Week 8

8.1 Degrees of freedom of a kernel estimator

Consider a parametric linear regression model relating a response variable Y to a set of predictor variables $X_1, X_2, \dots, X_p$:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

A simple summary of the "complexity" of the model is given by the degrees-of-freedom of the model, p, the number of predictors in the regression function.

Recall that, in such a parametric regression model, the sum of squared errors satisfies

$$\mathrm{E}\left(\sum_{j=1}^{n}(Y_j - \hat{Y}_j)^2\right) = n - (p+1)\sigma^2 \tag{8.1}$$

where $Y_1, Y_2, \dots, Y_n$ denote the observed values of Y and $\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n$ denote the "predicted values" from the regression:

$$\hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_{1j} + \cdots + \hat{\beta}_p X_{pj}, \quad j = 1, 2, \dots, n$$

where $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ are the least-squares estimators and $X_{1j}, X_{2j}, \dots, X_{pj}$ are predictor variables for the jth observation so that $Y_j - \hat{Y}_j$, $j = 1, 2, \dots, n$ are the residuals from the model.

We may use this idea to define a "degrees-of-freedom" corresponding to the local linear kernel estimator based on a specific choice of h.

Consider a nonparametric regression model for the pair of random variables (X, Y),

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

where ϵ is a random variable satisfying $\mathrm{E}(\epsilon|X) = 0$ and let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ denote independent random vectors each with the distribution (X, Y). Suppose further that $\mathrm{Var}(Y|X) = \sigma^2$, for some unknown constant $\sigma^2 > 0$, so that each Y_j has variance σ^2.

One way to define the degrees-of-freedom corresponding to a kernel estimator $\widehat{m}(\cdot)$ is to consider an estimator of σ^2 that is analogous to the estimator (8.1) used in parametric regression models. Let

$$Y_j - \widehat{m}(X_j), \quad j = 1, 2, \ldots, n$$

denote the residuals from the estimation procedure and consider the sum-of-squares

$$\sum_{j=1}^{n}(Y_j - \widehat{m}(X_j))^2.$$

It may be shown that the expected value of $\sum_{j=1}^{n}(Y_j - \widehat{m}(X_j))^2$ is approximately of the form $(n-D)\sigma^2$, where D does not depend on σ^2. By analogy with the estimator of σ^2 used in a parametric regression model, we may interpret D as a type of "degrees-of-freedom" for a local linear kernel estimator (or any type of kernel estimator), giving a simple measure of the complexity of the estimate $\widehat{m}(\cdot)$. Note that D depends on n, the values of $X_1, X_2, \ldots, X_n$, and the kernel $K(\cdot)$; however, the most important factor affecting D is the value of h.

In a sense, D is similar to h, the smoothing parameter, but with the added advantage of being easier to intepret because of the relationship to the concept of degrees-of-freedom in parametric regression models, which is well-understood. Note that, in a parametric regression model with p predictors, $D = p+1$, while the usual degrees-of-freedom for a regression model is taken to be p; hence, $D - 1$ corresponds to the "usual" definition of degrees-of-freedom. Also note that D need not be an integer.

In fact, in the function `sm.regression`, the degree of smoothing may be specified by choosing the desired degrees-of-freedom, rather than by specifying the desired value of h.

Example 8.1 Consider the fossil data analyzed in Example 7.1. To calculate the local linear kernel estimate corresponding to 5 degrees-of-freedom we may use the command

```
> sm.regression(age, sratio,  df=5,  ngrid=1000)
```

A plot of the kernel regression estimate calculated by the above command is given in Figure 8.1.

The value of h corresponding to a given value for degrees-of-freedom may be obtained from the component `$h` of the result of the `sm.regression` function; thus, using 5 degrees-of-freedom as above corresponds to a value of h of 3.316:

```
> sm.regression(age, sratio, df=5)$h
[1] 3.316
```

For comparison, $h_{cv} = 0.9303$:

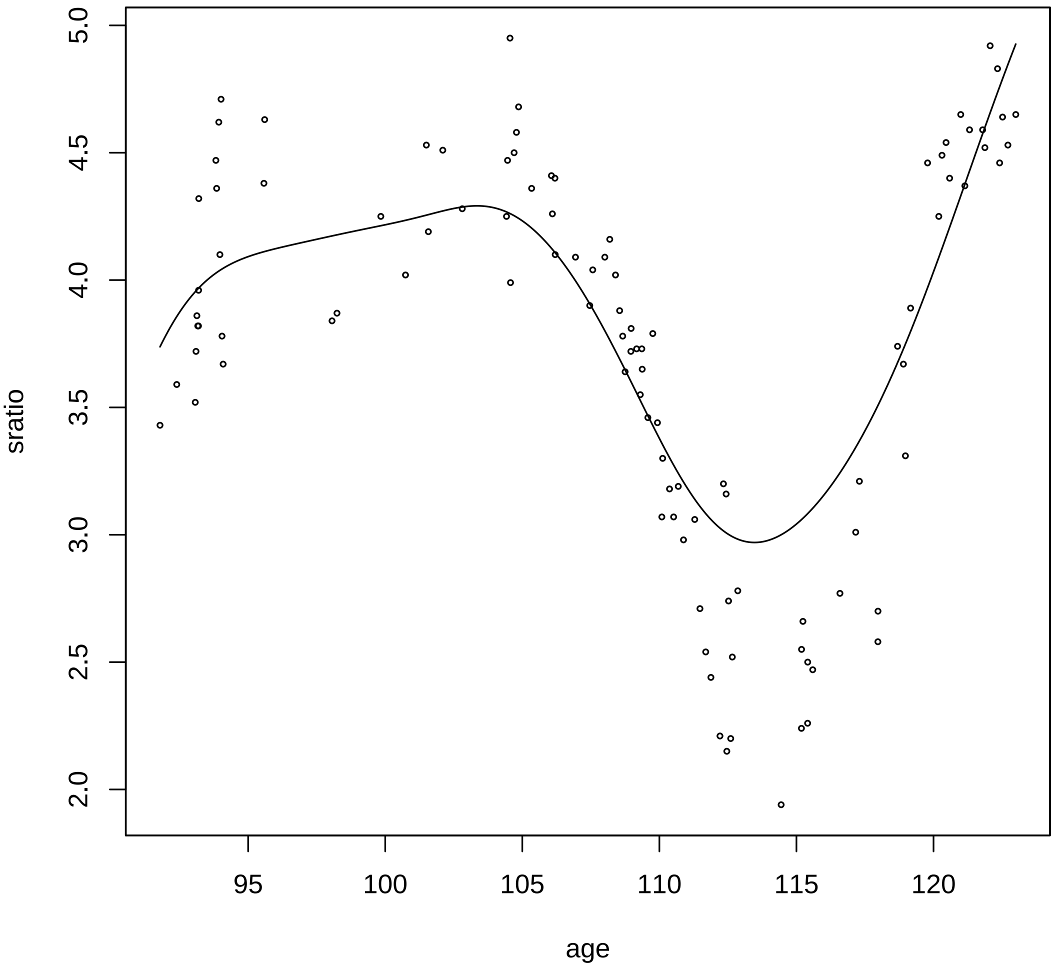

Figure 8.1: Local Linear Estimate of the Regression Function for the Fossil Data using 5 Degrees-of-Freedom

```
> sm.regression(age, sratio, method="cv")$h
[1] 0.9303
```

□

Although the function `sm.regression` may be used to find the value of h corresponding to a given value for the degrees-of-freedom, in some cases, we would like to find the degrees-of-freedom corresponding to a given value of h. For example, for the fossil data, it may be interesting to know the degrees-of-freedom corresponding to $h_{cv} = 0.9303$.

Of course, it is possible to use trial-and-error with a number of different values for the degrees-of-freedom to hone in on the value of h of interest. A more systematic approach is to use function `uniroot`, which finds the root of a given function, that is, it finds the argument of the function that returns the value 0. The following user-defined function `h2df` illustrates this idea.

```
> h2df<- function(h_target, x, y){
+  f<-function(df){
+  sm.regression(x, y, df=df, display="none")$h - h_target
+ }
+ hmin<-sm.regression(x, y, df=length(x), display="none")$h
+ hmax<-sm.regression(x, y, df=2.01, display="none")$h
+ if ((h_target<= hmax)&(h_target >= hmin)){
+  dfroot<-uniroot(f, c(2.01, length(x)), tol=0.005)$root
+  return(round(dfroot, 2))
+  }
+ else return("out of range")
+ }
```

The value of `df` used in the function `sm.regression` must be greater than 2 (the value corresponding to a linear regression model with a single predictor) and it is reasonable to only consider values of `df` less than or equal to n, the number of observations. Thus, only values of h between `hmin` and `hmax`, as defined in the function, are considered. The function `f` defined in `h2df` returns the value of h corresponding to a given value of `df`, minus the target value of h, `h_target`. The function `uniroot` then finds the root of `f`, the value of the value of `df` that yields the target value of h. The second argument of `uniroot` gives the range of values over which to search for the root and the third argument sets the tolerance of the procedure, here set to the relatively large value of `0.005` (because we do not need a very precise value for `df`). Then the result is rounded to 2 decimal places to reflect the tolerance used.

Example 8.2 For the fossil data, the degrees-of-freedom corresponding to $h_{cv} = 0.9303$ is 14.74:

```
> h2df(0.9303, age, sratio)
[1] 14.74
```

Larger values of h correspond to smoother estimates and, hence, fewer degrees-of-freedom (that is, less complexity of the estimated function) while smaller values of h correspond to greater degrees-of-freedom:

```
> h2df(1, age, sratio)
[1] 13.86
> h2df(2, age, sratio)
[1] 7.66
> h2df(3, age, sratio)
[1] 5.43
> h2df(0.75, age, sratio)
[1] 17.61
> h2df(0.5, age, sratio)
[1] "out of range"
```

Note that there is no value of `df` less than n (in this case 106) and greater than 2.01 corresponding $h = 0.5$. □

The definition of degrees-of-freedom presented in this section is only one of several definitions that are sometimes used in nonparametric regression. The other definitions are based on other analogies with parametric regression models; while all definitions give the same value for parametric models, when used in the nonparametric context, they generally give roughly similar, but different, values.

8.2 Estimation of the error variance

For the nonparametric regression model

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

where $\mathrm{E}(\epsilon|X) = 0$ and $\mathrm{Var}(\epsilon|X) = \sigma^2$; hence, we are assuming that the error variance is constant, in the sense that it does not depend on X. It is important to note that this is an important assumption that may or may not be appropriate for a given set of data. In this section, we consider estimation of σ^2.

Based on the result discussed in the previous section, one approach is to use residual-based estimators, such as

$$\frac{\sum_{j=1}^{n}(Y_j - \widehat{m}(X_j))^2}{n - D}, \tag{8.2}$$

where D is the degrees-of-freedom measure considered in the previous section.

However, such estimators are biased, with the level of bias depending on the value of the smoothing parameter h used in the estimator $\widehat{m}(\cdot)$. That is, even after adjusting for the degrees-of-freedom by including the factor D, this estimator is biased.

One reason for this bias is that the estimator $\widehat{m}(\cdot)$ is a biased estimator of $m(\cdot)$. It may be shown that the bias of the estimator (8.2) is of order $O(h^4)$, the order of the squared bias of $\widehat{m}(\cdot)$. For instance, if $h = O(n^{-\frac{1}{5}})$, the order of h that minimizes the MSE, the bias of the residual-based estimator (8.2) is of order $O(n^{-\frac{4}{5}})$.

For comparison, in a parametric regression model with p predictors, suppose we use the estimator of the error variance given by

$$\frac{\sum_{j=1}^{n}(Y_j - \hat{Y}_j)^2}{n} \tag{8.3}$$

where the Y_j are the response variables and the $\hat{Y}_j$ are the predictor values from the regression. That is, suppose we use the estimator based on the average of the squared residuals, but without any degrees of freedom adjustment. Then the expected value of this estimator is

$$\left(\frac{n-(p+1)}{n}\right)\sigma^2 = \left(1 - \frac{p+1}{n}\right)\sigma^2$$

so that the bias is

$$-\frac{p+1}{n}\sigma^2,$$

which is of order $O(1/n)$.

Thus, the order of the bias of the residual-based estimator in nonparametric regression, with a degrees-of-freedom adjustment, is larger than that of the residual-based estimator in parametric regression that does not include a degrees-of-freedom adjustment. That is, although the measure of degrees-of-freedom used in nonparametric regression gives useful summary of the properties of the estimator $\widehat{m}(\cdot)$, the degrees-of-freedom adjustment is not very effective in reducing bias when estimating σ^2.

An alternative approach is to base an estimator of σ^2 on Y-values corresponding to similar X-values. For instance, suppose that $X_j = X_k$. Then $(Y_j - Y_k)^2/2$ has expected value exactly equal to σ^2. In practice, we can't be sure that there will be sets of equal X-values; however, the same basic approach can be used on Y-values corresponding to X-values that are approximately equal.

Assume that the range of X is an interval of length L and that $X_1, X_2, \ldots, X_n$ are in increasing order, so that

$$X_1 \leq X_2 \leq \cdots \leq X_n$$

and suppose that it is reasonable to assume that there exists a constant $c > 0$ such that

$$X_j - X_{j-1} \leq c\frac{L}{n}, \quad j = 2, 3, \ldots, n. \tag{8.4}$$

Note that this condition is always satisfied for a given value of n (by choosing c large enough); hence, it is a statement about the hypothetical situation in which n increases. It requires that $X_1, X_2, \ldots, X_n$ are roughly evenly-spaced across the range of X. In particular, condition (8.4) would not be a reasonable assumption if there is an interval in the range of X such that values of X_j in that interval are, because of the nature of X, rare or impossible; this would be the case, for example, if the X_j take only integer values.

Note that, for $j = 2, 3, \ldots, n$,

$$\begin{aligned} \mathrm{E}\left((Y_j - Y_{j-1})^2 | X_1, X_2, \ldots, X_n\right) &= \mathrm{Var}(Y_j - Y_{j-1} | X_1, X_2, \ldots, X_n) + \mathrm{E}\left(Y_j - Y_{j-1} | X_1, X_2, \ldots, X_n\right)^2 \\ &= 2\sigma^2 + (m(X_j) - m(X_{j-1}))^2. \end{aligned}$$

Hence, the statistic

$$\frac{1}{2(n-1)} \sum_{j=2}^{n} (Y_j - Y_{j-1})^2 \tag{8.5}$$

has expected value

$$\sigma^2 + \frac{1}{2(n-1)} \sum_{j=2}^{n} (m(X_j) - m(X_{j-1}))^2.$$

Using a Taylor's series expansion of the form

$$m(X_j) - m(X_{j-1}) = m'(X_{j-1})(X_j - X_{j-1}) + \cdots ,$$

together with condition (8.4), under which $X_j - X_{j-1} = O(1/n)$, it follows that

$$\mathrm{E}\left(\frac{1}{2(n-1)} \sum_{j=2}^{n} (Y_j - Y_{j-1})^2 \,|\, X_1, X_2, \ldots, X_n \right) = \sigma^2 + O(\frac{1}{n^2}) \quad \text{as} \quad n \to \infty. \tag{8.6}$$

In a sense, in the estimator (8.5), $m(X_j)$ is estimated by Y_{j-1}. A better approach is to replace Y_{j-1} by a weighted average of Y_{j-1} and Y_{j+1}. The weights given to Y_{j-1} and Y_{j+1} depend on how close X_j is to X_{j-1} and X_{j+1}, using a procedure similar to that used in linear interpolation.

Define an estimator of σ^2 by

$$\hat{\sigma}_D^2 = \frac{1}{n-2} \sum_{j=2}^{n-1} \frac{(Y_j - b_j Y_{j-1} - c_j Y_{j+1})^2}{1 + b_j^2 + c_j^2},$$

where, for $j = 2, 3, \ldots, n-1$,

$$b_j = \frac{X_{j+1} - X_j}{X_{j+1} - X_{j-1}}$$

and

$$c_j = \frac{X_j - X_{j-1}}{X_{j+1} - X_{j-1}}.$$

Then, using a more complicated version of argument used to show (8.6), it may be shown that

$$\mathrm{E}\left(\hat{\sigma}_D^2 \,|\, X_1, X_2, \ldots, X_n\right) = \sigma^2 + O(\frac{1}{n^4}).$$

That is, $\hat{\sigma}_D^2$ is approximately unbiased as an estimator of σ^2; furthermore, the bias does not depend on the value of the smoothing parameter h. The estimator $\hat{\sigma}_D^2$ is sometimes described as the *second-differences* estimator of σ^2.

Example 8.3 Consider the fossil data analyzed in Example 8.1. To obtain an estimate of the error standard deviation σ using the second-differencing approach discussed in this section, we may use the command

```
> sm.sigma(age, sratio)$estimate
[1] 0.2357
```

This may be compared to the residual-based estimator based on using $h = h_{cv}$:

```
> yhat<-sm.regression(age, sratio, method="cv", eval.points=age)$estimate
> (sum((sratio-yhat)^2)/(106 - 14.74))^.5
[1] 0.2493
```

Here `yhat` contains the estimates $\widehat{m}(X_1), \ldots, \widehat{m}(X_n)$; the argument `eval.points = age` specifies that the function be estimated at the observed values of `age`. Recall that here $n = 106$ and the degrees-of-freedom corresponding to $h = h_{cv}$ is 14.74 (as calculated in Example 8.1. □

8.3 Confidence bands

When estimating the mean μ of a distribution based on a random sample $Y_1, Y_2, \ldots, Y_n$ we know that the sample mean $\bar{Y}$ is a point estimator of μ. Although the value of the point estimate gives useful information about the value of μ, it does not take into account the sampling variability of the point estimator. Hence, we generally consider a confidence interval for μ, such as the one given by

$$\bar{Y} \pm 1.96\frac{S}{\sqrt{n}}$$

where S^2 is the sample variance of $Y_1, Y_2, \ldots, Y_n$; this interval has an approximate coverage probability of 95%. The confidence interval has the property that we are "reasonably certain" that μ lies in the interval, in the sense that

$$\Pr(\bar{Y} - 1.96S/\sqrt{n} < \mu < \bar{Y} + 1.96S/\sqrt{n}) \doteq 0.95.$$

In this section, we consider a similar type of result for the unknown parameter $m(\cdot)$; however, the analysis is complicated by the fact that $m(\cdot)$ is a function. Thus, we consider a “pointwise approach”, in which we fix a value x in the range of X and then construct an approximate confidence interval for $m(x)$. Varying the value of x yields a type of “confidence band” for the function $m(\cdot)$.

For instance, we might construct an approximate 95% confidence interval $(L(x), U(x))$ for $m(x)$; then the functions $L(\cdot)$ and $U(\cdot)$ form 95% *pointwise confidence bands* for the function $m(\cdot)$. In interpreting such confidence bands, it is important to keep in mind that the 95% coverage probability refers to each $m(x)$ individually.

To compute pointwise confidence bands, we need to calculate an approximate confidence interval for $m(x)$ for each possible x. Recall that, in the case of a population mean μ, an approximate confidence interval for μ is based on the fact that

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}}$$

is approximately distributed according to a standard normal distribution. Here S denotes the sample variance of the observations and n is the sample size; “approximately distributed according to a standard normal distribution” refers to the convergence in distribution as n aproaches infinity. That is,

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}}$$

converges in distribution to a standard normal random variable Z, so that for any t, $-\infty < t < \infty$,

$$\mathrm{P}\left(\frac{\bar{Y} - \mu}{S/\sqrt{n}} \leq t\right) \to \mathrm{P}(Z \leq t) \quad \text{as } n \to \infty.$$

For a more general parameter θ, if an estimator $\hat{\theta}$ is such that

$$\hat{\theta} - \theta$$

is approximately normal with mean 0 and variance $\sigma_{\hat{\theta}}^2/n$, then an approximate confidence interval for θ may be based on the fact that

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}/\sqrt{n}}$$

is approximately distributed according to a standard normal distribution, where $\hat{\sigma}_{\hat{\theta}}$ is an estimator of $\sigma_{\hat{\theta}}$. For instance, this result leads to the approximate 95% confidence interval for θ given by

$$\hat{\theta} \pm 1.96 \frac{\hat{\sigma}_{\hat{\theta}}}{\sqrt{n}}.$$

We now consider the same approach to constructing an approximate confidence interval for $m(x)$, for a given value of x. Let $\hat{m}(\cdot)$ denote the local linear kernel estimator of $m(\cdot)$, based on choosing the smoothing parameter by cross-validation, and let $V(x)$ denote an estimator of the asymptotic variance of $\widehat{m}(x)$. Then, if

$$\frac{\hat{m}(x) - m(x)}{\sqrt{\hat{V}(x)}}$$

is approximately distributed according to a standard normal distribution, where $\hat{V}(x)$ is an estimator of $V(x)$, then we can construct an approximate confidence interval for $m(x)$ using the same basic approach used above for μ and θ.

Unfortunately, there is an important technical problem in implementing this approach. Recall the bias of $\hat{m}(x)$ is of order $O(n^{-\frac{2}{5}})$ and the variance of $\hat{m}(x)$ is of order $O(n^{-\frac{4}{5}})$. Therefore, we expect

$$\frac{\hat{m}(x) - m(x)}{\sqrt{\hat{V}(x)}} \doteq \frac{\hat{m}(x) - m(x)}{\sqrt{V(x)}}$$

to have expected value of order

$$\frac{O(n^{-\frac{2}{5}})}{\sqrt{O(n^{-\frac{4}{5}})}} = O(1).$$

That is, the expected value of

$$\frac{\hat{m}(x) - m(x)}{\sqrt{\hat{V}(x)}} \tag{8.7}$$

cannot be expected to converge to 0 as $n \to \infty$; it follows that its approximate distribution as $n \to \infty$ does not have mean 0. On the other hand, the standard normal distribution has mean 0. Hence, the approximate distribution of (8.7) cannot be standard normal.

Note that this issue does not arise in typical parametric approximate confidence intervals; in those cases, the bias of an estimator $\hat{\theta}$ of a parameter θ typically is of order $O(n^{-1})$ or smaller (for example, the bias of the maximum likelihood estimator of the variance σ^2 of a normal distribution is $-\sigma^2/n$) so that the expected value of

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}/\sqrt{n}}$$

is generally of order

$$\frac{O(n^{-1})}{O(n^{-\frac{1}{2}})} = O(n^{-\frac{1}{2}}) \quad \text{as} \quad n \to \infty$$

which approaches 0 as $n \to \infty$.

Hence, in order to construct an interval estimate for $m(x)$ we must deal with the fact that the mean of the distribution of

$$\frac{\hat{m}(x) - m(x)}{\sqrt{\hat{V}(x)}}$$

is not approximately 0.

There are three basic approaches that are used in these cases. Let

$$b(x) = \mathrm{E}\left(\hat{m}(x) - m(x)\right)$$

denote the bias of $\hat{m}(x)$ and let $\hat{b}(x)$ denote an estimator of $b(x)$. Then the expected value of

$$\hat{m}(x) - b(x) - m(x)$$

is 0 and the expected value of

$$\hat{m}(x) - \hat{b}(x) - m(x)$$

is generally of smaller order than $\sqrt{V(x)}$ so that the distribution of

$$\frac{\hat{m}(x) - \hat{b}(x) - m(x)}{\sqrt{\hat{V}(x)}}$$

is approximately standard normal. An approximate confidence interval for $m(x)$ can then be constructed in the usual way:

$$\hat{m}(x) - \hat{b}(x) \pm z_{\alpha/2}\sqrt{\hat{V}(x)}$$

where $z_{\alpha/2}$ is the appropriate quantile of the standard normal distribution.

The problem with this approach is that it is difficult to estimate $b(x)$. Also, the variance of $\hat{m}(x) - \hat{b}(x)$ is not necessarily the same as the variance of $\hat{m}(x)$ and, hence, it must be determined (and estimated).

A second approach is to use "undersmoothing" to reduce the bias of the estimator and to inflate the variance. Recall that the kernel estimator of $m(x)$ based on smoothing parameter h has bias of order $O(h^2)$ and variance of order $O(1/(nh))$. The optimal choice of h is of order $O(n^{-\frac{1}{5}})$, which yields a bias of order $O(n^{-\frac{2}{5}})$ and a variance of order $O(n^{-\frac{4}{5}})$. Thus, the bias and standard deviation are of the same order and, hence, their ratio is of order $O(1)$.

However, suppose we use a value of h that is smaller than the optimal choice. For instance, if $h = O(n^{-\frac{1}{4}})$, then the bias of the kernel estimator of $O(n^{-\frac{1}{2}})$ and the variance is of order $O(n^{-\frac{3}{4}})$. It follows that the ratio of the bias to the standard deviation is of order

$$\frac{O(n^{-\frac{1}{2}})}{O(n^{-\frac{3}{8}})} = O(n^{-\frac{1}{8}})$$

which approaches 0 as $n \to \infty$. It follows that, for such a choice of h,

$$\frac{\hat{m}(x) - m(x)}{\sqrt{\hat{V}(x)}}$$

is approximately standard normal.

The drawback of this approach is that standard methods of selecting h are designed to estimate the optimal value of h, which balances the orders of the bias and standard deviation of the estimator and it is not clear how to choose h to undersmooth enough so that the normal approximation to the distribution of

$$\frac{\hat{m}(x) - m(x)}{\sqrt{\hat{V}(x)}}$$

is valid, but not so much that $\hat{m}(x)$ is a poor estimator of $m(x)$.

The third approach is based on noting that $\hat{m}(x) - \mathrm{E}\,(\hat{m}(x))$ has an expected value exactly equal to 0 and, because $\mathrm{E}\,(\hat{m}(x))$ is non-random, the variance of $\hat{m}(x) - \mathrm{E}\,(\hat{m}(x))$ is identical to the variance of $\hat{m}(x)$. Hence, we expect that the distribution of

$$\frac{\hat{m}(x) - \mathrm{E}\,(\hat{m}(x))}{\sqrt{\hat{V}(x)}}$$

is approximately standard normal; then an approximate $(1 - \alpha) \times 100\%$ confidence interval for $\mathrm{E}\,(\hat{m}(x))$ is given by

$$\hat{m}(x) \pm z_{\alpha/2}\sqrt{\hat{V}(x)};$$

that is, instead of changing the form of the interval estimate, we interpret it as an interval for $\mathrm{E}\,(\hat{m}(x))$ rather than as an interval for $m(x)$.

The resulting confidence intervals, as functions of x, are often called "variability bands" because they reflect the variability in $\hat{m}(x)$ but they do not address any possible bias in $\hat{m}(x)$.

Example 8.4 Consider the fossil data. To add variability bands to the plot of the estimate $\widehat{m}(\cdot)$, the argument `se=T` should be included in the call to `sm.regression`. That is, the command

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

produces the plot given in Figure 7.6 but now with variability bands included; the result is given in Figure 8.2. □

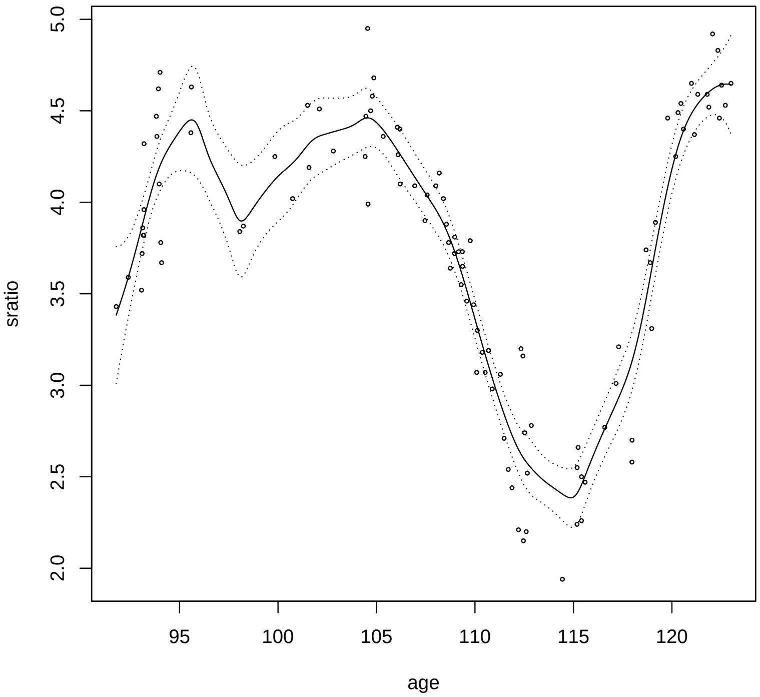

Figure 8.2: Local Linear Estimate of the Regression Function for the Fossil Data Together with Variability Bands

8.4 Applications of nonparametric regression

For a pair of random variables (X, Y), the nonparametric regression function $m(x) = \mathrm{E}(Y|X = x)$ gives the expected value of Y corresponding to a given value of X. Hence, like an estimate of a parametric regression model, an estimate $\widehat{m}(\cdot)$ of $m(\cdot)$ is useful for understanding the relationship between Y and X and for using X to predict the value of Y. In addition to these more direct uses of $\widehat{m}(\cdot)$, there are other uses of nonparametric regression estimates in which estimation of the function $\mathrm{E}(Y|X = x)$ is not the primary goal; in this section, we consider two of these.

Model checking in linear regression

When modeling the relationship between a response variable Y and a predictor variable X, the starting point is often a simple linear regression model of the form

$$Y = \alpha + \beta X + \epsilon \tag{8.8}$$

where ϵ is an unobserved random variable such that $\mathrm{E}(\epsilon|X) = 0$. Under this model, the relationship between Y and X is described by two parameters, α and β, with straightforward interpretations; for instance, β can be interpreted as a measure of the change in the expected value of Y corresponding to an increase in X of 1 unit.

However, the inferences based on such a model may be misleading if the true relationship between Y and X is nonlinear. Hence, a number of diagnostic methods have been proposed for determining the appropriateness of the simple linear regression model, such as residual plots and comparisons of the results from the model (8.8) to the result from a model including additional terms, such as a quadratic term of the form γX^2, for some nonzero value of γ. Here we consider an alternative approach based on comparing the results of the analysis based on a parametric regression model to those based on nonparametric regression analysis.

Consider a sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ from the distribution of (X, Y) so that, according to the model in (8.8),

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

Let $\hat{\alpha}$ and $\hat{\beta}$ denote the least-squares estimators of α and β, respectively, and let

$$e_j = Y_j - \hat{\alpha} - \hat{\beta} X_j, \quad j = 1, 2, \ldots, n$$

denote the residuals from the model. The fit of this model may be summarized by the "sum of squared errors", given by

$$\mathrm{SSE}_0 = \sum_{j=1}^{n} e_j^2.$$

Now consider using a nonparametric regression model for the relationship between Y and X,

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

where $\mathrm{E}(\epsilon | X) = 0$, and let $\widehat{m}(\cdot)$ denote the local-linear kernel estimator of $m(\cdot)$, with the smoothing parameter chosen by cross validation. The fit of this model may be summarized by the corresponding sum of squared errors

$$\mathrm{SSE}_1 = \sum_{j=1}^{n} (Y_j - \widehat{m}(X_j))^2.$$

Note that $\mathrm{SSE}_1 \leq \mathrm{SSE}_0$; however, if the linear model (8.8) fits the data well, we expect that

$$\mathrm{SSE}_0 \doteq \mathrm{SSE}_1.$$

Hence, the statistic

$$\frac{\mathrm{SSE}_0}{\mathrm{SSE}_1} \tag{8.9}$$

can be used as a measure of goodness-of-fit for the model $\alpha+\beta X$, with a large value indicating that the model does not fit the data well.

Under the assumption that the error terms $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent, normally distributed random variables, it is possible to derive an approximation to the distribution of the statistic (8.9) under the null hypothesis that the model

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

holds; that approximation may then be used to calculate a p-value for testing goodness-of-fit. Although the formal result applies only to normally-distributed data, it can be used as a rough guide more generally.

Example 8.5 A study was done on the relationship between the strength and elasticity of wood and its density. Samples of 75 different types of wood were analyzed; for each, the density (in g/cc), the modulus of elasticity (in kg per square mm) and the modulus of rupture (in kg per square mm) were measured. The question of interest is whether or not the modulus of elasticity and the modulus of rupture are each linearly related to the wood's density.

The variable `density` contains the measurements on density, the variable `elasticity` contains the measurements on the modulus of elasticity, and the variable `rupture` contains the measurements on the modulus of rupture. Figure 8.3 contains a plot of modulus of rupture versus density and Figure 8.4 contains a plot of the modulus of elasticity versus density. Both plots suggest that an approximate linear relationship is not unreasonable.

To test the hypothesis that a linear function explains relationship between the variables, we can use the function `sm.regression` with the argument `model = "linear"`. For example, to test the hypothesis that

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

where Y denotes the modulus of rupture and X denotes the density, we use the command

```
> sm.regression(density, rupture, method="cv", ngrid=1000, model="linear")
Test of linear model:  significance =  0.023
```

The small p-value suggests that the hypothesis that the linear relationship is appropriate is rejected; that is, according to these results, the relationship between modulus of rupture and density is not linear. The command also produces a plot of the local linear kernel estimate together with shading representing "reference bands" for the kernel estimate, calculated under the assumption that the null hypothesis holds; these bands indicate likely values for the kernel estimate if the linear relationship holds. The calculation of the reference bands

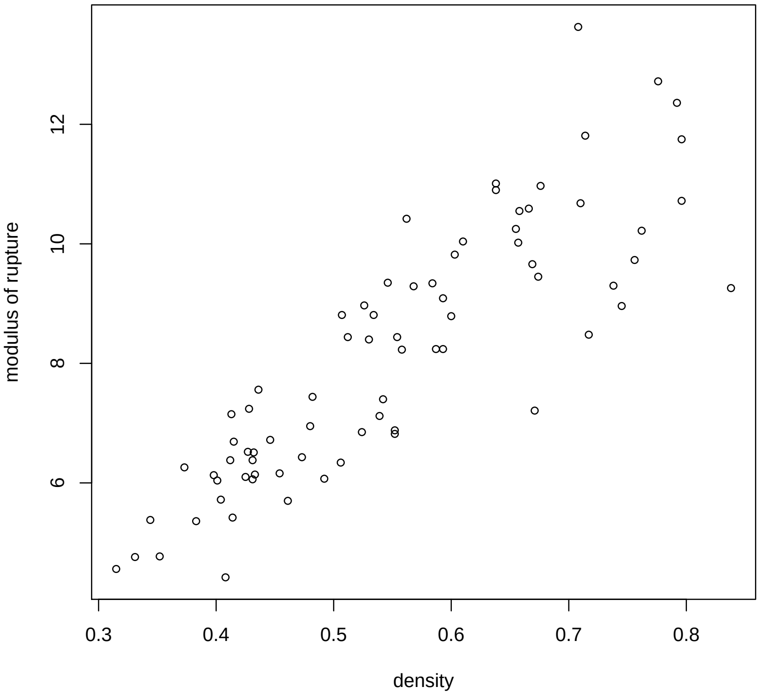

Figure 8.3: Modulus of Rupture vs Density for a Sample of 75 Woods

is based on the properties of $\widehat{m}(x) - \hat{\alpha} - \hat{\beta}x$, which are easily determined when the model $Y = \alpha + \beta X + \epsilon$ holds.

Given that the hypothesis of a linear relationship is rejected, we expect that the local linear kernel estimate will lie outside these reference bands, at least for some values of x; the relationship between the kernel estimate and the reference bands gives further information regarding the nonlinearity of the relationship. This plot is given in Figure 8.5; this plot shows that the relationship is nearly linear but there is some slight curvature.

The results of the test for elasticity are given by

```
> sm.regression(density, elasticity, method="cv", ngrid=1000, model="linear")
Test of linear model:  significance =  0.296
```

along with the plot in Figure 8.6. Thus, we do not reject the hypothesis that the relationship is linear and the plot shows that the kernel estimate lies inside the confidence bands for all x. □

Another test that can be performed using a `sm.regression` is a test of the hypothesis that

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

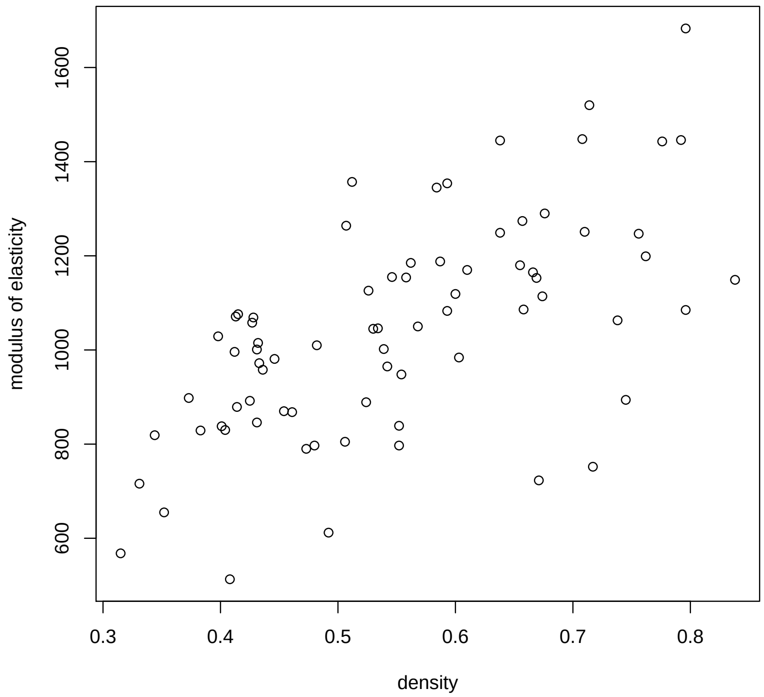

Figure 8.4: Modulus of Elasticity vs Density for a Sample of 75 Woods

Under this hypothesis, the conditional expected value of Y given X does not depend on X; to perform this test, we use the argument `model="no effect"`. When interpreting the results ot this test, it is important to keep in mind that, even if $\mathrm{E}(Y|X) = \alpha$ holds, it does not mean that Y and X are "unrelated" or independent; it could be that $\mathrm{Var}(Y|X)$ depends on X, for example.

Semiparametric regression models

Consider the case in which we are interested in the relationship between a response variable Y and a predictor variable X and that the relationship between Y and X is thought to be a linear one. Furthermore, suppose that there is a second predictor Z, possibly related to X, that also is believed to have an important effect on Y. Then we might consider a regression model with Z and X as predictors, such as

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

where β is the parameter of interest, describing the relationship between Y and X, for fixed values of Z.

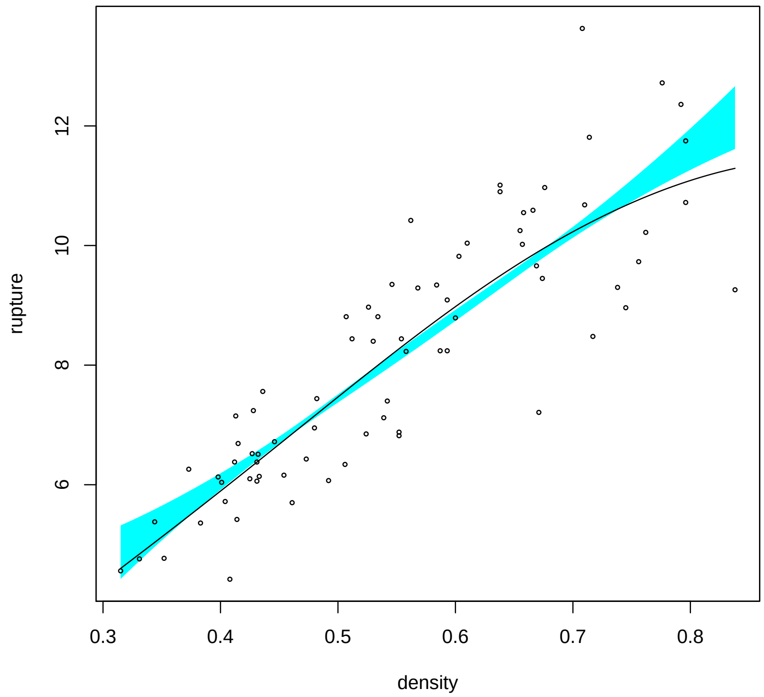

Figure 8.5: Kernel Estimate of the Function Relating Modulus of Rupture to Density with Confidence Bands for the Estimated Linear Relationship

However, suppose that the relationship between Y and Z may be nonlinear. In such cases, we might consider a *semiparametric regression model* of the form

$$Y = \beta X + m(Z) + \epsilon \tag{8.10}$$

where $m(\cdot)$ is an unknown function and the error term in the model, ϵ, satisfies $\mathrm{E}(\epsilon|X, Z) = 0$. In this model, the relationship between Y and X is a parametric one, described by the parameter β, which is of primary interest, while the relationship between Y and Z is a nonparametric one. This model is also called a "partially linear model" because it is a linear model with respect to X but not with respect to Z.

The challenge in estimating the parameters β and $m(\cdot)$ is to use a parametric-type method of estimation (such as least-squares) for β and a nonparametric-type method of estimation (such as kernel regression) for the unknown function $m(\cdot)$.

One way to achieve this is use a two-step approach to estimation. First, suppose that β is known. Then the model (8.10) may be written

$$Y - \beta X = m(Z) + \epsilon; \tag{8.11}$$

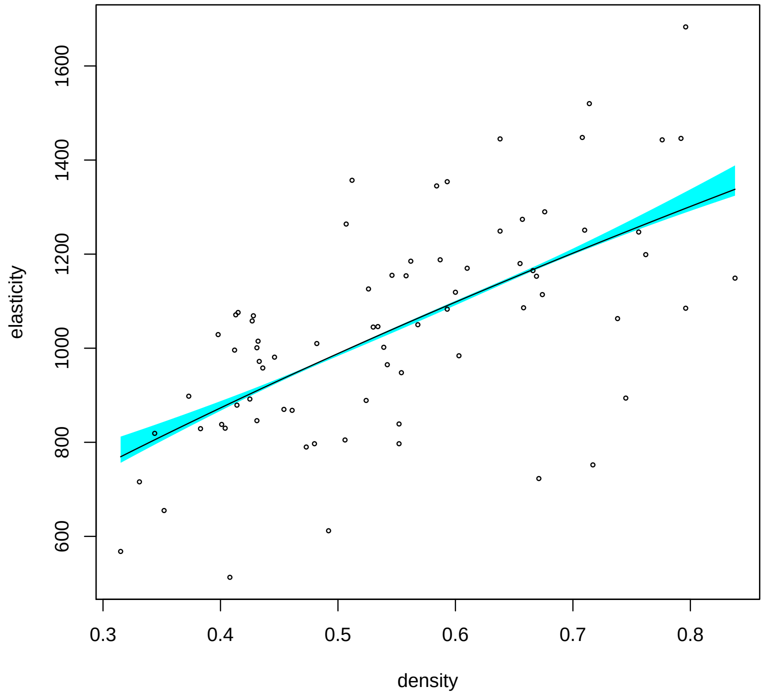

Figure 8.6: Kernel Estimate of the Function Relating Modulus of Elasticity to Density with Confidence Bands for the Estimated Linear Relationship

note that this is a standard nonparametric regression model, except that the response variable is $Y - \beta X$ rather than Y. Estimating $m(\cdot)$ in this model yields an estimator that depends on the value of the parameter β (as well as on the data).

Let $\widehat{m}_\beta(\cdot)$ denote a nonparametric regression estimator of $m(\cdot)$ in the model (8.11). We can now estimate β using least-squares in the model (8.10) substituting $\widehat{m}_\beta(\cdot)$ for $m(\cdot)$; that is, an estimator of β can be obtained by minimizing

$$\sum_{j=1}^{n}(Y_j - \beta X_j - \widehat{m}_\beta(Z_j))^2$$

where (X_j, Y_j, Z_j), $j = 1, 2, \ldots, n$ denote independent random vectors, each with the same distribution as (X, Y, Z).

Suppose that $\widehat{m}_\beta(\cdot)$ is a standard (i.e., "local constant") kernel estimator. Then

$$\widehat{m}_\beta(z) = \frac{\sum_{j=1}^{n}(Y_j - \beta X_j)K(\frac{z-Z_j}{h})}{\sum_{j=1}^{n} K(\frac{z-Z_j}{h})} = \frac{\sum_{j=1}^{n} Y_j K(\frac{z-Z_j}{h})}{\sum_{j=1}^{n} K(\frac{z-Z_j}{h})} - \beta\frac{\sum_{j=1}^{n} X_j K(\frac{z-Z_j}{h})}{\sum_{j=1}^{n} K(\frac{z-Z_j}{h})}.$$

Thus, we can write

$$\widehat{m}_\beta(Z_j) = \hat{Y}_j - \beta\hat{X}_j$$

where

$$\hat{Y}_j = \frac{\sum_{k=1}^{n} Y_k K(\frac{Z_j - Z_k}{h})}{\sum_{k=1}^{n} K(\frac{Z_j - Z_k}{h})}$$

and

$$\hat{X}_j = \frac{\sum_{k=1}^{n} X_k K(\frac{Z_j - Z_k}{h})}{\sum_{k=1}^{n} K(\frac{Z_j - Z_k}{h})}$$

denoted the “fitted values” from the nonparametric regressions of Y_j on Z_j and X_j on Z_j, respectively. A similar result holds if $\widehat{m}_\beta(\cdot)$ is a local-linear kernel estimator, although with different expressions for the fitted values $\hat{Y}_j$ and $\hat{X}_j$.

It follows that the estimator of β is the minimizer of

$$\sum_{j=1}^{n} (Y_j - \hat{Y}_j - \beta(X_j - \hat{X}_j))^2;$$

it follows that

$$\hat{\beta} = \frac{\sum_{j=1}^{n} (Y_j - \hat{Y}_j)(X_j - \hat{X}_j)}{\sum_{j=1}^{n} (X_j - \hat{X}_j)^2}.$$

This same approach can be used for more complex models that include several predictors $X_1, X_2, \ldots, X_p$ in the parametric component of the model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + m(Z) + \epsilon.$$

Example 8.6 A study was done of the efficacy of certain mildew treatments used on barley crops. A field was divided into 36 adjacent plots and each plot was assigned one of four treatments: one of three applications of the mildew control agent or no application, with the “no application” acting as a control. To ensure a degree of balance in the assignments, the 36 plots were divided in to 9 blocks of 4 plots and each treatment appears once in each block.

The variable `treat` contains the treatment for the 36 plots with $1, 2, 3$ denoting a specific treatment and 0 denoting that no treatment was applied:

```
> treat
 [1] 2 3 0 1 0 2 3 1 0 1 2 3 2 1 3 0 3 0 2 1 2 0 1 3 1 3 2 0 2 0 3 1 2 1 0 3
```

The response variable `yield` is the yield of grain in tons per hectare. A plot of the data is given in Figure 8.7, with the plotting symbol indicating the treatment used. Note that here the label “plot number” refers to the id number of the agricultural plot used in the experiment, ranging from 1 to 36.

To analyze the data, we can fit a model with yield as a qualitative predictor variable (i.e., a “factor”) which gives an estimate of the treatment effect relative to the control for each of the three treatments:

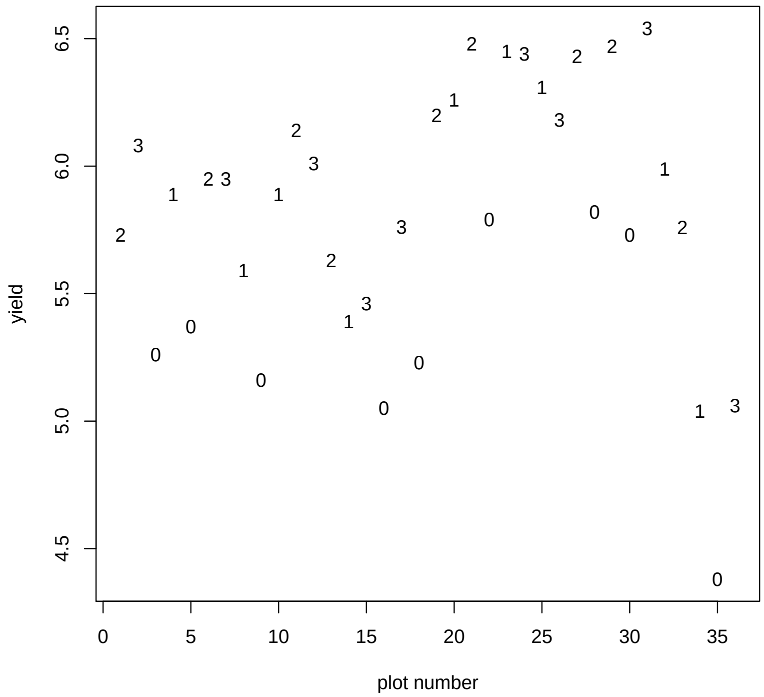

Figure 8.7: Crop Yields for the Different Treatments as a Function of Plot Position

```
> summary(lm(yield˜as.factor(treat)))
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.310      0.144   36.93  < 2e-16 ***
as.factor(treat)1    0.558      0.203    2.74  0.00989 **
as.factor(treat)2    0.778      0.203    3.82  0.00057 ***
as.factor(treat)3    0.632      0.203    3.11  0.00393 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.431 on 32 degrees of freedom
Multiple R-squared:  0.344,     Adjusted R-squared:  0.283
F-statistic: 5.61 on 3 and 32 DF,  p-value: 0.00332
```

However, there is some evidence that the fertility of the field varies across the plots. For instance, a plot of the residuals from the analysis above, along with a kernel estimate based on cross-validation, is given in Figure 8.8.

It is clear that the fertility pattern may influence the estimates of the treatment effects and, without accounting for fertility, the results on the effectiveness of the different treatments may be misleading. Also, much of the "unexplained variation", reflected in the estimate of the error standard deviation σ, may be attributable to the variation in fertility. Because this fertility pattern does not follow a simple parametric model, we might consider a semiparametric model in which the "treatment" predictor is modeled parametrically and the effect of the plot location is modeled nonparametrically.

Let X_1, X_2, X_3 denote indicator variables for the three non-control treatments so that, for example, $X_1 = 1$ if the plot received treatment 1 and $X_1 = 0$ otherwise, and let Z denote the plot id number (1 to 36). Then we might consider the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + m(Z) + \epsilon$$

where Y denotes the yield of the plot. Under this model, the expected yield for the control is $\beta_0 + m(Z)$ and for treatment j it is $\beta_j + m(Z)$.

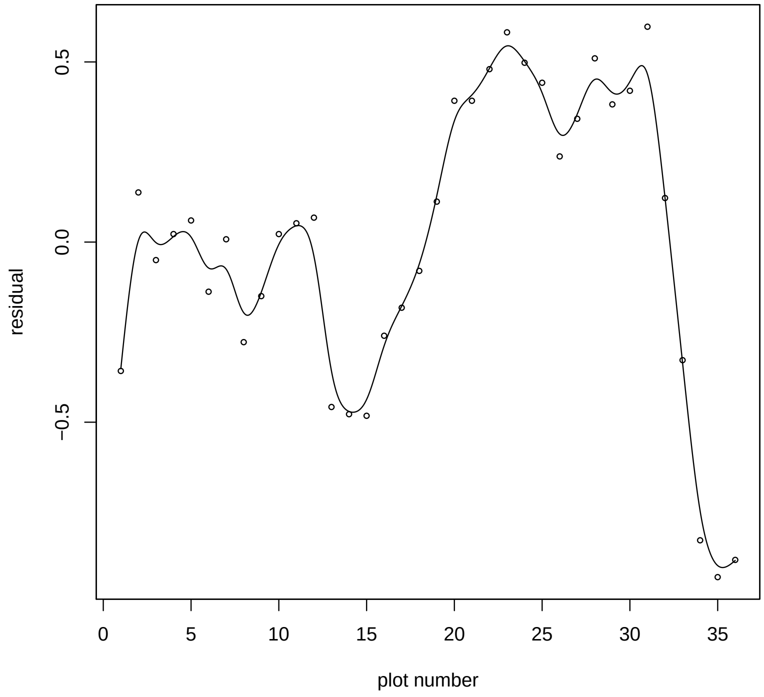

Figure 8.8: Residuals from the Analysis of Crop Yields with a Kernel Estimate

To fit such a model in R, we can use functions in the package "np". The variable `yield` contains the yield data, the variables `trt1`, `trt2`, and `trt3` are the indicator variables for treatments 1, 2, and 3, respectively, and `plotid` contains the plot id number.

Estimation of the parameters of the semiparametric model (8.10) uses a two-step procedure:

```
> library(np)
> barley.bw<-npplregbw(yield~trt1+trt2+trt3|plotid, regtype="ll")
> barley.res<-npplreg(barley.bw)
```

The first function, `npplregbw`, specifies the model using the usual R syntax, except that the variable appearing after the vertical line | are interpeted as the "nonparametric predictor" Z. The argument `regtype="ll"` specifies that the locally linear kernel estimator is to be used. The output from `npplregbw` is then used as the input to `npplreg`, which produces the parameter estimates and other information about the fitted model.

To obtain the parameter estimates for the coefficients of `trt1`, `trt2`, and `trt3` we use the following command

```
> coef(barley.res)
  trt1   trt2   trt3
0.5298 0.7091 0.6690
```

Including the argument `errors=T` returns the standard errors of the estimates:

```
> coef(barley.res, errors=T)
   trt1    trt2    trt3
0.09996 0.09998 0.09996
```

For example, an approximate 95% confidence interval for β_1 is given by

$$0.530 \pm 1.96(0.100) = (0.510, 0.550).$$

Additional information is provided by the `summary` function:

```
> summary(barley.res)


Partially Linear Model
Regression data: 36 training points, in 4 variable(s)
With 3 linear parametric regressor(s), 1 nonparametric regressor(s)


               y(z)
Bandwidth(s): 3.438


                   x(z)
```

```
Bandwidth(s): 757233413
              363809407
              329108142

                  trt1   trt2  trt3
Coefficient(s): 0.5298 0.7091 0.669

Kernel Regression Estimator: Local-Linear
Bandwidth Type: Fixed

Residual standard error: 0.1999
R-squared: 0.8457

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 1
```

Thus, the parameter estimates from the semiparametric regression model are generally similar to those based on the standard parametric regression model, with the biggest difference in the estimate of the effect of treatment 2. However, the standard errors of the estimates are much smaller for the semiparametric model because much of the variation in the data is now explained by the plot location; this feature is also reflected in the estimate of σ, which is 0.431 for the parametric model but only 0.200 for the semiparametric model.

To obtain a plot of the estimate of $m(\cdot)$ we may use the command

```
> barley1<-npplreg(barley.bw, exdat=matrix(0, 3600, 3),
+ ezdat=(1:3600)/100)$mean
```

which produces the fitted values for the “evaluation data” given by `exdat` and `ezdat`, which correspond to the variables modeled parametrically and nonparametrically, respectively. Hence, the evaluation data for `trt1`, `trt2`, and `trt3` are taken to be all 0s and the evaluation data for `plotid` is taken to be a fine grid of values from 1 to 36. It follows that `barley1` contains the values of $\widehat{m}(z)$ for a $1 \leq z \leq 36$. A plot of this estimate, generated by the command

```
> plot((1:3600)/100, barley1, type="l", xlab="plot", ylab="estimate of m")
```

is given in Figure 8.9. □

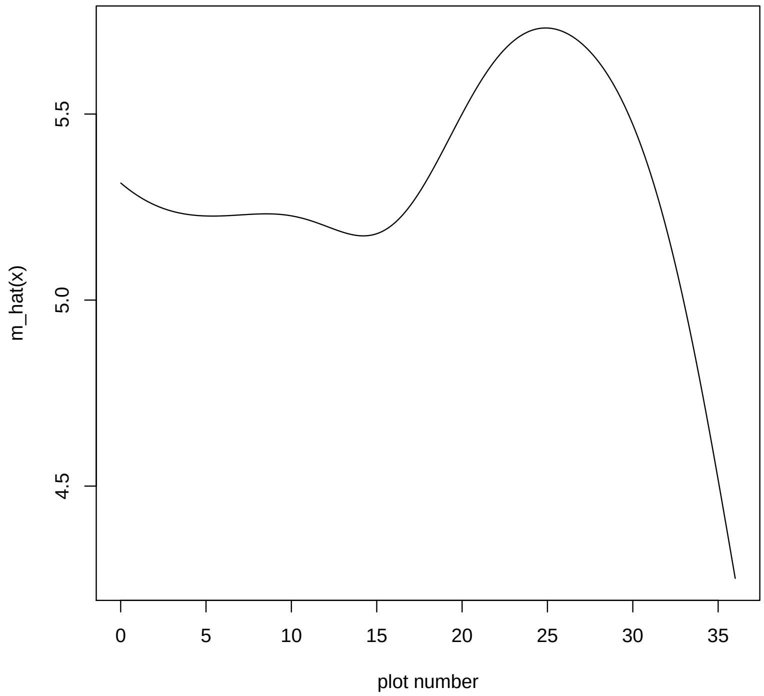

Figure 8.9: Function Estimate in the Semiparametric Model for the Analysis of Crop Yields

8.5 Exercises

8.1. Consider the dataset "runs" which contains the average runs scored per team per game in each year from 1900 to 2016 and which was used in Exercise 7.3.
Estimate the error standard deviation σ using the second-differences estimator

8.2. Consider the the dataset "geyser" which contains data on the time between eruptions and the length of the eruptions of the Old Faithful geyser and which was used in Exercise 7.2.

(a) For the regression function estimate calculated in Exercise 7.2 (based on the local linear estimator and cross validation), use the function `h2df` described in the notes to find the degrees-of-freedom of the estimate.

(b) Round off the degrees-of-freedom up to the next largest integer and denote that value by p. Use `lm` to estimate the parameters of the regression model

$$Y = \alpha + \beta_1 X + \cdots + \beta_{p-1} X^{p-1} + \epsilon$$

where Y duration of the eruption and X is the waiting time; note that, like the function estimate determined in Exercise 7.2, this model has p degrees of freedom.

(c) Using the estimates from part (b), plot the estimated regression line together with the data and compare the result to the estimate found in Exercise 7.2.

One way to plot the estimated regression line is to use the function `seq` to construct a vector `x`, starting at 43, ending at 108 and having step sizes of 0.1. Then use the command

```
> pred.y<-predict(reg.out, newdata=data.frame(waiting=x))
```

to calculate the predicted values corresponding to `x`. Here `reg.out` is a variable containing the output from `lm` and `waiting` is the variable containing the waiting times (i.e., the times between eruptions). Plotting `pred.y` versus `x` gives a plot of the estimated regression function.

8.3. The dataset "pulse" contains data on the heights and resting pulses of a sample of hospital patients. The purpose of this exercise is to try to determine if pulse is related to height.

(a) Use a local linear kernel estimator and cross-validation to estimate the nonparametric regression function relating pulse (the response variable) to height (the predictor variable). Plot the estimate together with the data. Based on this plot, does it appear that pulse is related to height?

(b) Compute the p-value for the test the hypothesis that conditional expected value of pulse given height does not depend on height; use the function `sm.regression` with the argument `model = "no effect"`. Based on this result, does it appear that pulse and height are related? Why or why not?

8.4. The dataset "electricity" contains data on the electricity usage of a house in Westchester County, New York. There are two variables: *usage*, the electricity usage (in kilowatt-hours) for a given month, and *temp*, the average temperature for the month in degrees F. The goal of the analysis is to model electricity usage as a function of temperature.

(a) Using a local linear kernel estimator, estimate and plot the nonparametric regression function relating usage to temp; choose the value of the smoothing parameter using cross-validation.

(b) Compute the p-value for the test the hypothesis that the relationship between usage and temp is linear. Based on this result, does it appear that the relationship between usage and temp is linear?

8.5. Repeat the analysis in the previous exercise, using log(usage) in place of usage. Does it appear that the relationship between log(usage) and temp is linear?

8.6. The dataset "trawl" contains data from a survery of the fauna on the sea bed in an area lying between the coast of northern Queensland and the Great Barrier Reef. In this exercise, we consider the relationship between the amount of fish (expressed in terms of score based on log-weight), given in the variable Score1, and the location at which the fish were captured, described by the variables Longitude and Latitude.

(a) Estimate the nonparametric regression function relating Score1 to Longitude; use a local-linear kernel estimate and cross-validation. Plot the data together with the function estimate. Comment on the nature of the relationship (i.e., linear or nonlinear).

(b) Repeat part (a) for the model relating Score1 to Latitude.

(c) Let Y denote Score1, let X denote Latitude, and let Z denote Longitude; consider a semiparametric regression model of the form

$$Y = \beta X + m(Z) + \epsilon.$$

Following the approach used in Example , use the functions `npplregbw` and `npplreg` in package "np" to estimate the parameters of the model. Plot the function estimate, using the procedure described in Example , setting `ezdat` to be the vector given by `seq(142.8, 144, 0.01)` and taking `exdat` to a vector of the same length as `ezdat`, with all values equal to 0.

(d) Find the standard error of $\widehat{\beta}$ and use the standard error to construct an approximate 95% for β. Compare the result to the approximate confidence interval obtained for the slope parameter in a linear regression model relating Score1 and Latitude (ignoring Longitude).