Cubic Splines

Instead of approximating $E(Y|x)$ with a piecewise linear function (e.g., a continuous linear spline), it might be better to use a piecewise cubic function. An approximation of the form

$$\widehat{E(Y|x)} = \begin{cases} b_0 + b_1 x + b_2 x^2 + b_3 x^3, & x \leq \xi, \\ b_4 + b_5 x + b_6 x^2 + b_7 x^3, & x > \xi, \end{cases}$$

is one possibility, and can be based on the basis functions

$$h_0(x) = 1,$$
$$h_1(x) = x,$$
$$h_2(x) = x^2,$$
$$h_3(x) = x^3,$$
$$h_4(x) = I_{(0,\infty)}(x-\xi),$$
$$h_5(x) = (x-\xi)_+,$$
$$h_6(x) = (x-\xi)_+^2,$$
& $$h_7(x) = (x-\xi)_+^3,$$

but with this choice one has to estimate eight uhknown parameters, and the approximating function may be discontinuous at ξ.

If one removes h_4 from the set, the resulting function will be continuous, but it may not be very smooth, and there are still 7 parameters to estimate.

If one uses the basis set

$$h_0(x) = 1,$$
$$h_1(x) = x,$$
$$h_2(x) = x^2,$$
$$h_3(x) = x^3,$$
& $$h_4(x) = (x - \xi)^3_+,$$

then there are only five parameters to estimate, and not only is the resulting piecewise cubic function continuous, but it is also smooth — it has continuous first and second derivatives. (Note: ξ is considered to be fixed, and is not something to be estimated.) See p. 119 of HTF for some nice plots (Fig. 5.2).

The term *cubic spline* generally refers to a continuous piecewise cubic function with continuous first and second derivatives.

$$h_0(x) = 1,$$
$$h_1(x) = x,$$
$$h_2(x) = x^2,$$
$$h_3(x) = x^3,$$
$$h_4(x) = (x - \xi_1)_+^3,$$
&
$$h_5(x) = (x - \xi_2)_+^3$$

is a basis set for a cubic spline having knots at ξ_1 and ξ_2.

When using OLS regression to fit a cubic spline approximation of $E(Y|x)$, the variance of the estimator can be annoyingly large if x is near an end of the range of the x values in the data. If one instead fits a natural cubic spline (aka restricted cubic spline), which is like an ordinary cubic spline except that it has linear segments beyond the boundary knots (which are the smallest and largest knots), then the variance near the ends can be appreciably reduced. The bias near the ends can be increased, but this is somewhat offset by the fact that a natural cubic spline having the same number of parameters as a cubic spline can have four more knots than the cubic spline, and these additional knots help to reduce bias, and by placing knots near the smallest and largest x values, they can serve to control the bias near the ends.

Frank Harrell, author of *Regression Modeling Strategies*, strongly advocates using natural cubic splines to represent a variable in a regression model when there is adequate data. Since a natural cubic spline having k knots contributes k parameters which have to be estimated, one could choose between a natural cubic spline having four knots (and so three cubic pieces and two linear pieces) and a single cubic polynomial expansion to represent a variable. Or instead of a single 4th-degree polynomial, one could use a natural cubic spline having five knots (and so four cubic pieces and two linear pieces) and contribute the same number of parameters to the model. (Does this seem too good to be true?) In a multiple regression setting, one version of a *generalized additive model* (GAM) represents the effect of each continous variable with an appropriate spline term (making adjustments to not include an intercept term more than once).

It should be noted that the basis sets given previously for cubic splines are not the ones preferred in practice. Other choices, which are equivalent theoretically, are favored due to numerical considerations. (I think it best to avoid the grubby details at this time (and my advice will be the same next year).)