MA 751
M. Kon

Data Assignment 2
Due Tues. April 5

(a) Analyze the ozone data (this is not the 'Los Angeles ozone data') on the website of the Hastie book. Do a predictive analysis of ozone in terms of the other three variables using regression splines with fixed knots. Use the example from the South African heart disease data (Section 5.2.2) as a very rough model for your analysis.

(b) Now try natural splines (with fixed knots).

(c) Now try the same things using smoothing splines. This time do three analyses, of ozone in terms of each other variable individually.
(Optional) How could you do an analog of part (b) above using smoothing splines?

You can use your own program, or implement existing functions, e.g., in R or Matlab. Please include all graphs/diagrams and the code or commands you have used.

Please also explain in detail what the algorithm you used does - i.e., what the idea and pseudocode are.

Suggestions

(a) For this part you should try to include the confidence band around the estimated functions. Note your model here is

$$Y = \theta_0 + \mathbf{h}_1(X_1)^T\boldsymbol{\theta}_1 + \mathbf{h}_2(X_2)^T\boldsymbol{\theta}_2 + \mathbf{h}_3(X_3)^T\boldsymbol{\theta}_3$$

$$= \theta_0 + f_1(X_1) + f_2(X_2) + f_3(X_3),$$

where $f_i(X_i) = \mathbf{h}_i(X_i)^T\theta_i$. Assuming you arrange your spline functions in the same way as the example (but here for standard regression, not natural, splines), then, e.g.,

$$\mathbf{h}_1(X_1) = \begin{bmatrix} h_{11}(X_1) \\ h_{12}(X_1) \\ h_{13}(X_1) \\ h_{14}(X_1) \end{bmatrix}, \quad \boldsymbol{\theta}_1 = \begin{bmatrix} \theta_{11} \\ \theta_{12} \\ \theta_{13} \\ \theta_{14} \end{bmatrix}.$$

assuming you want a combination of 4 natural spline basis functions for your first coordinate X_1. Thus the first coordinate function in your model would be

$$\mathbf{h}_1(X_1)^T\boldsymbol{\theta}_1 = \theta_{11}h_{11}(X_1) + \ldots + \theta_{14}h_{14}(X_1).$$

with similar forms for the other ones.

Putting things together note that you can consider the union of all the above functions $\{h_{ij}(X_j)\}_{i,j}$ as a single large basis $\{h_k(X)\}_k$ where $X = (X_1, X_2, X_3)$; (note in reality all the functions $h_k(X_1, X_2, X_3)$ are functions of single variables X_i like the ones listed above). If you add the constant function to this basis, then all of the standard basis methods we have used will apply.

However, there is one caveat: at this point the basis functions $\{h_k(X)\}$ are not linearly independent. This is because the constant function 1 can be represented in multiple ways by these (i.e. as a function of X_1 or of X_2, etc.). Notice in fact that the functions $h_{11}, \ldots, h_{14}$ (or however many you have in the first group) should contain a constant function among them (since the constant function is one of the functions in a standard regression spline basis; see, e.g., eq. (5.3)). The best thing to do is to eliminate the contant function from each group, and include all constants in the single term θ_0.

With this new set of basis functions you can proceed with the standard methods basis methods in Chapter 5 (see, e.g. eq. (5.15) which holds for any basis). Specifically (including now the constant function as one of the h_k), we have

$$Y = \mathbf{h}(\mathbf{x})^T \boldsymbol{\theta} = f(\mathbf{x}),$$

so that from our linear model the solution follows exactly as in standard linear regression, in which $h_k(X) = X_k$.

Thus we have

$$\widehat{\mathbf{y}} = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{y} \equiv \mathbf{H}\mathbf{y},$$

where

$$B_{ki} = h_k(\mathbf{x}_i),$$

and **H** is the well-known hat matrix.

You can also make plots similar to those in Figure 5.4. Similarly to what is done in the text in section 5.2.2 for logistic regression, note

$$\widehat{\theta} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{y} \equiv \mathbf{J}\mathbf{y},$$

where

$$\mathbf{J} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T.$$

Therefore

$$\widehat{\Sigma} = V(\widehat{\theta}) = \mathbf{J}V(\mathbf{y})\mathbf{J}^T,$$

where $V(\mathbf{y}) = \sigma^2 I$ is the covariance matrix of the responses y_i at the data points $\mathbf{x}_i$.

Now plot the estimated function $f_1(X_1)$ in a way similar to the plots in Figure 5.4. To obtain error bars, follow a procedure similar to that in Section 5.2.2. Namely, for $f_1(X_1)$, show you can obtain

$$V(\widehat{f}_1(X_1)) = \mathbf{h}_1^T(X_1)\widehat{\Sigma}_{11}\mathbf{h}_1(X_1),$$

where $\widehat{\Sigma}_{11}$ is the sub-matrix of $\widehat{\Sigma}$ containing only the four rows and columns corresponding to the functions $f_{11}(X_1), \ldots, f_{14}(X_1)$. Now compute error bars above and below the curve at different points X_1 to obtain the error region above and below the curve. This can be done also for $f_2(X_2)$ and $f_3(X_3)$.

(b) Note that though you are using natural splines, the knots will still be fixed. You will still want to eliminate the constant function from the natural spline basis in each coordinate.

(c) Now recall that smoothing splines are nothing but natural splines with knots at the data points. Here you just need to do three univariate approximations of the ozone level. For the second (optional) part of the question, recall that smoothing splines have a basis set defined in the text. Note you will now have three regularization parameters λ_1, λ_2, λ_3.

To estimate the ozone as a function of all three other variables, you can use the iterative method described in section 9.1.1. Recall though you will need adjust your basis functions so that they have 0 means here as well, for the same reason.