**Chapter 7: Moving beyond linearity**

* However, standard linear regression can have significant limitations in terms of predictive power. This is because the linearity assumption is almost always an approximation, and sometimes a poor one.
* Polynomial regression extends the linear model by adding extra predictors, obtained by raising each of the original predictors to a power. For example, a cubic regression uses three variables, X, X2, and X3, as predictors
* Step functions cut the range of a variable into K distinct regions in order to produce a qualitative variable. This has the effect of fitting a piecewise constant function
* Regression splines are more flexible than polynomials and step functions, and in fact are an extension of the two. They involve dividing the range of X into K distinct regions. Within each region, a polynomial function is fit to the data. However, these polynomials are constrained so that they join smoothly at the region boundaries, or knots.
* Smoothing splines are similar to regression splines, but arise in a slightly different situation. Smoothing splines result from minimizing a residual sum of squares criterion subject to a smoothness penalty.
* Local regression is similar to splines, but differs in an important way. The regions are allowed to overlap, and indeed they do so in a very smooth way
* Generalized additive models allow us to extend the methods above to deal with multiple predictors

*Polynomial regression*

* For large regression enough degree d, a polynomial regression allows us to produce an extremely non-linear curve.
* Generally speaking, it is unusual to use d greater than 3 or 4 because for large values of d, the polynomial curve can become overly flexible and can take on some very strange shapes.

*Step functions*

* Using polynomial functions of the features as predictors in a linear model imposes a global structure on the non-linear function of X. We can instead use step functions in order to avoid imposing such a global structure. Here step function we break the range of X into bins, and fit a different constant in each bin. This amounts to converting a continuous variable into an ordered categorical variable
* Unfortunately, unless there are natural breakpoints in the predictors, piecewise-constant functions can miss the action

*Part 1: Regression slines*

*Piecewise polynomials*

* Instead of fitting a high-degree polynomial over the entire range of X, piecewise polynomial regression involves fitting separate low-degree polynomials piecewise polynomial regression over different regions of X.
* In other words, we fit two different polynomial functions to the data, one on the subset of the observations with xi < c, and one on the subset of the observations with xi ≥ c.
* Using more knots leads to a more flexible piecewise polynomial. In general, if we place K different knots throughout the range of X, then we will end up fitting K + 1 different cubic polynomials.

*Constraints and splines*

* To remedy the problem of too flexible curve, we can fit a piecewise polynomial, under the constraint that the fitted curve must be continuous.
* In other words, we are requiring that the piecewise polynomial be not only continuous when age=50, but also very smooth.
* Just as there were several ways to represent polynomials, there are also many equivalent ways to represent cubic splines using different choices of basis functions in (7.9). The most direct way to represent a cubic spline using (7.9) is to start off with a basis for a cubic polynomial—namely, x, x2, x3—and then add one truncated power basis function per knot.
* Unfortunately, splines can have high variance at the outer range of the predictors—that is, when X takes on either a very small or very large value.
* A natural spline is a regression spline with additional boundary constraints: the natural function is required to be linear at the boundary (in the region where X is spline smaller than the smallest knot, or larger than the largest knot). This additional constraint means that natural splines generally produce more stable estimates at the boundaries.

*Choosing the number and location of knots*

* The regression spline is most flexible in regions that contain a lot of knots, because in those regions the polynomial coefficients can change rapidly.
* One option is to try out different numbers of knots and see which produces the best looking curve. A somewhat more objective approach is to use cross-validation, as discussed in Chapters 5 and 6. W

*Comparison to polynomial regression*

* Regression splines often give superior results to polynomial regression. This is because unlike polynomials, which must use a high degree (exponent in the highest monomial term, e.g. X15) to produce flexible fits, splines introduce flexibility by increasing the number of knots but keeping the degree fixed. Generally, this approach produces more stable estimates. Splines also allow us to place more knots, and hence flexibility, over regions where the function f seems to be changing rapidly, and fewer knots where f appears more stable

*Smoothing splines*

* We have seen that a smoothing spline is simply a natural cubic spline with knots at every unique value of xi. It might seem that a smoothing spline will have far too many degrees of freedom, since a knot at each data point allows a great deal of flexibility. But the tuning parameter λ controls the roughness of the smoothing spline, and hence the effective degrees of freedom.
* In the context of smoothing splines, why do we discuss effective degrees of freedom instead of degrees of freedom?
* Although a smoothing spline has n parameters and hence n nominal degrees of freedom, these n parameters are heavily constrained or shrunk down. Hence dfλ is a measure of the flexibility of the smoothing spline—the higher it is, the more flexible (and the lower-bias but higher-variance) the smoothing spline. The definition of effective degrees of freedom is somewhat technical.
* In fitting a smoothing spline, we do not need to select the number or
* location of the knots—there will be a knot at each training observation, x1, . . . , xn. Instead, we have another problem: we need to choose the value of λ.
* It should come as no surprise that one possible solution to this problem is cross-validation. In other words, we can find the value of λ that makes the cross-validated RSS as small as possible. It turns out that the leave- one-out cross-validation error (LOOCV) can be computed very efficiently for smoothing splines

*Local regression*

* Local regression is a different approach for fitting flexible non-linear functions, which involves computing the fit at a target point x0 using only the nearby training observations.
* Local regression is sometimes referred to as a memory-based procedure, because like nearest-neighbors, we need all the training data each time we wish to compute a prediction.
* However, local regression can perform poorly if p is much larger than about 3 or 4 because there will generally be very few training observations close to x0.

*Part Four: Generalised Additive Models*

* The approaches above can be seen as extensions of simple linear regression, in that they deal with the relationship between only two variables, X and Y. Here we explore the problem of flexibly predicting Y on the basis of several predictors, X1, . . . , Xp. This amounts to an extension of multiple linear regression
* Generalized additive models (GAMs) provide a general framework for extending a standard linear model by allowing non-linear functions of each ofthevariables,whilemaintainingadditivity. Justlikelinearmodels,GAMs can be applied with both quantitative and qualitative responses
* GAMs provide a useful compromise between linear and fully nonparametric models.

*Pros of GAMs*

* GAMs allow us to fit a non-linear fj to each Xj, so that we can automatically model non-linear relationships that standard linear re- gression will miss.
* Because the model is additive, we can still examine the effect of each Xj on Y individually while holding all of the other variables fixed.
* The smoothness of the function fj for the variable Xj can be summarized via degrees of freedom.

*Cons of GAMs*

* The main limitation of GAMs is that the model is restricted to be additive. With many variables, important interactions can be missed. However, as with linear regression, we can manually add interaction terms to the GAM model