Statistical Modeling with Spline Functions
Methodology and Theory

Mark H. Hansen
University of California at Los Angeles

Jianhua Z. Huang
University of Pennsylvania

Charles Kooperberg
Fred Hutchinson Cancer Research Center

Charles J. Stone
University of California at Berkeley

Young K. Truong
University of North Carolina at Chapel Hill

January 5, 2006

3
Linear Models

In this chapter, we consider linear regression models that use ordinary least squares as a fitting criterion. The simplicity of this estimation problem allows us to focus on issues unique to spline modeling. We first treat the case of a single, univariate predictor, a problem that is often referred to as *curve estimation* or *scatterplot smoothing.* We introduce splines as natural extensions of ordinary polynomial regression. Our approach is essentially data analytic, illustrating features that make splines ideal for statistical applications. We examine the smoothing parameters implicit in these models, exploring in particular how the positioning of breakpoints or knots affects the overall smoothness of a spline fit. Heuristics involving local bias-variance considerations make these ideas concrete.

Early applications of splines typically relied on prior information or subject knowledge to place breakpoints. In practice, we rarely receive such guidance, and hence recent work has focused on designing automatic knot placement algorithms. Familiar techniques from variable selection have inspired many such schemes, often by simply treating a number of candidate spline basis functions as potential predictors in a simple linear model. The connection between spline smoothing and (parametric) data analytic tools for model building will appear throughout the text as we incorporate flexible elements into more elaborate estimation schemes. A clear advantage of this programme is that we can apply our intuition concerning regression diagnostics directly to curve estimation. This connection breaks down somewhat, however, over questions of inference for spline models, a topic we will address in the context of a particular example.

In this chapter, we will also examine problems with more than one predictor. In traditional applications of multivariate regression, an analysis of variance (ANOVA) decomposition can be an important tool for assessing the dependence between (subsets of) input variables and the response. Borrowing this construction, we introduce the concepts of "main effects" and "interactions" for multivariate spline models, where individual ANOVA components are now smooth functions of the predictors. Tensor products of univariate spline spaces provide a formal way to describe the resulting decomposition. We next present relatively straightforward extensions of the adaptation procedure developed for curves. The presence or absence of ANOVA components is determined adaptively subject to familiar constraints on the order in which main effects and interactions can be added or removed from a model. As we will see, the power of such schemes comes from their ability to identify important structure in large, high-dimensional problems, and to express often complicated dependencies in an interpretable way. The ease with which polynomial splines carry us from curve estimation to multivariate problems is somewhat surprising, making this approach relatively unique among competing techniques for nonparametric regression.

3.1 Examples

We begin with two applications of spline-based methods, one for simple curve estimation and one involving a high-dimensional regression function. We briefly present the data and describe the fits, and then we use the analysis to motivate a general methodology for modeling with splines.

3.1.1 Smoothing and extrinsic catastrophists

The data set under study was collected to test several hypotheses about the catastrophic events that occurred approximately 65 million years ago. This point marks a division between two geologic time periods, the Cretaceous (from 66.4 to 144 million years ago) and the Tertiary (spanning from about 1.6 to 66.4 million years ago). Earth scientists believe that the boundary between these periods is distinguished by tremendous changes in climate that accompanied a mass extinction of over half of the species inhabiting the planet at the time. Recently, the composition of stronium (Sr) isotopes in sea water has been used to evaluate several hypotheses about the cause of these extreme events. The quantity plotted in Figure 3.1 is related to the isotopic makeup of Sr measured for the shells of marine organisms (foraminifera). The vertical line in the middle of this figure denotes the Cretaceous-Tertiary boundary, referred to as the KTB. The apparent peak exhibited by these data at the KTB has been used by researchers to argue

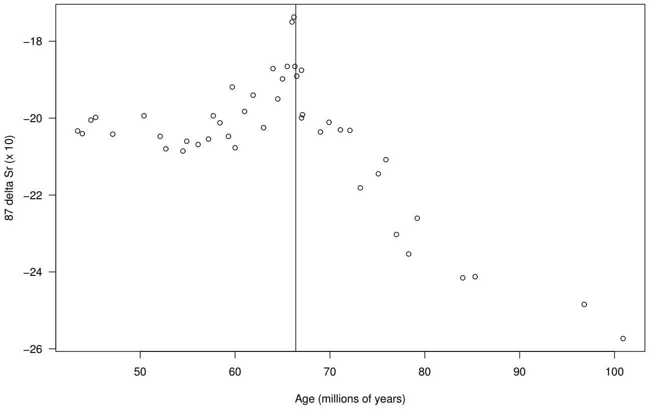

FIGURE 3.1. Tracking the standardized ratio, ${}^{87}\delta\,\mathrm{Sr}$, of strontium isotopes ${}^{87}\mathrm{Sr}$ to ${}^{86}\mathrm{Sr}$ present in shells of marine organisms. Data are taken from 4 sites listed in Table 2 of Hess et al (1986).

that the disastrous climate changes could not have resulted solely from increased volcanic activity or the general drop in sea level that occurred at the time. Instead, this peak is said to support the theory that one or more meteors collided with the earth, causing short term acid rain, the emission of poisonous gases, and an overall cooling. Scientists favoring this hypothesis have been labeled "extrinsic catastrophists."

The data in Figure 3.1 represent a (standardized) ratio of strontium-87 isotopes (${}^{87}\mathrm{Sr}$) to strontium-86 isotopes (${}^{86}\mathrm{Sr}$) contained in the shells of foraminifera fossils taken from cores collected by the Deep Sea Drilling project (Hess et al., 1986). Let ${}^{87}\mathrm{Sr}/{}^{86}\mathrm{Sr}_{\text{ sample}}$ denote the isotopic Sr ratio for a given sample. The Sr composition of each selected shell records the ratio of ${}^{87}\mathrm{Sr}$ to ${}^{86}\mathrm{Sr}$ in the oceans at the time the shell was formed. Also, because the time Sr remains in sea water is much longer than the mixing time of the oceans, it is believed that this ratio does not depend on the location where the fossils were found. Finally, for each sample a standardized ratio was computed via

$$
{}^{87}\delta\,\mathrm{Sr} = \left(\frac{{}^{87}\mathrm{Sr}/{}^{86}\mathrm{Sr}_{\text{ sample}}}{{}^{87}\mathrm{Sr}/{}^{86}\mathrm{Sr}_{\text{ sea water}}} - 1 \right) \times 10^5
$$

where we let ${}^{87}\mathrm{Sr}/{}^{86}\mathrm{Sr}_{\text{ sea water}}$ denote the isotopic concentration of Sr for modern sea water. Earth scientists expect that ${}^{87}\delta\,\mathrm{Sr}$ is a smoothly-varying function of time, depending on "gradual" effects like the rate of runoff from rivers and the introduction of Sr from deep-sea ridges. The scatter in Figure 3.1 is thought to be primarily measurement error. Martin and

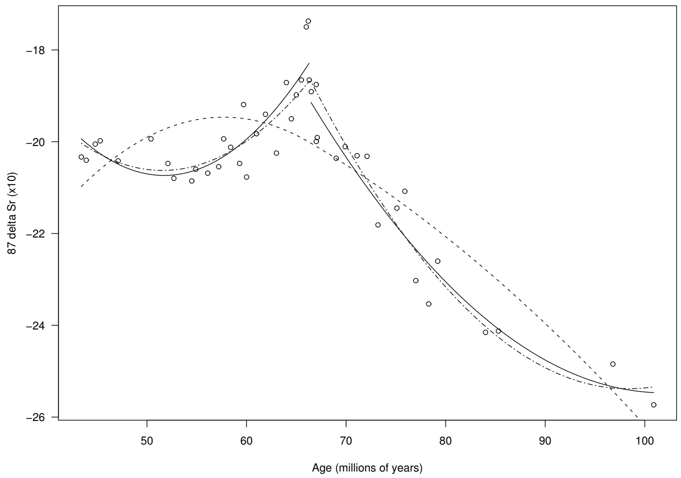

FIGURE 3.2. Several different fits to the isotope data. The solid line is discontinuous and has 6 degrees of freedom. The dashed curve is continuously differentiable and involves 4 degrees of freedom.

Macdougall (1991) outline how $^{87}\delta\,\mathrm{Sr}$ values are obtained from foraminifera fossils and detail several sources of measurement error.

Hess et al. (1986) present a complete collection of $^{87}\delta\,\mathrm{Sr}$ values that extend from about 100 million years ago to the present. Following Hallam and Wignall (1997) and Macdougall (1988), however, we focus on strontium concentrations around the transition between the Cretaceous and Tertiary periods. To obtain coverage similar to these other studies, we selected measurements from the sites in Hess et al. (1986) labeled 366, 305, 356 and 577. In Figure 3.1, we have 45 points ranging in age from 43.4 to 100.9 million years ago.

In Figure 3.2, we present three different fits to the isotope data. The curves are each *splines*, piecewise polynomials that satisfy certain smoothness constraints. For example, the solid curve is discontinuous, with a break positioned at the KTB, while the dashed curve is continuously differentiable. The models exhibit differ in their ability to capture the features evident in the $^{87}\delta\,\mathrm{Sr}$ data. The dashed curve seems to smooth over the peak at the KTB and in the process produce a rather poor fit everywhere. In addition, one can question whether a jump occurs at the KTB, as exhibited by the solid line, or if the simple bend in the dot-dashed line is sufficient. In terms of degrees of freedom, the solid line represents two separate quadratic polynomials, one fit to data on either side of the KTB, and hence has 6 parameters. By forcing the quadratics to be continuously differentiable across the KTB, we impose two “smoothness constraints” on the fit (one to ensure continuity and then one to ensure continuity of the

Model	DF	F	P-value
Quadratic polynomial	3	159	0.000
Discontinuous second derivative at KTB	1	10	0.002
Discontinuous first derivative at KTB	1	111	0.000
Discontinuous at KTB	1	7	0.011

TABLE 3.1. Judging smoothness across the KTB with a classical ANOVA table.

first derivative) and drop the number of free parameters to 4. The dashed curve is the result. In between these two in terms of smoothness is the dot-dashed curve which is continuous across the KTB, but has a discontinuity in its first derivative at that point.

Later in the chapter we will develop criteria to evaluate how much smoothness is required across a given breakpoint using simple t tests or F tests from classical regression analysis. As it turns out, each constraint corresponds to a single degree of freedom in the model and testing for its presence is equivalent to deciding if that parameter is different from zero. In Table 3.1, we present an ANOVA decomposition that separates an overall quadratic fit from various elaborations that reduce the smoothness of the estimated $^{87}\delta\,\mathrm{Sr}$ curve across the KTB. We computed the F-statistics using a hierarchy of conditions that relaxes the continuity in higher derivatives first. This kind of ordering turns out to be quite natural from a regression analysis perspective as well, a fact we will return to in Section 3.3. Note that all of the breaks seem highly significant, providing some evidence for a large drop in $^{87}\delta\,\mathrm{Sr}$ across the KTB.

While inference for the $^{87}\delta\,\mathrm{Sr}$ data naturally focuses on the behavior of the curve at the KTB, in many smoothing situations we do not know what sort of structures might be present in the data. Instead, we would want some way of automatically identifying prominent features. In this chapter, we will discuss methods that adapt to peaks and valleys by introducing breakpoints like those at the KTB in Figure 3.2. These techniques use regression diagnostics similar to the ANOVA decomposition in Table 3.1 to identify locations that require more or less flexibility. As an example, we applied one such scheme to the $^{87}\delta\,\mathrm{Sr}$ data; the "best fitting" curve is plotted with a solid line in Figure 3.3. Given several different fits, we judged the "best" using a common (regression) model selection criterion, an approach we will outline in detail in Section 3.3.3. The curve itself is a so-called *natural cubic spline*. For the moment, all that we need to know is that this curve represents a piecewise cubic polynomial which is twice continuously differentiable, and that it has jumps in its third derivative at the three points indicated by arrows in the plot. The fact that breakpoints cluster in the region of the KTB indicates the need for extra flexibility there. The grey bands in this plot represent (asymmetric) pointwise 95% confidence bands obtained by a simple bootstrap procedure suggested by

Friedman and Silverman (1989) and discussed in Section 3.8.1. A total of 200 bootstrap samples were used to create this plot.

In the lower panel of Figure 3.3, we present a closeup of KTB together with a boxplot of the locations of the peaks (ages at which the peaks occurred) in the 200 bootstrap samples. The run of four positive residuals to the left of the KTB provides us with only modest evidence for a stronger peak; but from the boxplot we do see a fair amount of uncertainty in the placement of the maximum. (The variability of a spline fit has been characterized considering runs of errors of a common sign; see Section 3.6.1 for more details.) This kind of observation might lead us to explore further, possibly with the analysis in Figure 3.2. Granted we have only 45 data points, the peak is identified mainly by the two largest points just to the left of the KTB. In Section 3.8.2, we will introduce an auxiliary data set collected by Martin and Macdougall (1991) to better resolve the peak.

Clearly, modeling with splines provides us with considerable flexibility in constructing an estimate of the $^{87}\delta\,\mathrm{Sr}$ values. How we decide on an appropriate model is the subject of this chapter. For the moment, it is only important to recognize that several different kinds of spline fits can be constructed, and we can evaluate each on the basis of its ability to describe the underlying data. Typically, our choice of spline model involves some implicit assumptions about the smoothness of the function we are estimating. In the later sections of this chapter, we will translate these assumptions into explicit constraints on a general methodology. As we will see, one strength of the spline approach is that these constraints can be cast in terms of familiar stepwise approaches to building regression models.

Martin and Macdougall (1991), McArthur, Thirlwall, Engkilde, Zinsmeister and Howarth (1998) and Howarth and McArthur (1997) consider the geological implications of the slopes just to the left and the right of the KTB. The approach from the right indicates an introduction of strontium isotopes into the planet's oceans. Martin and Macdougall (1991) argue that the only explanation for an extremely sharp rise in strontium isotopes is a meteor impact that sets off hundreds of years of acid rain.[1] The narrowness of the peak relative to other features found in the $^{87}\delta\,\mathrm{Sr}$ curve at other periods (from the present, reaching as far back as 200 million years ago) could tell us something about the plausibility of this argument. Strontium isotopes dissipate at a fixed rate, and hence the rate of decline on the left of the KTB was argued by McArthur et al. (1998) and Howarth and McArthur (1997) to be too steep to be believable. Questions related to possible deficiencies in the underlying data are beyond the scope of this text and the reader is referred to McArthur et al. (1998).

[1] A meteor striking the earth could produce enough airborne soot to trigger long periods of acid rain, and the runoff or "continental weathering" would be responsible for the extra isotopes.

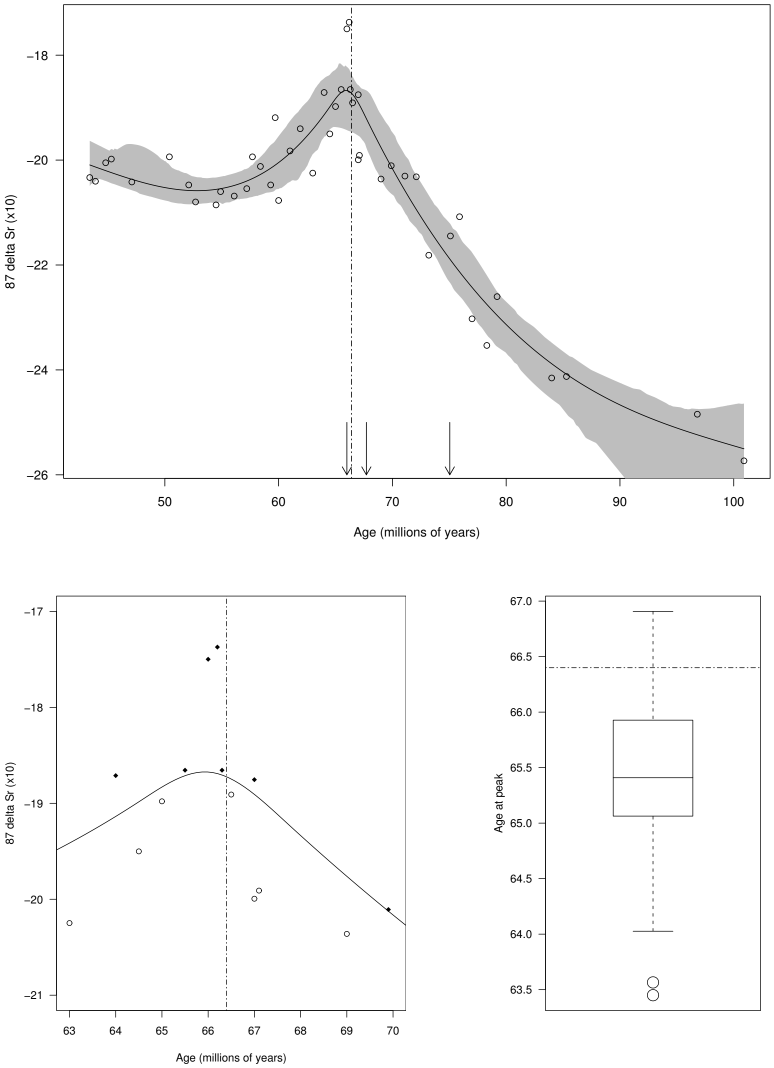

FIGURE 3.3. Upper: Modeling the standardized ratio $^{87}\delta\,$Sr with a (natural) cubic spline. Gray regions represent (asymmetric) 95% pointwise bootstrap confidence intervals. Lower: On the left, we plot the fit only in the neighborhood of the KTB. Black points indicate points that are larger than the fit (positive residuals) and open circles indicate points that are smaller than the fit (negative residuals). On the right we present a boxplot representing the distribution of times at which the peak occurred in our bootstrap samples.

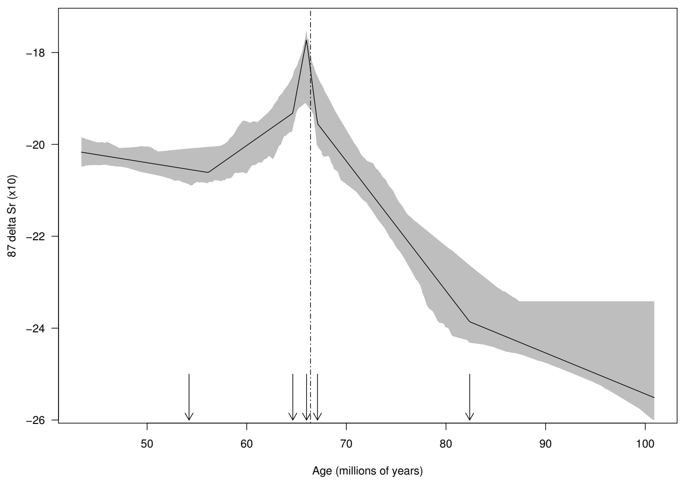

FIGURE 3.4. Piecewise linear spline fit to the ${}^{87}\delta\,\mathrm{Sr}$ data with pointwise (asymmetric) 95% confidence bands. Breakpoints are marked with arrows.

So far, we have seen examples of several different kinds of spline spaces. In each case, we are working with piecewise polynomials that satisfy certain smoothness constraints across a set of breakpoints distributed within the support of the data. In curve fitting, the kind of analysis we performed with the (natural) cubic splines is quite common as a general purpose methodology. When the data do not seem to support such smooth models, we might consider other kinds of splines. In Figure 3.4, we present an example of a *linear spline*, or a continuous, piecewise linear function. In this figure we see the tradeoffs between the two kinds of fits; while the linear spline might resolve a sharp peak more effectively, it often smoothes over more subtle structure like the bend just to the left of the KTB or the subtle drop to the right. The grey band represents pointwise (asymmetric) 95% confidence intervals for the regression function using the same bootstrap procedure as above. Notice that our fit is pinned near the edge of the confidence interval at the KTB. This could suggest that our actual piecewise linear model has "over-fit" the data in the sense that it is tracking the two large points to the left of the KTB too closely. The raggedness of the lower confidence band is one byproduct of the "stiffness" or simplicity of the piecewise linear fit. Still, even with these caveats, the overall shape of the curve has been captured faithfully. In the next subsection, we will see that for multivariate problems, such piecewise linear components are excellent building blocks.

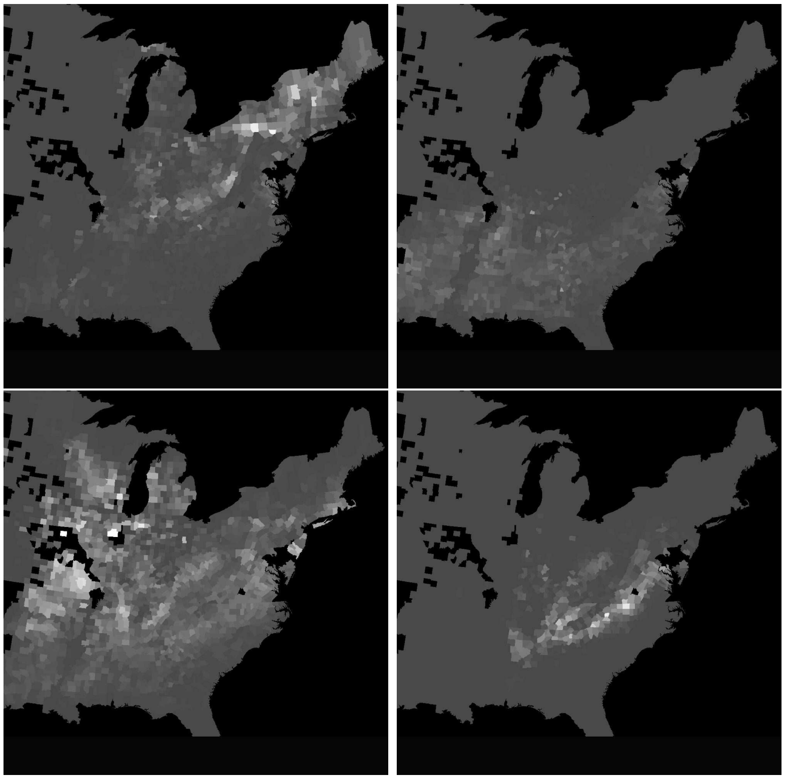

FIGURE 3.5. Abundance (IV) of four tree species. From left to right, top to bottom: American beech, southern red oak, white oak and Virginia pine. Lighter pixels indicate counties with greater IV values, black pixels indicate missing or absent data.

3.1.2 Global warming and tree migration

We now consider another, more contemporary (and avoidable?), environmental disaster, namely the buildup of so-called greenhouse gases and the accompanying trend of global warming. Researchers predict that by the end of the twenty-first century we could see a doubling of current CO_2 levels, triggering a temperature increase of between 1 and 4.5 degrees Celsius. What impact will this have on the living systems around us? In this section, we consider a large data set assembled by the United States Department of Agriculture (USDA) Forest Service to assess how such changes in climate will affect the distribution of various tree species in the eastern portion of the United States. The USDA study examined more than 2,100 eastern counties with data characterizing the local climate, soils, and elevation. The Forest Inventory Analysis (FIA) database provided figures on the range and abundance of various tree species.

Historical data suggests that climatic warming can shift the "optimal habitats" for various tree species and that, over time, such shifts can result in a kind of "migration" of trees toward the more favorable conditions. For most species, this movement has been northward and has occurred over thousands of years. By contrast, recent studies of global warming predict dramatic changes in climate that significantly alter the optimal habitats of many species in a relatively short period of time. To quantify the impact of such rapid changes, the USDA Forest Service first constructed a model describing the abundance of trees as a function of various climatic and soil factors. The USDA then substituted predictions for the climatic variables under different warming scenarios to study habitat shifts species-by-species. The implication that we might see actual tree migrations as a result of these changes is less obvious, again because of the short time periods involved in the warming scenarios. In addition, the open landscape of the modern US is much more fragmented (crossed with highways, homes and shopping malls, for example), complicating the migration of species in certain areas. Still, with these caveats in mind, it is useful to assess how the forests in the US might be impacted by projected environmental conditions.

As mentioned above, much of the data to support this study comes from the FIA. The FIA contains measurements on over 3 million trees, organized around 100,000 forested plots. For each plot researchers calculated an importance value (IV) for 8 different tree species,

$$\mathrm{IV}(s) = 100\frac{\mathrm{area}(s)}{\sum_x \mathrm{area}(x)} + 100\frac{\mathrm{stems}(s)}{\sum_x \mathrm{stems}(x)}\,, \qquad (3.1.1)$$

where area(s) is the total basal area of trees of species s in a given plot and stems(s) is a count of the number of trees of species s in a given plot. In a stand of trees consisting of essentially one species, the IV value will be near 200. Our data set consists of county-level averages of importance values together with the environmental and land use statistics, also summarized by county. The IV values for four species are plotted in Figure 3.5.

To predict IV for a given tree species, a series of variables was collected that describe the habitat of the forests in each of the 2,100 study counties. From an original list of 100 potential predictor variables, the researchers focused on 28. In Table 3.2 we present these covariates, listing the seven climate-related variables first. The remaining factors deal with the condition of the soil in each county as well as measures relating to the elevation. Unlike the curve estimation problem in the previous section, model building for the IV data means also having to identify *which* variables are important as well as the general shape of the dependence. This extra complexity is the first we will see of the so-called "curse of dimensionality."

We begin by considering just the IV values for the American beech (*Fagus grandifolia*) because Iverson and Prasad (1999, 1998) report success in constructing models based on the variables in Table 3.2. In Prasad and Iverson (2000), two kinds of spline-based smoothers were studied with four

Climate factors (monthly means, 1984–1987)	
AVGT	Temperature (C)
MAYSEPT	Mean May–September temperature (C)
JANT	Mean January temperature (C)
JULT	Mean July temperature (C)
PPT	Annual precipitation (mm)
PET	Potential evapotranspiration (mm/month)
JARPPET	July–August ratio of precipitation to PET

Soil factors: Habitat	
PH	Soil pH
TAWC	Total available water capacity (cm, to 152 cm)
OM	Organic matter content (% by weight)
PERM	Soil permeability rate (cm/hour)
BD	Soil bulk density (g/cm3)
CEC	Cation exchange capacity
ROCKDEP	Depth to bedrock (cm)
ERODFAC	Soil erodibility factor, rock fragments free (susceptibility of soil erosion to water movement)
SLOPE	Soil slope (percent) of a soil component
PROD	Potential soil productivity, (m3 of timber/ha)

Soil factors: Texture	
TEXFINE	Percent passing sieve No. 200 (fine)
TEXCOARS	Percent passing sieve No. 10 (coarse)
ROCKFRAG	Percent weight of rock fragments 8–25 cm
CLAY	Percent clay (< 0.002 mm size)

Soil factors: Soil orders	
ALFISOL	Alfisol (%)
ULTISOL	Ultisol (%)
INCEPTSL	Inceptisol (%)
SPODOSOL	Spodosol (%)
MOLLISOL	Mollisol (%)

Elevation	
MAXELV	Maximum elevation (m)
MINELV	Minimum elevation (m)

TABLE 3.2. Candidate variables considered for predicting abundance of tree species in the eastern United States. The first group consists of a series of climate factors generated by the United States Environmental Protection Agency (EPA) Environmental Research Laboratory (1993). The second class of variables relates to soil conditions in each county and was collected by the USDA Soil Conservation Service (now Natural Conservation Service). This class is further divided into groups of factors relating to the habitat, the soil texture and the soil orders for each county. Finally, elevation (the minimum and maximum for each county) was obtained from the US Geological Survey. The data were provided by the USDA Forest Service.

different tree species. After some preliminary exploratory analysis, we decided to model the square root of IV and focus our attention on the middle region of the eastern half of the US consisting of 1093 counties. (After some experimentation, these choices tended to remove some of the skew in the residuals.) The spline procedure we applied is similar to the curve fitting routine outlined for the $^{87}\delta\,\mathrm{Sr}$ data: We generated a number of different candidate spline models with different complexities and employed a model selection criterion to identify the "best" or final fit. For this example, we chose to work with linear splines, and applied a methodology known as PolyMARS. In Figure 3.6 we present the spline components from this model involving just the climate variables (AVGT, JULT and PPT). In each case, dependence of IV on the selected variable involves a single breakpoint or bend. In all, the model makes use of 15 of the candidate predictors; 9 appear as simple linear effects (no breakpoints) and the remaining 6 are all simple broken lines as in Figure 3.6. We will have more to say about this model and the stepwise process used to construct it later in the chapter.

In the lower right corner of Figure 3.6 we present a simple quantile-quantile plot of the residuals from the spline fit against the standard normal distribution. Common regression diagnostics did not suggest any gross misfit, although the R^2 value is only 58%. Iverson and Prasad (2000) report similar values for their fits to the data, in part suggesting that this is a difficult estimation problem. In Figure 3.7 we examine the fit to the IV data a bit more closely. In the upper left corner, we present the raw importance values. Predictions under current climate conditions are mapped out in the upper right panel of this figure. The model has trimmed some of the peaks (the light patches in these images) and has smoothed out areas in the Ohio valley. Still, as was true for our piecewise-linear fit to the $^{87}\delta\,\mathrm{Sr}$ values in the previous section, the broad structures appear to be reproduced by the model.

Using this fit, we now consider two different scenarios for climate change in North America. In both cases, we generate IV values for the American beech using the same county-by-county values for the soil factors, but replacing the climate variables with predictions from one of two different models for weather conditions associated with a doubling of CO_2 levels. The lower left image in Figure 3.17 is based on predictions from the Hadley Centre for Climate Prediction and Research, part of the Meteorological Office of the British Government. The lower right image is based on data from the Canadian Centre For Climate (CCC) Prediction and Analysis. The climate predictions themselves are taken to be county-wise, 30 year averages from the period 2071–2100. The Hadley scenario predicts increases in the temperature variables of about 2°C in each county, while CCC predicts a jump of 5°C per county. In terms of precipitation, the Hadley model anticipates county increases of between 200 and 400mm (with only 3 of the 1093 counties experiencing less rainfall). The CCC model, on the other

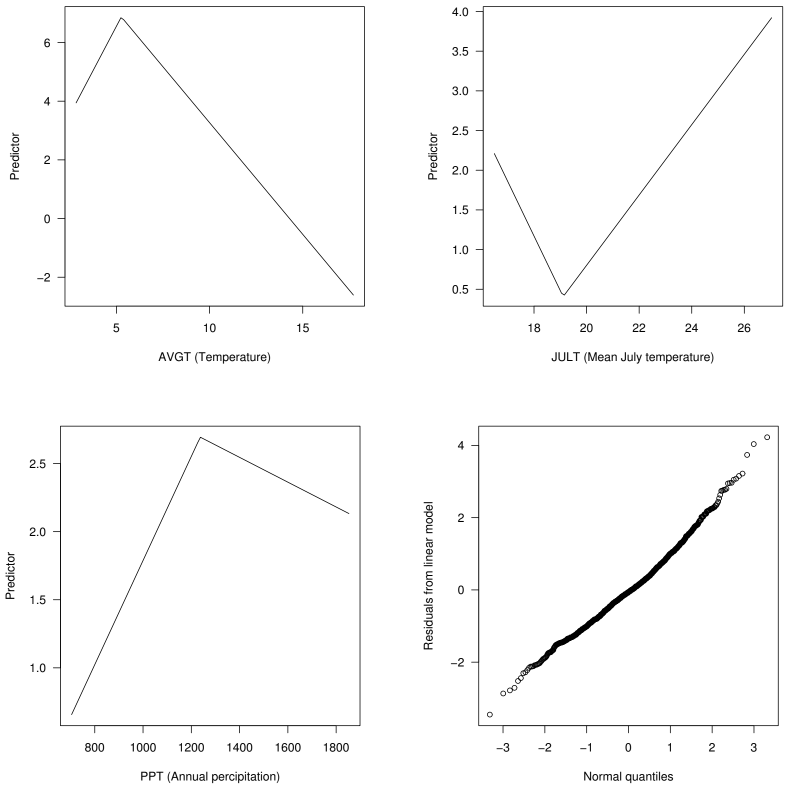

FIGURE 3.6. Some of the functional components (based on the climate variables) in the spline fit to the square root of the IV values computed for the American Beech.

hand, suggests that about a third of the counties will have less rainfall, and that the county-specific shifts will be between −200 and 200 mm. Geographically, Hadley predicts much greater rainfall in the southeast, near the mountains of Tennessee and the coast of North Carolina, while CCC predicts modestly more rainfall in the Great Lakes region.

For each of Hadley and CC, we extract county-by-county values of AVGT, JULT and PPT and supply them (together with the soil variables) to the partially exhibited model in Figure 3.6. Comparing the upper and lower rows in Figure 3.7, we see that both climate predictions indicate a dramatic shift in the regions with high IV values for the American beech. The trend is northward in each case, with the CCC model predicting a virtual elimination of the species from the US. To quantify this shift, we considered the percent change in an area-weighted IV score for each of the fitted models under the two climate scenarios. Under our spline model, we see

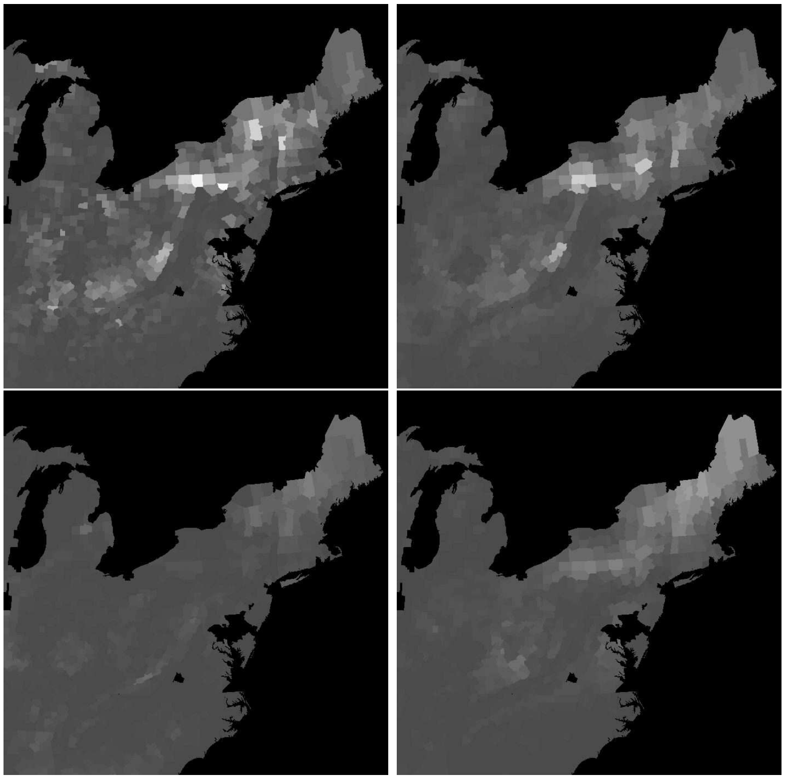

FIGURE 3.7. Predicted abundance of the American beech using the linear spline model. Upper left is the raw data; upper right is the fit. The lower set of plots are predictions under two different global warming scenarios; CCC is left, Hadley is right.

a 35% drop in the area-weighted IV scores under the Hadley model, with a 68.5% drop under the CCC predictions. In Section 3.4 we will extend this analysis to other tree species and even extend the spline modeling to work with multiple IV values at one time. Finally, we consider a censored regression model that is more appropriate for these data.

3.2 Regression modeling and approximation spaces

Given a set of (potentially relevant) predictor variables $X_1, \ldots, X_d$ and a univariate response Y, our interest is in describing the dependence of Y on $\boldsymbol{X} = (X_1, \ldots, X_d)$. Formally, we want to capture the major features evident in the *regression function*

$$f(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x}), \qquad \boldsymbol{x} \in \mathcal{X}, \tag{3.2.1}$$

where $\mathcal{X}$ is a (possibly unbounded) subset of $\mathbb{R}^d$. Throughout this chapter we will assume that the conditional variance of Y given $\boldsymbol{X}$ is constant; that is,

$$\operatorname{var}(Y|\boldsymbol{X}=\boldsymbol{x})=\sigma^2 \qquad \boldsymbol{x}\in\mathcal{X}\,. \tag{3.2.2}$$

Equivalently, we can write

$$Y=f(\boldsymbol{X})+\epsilon, \tag{3.2.3}$$

where $E(\epsilon|\boldsymbol{X})=0$ and $\operatorname{var}(\epsilon|\boldsymbol{X})=\sigma^2$. As an example, note that the *normal regression model*

$$Y=f(\boldsymbol{X})+\epsilon, \quad \epsilon\sim N(0,\sigma^2)\,, \tag{3.2.4}$$

where ϵ is independent of $\boldsymbol{X}$, yields a dependence of the form (3.2.1) and (3.2.2).

We gain insight into the important features in the relationship between $\boldsymbol{X}$ and Y by entertaining various descriptions of our models for f. Through this exercise, we might identify the width and height of peaks or perhaps simply explore the overall shape of f in some neighborhood, finding areas of sharp increase or regions exhibiting little curvature. This was the case with the ${}^{87}\delta\,\mathrm{Sr}$ data in the previous section. Sometimes the need to estimate f arises when investigators have to decide among various explanations for a physical phenomenon, explanations that might be indistinguishable from the standpoint of existing subject-knowledge or scientific theory. Exhibiting some aspect of f may then imply the confirmation or revision of a given theory. This was the case with the importance values IV for the American beech: The dependence on certain climate variables helped us understand the "migration" patterns we could expect from this species under global warming. The data to support such investigations are typically a set of n paired observations $(\boldsymbol{X}_1,Y_1),\ldots,(\boldsymbol{X}_n,Y_n)$. These can be either a random sample from the joint distribution of $(\boldsymbol{X},Y)$, as is the case in observational studies; alternatively, the input values $\{\boldsymbol{X}_i\}$ can be fixed, as in a designed experiment. For the moment, we assume that the inputs are distinct.

In the next three sections, we will review the basic components of regression analysis. We start with the least squares criterion and the computation of regression estimates; we then consider the bias-variance tradeoff that arises in the context of spline modeling (or any time in which we work with so-called approximation spaces); and finally, we examine simple model selection criteria, rules for comparing different fits and determining a single "best" model from the available candidates. Those familiar with regression analysis can safely skim the material in Section 3.2.1 and skim or skip Sections 3.2.2 and 3.2.3 on first reading.

3.2.1 Linear spaces and ordinary least squares

Our approach to estimating f involves the use of finite-dimensional linear spaces, and in particular one of the approximation spaces presented in Chapter 2. (We will mainly be concerned with polynomials, piecewise polynomials, or splines; the separate fits in Figure 3.2 is a good example of the different spaces we might consider). A *linear model* for the regression function (3.2.1) consists of a J-dimensional linear space $\mathbb{G}$ having as a basis the functions

$$B_j(\boldsymbol{x}), \qquad j = 1, \ldots, J, \tag{3.2.5}$$

defined for $\boldsymbol{x} \in \mathcal{X}$. Each member of $g \in \mathbb{G}$ can be written uniquely as a linear combination

$$g(\boldsymbol{x}) = g(\boldsymbol{x}; \boldsymbol{\beta}) = \beta_1 B_1(\boldsymbol{x}) + \cdots + \beta_J B_J(\boldsymbol{x})\,, \qquad \boldsymbol{x} \in \mathcal{X}\,, \tag{3.2.6}$$

for a unique value of the coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^{\mathrm{T}}$. The space $\mathbb{G}$ represents a linear model, and each $g \in \mathbb{G}$ is a candidate description for f. In the next section, we will consider models $\mathbb{G}$ that are constructed from polynomials or piecewise polynomials.

We choose between the competing functions $g \in \mathbb{G}$ on the basis of the *ordinary least squares* (OLS) criterion

$$\rho(g) = \sum_{i=1}^{n} [\, Y_i - g(\boldsymbol{X}_i)\,]^2\;, \qquad g \in \mathbb{G}\,. \tag{3.2.7}$$

The function

$$\widehat{g} = \operatorname*{argmin}_{g \in \mathbb{G}} \rho(g) \tag{3.2.8}$$

minimizing this criterion is referred to as the OLS estimate of f in $\mathbb{G}$.

Computationally, we solve this problem by rewriting the OLS criterion in terms of the parameter vector $\boldsymbol{\beta}$ as

$$\begin{aligned}\rho(\boldsymbol{\beta}) &= \sum_{i=1}^{n} [\, Y_i - g(\boldsymbol{X}_i; \boldsymbol{\beta})\,]^2 \\ &= \sum_{i=1}^{n} [\, Y_i - \beta_1 B_1(\boldsymbol{X}_i) - \cdots - \beta_J B_J(\boldsymbol{X}_i)\,]^2\,,\end{aligned} \tag{3.2.9}$$

we find that

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^J} \rho(\boldsymbol{\beta}), \tag{3.2.10}$$

and that $\widehat{g} = g(\boldsymbol{x}; \widehat{\boldsymbol{\beta}})$. We obtain the OLS estimate $\widehat{\boldsymbol{\beta}}$ by solving the so-called *normal equations*

$$(\mathbf{B}^{\mathrm{T}}\mathbf{B})\widehat{\boldsymbol{\beta}} = \mathbf{B}^{\mathrm{T}}\boldsymbol{Y}, \tag{3.2.11}$$

where $\mathbf{B}$ is the $n \times J$ design matrix with elements $[\mathbf{B}]_{ij} = B_j(\boldsymbol{X}_i)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. If we solve for $\widehat{\boldsymbol{\beta}}$ in the above expression we find

$$\widehat{\boldsymbol{\beta}} = (\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\boldsymbol{Y}\,. \tag{3.2.12}$$

Finally, we define the residual sum of squares RSS associated with $\mathbb{G}$ to be the value of the OLS criterion for $\widehat{g}$; that is,

$$\mathrm{RSS}(\mathbb{G}) = \min_{g\in\mathbb{G}} \rho(g) = \rho(\widehat{g}) = \sum_{i=1}^{n} [Y_i - \widehat{g}(\boldsymbol{X}_i)]^2 \,. \tag{3.2.13}$$

We will study some properties of RSS in Section 3.2.3.

Connecting with classical treatments on regression, under the normal model (3.2.4), $\rho(g)$ is proportional to the log-likelihood for g or, equivalently, $\boldsymbol{\beta}$. If we assume that f is contained in $\mathbb{G}$, then there is a "true parameter" $\boldsymbol{\beta}^*$ such that $f = g(\boldsymbol{x};\boldsymbol{\beta}^*)$. The reader can consult Cook and Weisberg (1999), Seber (1977) or Rao (1973) for basic background material on classical regression modeling and the geometry of OLS estimation. We have chosen to work with approximation spaces (piecewise polynomials or splines) because they can adapt to a variety of different features; they can provide a good description of an unknown regression function. Therefore, we are rarely comfortable with the assumption that $f \in \mathbb{G}$; and instead, we assume that f can be reasonably well *approximated* by some member of $\mathbb{G}$. Still, the least squares criterion is a useful measure of misfit and is frequently applied in such cases. In the next section, we examine some unique aspects of regression analysis involving approximation spaces.

3.2.2 The bias-variance tradeoff

In this subsection, we derive one version of the bias-variance tradeoff for regression estimators based on linear spaces. In Sections 3.3.2 and 3.4.2 we will apply this result for univariate and multivariate splines, respectively. For simplicity, we assume the input variables, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, are fixed. While it is possible to derive versions of these results more generally, we delay a complete treatment until Chapter 11. For the moment, the reader can either think of the data as coming from a sequence of designed experiments, or instead think of the analysis as conditional on the values of the input data. Then, given the design points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, we make observations $Y_1, \ldots, Y_n$ according to the regression setup (3.2.3).

In this book, we are concerned with linear spaces $\mathbb{G}$ that are adaptable in the sense that they can describe a wide range of smooth functions. In Section 3.1, for example, we fit different smooth curves to the $^{87}\delta$ Sr data. In the literature on numerical analysis, it is common to quantify the notion of adaptability in terms of the achievable *approximation error.* To be precise, given a function f and a linear space $\mathbb{G}$, we define the distance

$$d(f,\mathbb{G}) = \min_{g\in\mathbb{G}} \|f-g\|_\infty, \tag{3.2.14}$$

where the norm on the right is defined by

$$\|f-g\|_\infty = \sup_{\boldsymbol{x}} |f(\boldsymbol{x}) - g(\boldsymbol{x})| \,.$$

The approximation error (3.2.14) associated with spaces of polynomials or piecewise polynomials is well known and will be used as motivation in the next section.

In statistical applications, the flexibility or approximation power of a linear space is (quite literally) only half the battle. When we use a linear space $\mathbb{G}$ to form an estimate of the unknown regression function f, we need a different notion of error, one that specifically incorporates the fact that we are working with noisy observations of f at some set of points. To this end, for any point $\boldsymbol{x}_0 \in \mathcal{X}$, we define the pointwise error

$$E\Big(\big[f(\boldsymbol{x}_0) - g(\boldsymbol{x}_0; \widehat{\boldsymbol{\beta}})\big]^2\Big),$$

where the expectation is taken with respect to $Y_1, \ldots, Y_n$. This quantity tells us on average how well we can expect to capture $f(\boldsymbol{x}_0)$ if we repeat the experiment multiple times, each time drawing new observations at our design points. We gain insight into this metric by decomposing it into two components:

$$\begin{aligned} &E\Big(\big[f(\boldsymbol{x}_0) - g(\boldsymbol{x}_0; \widehat{\boldsymbol{\beta}})\big]^2\Big) \\ &\qquad = \big[f(\boldsymbol{x}_0) - g(\boldsymbol{x}_0; \widetilde{\boldsymbol{\beta}})\big]^2 + E\Big(\big[g(\boldsymbol{x}_0; \widetilde{\boldsymbol{\beta}}) - g(\boldsymbol{x}_0; \widehat{\boldsymbol{\beta}})\big]^2\Big), \end{aligned} \tag{3.2.15}$$

where $E\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}$. This equality follows easily because $\mathbb{G}$ is a linear space and the representation in equation (3.2.6) holds. Rather than evaluate the model error at one point, we will instead take an average over all the design points. That means replacing x_0 in (3.2.15) with each of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and forming the *model error*

$$\mathrm{ME}(\mathbb{G}) = \frac{1}{n}\sum_{i=1}^{n} E\Big(\big[f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i; \widehat{\boldsymbol{\beta}})\big]^2\Big). \tag{3.2.16}$$

(Since the expectation in the above expression is with respect to the data $Y_1, \ldots, Y_n$ conditional on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, this quantity is often referred to as an *in-sample* measure of error.) We can gain insight into the structure of this quantity with the simple decomposition

$$\begin{aligned} \mathrm{ME}(\mathbb{G}) = &\frac{1}{n}\sum_{i=1}^{n} E\Big(\big[f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i; \widetilde{\boldsymbol{\beta}})\big]^2\Big) \\ &\qquad + \frac{1}{n}\sum_{i=1}^{n} E\Big(\big[g(\boldsymbol{x}_i; \widetilde{\boldsymbol{\beta}}) - g(\boldsymbol{x}_i; \widehat{\boldsymbol{\beta}})\big]^2\Big). \end{aligned} \tag{3.2.17}$$

The expression on the right is often referred to as a *bias-variance decomposition.* To make this terminology precise, we consider each component in a bit more detail.

We begin with the second sum on the right in (3.2.17), which we think of as a variance term. Using the expression in (3.2.12), we find that $\widehat{\boldsymbol{\beta}}$ has mean

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\,,$$

where $= [\mathbf{f}(\boldsymbol{x_1}), \dots, \mathbf{f}(\boldsymbol{x_n})]$ and variance-covariance matrix $\sigma^2(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}$. Now, if we let $\boldsymbol{b}(\boldsymbol{x}_0)$ denote the $1 \times n$ row vector $(B_1(\boldsymbol{x}_0), \dots, B_J(\boldsymbol{x}_0))$ for any point $\boldsymbol{x}_0 \in \mathcal{X}$, then we can rewrite the second term on the right in (3.2.15) as

$$E\Big(\big[g(\boldsymbol{x}_0;\widetilde{\boldsymbol{\beta}}) - g(\boldsymbol{x}_0;\widehat{\boldsymbol{\beta}})\big]^2\Big) = \sigma^2 \boldsymbol{b}(\boldsymbol{x}_0)\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)^{-1}[\boldsymbol{b}(\boldsymbol{x}_0)]^{\mathrm{T}}\,.$$

By summing this expression over the design points we find that the variance component of the model error is given by

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{b}(\boldsymbol{x}_i)\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)^{-1}\boldsymbol{b}(\boldsymbol{x}_i)^{\mathrm{T}} = \frac{\sigma^2}{n}\mathrm{trace}\left[\mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\right] = \frac{J\sigma^2}{n} \qquad (3.2.18)$$

from the properties of the so-called hat matrix, $\mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}$ (see Rao, 1973, Chapter 4; Seber, 1977; or Cook and Weisberg, 1999).

To get a handle on the first term in (3.2.17), note that from the expression for $\widehat{\boldsymbol{\beta}}$ (3.2.12) we can write $\widetilde{\boldsymbol{\beta}}$ as

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{f}\,. \qquad (3.2.19)$$

where $\mathbf{f} = (f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n))^{\mathrm{T}}$ is an $n \times 1$ column vector evaluating f at the design points. In fact, comparing this with (3.2.11), we find that $\widetilde{\boldsymbol{\beta}}$ minimizes the measure

$$\sum_{i=1}^{n}\left[f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i;\boldsymbol{\beta})\right]^2\,. \qquad (3.2.20)$$

over all coefficient vectors $\boldsymbol{\beta}$. Therefore, $g(\boldsymbol{x};\widetilde{\boldsymbol{\beta}})$ is obtained by solving an OLS problem in $\mathbb{G}$ with response $\mathbf{f}$; that is,

$$\frac{1}{n}\sum_{i=1}^{n}\Big[f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i;\widetilde{\boldsymbol{\beta}})\Big]^2 = \min_{g\in\mathbb{G}}\frac{1}{n}\sum_{i=1}^{n}\left[f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i)\right]^2. \qquad (3.2.21)$$

Letting $g^* = \mathrm{argmin}_{g\in\mathbb{G}}\,\|f - g\|_\infty$, we conclude from (3.2.21) that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\Big[f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i;\widetilde{\boldsymbol{\beta}})\Big]^2 &\le \frac{1}{n}\sum_{i=1}^{n}\left[f(\boldsymbol{x}_i) - g^*(\boldsymbol{x}_i)\right]^2 \\ &\le \frac{1}{n}\sum_{i=1}^{n} d^2(f,\mathbb{G}) = d^2(f,\mathbb{G}).\end{aligned}$$

Combining this with (3.2.18), we have the following result.

Proposition 3.2.1. *The the model error can be bounded:*

$$\begin{aligned}\mathrm{ME}(\mathbb{G}) &= \frac{1}{n}\sum_{i=1}^{n} E\Big(\big[f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i;\widehat{\boldsymbol{\beta}})\big]^2\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n} E\Big(\big[f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i;\widetilde{\boldsymbol{\beta}})\big]^2\Big) + \frac{J\sigma^2}{n} \qquad (3.2.22) \\ &\le d^2(f,\mathbb{G}) + \frac{J\sigma^2}{n}\,. \qquad (3.2.23)\end{aligned}$$

The last expression is referred to as a bias-variance decomposition; its first component represents the squared bias *and the second is the* variance *of estimates from* $\mathbb{G}$*. The bias is controlled by the approximation error of* $\mathbb{G}$*, while the variance term simply records the number of terms used in our model, which is the dimension of* $\mathbb{G}$*.*

The model error in Proposition 3.2.1 is one component of the *prediction error*

$$\mathrm{PE}(\mathbb{G}) = \frac{1}{n}\sum_{i=1}^{n} E_{\boldsymbol{Y}} E_{\boldsymbol{Y}^*}\Big(\big[Y_i^* - g(\boldsymbol{x}_i;\widehat{\boldsymbol{\beta}})\big]^2\Big)\,, \qquad (3.2.24)$$

where Y_i^* is a new observation from the model (3.2.3) for $i = 1, \ldots, n$. The inner expectation is with respect to the new data points $\boldsymbol{Y}^* = (Y_1^*, \ldots, Y_n^*)$, and it records the error in our predictions based on models from $\mathbb{G}$. In this expression we have also made explicit the fact that the outer expectation is with respect to the sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ used to construct $\widehat{\boldsymbol{\beta}}$. Notice that this is the same expectation that appears in our definition of model error, (3.2.17). It is not hard to show that the prediction error can be written as

$$\mathrm{PE}(\mathbb{G}) = \sigma^2 + \mathrm{ME}(\mathbb{G})\,. \qquad (3.2.25)$$

The first term is often referred to as an *irreducible error* in the sense that we cannot eliminate it by our choice of model space $\mathbb{G}$. In various chapters of this book, we will alternate between methods that select models based on estimates of either PE or ME.

3.2.3 How smooth? Some simple model selection criteria

In our presentation on the $^{87}\delta\,$Sr data, we needed to evaluate different models, which represented different degrees of smoothness across the KTB. Initially, we relied on simple F-statistics to help us decide the question. Throughout this chapter, we will encounter the problem of deciding which aspects of the data represent structures in our regression function and which are artifacts of noise. In the case of the $^{87}\delta\,$Sr data, researchers are interested in behavior of the function at a single point, the KTB. In most situations, however, we do not have such refined hypotheses and the F-statistic approach breaks down. Instead, we will compare not just two different linear models but possibly hundreds.

Estimates of model error and related criteria

To guide our search for linear models that are well supported by the data, we will employ a *model selection criterion.* One of the first of these is based on model error (3.2.16). Obviously, $\mathrm{ME}(\mathbb{G})$ is not directly useful as a metric for evaluating $\mathbb{G}$ as it depends on the unknown function f. Instead, Mallows (1973, 1994) defined an estimate of ME known as Mallows' C_p. For a J-dimensional linear space $\mathbb{G}$ with residual sum of squares RSS, the criterion is defined by

$$C_p = \frac{\mathrm{RSS}}{\sigma^2} - n + 2J\,, \tag{3.2.26}$$

where we have assumed for the moment that the error variance σ^2 is known. Extending slightly the analysis leading to (3.2.22), we find that

$$E(\mathrm{RSS}) = \sum_{i=1}^{n} E\Big(\big[f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i;\widetilde{\boldsymbol{\beta}})\big]^2\Big) + \sigma^2(n-J) \tag{3.2.27}$$

$$= n\mathrm{ME} + \sigma^2(n-2J) \tag{3.2.28}$$

so that C_p is an unbiased estimate of $\frac{n}{\sigma^2}\mathrm{ME}$.

Therefore, when σ^2 is known, we can use C_p as a (scaled) estimate of ME. Given two linear spaces $\mathbb{G}_1$ and $\mathbb{G}_2$ with RSS_1 and RSS_2, respectively, we would prefer the model with the smaller value of C_p. Suppose $\mathbb{G}_1 \subset \mathbb{G}_2$ and that $\mathbb{G}_1$ has J parameters while $\mathbb{G}_2$ has $J+1$ parameters. The C_p value for $\mathbb{G}_1$ is smaller than that for $\mathbb{G}_2$ if

$$\frac{\mathrm{RSS}_2}{\sigma^2} - n + 2(J+1) > \frac{\mathrm{RSS}_1}{\sigma^2} - n + 2J \quad \text{or} \quad \frac{\mathrm{RSS}_2 - \mathrm{RSS}_1}{\sigma^2} > 2\,.$$

It is possible to go farther and interpret this result in terms of classical hypothesis testing, and even develop the F tests alluded to at the beginning of the chapter (substituting an estimate for σ^2 based on the residual sum of squares associated with $\mathbb{G}_2$). In situations where we believe that the unknown regression function f is actually contained in one of the linear spaces we are considering, this approach is sensible, and certainly a mainstay of classical statistics.

In our applications, however, we are entertaining linear spaces because they have good approximation properties, and we do not believe f is actually a member of any of these spaces. Hence we prefer to consider criteria like C_p as estimates of model error. Akaike (1973) arrived at the same measure, but took an information-theoretic approach. He considered a general likelihood and approximated the Kullback-Liebler divergence between the data-generating model and its maximum likelihood estimate using $\mathbb{G}$. This criterion is given in general by

$$-2 \max_{g \in \mathbb{G}} \ell(g) + 2J \,, \tag{3.2.29}$$

where $\ell(\mathbb{G})$ is the maximized log-likelihood computed for the model $\mathbb{G}$. For the normal linear model, this becomes essentially C_p or

$$\frac{\text{RSS}}{\sigma^2} + 2J \qquad \text{when } \sigma^2 \text{ is known}$$

and

$$n \log \text{RSS} + 2J \quad \text{when } \sigma^2 \text{ is unknown.}$$

The form of AIC (3.2.29) has been modified by several authors to incorporate more or less penalty on dimension, depending on the characteristics of the function under study, the smoothness of the splines, and the type of search used to identify the final model; see Section 3.6.3 for a simulation to illustrate the point. In general, these alterations yield a criterion of the form

$$-2 \max_{g \in \mathbb{G}} \ell(g) + p(n)J \,, \tag{3.2.30}$$

where $p(n)$ is some increasing function of sample size. For the normal linear model with σ^2 unknown, this is simply

$$n \log \text{RSS} + p(n)J. \tag{3.2.31}$$

One such example is the Bayesian Information Criterion (BIC) developed by Schwarz (1978)

$$-2 \max_{g \in \mathbb{G}} \ell(g) + (\log n)J \,. \tag{3.2.32}$$

Schwarz derived this formula as an approximation to a posterior distribution on model classes and not as an estimate of prediction error. The interested reader is referred to Chapter 10 for more discussion of this.

Estimates of prediction error

Many model selection schemes estimate the prediction error and decide between competing linear spaces on the basis of this estimate. A *test set* is used in the simplest such approach. Here we set aside an independent set of observations $(\boldsymbol{X}_1^*, Y_1^*), \ldots, (\boldsymbol{X}_K^*, Y_K^*)$ and use them to estimate PE as

$$\widehat{\text{PE}}_{\text{TS}}(\mathbb{G}) = \frac{1}{K} \sum_{i=1}^{K} \left[Y_i^* - g(\boldsymbol{X}_i^*; \widehat{\boldsymbol{\beta}}) \right]^2 . \tag{3.2.33}$$

Note that in this formulation, our new observations are not necessarily taken at the original design points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. The definition and analysis we followed for ME and PE that held the design points fixed can be relaxed and we can instead take our design points to be observations from a random variable $\boldsymbol{X}$. In that case, we take an expectation with respect to the distribution of the pair $(\boldsymbol{X}, Y)$ in (3.2.16) and (3.2.24), replacing the average over the fixed design points. We cover this approach in greater depth in Chapter 11.

When a test set is not available, *leave-one-out cross validation* has been proposed as another estimate of the prediction error. Here

$$\widehat{\mathrm{PE}}_{\mathrm{CV}}(\mathbb{G}) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - g(\boldsymbol{x}_i; \widehat{\boldsymbol{\beta}}^{(-i)})\right]^2. \tag{3.2.34}$$

where $\widehat{\boldsymbol{\beta}}^{(-i)}$ is the least squares fit in $\mathbb{G}$ omitting the point $(\boldsymbol{X}_i, Y_i)$. The omitted point is now a test set of size one. For OLS, this leave-one-out cross-validated estimate of prediction error can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\left[Y_i - g(\boldsymbol{x}_i; \widehat{\boldsymbol{\beta}})\right]^2}{D_i^2}, \tag{3.2.35}$$

where D_i is the ith diagonal entry of $\boldsymbol{I} - \mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}$. Note that in this expression, we are now working with the ordinary residuals and not the residual from a fit that dropped the ith point. The derivation of this fact can be found in Miller (1990).

An adjusted RSS criterion known as generalized cross-validation was proposed by Craven and Wahba (1975):

$$\mathrm{GCV} = \frac{1}{n}\frac{\mathrm{RSS}}{\left(1 - \frac{J}{n}\right)^2}. \tag{3.2.36}$$

The original motivation for this scheme was as an approximation to leave-one-out cross validation in the context of smoothing splines. For OLS, it is obtained by replacing the D_i in (3.2.35) with their average. Recalling that the sum of the diagonal elements of $\mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}$ is just J, the average of the D_i is $(1 - J/n)$. Making this substitution in (3.2.35) yields (3.2.36). Similar adjustments for the penalty implicit in this expression have also been applied to the GCV criterion, where one might write

$$\frac{1}{n}\frac{\mathrm{RSS}}{\left(1 - \frac{p(n)J}{n}\right)^2}. \tag{3.2.37}$$

to again add a higher penalty for increased dimensionality. Here we must restrict consideration to models having fewer than $n/p(n)$ basis elements. The criterion (3.2.37) is used in MARS (Friedman, 1990), Hybrid adaptive splines (Luo and Wahba, 1995) and PolyMars (Stone, et al., 1997) for comparing different linear spaces.

We can relate this criterion to C_p by noting that

$$\log(1-x) \approx -x$$

for small x. Applying this to (3.2.37), we get that

$$\log\left(\frac{1}{n}\frac{\text{RSS}}{\left(1-\frac{p(n)J}{n}\right)^2}\right) \approx -\log n + \log \text{RSS} + 2\,\frac{p(n)J}{n}\,. \tag{3.2.38}$$

This expression is proportional to the rule (3.2.31), providing we pick our penalties properly. (Actually, the two expressions also differ by an additive constant that depends only on n and hence will not change our model comparisons.)

3.3 Curve estimation

3.3.1 From polynomials to splines

For the moment, the regression function f given in (3.2.1) depends on a single, real-valued predictor X (and hence we drop the boldface notation) ranging over some (possibly infinite) subinterval $\mathcal{X} \subset \mathbb{R}$ of the real line. Therefore, the dependence of the mean of Y on X is given by

$$f(x) = E(Y|X=x), \qquad x \in \mathcal{X} \subset \mathbb{R}. \tag{3.3.1}$$

This setup is often referred to as scatterplot smoothing. The properties of OLS estimates based on (fixed, univariate) spline spaces $\mathbb{G}$ are most easily characterized in this simple context. After illustrating the implicit smoothing parameters that control the flexibility of a spline model, we take up the topic of adaptation. Familiar model selection schemes translate almost immediately into easily understood smoothing methods.

Throughout this section, we will return to the data in Figure 3.1 as we discuss various curve estimation procedures. The ability to resolve the apparent peak near the KTB is a good benchmark for comparing different models. In motivating the applicability of spline models, we borrow ideas and notation from Chapter 2. We have tried to keep the material somewhat self-contained, however, so that a reader unfamiliar with (or uninterested in) detailed results from approximation theory can easily follow our discussion. Splines have been used extensively as flexible tools for data analysis, so we have chosen a somewhat pragmatic introduction to the subject.

Polynomials

Recall the univariate regression setup given in (3.3.1). In many practical applications, standard exploratory analysis reveals that a simple linear relationship between the predictor X and the observed output Y is inadequate;

that is, as a description of the regression function f, the model

$$g(x;\boldsymbol{\beta}) = \beta_1 + \beta_2 x, \qquad x \in \mathcal{X}\,,$$

ignores important features in the data. This is certainly the case for the values of $^{87}\delta\,\mathrm{Sr}$ plotted in Figure 3.1. To overcome such deficiencies, we might consider a more flexible polynomial model. Let $\mathcal{P}_k$ denote the linear space of polynomials in x of *order* (at most) k defined as

$$g(x;\boldsymbol{\beta}) = \beta_1 + \beta_2 x + \beta_3 x^2 + \cdots + \beta_k x^{k-1}, \qquad x \in \mathcal{X}. \tag{3.3.2}$$

for the parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k) \in \mathbb{R}^k$. Note that the space $\mathcal{P}_k$ consists of polynomials having *degree* $k-1$.

In exceptional cases, we have reason to believe that the regression function f is in fact a high-order polynomial. This *parametric* assumption could be based on physical models describing how the data were generated. For historical values of $^{87}\delta\,\mathrm{Sr}$, however, we consider polynomials simply because we believe f to be *smooth.* We recall from elementary analysis that polynomials are good at approximating well-behaved functions in reasonably tight neighborhoods (the classical Taylor expansions). If f is not exactly given by (3.3.2), then our estimates will be biased by an amount that reflects the approximation error incurred by a polynomial model. (This is just the analysis in Section 3.2.2.) In Chapter 11 we formalize the assumption of smoothness for f. In short if all we are able to say about a function is that it is smooth, we are naturally led to a *nonparametric* problem; that is, we cannot express f as a member of any finite dimensional linear space. The interested reader is referred to Chapters 11 and 12 for more discussion of these ideas.

Given (X_i, Y_i), $i = 1, \ldots, n$, a unique OLS estimate $\widehat{\boldsymbol{\beta}}$ exists provided there are at least k unique values in $\{X_i\}$. We should note, however, that the basis of *monomials*

$$B_j(x) = x^{j-1} \qquad j = 1, \ldots, k\,, \tag{3.3.3}$$

given in (3.3.2) is not well suited for numerical calculations. While convenient for analytic manipulations (differentiation, integration), this basis is *ill-conditioned* for k larger than 8 or 9; that is, the matrix operations needed to compute the coefficient estimates in (3.2.10) are prone to rounding and other numerical errors. We make this notion precise in Section 3.7.

Returning to our $^{87}\delta\,\mathrm{Sr}$ data in Figure 3.1, we see that they exhibit a local minimum between 50 and 60 million years ago, reach a peak at the KTB, and then decrease rather quickly (toward 100 million years). By simply counting features, we expect that at least a cubic polynomial is needed for a good fit (that is, $k \geq 4$). Several OLS estimates are plotted in Figure 3.8. The cubic (order 4) and even quintic (order 6) models are clearly inadequate. While larger values of k yield greater adaptability to resolve

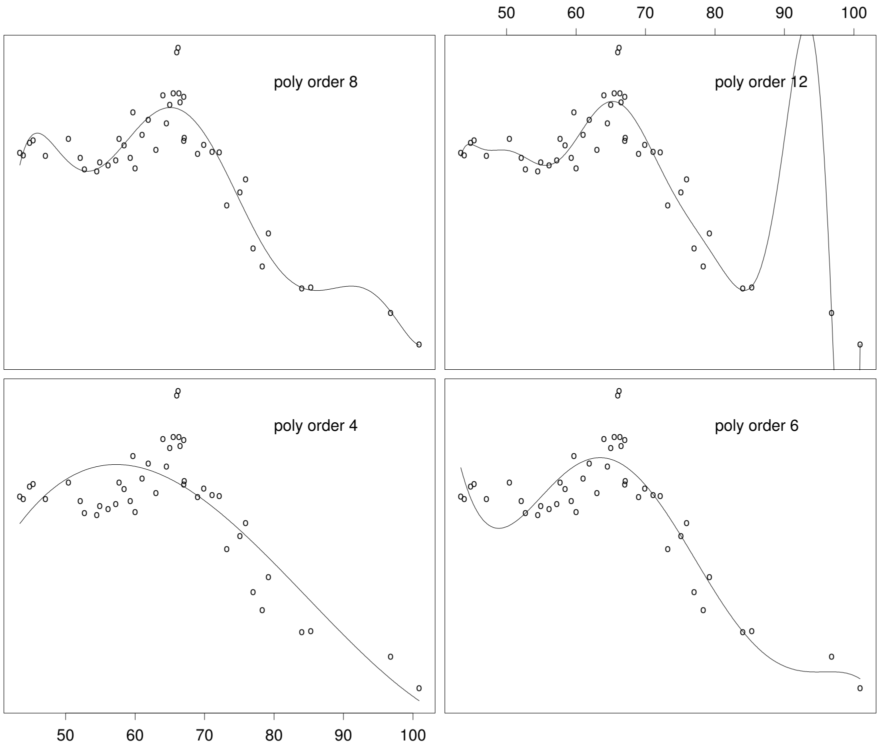

FIGURE 3.8. Simple polynomial fits to the $^{87}\delta\,\mathrm{Sr}$ dataset. Increasing the order of a polynomial adds flexibility, but it can yield wild features in regions with few data points.

features like the peak at the KTB, the fit degrades dramatically in other regions. For example, the data are thin near the 90 million year mark, and the highest order fit oscillates wildly. Similarly, a spurious peak appears at 45 million years for $k = 8$. This feature slowly disappears for $k \geq 9$ (see the upper right hand panel of Figure 3.8). Although we considered polynomials because they represent a space of smooth, approximating functions, it seems that they are in some sense "too smooth." To overcome this problem, we are led to models with large order that can seriously overfit the data. For more information on the difficulty in working with high-order polynomials see Schumaker (1981).

In Chapter 2, we derived a result from approximation theory known as Jackson's Theorem. This result is basically a refinement of the classical Taylor expansion for smooth functions. It says that if a function f defined on the interval $\mathcal{X} = [a, b]$ has bounded pth derivative, then

$$\mathrm{dist}(f, \mathcal{P}_k) \leq C\Big(\frac{b-a}{2k}\Big)^p, \tag{3.3.4}$$

for $k > p$, where C is a constant that depends only on p and properties of the pth derivative of f. Clearly, we have the same effect on the approximation error in (3.3.4) by dividing $[a, b]$ into, say, 5 equal pieces and approximating f with separate cubic polynomials ($k = 4$) in each as we do by approximating f by a single polynomial of order $5 \times 4 = 20$ in the entire interval $[a, b]$. Each of these options has 20 degrees of freedom (unknown parameters to be estimated).

Recall that the distance between a function f and a linear space $\mathbb{G}$, $\text{dist}(f, \mathbb{G})$, records how well f can be approximated by *some* member of $\mathbb{G}$. The result above demonstrates that the two spaces (a global polynomial of order 20 or 5 cubic pieces) are equally flexible and, if we knew f, we could find some function of each type that is close to f. What we have seen in the previous section, however, is that it can be very hard *based on data* to find these functions; identifying a polynomial of order 20 that closely tracks f when we are only privy to 45 noisy observations leads to possibly wild effects. (It turns out that these effects can happen even if we observe samples of f without noise; the issue is less about noise than it is about the difficulty of working with polynomials; see Schumaker, 1981 for more details.) The use of lower-order polynomials piecewise turns out to be a powerful idea which we now examine in more detail.

Piecewise polynomials

Given a sequence $a = t_0 < t_1 < \cdots < t_m < t_{m+1} = b$, we construct $m + 1$ (disjoint) intervals

$$\mathcal{X}_l = [t_{l-1}, t_l), \quad 1 \le l \le m \text{ and } \mathcal{X}_{m+1} = [t_m, t_{m+1}],$$

whose union is $\mathcal{X} = [a, b]$. Set $\boldsymbol{t} = (t_1, \ldots, t_m)$ and let $\mathcal{PP}_k(\boldsymbol{t})$ denote the space of piecewise polynomials of order k defined on $\cup_m \mathcal{X}_m$. Then each function $g \in \mathcal{PP}_k(\boldsymbol{t})$ is of the form

$$g(x) = \begin{cases} g_1(x) = \beta_{1,1} + \beta_{1,2}x + \cdots + \beta_{1,k}x^{k-1}, & x \in \mathcal{X}_1 \\ \quad\vdots \qquad\qquad \vdots \qquad\qquad \vdots & \quad\vdots \\ g_{m+1}(x) = \beta_{m+1,1} + \beta_{m+1,2}x + \cdots + \beta_{m+1,k}x^{k-1}, & x \in \mathcal{X}_{m+1}, \end{cases} \tag{3.3.5}$$

and for the parameter vector $\boldsymbol{\beta} = (\beta_{1,1}, \ldots, \beta_{1,k}, \ldots, \beta_{m+1,1}, \ldots, \beta_{m+1,k})$ we write $g(x) = g(x; \boldsymbol{\beta})$. The least squares criterion (3.2.9) is again used to form an estimate of f based on observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. This time, our solution is obtained by a series of ordinary polynomial regressions of the form (3.2.11), one for each interval. We are guaranteed a solution provided that $X_1, \ldots, X_{m+1}$ each contain k or more distinct values of $\{X_i\}$. As mentioned above, we envision taking k small so that we avoid problems of numerical instability. This is but the first advantage of a piecewise approach.

For Figure 3.9 we constructed piecewise linear and quadratic models having the same degrees of freedom as the corresponding polynomial fits in

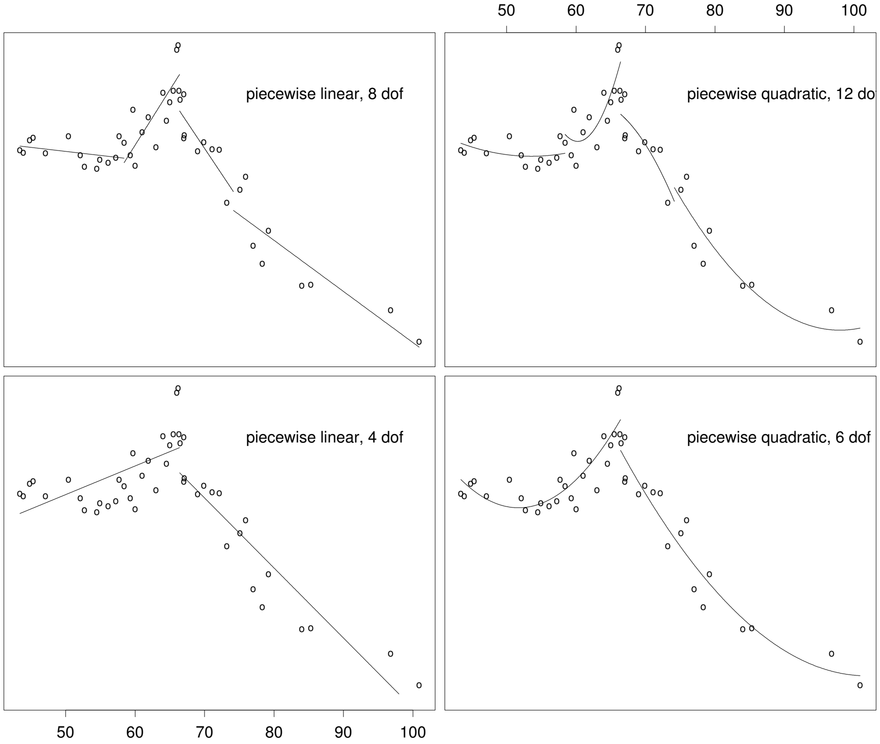

FIGURE 3.9. Piecewise polynomial fits to the $^{87}\delta\,\mathrm{Sr}$ dataset. The number of degrees of freedom in each panel matches that for Figure 3.8. While the approximation power in each case is essentially the same, these estimates appear to track the data better.

Figure 3.8. In comparing these two figures, we find our second advantage to piecewise modeling. The fits from $\mathcal{PP}_k$ with equal degrees of freedom exhibit far fewer wild oscillations than their global counterparts and track trends in the the data much more reliably. For the plots in the bottom row of this figure, the range of the data $[43.4, 100.9]$ was divided into two intervals taking the KTB as a dividing point and piecewise linear (left panel) and quadratic (right panel) polynomials were fit to the $^{87}\delta\,\mathrm{Sr}$ data. Each of these intervals was further divided in half, producing the (4-piece) fits in the upper row. By design, the piecewise polynomial spaces underlying the fits in the left column are subspaces of the models in the right column (piecewise linear functions are also piecewise quadratics providing they are defined relative to the same intervals). This nesting also occurs between the top and bottom rows (each function in the 2-piece spaces is trivially a member of the 4-piece space). Using these relationships, classical F tests have been

suggested to assess the quality of competing piecewise polynomial fits. We will return to this topic in our discussion of splines.

Piecewise polynomials have a long tradition in statistics. Beginning in the late 1950's, econometricians studied these models under the title of "multiphase" or "switching" regressions (Quandt 1958, 1960; Sprent 1961). At that time, many authors viewed $\mathcal{PP}_k$ as a means of constructing parametric descriptions of structural changes in the regression function. As such, f was assumed to be of the form (3.3.5), where the endpoints of the intervals $\mathcal{X}_1, \ldots, \mathcal{X}_{m+1}$ were either given or estimated. For example, taking this approach with our $^{87}\delta\,\mathrm{Sr}$ data, we might posit a model with two quadratic regimes, one before and one after the KTB (as in the lower right hand panel of Figure 3.9). The most natural question about a multiphase regression model is whether or not the fitted structural changes are real. One might also question the order of the polynomial needed in each regime. For known breakpoints, these hypotheses are easily evaluated for fixed k and $\mathcal{X}_1, \ldots, \mathcal{X}_{m+1}$ using the F tests described above. Analyses of this kind can be found in Poirier (1973), Wold (1974) and Smith (1979, 1982ab), among others.

Splines

As mentioned previously, the curves in Figure 3.9 were formed by OLS estimates using nested, piecewise polynomial spaces. Take, for example, the two quadratic models in the right hand column of this figure. These two fits differ by 6 degrees of freedom. One can easily question the need for the breaks at 60 and 75 million years. Would a continuous estimate of $^{87}\delta\,\mathrm{Sr}$ suffice? What about continuous first derivatives? We now consider subspaces of $\mathcal{PP}_k(\boldsymbol{t})$ that satisfy these types of constraints. First, we need to introduce notation describing the smoothness of functions in $\mathcal{PP}_k(\boldsymbol{t})$. For each breakpoint t_l, $1 \leq l \leq m$, the size of the discontinuity in the jth derivative of $g \in \mathcal{PP}_k(\boldsymbol{t})$ at t_l is given by

$$\mathrm{jump}_{t_l}(D^j g) = D^j g_{l+1}(t_l) - D^j g_l(t_l), \tag{3.3.6}$$

where g_l and g_{l+1} are the two polynomial pieces meeting at t_l defined in (3.3.5).

Working directly with (3.3.5) and (3.3.6), we can rewrite a continuity restriction of the form $\mathrm{jump}_{t_l}(D^j g) = 0$ as a simple linear constraint in the coefficients $\boldsymbol{\beta}$. For example, to remove the 6 degrees of freedom separating the upper and lower piecewise quadratic fits in Figure 3.9, we derive continuity constraints (3 for each of the two extra breakpoints) of the form $C\boldsymbol{\beta} = 0$, where C is a 6×12 matrix and $\boldsymbol{\beta}$ is the coefficient vector corresponding to the 4-piece model. This leads to an ordinary F test to decide whether the extra flexibility is necessary. Here, the F statistic corresponds to a P-value of 0.60, indicating that the 2-piece model is sufficient.

Rather than remove a break completely, we may instead prefer only to force the curves to be continuous, or perhaps to have one continuous derivative. These restrictions again yield linear constraints of the form $C\boldsymbol{\beta} = 0$. For example, by imposing a single constraint, we can make the 2-piece, quadratic fit in the lower right hand panel of Figure 3.9 be continuous. Therefore, the effect of removing the jump at the KTB can be evaluated by another F test. In this case, the F statistic has a P-value of 0.0007, strong evidence of a sharp peak in ${}^{87}\delta\,\mathrm{Sr}$. In both this and the previous application of classical hypothesis tests for smoothing a piecewise polynomial model, we have intentionally avoided specifying the matrices C. Our purpose has been to make a connection between tools for parametric model building and nonparametric estimation with piecewise polynomials. While deriving the necessary constraints is a straightforward exercise, it should be clear that this approach becomes needlessly complex for large curve fitting problems. In particular, as we introduce automated procedures for locating breakpoints, constrained fits become impractical.

Instead, we now introduce an alternative basis for piecewise polynomials. While (slightly) more problematic from a computing standpoint, this representation leads naturally to the definition of polynomial splines. Given a sequence of breakpoints $\boldsymbol{t} = (t_1, \ldots, t_m)$ and an order k, we can write $g \in \mathcal{PP}_k(\boldsymbol{t})$ in the *truncated power basis*,

$$\begin{aligned} g(x) = \quad & \beta_{0,1} + \beta_{0,2}x + \cdots + \beta_{0,k}x^{k-1} + \\ & \beta_{1,1}(x-t_1)_+^0 + \beta_{1,2}(x-t_1)_+ + \cdots + \beta_{1,k}(x-t_1)_+^{k-1} + \\ & \quad \vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad\qquad \vdots \\ & \beta_{m,1}(x-t_m)_+^0 + \beta_{m,2}(x-t_m)_+ + \cdots + \beta_{m,k}(x-t_m)_+^{k-1} \end{aligned} \tag{3.3.7}$$

where $(\,\cdot\,)_+ = \max(\,\cdot\,, 0)$. Expressed in this way, the coefficients $\beta_{1,1}, \ldots, \beta_{m,1}$ record the size of the discontinuities in g at the points $t_1, \ldots, t_m$, respectively. Similarly, the jumps in g' at these points are $\beta_{1,2}, \ldots, \beta_{m,2}$. In general, for $g \in \mathcal{PP}_k(\boldsymbol{t})$,

$$\mathrm{jump}_{t_l}(D^j g) = j!\,\beta_{l,j+1} \qquad \text{for } 0 \le j \le k-1\,.$$

The continuity constraints discussed previously now only involve setting one or more elements of $\boldsymbol{\beta}$ equal to zero. For example, fixing $\beta_{1,l} = 0$ guarantees that g is continuous across t_l. Equivalently, we drop the basis element $(x - t_l)_+^0$ from the model (3.3.7).

While the truncated power basis (3.3.7) is extremely natural for specifying the properties of piecewise polynomials, it can be disastrous for numerical computations. Recall that with the piecewise specification (3.3.5), fitting a model from $\mathcal{PP}_k(\boldsymbol{t})$ involved $m+1$ separate polynomial regressions of the form (3.2.11). The normal equations associated with the truncated power basis, however, are more complicated. Given data (X_i, Y_i), $i = 1, \ldots, n$, the OLS estimate of the parameter vector

$$\boldsymbol{\beta} = (\beta_{0,1}, \ldots, \beta_{0,k}, \ldots, \beta_{m,1} \ldots, \beta_{m,k}) \tag{3.3.8}$$

in (3.3.7), requires working with a full $(m+1)k \times (m+1)k$ system of equations $\mathbf{B}^{\mathrm{T}}\mathbf{B}\boldsymbol{\beta} = \mathbf{B}^{\mathrm{T}}\boldsymbol{Y}$ where

$$[\mathbf{B}]_{i,j} = B_j(X_i), \quad 1 \le i \le n \text{ and } 1 \le j \le (m+1)k\,, \tag{3.3.9}$$

and B_j denotes the basis function in (3.3.7) corresponding to the jth coefficient in the parameter vector $\boldsymbol{\beta}$ (3.3.8).

We are now in a position to define polynomial splines. As mentioned before, splines are elements of $\mathcal{PP}_k(\boldsymbol{t})$ that satisfy certain smoothness or continuity constraints. Each spline space will take as its basis a subset of the elements in (3.3.7). With each breakpoint t_l, we associate an integer s_l that counts the number of continuity restrictions enforced across t_l. By setting $s_l = 0$ we allow a jump at t_l: $s_l = 1$ specifies continuity at t_l; $s_l = 2$ provides one continuous derivative; $s_l = 3$ results in a continuous second derivative at t_l; and so on. Any value of s_l larger than $k-1$ forces the polynomial pieces on either side of t_l to be the same, removing the breakpoint. To avoid such degeneracies, we require that $0 \le s_l \le k-1$. Set $s = (s_1, \ldots, s_m)$ and let $\mathcal{S}_k(\boldsymbol{t}, s) \subset \mathcal{PP}_k(\boldsymbol{t})$ be such that, for $g \in \mathcal{S}_k(\boldsymbol{t}, s)$,

$$g(x) = \beta_{0,1} + \beta_{0,2}x + \cdots + \beta_{0,k}x^{k-1} + \sum_{l=1}^{m}\sum_{j=1}^{k-s_l} \beta_{l,j}(x-t_l)_+^{j+s_l-1} \tag{3.3.10}$$

Aside from a renumbering of the coefficients, the difference between this expression and (3.3.7) is that we have left out the first s_l basis functions, $(x-t_l)_+^j$, $0 \le j \le s_l - 1$, associated with the breakpoint t_l. We refer to $\mathcal{S}_k(\boldsymbol{t}, s)$ as a space of *polynomial splines*, and the breakpoints $\boldsymbol{t}$ are commonly called *knots*.

We recover the space of piecewise polynomials by choosing $s_l = 0$ for $1 \le l \le m$. For a fixed k and $\boldsymbol{t}$, the "smoothest" spline space corresponds to setting $s_l = k-1$ for $1 \le l \le m$. Because this specification is extremely popular in statistical applications, we write $\mathcal{S}_k(\boldsymbol{t})$ for this special case. The spaces $\mathcal{S}_k(\boldsymbol{t})$ for $k = 2$, 3 and 4 are known, respectively, as linear splines, quadratic splines and cubic splines. Because of their importance in this book we write out explicitly the truncated power basis for the space of linear splines with knots $\boldsymbol{t}$

$$1, x, (x-t_1)_+, \ldots, (x-t_m)_+\,.$$

The cubic spline space with knot sequence $\boldsymbol{t}$ has a basis of the form

$$1, x, x^2, x^3, (x-t_1)_+^3, \ldots, (x-t_m)_+^3\,.$$

Cubic splines are frequently used in penalized regression problems, where a penalty is placed on the "roughness" of the spline curve (Wahba, 1990; O'Sullivan, 1987; Eilers and Marx, 1997; and Ruppert, Carrol and Wand, 1999). We will return to these so called *smoothing splines* in Section **??**.

In terms of the flexibility or approximation power of splines, we observe that a spline space performs identically to a space of piecewise polynomials when the underlying function f is sufficiently smooth. Thus, for functions f with three continuous derivatives, $\mathrm{dist}(f, \mathcal{PP}_4(\boldsymbol{t}))$ is essentially the same as $\mathrm{dist}(f, \mathcal{S}_4(\boldsymbol{t}))$. The extra degrees of freedom associated with the discontinuities across the break points do not help us when approximating a smooth curve (and will only add to extra variance when it comes to our OLS estimates). In the pages that follow, it will be useful to have the approximation rate for a very simple example: Assume f has p continuous derivatives over the interval $\mathcal{X} = [a, b]$. Then for $k > p$, the approximation rate for the spline space $\mathcal{S}_k(\boldsymbol{t})$ associated with a vector $\boldsymbol{t} = (t_1, \ldots, t_m)$ of m equally spaced knots in $[a, b]$ is given by

$$d(f, \mathbb{G}) = C\Big(\frac{1}{m}\Big)^p, \tag{3.3.11}$$

where m is the number of knots and C is a positive constant that depends only on properties of the pth derivative of f and $b - a$ (technically, we require further conditions on the pth derivative of f, but we leave these until Chapter 11). For piecewise polynomials, we saw a similar result in (3.3.4), Jackson's Inequality. The intersted reader can find more results of this kind in Chapter 2 as well as Schumaker (1981).

3.3.2 Model error for fixed-knot splines

Using the approximation rates described in the previous section, we now revisit the bias-variance tradeoff for regression using approximation spaces, focusing our discussion on splines. We assume that the unknown regression function f has p continuous derivatives over the interval $\mathcal{X} = [a, b]$ and that we will form an estimate from the linear space $\mathcal{S}_k(\boldsymbol{t})$, where $\boldsymbol{t}$ is a vector of equally spaced knots over $[a, b]$. Substituting the rate (3.3.11) this into (3.2.23), we have

$$\frac{1}{n}\sum_{i=1}^{n} E\Big(\big[f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i; \widehat{\boldsymbol{\beta}})\big]^2\Big) \leq C^2\Big(\frac{1}{m}\Big)^{2p} + \frac{(m+k)\sigma^2}{n},$$

where we recall that $\mathcal{S}_k(\boldsymbol{t})$ is an $(m + k)$-dimensional space. The tension between bias and variance is now clear: the more knots we use, the better the approximation, but the larger the variance. A simple heuristic is that for equally spaced knots, the bias decreases like a power of $1/m$, while the variance depends roughly on m/n, the average number of points per interval separating adjacent pairs of knots. We refer to this quantity as the *span* of the interval.

We can strike a balance between the two effects, bias and variance, by minimizing the model error with respect to m. Differentiating with respect to m and solving leads to

$$m = \left[2pC^2\left(\frac{n}{\sigma^2}\right)\right]^{\frac{1}{2p+1}} \sim n^{\frac{1}{2p+1}} .$$

Substituting this value into the bias-variance decomposition, we find that

$$\frac{1}{n}\sum_{i=1}^{n} E\Big(\big[f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i;\widehat{\boldsymbol{\beta}})\big]^2\Big) \sim n^{-\frac{2p}{2p+1}} . \tag{3.3.12}$$

Clearly, as we collect more data (as n gets larger), we should entertain larger and larger knot sequences; we do not believe that f belongs to any given spline space, but we entertain more and more complex representations as our sample size grows. This is consistent with our notion of a *nonparametric* problem mentioned earlier.

In Chapter 11, we will see that this *rate of convergence* holds in much more general situations. Assuming $X_1, \ldots, X_n$ are independent and identically distributed observations from some density on $[a, b]$, then it is possible to bound

$$E_{X,Y}\Big(\big[f(X) - g(X;\widehat{\boldsymbol{\beta}})\big]^2\Big) ,$$

in probablility, where X, Y are new observations independent of the data we used in our OLS estimates. The bound is essentially the one derived in (3.3.12), but requires more analysis. This kind of asymptotic study has been an important motivation for many methodological innovations using splines. In a series of papers, Stone (1980,1982, 1985, 1986) demonstrates rates of this kind for multivariate problems and in so doing opened the door for techniques like those illustrated in Section 3.1.2 on the importance values for the American beech.

We will return to this historical motivation in Section 3.4.2 when we take up multivariate problems. Before leaving this topic, however, we should note that the results of this section, and in particular (3.3.12) and the more sophisticated rate result alluded to above, were derived with fixed knot sequences; that is, knots simply fill up the domain $[a, b]$ evenly in some sense. In the earliest methodologies, this was also the case; knots were placed either evenly or some prior knowledge led researchers to specific configurations. In addition to the regression examples cited above, Stone and Koo (1986a,b) applied a similar strategy for logistic regression and density estimation. An important long-term goal of the development of theoretical results like those in Stone (1980,1982,1985,1986) for nonadaptive methodologies was to motivate the equally challenging development and implementation of corresponding practically useful adaptive methodologies and to make such methodologies seem less ad hoc. Conversely, the development and implementation of such adaptive methodologies has motivated the further development of theories for nonadaptive versions of these and similar methodologies.

In the next section, we will explore the use of adaptively placed knots, letting the data indicate areas needing more or less flexibility. Theoretical results for this case have been derived in Huang and Stone (2002), and are also presented in Chapter 12.

3.3.3 Adaptive knot placement

Having introduced splines through a regression-modeling framework, we now consider the issue of knot placement in more detail. We begin with some notation. First, we rewrite the least squares criterion (3.2.9) making $\boldsymbol{t}$ explicit:

$$\rho(\boldsymbol{t},\boldsymbol{\beta}) = \sum_{i=1}^{n} [\, Y_i - g(X_i;\boldsymbol{\beta})\,]^2 \quad \text{for} \quad g(x;\boldsymbol{\beta}) \in \mathcal{S}_k(\boldsymbol{t})\,, \tag{3.3.13}$$

where for simplicity we take $\boldsymbol{t} = (t_1, \ldots, t_m)$ and $t_1 < t_2 < \cdots < t_m$. Now, we know for any fixed $\boldsymbol{t}$ that the OLS estimate in $\mathcal{S}_k(\boldsymbol{t})$ minimizes $\rho(\boldsymbol{t},\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. By substituting the corresponding coefficient estimate $\widehat{\boldsymbol{\beta}}$ into this expression, we derive a criterion that depends only on $\boldsymbol{t}$:

$$\rho(\boldsymbol{t}) = \sum_{i=1}^{n} \left[Y_i - g(X_i;\widehat{\boldsymbol{\beta}}\,)\right]^2 \qquad \text{for} \quad g(x;\widehat{\boldsymbol{\beta}}\,) \in \mathcal{S}_k(\boldsymbol{t})\,. \tag{3.3.14}$$

In statistical terminology, we might consider the elements of $\boldsymbol{\beta}$ to be nuisance parameters and view $-\rho(\boldsymbol{t})$ as a kind of *profile log-likelihood* for $\boldsymbol{t}$; our interest focuses on finding $\boldsymbol{t}^* = \operatorname{argmin}_{\boldsymbol{t}}\, \rho(\boldsymbol{t})$. In the numerical analysis literature, this set up is referred to as curve fitting or approximation with *free-knot splines* because the breakpoint locations are free parameters.

As you add more predictors to an OLS fit, the residual sum of squares decreases. Therefore, we cannot use RSS to compare spline spaces; this metric will always favor larger models. Instead, following the discussion in Section 3.2.3, we want to identify a knot sequence $\boldsymbol{t}$ that minimizes a model selection criterion. For a given penalty $p(n)$, the generalized AIC criterion is simply

$$-\log \mathrm{RSS}(\boldsymbol{t}) + p(n)J\,,$$

where, if a knot sequence $\boldsymbol{t}$ has m elements, the dimension of $\mathcal{S}_k(\boldsymbol{t})$ is just $J = k + m$. For each fixed dimension, we are again looking for the knot sequence with the smallest residual sum of squares. Unfortunately, these kinds of problems are quite difficult numerically. For a fixed number of knots, the criterion $\rho(\boldsymbol{t})$ depends nonlinearly on $\boldsymbol{t}$ and in fact is known to have local minima: for any two free knots $\rho(\boldsymbol{t}) = \rho(t_1, t_2)$ is symmetric along any normal to the line $t_1 = t_2$, and hence has zero derivative there (recall the discussion of repeated knots at the end of Section **??**). Standard optimization techniques will have difficulty with this kind of problem.

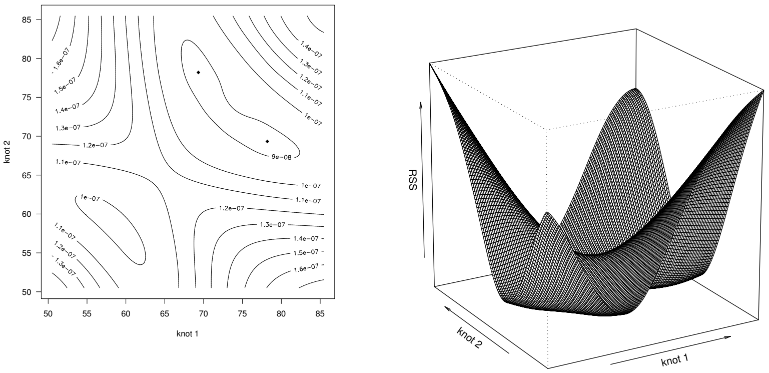

FIGURE 3.10. Fitting a two-knot cubic spline model. Contour and perspective plots of RSS($\boldsymbol{t}$) for $\boldsymbol{t} = (t_1, t_2)$. The minimum is marked with with black points on the contour plot.

To make this more concrete, in Figure 3.10 we present RSS($\boldsymbol{t}$) for just two knots $\boldsymbol{t} = (t_1, t_2)$ using $\mathcal{S}_4(\boldsymbol{t})$, or cubic splines. The data are again the $^{87}\delta\,$Sr values. From the contour plot, we notice that the RSS surface is in fact symmetric. The black points correspond to the knot sequence with the minimum RSS, namely $\widehat{t} = (69.3, 78.2)$. The value obtained is 8.8×10^{-8}. A pair of local minima occur at $\boldsymbol{t} = (56.7, 60.2)$ with a residual sum of squares value of 9.9×10^{-8}. Even with just two knots, we get a sense of how difficult this optimization problem can become. Several suggestions have been made to improve the numerical properties of free-knot splines through some kind of penalization. This regularization also tends to improve the statistical performance of the estimator as well. We will discuss two such schemes in Section **??**.

For the rest of this section, we will consider strategies designed only approximately to minimize $\rho(\boldsymbol{t})$. In particular we consider stepwise approaches that solve the problem one knot at a time. The reader will notice similarities with stepwise methods for variable selection in classical linear models. These methods were used to create the fits in Figures 3.3 and 3.4 at the beginning of the chapter.

Stepwise addition

As its name suggests, under this method, we introduce knots sequentially, placing each knot so that it produces the greatest drop in RSS. In Figure 3.11 we illustrate this process for cubic splines. Consider the top curve in the leftmost plot. For each value of t_1, we plot $\rho(t_1) = \text{RSS}(t_1)$, the residual sum of squares associated with the model

$$g(x;\widehat{\boldsymbol{\beta}}) = \widehat{\beta}_{0,1} + \widehat{\beta}_{0,2}x + \widehat{\beta}_{0,3}x^2 + \widehat{\beta}_{0,4}x^3 + \widehat{\beta}_{1,2}(x-t_1)_+^3 \,,$$

where $g(x;\widehat{\boldsymbol{\beta}}) \in \mathcal{S}_4(t_1)$ and $\widehat{\boldsymbol{\beta}}$ is the OLS estimate of $\boldsymbol{\beta} = (\beta_{0,1},\ldots,\beta_{1,2})$. Here we have used the truncated power basis representation of a cubic spline space; this construction was used to motivate stepwise knot placement as early as Smith (1982). In fact, the work in Smith (1982) inspired many later polynomial spline algorithms, such as TURBO (Friedman and Silverman, 1988), MARS (Friedman 1990), and the Polyclass, Polymars, Logspline, LSPEC, and Triogram algorithms discussed in Chapters 5–9.

As t_1 runs through the range of the data (from 40 to 100 million years ago) we sweep out a smooth curve in $\mathrm{RSS}(t_1)$. Given the form of the function g and the OLS criterion, this curve has to have two continuous derivatives in t_1. The best location for a single knot is marked in Figure 3.11 with a vertical line and labeled “knot 1.” Let $\widehat{t}_1$ denote this point. The single-knot model is shown in the top panel of the right column of the figure. Next, we set $\boldsymbol{t} = (\widehat{t}_1, t_2)$ and now define $\mathrm{RSS}(t_2)$ to be the residual sum of squares corresponding to the model

$$\widehat{\beta}_{0,1} + \widehat{\beta}_{0,2}x + \widehat{\beta}_{0,3}x^2 + \widehat{\beta}_{0,4}x^3 + \widehat{\beta}_{1,2}(x-\widehat{t}_1)_+^3 + \widehat{\beta}_{2,2}(x-t_2)_+^3$$

when $t_2 \neq \widehat{t}_1$ and

$$\widehat{\beta}_{0,1} + \widehat{\beta}_{0,2}x + \widehat{\beta}_{0,3}x^2 + \widehat{\beta}_{0,4}x^3 + \widehat{\beta}_{1,2}(x-\widehat{t}_1)_+^3 + \widehat{\beta}_{2,2}(x-\widehat{t}_1)_+^2 \qquad (3.3.15)$$

otherwise. Keep in mind that when knots are repeated, we lower the degree of smoothness across the breakpoint. In both cases, $\widehat{\boldsymbol{\beta}}$ is the OLS estimate of $\boldsymbol{\beta} = (\beta_{0,1},\ldots,\beta_{2,2})$.

The second curve (dashed) in Figure 3.11 sweeps out $\mathrm{RSS}(t_2)$, which again appears quite smooth. The minimizing point, $\widehat{t}_2$, is marked “knot 2” and the associated fit is in the second panel at the right in this figure. The new knot sequence is now taken to be $(\widehat{t}_1, \widehat{t}_2, t_3)$, and we look for the single addition that drops the residual sum of squares the most. Figure 3.11 illustrates this process, sequentially adding 5 knots for the $^{87}\delta\,\mathrm{Sr}$ data. The curve in Figure 3.3 at the beginning of the chapter was constructed in this way, where the largest fit had 10 knots. Most readers will recognize this scheme as a variant of the classical stepwise addition of variables for building regression models. While a simple idea, it has been used successfully in non-parametric regression schemes like that of Smith (1982), Friedman and Silverman (1989), Friedman (1990), Luo and Wahba (1995) and Stone et al. (1997).

The smoothness of the $\rho(\cdot)$ curves in Figure 3.11 should not be surprising since the criteria inherit their smoothness from the underlying spline spaces.

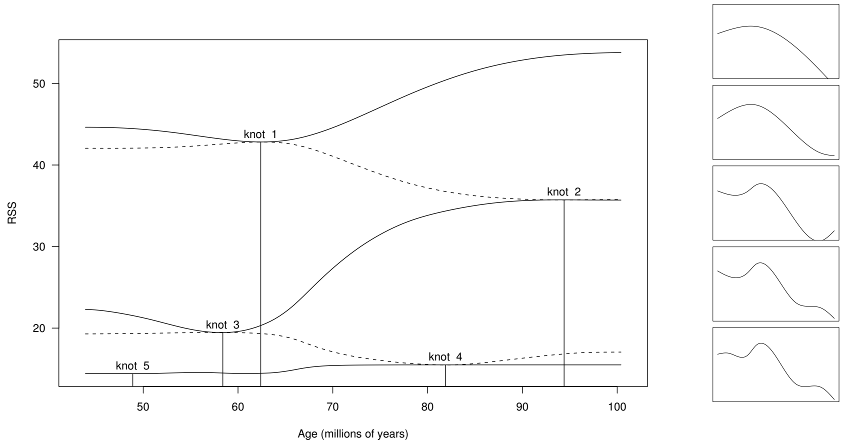

FIGURE 3.11. Illustrating the process of stepwise addition. Each curve sweeps out the residual sum of squares for a model with one free knot location. The five plots on the right correspond to adding knots 1 through 5, respectively.

For the first knot t_1 we have

$$\begin{aligned}\rho(t_1) &= \sum_{i=1}^{n} \left[Y_i - g(X_i;\widehat{\boldsymbol{\beta}})\right]^2 \\ &= \sum_{i=1}^{n} \left[Y_i - \widehat{\beta}_{0,1} + \widehat{\beta}_{0,2}X_i + \widehat{\beta}_{0,3}X_i^2 + \widehat{\beta}_{0,4}X_i^3 + \widehat{\beta}_{1,2}(X_i - t_1)_+^3\right]^2 .\end{aligned}$$

Since functions $g \in \mathcal{S}_4(t_1)$ have two continuous derivatives, so does the profile log-likelihood $\rho(t_1)$, with breaks in the third derivative at each of the data points $X_1, \ldots, X_n$. Given this smoothness, we can apply various heuristic schemes to locate the optimal value of t_1. For example, we can initially evaluate $\rho(\cdot)$ at several points in the interval $\mathcal{X}$ and further refine our search based on a smooth fit to these values. This strategy is employed in the context of density estimation discussed in Chapter 6. After the first knot $\widehat{t}_1$ is added, the curve for RSS(t_2) will again be smooth. Although it is not immediately obvious from our use of truncated powers and the change in basis in (3.3.15), RSS(t_2) still has two continuous derivatives in t_2. The reader is referred to Chapter 2 or Schumaker (1981) for these and other properties of spline spaces.

By considering just one knot at a time, we have sidestepped some of the numerical difficulties in working with the full sequence $\boldsymbol{t}$. Of course, it is unlikely that a sequential approach to determining $\boldsymbol{t}$ will actually identify the knots $\boldsymbol{t}^*$ that minimize $\rho(\boldsymbol{t})$. In Chapter 10, we study the tradeoffs involved in terms of bias, variance and computation time.

Stepwise deletion

Continuing our analogy with stepwise procedures for building regression models, we recall that forward addition is not the only strategy for uncovering interesting structures in a data set. Backward deletion from a larger model also proves to be a useful scheme. In the spline context, suppose we have a knot sequence $\boldsymbol{t} = (t_1, \ldots, t_m)$ that does not contain any repeated knots. The cubic spline associated with $\boldsymbol{t}$ can be written as

$$1, x, x^2, x^3, (x - t_1)_+^3, \ldots, (x - t_m)_+^3 \,, \tag{3.3.16}$$

in terms of the truncated power basis. Since each knot is associated with a single basis function, we can treat the individual spline terms (the truncated monomials) individually and drop them from the model just as we would predictor variables in an ordinary linear regression setup. In this case, we would consider eliminating the terms that create the smallest increase in the residual sum of squares. The DKCV (delete knot, cross-validate) procedure of Breiman (1990) starts with a rich set of knots, removing them one at a time.

If any of the knots in $\boldsymbol{t}$ were repeated, then we have an expansion of the form (3.3.7) in terms of the truncated power basis. In this case, we want to be careful during stepwise deletion to remove the a term of the form $(x - t)_+^{j_1}$ before $(x - t)_+^{j_2}$ when $j_1 < j_2$. This hierarchy makes sense given how we have constructed our spline spaces. Removing terms $(x - t)_+^j$ in the order of j means we are reintroducing smoothness across the breakpoint beginning with discontinuities in the lowest derivatives first. This is just one case in which the analogy with variable selection is not to be taken too literally; it is only a device to simplify the computations.

As another example, given the basis in (3.3.16), we should not apply stepwise deletion blindly and remove any of the pure monomial terms while there are still truncated basis elements in the model. In Chapter 2 and the previous subsection, we have motivated an ordering of approximation spaces in terms of their flexibility (their ability to capture important features of a function). In terms of a range of models achievable by variable addition and deletion from the cubic splines, this ordering is $\mathcal{P}_0 \subset \mathcal{P}_2 \subset \mathcal{P}_3 \subset \mathcal{P}_4 \subset \mathcal{S}_4(\boldsymbol{t})$ for any knot sequence $\boldsymbol{t}$; that is, we begin with linear, quadratic and cubic polynomials and then finish with any cubic spline model. Therefore, if we view variable addition and deletion as tools for constructing candidate approximation spaces, we need to restrict the selection process. We should only consider candidate "alterations" that make sense from the standpoint of increasing or decreasing the approximation power of the underlying linear space. (For a discussion about the pitfalls of not following this hierarchy, the reader is referred to Stone et al., 1997.)

```
Specify an initial spline model
Stepwise addition: Continue until the maximum model size is reached or
  no candidates can be found
    Decide which basis functions are candidates for addition
    Add the best candidate
    Fit the model
    Evaluate the model
    If the model is better than the best previous one, save it
Stepwise deletion: Continue until the minimum model size is reached
    Decide which basis function can be removed from the model
    Remove the one that is the worst predictor
    Fit the model
    Evaluate the model
    If the model is better than the best previous one, save it
Output the best fit
```

FIGURE 3.12. Adaptive knot placement via stepwise addition and deletion.

A simple recipe

Passes of stepwise addition and deletion can be used in concert to produce a series of nested fits. In Figure 3.12 we present a simple strategy for adaptively placing knots in this way. The algorithm is worded somewhat generally because we will use the same prescription several times in the course of the text in much more elaborate modeling situations. Model evaluation can be based on one of the selection criterion discussed in Section 3.2.3 or some other appropriate measure. Candidate models for both addition and deletion involve the basic hierarchy between polynomial models and splines discussed above. For the final model, the “best fitting linear space,” in Figure 3.3, we added knots sequentially until we reached a model of size 10. We then removed knots one at a time, producing a chain of 19 models. The evaluation criterion BIC (3.2.32) was used to decide which of the 19 models “best” fit the data. The entire process is plotted in Figure 3.13.

The simple recipe outlined here is nearly identical to that used to produce the fit given in Figures 3.3 and 3.13. In Section 3.6, we will see that we can control the variance of our adaptive estimate in two ways; one is by assigning a minimum number of data points between consecutive knot points. We have chosen 3 for this example, meaning that at each stage in the addition process we are only searching in those locations where a new knot has at least three points separating it and a knot already in the model. Such points are “viable candidates” in the recipe in Figure 3.12. In Section 3.6 will also see that boundary constraints can help reduce variance in our estimate outside the support of our data. The fit in Figures 3.3 and 3.13 uses so-called

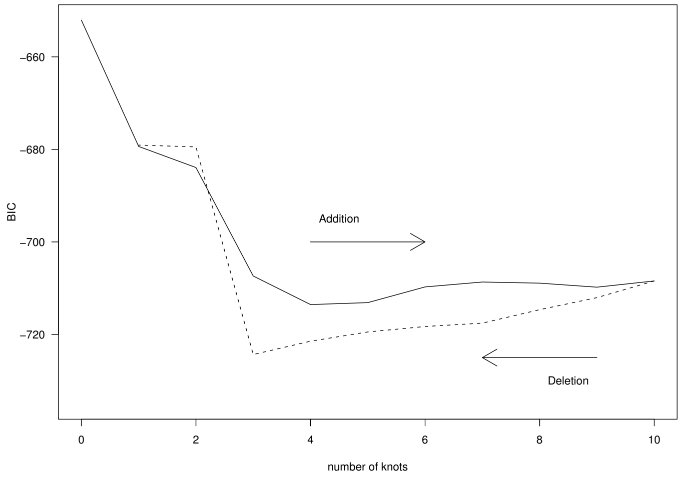

FIGURE 3.13. Stepwise addition and deletion for the $^{87}\delta\,$Sr data using (natural) cubic splines. The evaluation criterion BIC is plotted against the number of knots in the spline space.

tail-linear constraints at the smallest and largest data points in the $^{87}\delta\,$Sr data. This changes our spline space from the ordinary cubic splines to the *natural splines.* These alterations are spelled out in Section 3.6, but do not really change the general methodology. The basic ingredients – stepwise addition, stepwise deletion, and then model selection – are at the heart of several popular spline-based techniques, including MARS (multivariate adaptive regression splines) by Friedman (1990) and PolyMARS described in Bose, Kooperberg and Stone (1997).

Amazingly, this recipe can be applied in a wide range of estimation contexts and not just regression. First, the initial model can be a simple spline space with a modest number of knots (as is the case in Chapter 5 for the Logspline density estimator) or perhaps a polynomial space (constant functions are used in the PolyMARS algorithm in the next subsection). The "evaluation" steps in both the addition and deletion phases do not necessarily require refitting models with the candidate alterations. In the regression context, there are reasonably efficient schemes for computing the change in RSS without first finding the new OLS fit. For more elaborate modeling situations, however, some form of iteration is required to fit a spline model (recall the Newton–Raphson iterations in Chapter 2). In Chapter 4, we will derive simple Taylor expansions that allow us to approximate the change in the objective function (usually a log-likelihood) without computing new parameter estimates. These methods really get their power from the concavity of the likelihood as a function of the coefficients of the spline basis elements. For stepwise addition, we obtain a series of Rao statistics for the candidate variables, while deletion is based on a collection of Wald statistics. Later in this chapter we discuss various computational schemes for regression, illustrating the tradeoff between our choice of basis and the speed of computations.

Finally, the algorithm in Figure 3.12 is designed to produce a single fit. Even if it were computationally feasible to find the best spline fit according to some selection criterion, we would still want to examine a few "reasonable" models as well. Naturally, the same is true for automated variable selection routines from standard regression analysis (Mallows, 1973). In Chapter 10, we examine Bayesian and other sampling-based methods for generating several good-fitting models, highlighting the tradeoff between simplicity of interpretation and predictive performance.

3.4 Multivariate models

In this section, we consider various methods for estimating a regression function $f(\boldsymbol{x})$, originally defined in (3.2.1), where the vector $\boldsymbol{x} \in \mathbb{R}^d$ is comprised of one or more candidate predictors $\boldsymbol{x} = (x_1, \ldots, x_d)$. Data analysts will appreciate the fact that this is a challenging task even when the vector of predictors is of modest size (consisting of, say, 4 or 5 variables), much less when there are 28 candidate predictors as in the case of the importance values in our tree species example. Despite advances in graphical methods, it can be difficult even to visually assess the relationship between inputs and outputs, much less capture the dependence mathematically. In part, our ability to ascertain the structure in multivariate problems is hampered by what Bellman (1961) referred to as the *curse of dimensionality.* In the 40 years since it was originally described, the curse has been broadly applied to a number of difficulties that scale exponentially with dimension. When estimating a function of several variables, we will contend with several consequences of Bellman's observation and propose an adaptive strategy partially to overcome these difficulties.

3.4.1 From multivariate polynomials to splines

Multivariate polynomials

The simplest regression model for the d predictors is given by

$$g(\boldsymbol{x};\boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + \cdots \beta_d x_d,$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)$. If diagnostic plots or some other model assessment tool suggest that this model is inadequate, we might consider adding higher order terms and possibly interactions between the variables: In the first case, we add monomial terms like x_1^2 or x_1^3, while in the latter, we entertain products like x_1x_2. Our statistical modeling approach naturally leads us to consider spaces of *multivariate polynomials*, which are linear combinations of product terms of the form

$$x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d} \,. \tag{3.4.1}$$

Multivariate polynomials also arise in approximation theory. The reader is no doubt familiar with multivariate versions of Taylor expansions for smooth functions. In that context, we might consider how closely we can approximate a given smooth function with multivariate polynomials of a given order. For example, the space of multivariate polynomials with (coordinate) order at most k is made up of all terms (3.4.1) where $k_l < k$, $l = 1, \ldots, d$, and hence it has dimension k^d.

As with the univariate polynomials in Section 3.3, increasing k increases the flexibility of the space, an effect we can quantify through the associated approximation rate achievable for smooth functions. Unfortunately, the complications of working with high-order polynomials carry over from the univariate to the multivariate case as well. In fact, things become much worse. The wild excursions we saw in Section 3.3 when there were gaps in the input data are even more prevalent for a multivariate domain, simply because data tend to appear more sparse in higher-dimensional settings. To make this precise, suppose that the vector $\boldsymbol{X}$ of predictors is uniformly distributed over a d-dimensional unit cube. Take as our gap a subcube of side-length δ. The chance that in a sample of size n, no points fall into this gap is $(1-\delta^d)^n$. For 1,000 points, we can expect a hole with $\delta = 0.5$ with probability essentially zero for dimensions $d = 1, \ldots, 5$. However, for $d = 12$ the probability is 0.78, and it is practically one for $d = 20$ and larger. This is one manifestation of Bellman's curse.

We encounter a second form of the curse when we work with multivariate polynomials with different orders k. As noted earlier, the space of polynomials with (coordinate) order at most k has dimension d^k. For even modest orders, the number of parameters balloons as we increase k. Aside from perhaps deriving theoretical approximation rates, using a global polynomial of a fixed order is simply not practical. This is precisely why most statisticians will instead be familiar with modeling schemes that introduce monomials and products of monomials in a sequential fashion; adding interactions or higher order terms a few at at time. As it turns out, this kind of stepwise approach will help to offset the curse. Before developing this idea further, we first introduce some basic concepts for working with multivariate functions.

Tensor products of linear spaces and simple ANOVA representations

In moving from univariate to multivariate polynomials, we considered products of monomials in the different variables. This operation is a simple building block for constructing multivariate approximation spaces from univariate ones. Formally, we consider products of basis functions from each univariate space, and form what is known as the *tensor product.* Consider linear spaces $\mathbb{G}_1, \ldots, \mathbb{G}_d$ of dimensions $J_1, \ldots, J_d$ in the variables $x_1, \ldots, x_d$, respectively. Let $\mathbb{G}_l$ have basis functions $B_{l,1}, \ldots, B_{l,J_l}$. Then we define $\mathbb{G}_1 \otimes \cdots \otimes \mathbb{G}_d$ to be the space of all functions of the form

$$g(\boldsymbol{x}) = g(x_1, \ldots, x_d) = \sum_{l_1=1}^{J_1} \cdots \sum_{l_d=1}^{J_d} \beta_{l_1,\ldots,l_d} B_{1,l_1}(x_1) \cdots B_{d,l_d}(x_d), \quad (3.4.2)$$

where we collect the coefficients into a vector $\boldsymbol{\beta} = (\beta_{1,1}, \ldots, \beta_{d,J_d})$. The dimension of this space is $J_1 \cdots J_d$.

Suppose each of the spaces $\mathbb{G}_1, \ldots, \mathbb{G}_d$ consists of polynomials of order k. For example, let $\mathbb{G}_1$ have the basis $1, x_1, x_1^2, \ldots, x_1^{k-1}$ and do the same for the other spaces. Then the tensor product $\mathbb{G}_1 \otimes \cdots \otimes \mathbb{G}_d$ consists of the polynomials of coordinate order k given in (3.4.4). Before considering constructions based on piecewise polynomials or splines, we first introduce a general decomposition of tensor product spaces using ideas from classical regression modeling.

Suppose each $\mathbb{G}_l$ contains the constant function, and in fact, let $B_{l,1} = 1$ for $l = 1, \ldots, d$. By analogy with our polynomial models, we can pull out terms from (3.4.2) and set

$$g_1(x_1) = \sum_{l_1=2}^{J_1} \beta_{l_1,1,\ldots,1} B_{1,l_1}(x_1)\,.$$

We can do the same for the other variables, defining $g_l(x_l)$ for all $1 \le l \le d$. Put another way, each of these terms involve only one variable. We can then proceed to collect terms that involve only two variables. For x_1 and x_2, for example, we can set

$$g_{1,2}(x_1, x_2) = \sum_{l_1=2}^{J_1} \sum_{l_2=2}^{J_2} \beta_{l_1,l_2,1,\ldots,1} B_{1,l_1}(x_1) B_{2,l_2}(x_2)\,.$$

If we proceed in this way, we can re-express each function $g \in \mathbb{G}_1 \otimes \cdots \otimes \mathbb{G}_d$ in the tensor product space (3.4.2) as a sum of terms

$$\begin{aligned} g(\boldsymbol{x}) = g_0 + \sum_{j_1} g_{j_1}(x_{j_1}) + \sum_{j_1<j_2} g_{j_1,j_2}(x_{j_1}, x_{j_2}) \\ + \sum_{j_1<j_2<j_3} g_{j_1,j_2,j_3}(x_{j_1}, x_{j_2}, x_{j_3}) + \cdots\,, \quad (3.4.3) \end{aligned}$$

where we take $g_0 = \beta_{1,\ldots,1}$, the constant function.

The expansion in (3.4.3) is similar to constructions found in the classical analysis of variance for contingency tables; hence we refer to it as an *ANOVA decomposition* for g. It has a natural interpretation in terms of main effects and interactions. It also builds nicely on our intuition from regression modeling with polynomials. Throughout this book, we will see various expansions of this form in both applications and theoretical work.

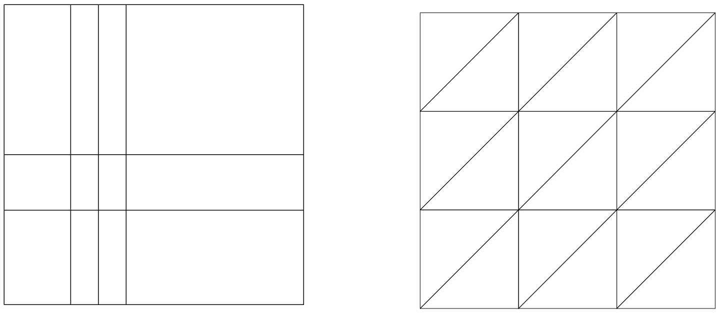

FIGURE 3.14. Left: Breaklines for a piecewise polynomials space with knots $\boldsymbol{t}_1 = (0, 0.2, 0.3, 0.4, 1)$ (horizontal axis) and $\boldsymbol{t}_2 = (0, 0.3, 0.5, 1)$ (vertical axis). Right: An alternate construction of "pieces" for a bivariate, piecewise polynomial.

Piecewise polynomials and splines

In curve estimation, we found that knot placement controls the flexibility of piecewise polynomial and spline spaces. For example, by introducing knots near the KTB we improved our ability to resolve the peak in ${}^{87}\delta\,\mathrm{Sr}$. For multivariate models based on tensor products, flexibility is inherited from the separate spaces $\mathbb{G}_1, \ldots, \mathbb{G}_d$. We have seen that tensor products of univariate polynomials of order k give rise to multivariate polynomials of coordinate order k. For piecewise polynomials, assume that each x_j ranges over an interval $[a_j, b_j]$ and let $\boldsymbol{t}_j = (t_{j,0}, \ldots, t_{j,m_j+1})$ denote a knot sequence dividing the interval into $m_j + 1$ subintervals $\mathcal{X}_{j,l} = [t_{j,l-1}, t_{j,l})$. Then, using the expression in (3.3.5) for the space of piecewise polynomials, we find that the tensor product space is made up of polynomials of the form

$$x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d} \qquad \text{for } k_j < k \tag{3.4.4}$$

on each interval $\mathcal{X}_{j_1,\ldots,j_d} = \mathcal{X}_{1,j_1} \times \cdots \times \mathcal{X}_{d,j_d}$. In short, the "pieces" associated with this space of multivariate piecewise polynomials are hyperrectangles. In the lefthand panel of Figure 3.14, we take $d = 2$ and exhibit the rectangular regions corresponding to $\boldsymbol{t}_1 = (0, 0.2, 0.3, 0.4, 1)$ and $\boldsymbol{t}_2 = (0, 0.3, 0.5, 1)$.

When moving from piecewise polynomials to splines in Section 3.3, we introduced constraints that forced the functions to join smoothly at the breakpoints. For tensor product models, this means that the lines in Figure 3.14 no longer delineate regions in which we fit separate multivariate polynomials of order k. Instead, they indicate places where surfaces in $\mathbb{G}_1 \otimes \mathbb{G}_2$ can have discontinuities in various partial derivatives. For example, if $\mathbb{G}_1$ and $\mathbb{G}_2$ consist of cubic splines with knot sequences $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$

specified above, then for $g \in \mathbb{G}_1 \otimes \mathbb{G}_2$ the partial derivatives

$$\frac{\partial^{k_1+k_2}}{\partial x_1^{k_1} \partial x_2^{k_2}} g(x_1, x_2)$$

are continuous for $k_1, k_2 < 3$. However, the third partial derivatives in each variable have discontinuities along the lines indicated in Figure 3.14. Clearly our notion of knots representing increased flexibility locally is somewhat different for tensor products.

In Section 3.3 we found that we could increase the approximation power of polynomials either by increasing their order or by breaking the domain into separate pieces and fitting a relatively low-order polynomial in each piece. The same is true in a multivariate setting. Using the tensor product construction, we can either take $\mathbb{G}_1, \ldots, \mathbb{G}_d$ to be polynomial spaces and increase the coordinate order k in (3.4.4); or we could take each space to consist of piecewise polynomials with breaks at $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_d$ and let the number of pieces (equivalently, the number of breakpoints) increase. To make this precise, let $f = f(x_1, \ldots, x_d)$ be a function of d input variables and assume that each x_i has the same domain, namely, $x_i \in [a, b]$. For some integer p, assume that f is p times continuously differentiable on $\mathcal{X} = [a, b]^d$. Now, if we let each $\mathbb{G}_i$ be a spline space $\mathcal{S}_k(\boldsymbol{t})$ with m equally spaced knots, where $k > p$, then the approximation rate for $\mathbb{G} = \mathbb{G}_1 \otimes \cdots \otimes \mathbb{G}_d$ is given by

$$\operatorname{dist}(f, \mathbb{G}) \le C\left(\frac{1}{m}\right)^p, \tag{3.4.5}$$

just as it was in the univariate case (3.3.11). (This result actually requires extra conditions on certain partial derivatives of f; we leave these technicalities to Chapter 11.) Note that, as also true for univariate splines, this result holds for $\mathcal{S}_k(\boldsymbol{t})$ as well as for $\mathcal{PP}_k(\boldsymbol{t})$ assuming f is sufficiently smooth; that is, the extra degrees of freedom in a piecewise polynomial space do not help when approximating f.

The use of hyper-rectangles to define the structures in a multivariate piecewise polynomial or spline fit is by no means the only option. In the literature on approximation theory, there are numerous examples of alternative constructions. For bivariate predictors, $d = 2$, one might consider triangles, or for $d = 3$, simplicies. An example of such a structure is given on the right in Figure 3.14. While it is fairly easy to conceive of techniques for fitting with piecewise polynomials defined over triangles, it is harder to envision a simple construction for splines. Forcing the different pieces to join smoothly along the lines in Figure 3.14 happens automatically for tensor products and the rectangular mesh; for triangles the process becomes harder. At this point, we merely hint at the idea that there are many more multivariate approximation spaces than those based on tensor products. We take up the topic in more detail in Chapter 9.

3.4.2 Model error and functional ANOVA

We now quantify the impact that the curse of dimensionality has on the estimation of multivariate functions. First, recall the bias-variance trade-off defined in Section 3.2.2. In analyzing curve estimation, we found that the variance component of the model error for a spline estimate depended roughly on the average span, the average number of data points separating each knot: The larger the span, the lower the variance. We will show that the same result holds for $d \geq 1$ when we use tensor products, but now we consider the size of hyper-rectangles defined by the knot sequences. The curse confounds our ability to estimate a function of several variables accurately by making even moderately large data sets appear sparse. At the beginning of this section, we noted that randomly distributing design points in the unit cube leaves (subcube) gaps with large edge lengths with high probability for even moderately sized d. Therefore, to achieve roughly the same variance for n data points, we have to consider models involving relatively fewer "pieces" as the dimension d increases.

This observation can be made rigorous by examining the model error attainable by splines (or any other nonparametric method). Assuming the standard regression setup (3.2.3) with unknown regression function f, recall the result from Proposition 3.2.1:

$$\mathrm{ME}(\mathbb{G}) = \frac{1}{n}\sum_{i=1}^{n} E\Big(\big[f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i;\widehat{\boldsymbol{\beta}})\big]^2\Big) \leq d^2(f,\mathbb{G}) + \frac{J\sigma^2}{n}\,. \qquad (3.4.6)$$

Let's take each $\mathbb{G}_i$ to be a spline space of order k with m equally spaced knots and define $\mathbb{G}$ to be the tensor product space $\mathbb{G}_1 \otimes \cdots \otimes \mathbb{G}_d$. Then we know that the approximation rate is $C(1/m)^p$. We also recall that the dimension of $\mathbb{G}$ is just the product of the dimensions of $\mathbb{G}_1, \ldots, \mathbb{G}_J$, which in this case is $(m+k)^d$. Therefore, substituting these values into (3.4.6), we find that

$$\mathrm{ME}(\mathbb{G}) \leq C^2\Big(\frac{1}{m}\Big)^{2p} + \frac{(m+k)^d\sigma^2}{n} \sim \Big(\frac{1}{m}\Big)^{2p} + \frac{m^d}{n}\,. \qquad (3.4.7)$$

The last expression is a minimum if we take $m \sim n^{1/(d+2p)}$, and then we have

$$\mathrm{ME}(\mathbb{G}) \sim n^{\frac{-2p}{2p+d}}\,. \qquad (3.4.8)$$

The impact of the dimension of our input space is now clear. To put this in perspective, if we assume $p = 2$, then to achieve the same rate of decay for ME when estimating a univariate function with $n = 500$ data points, we would need $n^2 = 250K$ data points for a function of $d = 6$ variables, and $n^3 = 125M$ points for a function of $d = 11$ variables—a clear sign of the curse at work.

We can partially overcome this problem by borrowing intuition from the decomposition in (3.4.3). In regression analysis, it is common to consider only models that consist of, say main effects and pairwise interactions. Suppose we truncate the expansion (3.4.3) and consider only terms involving at most 2 variables; to be precise, we set equal to zero all the coefficients $\beta_{l_1,\ldots,l_d}$ for which more than 2 of $l_1, \ldots, l_d$ are not equal to one. As a modeling tool, this space still retains a certain degree of flexibility. In addition, we are able display the main effects and the interactions using standard plotting routines for curves and surfaces. This means we can rely on common graphical tools to explore the model and diagnose possible misfit. The gains in interpretability certainly help ameliorate the curse of dimensionality from a practical standpoint.

In fact, there are also theoretical gains to be had by considering only low-order interactions. Suppose we truncate (3.4.3) to include only main effects. This is often referred to as an *additive model.* As we will see in Chapter 11, the model error associated with this reduced space is given by

$$\mathrm{ME}(\mathbb{G}) = \frac{1}{n}\sum_{i=1}^{n} E\Big(\big[f^*(\boldsymbol{x}_i) - g(\boldsymbol{x}_i;\widehat{\boldsymbol{\beta}})\big]^2\Big) \sim n^{\frac{-2p}{2p+1}}\,, \tag{3.4.9}$$

where the target of our estimation procedure is no longer f, the unknown regression function, but the "best" approximation f^* to f of the form

$$f^*(\boldsymbol{x}) = f_0^* + \sum_{j=1}^{d} f_j^*(x_j)\,. \tag{3.4.10}$$

Comparing this expression to (3.3.12), we find that, in terms of model error, estimating f^* is no harder than estimating a function of just one variable. That is, we do not have the same explosion in data required to estimate this function as we did for f itself. Of course, we cannot guarantee that f^* is a reasonable approximation to f, so it is common to include higher-order terms in the expansion. In general, if we truncate (3.4.3) to include all l-factor interactions, then

$$\mathrm{ME}(\mathbb{G}) = \frac{1}{n}\sum_{i=1}^{n} E\Big(\big[f^*(\boldsymbol{x}_i) - g(\boldsymbol{x}_i;\widehat{\boldsymbol{\beta}})\big]^2\Big) \sim n^{\frac{-2p}{2p+l}}\,, \tag{3.4.11}$$

where f^* is the best approximation to f of the form

$$\begin{aligned} f^*(\boldsymbol{x}) = f_0 \;+ \sum_{\text{1-factor terms}} f^*_{j_1}(x_{j_1}) \;+ \sum_{\text{2-factor terms}} f^*_{j_1,j_2}(x_{j_1}, x_{j_2}) \\ + \cdots + \sum_{l\text{-factor terms}} f^*_{j_1,\cdots,j_l}(x_{j_1}, \ldots, x_{j_l}). \end{aligned}$$

In Chapter 11 these results will be spelled out in much more detail. For the moment, we mention them to indicate that in addition to the practical value in considering only a portion of a tensor product space, there is a considerable theoretical justification.

In our discussion of splines, we have alternated between data analysis, methdology and theory. In fact, the development of asymptotic theory in the context of spline modeling has fueled a great deal of the innovative methodological work. First, Stone (1980, 1982) quantified the curse of dimensionality for the problem of function estimation; that is, the *optimal* pointwise (squared error) or global (integrated squared error) is shown to be of the form $n^{-2p/(2p+d)}$, where p is the number of bounded derivatives assumed on the regression function and d is the number of covariates. In other words, no matter what method we might use to estimate f, this is the best rate we can expect to achieve. In Stone (1982), the possibility was raised that if the regression function is the sum of p-times differentiable functions in the individual covariates, then the optimal rate of convergence would be $n^{-2p/(2p+1)}$, which would ameliorate the curse of dimensionality. This possibility was verified in Stone (1985). In this paper the regression function was modeled as the sum of polynomial splines as in (3.4.10).

In Stone (1986), the results in Stone (1985) were extended to logistic regression, Poisson regression and other concave generalized linear models. Here the canonical regression function was modeled as the sum of polynomial splines in the individual covariates, with maximum likelihood being used to fit the unknown parameters. Further theoretical results involving polynomial splines are found in Stone (1989, 1990, 1991b), Hansen (1994) and Huang (1998, 1998, 2001), among others. These results are discussed in detail in Chapters 11 and 12. As mentioned earlier, many of these theoretical results, including the simpleminded bound in (3.4.11) are derived for spline spaces based on essentially equally-spaced knots. In more recent work, this restriction has been relaxed, and in Huang and Stone (2002) properties of free-knot splines are discussed. In the next section, we explore adaptation for multivariate regression problems, where we let the data guide us not only in choosing knots, but also which terms in the ANOVA expansion to include.

3.4.3 Adaptation for multivariate splines

At this point, spline modeling seems to involve a dizzyingly large number of choices, from both the number and location of knots on individual variables to the kinds of interactions we should include in the ANOVA-style decomposition. In some sense, this is another manifestation of Bellman's curse: in modeling multivariate functions, we have an exponentially increasing number of decisions. Performing an exhaustive search for models among all the possibilities is simply infeasible. To sidestep some of these issues, we look to common statistical practice. When faced with a large number of variables, we often opt for some kind of stepwise modeling approach, adding predictors one at a time to optimize an overall selection criterion

	Estimate	SE	t	P-value
(Intercept)	2.6229	0.5012	5.233	.000
AVGT	−0.3884	0.0242	−16.047	.000
PPT	0.0012	0.0003	3.707	.000
JARPPET	−0.7193	0.2114	−3.403	.001
PERM	−0.1227	0.0211	−5.824	.000
CEC	−0.0375	0.0069	−5.449	.000
ROCKDEP	0.0540	0.0067	8.074	.000
SLOPE	0.0570	0.0062	9.169	.000
PROD	0.1648	0.0354	4.660	.000
ALFISOL	−0.0055	0.0012	−4.610	.000
SPODOSOL	−0.0072	0.0018	−3.939	.000
MOLLISOL	−0.0110	0.0027	−4.121	.000
MAXELV	−0.0008	0.0002	−5.056	.000
MINELV	−0.0018	0.0004	−4.240	.000

TABLE 3.3. Simple linear fit to the IV data.

like BIC (3.2.32). Elaborations of a simple linear fit

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$$

might include interactions or higher order terms, again added sequentially according to a selection criterion. As we will see, the same kind of adaptation can take place in the context of spline modeling. While this recipe reduces the complexity of our search for good-fitting spline models, we are certainly not guaranteed to find the "best" set of choices (knots, locations, interactions). Again, we have made a compromise to undercut the curse of dimensionality.

To motivate a general purpose strategy for modeling with multivariate splines, we reconsider the IV values for the American beech introduced previously. Recall that in the FIA dataset, we have 28 possible predictor variables. As was done at the beginning of the chapter, we still work with the square root of IV and focus our attention on 1093 counties in the middle region of the eastern half of the US. To see how we will introduce spline elements into this multivariate setting, we first fit a simple regression model in which all of the covariates appear linearly. In this fit, 13 of the 28 variables are not statistically significant (individually at the 5% level). From here, simple backward deletion was performed to trim off irrelevant predictors. A final model was chosen using the BIC criterion (3.2.32).

In Table 3.3, we present some simple summary statistics from this fit. Note that three of the climatic variables (AVGT, PPT and JARPPET) remain after backward deletion. While the R^2 value for this simple fit is rather low, just 41%, this value is not out of the range for the 80 or so tree species considered by Iverson and Prasad (1999). To assess systematic deficiencies in the model, we examined several diagnostic plots. In Figure 3.15,

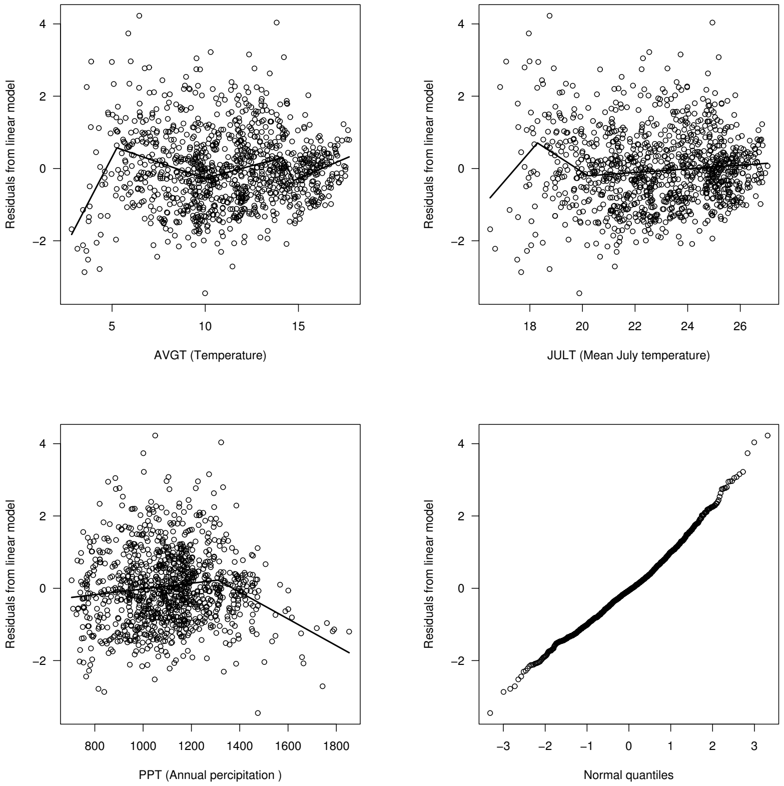

FIGURE 3.15. Diagnostic plots associated with a simple linear fit to the IV values computed for the American Beech.

we present residual plots from the model in Table 3.3 against three climatic variables that Prasad and Iverson (2000) found to be important for predicting abundance of the American beech (two of the variables, JULT and PPT, being among the factors remaining after backward deletion). We also include a normal quantile plot of the residuals. The lines in the three plots were fit using the stepwise addition and deletion scheme outlined in the previous section, but with continuous, piecewise linear functions. We include them as crude indicators of regions where the model is missing possible structure.

Additive models

Starting from these scatter plots, we might consider model elaborations that introduce extra flexibility associated with each variable.

For example, it is common to add higher order monomials like quadratics in, say, a stepwise-forward manner. Simple diagnostics like those in

Figure 3.15 can be used to guide the process, and perhaps a model selection criterion like BIC should be consulted as a more formal benchmark. Broadly, this kind of approach might take us from a simple linear model

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d \tag{3.4.12}$$

for $f(\boldsymbol{x}) = f(x_1, \ldots, x_d)$ to a more general expansion of the form (3.4.3) involving only main effects; that is, the additive model

$$g_0 + g_1(x_1) + \cdots + g_d(x_d), \tag{3.4.13}$$

where g_0 is just a constant.

To ground our discussion leading to (3.4.13), we explore how spline functions like those in the diagnostic plots of Figure 3.15 can be used in a general procedure for fitting additive models. Here we return to the algorithm in Figure 3.12 but consider several input variables at once. As with simple curve estimation, we want to make sure we add spline basis elements for a variable, terms of the form $(x_k - t)_+$, only after we have added the simple variable x_k itself. We can formally express these constraints in terms of the candidate basis functions available for each step of addition

- x_k, $k = 1, \ldots, d$; and
- $(x_k - t_{k,m})_+$ if x_k is already a basis function in the model.

We performed this addition process with the IV data starting from a constant fit, and proceeding until we had added a total of 25 terms.

Then, following the recipe in Figure 3.12, we performed backward deletion. Borrowing from our experience with curve estimation, we want to impose some restrictions on the deletion process, making sure we remove all the spline elements for a variable x_k before we remove the simple term x_k itself. Again, we can formally word this in terms of the candidate basis functions at each step of the deletion process

- $(x_k - t_{k,m})_+$; and
- x_k if there are no terms of the form $(x_k - t_{k,m})_+$ in the model.

From our 25-term additive fit to the IV data, we then conduct backward deletion down to a constant fit. We then select the best model according to BIC. When we do this, we find functional dependencies on three of the climatic variables plotted in Figure 3.16. As indicated in the discussion of Figure 3.15, these three variables were all found to be important by Prasad and Iverson (2000). For the most part, these curves are all consistent with the missing structures found in the residual plots of Figure 3.15. (They are also very similar to those in the final model plotted in Figure 3.6 earlier in the chapter.)

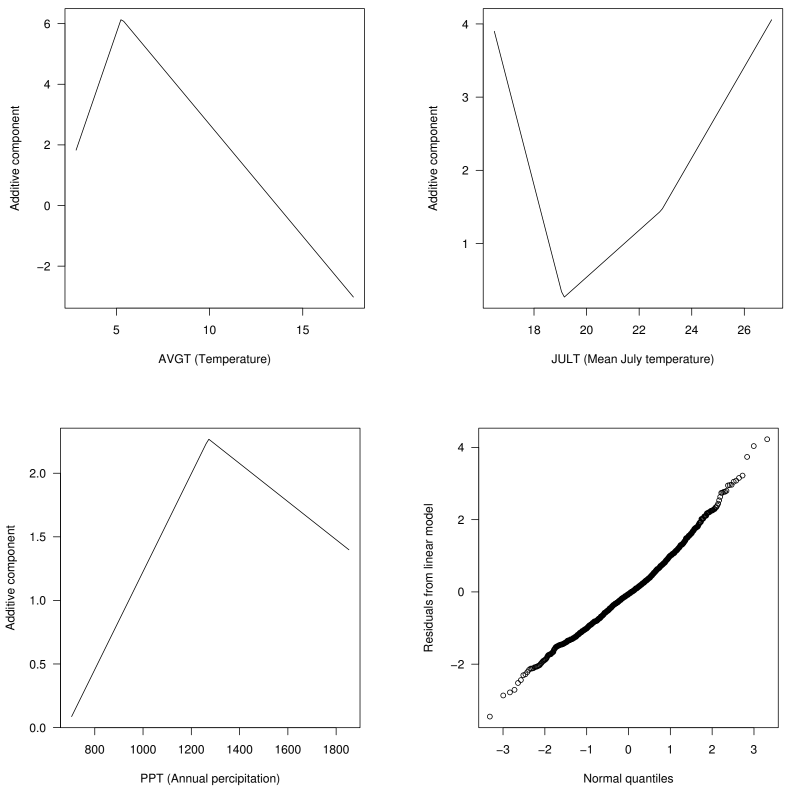

FIGURE 3.16. Functional components (based on the climate variables) in the additive fit to the square root of the IV values computed for the American Beech.

Two-factor interactions

Finally, we generalize this procedure one step further. Given a significant lack of fit in a model of the form (3.4.12), we might consider adding interactions between the covariates. Borrowing from this idea, we consider a model that allows for interactions involving not only linear terms, but also some of the nonlinear elements plotted in Figure 3.6. We can think of moving from the simple linear model

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$$

to a more general expansion involving two-factor interactions; that is, truncating (3.4.3) to include only

$$g_0 + g_1(x_1) + \cdots + g_d(x_d) + \sum_{k_1 < k_2} g_{k_1,k_2}(x_{k_1}, x_{k_2})\,. \tag{3.4.14}$$

Just as with additive models, we have to determine which interactions are important and which are not.

We now present a simple spline procedure that fits this kind of model. We again build on the algorithm in Figure 3.12. In this case, the set of candidates at each stage involve a hierarchy of linear terms, spline terms and products. We impose this hierarchy to create a tensor product space of splines. Therefore, we want to maintain the order of linear terms x_k before spline terms $(x_k - t)_+$. In addition, we have to impose a constraints on the kinds of interactions that can be included at each step.

To be precise, the candidate basis functions are:

- x_k, $k = 1, \ldots, d$;
- $(x_k - t_{k,m})_+$ if x_k is already a basis function in the model;
- $x_{k_1} x_{k_2}$ if x_{k_1} and x_{k_2} are already basis functions in the model;
- $x_{k_1}(x_{k_2} - t_{k_2,m})_+$, if $x_{k_1} x_{k_2}$ and $(x_{k_2} - t_{k_2,m})_+$ are in the model;
- $(x_{k_1} - t_{k_1,m_1})_+(x_{k_2} - t_{k_2,m_2})_+$ if $x_{k_1}(x_{k_1} - t_{k_1,m_1})$ and $(x_{k_1} - t_{k_1,m_1})_+ x_{k_2}$ are in model.

In short, we add the linear effects (monomials) of a variable first, then spline terms; interactions are entered first with the simple linear effects (monomials) and then between splines and monomials and finally between spline elements of a different variable.

For the IV data, we added terms until the model included 25 terms and then performed backward deletion. In Table 3.4 we display the individual terms in the model and their complexity. All but one of the interactions are between linear effects, and these would have been found by operating as usual with a linear model. The one complex interaction is between temperature and elevation.

Predictions from the IV models

Using the models fitted so far, we now consider different scenarios for climate change in North America. As we did at the beginning of the chapter, we generate IV values for the American beech using the same county-by-county values for the soil factors, but replacing the climate variables with predictions with one of two different models for weather conditions associated with a doubling of CO_2 levels. The three rightmost images in the top row of Figure 3.18 are based on the Hadley predictions, while the lower images are based on data from the CCC. The three two-image columns correspond to predictions from the simple linear regression (left), the additive spline model (middle) and the interaction spline model (right).

As done previously, we quantify the drift northward of regions with high IV values through the percent change in an area-weighted IV score for each of the fitted models under the two climate scenarios. The entries in

		Interactions				
	Term	SLOPE	CEC	ROCKDEP	INCEPTSL	MINELV
AVGT	spline	1				2
JULT	spline	1			1	
PPT	spline					
TAWC	linear					
PERM	linear					
CEC	linear					
ROCKDEP	linear					
SLOPE	linear					
PROD	linear				1	
TEXFINE	spline					
ROCKFRAG	spline		1	1	1	
ALFISOL	linear					
INCEPTSL	linear					
MOLLISOL	linear					
MINELV	spline					

TABLE 3.4. Spline fit to the American beech IV values. The second column describes whether the variable enters linearly or as a spline function. The remaining columns describe the interactions. A “1” denotes an interaction between linear terms and a “2” denotes an interaction with a spline term.

Table 3.5 correspond to the 2×3 grid of images on the right in Figure 3.18. Under the interaction spline model of Table 3.4 we recognize the 35% drop in the area-weighted IV scores under the Hadley model and the 68.5% drop under the CCC predictions.

Scenario	Linear	Additive Spline	Interaction Spline
Hadley	54.9%	69.0%	65.0%
CCC	18.4%	18.4%	31.5%

TABLE 3.5. Ratio of the area-weighted IV score under each climate scenario using the three statistical models for the area-weighted IV score for the raw data.

3.5 A survey of multivariate spline methods

3.6 Properties of spline estimates

3.6.1 *Knot spacing*

The bias-variance discussion in Section 3.2.2 also has implications for knot spacing. In Figure 3.19, we present plots of the pointwise bias and variance obtained from simple OLS estimates based on regularly spaced knot sequences. The input values are 100 equally spaced points in the interval $[0, 10]$, and the response is generated via

$$Y = 2.5 \left[\cos(\pi x) + \phi(20x - 12.5)/6 \right] + \epsilon \tag{3.6.1}$$

where ϕ is the standard normal density function and ϵ is normally distributed with mean 0 and $\sigma^2 = 1$. The individual panels summarize the

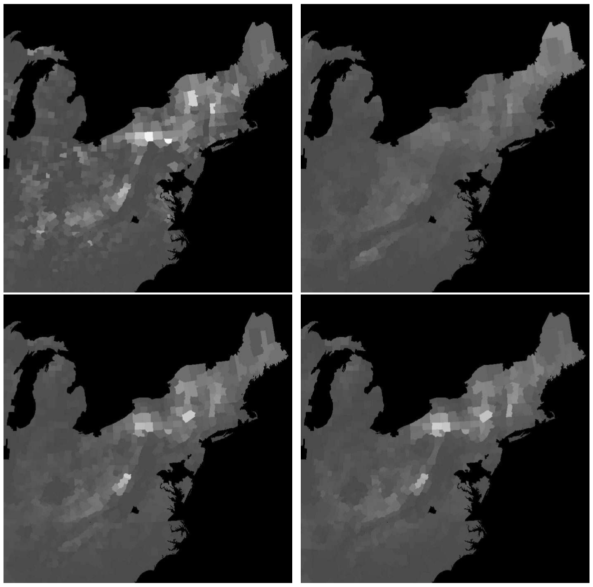

FIGURE 3.17. Predicted abundance of the American beech using four different models. Upper left is the raw data; upper right is the simple linear fit; lower left is the additive linear-spline model; and lower right is the interaction spline model.

pointwise behavior of estimates based on spaces of cubic splines, where the knots are indicated by vertical lines. The plots in each row have the same vertical axis, so that we can directly compare how the (squared) bias (left column) and variance (right column) behave as the number of knots increases. Because the number of points between knots is constant, the pointwise variances are roughly constant across the interval. Increasing the number of knots reduces the squared bias at the peak ($x = 6.25$) from 4.0 to 0.4, while the variance increases from 0.04 to 0.1. Away from the peak, the bias drops more quickly and is eventually overtaken by the variance.

By making a more careful accounting of the hidden constants in expressions like (3.2.23) for spline models, Agarwal and Studden (1980) show that the optimal distribution of knots (in terms of mean squared error) is related to both the derivatives of f (its local roughness, coming from the local approximation error given by Jackson's Inequality) and the placement of the inputs $\{X_i\}$. The simulation results of Figure 3.19 support this result. Consider the fit with three knots (the middle row of Figure 3.19). We have paid a huge price, in terms of squared bias, for not locating the breakpoints near the peak. When knots are placed at 5.0, 6.25 and 7.5, the maximum squared bias is a full order of magnitude smaller than it is for the arrangement in Figure 3.19. This is a problem in general with nonadaptive knot placement schemes.

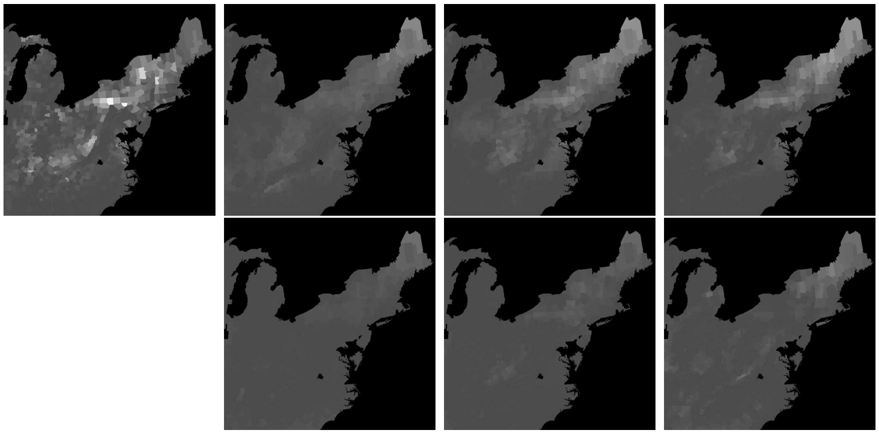

FIGURE 3.18. Abundance values for the American beech under two different climate scenarios.

When the underlying curve is sufficiently "regular" in some sense, the nonadaptive knot placement schemes are generally not too bad provided that we have a rich-enough collection of knots. For example, the fit represented at the bottom of Figure 3.19 smoothes over the peak at $x = 6.25$, leading to a large squared bias at that point. After experimenting with different knot locations, we found that any single-breakpoint model fails to capture some feature in the data, so that the resulting fits all have essentially the same overall squared bias. While our immediate goal is to describe how knot spacing affects the bias-variance tradeoff for a spline model, implicit in this discussion is the understanding that any adaptive scheme must determine *how many* knots to use as well as where they should be placed.

Friedman and Silverman (1989) and Friedman (1990) characterize the effect of knot spacing on the variance (3.2.23) by examining the sensitivity of the estimate to "runs" of positive or negative errors. To explain their reasoning, write the linear regression model in the form

$$Y_i = f(X_i) + \epsilon_i \quad \text{for } 1 \leq i \leq n, \tag{3.6.2}$$

where we now assume that the errors ϵ_i have a symmetric distribution. Let $\mathbb{G}$ be the space of linear splines defined for some knot sequence $\boldsymbol{t}$, and let $g(x; \widehat{\boldsymbol{\beta}})$ denote the OLS estimate of (3.2.12). The fitted curve will be *resistant* to a series of L consecutive positive or negative errors ϵ_i provided that the knots are adequately separated. If not, the function $g(x; \widehat{\boldsymbol{\beta}})$ will follow the run and exhibit spurious features.

Think of any curve $g(x; \boldsymbol{\beta}) \in \mathbb{G}$ as interpolating the data points $t_l, g(t_l, \boldsymbol{\beta})$ with a broken line (where $1 \leq l \leq m$). Viewed in this way, it should be clear that changing the value of $g(t_l, \boldsymbol{\beta})$ affects the curve only in the neighboring intervals $[t_{l-1}, t_l]$ and $[t_l, t_{l+1}]$. As a result, if a series of positive errors

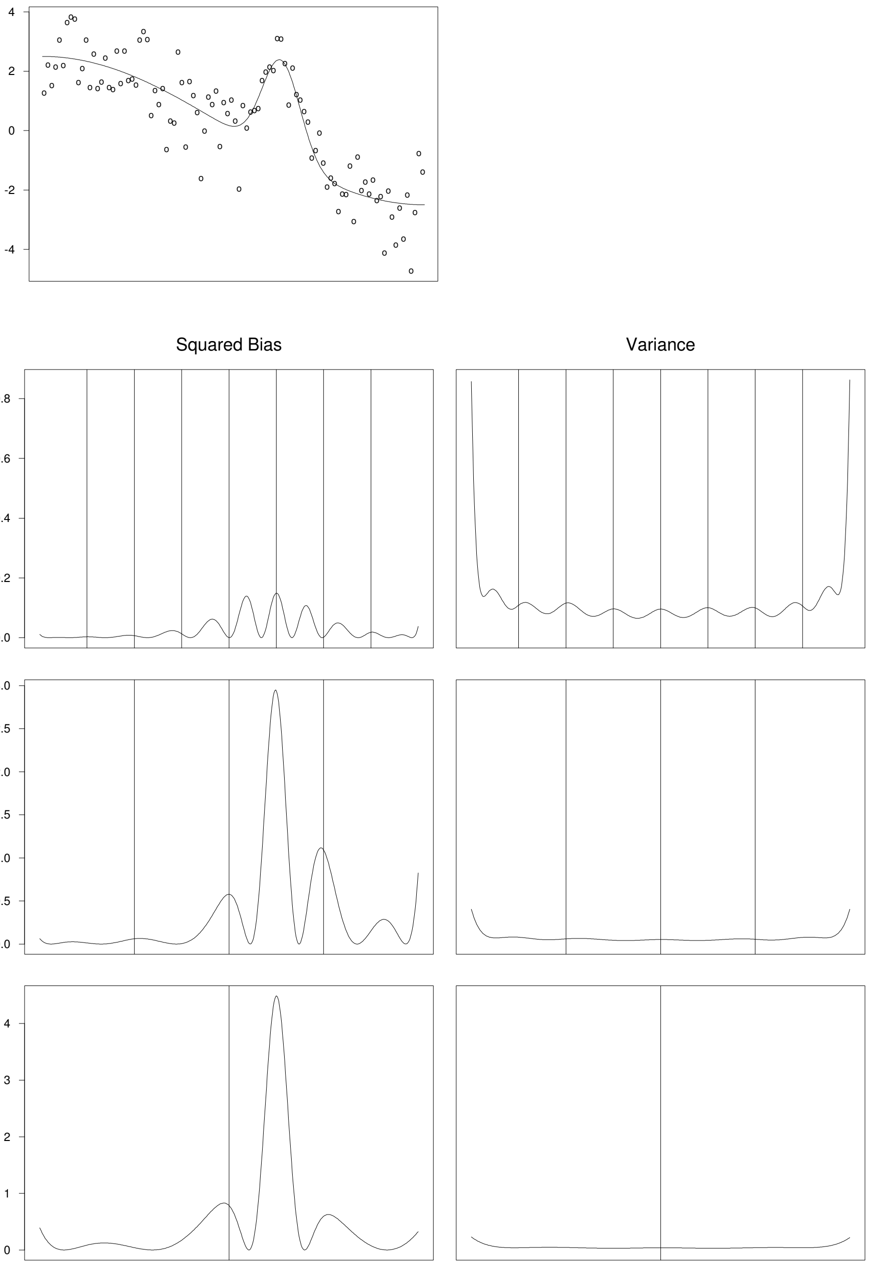

FIGURE 3.19. The bias-variance tradeoff and knot spacing. The upper leftmost graph illustrates a sample data set, with the true regression function plotted as a solid curve. The input data are equally spaced and $n = 100$. In the lower 3 rows, we plot the pointwise squared bias (left column) and the pointwise variance (right column) of OLS estimates using $\mathcal{S}_3(\boldsymbol{t})$, where the knot points $\boldsymbol{t}$ are marked by vertical lines. Each curve is based on 1,000 simulations.

covers (at least) the three knots t_{l-1}, t_l, and t_{l+1}, the fitted curve $g(x;\widehat{\boldsymbol{\beta}})$ will track the run and overshoot f in a neighborhood of t_l. It is free to do so because the fit in other regions is not influenced by changes near t_l. As a new sample from (3.2.3) will be just as likely to have a similar run of negative errors, we see that such effects add to the local variability of $g(x;\widehat{\boldsymbol{\beta}})$. Depending on the overall knot configuration, the same problems can occur (but to a lesser degree) when a run traps only two knots.

The distribution of run lengths L (assuming a symmetric error distribution) is easily derived by a simple coin tossing argument. Let L_α denote the $1-\alpha$ quantile of L, which we can approximate by

$$L_\alpha = -\log_2\Big[-\frac{1}{n}\ln(1-\alpha)\Big]$$

for $\alpha < 0.1$ and $n > 10$. Friedman and Silverman (1989) suggest that the knots for a linear spline fit should be positioned with (at minimum) $L_\alpha/3$ data points between neighbors, where L_α represents an improbably long run length. To be conservative, Friedman and Silverman (1989) recommend separating knots in a linear spline fit with

$$M = M(n,\alpha) = L_\alpha/2.5\,. \tag{3.6.3}$$

data points, where α is some value between 0.05 and 0.01. This prescription should provide resistance to all but the most infrequent conspiracies of the error process. We refer to M as the *minimal span* acceptable for a spline smooth. Using the fact that splines of order k have support on k neighboring knot intervals, we can easily extend this argument to quadratic and cubic splines as well.

3.6.2 Boundary conditions

Our discussion has been restricted to approximation spaces defined on a finite interval $[a,b]$. In many statistical applications, this assumption is unnatural, and we would like to make predictions beyond the endpoints a and b. Let g be our original spline function and denote by $\tilde{g}$ its extrapolation to $\mathbb{R}$. In the simplest scheme, we carry the leftmost and rightmost polynomial pieces of g into $(-\infty,a]$ and $[b,\infty)$, respectively. In terms of (3.3.5), this becomes

$$\tilde{g}(x) = \begin{cases} g_0(x), & x < t_0; \\ g(x), & a \le x \le b; \\ g_{m+1}(x), & x > t_{m+1}. \end{cases} \tag{3.6.4}$$

As we have seen in Figure 3.19, the variance of our estimate grows as we approach the boundary. Predictions beyond the interval $[a,b]$ suffer from the same variance inflation.

A surprisingly effective alternative to (3.6.4) can control both effects at once. Rather than extend g with cubic polynomials, consider linear

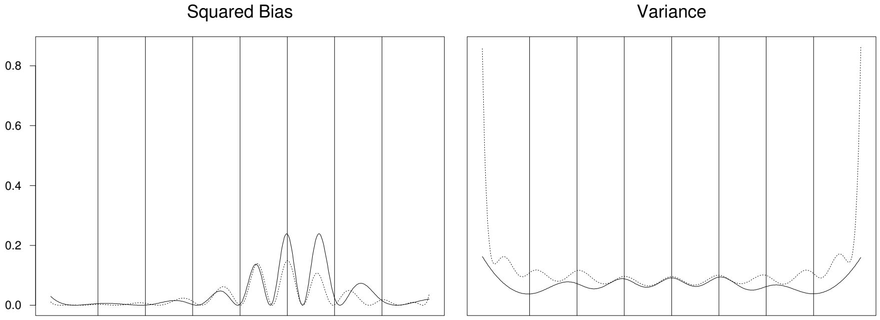

FIGURE 3.20. Quantifying the effect of the natural spline boundary conditions.

extrapolation outside of $[a, b]$ — a so-called *tail-linear constraint.* To be more precise, replace the cubic polynomials in $(-\infty, a]$ and $[b, \infty)$ with linear functions and add the condition that $\tilde{g}$ blend smoothly across the boundary points a and b. We achieve this by imposing the constraints $\tilde{g}''(a) = \tilde{g}''(b) = 0$ and $\tilde{g}'''(a) = \tilde{g}'''(b) = 0$. The resulting linear space consists of the so-called *natural cubic splines.* By extrapolating with linear functions, we reduce the variance of predictions made outside the central interval $[a, b]$, while enforcing the smoothness constraint serves to reduce the variance near the boundaries within $[a, b]$. To see this in action, we have repeated a portion of the simulations described in Figure 3.19. The two panels of Figure 3.20 illustrate the impact that the boundary conditions have had on our fit. The natural spline boundary conditions are applied to the leftmost and rightmost knots, so our fits are linear in the leftmost and rightmost intervals (as marked by vertical lines). Clearly, the large edge effects in the variance plot (right) are brought down nearly to the level inside the interval, while the squared bias (left) has seen a slight increase near the right boundary. In general, the extra bias is to be expected as the boundary conditions reduce the approximation rate from $\overline{\Delta}^4$ (**??**) across the interval $\mathcal{X}$ down to $\overline{\Delta}^2$ near the edges (de Boor, 1978). Because we are interested in balancing squared bias and variance, we are willing to pay a small price for the overall improved accuracy of the constrained fit.

Given that we are imposing 4 constraints on the coefficients (two on each of the second and third derivatives at a and b), unless we have at least three knots, our solution space will consist of only the linear functions $\mathcal{P}_2$. In many applications, it is common to place knots at the boundary points $t_0 = a$ and $t_{M+1} = b$, in which case we can think of placing at least $1 \leq M$ knots inside $[a, b]$. For example, take $M = 1$ so that $\boldsymbol{t} = (a, t, b)$. In terms of the truncated power basis, any function $g \in \mathcal{S}_4(\boldsymbol{t})$ can be written as

$$g = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \gamma_3 x^3 + \beta_0 (x-a)_+^3 + \beta_1 (x-t)_+^3 + \beta_2 (x-b)_+^3 \,.$$

Enforcing the boundary conditions at a has us drop x^2 and x^3 from the basis set. The two conditions at b then involve the coefficients β_0, β_1 and β_2 of the truncated polynomials:

$$\beta_0 + \beta_1 + \beta_2 = 0 \quad \text{and} \quad \beta_0 + \beta_1(1-t) = 0\,.$$

Solving these equations yields that we can write any natural cubic spline g^* in the form

$$\begin{aligned} g^* &= \gamma_0 + \gamma_1 x + \beta\left[(1-t)(x-a)^3_+ (x-t)^3_+ - t(x-b)^3_+\right] \\ &= \gamma_0 + \gamma_1 x + \beta h(x,t)\,, \end{aligned}$$

where $h(x,t)$ is a twice continuously differentiable function of t. Therefore, we see that the space of natural cubic splines with knots at $\boldsymbol{t} = (a,t,b)$ with tail-linear constraints enforced at the boundaries of the interval $[a,b]$ differs from the space of linear functions by the addition of a single degree of freedom. We will encounter this again in Section 3.7.3.

The use of tail-linear constraints to control for the variance at the boundaries of the data appears in Stone and Koo (1986a,b), Friedman and Silverman (1988), and Breiman (1993).

3.6.3 Degrees of freedom associated with knot placement

In Section 3.2.3 we discussed model selection criteria for comparing linear spaces. We introduced a penalty $p(n)$ that could be used to add an additional cost for each adaptively selected basis function. To make this clear, consider $\mathcal{S}_2(t)$, the space of continuous, piecewise linear functions with just a single knot t. Assume that our inputs are restricted to the interval $[0,1]$. Given data points $(X_1,Y_1),\ldots,(X_n,Y_n)$, we can define the residual sum of squares $\mathrm{RSS}(t)$ for any knot point t by

$$\mathrm{RSS}(t) = \sum_{i=1}^{n}\left[Y_i - g(X_i;\widehat{\boldsymbol{\beta}})\right]^2 \quad \text{for } g(x;\widehat{\boldsymbol{\beta}}) \in \mathcal{S}_2(t)\,,$$

where $\widehat{\boldsymbol{\beta}}$ is the OLS estimate of $\boldsymbol{\beta}$ in $\mathcal{S}_2(t)$. We let $\widehat{t}$ denote the breakpoint for which $\mathrm{RSS}(t)$ is a minimum. Next, let RSS_0 denote the residual sum of squares corresponding to the space $\mathcal{P}_2$, the collection of polynomials that are linear in x:

$$\mathrm{RSS}_0 = \sum_{i=1}^{n}\left[Y_i - g_0(X_i;\widehat{\boldsymbol{\beta}}_0)\right]^2 \quad \text{for } g_0(x;\widehat{\boldsymbol{\beta}}_0) \in \mathcal{P}_2\,.$$

Assuming that the error variance σ^2 is known, the selection criterion AIC suggests we favor $\mathcal{P}_2$ over $\mathcal{S}_2(t)$ providing $\mathrm{RSS}_0 - \mathrm{RSS}(t) > 2\sigma^2$.

Under the hypothesis that $f \in \mathcal{P}_2$, or that the unknown regression function is well described by a line, then the difference $V(t) = \mathrm{RSS}_0 - \mathrm{RSS}_t$ has

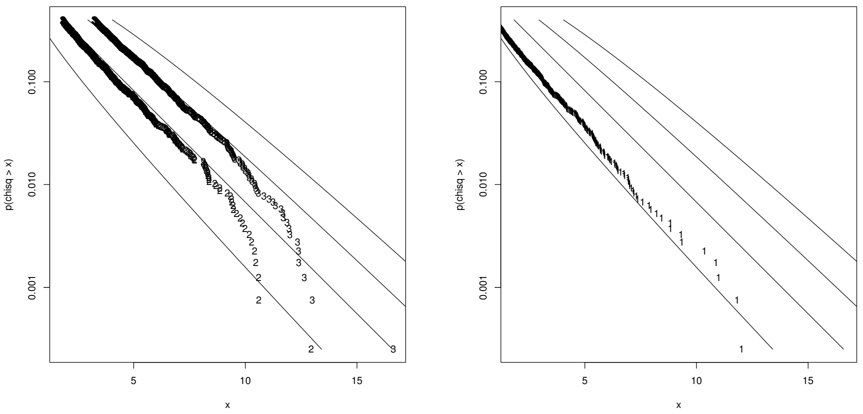

FIGURE 3.21. In each plot, the solid lines denote tail probabilities for χ^2 distributions with 1, 2, 3 and 4 degrees of freedom (arranged from the lower left to the upper right, respectively).

a χ^2 distribution with one degree of freedom. The AIC criterion suggests a classical hypothesis test with critical value 2 (or a level set at 15.73%). This holds when comparing any fixed space $\mathcal{S}_2(t)$ to $\mathcal{P}_2$. But when we conduct adaptive knot placement, we are not considering just a single t, but the t that minimizes $\mathrm{RSS}(t)$. Put another way, we are seeking $\widehat{t}$ that *maximizes the drop* in residual sum of squares

$$V = \max_t V(t) = \max_t \left[\mathrm{RSS}_0 - \mathrm{RSS}(t)\right] . \tag{3.6.5}$$

We can approximate the tail probabilities of V using a χ^2 distribution, but the degree of freedom is somewhat larger than 1.

In Figure 3.21 we plot the tail probabilities for four χ^2 distributions; as we move from lower left to upper right, these lines correspond to 1, 2, 3 and 4 degrees of freedom, respectively. The upper set of points (plotted with the symbol "3") are empirical quantiles from the distribution of V. To be more precise, we generated data $(X_1, Y_1) \ldots, (X_{100}, Y_{100})$ from the model

$$Y_i = \frac{i}{100} + \epsilon_i \qquad \text{for } i = 1, \ldots, 100 \tag{3.6.6}$$

where the error terms ϵ_i are independent standard normal random variables. We then computed RSS_0 and $\mathrm{RSS}(t)$ for $t \in (0, 1)$ and computed V according to (3.6.5). We repeated this 2,000 times and present the empirical quantiles in Figure 3.21. Notice that these points are reasonably well described by a χ^2 distribution with *three degrees of freedom.*

The lower set of points in this figure, marked with the symbol "2," are also derived from the simulation setup in (3.6.6), but this time we restricted the interval over which we searched for our knot location; that is, we set

$$V = \max_{t \in (0.2, 0.8)} \left[\mathrm{RSS}_0 - \mathrm{RSS}(t) \right] .$$

By restricting our search to the central 60% of the data, the empirical quantiles of V are now reasonably described by a χ^2 distribution with only *two degrees of freedom.* This is another quantification of the effects we have noted at the boundaries of the support of the data. This effect as first cited by Hinkley (1969) in the context of linear switching regressions. Owen (1991) and Luo and Wahba (1997) present mathematical treatments that directly quantify the degrees of freedom and the range over which we search for $\widehat{t}$.

Under the normal linear model, standard asymptotic theory suggests that if the log-likelihood is a smooth function of (t, β), then V should be roughly χ^2 with two degrees of freedom. Unfortunately, these results do not apply for $\mathcal{S}_2(t)$ because the log-likelihood involves terms of the form $(X_i - t)_+$ and hence is not differentiable at the inputs $X_1, \ldots, X_n$. Suppose, instead, that we work with cubic splines $\mathcal{S}_4(t)$. In this case, the log-likelihood has two continuous derivatives (see the surface in Figure 3.10 for an example). Now, let $\mathrm{RSS}(t)$ denote the residual sum of squares from fitting a natural cubic spline model (enforcing the tail-linear constraints at the boundaries of the data 0 and 1) with a single knot at t and again let RSS_0 denote the fit from $\mathcal{P}_2$, the space of linear functions. Given the boundary conditions, the natural cubic spline with a knot at t can be written in the form

$$\gamma_0 + \gamma_1 x + \beta h(x, t)$$

where $h(x, t)$ is given in the previous section. Since this space is a one-basis function elaboration of $\mathcal{P}_2$ and it is still appropriate to consider a break at t using $V(t) = \mathrm{RSS}_0 - \mathrm{RSS}(t)$. The points in the righthand panel of Figure 3.21 are empirical quantiles of V simulated using (3.6.6) but this time comparing the simple linear fit to the space of natural cubic splines with knots at 0, t and 1. Notice that in this case, the tail probabilities of V is well described by a χ^2 distribution with just over one degree of freedom. This is true whether we search for t in all of $(0, 1)$ or in just the central 60%, (0.2,0.8). Luo and Wahba (1997) explain this fact theoretically, and contrast it to the case of linear splines above.

This general analysis provides some justification for different values of $p(n)$ when applying standard model selection criterion like AIC and GCV. The general message is that the more elaborate the search (the larger the interval we are scanning for potential knot locations) and the rougher the spline space (in terms of the number of continuous derivatives), we can be justified in choosing a larger value of $p(n)$.

3.7 Representation and computation

Throughout this chapter, we have focused on the methodological aspects of spline modeling in the regression context. Sample data analysis and theoretical sketches have been used to illustrate the properties of splines. We now consider a few of the practicalities. In particular, we discuss the interplay between our choice of basis and the feasibility of model fitting (for a fixed spline space) and model selection (for adaptive knot placement).

3.7.1 Selecting a basis

Some history

For decades, numerical analysts have studied the computational issues involved in solving least squares problems. While available computing power has experienced exponential gains in terms of memory and speed, some of the fundamental concepts relating to precision and stability are still relevant. Recall our regression setup in Section 3.2.1 with a J-dimensional space $\mathbb{G}$ having basis functions $B_1, \ldots, B_J$, so that any $g \in \mathbb{G}$ can be written

$$g(\boldsymbol{x}; \boldsymbol{\beta}) = \beta_1 B_1(\boldsymbol{x}) + \cdots + \beta_J B_J(\boldsymbol{x}) . \tag{3.7.1}$$

Given a sample $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ of size n, we consider finding a value of $\boldsymbol{\beta}$ that minimizes the squared loss

$$\rho(\boldsymbol{\beta}) = \sum_{i=1}^{n} [Y_i - g(\boldsymbol{X}_i; \boldsymbol{\beta})]^2 = \|\boldsymbol{Y} - \mathbf{B}\boldsymbol{\beta}\|^2, \tag{3.7.2}$$

where $\mathbf{B}$ is the $n \times J$ design matrix with elements $[\mathbf{B}]_{ij} = B_j(\boldsymbol{X}_i)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. Let $\widehat{\boldsymbol{\beta}}$ denote the *true solution* to this problem, and $\widetilde{\boldsymbol{\beta}}$ the *computed solution*. There are two kinds of errors that would cause $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ to differ: *input error* and *roundoff error*. The first relates to operations on our data prior to performing the least squares operation. Data collection and various preprocessing steps may limit the accuracy of the data we have to work with. In effect, this means that $\mathbf{B}$ and $\boldsymbol{Y}$ may include errors that are much larger than the working precision of our computer. Typically, there is very little we can do about this kind of error unless we are directly involved in the data collection.

From the point of view of basis selection, our main concern is roundoff error. Each computer operation can only be executed with finite accuracy known as machine precision. While solving the least squares problem, the accumulation of these errors can result in large differences between $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$. The *condition number* of the matrix $\mathbf{B}$ in (3.7.2), denoted by $\kappa(\mathbf{B})$, describes the sensitivity of $\widetilde{\boldsymbol{\beta}}$ to roundoff errors. It is defined as the ratio

$$\kappa(\mathbf{B}) = \sqrt{\lambda_1(\mathbf{B})/\lambda_J(\mathbf{B})} \, , \tag{3.7.3}$$

where $\lambda_1(\mathbf{B})$ is the largest and $\lambda_J(\mathbf{B})$ is the smallest singular value of $\mathbf{B}$. If we let ϵ_M denote the precision of our computer, then the relative error in $\widetilde{\boldsymbol{\beta}}$ due to roundoff errors is given by

$$\frac{\|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_2}{\|\widehat{\boldsymbol{\beta}}\|_2} \leq \epsilon_M \kappa^2(\mathbf{B}) \,,$$

If $\mathbf{B}$ and $\boldsymbol{Y}$ are only known to a certain accuracy, say $\epsilon_D > \epsilon_M$, then the same bound holds but with ϵ_M replaced by ϵ_D.

Consider now a univariate predictor space (the curve fitting problem of Section 3.3, and in particular, let $X_1, \ldots, X_n$ be equally spaced over the interval $[0, 1]$. Then, if we represent the polynomials of order $k+1$ in terms of the simple power basis, $1, x, x^2, \ldots, x^k$, we can show that the condition number on the design matrix is $O(e^{3.5k})$. In short, k does not have to be very big before the system (3.7.2) becomes unstable. The essential problem is that for even moderate sized k, the basis functions are nearly collinear. This is not really a fact about sample size as the problem (and the bounds) hold for all sufficiently large n. (Note that it also does not improve if we scale the functions in some way.) To clear up this instability when working with polynomials, most computing packages like S-Plus and R make use of the so-called (orthogonal) Chebyshev polynomials (Schumaker, 1981). For a formal definition of the condition number, the reader is referred to Golub and Van Loan (1996), while complications specific to polynomial regression are discussed in Seber (1977).

Conditioning and splines

Why should we be concerned with conditioning? In Section 3.3 we found that from a methodological perspective, the truncated power basis was ideal for adaptive knot placement: each knot is associated with a single basis function, and adding or deleting that function was equivalent to inserting or removing the knot. Unfortunately, directly implementing this approach can result in badly conditioned design matrices as the order of the spline space k increases or as the number of knots M increases. In both cases, the basis functions become increasingly collinear, no matter what the value of n. We will spell this out in more detail shortly. The competing representation, the B-spline basis, was designed to avoid these difficulties and has nearly constant κ as a function of M. To make the tradeoffs precise, consider a piecewise-linear model, or rather the spline space $\mathcal{S}_2(\boldsymbol{t})$. Recall that, given the knots $\boldsymbol{t} = (t_1, \ldots, t_M)$, the truncated power basis for this space consists of the ramp functions $(x - t_m)_+$; while the B-spline basis is made up of "tent" functions $B(x; t_m, t_{m+1}, t_{m+2})$, where

$$B(x; d, e, f) = \begin{cases} (x-d)/(e-d) & x \in [d, e)\,, \\ (f-x)/(f-e) & x \leq [e, f]\,. \end{cases} \tag{3.7.4}$$

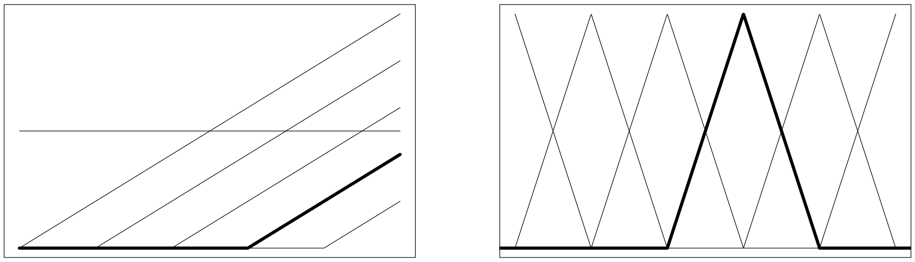

FIGURE 3.22. Two different bases for univariate linear splines: the truncated power basis (left) and the B-spline basis (right).

The two basis sets are given in Figure 3.22 for a collection of equally spaced knots. It is not hard to convince oneself that these two sets of functions do in fact form a basis for $\mathcal{S}_2(\boldsymbol{t})$.

Computationally, the main advantage of the power basis terms is their simplicity; each knot t_m is associated with a single basis term $B_m = (x - t_m)_+$, and by introducing or deleting B_m we remove the flexibility at that point. The main advantage of the B-splines is their local support. That is, each term takes on nonzero values only in an interval $[t_{m-1}, t_{m+1}]$. Associate the basis element B_m with the tent function centered at t_m, $m = 0, \ldots, M+1$, where $t_0 = a$ and $t_{M+1} = b$ are the boundaries of the interval $\mathcal{X}$. The design matrix $\mathbf{B}$ associated with the OLS problem (3.7.2) with the B-spline basis is *tridiagonal*, meaning that all of the entries $[\mathbf{B}]_{ij}$ for which $|i - j| > 2$ are zero. This represents a considerable computational savings over the power basis for large models. It is also responsible for the low condition number of the basis.

In Figure 3.23 we present the ratio of the condition number of the truncated power basis to that of the B-spline basis as a function of the number M of equally spaced knots in the interval $[0, 1]$ for cubic (top curve), quadratic (middle curve) and linear (bottom curve) splines. Our design matrix is formed from $n = 1000$ equally spaced points in the interval. While the growth is not quite exponential in the number of knots, it is considerable, even for the linear splines. This holds up no matter how large n becomes; it is more about the similarity (formally, the collinearity) between the basis elements as the knots get close. Adaptive procedures serve to ameliorate the problems with conditioning in the sense that knots will not be proposed if they make the resulting design matrix too ill-conditioned. To make this idea precise, we first discuss some computational details behind stepwise addition.

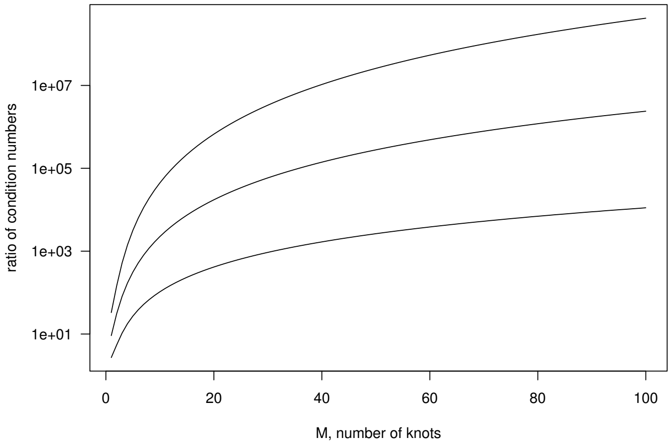

FIGURE 3.23. The ratio of the condition number for the truncated power basis to the condition number for the B-spline basis as a function of the number of knots M. The upper curve corresponds to cubic splines, the middle curve to quadratic splines, and the lower curve to linear splines.

3.7.2 Implementing stepwise addition

Direct update of the normal equations

Suppose we have an initial model consisting of the basis functions $B_i(\boldsymbol{x})$, $i = 1, \ldots, J$. Let $\mathbf{B}_0$ denote the design matrix corresponding to this set so that $[\mathbf{B}_0]_{ij} = B_i(\boldsymbol{X}_j)$. During stepwise addition, we entertain single-term additions to this model by adding a new function $B_t(\boldsymbol{x})$ where t might be a discrete variable (taking values only at order statistics of the input data) or it can range continuously over a subset of the input space. For example, $B_t(\boldsymbol{x})$ might be the linear spline basis element $(x - t)_+$ corresponding to a knot at t. Let $\mathbf{B}_1$ denote the new $n \times (J+1)$ design matrix formed from $\mathbf{B}_0$ by appending a new $(n \times 1)$ column vector $\mathbf{B}_t$, where $[\mathbf{B}_t]_j = B_t(\boldsymbol{X}_j)$; that is, $\mathbf{B}_1 = (\mathbf{B}_0, \mathbf{B}_t)$.

To determine the effect of the additional basis function, we first compute $\widehat{\mathbf{B}}_t$, the OLS estimate of the new basis function onto the span of the columns of $\mathbf{B}_0$. If we let $\widehat{\boldsymbol{\beta}}_{0t}$ denote the OLS coefficient estimate for this fit, then $\widehat{\boldsymbol{\beta}}_{0t} = (\mathbf{B}_0^{\mathrm{T}}\mathbf{B}_0)^{-1}\mathbf{B}_0^{\mathrm{T}}\mathbf{B}_t$ and $\widehat{\mathbf{B}}_t = \widehat{\boldsymbol{\beta}}_{0t}\mathbf{B}_0$. Finally, let $R_t^2 = \|\mathbf{B}_t - \widehat{\mathbf{B}}_t\|^2$ denote the residual sum of squares associated with this fit. Then, following Rao (1973), we have the equality

$$(\mathbf{B}_1^{\mathrm{T}}\mathbf{B}_1)^{-1} = \begin{pmatrix} (\mathbf{B}_0^{\mathrm{T}}\mathbf{B}_0)^{-1} + \frac{1}{R_t^2}\widehat{\boldsymbol{\beta}}_{0t}\widehat{\boldsymbol{\beta}}_{0t}^{\mathrm{T}} & -\frac{1}{R_t^2}\widehat{\boldsymbol{\beta}}_{0t} \\ -\frac{1}{R_t^2}\widehat{\boldsymbol{\beta}}_{0t}^{\mathrm{T}} & \frac{1}{R_t^2} \end{pmatrix} \tag{3.7.5}$$

which partitions $(\mathbf{B}_0^{\mathrm{T}}\mathbf{B}_0)^{-1}$ into the $J \times J$ block in the upper left, a $J \times 1$ row at the lower left, a $1 \times J$ column at the upper right, and a scalar in the lower right. Returning to our original problem, if we let $\widehat{\boldsymbol{\beta}}_0$ denote the estimated coefficients for the regression model based on $\mathbf{B}_0$, and $\widehat{\boldsymbol{\beta}}_1$ the coefficients for $\mathbf{B}_1$, then the (3.7.5) yields

$$\widehat{\boldsymbol{\beta}}_1 = \left(\begin{array}{c} \widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}_{0t} \frac{\mathbf{B}_t^{\mathrm{T}}(\boldsymbol{Y} - \mathbf{B}_0\widehat{\boldsymbol{\beta}}_0)}{R_t^2} \\ \\ \frac{\mathbf{B}_t^{\mathrm{T}}(\boldsymbol{Y} - \mathbf{B}_0\widehat{\boldsymbol{\beta}}_0)}{R_t^2} \end{array} \right) . \tag{3.7.6}$$

This well-known formulation is ideal for stepwise computations: adding a new basis function involves the coefficients at the current step $\widehat{\boldsymbol{\beta}}_0$, the residuals at the current step $\boldsymbol{Y} - \mathbf{B}_0\widehat{\boldsymbol{\beta}}_0$, and a single OLS estimate of the new function onto the span of the existing basis functions.

For model selection, our main interest is in the residual sum of squares for projecting $\boldsymbol{Y}$ onto $\mathbf{B}_1$. Again using (3.7.5), we find that

$$\mathrm{RSS}_0 - \mathrm{RSS}_1 = \frac{\left[\mathbf{B}_t^{\mathrm{T}}(\boldsymbol{Y} - \mathbf{B}_0\widehat{\boldsymbol{\beta}}_0)\right]^2}{R_t^2} , \tag{3.7.7}$$

where RSS_0 and RSS_1 are the residual sums of squares associated with the basis set $\mathbf{B}_0$ and $\mathbf{B}_1$, respectively. At each stage in stepwise addition, we examine all the viable candidate basis functions $B_t(\boldsymbol{x})$ by computing the drop (3.7.7). The function resulting in the greatest drop in the residual sum of squares is then added to the model and the overall fit is updated using (3.7.6).

With this approach, we construct a solution to the regression problem sequentially. This scheme has several advantages: first, while solving the so-called normal equations usually takes $O(J^3)$ operations, each individual update involves only $O(J^2)$ operations. This means we can audition candidate functions B_t more quickly. Next, we can control the conditioning of the least squares problem by avoiding terms that are nearly collinear with those already in the model. This effect is measured by the term R_t^2, the residual sum of squares obtained by projecting the new basis element $\mathbf{B}_t$ onto the span of $\mathbf{B}_0$. By setting an overall tolerance for the this term, we bound the condition number of the design matrices at each stage. This bounding is similar to the pivoting scheme mentioned above in connection with the implementation of OLS in packages like S-Plus and R. We also observe that these computations are formally identical to the so-called sweep operations used by Breiman (1993).

Progressive orthogonalization

Stone (1990) suggests an alteration to this scheme that drops the computation for each addition from $O(J^2)$ to $O(J)$. Basically, the idea is to orthogonalize the variables as they are entered. In terms of the notation above, this means that instead of entering the basis vector $\mathbf{B}_t$, we add the residual $\mathbf{B}_t - \widehat{\mathbf{B}}_t$. In so doing, we guarantee that each term is orthogonal to those already in the model. Notice that at each stage, the design matrix is now orthogonal; that is, if we let $\widetilde{\mathbf{B}}_0$ denote the design matrix formed by adding J terms under this new scheme, then $\widetilde{\mathbf{B}}_0^{\mathrm{T}}\widetilde{\mathbf{B}}_0 = I_{J\times J}$, the $J \times J$ identity matrix. That means that the effect of $\mathbf{B}_t$ onto $\widetilde{\mathbf{B}}_0$ is obtained by the single matrix multiplication $\widetilde{\mathbf{B}}_0^{\mathrm{T}}\mathbf{B}_t$. We note that while the speedup in computation alone recommends progressive orthogonalization, the method also simplifies the corresponding programming task.

Cholesky updates

Using the relation in (3.7.5), we have derived a simple scheme for updating the ingredients necessary to audition a large number of candidate basis functions B_t. The idea was to solve the normal equations in stages. Similar in spirit, Friedman (1990) suggests forming the Cholesky decomposition of $\mathbf{B}_0^{\mathrm{T}}\mathbf{B}_0$, and using well-known updating algorithms to audition new basis functions. The Cholesky decomposition of a symmetric, nonsingular matrix $\boldsymbol{A}$ is given by the product $\mathbf{L}\mathbf{L}^{\mathrm{T}}$, where $\mathbf{L}$ is a lower-triangular matrix. By setting $\boldsymbol{A} = \mathbf{B}_0^{\mathrm{T}}\mathbf{B}_0$, this approach to solving the normal equations means that we need only invert two triangular systems. Typically, finding the Cholesky decomposition requires $O(J^3)$ operations. However, it is possible to update this decomposition by adding a single row and column to $\boldsymbol{A}$ in only $O(J^2)$ steps, a process we describe in the next paragraph. The basic matrix computations for the Cholesky decomposition and its updates can be found in Golub and Van Loan (1989).

Specialized computations

So far, we have considered a generic setting where we have not used any facts about the basis functions being employed. When a fixed number of candidate knots are to be considered, many of the ingredients for the schemes above can be computed prior to the selection process. For example, the direct and Cholesky updates involve inner-products with candidate variables and the response. These can be computed and stored prior to the adaptive phase of the algorithm, reducing computation time during the addition process. Friedman and Silverman (1989) and Friedman (1990) present some quick updating formulae for building up the set of inner products when working with linear splines and the truncated power basis. They propose visiting the knots in decreasing order and using the fact that for $t \le u$,

$$(x-t)_+ - (x-u)_+ = \begin{cases} 0, & x \le t, \\ x-t, & t < x < u, \\ u-t, & x \ge u, \end{cases}$$

By structuring the computations in this way and precomputing the inner-products for a fixed set of candidate knot locations, we need only $O(J^2)$ operations per Cholesky update rather than (J^3).

When working with B-splines, we can reduce computations by taking into account the fact that these basis functions have small support. In the case of linear splines, we observed above that the entries $[\mathbf{B}_0^{\mathrm{T}}\mathbf{B}_0]_{ij}$ are zero for all $|i-j| > 2$. In the literature on matrix computations, such a matrix is referred to as *band-limited* because the non-zero entries are restricted to a narrow band around the main diagonal. Golub and Van Loan (1989) discuss band-limited versions of the Cholesky decomposition and its updates that improve even further over the $O(J^2)$ operations.

For linear splines, the savings incurred for moderate to large problems far outweigh the speed-ups suggested by the updating schemes in Friedman and Silverman (1989) for the truncated power basis. During knot addition, for example, should we want to enter a knot at point t_l, we can add a B-spline basis function from the mesh with t_l added (one of the basis elements with support that covers t_l) rather than $(x-t_l)_+^{k-1}$. Short cuts of this type are discussed in more detail in Chapter 9 in the context of bivariate linear splines defined over triangulations in the plane.

3.7.3 Connection to smoothing splines

We close this section with one last basis representation, this time coming from the literature on so-called *smoothing splines*. For the moment, we restrict ourselves to 1-dimensional curve fitting where the regression function is known to be defined in the interval $\mathcal{X} = [0,1]$. Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ we formulate a penalized least-squares criterion

$$\sum_{i=1}^{n} [Y_i - h(X_i)]^2 + \int_0^1 (h''(x))^2\, dx \tag{3.7.8}$$

for functions h. The smoothing spline estimate of f is the minimizer $\hat{h}$ of (3.7.8). In Section **??** we find that the solution is a natural cubic spline; that is, a piecewise cubic function with continuous second derivatives satisfying the tail-linear constraints (3.6.4) specified above. While smoothing splines have a rich, detailed history, our current interest in the construction is that it provides us with another basis for the natural cubic splines.

To make this precise, define the functions

$$k_1(x) = x - 1/2\,, \quad k_2(x) = \left(k_1^2(x) - 1/12\right)/2\,, \tag{3.7.9}$$

and

$$k_4(x) = \left(k_1^4(x) - k_1^2(x)/2 + 7/240\right)/24\,. \tag{3.7.10}$$

Then, given a knot sequence $0 = t_0 < t_1, \ldots, t_M < t_{M+1} = 1$, Craven and Wahba (1979) show that a natural cubic spline can be written as

$$h(x) = \alpha_0 + \alpha_1 x + \beta_1 R(x, t_1) + \cdots + \beta_m R(x, t_M) \tag{3.7.11}$$

where the *kernel functions* $R(\cdot,\cdot)$ are given by

$$R(x,t) = k_2(x)k_2(t) - k_4(|x-t|)\,. \tag{3.7.12}$$

From the previous section, we recall that the natural cubic splines with the indicated knot sequence should have dimension M, and yet the expansion in (3.7.11) involves $M+2$ functions. Also, notice that because of the leading term in k_4, the kernel functions are in fact quartics.

We obtain the cubic natural splines from expansions of the form (3.7.11) by imposing two linear constraints on the coefficients $\beta_1,\ldots,\beta_M$. Interestingly, these turn out to be precisely the constraints given in (**??**). To see that this works, we can reexpress k_4 as

$$k_4(x) = \left[x^4 - \left(x^3 + x^2 k_1(x) + x k_1^2(x) + k_1^3(x)\right)/2\right]/24\,.$$

This means that in the kernel function $R(x,t)$ we have terms like $(x-t)^4$. Taking three derivatives of $R(x,t)$ with respect to x, we have

$$\frac{\partial^3}{\partial x^3} R(x,t) = (-(x-t) + I_{x>t} - I_{x>t})/2\,. \tag{3.7.13}$$

Now, consider the third derivative of $h(x)$ in (3.7.8) and notice that each kernel function contributes a term of the form $(x-t_m)$. The constraints act to eliminate these. Note that that for $x=1$

$$\sum_m \beta_m(1-t_m) = 0 \quad \text{implies} \quad \sum_m \beta_m = \sum_m \beta_m t_m$$

while for $x=0$,

$$\sum_m \beta_m(0-t_m) = 0 \quad \text{implies} \quad \sum_m \beta_m t_m = 0\,.$$

Therefore, the sum $\sum_m \beta_m(x-t_m)$ is zero for any $x \in [0,1]$.

While the truncated power basis required the two constraints to satisfy the boundary conditions at the rightmost interval, the kernel functions that make up this basis all satisfy the tail-linear property given in (**??**). Instead, the kernel basis requires the constraints to eliminate the quartic component that helps each kernel individually satisfy the tail-linear conditions. As with the truncated power basis, each kernel function is associated with a single knot. As we see from (3.7.13), the function $R(x,t)$ has a break in its second derivative at t, and the jump is 1. That means that in the expansion (3.7.11), the jump at knot t_m is β_m.

Luo and Wahba (1997) use these basis functions in a stepwise addition scheme, ignoring the constraints required to make them cubics. We introduce them here because many generalizations now exist for different kinds of kernels. In Section **??** we will see how this general approach has taken shape in the context of support vector machines and other tools coming from the data mining and machine learning literature.

3.8 A second look at the examples

3.8.1 *Assessing uncertainty in curve fitting*

Consider again the fit to the ${}^{87}\delta\,\mathrm{Sr}$ data given in Figure 3.3 near the beginning of the chapter. The fit was in a linear space $\mathbb{G}$ of cubic splines with tail-linear constraints, and it can be represented as

$$\widehat{g}(x) = \sum_{j=1}^{J} \widehat{\beta}_j B_j(x)\,,$$

for OLS estimates $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_J)$. In this case, we have a cubic spline space with five knots (three interior and two at the boundaries, the smallest and largest age values) and we have enforced two tail-linear constraints at each of the boundary knots. By imposing $2 \times 2 = 4$ constraints on a space of dimension 9 (cubic splines with three knots), our final model has $J = 9 - 4 = 5$ degrees of freedom. Since we are working with OLS estimates, we can apply usual regression theory and estimate the variance-covariance matrix for $\widehat{\boldsymbol{\beta}}$ as $s^2(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}$, where $\mathbf{B}$ is the usual design matrix and $s^2 = \mathrm{RSS}/(n-J)$. From here, we can define pointwise confidence intervals. For any point x_0, we have

$$\widehat{g}(x_0) \pm 2\sqrt{s^2 \boldsymbol{b}(x_0)^{\mathrm{T}}(\mathbf{B}^{\mathrm{T}}\mathbf{B})\boldsymbol{b}(x_0)},$$

where $\boldsymbol{b}(x_0)$ is the vector $(B_1(x_0), \ldots, B_J(x_0))^{\mathrm{T}}$. If $f \in \mathbb{G}$ and we have not done any knot selection, this interval would be correct. In Figure 3.24 we overlay this standard OLS pointwise confidence interval on top of the bootstrap estimate from Figure 3.3. While similar in some places, we see that the regression interval is often too narrow, especially in areas with strong features like the KTB.

The pointwise confidence intervals in 3.3 were derived using a bootstrap procedure. We constructed bootstrap samples by drawing 45 data points from the ${}^{87}\delta\,\mathrm{Sr}$ data with replacement. For each bootstrap sample, we rang the greedy stepwise algorithm, adding knots to a maximum model of 10 breakpoints and then conducting stepwise deletion. BIC was used to select the best model for this sample. We repeated this for each of 200 samples and plotted the 97.5 and 2.75 percentiles as curves. These are the black bands in Figure 3.24.

There has to be more to say here; if we use a model selection criterion to gauge the amount of smoothing, we're trading off the bias and variance... so...

3.8.2 *Test set prediction error*

In Section 3.2.3, we motivated the use of selection criterion like AIC and C_p. While these criteria provide reasonable protection against overfitting

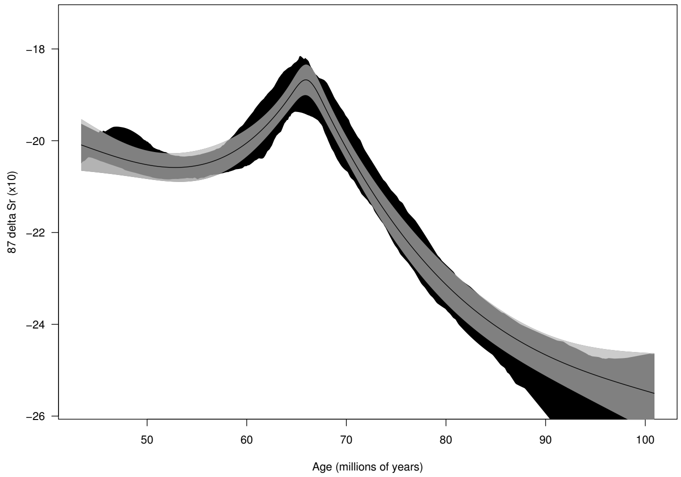

FIGURE 3.24. Comparing standard pointwise confidence intervals to a bootstrap procedure.

when tuned properly (through a good choice of $p(n)$), we will now explore out-of-sample estimates. Recall that the test set estimate of prediction error is given by

$$\widehat{\mathrm{PE}}_{\mathrm{TS}}(\mathbb{G}) = \frac{1}{K}\sum_{i=1}^{K}\left[Y_i^* - g(\boldsymbol{X}_i^*;\widehat{\boldsymbol{\beta}})\right]^2. \tag{3.8.1}$$

Let us reconsider the $^{87}\delta\,\mathrm{Sr}$ data and examine model selection using a test set. Martin and Macdougall (1991) present a follow up study with a series of observations taken from shell samples having dates very close to the KTB. We will combine the 39 points from their site 356 (having “quality” measures of 1 or 2, indicating clean measurements) with their original 45 points to create a dataset with 84 observations. We then divided the dataset into a training set (two thirds of the available data or 56 observations) and a test set (one third or 28 observations).

We then conducted stepwise addition with a natural spline basis to 10 knots, followed by backward deletion. In Figure 3.25 we present the best model suggested by the test set (solid line, 6 knots) compared with the best suggested by BIC using the full 84 data points (dashed line, 10 knots). In this case, the BIC fit consists of more knots, and specifically more in the neighborhood of the KTB. In both cases, the extra data at the KTB has allowed us to resolve the peak more completely.

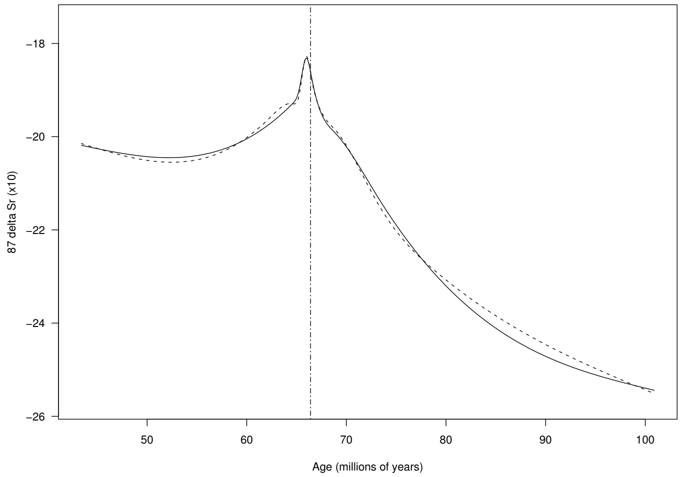

FIGURE 3.25. Two fits based on

3.8.3 Multivariate responses

The abundance score (3.1.1) for a given species is normalized by the prevalence of other kinds of trees in the same forest. For regions with just a small number of species, the abundance scores will be correlated. This implies that there may be some advantage to "borrowing strength" between species when modeling the IV. To formalize this idea, for each input vector $\boldsymbol{X}_i$, we consider a vector of responses (Y_{1i}, Y_{2i}, Y_{3i}) for $i = 1, \ldots, n$. As was done at the beginning of this chapter, we will model the response vector using linear spaces, typically the approximation spaces. Let $\mathbb{G}$ denote a J-dimensional linear space with basis

$$B_j(\boldsymbol{x}), \qquad j = 1, \ldots, J, \tag{3.8.2}$$

defined for $\boldsymbol{x} \in \mathcal{X}$. Let $g = (g_1(\boldsymbol{x}), g_2(\boldsymbol{x}), g_3(\boldsymbol{x}))$ denote three functions in $\mathbb{G}$. We evaluate how well they fit our data using an extension of the OLS criterion:

$$\rho(g) = \sum_{m=1}^{3} \sum_{i=1}^{n} \left[\, Y_{mi} - g_m(\boldsymbol{X}_i)\,\right]^2 , \qquad g \in \mathbb{G}\,. \tag{3.8.3}$$

Each of the functions g_m can be expressed in terms of the basis for $\mathbb{G}$ as

$$g(\boldsymbol{x}; \boldsymbol{\beta}_m) = \beta_{m1} B_1(\boldsymbol{x}) + \cdots + \beta_{mJ} B_J(\boldsymbol{x})$$

for some parameter vector $\boldsymbol{\beta}_m$. Combining these parameters across the different models into one vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$, we can rewrite the OLS

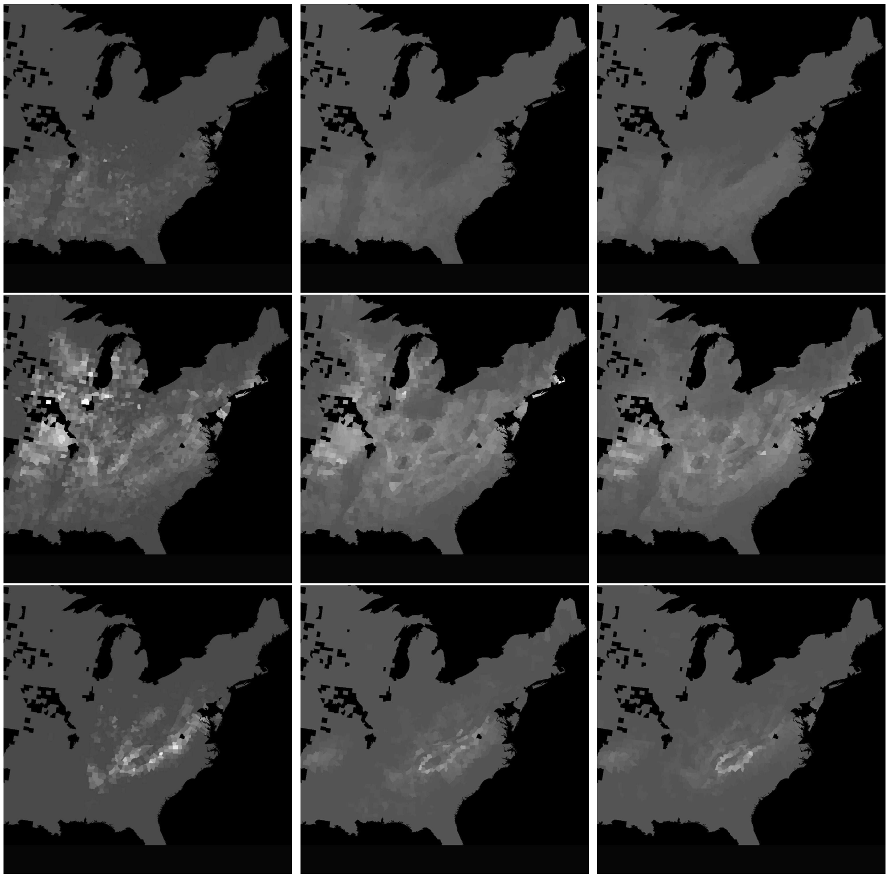

FIGURE 3.26. Raw IV abundance values together with predictions from two different spline methods. Raw values (left column) for the southern red oak, the white oak and the virginia pine (top, middle and bottom); predictions from a cross-validated spline fit; and predictions from a "joint" model that treats the IV values for all three species simultaneously.

criterion as

$$\begin{aligned}\rho(\boldsymbol{\beta}) &= \sum_{m=1}^{3}\sum_{i=1}^{n}\left[\, Y_{mi} - g(\boldsymbol{X}_i; \boldsymbol{\beta}_m)\,\right]^2 \\ &= \sum_{m=1}^{3}\sum_{i=1}^{n}\left[\, Y_{mi} - \beta_{m1}B_1(\boldsymbol{X}_i) - \cdots - \beta_{mJ}B_J(\boldsymbol{X}_i)\,\right]^2 \qquad (3.8.4)\end{aligned}$$

and our estimator solves

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^J}\,\rho(\boldsymbol{\beta})\,.$$

We obtain the OLS estimate $\widehat{\boldsymbol{\beta}}$ by solving the usual normal equations

$$(\mathbf{B}^{\mathrm{T}}\mathbf{B})\widehat{\boldsymbol{\beta}} = \mathbf{B}^{\mathrm{T}}\boldsymbol{Y}, \tag{3.8.5}$$

where $\mathbf{B}$ is the $n \times J$ design matrix with elements $[\mathbf{B}]_{ij} = B_j(\boldsymbol{X}_i)$, and $[\boldsymbol{Y}]_{mj} = Y_{mj}$. Comparing this with (3.2.11), we see that the estimate of $\boldsymbol{\beta}_m$ is simply the OLS estimate using $\mathbb{G}$ using the data $Y_{m1}, \ldots, Y_{mn}$ and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

When adaptively selecting $\mathbb{G}$, we can leverage the existence of general structures common to the different tree species by placing knots and introducing interactions using the criterion (3.8.3). That is, at each step in the addition process, we introduce a basis function that has the greatest drop in (3.8.3). In this case, we are now estimating 3 more parameters (one for each element in $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$). Similarly, during backward deletion, we now drop that element from a linear space that creates the least rise in (3.8.3), again reducing the model by 3 degrees of freedom each time.

In the last column of Figure 3.26, we present the predictions from this jointly derived model. As was done with the middle column, we again select between the collection of models using a test set estimate of prediction error. In Table 3.6 we see the wide range of model sizes obtained by simple marginal fitting. In terms of parameters, the combined model is only modestly larger than the smallest models, and much smaller than that fit for the white oak (22 versus 48 degrees of freedom). Even so, the prediction error estimates are not very far off, although they are consistently better for the separate fits. In the case of the white oak, we seem to lose very little in terms of prediction error for having such a drastically smaller model. Among the 22 variables in the combined model, only one factor involving the climate variables appears, namely JANT, the average January temperature for the county. In Figure 3.27 we plot out the linear components of this model. Note the different effect a warm winter has on tree abundance for the three species.

	J	PE	J	PE
Southern red oak	16	5.7	22	5.7
White oak	48	43.8	22	50.5
Virginia pine	19	12.4	22	13.1

TABLE 3.6. Model sizes and test set estimates of prediction error for three tree species, fitting separate and combined models.

3.9 Conclusion