15

Functional linear models for scalar responses

15.1 Introduction

In this chapter, we consider a linear model defined by a set of functions, but where the response variable is scalar or multivariate. This contrasts with Chapter 13, where the responses and the parameters were functional, but, because of the finite and discrete covariate information, the linear transformation from the parameter space to the observation space was still specified by a *design matrix* $\mathbf{Z}$ as in the conventional multivariate general linear model

$$\mathbf{y} = \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon} \; . \tag{15.1}$$

We now consider a functional extension of linear regression where the prediction of the scalar values y_i is based on functions z_i. This problem is of interest in its own right, and also raises issues about more complicated problems in subsequent chapters.

For illustration, let us predict total annual precipitation for a Canadian weather station from the pattern of temperature variation through the year. To this end, let $y_i = \texttt{LogPrec}_i$ be the logarithm of total annual precipitation at weather station i, and let $z_i = \texttt{Temp}_i$ be its daily temperature function. We now replace the regression vector $\mathbf{b}$ in (15.1) by a function β, so that the model now takes the form

$$\texttt{LogPrec} = \alpha + \int_0^T \texttt{Temp}(s)\beta(s)\, ds + \boldsymbol{\epsilon} \; . \tag{15.2}$$

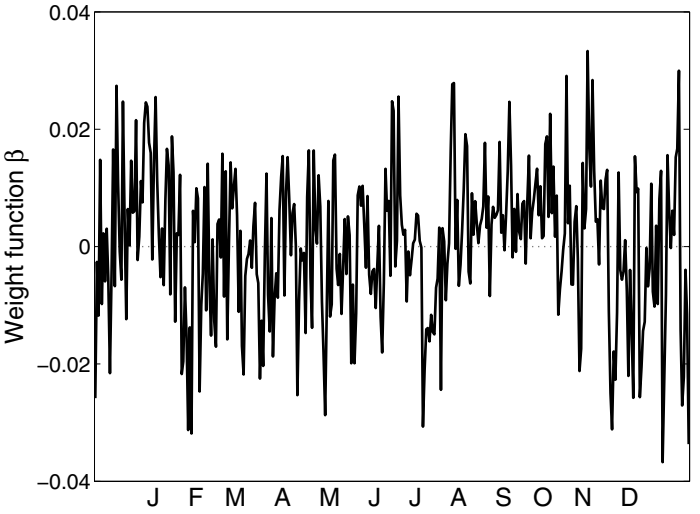

Figure 15.1. The weight function β that allows perfect prediction of log total annual precipitation from observed annual pattern of temperature.

We see that the summation implied in the matrix product **Zb** in (15.1) is now replaced by an integration over a continuous index s in (15.2).

15.2 A naive approach: Discretizing the covariate function

It might occur to us to treat the values of temperature at each observation point as a separate covariate, and then just proceed with ordinary multiple regression. This would certainly get us into trouble! To see why, suppose that $\texttt{Temp}_{ij}$ is the entry for the temperature at station i on day j, and we wish to predict $\texttt{LogPrec}_i$ by

$$\texttt{LogPrec}_i = \alpha + \sum_{j=1}^{365} \texttt{Temp}_{ij}\beta_j + e_i \; i = 1, 2, \ldots, 35. \tag{15.3}$$

We can view this as a finely discretized version of the functional model being considered. This is a system of 35 equations with 366 unknowns. Even if the coefficient matrix is of full rank, there are still infinitely many sets of solutions, all giving a perfect prediction of the observed data. Figure 15.1 plots the b_j's for one such solution, and it is hard to imagine that we can make much practical use out of such a result.

Returning to the functional model (15.2), we now understand that the regression coefficient function β is bound to be under-determined on the

basis of any finite sample (z_i, y_i). This is because, essentially, we have an infinite number of parameters $\beta(s)$ available by discretizing s finely enough, but a finite number of conditions $y_i = \alpha + \int z_i\beta$ to approximate. Usually it is possible to find $\hat{\alpha}$ and $\hat{\beta}$ to reduce the residual sum of squares (15.2) to zero. Furthermore, if β^* is any function satisfying $\int z_i\beta^* = 0$ for $i = 1, \ldots, N$, then adding β^* to $\hat{\beta}$ does not affect the value of the residual sum of squares.

In the weather data example, a possible approach is to reduce the number of unknowns in problem (15.3) by considering the temperatures on a coarser time scale. It is unlikely that overall precipitation is influenced by details of the temperature pattern from day to day, and so, for example, we could investigate how the 12-vectors of monthly average temperatures can be used to predict total annual precipitation. If $\mathbf{Z}$ is the 35×12 matrix containing these values, we can then fit a model of the form $\hat{y} = \hat{\alpha} + \mathbf{Z}\hat{\beta}$, where $\hat{y}$ is the vector of values of log annual precipitation predicted by the model, and $\hat{\beta}$ is a 12-vector of regression parameter estimates. Since the number of parameters to be estimated is now only 13, and thus less than the number of observations $N = 35$, we can use standard multiple regression to fit the model by least squares.

We can summarize the fit in terms of the conventional $R^2 = 1 - \texttt{SSE}/\texttt{SSY}$ measure, and this is 0.84, indicating a rather successful fit, even taking into account the 13 parameters in the model. The corresponding F-ratio is 9.8 with 12 and 22 degrees of freedom, and is significant at the 1% level. The standard error estimate is 0.34, as opposed to the standard deviation of the dependent variable of 0.69.

Figure 15.2 presents the estimated regression function β, obtained by interpolating the individual estimated coefficients $\hat{\beta}_j$ as marked on the figure. It is still not easy to interpret this function directly, although it clearly places considerable emphasis on temperature in the months of April, May, August and September. The lack of any very clear interpretation indicates that this problem raises statistical questions beyond the formal difficulty of fitting an under-determined model. In any case, the model certainly uses up a rather large proportion of the 35 degrees of freedom available in the data.

Since the space of functions satisfying (15.2) is infinite-dimensional, no matter how large our sample size N is, minimizing the residual sum of squares cannot, of itself, produce a meaningful or consistent estimator of the parameters β in the model (15.2). Consequently, to provide an estimate of $\hat{\beta}$ that we can interpret or otherwise use, or even just identify uniquely, we must use some method of regularization, and this is discussed in the following sections.

In short, penalizing roughness when a functional covariate is involved is no longer cosmetic, but an essential aspect of finding a useful solution. We have already seen this issue discussed in Section 11.5 in functional canonical correlation analysis, and we will consider it again in the next chapter.

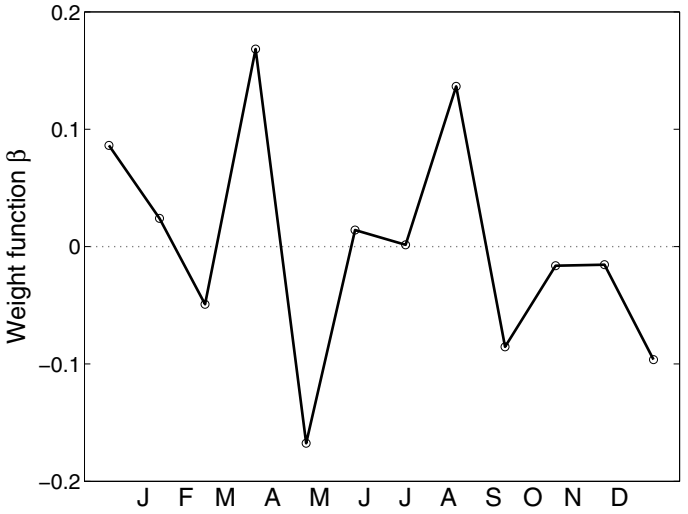

Figure 15.2. The regression function β for the approximation of annual mean log precipitation by the temperature profiles for the Canadian weather stations.

15.3 Regularization using restricted basis functions

To reduce the degrees of freedom in the model still further, we now expand the regression function β in terms of a set of basis functions $\theta_k(s)$, and the Fourier basis is the logical choice here because of the the underlying smoothness and stationarity of the seasonal variation in temperature. Let $\boldsymbol{\theta}$ be a vector of Fourier basis functions of length K_β, so that

$$\beta(s) = \sum_{k}^{K_\beta} b_k \theta_k(s) \quad \text{or} \quad \beta = \boldsymbol{\theta}'\mathbf{b}. \tag{15.4}$$

We choose some suitably large K_β that does not entail any significant loss of information, but hopefully keeps K_β small enough so that we can reasonably interpret β.

At the same time, let us assume that the covariate functions $\texttt{Temp}_i$ are also expanded in terms of Fourier basis vector $\boldsymbol{\psi}$ of length K_z, so that

$$\texttt{Temp}_i(s) = \sum_{k}^{K_z} c_{ik}\psi_k(s) \quad \text{or} \quad \texttt{Temp}(s) = \mathbf{C}\boldsymbol{\psi}(s)\ , \tag{15.5}$$

where coefficient matrix $\mathbf{C}$ is N by K_z. For the monthly and daily temperature data, for example, K_z would be 12 and 365, respectively.

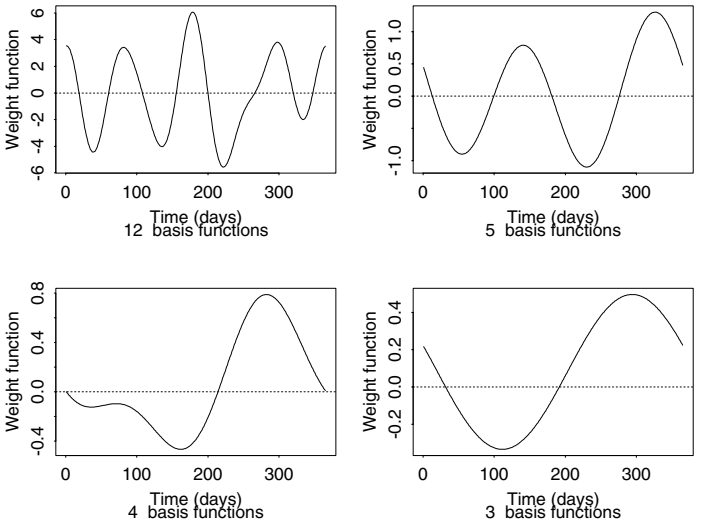

Figure 15.3. Estimated regression weight functions β using $K_\beta = 12$, 5, 4 and 3 basis functions.

Now the model can be expressed as

$$\hat{y}_i = \int_0^T \texttt{Temp}(s)\beta(s)\,ds = \int_0^T \mathbf{C}\psi(s)\theta(s)'\mathbf{b}\,ds = \mathbf{C}\mathbf{J}_{\psi\theta}\mathbf{b}\ , \tag{15.6}$$

where K_z by K_β matrix $\mathbf{J}_{\psi\theta}$ is defined by

$$\mathbf{J}_{\psi\theta} = \int \boldsymbol{\psi}(s)\boldsymbol{\theta}'(s)\,ds\ . \tag{15.7}$$

We can further simplify notation by defining the $(K_\beta + 1)$-vector $\boldsymbol{\zeta} = (\alpha, b_1, \ldots, b_K)'$ and defining the coefficient matrix $\mathbf{Z}$ to be the $N \times (K_\beta + 1)$ matrix $\mathbf{Z} = [\mathbf{1}\ \ \mathbf{C}\mathbf{J}_{\psi\theta}]$. Then the model (15.1) becomes simply

$$\hat{\mathbf{y}} = \mathbf{Z}\hat{\boldsymbol{\zeta}} \tag{15.8}$$

and the least squares estimate of the augmented parameter vector $\boldsymbol{\zeta}$ is the solution of the equation

$$\mathbf{Z}'\mathbf{Z}\hat{\boldsymbol{\zeta}} = \mathbf{Z}'\mathbf{y}\ . \tag{15.9}$$

A convenient method of regularization that we used in Chapter 4 is to truncate the basis by choosing a value $K_\beta < K_z$. We can then fit $\boldsymbol{\zeta}$ by least squares, and the problem is now a standard multiple regression problem.

Figure 15.3 shows the result of carrying out this procedure for the daily weather data with varying numbers K_β of basis functions. The choice $K_\beta =$

12 is intended to correspond to the same amount of discretization as using monthly average data, and we can see that the weight function is similarly uninformative. To obtain results more likely to be meaningful, we have to use a much smaller number of basis functions, and, by considering the graphs for $K_\beta = 4$ and $K_\beta = 3$, it appears that a predictor for high precipitation is a relatively high temperature towards the end of the year.

But the model complexity increases in discrete jumps as K_β varies from three to five, and we might want finer control. Also, to obtain reasonable results, β must be rigidly constrained to lie in a low-dimensional parametric family, and we may worry that we are missing important features in β as a consequence. Section 15.4 develops a more flexible approach making use of a roughness penalty method.

15.4 Regularization with roughness penalties

The estimated function $\hat{\beta}$ in Figure 15.1 illustrates that fidelity to the observed data, as measured by the residual sum of squares, is not the only aim of the estimation. The roughness penalty approach makes explicit the complementary, possibly even conflicting, aim of avoiding excessive local fluctuation in the estimated function.

To this end, we can define the penalized residual sum of squares

$$\texttt{PENSSE}_\lambda(\alpha, \beta) = \sum_{i=1}^{N} [y_i - \alpha - \int z_i(s)\beta(s)\, ds]^2 + \lambda \int [L\beta(s)]^2\, ds \ , \quad (15.10)$$

where L is a linear differential operator that is suitable for the problem. In this situation, it is reasonable to expect that regression function β will be periodic, just like the average temperature function that it multiplies. Consequently, it seems appropriate to choose *harmonic acceleration* as the type of roughness to penalize. That is, we choose

$$L\beta = (\frac{2\pi}{365})^2 D\beta + D^3\beta$$

so that in the limit, as $\lambda \to \infty$, the regression function will approach a shifted sinusoid. Sections 15.5 and 15.7 discuss the algorithmic aspects of minimizing (15.10).

We can choose the smoothing parameter λ either subjectively or by an automatic method such as cross-validation. To apply the cross-validation paradigm in this context, let $\alpha_\lambda^{(-i)}$ and $\beta_\lambda^{(-i)}$ be the estimates of α and β obtained by minimizing the penalized residual sum of squares based on all the data except (z_i, y_i). We can define the cross-validation score as

$$\texttt{CV}(\lambda) = \sum_{i=1}^{N} \left[y_i - \alpha_\lambda^{(-i)} - \int z_i(s)\beta_\lambda^{(-i)}(s)\right]^2 ds \quad (15.11)$$

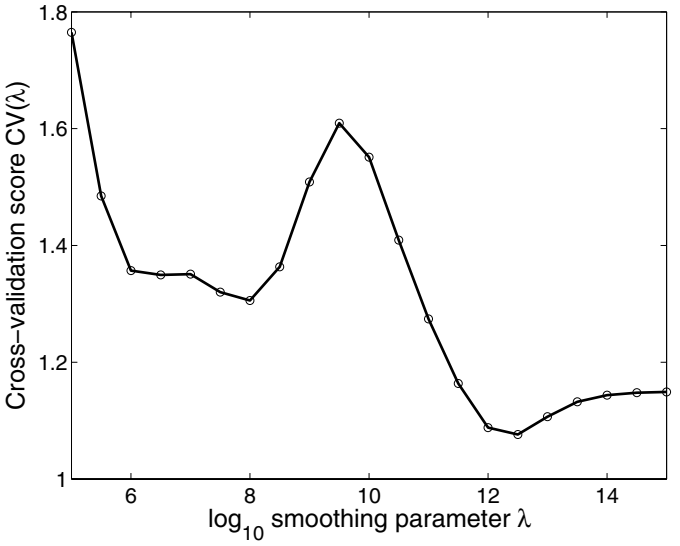

Figure 15.4. The cross-validation score function $\texttt{CV}(\lambda)$ for fitting log annual precipitation by daily temperature variation, with a penalty on the size of harmonic acceleration. The logarithm of the smoothing parameter is taken to base 10.

and minimizing $\texttt{CV}(\lambda)$ over λ gives an automatic choice of λ. In practice, there are efficient algorithms for calculating the cross-validation score, and Section 15.6 discusses these.

We used 65 basis functions to represent the temperature curves and 35 Fourier basis functions to represent β. With this number of basis functions for β, it would be possible to exactly fit the data from the 35 weather stations. However, we wanted to see how well cross-validation would help us in arriving at a reasonable fit by penalizing harmonic acceleration. Figure 15.4 plots the cross-validation score against the logarithms of various values of λ. The plot shows two distinct minima over the range of values plotted. Not shown, however, is the fact that fitting the data exactly or nearly exactly actually gave a smaller cross-validation score than either of these minima. We chose $\lambda = 10^{12.5}$ for the final fit, corresponding to the lower minimum in the plot.

Figure 15.5 shows the estimated regression function along with pointwise 95% confidence limits. The confidence intervals in the earlier summer months contain zero, suggesting that the influence of temperature on precipitation in that period is not important. However, we see a strong peak in the late fall followed by a valley in the early spring. This pattern is, in effect, computing a contrast between fall and early spring temperatures, with more emphasis on the autumn. This pattern favors weather stations that are comparatively warm in October and cool in spring, and where, moreover, spring comes early. This is just what we saw in Chapter 7 for

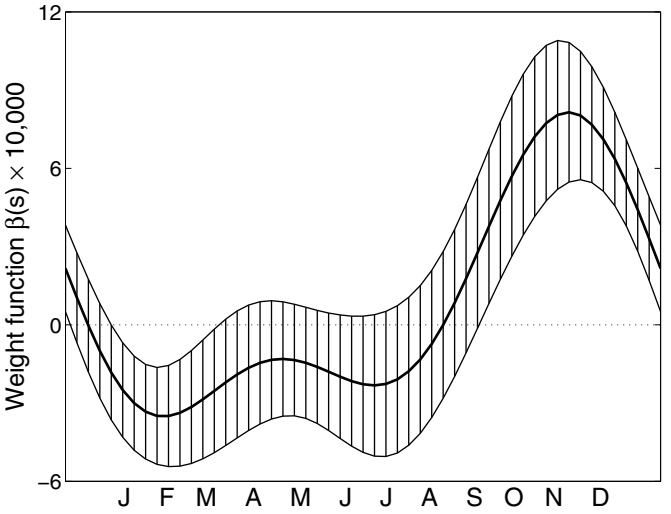

Figure 15.5. The estimated weight function for predicting the log total annual precipitation from the daily temperature pattern. The estimate was constructed by the penalizing the size of harmonic acceleration, with the smoothing parameter $\lambda = 10^{12.5}$ chosen by cross-validation.

the Pacific and Atlantic stations with marine climates, where the seasons are later than average and the fall weather is warm relative to the inland stations.

In Figure 15.6, we have plotted the observed values y_i against the fitted values $\hat{y}_i$ obtained using this functional regression. The squared correlation between the predicted and actual values in the plot is 0.75. This simple regression diagnostic seems to confirm the model assumptions. However, we didn't do so well for Kamloops, whose predicted value of about 2.9 is well above its actual value of a bit under 2.5. But Kamloops is deep in the Thompson River valley, and the rain clouds usually just pass on by. Section 15.6 describes another diagnostic plot.

15.5 Computational issues

A basis function approach has appeal because it is especially simple to apply, and moreover some problems in any case suggest a particular choice of basis. The periodic nature of the temperature and precipitation data, for example, seems naturally to call for the use of a Fourier series basis. Our first strategy is therefore to represent the regularized fitting problem in terms of a basis function expansion, and then to apply the concept of regularization to this representation.

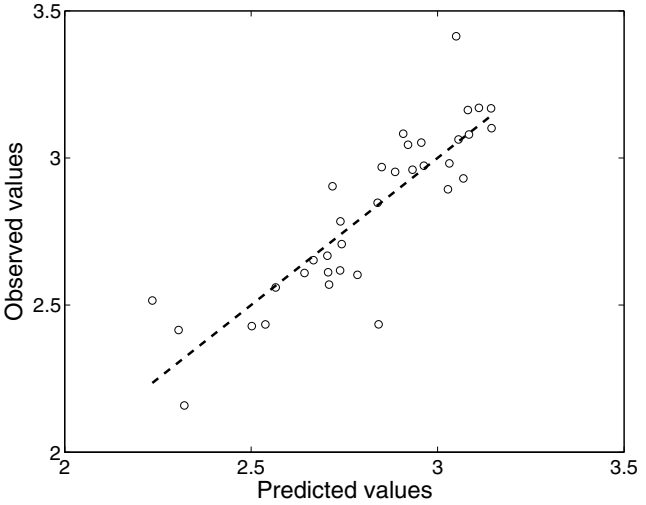

Figure 15.6. Observed values y_i of log annual precipitation plotted against the values $\hat{y}_i$ predicted by the functional regression model with the smoothing parameter chosen by cross-validation. The straight line corresponds to zero residuals.

15.5.1 Computing the regularized solution

Suppose that we expand the covariate functions z_i to K_z terms relative to basis functions ψ_m and the regression function β to K_β terms relative to basis functions θ_k, as in (15.5) and (15.4), respectively. Define a matrix $\mathbf{R}$ as

$$\mathbf{R} = \int [D^2\boldsymbol{\phi}(s)][D^2\boldsymbol{\phi}'(s)]\, ds \ . \tag{15.12}$$

In the Fourier case, note that $\mathbf{R}$ is diagonal, with diagonal elements ω_k^4 as in Section 9.4.1. In general, the penalized residual sum of squares can be written as

$$\texttt{PENSSE}_\lambda(\alpha, \beta) = \|\mathbf{y} - \alpha - \mathbf{C}\mathbf{J}_{\psi\theta}\mathbf{b}\|^2 + \lambda \mathbf{b}'\mathbf{R}\mathbf{b}. \tag{15.13}$$

where $\mathbf{J}_{\psi\theta}$ was defined in (15.7). As before, we deal with the additional parameter α by defining the augmented vector $\boldsymbol{\zeta} = (\alpha, \mathbf{b}')'$, and at the same time use $\mathbf{Z}$ as the $N \times (K_z + 1)$ coefficient matrix $[\mathbf{1} \ \ \mathbf{C}\mathbf{J}_{\psi\theta}]$. Finally, let the penalty matrix $\mathbf{R}$ be augmented by attaching a leading column and row of $K_z + 1$ zeros to yield $\mathbf{R}_0$. In terms of these augmented arrays, the expression (15.13) further simplifies to

$$\texttt{PENSSE}_\lambda(\boldsymbol{\zeta}) = \|\mathbf{y} - \mathbf{Z}\boldsymbol{\zeta}\|^2 + \lambda \boldsymbol{\zeta}'\mathbf{R}_0\boldsymbol{\zeta}. \tag{15.14}$$

It follows that the minimizing value $\hat{\boldsymbol{\zeta}}$ satisfies

$$(\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{R}_0)\hat{\boldsymbol{\zeta}} = \mathbf{Z}'\mathbf{y}. \tag{15.15}$$

15.5.2 Computing confidence limits

We can again follow the procedure that we used in previous chapters to compute sampling standard errors for the coefficients in $\mathbf{b}$ and the intercept α in the composite parameter vector $\boldsymbol{\zeta}$. Things are simpler here in one sense since there is no intermediate step of smoothing the response variable. Consequently, we can drop the mapping `y2cMap`.

The matrix corresponding to `y2bMap` can be simply lifted from (15.15), and is $(\mathbf{Z}'\mathbf{Z}+\lambda\mathbf{R}_0)^{-1}\mathbf{Z}'$. The variance-covariance matrix Σ_e computed from the residuals is now a scalar estimate σ_e^2 of the mean squared residual. The sampling variance of $\hat{\boldsymbol{\zeta}}$ is given by

$$\mathtt{Var}[\hat{\boldsymbol{\zeta}}] = \sigma_e^2(\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{R}_0)^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{R}_0)^{-1} \ . \tag{15.16}$$

15.6 Cross-validation and regression diagnostics

We have already noted the possibility of choosing the smoothing parameter λ by cross-validation. Various economies are possible in calculating the cross-validation score $\mathtt{CV}(\lambda)$ as defined in (15.11).

Let $\mathbf{S}$ be the so-called *hat matrix* of the smoothing procedure which maps the data values y to their fitted values $\hat{y}$ for any particular value of λ. A calculation described, for example, in Section 3.2 of Green and Silverman (1994), shows that the cross-validation score satisfies

$$\mathtt{CV}(\lambda) = \sum_{i=1}^{N}\left(\frac{y - \hat{y}_i}{1 - S_{ii}}\right)^2 .$$

If N is large and we are considering an expansion in a moderate number K of basis functions, then we can find the diagonal elements of $\mathbf{S}$ directly from

$$\mathbf{S} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{R})^{-1}\mathbf{Z}'.$$

From $\mathbf{S}$, we can also compute an indicator of the effective degrees of freedom used up in the approximation. Either trace $\mathbf{S}$ or trace $\mathbf{S}^2$ were recommended for this purpose by Buja, Hastie, and Tibshirani (1989). For the fit in Figure 15.5, defined by minimizing the cross-validation criterion, the effective degrees of freedom are estimated to be trace $\mathbf{S} = 4.7$.

Another important use of the hat matrix $\mathbf{S}$ is in constructing various regression diagnostics. The diagonal elements of the hat matrix are often called *leverage values*; they determine the amount by which the fitted value $\hat{y}_i$ is influenced by the particular observation y_i. If the leverage value

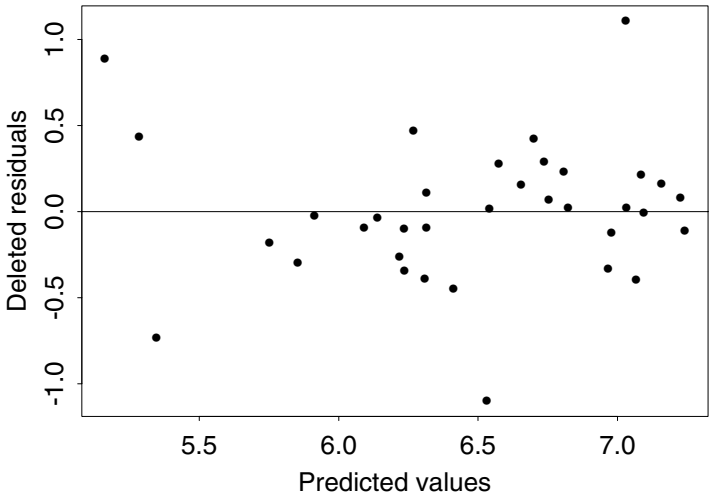

Figure 15.7. Deleted residuals from the fitted prediction of log annual precipitation from overall temperature pattern.

is particularly high, the fitted value needs to be treated with some care. Two standard ways of assessing the regression fit are to examine the raw residuals $y_i - \hat{y}_i$ and the *deleted residuals* $(y_i - \hat{y}_i)/(1 - S_{ii})$; the latter give the residual between y_i and the value predicted from the data set with case i deleted. We refer readers to works on regression diagnostics such as Cook and Weisberg (1982).

Figure 15.7 shows a plot of deleted residuals against fitted values for the log precipitation and temperature example, with the smoothing parameter chosen by cross-validation. The three observations with small predicted values have somewhat larger leverage values (around 0.4) than the others (generally in the range 0.1 to 0.2). This is not surprising, given that they are somewhat isolated from the main part of the data.

15.7 The direct penalty method for computing β

We now turn to a more direct way of using the roughness penalty approach that computes $\hat{\beta}$ direction without using basis functions. Our first task is to show how we can set up this approach as a two-stage process involving: (1) minimizing a simple quadratic expression to obtain the vector of values $\hat{y}$ approximating the data vector y, and (2) computing the smoothest linear functional interpolant of these values.

15.7.1 Functional interpolation

We have already seen that the observed data can in general be fitted exactly by an infinite number of possible parameter choices (α, β). In some contexts, it may be of interest to define a functional interpolant $(\tilde{\alpha}, \tilde{\beta})$ to the given data by the smoothest parameter function choice that fits the data exactly. In any case, we need to consider this problem in defining the technique used to compute the estimate for β in Figure 15.5. Therefore, we require that estimate $(\tilde{\alpha}, \tilde{\beta})$ minimizes $\|D^2\beta\|^2$ subject to the N constraints

$$y_i = \tilde{\alpha} + \langle x_i, \tilde{\beta} \rangle. \tag{15.17}$$

The functional interpolant is the limiting case of the regularized estimator as $\lambda \to 0$. In fact, the curve $\tilde{\beta}$ resulting from interpolating the weather data is identical to that shown in Figure 15.1.

We can consider this minimization problem (15.17) as a way of quantifying the roughness or irregularity of the response vector y relative to the observed functional covariates x_i. More generally, if $z_1, \ldots, z_N$ is any sequence of values, then we can define the roughness of z relative to the functional covariates x_i as being the roughness of the smoothest function β_z such that

$$z_i = \alpha_z + \langle x_i, \beta_z \rangle$$

for all i, for some constant α_z. This method of defining the roughness of a variate z_i will be of considerable conceptual and practical use later.

15.7.2 The two-stage minimization process

Section 15.7.3 shows that we can define an order N matrix $\mathbf{R}$ in such a way that the roughness of a variate z can be expressed as the quadratic form

$$\int [D^2\beta(s)]^2 \, ds = \mathbf{b}'\mathbf{R}\mathbf{b}.$$

Assuming this to be true for the moment, we can conceptualize the smoothing problem as being solved by dividing the minimization of the penalized residual sum of squares into two stages:

Stage 1: Find predicted values $\hat{y}$ that minimize $\texttt{PENSSE}_\lambda(\hat{y}) = \sum_i (y_i - \hat{y}_i)^2 + \lambda \hat{\mathbf{y}}'\mathbf{R}\hat{\mathbf{y}}$, the solution to which is

$$\hat{\mathbf{y}} = (\mathbf{I} + \lambda\mathbf{R})^{-1}\mathbf{y}.$$

Stage 2: Find the smoothest linear functional interpolant (α, β) satisfying

$$\hat{y}_i = \alpha + \int x_i(s)\beta(s) \, ds. \tag{15.18}$$

This two-stage procedure does indeed minimize $\texttt{PENSSE}_\lambda(\alpha, \beta)$ by the following argument. Write the minimization problem as one of first minimizing $\texttt{PENSSE}_\lambda(\alpha, \beta)$ as a function of (α, β) but with $\hat{y}$ fixed, and then

minimizing the result with respect to $\hat{y}$. Formally, this is

$$\begin{aligned} \min_{\hat{y}} \ & [\min_{\alpha,\beta}\{\texttt{PENSSE}_\lambda(\alpha,\beta)\}\ ] \\ = \ & \min_{\hat{y}}\{\sum(y_i-\hat{y}_i)^2+\lambda\min_{\beta}\int[D^2\beta(s)]^2\,ds\}, \end{aligned} \tag{15.19}$$

where the inner minimizations over α and β are carried out keeping the values of the linear functionals $\hat{y}_i$ as defined in (15.18) fixed.

But according to our assumption, these inner minimizations yield (α,β) as the smoothest functional interpolant to the variate $\hat{y}$, so we may now write the equation as

$$\texttt{PENSSE}_\lambda(\alpha,\beta)=\min_{\hat{y}}\{\sum(y_i-\hat{y}_i)^2+\lambda\hat{\mathbf{y}}'\mathbf{R}\hat{\mathbf{y}}\}. \tag{15.20}$$

Setting aside the question of how $\mathbf{R}$ is defined for a moment, one of the advantages of the roughness penalty approach to regularization is that it allows this conceptual division to be made, in a sense uncoupling the two aspects of the smoothing procedure. However, it should not be forgotten that the roughness penalty is used in the construction of the matrix $\mathbf{R}$, and so the functional nature of the covariates x_i, and the use of $\int(D^2\beta)^2$ to measure the variability of the regression coefficient function β, are implicit in both stages set out above.

We can think of the two-stage procedure in two ways: First as a practical algorithm in its own right, and second as an aid to understanding and intuition. We also see in subsequent chapters that it has wider implications than those discussed here.

In order to use the algorithm in practice, it is necessary to derive the matrix $\mathbf{R}$, and we now show how to do this.

15.7.3 Functional interpolation revisited

In this section, we present an algorithmic solution to the linear functional interpolation problem presented in Stage 2 in the two-stage procedure set out in Section 15.7.2. That is, it is of interest to find the smoothest functional interpolant $(\tilde{\alpha},\tilde{\beta})$ to a specified N-vector $\hat{y}$ relative to the given covariates $z_i, i=1,\ldots,N$. For practical purposes, our algorithm is suitable for the case where the sample size N is moderate, where matrix manipulations of $N\times N$ matrices do not present an unacceptable computational burden.

Let matrix $\mathbf{Z}$ be defined in terms of the functional covariates z_i as described in Section 15.3. In terms of basis expansions, we wish to solve the problem

$$\min\{\boldsymbol{\zeta}'\mathbf{R}\boldsymbol{\zeta}\} \ \text{ subject to } \ \mathbf{Z}\boldsymbol{\zeta}=\hat{\mathbf{y}}. \tag{15.21}$$

We first define some more notation. By rotating the basis if necessary, assume that the first M_0 basis functions ϕ_ν span the space of all functions f that have roughness $\int (D^2 f)^2 = 0$. In the Fourier case, this is true without any rotation: The only periodic functions with zero roughness are constants, so $M_0 = 1$, and the basis ϕ_ν consists of just the constant function.

Let $\mathbf{K}_2$ be the matrix obtained by removing the first M_0 rows and columns of $\mathbf{K}$. Then $\mathbf{K}_2$ is strictly positive-definite, and the rows and columns removed are all zeroes. In the Fourier case, $\mathbf{K}_2$ is diagonal.

Corresponding to the above partitioning, let $\mathbf{Z}_1$ be the matrix of the first $M_0 + 1$ columns of $\mathbf{Z}$, and let $\mathbf{Z}_2$ be the remaining columns. Defining $\mathbf{P}$ to be the $N \times N$ projection matrix $\mathbf{P} = \mathbf{I} - \mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'$ permits us to define $\mathbf{Z}^* = \mathbf{P}\mathbf{Z}_2$. In the periodic case, $\mathbf{Z}_1$ has columns $(1, \ldots, 1)$ and $(\bar{x}_1, \ldots, \bar{x}_N)$, where $\bar{x}_i = \int z_i(s)\, ds$ for each i. Thus $\mathbf{P}$ is the $N \times N$ matrix that projects any N-vector z to its residuals from its linear regression on $\bar{x}_i$.

Continuing with this partitioning process, let $\boldsymbol{\zeta}_1$ be the vector of the first $M_0 + 1$ components of $\boldsymbol{\zeta}$, and let $\boldsymbol{\zeta}_2$ be the remaining components of $\boldsymbol{\zeta}$. Then the constraint

$$\mathbf{Z}\boldsymbol{\zeta} = \mathbf{Z}_1\boldsymbol{\zeta}_1 + \mathbf{Z}_2\boldsymbol{\zeta}_2 = \hat{\mathbf{y}}$$

implies, by multiplying both sides by $\mathbf{Z}'$, that

$$\mathbf{Z}_1'\mathbf{Z}_1\boldsymbol{\zeta}_1 + \mathbf{Z}_1'\mathbf{Z}_2\boldsymbol{\zeta}_2 = \mathbf{Z}_1'\hat{\mathbf{y}}. \tag{15.22}$$

Solving for ζ_1 alone gives

$$\boldsymbol{\zeta}_1 = (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'(\hat{\mathbf{y}} - \mathbf{Z}_2\boldsymbol{\zeta}_2) \text{ and } \mathbf{Z}_1\boldsymbol{\zeta}_1 = (\mathbf{I} - \mathbf{P})(\hat{\mathbf{y}} - \mathbf{Z}_2\boldsymbol{\zeta}_2). \tag{15.23}$$

In the periodic case, equation (15.23) indicates that $\boldsymbol{\zeta}_1$ is obtained by linear regression of the values $\hat{\mathbf{y}} - \mathbf{Z}_2\boldsymbol{\zeta}_2$ on the vector with components $\bar{x}_i$. Thus, once $\boldsymbol{\zeta}_2$ has been determined, we can find $\boldsymbol{\zeta}_1$.

Now substitute solution (15.23) for $\boldsymbol{\zeta}_1$ back into the constraint (15.22) and rearrange to show that we can find $\boldsymbol{\zeta}_2$ by solving the minimization problem

$$\min_{\boldsymbol{\zeta}_2}\{\boldsymbol{\zeta}_2'\mathbf{K}_2\boldsymbol{\zeta}_2\} \text{ subject to } \mathbf{Z}^*\boldsymbol{\zeta} = \mathbf{P}\hat{\mathbf{y}} \tag{15.24}$$

using the fact that $\boldsymbol{\zeta}'\mathbf{K}\boldsymbol{\zeta} = \boldsymbol{\zeta}_2'\mathbf{K}_2\boldsymbol{\zeta}_2$.

Let $\mathbf{R}$ be defined as the Moore-Penrose g-inverse

$$\mathbf{R} = (\mathbf{Z}^*\mathbf{K}_2^{-1}\mathbf{Z}^{*\prime})^+. \tag{15.25}$$

The solution of the minimization (15.24) is then given by

$$\boldsymbol{\zeta}_2 = \mathbf{K}_2^{-1}\mathbf{Z}^{*\prime}\mathbf{R}\hat{\mathbf{y}} \tag{15.26}$$

and the minimum value of the objective function $\boldsymbol{\zeta}'\mathbf{R}\boldsymbol{\zeta}$ is therefore

$$\boldsymbol{\zeta}'\mathbf{R}\boldsymbol{\zeta} \quad = \quad \boldsymbol{\zeta}_2'\mathbf{K}_2\boldsymbol{\zeta}_2$$

$$\begin{aligned} &= \hat{\mathbf{y}}'\mathbf{R}\mathbf{Z}^{*\prime}\mathbf{K}_2^{-1}\mathbf{K}_2\mathbf{K}_2^{-1}\mathbf{Z}^*\mathbf{R}\hat{\mathbf{y}} \\ &= \hat{\mathbf{y}}'\mathbf{R}\mathbf{R}^+\mathbf{R}\hat{\mathbf{y}} \\ &= \hat{\mathbf{y}}'\mathbf{R}\hat{\mathbf{y}}. \end{aligned} \tag{15.27}$$

This is the assumption we made above in defining the two-step procedure, and moreover we have now defined the matrix $\mathbf{R}$.

We can now sum up this discussion by setting out an algorithm for functional interpolation as follows:

Step 1: Calculate matrices $\mathbf{P} = \mathbf{I} - \mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'$ and $\mathbf{Z}^* = \mathbf{P}\mathbf{Z}_2$. In effect, the columns of $\mathbf{Z}^*$ are the residuals from a standard regression of the corresponding columns of $\mathbf{Z}_2$ on the design matrix $\mathbf{Z}_1$.

Step 2: Compute $\mathbf{R}$ as defined in (15.25) above.

Step 3: Compute $\boldsymbol{\zeta}_2$ from (15.26) and use (15.23) to find $\boldsymbol{\zeta}_1$.

Of course, if all we require is the roughness of $\boldsymbol{\zeta}$, then we can find $\hat{\mathbf{y}}'\mathbf{R}\hat{\mathbf{y}}$ from (15.25) without actually calculating $\boldsymbol{\zeta}$.

Finally, returning now to our two-stage technique for smoothing, we can now carry out the first step by solving the equation

$$(\mathbf{I} + \lambda\mathbf{R})\hat{\mathbf{y}} = \mathbf{y}.$$

Note, by the way, that if $\mathbf{R}$ is either diagonal (as for the Fourier basis) or band-structured (as for the B-spline basis), that this solution is rapidly computable, and hence trying out various values for λ is quite feasible.

If we are dealing with a large data set by truncating or restricting the basis expansion to a reasonable dimensionality K as described in Section 15.3, then we only wish in general to assess the roughness of variates of the form $\mathbf{Z}\zeta$ for known $\boldsymbol{\zeta}$ with $\boldsymbol{\zeta}_j = 0$ for $j > m$. It is usually more appropriate to calculate $\boldsymbol{\zeta}'\mathbf{R}\boldsymbol{\zeta}$ for such variates directly if it is needed.

15.8 Functional regression and integral equations

Functional interpolation and regression can be viewed as a different formalization of a problem already considered in detail in Chapter 6, that of reconstructing a curve given certain indirect observations. Suppose that g is a curve of interest, and that we have noisy observations of a number of linear functionals $l_i(g)$. Such a problem was explored by Engle, Granger, Rice and Weiss (1986); see also Section 4.7 of Green and Silverman (1994). The problem involved in reconstructing the effect of temperature t on electricity consumption, so that $g(t)$ is the expected use of electricity per consumer on a day with average temperature t. Various covariates were also considered, but these need not concern us here.

Electricity bills are issued on various days and always cover the previous 28 days. For bills issued on day i, the average consumption (after correcting

for covariates) would be modelled to satisfy

$$\frac{1}{28}\mathrm{E}Y_i = \langle \theta_i, g \rangle,$$

where θ_i is the probability density function of temperature over the previous 28 day period. By setting $z_i = 28\theta_i$ and $\beta = g$, we see that this problem falls precisely into the functional regression context, and indeed the method used by the original authors to solve it corresponds precisely to the regularization method we have set out.

More generally, regularization is a very well-known tool for the solution of integral equations; see, for example, Section 12.3 of Delves and Mohamed (1985).

15.9 Further reading and notes

The subject of this chapter is probably the area in functional data analysis that has undergone the most development since the publication of the first version of this volume. The STAPH group that meets regularly at Paul Sabatier University in Toulouse has been especially active in terms of both applications and theory. To learn more about their work, consult the website `http://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html`.

Cardot, Faivre and Goulard (2003) predicted type of land use based on the evolution of the reflectance of a parcel of land in a specified wavelength over time as measured by satellite imagery. They also used functional principal components analysis to reduce the dimensionality of the reflectance curves prior to estimating the functional linear model, an approach first developed in Cardot, Ferraty and Sarda (1999) and discussed further in Cardot, Ferraty and Sarda (2003). Cardot, Goia and Sarda (2004) developed a test of the hypothesis that there is no effect on the outcome variable by the predictor variable, and Cardot, Ferraty, Mas and Sarda (2004) report further developments. Cardot, Faivre and Maisongrande (2004) use a mixed effects formulation of this model. Ferraty, Goia and Vieu (2002) forecast United States monthly electricity consumption, and Ferraty and Vieu (2002) predict the fat content of meat samples from spectrometric curves. Cardot (2002) used a roughness penalty that is similar to that used by Eilers and Marx (1996).

Escabias, Aguilera and Valderrama (2004), James (2002) and Cardot and Sarda (2004) look at the larger problem of how to adapt the generalized linear model to the presence of a functional predictor variable, and offer a number of examples, including the situation considered here of a continuous dependent variable. Escabias et al. (2004) combine the functional linear model with principal components analysis to reduce the dimensionality of the covariate space. James (2002) also describes an interesting method for estimating the between-curve variation as well as the

within-curve structure. Müller and Stadtmüller (2004) also investigate that they call the *generalized functional linear model*. James and Hastie (2001) consider linear discriminant analysis where at one of the of independent variables used for prediction is a function, and where the curves are irregularly sampled. Ratcliffe, Leader and Heller (2002) and Ratcliffe, Heller and Leader (2002) use the functional covariate foetal heart rate to model continuous and binary outcome variables.