- *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, X^2, and X^3, as predictors. This approach provides a simple way to provide a non-linear fit to data.

- *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*. Provided that the interval is divided into enough regions, this can produce an extremely flexible fit.

- *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.

7.1 Polynomial Regression

7.2 Step Functions

7.3 Basis Functions

- the idea is to have at hand a family of functions or transformations that can be applied to a variable X: $b_i(X)$, $i = 1, 2, 3, \dots, k$ ($\underbrace{b_i(\ )}_{\text{basis functions}}$ being functions or transformations)

  $y_i = \beta_0 + \beta_1 b_1(X_i) + \beta_2 b_2(X_i) + \dots + \beta_K b_K(X_i) + \varepsilon_i$

- We may choose any functions as our basis functions

7.4 Regression Splines

7.4.1 Piecewise Polynomials

- involves fitting separate low-degree polynomials over different regions of X.

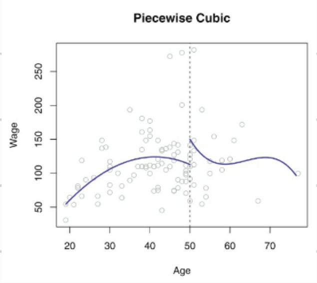
=> $y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_K X_i^k$, where the coefficients $\beta_0, \beta_1, \dots, \beta_K$ differ in different parts of the range of X.

$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$

Knots: the points where the coefficients change

(usually set as quantiles (1^{st}, 2^{nd}, 3^{rd} quartiles)

- As the number of knots increases, the model complexity also increases.
- if we place K different knots, then we will end up fitting K+1 different cubic polynomials.



- the function may not be continuous at each knot.

7.4.2 Constraints and Splines

- In cases there are jumps at knots, we can fit a piecewise polynomial under the constraint that the fitted curve must be continuous.
- Each constraint that we impose on the piecewise cubic polynomials effectively frees up one degree of freedom, by reducing the complexity of the resulting piecewise polynomial fit.

7.4.2 The Spline Basis Representation

- It turns out that we can use the basis model to represent a regression spline.
  $\Rightarrow$ $y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \varepsilon_i$
- start off with a basis for a cubic polynomial and then add one truncated power basis function per knot.
  $\Rightarrow$ $h(x, \xi) = (x-\xi)^3_+ = \begin{cases} (x-\xi)^3, & x > \xi \\ 0, & \text{otherwise} \end{cases}$, where ξ is the knot.
- In cubic spline case, we may add $\beta_K h(x, \xi_K)$ terms to the cubic model. Then, we perform least squares regression with an intercept and $3+K$, so this amounts to estimating a total of $K+4$ regression coefficients, and fitting a cubic spline with K knots uses $K+4$ degrees of freedom.
- Unfortunately, splines can have high variance at the outer range of the predictors.
  $\Rightarrow$ "Natural Splines" is a regression spline with additional boundary constraints.
  - the function where X's are smaller than the smallest knot or larger than the largest knot is required to be linear at the boundary.

7.4.4 Choosing the Number and Locations of the Knots

- specifying n degrees of freedom means the model has n-1 knots

7.4.5 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. X^{15}) 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. Figure 7.7 compares a natural cubic spline with 15 degrees of freedom to a degree-15 polynomial on the Wage data set. The extra flexibility in the polynomial produces undesirable results at the boundaries, while the natural cubic spline still provides a reasonable fit to the data.

7.5 Smoothing Splines

7.5.1 An Overview of Smoothing Splines

- In fitting a model, we want to find the function $g(x)$ that minimizes $RSS = \sum_{i=1}^{n}(y_i - g(x))^2$. However, if we don't put any constraints on $g(x)$, then we can always make $RSS = 0$ simply by increasing the complexity of the model.

- Smoothing Splines

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \underbrace{\lambda \int g''(t)^2 \, dt}_{\text{penalty}},$$ where λ is a nonnegative tuning parameter.

* $g'(t)$ measures the slope of a function t, and $\underbrace{g''(t)}_{\text{measure of roughness}}$ corresponds to the amount by which the slope is changing.

** $g''(t)$ is large in absolute value if $g(t)$ is very wiggly near t.

특정 점 주위에 model의 variance가 높은 애들 순서로 coefficient를 제한하는건가?

7.5.2 Choosing the Smoothing Parameter λ

- 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.

7.7 Generalized Additive Models

- provide a general framework for extending a standard linear model by allowing non-linear functions of each of the variables, while maintaining additivity.
- can be used for both quantitative / qualitative data.

7.7.1 GAMs for Regression Problems

A natural way to extend the multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

in order to allow for non-linear relationships between each feature and the response is to replace each linear component $\beta_j x_{ij}$ with a (smooth) non-linear function $f_j(x_{ij})$. We would then write the model as

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \\ &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i. \end{aligned} \tag{7.15}$$

This is an example of a GAM. It is called an *additive* model because we calculate a separate f_j for each X_j, and then add together all of their contributions.

- backlifting :

  ⇒ fits a model involving multiple predictors by repeatedly updating the fit for each predictor in turn, holding the others fixed.

Pros and Cons of GAMs

▲ GAMs allow us to fit a non-linear f_j to each X_j, so that we can automatically model non-linear relationships that standard linear regression will miss. This means that we do not need to manually try out many different transformations on each variable individually.

▲ The non-linear fits can potentially make more accurate predictions for the response Y.

▲ Because the model is additive, we can still examine the effect of each X_j on Y individually while holding all of the other variables fixed. Hence if we are interested in inference, GAMs provide a useful representation.

▲ The smoothness of the function f_j for the variable X_j can be summarized via degrees of freedom.

◆ 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 by including additional predictors of the form $X_j \times X_k$. In addition we can add low-dimensional interaction functions of the form $f_{jk}(X_j, X_k)$ into the model; such terms can be fit using two-dimensional smoothers such as local regression, or two-dimensional splines (not covered here).

~~7.7.2 GAMs for Classification Problems~~