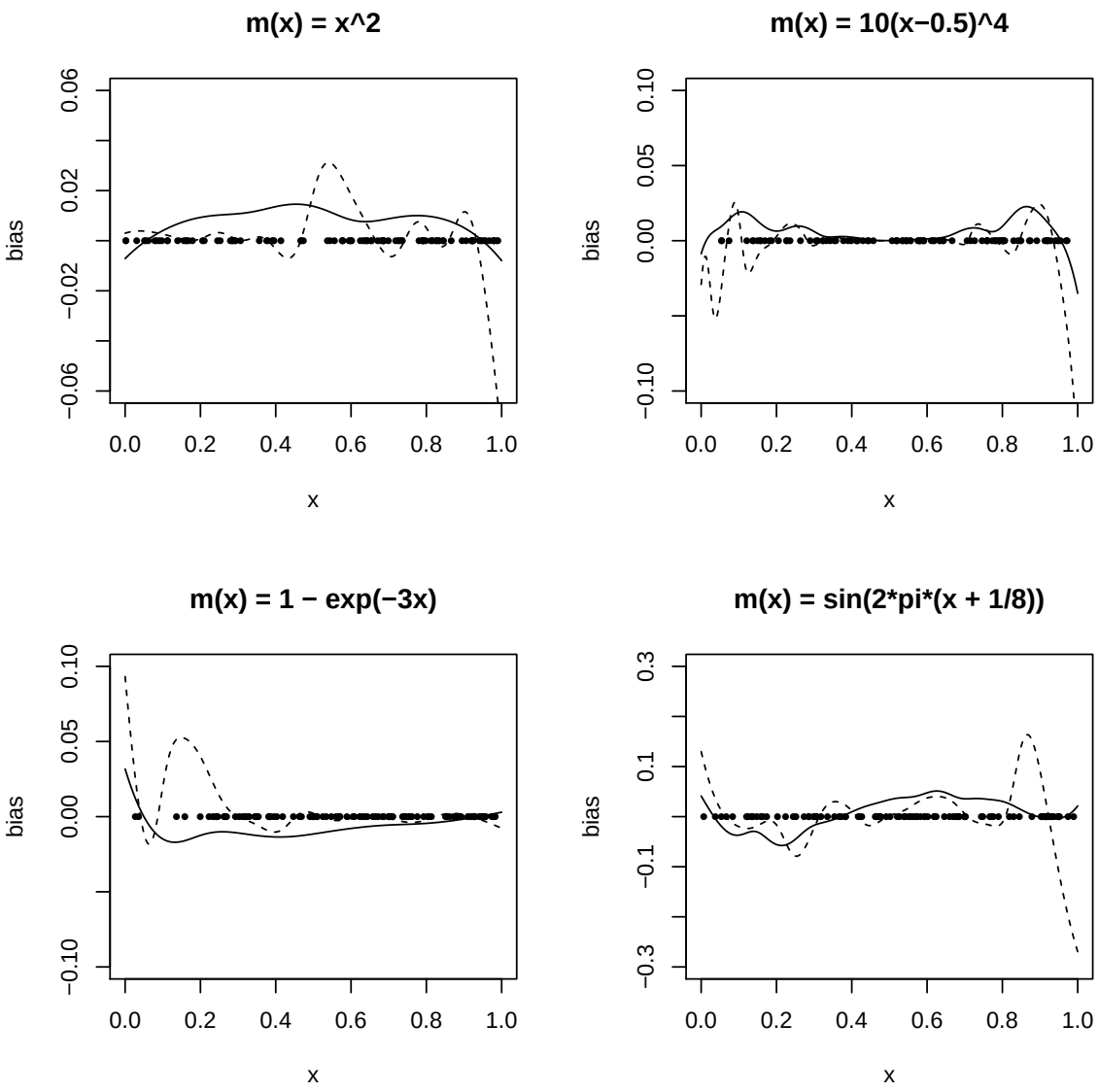

Figure 7.8: Bias of the Local Linear and Local Constant Kernel Estimators

That is, the bias of $\widehat{m}_L(\cdot)$ is order $O(h^2)$ for all x, even those value near the boundary of the range of X.

Often, when comparing two estimators, the one with the smaller bias has the larger variance; this is true here, but difference between the variance of the local linear estimator and the variance of the local constant estimator is relatively minor. In the interior of the range of X, the expansion of the conditional variance of $\widehat{m}_L(x)$ and the expansion of the conditional variance of $\widehat{m}(x)$ both have leading term

$$\frac{K_2\sigma^2(x)}{p_X(x)}\frac{1}{nh}$$

so that they are approximately equal. For values of x near the boundary, both estimators

have a conditional variance with an expansion with the leading term of the form $c(x)/(nh)$, for $c(x)$ not depending on n or h, although the constant is about three times larger for the local linear estimator than it is for the local constant estimator. This is a small increase in the variance, as compared to the bias where the *order* of the bias of the local linear estimator is $O(h^2)$ near the boundary while the bias of the local constant estimator is of order $O(h)$ near the boundary. Hence, the ratio of the bias of the local linear estimator to the bias of the local constant estimator approaches 0 as $n \to \infty$.

Thus, the local linear estimator $\widehat{m}_L(\cdot)$ is generally preferable to the standard kernel estimator and, hence, throughout the remainder of these notes, the only kernel estimator we will use is the local linear kernel estimator. Thus, for convenience, we will denote this estimator by $\widehat{m}(\cdot)$ and refer to it simply as the "kernel estimator".

7.7 Exercises

7.1. The data set "motor" contains data based on a simulated motorcycle accident used to test helmets. The variables are the time after impact ("time"), in milliseconds, and the head acceleration of a crash-test dummy ("accel").

(a) Find the local linear kernel regression estimator of $E(\text{accel} \mid \text{time})$, choosing the smoothing parameter h subjectively, balancing the goal of having a smooth function with the need to have the function accurately represent the relationship between the variables. Plot the estimated regression function together with the data.

(b) Repeat part (a) using cross-validation to choose the smoothing parameter. Plot the estimated regression function together with the raw data; compare the estimate to the one found in part (a).

7.2. The dataset "geyser" contains data on the time between eruptions ("waiting"), along with the length of the eruptions ("duration"), of the Old Faithful geyser at Yellowstone National Park, over the period from August 1 to August 15, 1985. Recall that we estimated the (bivariate) density of waiting and eruption in a Week 6 exericse.

Estimate the regression function relating duration (the response variable) to waiting (the predictor variable). Use a local linear kernel estimator and use cross-validation to choose the smoothing parameter. Plot the estimated regression function with the data and comment on how the duration of an eruption is related to the time since the last eruption.

7.3. The amount of scoring in baseball has changed over the years with changes in rules, athletic training, etc. Such changes complicate comparisons of players and teams from

different eras; hence, it is often useful to standardize certain results using some measure of scoring in a given year.

The dataset "runs" contains the average runs scored per team per game in each year from 1900 to 2016 (numbered 0 to 116).

(a) Calculate the local linear kernel estimate of the regression function relating runs to year. Choose the value of the smoothing parameter using cross-validation. Give a plot of the estimate together with the data.

(b) Give the value of the estimated regression function for the years 1920, 1940, 1960, 1980, 2000, and 2016.

(c) Repeat parts (a) and (b), taking the value of the smoothing parameter to be 1.5 times the cross-validation value (use `hmult`). How do the estimates found here for 1920, 1940, 1960, 1980, 2000, and 2016 compare to those found in part (b)?

7.4. In Section 7.3, we saw that the value of h that minimizes the MSE is of order $O(1/n^{\frac{1}{5}})$ as $n \to \infty$. The purpose of this exercise is to investigate whether the cross-validation choice of h if of this order.

One approach to this issue is to derive analytically the properties of h_{cv}; however, such an analysis is quite complicated. Hence, here we consider a more empirical approach. This method is useful for investigating the order of a term, without doing a lot of theoretical calculations.

(a) For $n = 100$, use the function `runif` to draw n random variates from a uniform distribution on the interval $(0, 1)$ and assign the values to `x`.

(b) Construct a vector `y` using

```
> y <- x^2 + rnorm(n, sd=1/2)
```

That is, the true regression function here is $m(x) = x^2$. Other functions could be used here, with similar results.

(c) Use the function `sm.regression` to obtain the cross-validation value of the smoothing parameter h corresponding to the values of `x` and `y`, using a local constant estimator. Thus, the value of h is given by

```
> sm.regression(x, y, method="cv",  poly.index=0)$h
```

(d) Repeat parts (b) and (c) four times (so you have five values) and average the results. This gives an estimate of h_{cv} for $n = 100$; call this estimate $\widehat{h}_{100}$.

(e) Repeat parts (a) - (d) for $n = 500, 1000, 5000$ and 10000 to obtain $\widehat{h}_{500}, \widehat{h}_{1000}, \widehat{h}_{5000}$ and $\widehat{h}_{10000}$.

(f) If $\widehat{h}_n$ is of order $O(1/n^\beta)$, for some $\beta > 0$, we expect that

$$\widehat{h}_n \doteq \frac{c}{n^\beta}$$

for some constant c_0 or, equivalently,

$$\log\left(\widehat{h}_n\right) \doteq \log(c) - \beta \log(n).$$

Hence, for $n = 100, 500, 1000, 5000, 10000$, plot $\log\left(\widehat{h}_n\right)$ versus $\log(n)$. Does the relationship between $\log\left(\widehat{h}_n\right)$ and $\log(n)$ appear to be approximately linear?

(g) Using least-squares regression (i.e., the function `lm`) estimate the parameter β in the model

$$\log\left(\widehat{h}_n\right) = \log(c) - \beta \log(n) + \epsilon.$$

If h_{cv} is of order $O(1/n^{\frac{1}{5}})$, then we expect β to be approximately 0.2. Are your results consistent with the hypothesis that $h_{cv} = O(1/n^{\frac{1}{5}})$? Why or why not?