7. Moving Beyond Linearity

So far, we have mostly focused on linear models.

However, linearity assumption can have significant limitations in terms of predictive power.

Summary

- Polynomial regression
- Step functions
- Regression Splines
- Smoothing Splines
- Local Regression
- Generalized Additive Models

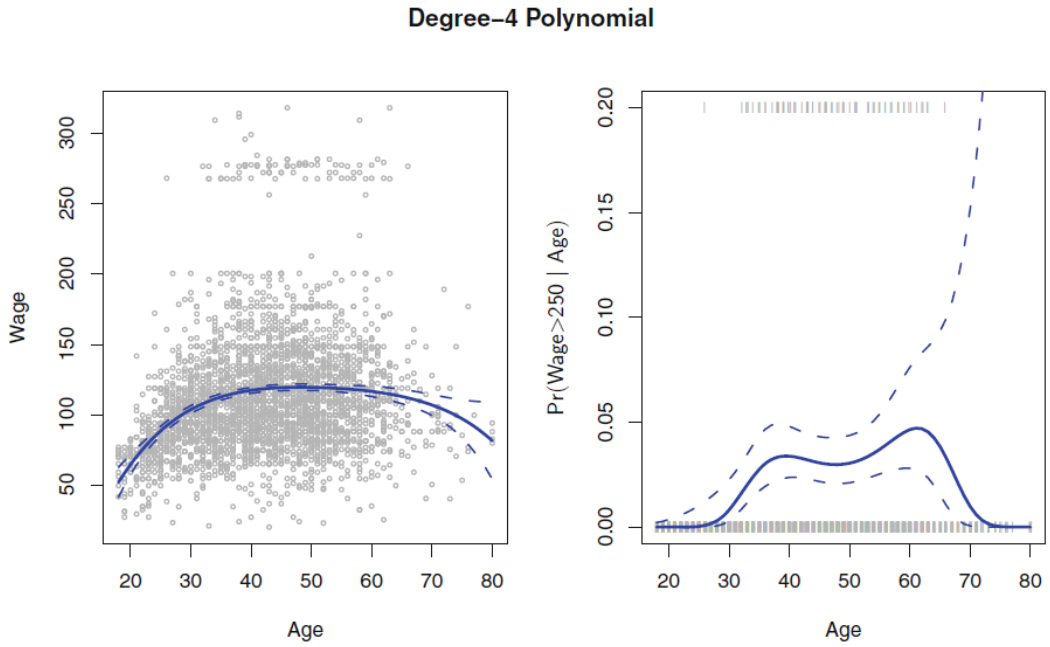
7.1 Polynomial Regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$ (standard linear model)

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i$$ (polynomial model)

Generally, 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.



$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4.$$ (polynomial model)

$$\Pr(y_i > 250|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d)}.$$ (logistic model)

What is the variance of the fit, i.e. Variance of estimated f(x0)?

7.2 Step functions

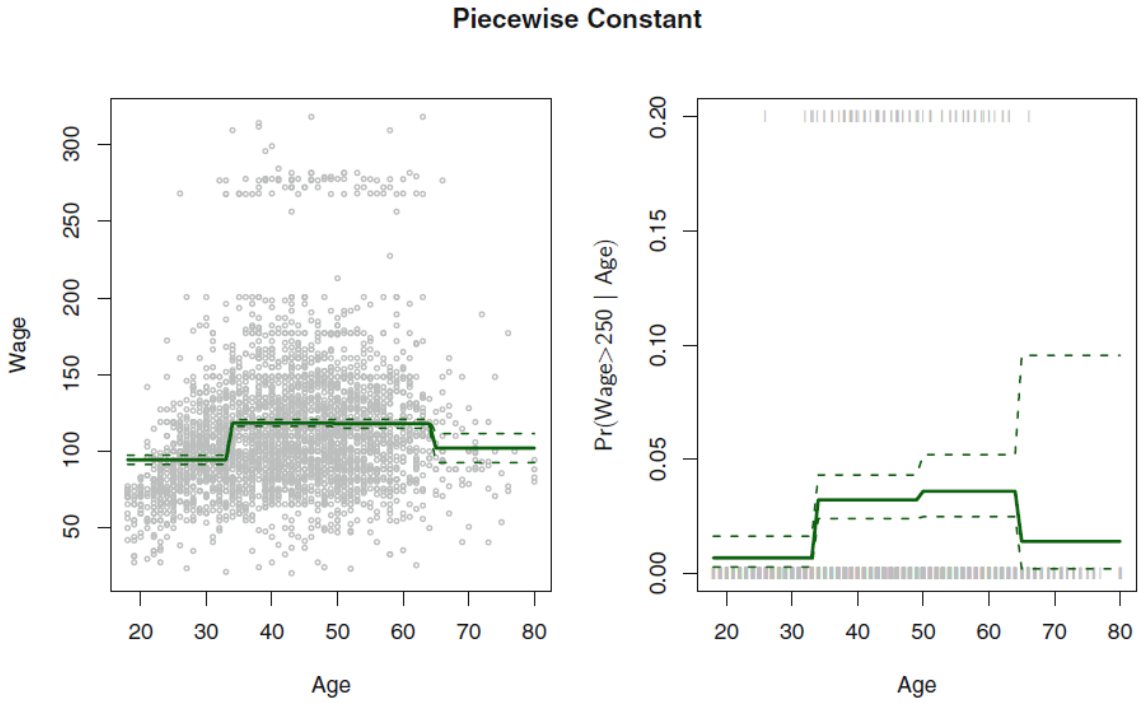
$$\begin{aligned} C_0(X) &= I(X < c_1), \\ C_1(X) &= I(c_1 \le X < c_2), \\ C_2(X) &= I(c_2 \le X < c_3), \\ &\vdots \\ C_{K-1}(X) &= I(c_{K-1} \le X < c_K), \\ C_K(X) &= I(c_K \le X), \end{aligned}$$

where $I(\cdot)$ is an *indicator function*

Create cutpoints c1, c2, . . . , cK in the range of X, and then construct K + 1 new variables

These are sometimes called dummy variables.

$$C_0(X) + C_1(X) + \ldots + C_K(X) = 1$$



$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \ldots + \beta_K C_K(x_i) + \epsilon_i$$ (polynomial model)

$$\Pr(y_i > 250|x_i) = \frac{\exp(\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i))}{1 + \exp(\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i))}$$ (logistic model)

7.3 Basis Functions

Polynomial and piecewise-constant regression models are special cases of a basis function approach.

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \ldots + \beta_K b_K(x_i) + \epsilon_i$$

$b_1(X), b_2(X), \ldots, b_K(X)$ are transformations of variable X.

For polynomial, $b_j(x_i) = x_i^j$

For step functions, $b_j(x_i) = I(c_j \leq x_i < c_{j+1})$

7.4. Regression Splines

7.4.1 Piecewise Polynomials

Separate low-degree polynomials over different regions of X.

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.

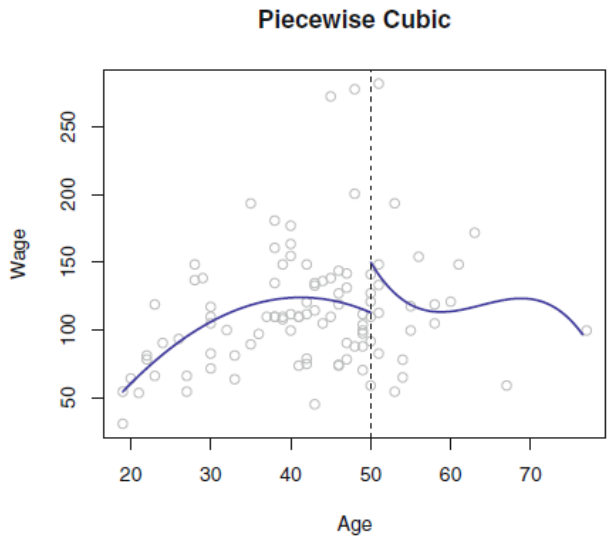
$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$

(A piecewise cubic polynomial with a single knot at a point c)

There are 8 coefficients. Each of these polynomial functions can be fit using least squares

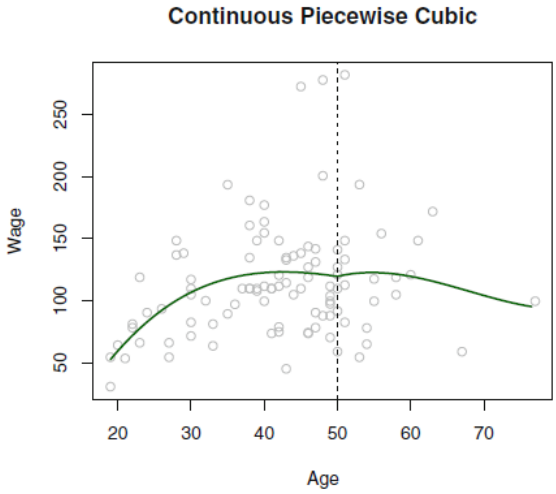
If we place K different knots throughout the range of X, then we will end up fitting K + 1 different cubic polynomials.

문제점! discontinuous

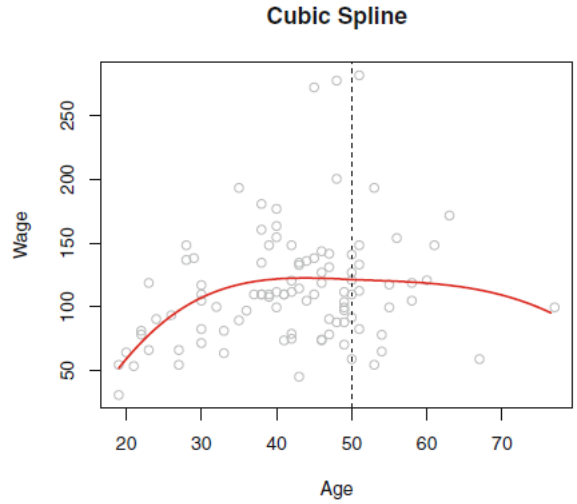


7.4.2 Constraints and Splines

Constraint 1 : continuous at age 50



Constraint 2 : smooth(the first and second derivatives of the piecewise polynomials are continuous)



This plot is called a cubic spline. In general, a cubic spline with K knots uses a total of 4 + K degrees of freedom.

- Degree of Freedom

The general definition of a *degree-d* spline is that it is a piecewise *degree-d* polynomial, with continuity in derivatives up to *degree d − 1* at each knot.

7.4.3. Splines Basis Representation

How can we fit a piecewise *degree-d* polynomial under the constraint?

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i$$

(cubic spline with K knots)

The most direct way to represent a cubic spline is to start off with a basis for a cubic polynomial—namely, x, x^2, x^3—and then add one truncated power basis function per knot.

$$h(x,\xi) = (x-\xi)_+^3 = \begin{cases} (x-\xi)^3 & \text{if } x > \xi \\ 0 & \text{otherwise} \end{cases}$$

(truncated power basis function)

Because of

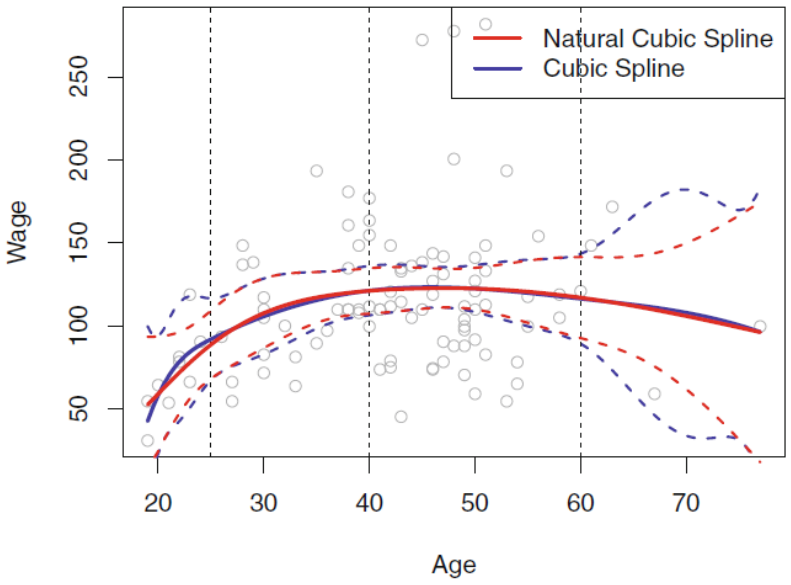
h(x) : smooth

1차 미분 : constant

2차 미분 : zero

So, there are K+3 variables, $X, X^2, X^3, h(X,\xi_1), h(X,\xi_2), \ldots, h(X,\xi_K)$

With intercept term, this uses K+4 degrees of freedom.



7.4.4 Choosing the Number and Locations of the Knots

How many knots? Where should we place the knots?

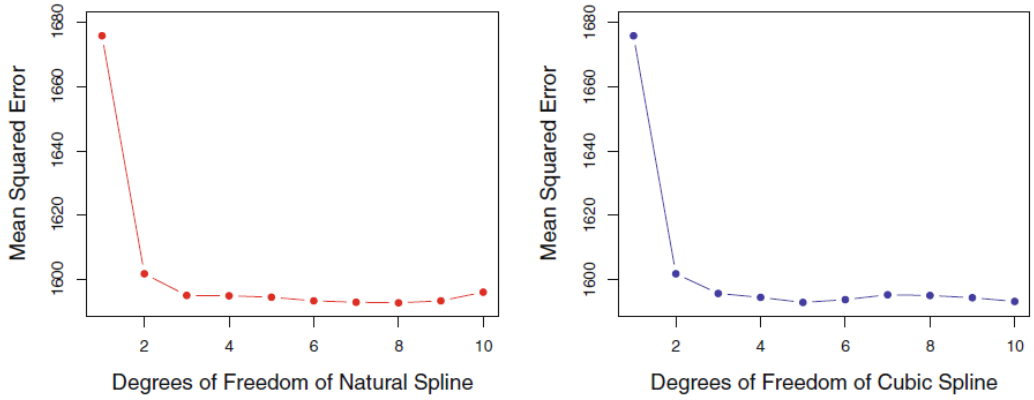
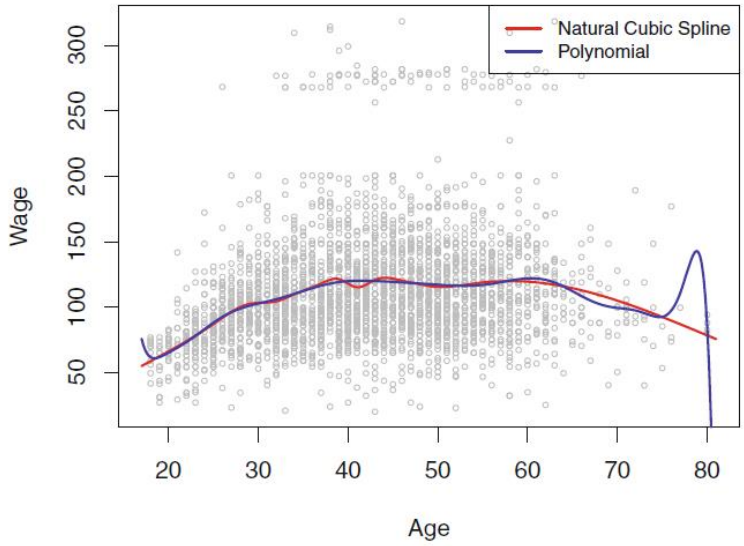


FIGURE 7.6. *Ten-fold cross-validated mean squared errors for selecting the degrees of freedom when fitting splines to the* `Wage` *data. The response is* `wage` *and the predictor* `age`. Left: *A natural cubic spline.* Right: *A cubic spline.*

Try out different numbers of knots and see which produces the best looking curve. A somewhat

More objective approach is to use **cross-validation**

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

Generally, spline produces more stable estimates.

7.5 Smoothing Splines

7.5.1 An Overview of Smoothing Splines

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

where λ is a nonnegative *tuning parameter* (another way to represent splines)

The function g that minimizes above equation is known as a **smoothing spline.**

"Loss + Penalty"

Smoothing Splines vs. Natural cubic spline

- with knots at all x
- it is not the same
- the value of the tuning parameter λ controls the level of shrinkage.

Smoothing spline is simply a natural cubic spline with knots at every unique value of x

7.5.2 Choosing the Smoothing Parameter λ

Problem of df -> effective degrees of freedom

$$\hat{\mathbf{g}}_\lambda = \mathbf{S}_\lambda \mathbf{y},$$

$$df_\lambda = \sum_{i=1}^{n} \{\mathbf{S}_\lambda\}_{ii},$$

(effective degrees of freedom)

is a measure of the flexibility of the smoothing spline—the higher it is, the more flexible

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

$$\mathrm{RSS}_{cv}(\lambda) = \sum_{i=1}^{n}(y_i - \hat{g}_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^{n}\left[\frac{y_i - \hat{g}_\lambda(x_i)}{1 - \{\mathbf{S}_\lambda\}_{ii}}\right]^2$$ (efficient way of calculating)

Proof 1)

$$\begin{aligned} Q(\beta) &= (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) \\ &= \mathbf{Y}^T\mathbf{Y} - \beta^T\mathbf{X}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta \\ &= \mathbf{Y}^T\mathbf{Y} - 2\beta^T\mathbf{X}^T\mathbf{Y} + \beta^T\mathbf{X}^T\mathbf{X}\beta \\ \frac{\partial}{\partial\beta}Q(\beta) &= -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\beta = \mathbf{0} \end{aligned}$$

Thus

$$\mathbf{b} = \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \underbrace{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{\mathbf{H}}\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

Proof 2)

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_{i(i)})^2 = \sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1-h_{ii}}\right)^2$$

proof. for $y_k - y_{k(k)} = \frac{y_k - \hat{y}_k}{1-h_{kk}}$

① Consider $y = f(\underline{x}) + \varepsilon = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1}x_{p-1} + \varepsilon$ — (A) (linear reg.) with dataset $(\underline{x}_1, y_1), \cdots, (\underline{x}_n, y_n)$ (x_i's are vectors.)

② LSE gives $\hat{f} = \underset{f}{\operatorname{argmin}}\{\sum(y_i - f(\underline{x}_i))^2\} = Hy$ — (B)

i.e. $\sum(y_i - \hat{f}(\underline{x}_i))^2 \le \sum(y_i - f(\underline{x}_i))^2$ for $\forall f$ of the form (A) : LS의 optimal 조건이 가장 작다.

③ Now consider the fitted regression without the k-th case. That is, $\hat{f}^{(k)} = \underset{f}{\operatorname{argmin}}\{\sum_{i\neq k}(y_i - f(\underline{x}_i))^2\}$ — (C)

We further denote $\hat{f}^{(k)}(\underline{x}_k) = \hat{y}_{k(k)}$

④ Then, (k번째 빼고의 SSE.)

$$\sum_{i\neq k}(y_i - \hat{f}^{(k)}(\underline{x}_i))^2 + (\hat{y}_{k(k)} - \hat{f}^{(k)}(\underline{x}_k))^2 \le \sum_{i\neq k}(y_i - f(\underline{x}_i))^2 \quad \text{for } \forall f \text{ of the form (A), definition of LSE}$$

($\hat{f}^{(k)}(\underline{x}_k) - \hat{y}_{k(k)} = 0$)

$$\le \sum_{i\neq k}(y_i - f(\underline{x}_i))^2 + (\hat{y}_{k(k)} - f(\underline{x}_k))^2$$

(≥ 0, LS가 아닌 다른 추정량)

⑤ The inequality in ④ means

$$\sum_{i\neq k}(y_i - \hat{f}^{(k)}(\underline{x}_i))^2 + (\hat{y}_{k(k)} - \hat{f}^{(k)}(\underline{x}_k))^2 \le \sum_{i\neq k}(y_i - f(\underline{x}_i))^2 + (\hat{y}_{k(k)} - f(\underline{x}_k))^2 \quad \text{for } \forall f \text{ of the form (A)} \quad \text{— (D)}$$

($\hat{f}^{(k)}(\underline{x}_k) = 0$, ≥ 0)

This implies that $\hat{f}^{(k)}$ is the LSE on the data with the k-th case (x_k, y_k) replaced with $(x_k, \hat{y}_{k(k)})$

($\sum_{i\neq k} \to \sum_{i=1}^{n}$ 을 위해 작업.)

⑥ In other words, if we define, $\tilde{y} = (\tilde{y}_1, \cdots, \tilde{y}_n)' = (y_1 \cdots \hat{y}_{k(k)} \cdots y_n)'$ (k번째만 바꿈 $y \to \tilde{y}$)

then (D) can be $\sum_{i=1}^{n}(\tilde{y}_i - \hat{f}^{(k)}(\underline{x}_i))^2 \le \sum_{i=1}^{n}(\tilde{y}_i - f(x_i))^2$ — (E) for $\forall f$ of the form (A)

$\Rightarrow \hat{f}^{(k)} = \underset{f}{\operatorname{argmin}}\{\sum_{i=1}^{n}(\tilde{y}_i - f(x_i))^2\} = H\tilde{y}$ — (F)

Note that H is the same as in (B), because $\underline{x}_k$ does not change.

⑦ Finally, $\hat{f}(\underline{x}_k) - \hat{f}^{(k)}_{k(k)} = [Hy]_k - [H\tilde{y}]_k = [H(y-\tilde{y})]_k$

$= y_k - \hat{y}_{k(i)}$

$$= \left[\begin{pmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & & \\ h_{n1} & \cdots & h_{nn} \end{pmatrix}\begin{pmatrix} 0 \\ \vdots \\ y_k - \hat{y}_{k(k)} \\ \vdots \\ 0 \end{pmatrix}\right]_k = h_{kk}\cdot(y_k - \hat{y}_{k(k)})$$

$y_k \to \hat{y}_{k(k)}$

$\therefore y_k - \hat{f}^{(k)}(\underline{x}_k) = \frac{y_k - \hat{f}(x_k)}{1-h_{kk}}$

7.6 Local Regression

Algorithm 7.1 *Local Regression At* $X = x_0$

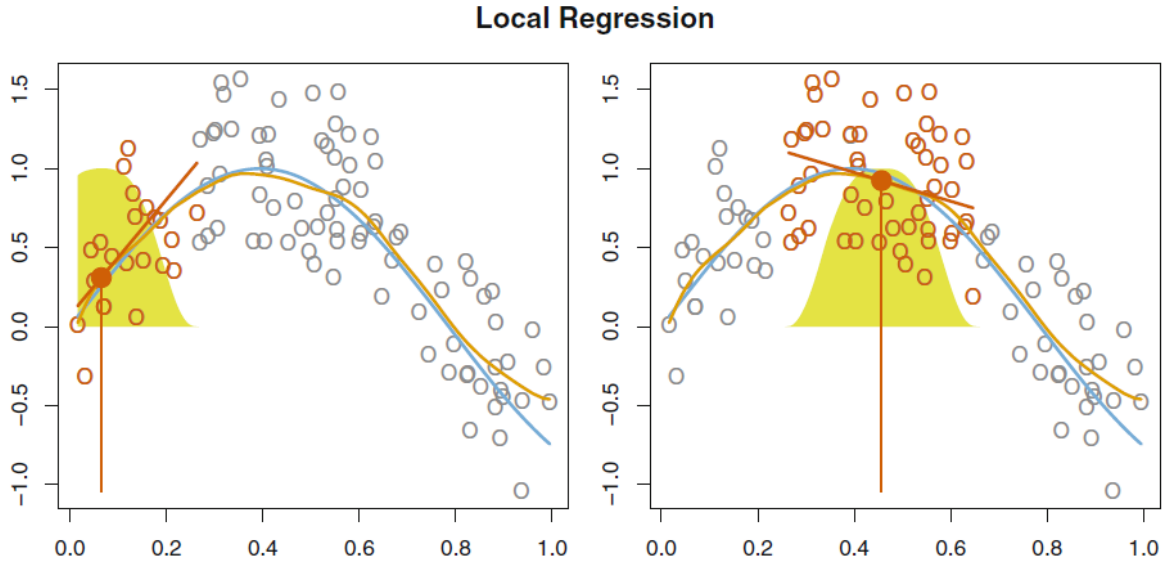
1. Gather the fraction $s = k/n$ of training points whose x_i are closest to x_0.

2. Assign a weight $K_{i0} = K(x_i, x_0)$ to each point in this neighborhood, so that the point furthest from x_0 has weight zero, and the closest has the highest weight. All but these k nearest neighbors get weight zero.

3. Fit a *weighted least squares regression* of the y_i on the x_i using the aforementioned weights, by finding $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2. \tag{7.14}$$

4. The fitted value at x_0 is given by $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.



- Near boundary
- Weighted least squares regression

7.7 Generalized Additive Models

Now, extension of multiple linear regression.

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

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

Example)

$$\texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \epsilon$$

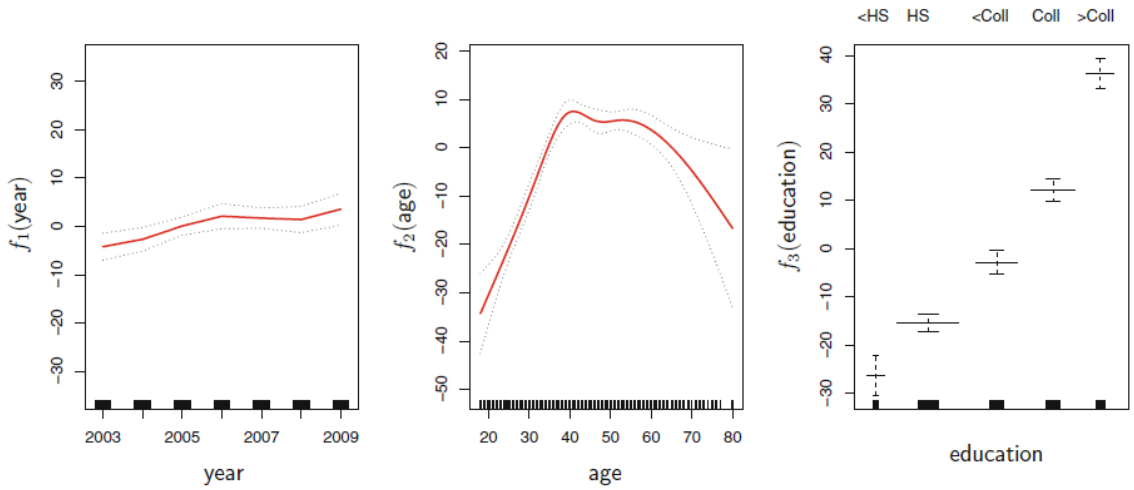


FIGURE 7.11. *For the* `Wage` *data, plots of the relationship between each feature and the response,* `wage`*, in the fitted model (7.16). Each plot displays the fitted function and pointwise standard errors. The first two functions are natural splines in* `year` *and* `age`*, with four and five degrees of freedom, respectively. The third function is a step function, fit to the qualitative variable* `education`*.*

Pros and Cons of GAMs

Before we move on, let us summarize the advantages and limitations of a GAM.

- ▲ 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).

For fully general models, we have to look for even more flexible approaches such as random forests and boosting, described in Chapter 8. GAMs provide a useful compromise between linear and fully nonparametric models.