7 Additive and generalised additive modelling

7.1 Introduction

When the data under investigation do not show a clear linear relationship, then additive modelling is a suitable alternative to linear regression. Figure 7.1 shows a scatterplot for two variables of the RIKZ data: species richness and grain size. See Chapter 27 for details on these data. A first look at the graph suggests there is a non-linear relationship between richness and grain size. Sites with large grain sizes seem to have a fairly constant but low species richness, with richness increasing as grain size decreases. Applying a linear regression model with richness as the response variable and grain size as the explanatory variable gives residuals showing a clear pattern, indicating a serious model misspecification.

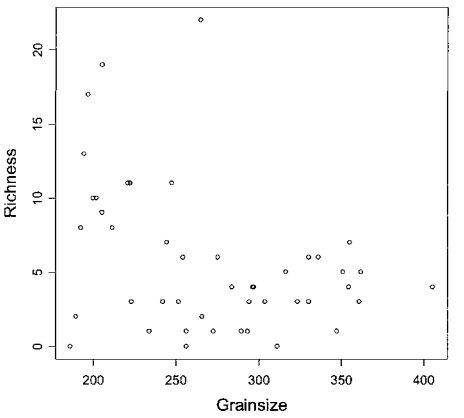

Figure 7.1. Scatterplot of species richness versus grain size for the RIKZ data.

If a data exploration suggests a clear non-linear relationship between the response variable and explanatory variable, then there are various options. First, you could apply a data transformation on the response variable or the explanatory variable (or both) to try to linearise the relationship. In this case, a square root or logarithmic transformation on grain size might produce a linear relationship be-

tween the two variables (see also the weight-length wedge clam example in Chapter 4) .

Another option is to model the non-linear relationship using interaction terms between explanatory variables. Suppose the samples with lower grain size were collected during the first few weeks of June 2002, and the other samples were collected towards the end of the month. Then adding a week-grain size interaction term will allow modelling of the non-linear relationship. The problem with interaction terms is deciding which interaction terms to use, especially if there are more than three explanatory variables.

Another alternative is to apply a smoothing method such as additive modelling or generalised additive modelling (GAM). These methods use smoothing curves to model the relationship between the response variable and the explanatory variables. Both allow for non-linear relationships, a common feature of many ecological datasets, and can be used to verify the results of linear regression or GLM models. This can provide confidence that applying a linear regression or GLM is the correct approach. In fact, you can consider linear regression and GLM as special cases of additive and generalised additive models. Indeed the smoothing equivalents of the bivariate and multiple linear regression models could be called bivariate additive models and multiple additive models. However, you are unlikely to see these names in the literature. Good GAM references are Hastie and Tibshirani (1990), Bowman and Azzalini (1997), Schimek (2000), Fox (2002a,b), Rupert et al. (2003), Faraway (2006) and Wood (2006). Although, all are fairly technical, Fox and Faraway are mathematically the simplest.

With linear regressions, we discussed several potential problems such as violation of homogeneity and negative fitted values for count data. As solutions we introduced the Poisson model for count data, and the logistic regression model for presence–absence data. This led to the general framework of generalised linear modelling. The same can be done for smoothing models. For count data, generalised additive models using the Poisson distribution with a log link function are used, and for presence–absence data, we use GAM with the binomial distribution and logistic link function. To understand the additive modelling text below, the reader needs to be familiar with linear regression as summarised in Chapter 5. And understanding GAM requires a detailed knowledge of additive modelling (discussed in this chapter) and GLM, which we discussed in Chapter 6.

Underlying principles of smoothing

The underlying principle of smoothing is relatively simple and is visualised in Figure 7.2 Figure 7.2-A shows a scatter plot of 100 artificial data points. Instead of having a line that goes through each point, we want to have a line that still represents the data reasonably well, but is reasonably smooth. The motivation for smoothing is simply because smooth lines are easier to interpret. A straight line is an extreme example of a smoothed line and although easy to interpret, it might not be the most representative fit of the data. The other extreme is a line that goes through every data point, which although providing a perfect fit may lack any smoothing and therefore be difficult to interpret. Before discussing how smooth

we need the curve, we need to discuss the mathematical mechanism that provides the smoothing. In Figure 7.2-B a box is drawn around an arbitrary chosen target value of $X = 40$. At this value of X, we would like to obtain a smoothing value (a Y-value). The easiest option is to take the mean value of all samples in the box. Moving the box along the X gradient will give an estimated value of the smoother at each target value. The resulting smoother of this technique is called a running-*mean* smoother. Another option is to apply a linear regression using only the points in the box. The smoothed value is then the fitted value at the target value of $X = 40$. The fitted line within the box is shown in panel C. Moving the box along the X gradient gives the running-*line* smoother. As well as the running-mean and running-line smoothers, a range of other methods are available, for example: bin smoother, moving average smoother, moving median smoother, nearest neighbourhood smoother, locally weighted regression smoother (LOESS), Gaussian kernel smoother, regression splines, smoothing splines, among others. The mathematical background of most of these smoothers can be found in Hastie and Tibshirani (1990).

An important smoother is the LOESS smoother (Figure 7.2-D). This smoother works like the running-line smoother except weighting factors for the regression are used within the box. The further away a point is from the target value, the less weight it has. Most statistics programmes have routines for LOESS smoothing and smoothing splines, and we discuss these smoothers in detail later.

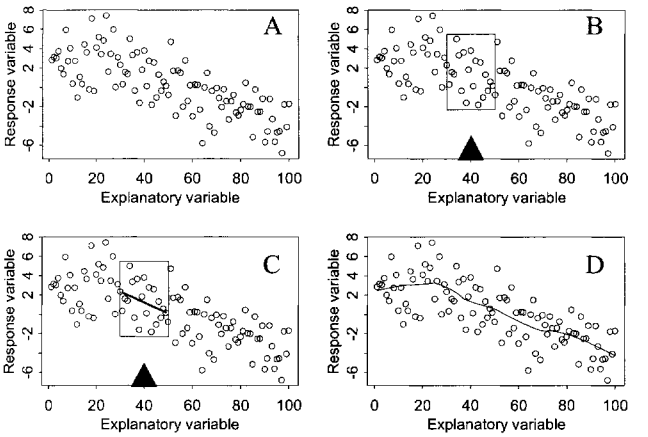

Figure 7.2. Underlying principle of smoothing. A: scatter plot of artificial data. B: A box around the target value of $X = 40$. The width of the box is 20 points (from 30 to 50). C: One option to get an estimated value at $X = 40$; the line within the box is obtained by linear regression. D: LOESS smoother obtained by applying weighted regression in the box and shifting the box along the gradient; the default span width of 0.5 was used (span width will be discussed later).

A crucial point we ignored in the above discussion is the size of the box. Regardless of the smoother used, the researcher still has to choose the width of the box. If the smoother does not use a box (like splines), it will have something equivalent. However, for the moment, we will explain this problem in terms of box width (also called span width), as it is conceptually easier to follow. To illustrate the effect of span width, a LOESS smoother with a span width of 0.7 is shown in panel A of Figure 7.3. A span of 0.7 means that at each target value, the width of the box is chosen such that it contains 70% of the data. If the points along the X gradient are unequally spaced, as with this example, then the width will change for different target values. Panel B is the LOESS smoother using a span of 0.3, which allows for more variation. Panel C shows the LOESS smoother with the default value of 0.5. Finally, panel D is a smoothing spline with the default amount of smoothing (4 df, which we explain later). Each panel gives different information; the smoother in the upper left panel indicates a slow decrease in species richness until a grain size of 275, and then it maintains an approximately constant value. The smoother in panel B shows two peaks in species richness and includes the two samples with low richness and grain size. The smoothing spline (panel D) gives a slightly smoother curve compared with LOESS with a span 0.5 (panel C).

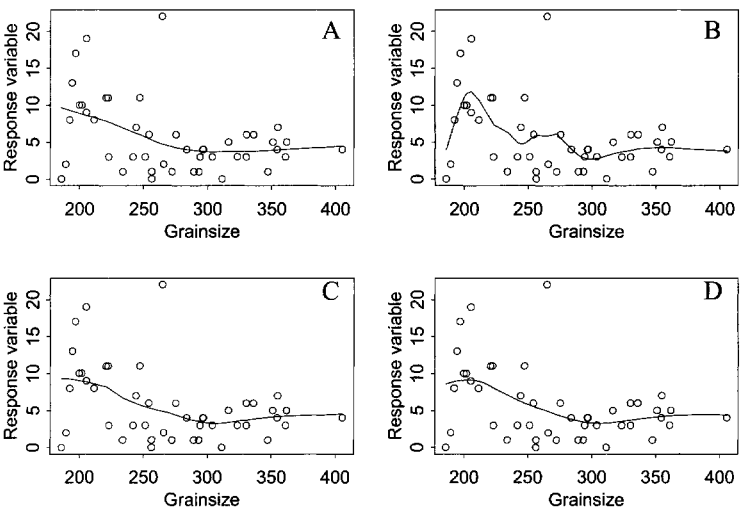

Figure 7.3. Four scatterplots and LOESS smoothers for species richness and grainsize for the RIKZ data. A: LOESS smoother with a span of 0.7. B: LOESS smoother with a span of 0.3. C: LOESS smoother with a span of 0.5, the default value in software packages like SPlus and R. D: Smoothing spline with 4 degrees of freedom (default value).

This arbitrary nature of span width might make smoothing methods confusing. However, if used with common sense, it is a useful method in the toolbox of any scientist. In the remaining part of this chapter we discuss the mathematical framework of additive modelling, the LOESS smoother and smoothing splines, the similarities with regression, model selection (including span width selection), model validation and extensions to the method for count data and presence–absence data.

7.2 The additive model

Before moving on to the additive model we begin by re-visiting the bivariate linear regression model given by

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad \text{and} \quad \varepsilon_i \sim N(0,\sigma^2)$$

where Y_i is the value of the response variable at sample i, X_i is the explanatory variable, and α and β are the population intercept and slope, respectively. The two most important assumptions are normality and homogeneity. The advantage of the linear regression model is that the relationship between Y and X is represented by the slope β. Hence, we only have to look at the estimated slope and its confidence interval to see whether there is a relationship between Y and X. The smoothing equivalent in an additive model is given by:

$$Y_i = \alpha + f(X_i) + \varepsilon_i \quad \text{and} \quad \varepsilon_i \sim N(0,\sigma^2)$$

The function $f()$ is the population smoothing function. This model is also called an additive model with one explanatory variable. So instead of summarising the relationship between Y and X using a slope parameter β, we use a smoothing function f. The disadvantage of this approach is that a graph is needed to visualise the function f(X). We will discuss later how to obtain estimates for the intercept and smoothing function.

The additive model $Y_i = \alpha + f(X_i) + \varepsilon_i$, is the equivalent of the bivariate linear regression model $Y_i = \alpha + \beta X_i + \varepsilon_i$. This equivalence allows us to present the underlying statistical principles of the additive model (Figure 7.4) just as we did for linear regression (Figure 5.5). In Figure 7.4, we have used species richness (R) and grain size from the RIKZ data to visualise the bivariate additive model. The observed richness values (points) and fitted values are plotted in the R-grain size space. Gaussian density curves are plotted on top of the fitted values indicating the range and likelihood of other realisations of richness against grain size. In the additive model we assume that R_i is normally distributed with expectation μ_i and variance σ^2, and $E[Ri] = \mu_i = \alpha + f(X_i)$. The only conceptual difference between Figure 7.4 here and Figure 5.5 from the regression chapter is the replacement of the regression curve with a smoothing curve. This similarity means that we can expect the additive model to share the same problems and solutions as we discussed for linear regression, and we discuss this later.

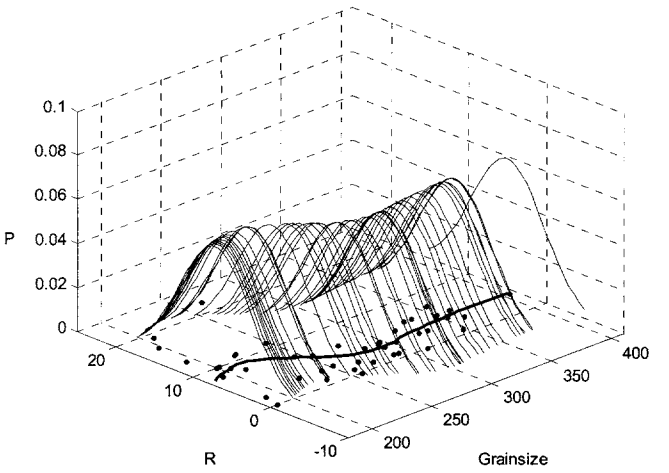

Figure 7.4. Visualisation of a bivariate additive model for the RIKZ data. The grain size richness relationship is modelled as a smoothing curve, and Gaussian density curves indicate the probability of other realisations.

7.3 Example of an additive model

To illustrate an additive model with only one explanatory variable, we use a Bahamas fisheries dataset (unpublished data from The Bahamas National Trust / Greenforce Andros Island Marine Study). The aim of this study was to find a relationship between reef fish and benthic habitat using data collected between February 2002 and September 2003. Here, the parrotfish density is used as the response variable and the explanatory variables are related to algae and coral cover. These are all nominal variables and include location, time (month), and survey method (method 1: point counts, method 2: transects). For this example we have only used the transect data (244 samples) with coral richness used as the explanatory variable. A scatterplot of the data is given in Figure 7.5-A. This shows that sample stations with the highest numbers of coral species also seem to have the highest densities of parrotfish. Dotplots and boxplots indicated there were no extreme observations. The additive model used is of the form:

$$\text{parrotfish}_i = \alpha + f(\text{coral richness}_i) + \varepsilon_i \quad \text{and} \quad \varepsilon_i \sim N(0,\sigma^2)$$

The smoothing function of coral richness is presented in panel B of Figure 7.5 and was estimated using a smoothing spline, which is explained later in this chapter. The numerical output (given below) shows that the intercept is equal to 6.45. This means that the fitted value at a particular station is obtained by

$$\text{parrotfish}_i = 6.45 + f(\text{coral richness}_i)$$

The smoothing function is the solid line in Figure 7.5-B. The confusing aspect of panel B is that both axes have labels that contain 'Coral richness'. The y-axis represents the value of the smoother, the software uses the notation s() instead of f(), and the x-axis contains the coral richness values. The shape of the smoothing curve indicates that for samples with a coral richness between 3 and 8, the expected parrotfish density is approximately 4.5 parrotfish (6.45 minus 2; the fitted value of the smoother minus the estimated intercept). The highest densities can be found at samples with a coral richness of about 12 (the density will be around 6.45 + 4 = 10.45), but there is a decline if the coral richness is higher.

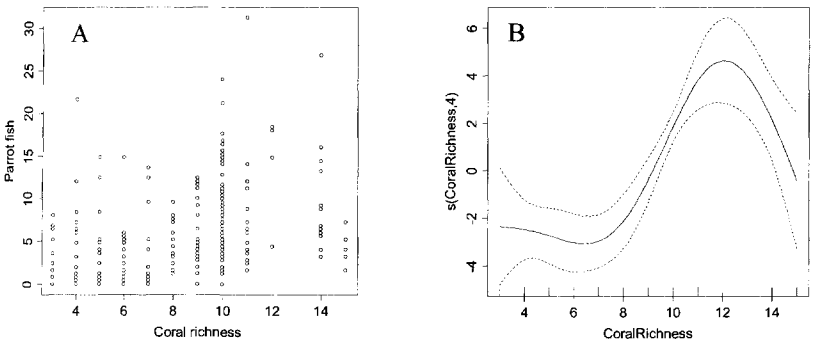

Figure 7.5. A: Scatterplot of coral richness against parrotfish density for the Bahama fisheries data. B: Smoothing function for the additive model applied on the Bahamas fisheries data. The solid line is the estimated smoother, and the dotted lines are point-wise 95% confidence bands.

The numerical output of the model is given by

Parametric coefficients:

	Estimate	std. err.	*t*-ratio	*p*-value
(Intercept)	6.44	0.311	20.74	<0.001

Approximate significance of smooth terms:

	edf	*F*	*p*-value
s(CoralRichness)	4	14.87	<0.001

R-sq.(adj) = 0.18 Deviance explained = 19.9%
Variance = 23.6 n = 244

This output indicates that the smoother is significant (an F-test is used); hence, there is a significant non-linear relationship between parrotfish and coral richness, provided a model validation would show that all assumptions are met. However, the explained variance (labelled as deviance by the software) is only 19.9%. This is the equivalent of the R^2 in linear regression. Adding more explanatory variables might improve the model. We will come back to model validation and model selection later in this chapter.

7.4 Estimate the smoother and amount of smoothing

The questions not yet addressed are (i) how do we estimate the smoother, and (ii) how to choose the optimal amount of smoothing. Starting with the first question, one of the simplest smoothers is the moving average smoother. Here is how it works. What is the mean value of the numbers 1, 2, 3, 4 and 5? The correct answer is 3, but to come to this you added them all up and divided by 5. Or in formula:

$$\text{mean} = \frac{1+2+3+4+5}{5} = \frac{1}{5}1+\frac{1}{5}2+\frac{1}{5}3+\frac{1}{5}4+\frac{1}{5}5$$

So, each value is multiplied with a weighting factor 1 over 5. In fact, we can easily extend the series with the values 0x6, 0x7, 0x8, etc. If we want to emphasise the importance of some values, then we can change the weighting factors. For the artificial data in Figure 7.2 we can use the following formula to estimate a smoothing value at the target value of $X = 40$:

$$\hat{Y}_{40} = a_{30}Y_{30} + \ldots + a_{40}Y_{40} + \ldots + a_{50}Y_{50}$$

The box contains 21 neighbouring points around the target value (X=40). If we choose the weighting factors a_i as 1 over 21, we just get the mean value (as in the example above). But suppose we only want the five neighbouring observations to have an influence on the target value. A possible choice of coefficients is

$$\hat{Y}_{40} = 0+\ldots+0+0.1*Y_{38}+0.2*Y_{39}+0.4*Y_{40}+0.2*Y_{41}+0.1*Y_{42}+0+\ldots+0$$

Hence, the fitted value at $X = 40$ is a weighted average of five neighbouring points. Note that this is the same principle as calculating the mean of the numbers 1 to 5. In matrix (or vector) notation, this can be written as

$$\hat{Y}_{40} = \begin{pmatrix} 0 & \cdots & 0 & 0.1 & 0.2 & 0.4 & 0.2 & 0.1 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_{37} \\ Y_{38} \\ Y_{39} \\ Y_{40} \\ Y_{41} \\ Y_{42} \\ Y_{43} \\ \vdots \\ Y_{100} \end{pmatrix} = s \times Y$$

where $s = (0,\ldots,0, 0.1, 0.2, 0.4, 0.2, 0.1\ 0, \ldots, 0)$ and $Y = (Y_1,\ldots,Y_{100})'$. In fact, the fitted values at all target values can be obtained using a single matrix multiplication:

$$\hat{Y} = S \times Y$$

The matrix S contains all weighting factors and $\hat{Y}$ all target values. Filling in all elements of S is not an exercise to do by hand, but it is straightforward using a computer. The estimated smoother can also be written as (the difference between fitted values and smoother is only the intercept):

$$\hat{f}(X) = S \times Y$$

We use the hat notation to indicate that it is an estimated function. So, where does the underlying theory like p-values, F tests, t-values, etc., come from in additive modelling? In the next couple of paragraphs we will show how the expression for the smoother is related to the expression for the linear regression model. This means that we have to give some mathematical detail based on matrix algebra. You may skip these paragraphs, or only try to catch the concepts, if you are not familiar with matrix algebra.

Using simple algebra, the fitted values of a linear regression model $Y = \beta X + \varepsilon$ (for ease of notation the intercept was written within β by setting the first column of X equal to one) can be written as

$$\hat{Y} = X(X'X)^{-1}X'Y$$

This is a standard formula obtained by ordinary least squares and can be found in any good linear regression textbook (e.g., Montgomery and Peck 1992). A weighted linear regression model gives the following formula for the fitted values:

$$\hat{Y} = X(X'WX)^{-1}X'WY$$

The matrix W contains the weights. Again, this formula can be found in any textbook that covers weighted linear regression. It is easy to see that this formula can be written as

$$\hat{Y} = HY, \qquad \text{where} \qquad H = X(X'WX)^{-1}X'W$$

H is called the hat matrix and was used in Chapters 4 and 5 to identify possible influential observations (leverage). The LOESS smoother applies a weighted linear regression on the data within each box. The LOESS smoothing value at the target X is then the fitted Y value for this X value. The LOESS weights are obtained as follows: Points outside the box have a weight of 0 and points inside the box have weights following a unimodal pattern centred at the target value. The formula for the weights can be found in Chapter 2 of Hastie and Tibshirani (1990). However, you do not need to know exactly how the weights are calculated, and we do not discuss it further. The important thing to remember is that LOESS smoothing can be written as $\hat{f}(X) = S \times Y$; all that is required is to fill in the elements of S, which we can leave to the computer.

Splines

Only a short introduction into splines is given here (based on Fox 2000, 2002b), and the interested reader is referred to Hastie and Tibshirani (1990) or Wood (2006) for more details. A polynomial model for the Bahama's Parrotfish example has the following form:

$$\text{parrotfish}_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \beta_4 X_i^4 + \ldots + \beta_p X_i^p + \varepsilon_i$$

where $\varepsilon_i \sim N(0,\sigma^2)$ and X_i is the coral richness. A cubic polynomial uses terms up to, and including, X_i^3. The fit of the cubic polynomial model is given in Figure 7.6-A. The main problem with polynomial functions is that the fit is often rather poor, even though it seems to perform reasonably well in this example. To improve the fit, we can split the data into two parts based on the values of coral richness, and fit a separate cubic polynomial model on each dataset. We arbitrarily decided to split the data at a richness value of eight and the fitted values for both datasets are presented in the Figure 7.6-B. If required, we could continue splitting the data and repeat the exercise until a satisfactory fit is achieved.

The points along the x-axis where we split the data are called knots. The problem with this approach is seen in Figure 7.6-B, where the lines are not continuous at the point where they should meet. In some cases, they might not meet at all, and this defeats the main aim of obtaining a smooth curve.

A solution is to use a cubic regression spline, which is a third-order polynomial function with the following conditions:

- The curves must join at each knot.
- The first derivative must be continuous at each knot.
- The second derivate must be continuous at each knot.

A *natural* cubic regression spline is like a cubic regression spline, but with an additional constraint that forces a linear fitted line beyond the smallest and largest X values. This aims to avoid spurious behaviour at the ends of the gradient. But, how do you know how many parameters are used? It is important to know this as we may be over-fitting the data. Assume K knots are selected. This means there are $K + 1$ datasets, with a cubic polynomial applied with four parameters (three slopes, one for each term X_i, X_i^2 and X_i^3, and one intercept), to each dataset. However, at each knot we have three restrictions (the curves must join and first and second derivatives must be continuous), plus two at the edges of the gradient. This means that the total number of parameters is $4(K + 1) - 3K - 2 = K + 2$. The problem is knowing how many knots to choose: a similar problem as deciding on the span width in LOESS.

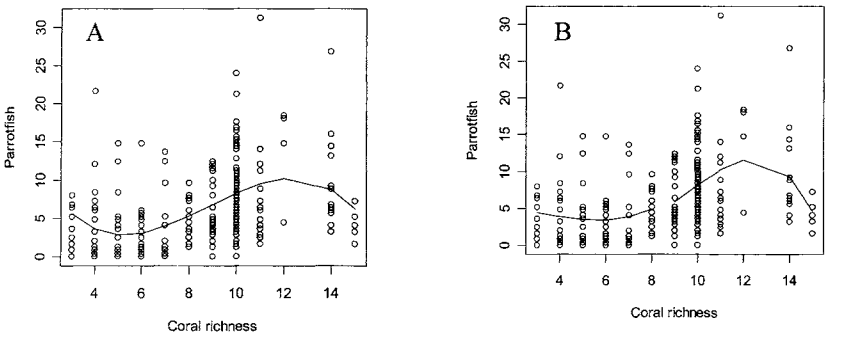

Figure 7.6. A: Fitted line obtained by a cubic polynomial model for the parrotfish of the Bahamas data. B: The data were split up in two parts (coral richness smaller and equal to eight, and larger than eight), and in each dataset a cubic polynomial model was applied.

Finally, we look at smoothing splines, starting with the formula called the penalised sum of squares:

$$SS(h) = \sum_{i=1}^{n} (Y_i - f(X))^2 + h \int_{x_{min}}^{x_{max}} f''(x)^2$$

This formula looks complex but its rationale is fairly simple (and is based on high school mathematics). The function $f()$ is again the smoother. The first part is a sum of residual squares. The less smooth the function $f()$, the smaller the residual sum of squares. The integral measures the smoothness of the function $f()$ and h is the penalty for non-smoothness. If $f()$ is rough, its second derivative will be large, and therefore the integral will be large as well. The aim is to minimise SS(h). Choosing a small value of h will give a low penalty to roughness, whereas choosing a large value gives a high penalty. Extremely small values, such as $h = 0$, will give a line that goes through every point, and a very large value of h will give a linear regression line. The function $f()$ that minimises SS(h) is a cubic smoothing spline, and the roughness penalty ensures that there are not too many parameters. Smoothing splines can also be written as $\hat{f} = SY$.

Degrees of freedom

We mentioned above that the fitted values in the linear regression model $Y = \beta X + \varepsilon$ are given by:

$$\hat{Y} = X(X'X)^{-1}X'Y$$

This is commonly written as:

$$\hat{Y} = HY, \qquad \text{where} \qquad H = X(X'X)^{-1}X'$$

The matrix H is called the hat matrix because it puts a hat on Y (it is estimated from the sample data) and, among other things, is used to calculate leverage. The degrees of freedom for the model is the number of parameters, and this is equal to the number of columns (variables) in X, assuming there are no problems of perfect collinearity that would require removing some collinear variables from the analysis. Another way to calculate the degrees of freedom is to use the rank or trace of the matrix H, where the trace is the sum of the diagonal elements. In linear regression, residuals can be calculated as observed values minus fitted values. In matrix notation, this is

$$e = Y - \hat{Y} = Y - HY = (I - H)Y$$

This shows the important role of the hat matrix in linear regression, with the degrees of freedom for the error given by df_{res}= rank($I - H$) or n − trace(H).

The reason for introducing this equation is because in additive modelling degrees of freedom are defined in a similar way. Our starting point is:

$$\hat{f} = SY$$

And we only have to replace H by S to obtain the degrees of freedom. Although there are some theoretical problems with this method, software like R and SPlus use this analogous approach, which also gives confidence bands for the smoothers:

$$\text{Covariance } \hat{f} = \sigma^2 SS'$$

The same issues exist for hypothesis tests in additive modelling (testing whether a smoother is equal to zero), which also only mimic linear regression, with no formal justification. However, Hastie and Tibshirani (1990) mention that empirical studies have shown that hypothesis tests are reasonably robust.

7.5 Additive models with multiple explanatory variables

Recall that the multiple linear regression model is given by

$$y_i = \alpha + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \qquad \text{and } \varepsilon_i \sim N(0,\sigma^2)$$

leading to an additive model for p explanatory variables being defined by

$$y_i = \alpha + f_1(x_{i1}) + \ldots + f_p(x_{ip}) + \varepsilon_i, \qquad \text{and } \varepsilon_i \sim N(0,\sigma^2)$$

Each function $f_j()$ is a smoothing curve for the population, and can be estimated by using, for example, a running-mean smoother, a LOESS smoother or a smoothing spline.

As an example, we again use the RIKZ data to investigate whether there is a relationship between species richness and the explanatory variables: temperature,

grain size and exposure. Exposure has only three values and is best modelled as a nominal variable with the results presented in a table rather than a graph. The additive model is of the form:

$$R_i = \alpha + \text{Exposure}_i + f_1(\text{Temperature}_i) + f_2(\text{Grain size}_i) + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$. The estimated smoothing curves for temperature and grain size, using the default amount of smoothing, are shown in Figure 7.7. The numerical information related to the intercept and exposure are shown in Table 7.1.

The fitted value for any particular sample receives a contribution from four terms: the intercept, exposure, temperature and grain size. The intercept is 17.68. Keep in mind that we are fitting species richness: the number of species recorded at a specific site. If a sample is from a site with exposure 3 (first class), a value of 0 is added. This is because the first exposure level is used as a baseline. So for sites with an exposure of 3, the exposure plus intercept is 17.68 + 0 (number of species). If a site has an exposure of 10, then 6.49 fewer species are counted. At sites with an exposure of 11, the richness is 19.46 lower, than the first class, with an exposure value of 3. The p-values suggests that exposure (at least for class 11) is highly significant at the 5% level. Later in this section, the F-test is used to obtain a single overall p-value for exposure. As well as exposure, the fitted values get some contribution from the smoothing curves for temperature and grain size. High temperature and low grain size is related to low richness, and high grain size and low temperature is related to higher richness (a larger number of species). Note that nothing stops the model from obtaining negative fitted values!

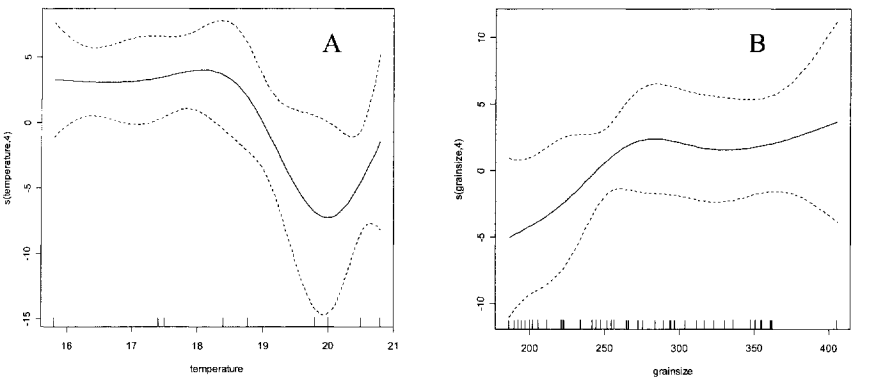

Figure 7.7. Smoothing curves for temperature (A) and grain size (B), obtained by additive modelling. The RIKZ dataset was used. The dotted lines are 95% confidence bands.

Table 7.1. Estimated parameters, standard errors, *t*-values and *p*-values of the parametric components in the additive model for the RIKZ data. Of the three nominal values for Exposure, the first exposure class is considered a baseline and adopts a value of zero. It is not reported in the table.

	Estimate	Std error	*t*-value	*p*-value
Intercept	17.32	3.40	5.09	<0.001
Exposure.10	−7.51	3.82	−1.96	0.06
Exposure.11	−18.67	4.81	−4.81	<0.001

Model selection

The question now is whether the smoothing terms are significantly differently from 0. The relevant numerical output produced is as follows.

```
Approximate significance of smooth terms:
                edf     F       p-value
s(temperature)  4       2.45    0.06
s(grain size)   4       0.56    0.69

R-sq.(adj) = 0.359  Deviance explained = 50.5%
GCV score = 21.845  Scale est. = 16.05    n = 45
Dispersion parameter= 16.05,
Deviance= 545.85, df= 11
```

For the Gaussian additive model that we are looking at here, the dispersion parameter (also denoted by the scale estimator in the numerical output) is the variance σ^2 (16.05). The deviance is equivalent to the residual sum of squares and explains 50.5% of the null deviance, which is the equivalent of the total sum of squares in linear regression. GCV is the cross-validation score and is explained later. The *p*-value for grain size indicates that the smoothing component is not significant, and could probably be dropped from the model.

As with linear regression, we can omit an explanatory variable and then compare the fit of the two nested models with each other: with and without the dropped variable. This can be done with an *F*-test, or using the AIC. The *F*-test uses the residual sum of squares of the full model (RSS_2) and compares them with residual sum of squares of the nested model (RSS_1):

$$F = \frac{(RSS_1 - RSS_2)/(df_2 - df_1)}{RSS_2 / df_{res}}$$

where df_j is the degrees of freedom in model j, df_{res} is $n - df_2$, and n is the number of samples. The *F*-ratio can be compared with an *F*-distribution with $df_2 - df_1$ and df_{res} degrees of freedom. The models compared are:

$$R_i = \alpha + \text{factor(Exposure}_i) + f_1(\text{Temperature}_i) + f_2(\text{Grain size}_i) + \varepsilon_i \quad (1)$$
$$R_i = \alpha + \text{factor(Exposure}_i) + f_1(\text{Temperature}_i) + \varepsilon_i \quad (2)$$
$$R_i = \alpha + \text{factor(Exposure}_i) + f_2(\text{Grain size}_i) + \varepsilon_i \quad (3)$$
$$R_i = \alpha + f_1(\text{Temperature}_i) + f_2(\text{Grain size}_i) + \varepsilon_i \quad (4)$$

Comparing models 1 (the full model) and 2 gives an F-ratio of 0.5631 ($p = 0.691$) indicating that grain size is not important. Comparing models 1 and 3 (leaving out temperature) gives an F-ratio of 2.45 ($p = 0.064$) and finally leaving out exposure gives a ratio of 7.52 ($p = 0.002$). This indicates that grain size is not significant; there is a weak temperature effect and a strong exposure effect.

The alternative is to calculate an AIC for each model. The AIC for an additive model is defined by

$$\text{AIC} = -2\log(\text{Likelihood}) + 2df \tag{7.1}$$

where df takes the role of the total number of parameters in a regression model (Chapter 5). For this example, using all three explanatory variables gives an AIC of 264.01. The model with exposure and temperature has an AIC of 258.9, the model with exposure and grain size 267.43, and grain size and temperature 276.5. The lower the AIC, the better the model. Hence, we can drop grain size from the model, giving a new model with the form:

$$R_i = \alpha + \text{factor}(\text{Exposure}_i) + f_{1i}(\text{Temperature}) + \varepsilon_i$$

The same process can be applied to this new model to see whether exposure or temperature can be dropped. As well as applying the selection procedure on the explanatory variables, it should also be applied on the amount of smoothing for each explanatory variable. This means that the model selection procedure in additive modelling not only contains a selection of the optimal explanatory variables but also the optimal degrees of freedom per variables.

The back-fitting algorithm for LOESS

We haven't discussed yet how to obtain numerical estimated of intercept, regression parameters and smoothers. For splines, this is rather technical and involves integrals and second-order derivates, and the interested reader is referred to Wood (2006) for technical details, but be prepared for some complicated mathematics. For LOESS it is all much easier, and the principle is sketched below. We start with the additive model in which only one smoother $f()$ is used.

For the additive model we have to estimate the intercept α, the smoothing function $f()$ and the variance σ^2. In linear regression, ordinary least squares can be used to give simple formulae that estimate both the intercept and the slope. In additive modelling, we need to use a so-called back-fitting algorithm to estimate the intercept α and the smoothing curve $f()$. The algorithm follows these basic steps:

- Estimate the intercept α for a given smoothing function $f()$.
- Estimate the smoothing function $f()$ for a given intercept α.

These two steps are applied until convergence is reached, and in more detail look like this:

1. Estimate the intercept as the mean value of observed values Y_i, $i = 1,..,n$.
2. Subtract the intercept from the observed values: $Y_i - \hat{\alpha}$, and estimate the smoothing curve $f()$, using any of the smoothing methods discussed above.

3. To make the smoother unique, centre it around zero (calculate the overall mean of the smoother and subtract it from each value of the smoother).
4. Estimate the intercept as the mean value of $Y_i - \hat{f}(X_i)$. The mean is taken over all observations *i*.
5. Repeat steps 2 to 4 until convergence.

By convention a hat is used to indicate an estimator of f and α. Although this is an extremely *ad hoc* mathematical procedure, in practice it works well and convergence is nearly always obtained in a few cycles. If the additive model contains more than one smoother, a generalisation of the back-fitting algorithm is used to identify the different smoothers and the intercept. It has the following form:

1. Obtain initial estimates for all components in the model by using for example random numbers.
2. Estimate the intercept as the mean value of the observed values Y_i, $i = 1,..,n$.
3. Estimate $f_1()$ by smoothing on $Y_i - \hat{\alpha} - \hat{f}_2(X_{i2}) - ... - \hat{f}_p(X_{ip})$.
4. Estimate $f_2()$ by smoothing on $Y_i - \hat{\alpha} - \hat{f}_1(X_{i1}) - \hat{f}_3(X_{i3}) - ... - \hat{f}_p(X_{ip})$.
5. ....repeated until....
6. Estimate $f_p()$ by smoothing on $Y_i - \hat{\alpha} - \hat{f}_1(X_{i1}) - ... - \hat{f}_{p-1}(X_{ip-1})$.
7. Repeat steps 1-5 until nothing changes anymore (convergence).

In each step, the smoothing functions are made unique by mean deletion (centring around zero).

7.6 Choosing the amount of smoothing

Instead of specifically choosing the span width in LOESS smoothing, or the value for the penalty term *h* in smoothing splines, we normally only choose a degree of freedom for a smoothing term. This defines the amount of smoothing required, and the software then chooses the required span width or penalty term (h) used for the smoothing spline. The smoothers are calibrated so that a smoother with one degree of freedom gives an approximate straight line. The default value in programmes like SPlus and R is for four degrees of freedom, which approximately coincides with the smoothing of a third-order polynomial. But is the default always the best choice? This can be checked by:

1. Using trial and error and looking at the graphs.
2. Looking at residuals and changing the degrees of freedom to see whether there are any residual patterns.
3. Using a modified version of the AIC, see equation (7.1).
4. Using cross-validation to estimate the amount of smoothing automatically.

We discuss each of these points next. The lattice graph shown in Figure 7.8 shows a range of smoothing curves for the parrotfish data. The lower left panel is the smoothing curve for the (only) explanatory variable coral richness using one

degree of freedom. The remaining panels show the smoothing curves with increasing degrees of freedom up to six. The shapes of the curves show there is little improvement after three or four degrees of freedom, and in this example, four degrees of freedom is the best choice. We can also look at the plot of residuals and if there is any discernable pattern choose to increase the degrees of freedom to allow for more variation in the smoother. This allows for stronger non-linear effects. When there are multiple explanatory variables, a plot of residuals versus each explanatory variable should be made. If any of these graphs shows a pattern, then the degrees of freedom of the smoother should be increased.

Earlier, we discussed using the AIC for selecting the best additive model and it can also be used to find the model with the optimal degrees of freedom. We can also use the F test to compare two nested models. A large F-ratio is evidence that the omitted explanatory variable is important. Note that a model with lower degrees of freedom can be seen as a nested model of a model with higher degrees of freedom. It is this principle that allows an additive model to be compared with a linear regression model (Fox 2000), because the linear regression can be seen as a nested model of the smoothing model.

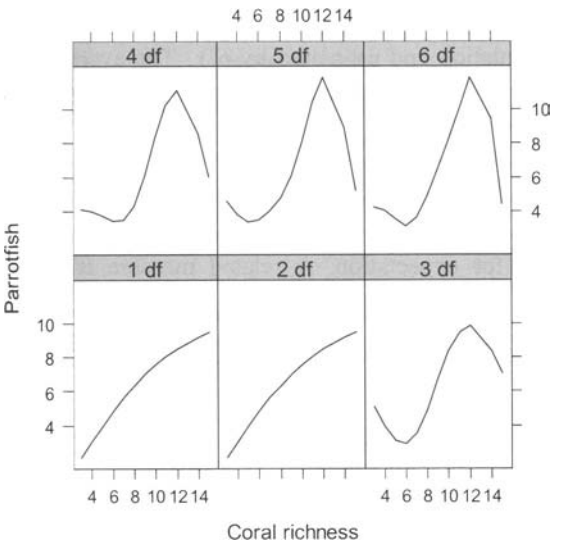

Figure 7.8. Smoothing curves for the explanatory variable coral richness in the Bahamas example. Each panel shows a smoothing curve using different degrees of freedom.

Staying with degrees of freedom, we now discuss cross-validation, which is a tool that automatically estimates the degrees of freedom for each smoother. In cross-validation, we leave out observation i, estimate a smoothing curve using the n - 1 remaining observations, predict the value at the omitted point X_i, and com-

pare the predicted value at X_i with the real value Y_i. The difference between the original and predicted value at X_i is given by:

$$Y_i - \hat{f}_\lambda^{-i}(X_i)$$

The notation "–i " in the smoother is used to indicate that observation i was omitted. The parameter λ refers to the amount of smoothing. This process is repeated for each observation. Adding up all squared residuals gives the cross-validation error:

$$CV(\lambda) = \frac{1}{n}\sum_{1=1}^{n}(Y_i - \hat{f}_\lambda^{-i}(X_i))^2$$

The cross-validation error is calculated for various values of the smoothing parameter λ. The value of λ that gives the lowest value for CV is considered the closest to the optimal amount of smoothing. The generalised cross-validation (GCV) is a modified version of the CV. To justify the use of the cross-validation, we introduce two quantities called MSE(λ) and PSE(λ). Remember, the linear regression model $Y = \alpha + \beta X + \varepsilon$ is a model for the entire population, and the same holds for the additive model $Y = \alpha + f(X) + \varepsilon$. The smoothing function $f()$ is also for the entire population and estimated by $\hat{f}()$. The average mean squared error (MSE) measures the difference between the real smoother and the estimated smoother, and is defined by

$$\text{MSE}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E[f(X_i) - \hat{f}_\lambda(X_i)]^2$$

where $E[]$ stands for expectation. A related measure is the average predicted squared error (PSE):

$$\text{PSE}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E[Y_i^* - \hat{f}_\lambda(X_i)]^2$$

Y_i^* is a predicted value at X_i. It can be shown that PSE = MSE + σ^2. The theoretical justification for using cross-validation is that E[CV(λ)] is approximately equal to PSE (Hastie and Tibshirani 1990).

To illustrate cross-validation, the following model was applied on the Bahamas parrotfish dataset:

$$\text{parrotfish}_i = \alpha + f(\text{coral richness}_i) + \varepsilon_i \quad \text{and} \quad \varepsilon_i \sim \text{N}(0, \sigma^2)$$

For the smoothing function, we first used two degrees of freedom and applied the cross-validation approach, and calculated CV(2). Then we used three degrees of freedom and obtained CV(3), continuing up to CV(10). A plot of CV versus degrees of freedom is given in Figure 7.9 and suggests that about four degrees of freedom is optimal. The R library mgcv allows for automatic application of the cross-validation method and gives a value of 4.37 degrees of freedom for coral richness.

You may expect that increasing the degrees of freedom always gives lower GCV values, but this is not the case as a high λ will ensure that observed values are fitted well, but it does not mean that omitted points are predicted well (Wood 2006).

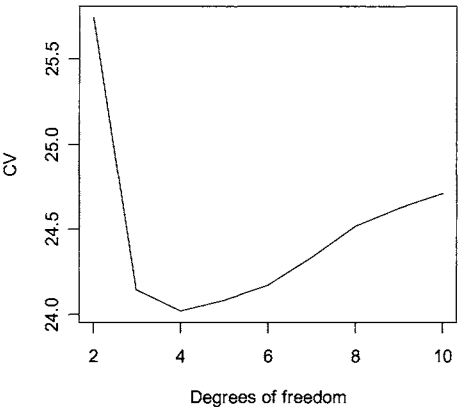

Figure 7.9. A plot of CV versus degrees of freedom for the Bahamas fisheries dataset.

7.7 Model selection and validation

This section works through the model selection and validation process using the squid data from Chapter 4. Recall from Section 4.1 that these data contain measurements for squid from various locations, months and years in Scottish waters. GSI is the Gonadosomatic index and is a standard index used by biologists.

Model selection

A data exploration using Cleveland dotplots and boxplots showed that there are no extreme observations. A coplot showed that month, year, sex and location effects might be expected. GSI is the response variable, with month, year, location and sex considered as explanatory variables. Of the explanatory variables only month has a reasonable number of different values (12). All other explanatory variables have less than five unique values, including year. In such situations, moving the box along the x-axis (Figure 7.2) will cause an error because there are not enough points in the box around a particular target value. If month is considered as a nominal variable, it will require 11 parameters because there are 12 months. It cannot be modelled as a parametric component because this would rank the dummy code 'twelve' for December higher than the 'one' for January, which would not make sense. An option is to model month as a smoother, and this re-

duces the number of parameters from 11 to 4, assuming the default amount of smoothing is used. The model we applied is in the form:

$$GSI_i = \alpha + f(\text{month}_i) + \text{year}_i + \text{location}_i + \text{sex}_i + \varepsilon_i$$

Year, location and sex are modelled as nominal variables, and there are 2644 observations in the dataset. The function $f()$ was estimated with a smoothing spline and in the first instance, we used four degrees of freedom. The fitted smoothing function for month is presented in the Figure 7.10-A, and it shows a clear seasonal pattern. In panel B the smoothing curve obtained by cross-validation is shown, and is considerably less smooth. The estimated degrees of freedom is approximately nine, which is two degrees of freedom less compared with using month as a nominal variable.

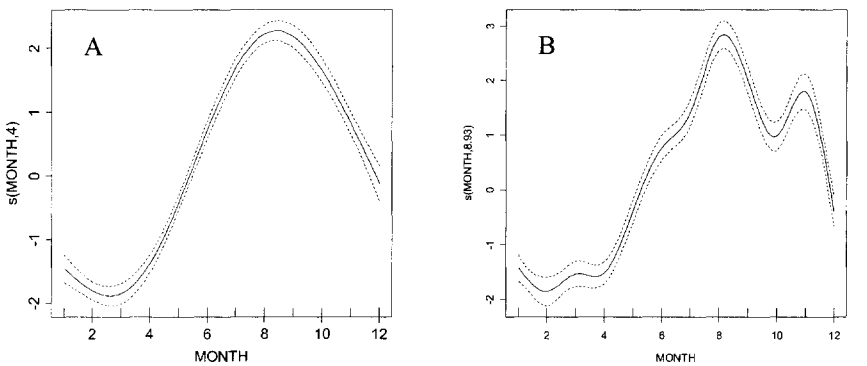

Figure 7.10. Smoothing curve for month for the squid data with 4 degrees of freedom (A) and 8.9 degrees of freedom (B).

Using the F-test to compare a full model with 8.9 degrees of freedom and a nested model with 4, 5, 6, 7 or 8 degrees of freedom confirms that 8.9 is indeed optimal. The numerical output of the model, with 8.9 degrees of freedom, is given below.

	Estimate	std. err.	t-ratio	p-value
Intercept	1.37	0.11	12.47	<0.001
factor(YEAR)2	0.14	0.12	1.22	0.22
factor(YEAR)3	−0.23	0.12	−1.88	0.05
factor(YEAR)4	−0.76	0.17	−4.27	<0.001
factor(Location)2	−0.33	0.18	−1.80	0.07
factor(Location)3	0.11	0.11	0.94	0.34
factor(Location)4	2.00	0.31	6.42	<0.001
factor(Sex)2	1.90	0.07	24.63	<0.001

For the smoothers we have the following numerical output

Approximate significance of smooth terms:

	edf	*F*-statistic	*p*-value
s(MONTH)	8.92	127.4	<0.001

R-sq.(adj) = 0.44. Deviance explained = 45%. GCV score = 3.93. Scale est. = 3.91. n = 2644. Dispersion parameter = 3.91. Deviance = 10275.3. df.residual (residual degrees of freedom) = 2627.07. df (n-df.residual) = 16.93. AIC according to formula: -2log(Likelihood) + 2df = 11128.29.

The explained deviance is 45%, which means that 45% of the total sum of squares is explained by the model. The estimated variance is $\sigma^2 = 3.91$ and the AIC is 11128.29. The smoothing term is significant at the 5% level and the *p*-values of the individual levels indicate the effects of location, year and sex are significant. The *F*-test can be used to obtain one overall *p*-value per nominal variable. Comparing the full model with a model without year gives an *F*-ratio of 14.34 ($p < 0.001$). Leaving out location gives $F = 16.79$ ($p < 0.001$). Both terms are highly significant in explaining GSI.

Model validation

As in linear regression, with additive models we need to verify the underlying assumptions of homogeneity and normality, and check for potential influential observations. Only if the model validation indicates that there are no problems, can we accept the smoothing curve in Figure 7.10-B, and the numerical output above. Figures 7.11 and 7.12 give a series of graphs, which can be used for model validation. Figure 7.11-A shows the fitted values against the observed values. Ideally, the points in Figure 7.11-A should lie on a straight line, but as the model only explains 45% of the variation in the data some discrepancies can be expected. Panel B illustrates the fitted values against the residuals, and these show a worrying violation of homogeneity. The concentration of points at the lower part of this panel are probably the samples with zero or low values. Panels C and D check on normality and suggest a lack of normality in our data. This violation of homogeneity and normality is enough justification to show the model is unsatisfactory. However, for the sake of this example, we will continue working through the model validation process. Panel E shows the leverage value for each sample, and this identifies a group of samples with considerably higher leverage than the rest. As none of the explanatory variables lend themselves to a transformation (there is no point in transforming a nominal variable and all months are sampled), there is little we can do about this particular problem. It might be caused by a set of samples only measured in one month in one location in one year. If this is indeed the case, it could be argued that they should be left out of the analysis, but a detailed data exploration using coplots and scatterplots is required before making this decision.

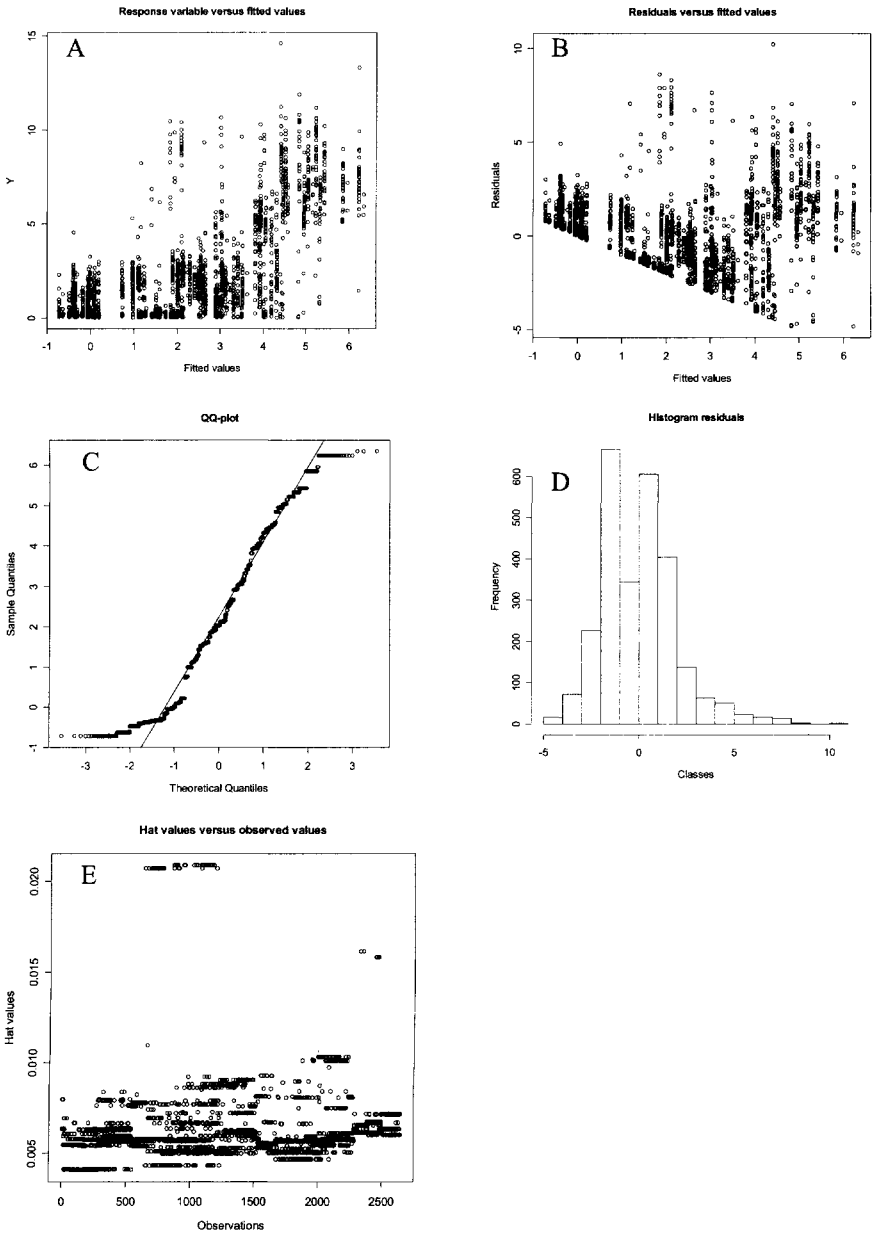

Figure 7.11. A: Observed values versus fitted values. B: Residuals versus fitted values. C: QQ-plot of residuals. D: Histogram of residuals. E: Leverage for each sample.

To detect a model misspecification, the residuals of the model can be plotted against the original explanatory variables (Figure 7.12). With nominal explanatory variables, this will be a boxplot, and for continuous variables, it will be a scatterplot. Under no circumstances should any of these plots show a pattern. One of the sex levels (Female) has considerably higher residuals (Figure 7.12A). The same holds for location one (Figure 7.12C). The scatterplot of residuals versus month shows that in months 2-6, the spread is smaller (Figure 7.12D). We cannot detect any residual-year relationship (Figure 7.12B). Although these graphs are trivial to make, the conclusions are critical; and they clearly show that the problem in the model fit is due to the female data in location 1 and that the heterogeneity is caused by a month (or more accurately: seasonal) effect. To show the importance of the data exploration, re-visit the coplot for these data in Figure 4.21. Most of the activity seems to occur in location one for females, and it might be advisable to analyse the female data from location one separately. An alternative approach is to use a *generalised* additive model using the Poisson distribution and log link function, and we look at this later.

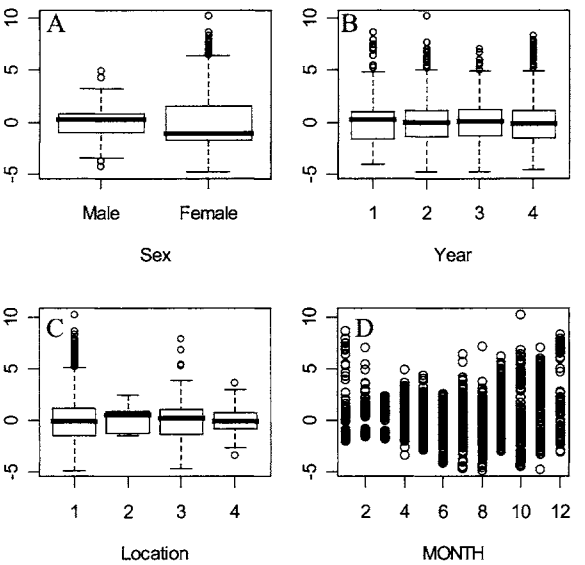

Figure 7.12. Residuals versus each explanatory variable. A: Residuals versus sex. B: Residuals versus year. C: Residuals versus location. D: Residuals versus month.

Concluding remarks on additive modelling

Although it lacks a theoretical justification, additive modelling is a useful data analysis tool. This is because it visualises the relationship between a response variable and multiple explanatory variables. And this 'let the data speak' approach

can be used to obtain important insights into how to proceed with a parametric analysis. For example, the shape of the smoothing curve for temperature in the additive model for the RIKZ data suggests continuing with a simple (polynomial) regression for temperature, as it is capable of capturing the same pattern.

The smoothing curves in additive models are obtained rather arbitrarily, and the hypothesis tests rely on approximations and follow regression methods without any formal justification. We are not aware of any additive modelling simulation study in which the effects of non-homogeneity, non-normality, non-fixed X, and dependency have been studied. On the other hand, ecological data tend to be very noisy, and (generalised) additive modelling may be the only tool available that can give useful results. As to the criticism on the lack of underlying theory and foundation for the use of hypothesis tests (and p-values), one can always apply bootstrapping to get more reliable confidence bands. See Davison and Hinkley (1997) or Clarke et al. (2003) for a discussion and examples of bootstrapping for smoothing methods.

As discussed above, additive modelling has the same problems with heterogeneity, negative fitted values and negative realisation as linear regression, and it cannot be used to analyse 0–1 data. To deal with this, generalised additive modelling (GAM) can be used, and this can be considered as an extension to additive modelling, just as GLM can be considered as an extension of linear regression.

7.8 Generalised additive modelling

In Section 7.1, additive modelling was introduced and we used additive modelling to investigate the relationship between species richness and temperature and exposure for the RIKZ data. The final additive model was of the form:

$$R_i \sim N(\mu_i, \sigma^2), \text{ where } E[R_i] = \mu_i = g(x_i) = \alpha + \text{Exposure}_i + f(\text{Temperature}_i)$$

This additive model is a convenient tool to obtain some insights into the relationships between the response variable and the explanatory variables. For example, it might show that temperature has a non-linear relationship with species richness. However, nothing stops the model from obtaining negative fitted values, and the Gaussian density curves on top of the fitted values suggest that realisations with negative values are equally possible. To prevent this, and in the same way as linear regression was extended to Poisson regression, we can extend the additive model to a Poisson additive model. Better named: a generalised additive model with log link function. The mathematical formula for a GAM model of the RIKZ data (using richness as the explanatory variable, exposure as a nominal explanatory variable and temperature as a smoothing function) is given by:

$$R_i \sim P(\mu_i) \quad \text{and} \quad E[R_i] = \text{Var}(R_i) = \mu_i$$

$$\text{Log}(\mu_i) = g(x_i), \quad \text{where} \quad g(x_i) = \alpha + \text{Exposure}_i + f(\text{Temperature}_i)$$

The smoothing curve for temperature is presented in Figure 7.13, and the numerical output is given by

	Estimate	std. err.	t-ratio	p-value
(Intercept)	3.18	0.35	9.16	<0.001
factor(exposure)10	−1.25	0.41	−3.08	<0.001
factor(exposure)11	−2.36	0.45	−5.21	<0.001

All parameters are significantly different from zero at the 5% level, including the smoother ($p < 0.001$). Just as in GLM, there may be the possibility of overdispersion (Chapter 6). Because overdispersion was suspected (based on the results of GLM), the quasi-Poisson GAM was applied. The overdispersion parameter was 2.6, and therefore, all standard errors were corrected with the square root of 2.6. After this correction, all parameters were still significantly differently from 0 at the 5% level.

The general GAM model using the Poisson distribution and the log link function is similar to the GLM Poisson model, except that the predictor function g(x) is given by

$$g(x_i) = \alpha + f_1(x_{i1}) + \ldots + f_p(x_{ip}).$$

The f_js are smoothing functions. It is also possible to have smoothing functions with parametric components, leading to semi-parametric models. In principle, GAM with a Poisson distribution follows the same requirements as GLM. We need to take into account overdispersion, and selecting the best model can be done by comparing deviances of nested models or by using the AIC. Just as additive modelling is mimicking linear regression, generalised additive modelling is mimicking generalised linear modelling. So, the tests for comparing models in GLM are also used in GAM: a Chi-square test if there is no overdispersion and an F-test if there is overdispersion.

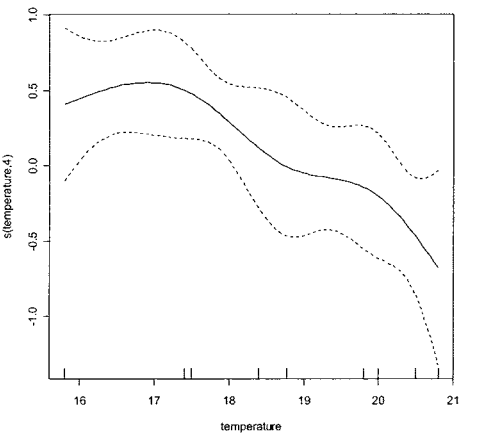

Figure 7.13. GAM for RIKZ data; the log link and Poisson distribution was used. The horizontal axis shows the temperature gradient and the vertical axis the contribution of the temperature smoother to the fitted values.

GAM for the squid data

The additive model applied on the squid data indicated violation of homogeneity, and therefore, a GAM model, using the Poisson distribution and log link function, was applied. Results indicated there was minor overdispersion (1.6), and a quasi-Poisson model was chosen. The smoothing curve is shown in Figure 7.14. Cross-validation was used to estimate the optimal degrees of freedom resulting in 8.8 df. All nominal variables were significantly differently from 0 at the 5% level and the model explained 50% of the deviance. As part of the model validation, deviance residuals were plotted against the explanatory variables (Figure 7.15). These residual indicate which samples contribute most to the deviance. As explained in the GLM section, the smaller the deviance the better. Observations from females in location one, in month one, have a relatively high contribution to the deviance, indicating a pattern in the residuals. This might be a reason not to accept the model as the 'best' model.

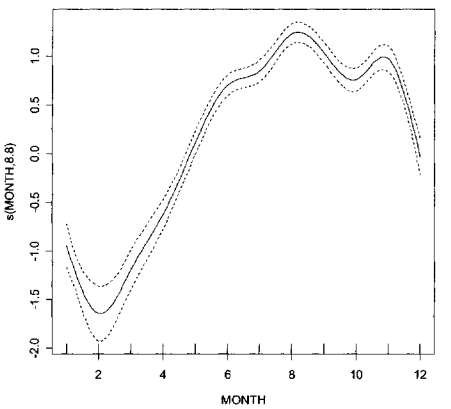

Figure 7.14. Smoothing curve for month in the GAM model with Poisson distribution and log-link.

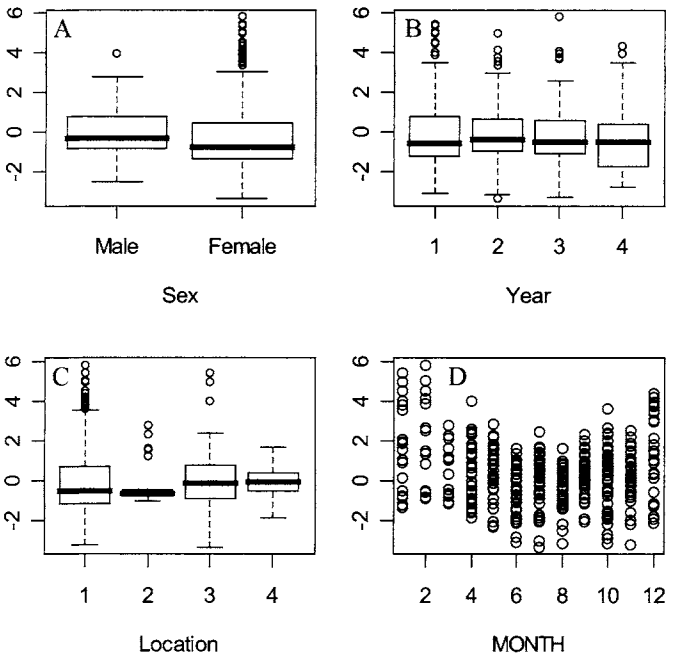

Figure 7.15. Deviance residuals versus explanatory variables. A: Residuals versus sex. B: Residuals versus year. C: Residuals versus location. D: Residuals versus month.

Presence–absence data

If the data are presence–absence data, the smoothing equivalent of a logistic regression model can be used. In this case there is only one observation per X value, and the formula becomes:

$$Y_i \sim \text{Bernoulli}(1, P_i) \quad \text{and} \quad \log\frac{P_i}{1-P_i} = \alpha + f_1(\mathrm{X}_{1i}) + \ldots + f_p(\mathrm{X}_{pi})$$

In case of data on proportions (multiple observations at the same X value), a binomial distribution should be used; see also the GLM section. Various examples of GAM applied to presence absence data and data on proportions are given in the case study chapters. It is also possible to have GAM models with interactions, and these are also discussed in the case study chapters.

7.9 Where to go from here

The additive modelling for the squid data showed that there was a different residual spread for females and for location 1. However, in the analysis we completely ignored the possible auto-correlation structure in the data. In the next chapter, we discuss using linear mixed modelling and additive mixed modelling that can deal with this problem.

Both modelling approaches allow for auto-correlation and multiple variances to be used (e.g., one for the male and one for the female data). The case study chapters contain several examples where we also show how to include interactions between smoothers and nominal explanatory variables.

These extensions can also be applied in GLM and GAM models and are called generalised linear mixed modelling (GLMM) and generalised additive mixed modelling (GAMM). These methods are not discussed in this book, but a good book for GLMM is Fitzmaurice et al. (2004), even though this has mainly medical examples, and for GAMM, Ruppert et al. (2003) or Wood (2006) are one of the few books available on this recently developed topic.

In the case studies we also present examples with interactions between smoothers and nominal variables, add auto-correlation, spatial correlation and model the heterogeneity. It is also possible to add random effects (Chapter 8).