Smoothing Splines

Cubic splines, with their cont. 1st and 2nd derivatives insuring smoothness (and saving a few parameters), tend to provide a decent approximation to $E(Y|x)$, provided that the number and locations of the knots are chosen well. In order to avoid the knot selection problem, one can use a *smoothing spline*, which is a natural cubic spline with knots at each observed value of the explanatory variable.

As motivation for smoothing splines, let $f_\lambda(x)$ be an approximation of $E(Y|x)$ which has two continuous derivatives and which minimizes

$$\sum_{i=1}^{n} [\, y_i - f(x_i)]^2 + \lambda \int [f''(t)]^2 dt,$$

where λ is a *smoothing parameter* which controls the degree of roughness allowed in the approximation of $E(Y|x)$. If $\lambda = 0$, f_λ can be any function that interpolates the data, including functions which are unrealistically wiggly. (If more than one value of Y is observed at a value x, the interpolation should go through $(x, \bar{y})$, where $\bar{y}$ is the sample mean of the response values observed at x.) If $\lambda = \infty$, f_λ must be linear (i.e., having the simple form $b_0 + b_1 x$), since the 2nd derivative must be 0 everywhere. For $\lambda \in (0, \infty)$, the minimizer is a natural cubic spline having a knot at each distinct x value, but with coefficients fitted to minimize

$$\sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int [f''(t)]^2 \, dt,$$

and not just the sum of the squared errors.

The goal is to choose λ so that the fitted $f_\lambda(x)$ is the best estimate of $E(Y|x)$, or the best prediction formula for future values of Y at given values of x. One can use cross-validation to estimate the value of λ which minimizes the expected squared prediction error for the prediction of response values for cases not used to fit the $f_\lambda(x)$. (We seek the value of λ which trades off between bias and variance in such a way as to minimize the mean squared prediction error. (Increasing λ decreases variance by limiting wiggleness, but this can increase bias.)) An *effective degrees of freedom* corresponding to the chosen value of λ can be determined, but I don't consider this to be too important — the main thing to focus on is that the selected value of λ is the one which was judged by cross-validation to give the best performance. (Note: If multiple observations of Y are given for any x value(s) in the training data, then a generalized cross-validation procedure should be used instead of a routine n-fold cross-validation.)

An extension to the multiple regression setting is not so trivial. For multiple regression, I recommend using fixed knots, and few of them.