Lecture 13: Non-linear Regression

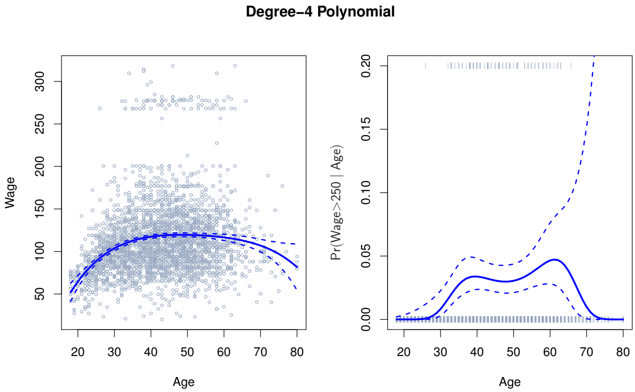
Reading: Section 5.2, 5.3, 5.4

GU4241/GR5241 Statistical Machine Learning

Linxi Liu
March 9, 2018

Non-linear regression

Problem: How do we model a non-linear relationship?



Left: Regression of wage onto age.
Right: Logistic regression for classes wage> 250 and wage≤ 250

Basis functions

Strategy:

- Define a model:

$$Y = \beta_0 + \beta_1 f_1(X) + \beta_2 f_2(X) + \cdots + \beta_d f_d(X).$$

Basis functions

Strategy:

- Define a model:

$$Y = \beta_0 + \beta_1 f_1(X) + \beta_2 f_2(X) + \cdots + \beta_d f_d(X).$$

- Fit this model through least-squares regression.

Basis functions

Strategy:

- Define a model:

$$Y = \beta_0 + \beta_1 f_1(X) + \beta_2 f_2(X) + \cdots + \beta_d f_d(X).$$

- Fit this model through least-squares regression.
- Options for $f_1, \ldots, f_d$:

Basis functions

Strategy:

- Define a model:

$$Y = \beta_0 + \beta_1 f_1(X) + \beta_2 f_2(X) + \cdots + \beta_d f_d(X).$$

- Fit this model through least-squares regression.
- Options for $f_1, \ldots, f_d$:
    1. Polynomials, $f_i(x) = x^i$.

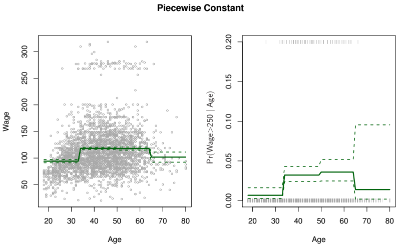
Basis functions

Strategy:

- Define a model:

$$Y = \beta_0 + \beta_1 f_1(X) + \beta_2 f_2(X) + \cdots + \beta_d f_d(X).$$

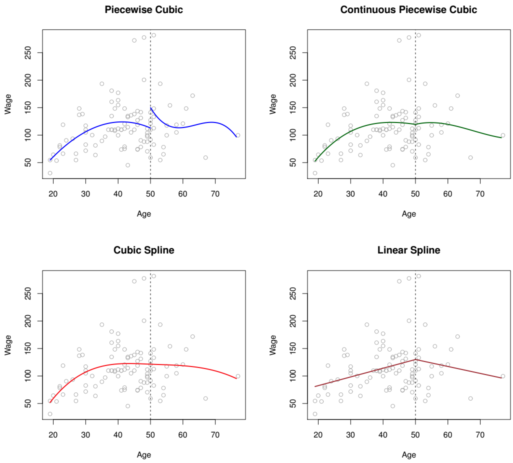
- Fit this model through least-squares regression.
- Options for $f_1, \ldots, f_d$:
  1. Polynomials, $f_i(x) = x^i$.
  2. Indicator functions, $f_i(x) = \mathbf{1}(c_i \leq x < c_{i+1})$.



Basis functions

- Options for $f_1, \dots, f_d$:

  3. Piecewise polynomials:



Cubic splines

- Define a set of knots $\xi_1 < \xi_2 < \cdots < \xi_K$.

Cubic splines

- Define a set of knots $\xi_1 < \xi_2 < \cdots < \xi_K$.
- We want the function $Y = f(X)$ to:

Cubic splines

- Define a set of knots $\xi_1 < \xi_2 < \cdots < \xi_K$.
- We want the function $Y = f(X)$ to:

  1. Be a cubic polynomial between every pair of knots ξ_i, ξ_{i+1}.

Cubic splines

- Define a set of knots $\xi_1 < \xi_2 < \cdots < \xi_K$.
- We want the function $Y = f(X)$ to:

  1. Be a cubic polynomial between every pair of knots ξ_i, ξ_{i+1}.
  2. Be continuous at each knot.

Cubic splines

- Define a set of knots $\xi_1 < \xi_2 < \cdots < \xi_K$.
- We want the function $Y = f(X)$ to:

  1. Be a cubic polynomial between every pair of knots ξ_i, ξ_{i+1}.
  2. Be continuous at each knot.
  3. Have continuous first and second derivatives at each knot.

Cubic splines

- Define a set of knots $\xi_1 < \xi_2 < \cdots < \xi_K$.
- We want the function $Y = f(X)$ to:

  1. Be a cubic polynomial between every pair of knots ξ_i, ξ_{i+1}.
  2. Be continuous at each knot.
  3. Have continuous first and second derivatives at each knot.

- It turns out, we can write f in terms of $K + 4$ basis functions:

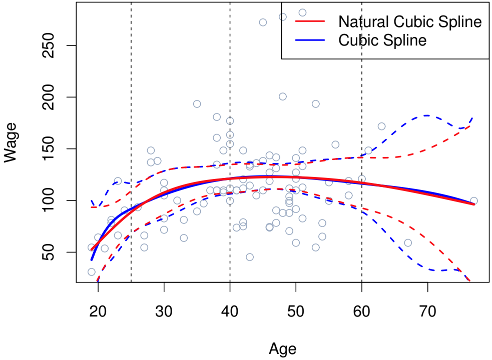
$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 h(X, \xi_1) + \cdots + \beta_{K+3} h(X, \xi_K)$$

where,

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

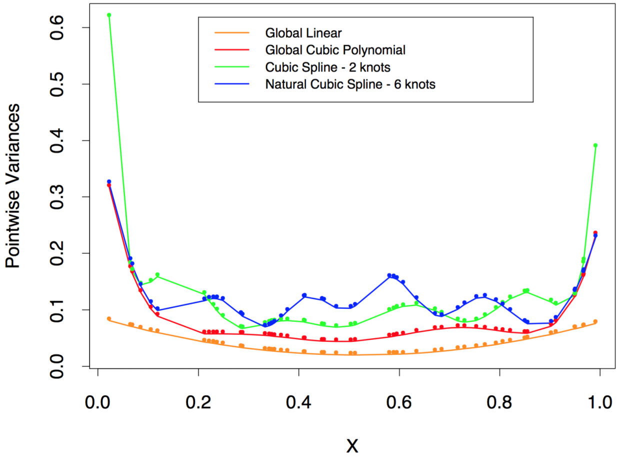
Natural cubic splines

Spline which is linear instead of cubic for $X < \xi_1$, $X > \xi_K$.



The predictions are more stable for extreme values of X.

Natural cubic splines vs. cubic splines



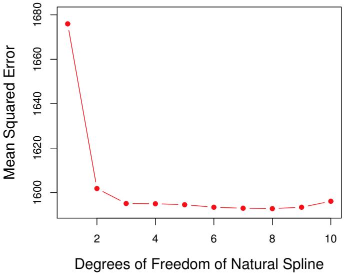
Choosing the number and locations of knots

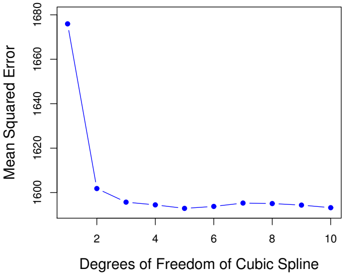
The locations of the knots are typically quantiles of X.

Choosing the number and locations of knots

The locations of the knots are typically quantiles of X.

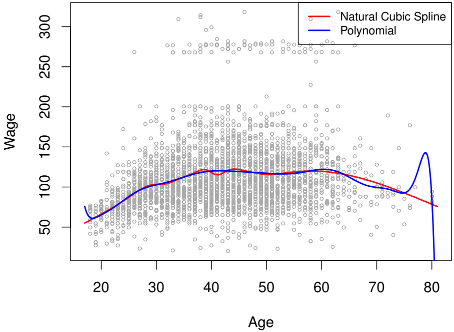
The number of knots, K, is chosen by cross validation:





Natural cubic splines vs. polynomial regression

- Splines can fit complex functions with few parameters.
- Polynomials require high degree terms to be flexible.
- High-degree polynomials can be unstable at the edges.



Smoothing splines

Find the function f which minimizes

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx$$

- The RSS of the model.
- A penalty for the roughness of the function.

Smoothing splines

Find the function f which minimizes

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx$$

- The RSS of the model.
- A penalty for the roughness of the function.

Facts:

- The minimizer $\hat{f}$ is a natural cubic spline, with knots at each unique sample point $x_1, \ldots, x_n$.
- Obtaining $\hat{f}$ is similar to a Ridge regression.

Natural cubic splines vs. Smoothing splines

Natural cubic splines

- Fix the locations of K knots at quantiles of X.

Smoothing splines

Natural cubic splines vs. Smoothing splines

Natural cubic splines

- Fix the locations of K knots at quantiles of X.
- Number of knots $K < n$.

Smoothing splines

Natural cubic splines vs. Smoothing splines

Natural cubic splines

- Fix the locations of K knots at quantiles of X.
- Number of knots $K < n$.
- Find the natural cubic spline $\hat{f}$ which minimizes the RSS:

$$\sum_{i=1}^{n}(y_i - f(x_i))^2$$

Smoothing splines

Natural cubic splines vs. Smoothing splines

Natural cubic splines

- Fix the locations of K knots at quantiles of X.
- Number of knots $K < n$.
- Find the natural cubic spline $\hat{f}$ which minimizes the RSS:

$$\sum_{i=1}^{n}(y_i - f(x_i))^2$$

- Choose K by cross validation.

Smoothing splines

Natural cubic splines vs. Smoothing splines

Natural cubic splines

- Fix the locations of K knots at quantiles of X.
- Number of knots $K < n$.
- Find the natural cubic spline $\hat{f}$ which minimizes the RSS:

$$\sum_{i=1}^{n}(y_i - f(x_i))^2$$

- Choose K by cross validation.

Smoothing splines

- Put n knots at $x_1, \ldots, x_n$.

Natural cubic splines vs. Smoothing splines

Natural cubic splines

- Fix the locations of K knots at quantiles of X.
- Number of knots $K < n$.
- Find the natural cubic spline $\hat{f}$ which minimizes the RSS:

$$\sum_{i=1}^{n}(y_i - f(x_i))^2$$

- Choose K by cross validation.

Smoothing splines

- Put n knots at $x_1, \ldots, x_n$.
- We could find a cubic spline which makes the RSS $= 0$ $\longrightarrow$ Overfitting!

Natural cubic splines vs. Smoothing splines

Natural cubic splines

- Fix the locations of K knots at quantiles of X.
- Number of knots $K < n$.
- Find the natural cubic spline $\hat{f}$ which minimizes the RSS:

$$\sum_{i=1}^{n}(y_i - f(x_i))^2$$

- Choose K by cross validation.

Smoothing splines

- Put n knots at $x_1, \ldots, x_n$.
- We could find a cubic spline which makes the RSS $= 0$ $\longrightarrow$ Overfitting!
- Instead, we obtain the fitted values $\hat{f}(x_1), \ldots, \hat{f}(x_n)$ through an algorithm similar to Ridge regression.

Natural cubic splines vs. Smoothing splines

Natural cubic splines

- Fix the locations of K knots at quantiles of X.
- Number of knots $K < n$.
- Find the natural cubic spline $\hat{f}$ which minimizes the RSS:

$$\sum_{i=1}^{n}(y_i - f(x_i))^2$$

- Choose K by cross validation.

Smoothing splines

- Put n knots at $x_1, \ldots, x_n$.
- We could find a cubic spline which makes the RSS $= 0$ $\longrightarrow$ Overfitting!
- Instead, we obtain the fitted values $\hat{f}(x_1), \ldots, \hat{f}(x_n)$ through an algorithm similar to Ridge regression.
- The function $\hat{f}$ is the only natural cubic spline that has these fitted values.

Deriving a smoothing spline

1. Show that if you fix the values $f(x_1), \ldots, f(x_n)$, the roughness

$$\int f''(x)^2 dx$$

is minimized by a natural cubic spline.

Deriving a smoothing spline

1. Show that if you fix the values $f(x_1), \ldots, f(x_n)$, the roughness

$$\int f''(x)^2 dx$$

is minimized by a natural cubic spline. Problem 5.7 in ESL.

Deriving a smoothing spline

1. Show that if you fix the values $f(x_1), \dots, f(x_n)$, the roughness

$$\int f''(x)^2 dx$$

   is minimized by a natural cubic spline. Problem 5.7 in ESL.

2. Deduce that the solution to the smoothing spline problem is a natural cubic spline, which can be written in terms of its basis functions.

$$f(x) = \beta_1 N_1(x) + \dots \beta_n N_n(x)$$

Deriving a smoothing spline

3. Letting $\mathbf{N}$ be a matrix with $\mathbf{N}(i,j) = N_j(x_i)$, we can write the objective function:

$$(y - \mathbf{N}\beta)^T(y - \mathbf{N}\beta) + \lambda\beta^T\Omega_{\mathbf{N}}\beta,$$

where $\Omega_{\mathbf{N}}(j,k) = \int N_j''(t)N_k''(t)dt$.

Deriving a smoothing spline

3. Letting $\mathbf{N}$ be a matrix with $\mathbf{N}(i,j) = N_j(x_i)$, we can write the objective function:

$$(y - \mathbf{N}\beta)^T(y - \mathbf{N}\beta) + \lambda\beta^T\Omega_{\mathbf{N}}\beta,$$

   where $\Omega_{\mathbf{N}}(j,k) = \int N_j''(t)N_k''(t)dt$.

4. By simple calculus, the coefficients $\hat{\beta}$ which minimize

$$(y - \mathbf{N}\beta)^T(y - \mathbf{N}\beta) + \lambda\beta^T\Omega_{\mathbf{N}}\beta,$$

   are $\hat{\beta} = (\mathbf{N}^T\mathbf{N} + \lambda\Omega_{\mathbf{N}})^{-1}\mathbf{N}^Ty$.

Deriving a smoothing spline

5. Note that the predicted values are a linear function of the observed values:

$$\hat{y} = \underbrace{\mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\Omega_{\mathbf{N}})^{-1}\mathbf{N}^T}_{\mathbf{S}_\lambda} y$$

Deriving a smoothing spline

5. Note that the predicted values are a linear function of the observed values:

$$\hat{y} = \underbrace{\mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\Omega_{\mathbf{N}})^{-1}\mathbf{N}^T}_{\mathbf{S}_\lambda} y$$

6. The **degrees of freedom** for a smoothing spline are:

$$\text{Trace}(\mathbf{S}_\lambda) = \mathbf{S}_\lambda(1,1) + \mathbf{S}_\lambda(2,2) + \cdots + \mathbf{S}_\lambda(n,n)$$

Choosing the regularization parameter λ

- We typically choose λ through cross validation.

Choosing the regularization parameter λ

- We typically choose λ through cross validation.
- Fortunately, we can solve the problem for any λ with the same complexity of diagonalizing an $n \times n$ matrix.

Choosing the regularization parameter λ

- We typically choose λ through cross validation.
- Fortunately, we can solve the problem for any λ with the same complexity of diagonalizing an $n \times n$ matrix.
- There is a shortcut for LOOCV:

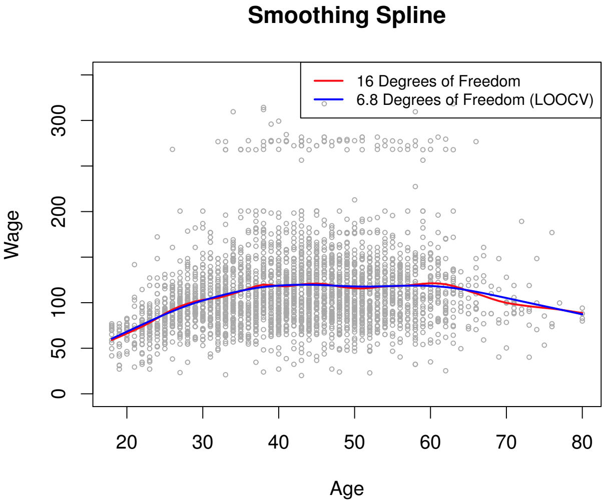
$$RSS_{\mathsf{loocv}}(\lambda) = \sum_{i=1}^{n} (y_i - \hat{f}_\lambda^{(-i)}(x_i))^2$$

Choosing the regularization parameter λ

- We typically choose λ through cross validation.
- Fortunately, we can solve the problem for any λ with the same complexity of diagonalizing an $n \times n$ matrix.
- There is a shortcut for LOOCV:

$$
\begin{aligned}
RSS_{\mathsf{loocv}}(\lambda) &= \sum_{i=1}^{n} (y_i - \hat{f}_\lambda^{(-i)}(x_i))^2 \\
&= \sum_{i=1}^{n} \left[\frac{y_i - \hat{f}_\lambda(x_i)}{1 - \mathbf{S}_\lambda(i,i)} \right]^2
\end{aligned}
$$

Choosing the regularization parameter λ



Natural cubic splines vs. Smoothing splines

Natural cubic splines

- Fix the locations of K knots at quantiles of X.
- Number of knots $K < n$.
- Find the natural cubic spline $\hat{f}$ which minimizes the RSS:

$$\sum_{i=1}^{n} (y_i - f(x_i))^2$$

- Choose K by cross validation.

Smoothing splines

- Put n knots at $x_1, \ldots, x_n$.
- We could find a cubic spline which makes the RSS $= 0$ $\longrightarrow$ Overfitting!
- Instead, we obtain the fitted values $\hat{f}(x_1), \ldots, \hat{f}(x_n)$ through an algorithm similar to Ridge regression.
- The function $\hat{f}$ is the only natural cubic spline that has these fitted values.