STA 5820
Chapter 7
Moving beyond linearity

Kazuhiko Shinki

Wayne State University

7. Moving Beyond Linearity

The most basic model to relate variable $\boldsymbol{x}$ and $\boldsymbol{y}$ is a linear model: $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x}$.

If a more complicated relationship is to be investigated, there are a number of ways to deal with nonlinear relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$.

7. Moving Beyond Linearity

7.1 Polynomial Regression

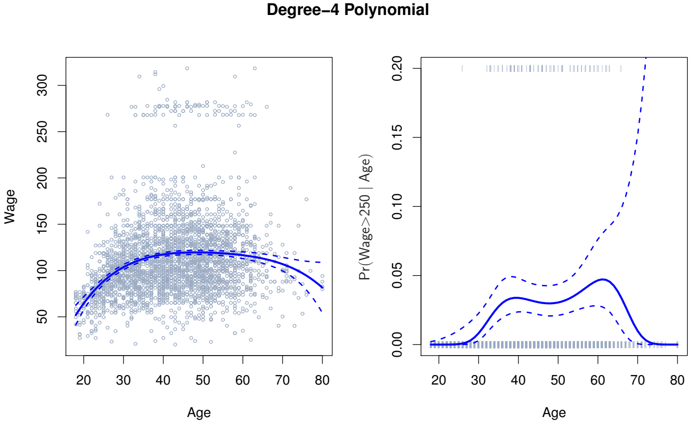
As we have discussed, a simple way to extend the linear relationship is the polynomial regression.

$$y = \beta_0 + \beta_1 x + \cdots + \beta_d x^d + \epsilon$$

Higher order terms can be also included in the logistic regression.

7. Moving Beyond Linearity

Figure 7-1: Polynomial regression (left). Logistic regression with polynomial function (right).



7.2 Step functions

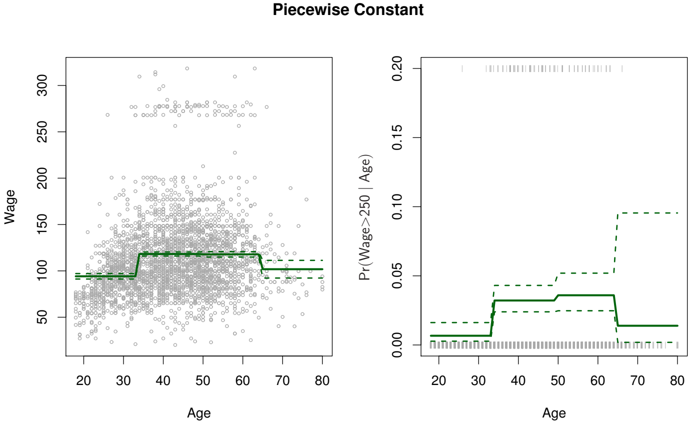
Alternatively, we can use step functions $C_k(x)$ instead of polynomials. I.e.,

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

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

However, it is difficult to determine knot points $c_1, \cdots, c_K$.

Figure 7-2: Step functions (left). Step functions for logistic regression (right).



7.3 Basis functions

More generally, we can use basis functions $b_1(x), \cdots, b_K(x)$ for regression. That is,

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

$b_1(x), \cdots, b_K(x)$ are any predetermined functions. This is more practical than the basis of the polynomial regression $x, \cdots, x^k$ in several aspects. For example, it can

- avoid multicollinearity by chosing an orthogonal basis $b_1(x), \cdots, b_K(x)$ given data, and
- include specific features such as periodicity and boundedness.

7.4 Regression splines

Step functions are polynomials with degree zero. As a general case of step functions, we can define piece-wise polynomial functions. For example,

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

Usually we want to make the entire function continuous. This imposes some constraint on coefficients.

The regression spline is a more sophisticated method to impose continuity as well as the same derivatives (of a specific order) at knot points so that the resulting fitted function is smooth enough.

7. Moving Beyond Linearity

A function which

- is a piece-wise cubic function,
- has K knots (where K is a fixed number), and
- is continuous, and has the same 1st and 2nd derivatives at knot points.

is called a cubic spline. This spline has an expression by basis functions:

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

where $b_1(x), \cdots, b_K(x)$ are cubic functions.

Why is it always possible for the function above to satisfy the three conditions?

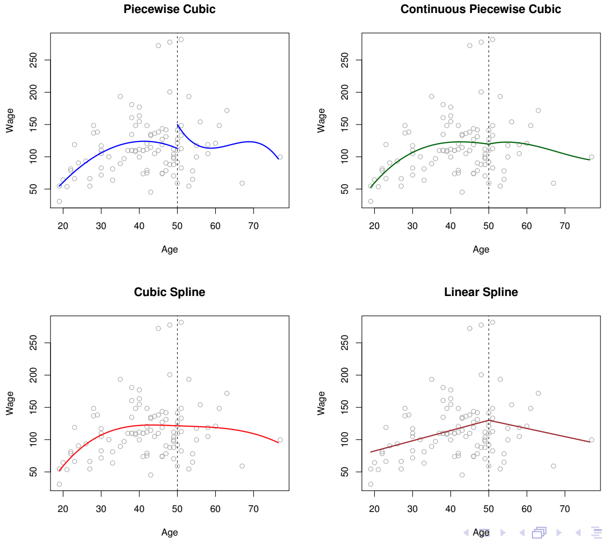
- The cubic function $y = g(x)$ for the most left section has three terms x, x^2, x^3 in addition to intercept.
- After passing each knot point ξ, the following basis function is needed:

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

In this way, the function satisfy all the conditions above.

7. Moving Beyond Linearity

Figure 7-3: Piecewise regressons and regression splines.



7.5 Smoothing splines

It is difficult to decide the number of knot points as well as the locations of knot points for regression splines. The smoothing spline solves this issue by 1) using knot points in every interval between two points, but 2) imposing penalty on smoothness.

The smoothing spline g is determined so that the following loss function is minimized:

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

where λ is a non-negative tuning parameter (i.e., a given constant). Note that the second derivative represents smoothness.

Effective degrees of freedom

Usually the degrees of freedom ***df*** (i.e., the number of parameters in the model) is a measure of model complexity. The measure of flexibility for smoothing splines is λ. What is the relationship between λ and ***df***?

Since the smoothing spline is fitting a regression model, each fitted value $\hat{y}$ is a linear combination of $y_1, \cdots, y_n$. Therefore, we can write the fitted value $\hat{g}_\lambda = (\hat{y}_1, \cdots, \hat{y}_n)'$ as

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

where S_λ is an n by n coefficient matrix, and $y = (y_1, \cdots, y_n)'$.

7. Moving Beyond Linearity

Define the effective degrees of freedom as

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

If we use too much information of $\boldsymbol{y_i}$ itself to determine $\boldsymbol{\hat{y}_i}$, then the fitted curve is ad hoc and the degrees of freedom is large.

On the other hand, if $\boldsymbol{\hat{y}_i = \bar{y}}$, we literally fit the model $\boldsymbol{y = \beta_0}$ so $\boldsymbol{df}$ should be 1. In fact, the effective degrees of freedom is 1, since all elements of $\boldsymbol{S_\lambda}$ is $\mathbf{1/n}$.

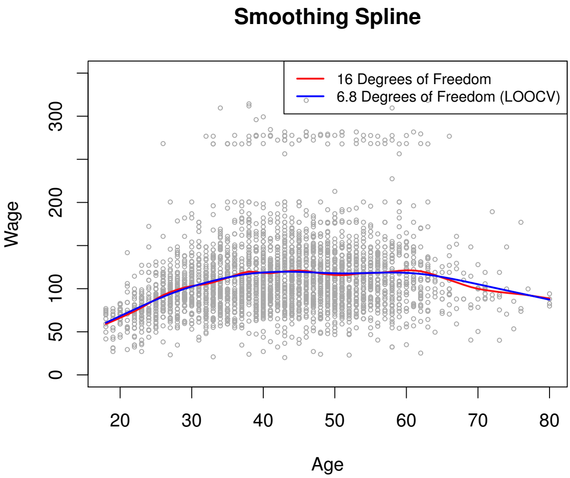
Optimal λ

The optimal λ (and the optimal effective ***df***) are determined by leave-one-out cross validation (LOOCV). Namely, choose λ which minimizes:

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

7. Moving Beyond Linearity

Figure 7-8: Smoothing splines.



7.6 Local regression

The local regression is a non-parametric version of regression.

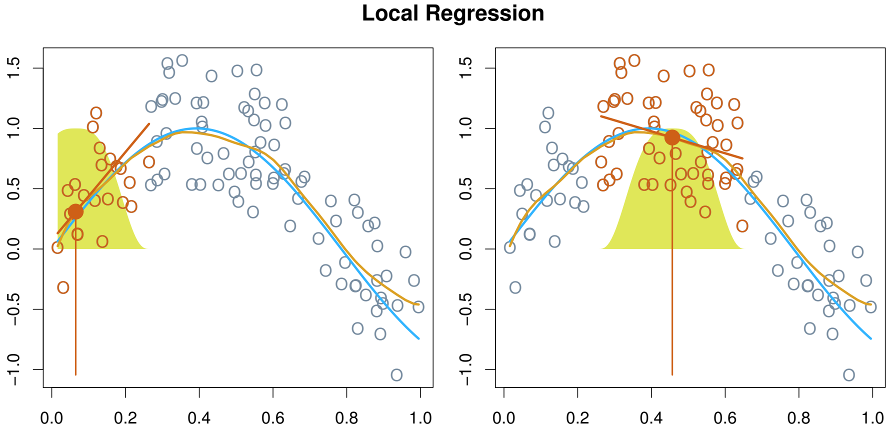
Suppose that we want to predict $\boldsymbol{y}$ for a given $\boldsymbol{x_0}$. The local regression fits a simple function (typically a straight line) only with a certain proportion ($\boldsymbol{s}$) of observations which are closest to $\boldsymbol{x_0}$. Also, observations closer to $\boldsymbol{x_0}$ are given higher weights.

A larger $\boldsymbol{s}$ makes $\hat{\boldsymbol{f}}(\boldsymbol{x})$ smoother, the variance of $\hat{\boldsymbol{f}}(\boldsymbol{x})$ smaller, but the bias of $\hat{\boldsymbol{f}}(\boldsymbol{x})$ larger.

The local regression needs fitting for each given $\boldsymbol{x}$, so we can not estimate the fitted curve at all possible $\boldsymbol{x}$ since there are infinitely many $\boldsymbol{x}$'s. Usually we only calculate $\hat{\boldsymbol{y}}$ for observed $\boldsymbol{x}$.

7. Moving Beyond Linearity

Figure 7-9: Image on how local regression is fitted.



More detailed algorithm of local regressions

To find the estimate $\hat{f}(x_0)$ by local regression:

1. Find $100s$ % of observations which are closest to x_0. Let $x_1, \cdots, x_k$ be such observations.
2. Assign a weight $K(x_i, x_0)$. for each x_i. $K(x_i, x_0)$ is larger when x_i is closer to x_0. $K(x_i, x_0) = 0$ if x_i is not among the k closest point to x_0.
3. Estimate the coefficients of a regression line $y = \beta_0 + \beta_1 x$ by minimizing

$$\sum_{i=1}^{n} K(x_i, x_0) \cdot (y_i - \beta_0 - \beta_1 x_i)^2.$$

4. $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

7.7 Generalized additive models (GAMs)

When there are multiple predictors $x_1, \cdots, x_p$, considering a general regression function $y = f(x_1, \cdots, x_p) + \epsilon$ is too demanding.

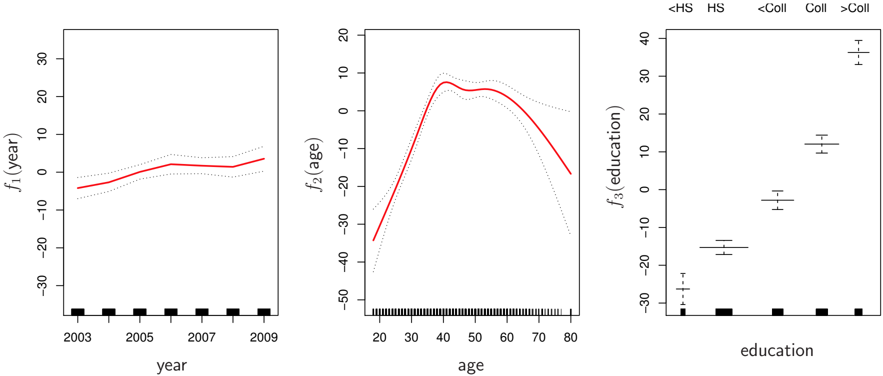
The GAM is a regression model ignoring interaction between predictors, i.e.,

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

Each function f_j can be estimated by smoothing spline for example.

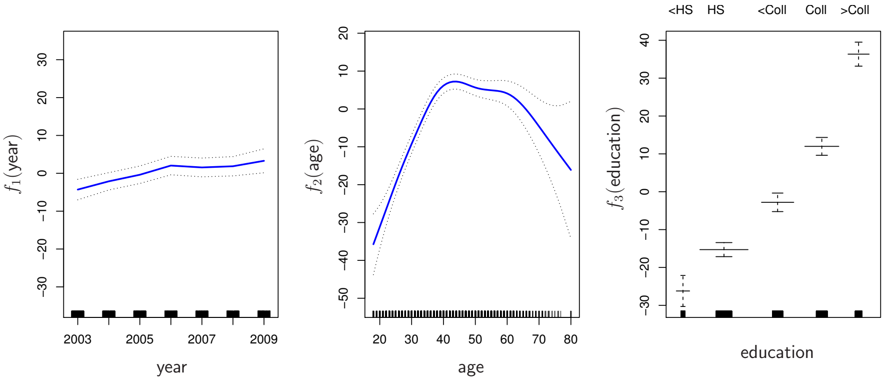
7. Moving Beyond Linearity

The following figures illustrate the regression results when ***wage*** is regressed on ***year*** and ***age*** and ***education***. $f_1(\textit{year})$, $f_2(\textit{age})$ and $f_3(\textit{education})$ are estimated by the least-square method with regression splines. (***Education*** is categorical, so an ANOVA is applied).



7. Moving Beyond Linearity

Same as the previous slide, but $f_1(year)$, $f_2(age)$ and $f_3(education)$ are estimated by smoothing splines.



GAMs for classification

Similarly to polynomial regression, GAMs works for logistic regression as well. Suppose that Y is a binary response variable (i.e., $Y = 0$ or 1), and let $p(X) := P(Y = 0)$. Then,

$$\log\left(\frac{1 - p(X)}{p(X)}\right) = \beta_0 + f_1(X_1) + \cdots + f_p(X_p).$$

$f_1(X_1), \cdots, f_p(X_p)$ can be estimated by regression or smoothing spline, for example.

Memo