Designing Nonlinear Features

Linear regression and beyond

Marek Petrik

4/11/2017

Linear Regression

- Can linear regression fit non-linear functions?

Linear Regression

- Can linear regression fit non-linear functions?
- Can logistic regression be used to compute non-linear decision boundaries?

Linear Regression

- Can linear regression fit non-linear functions?
- Can logistic regression be used to compute non-linear decision boundaries?
- What feature transformations do you know?

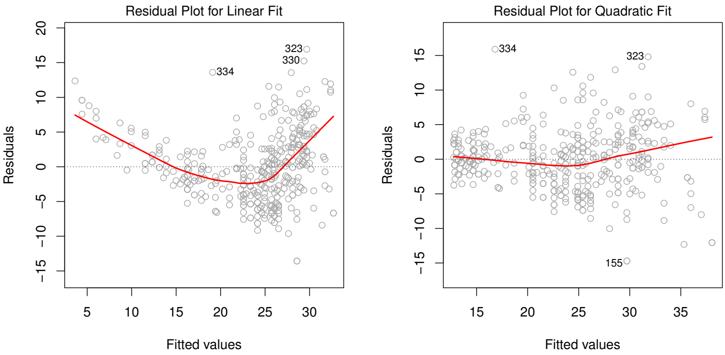
Linear Regression

- Can linear regression fit non-linear functions?
- Can logistic regression be used to compute non-linear decision boundaries?
- What feature transformations do you know?
- How is it related to kernels?

When to Fit Nonlinear Model?

When to Fit Nonlinear Model?

- Residual plot



Approaches to Nonlinear Feature Relationship

- We will cover:
  1. Polynomial regression
  2. Step functions
  3. Regression splines
  4. Smoothing splines
  5. Local regression
  6. Generalized additive models

Approaches to Nonlinear Feature Relationship

Today: Problems with a single variable

- We will cover:
  1. Polynomial regression
  2. Step functions
  3. Regression splines
  4. Smoothing splines
  5. Local regression
  6. Generalized additive models
- Others significant ones:
  1. Fourier Analysis
  2. Wavelets

Polynomial Regression

- Standard linear model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Polynomial function:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i + \beta_3 x_i + \ldots \beta_d x_i + \epsilon_i$$

Example Polynomial Regression

- Linear regression:

$$\text{mpg} = \beta_0 + \beta_1 \times \text{power}$$

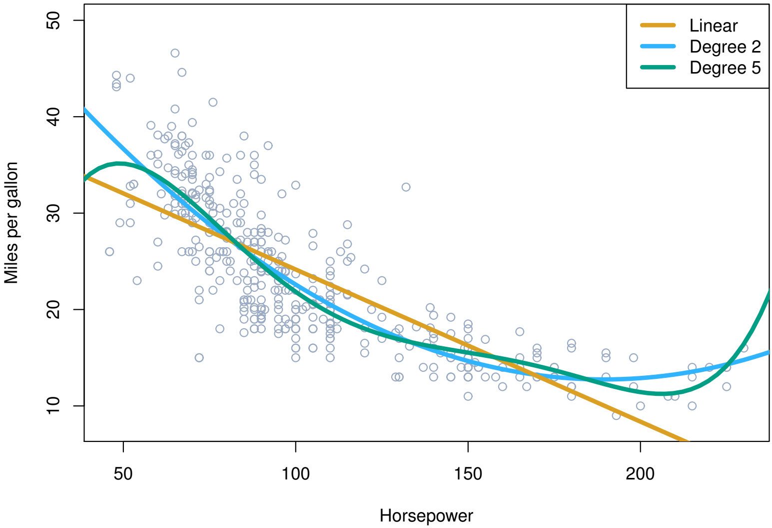
- Degree 2 (Quadratic):

$$\text{mpg} = \beta_0 + \beta_1 \times \text{power} + \beta_2 \times \text{power}^2$$

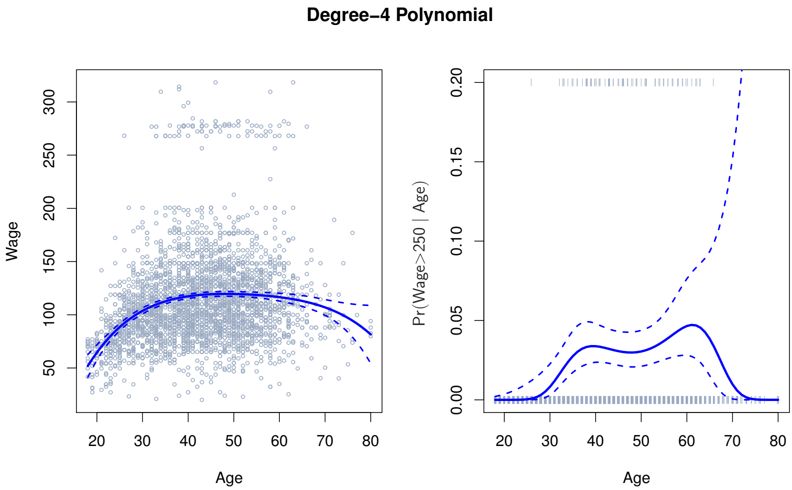
- Degree k:

$$\text{mpg} = \sum_{i=0}^{k} \beta_k \times \text{power}^k$$

Polynomial Functions



Polynomial Functions (Linear and Logistic)



Why Polynomial Regression is Insufficient?

- Does not account for **local** non-linearity
- Limited a-priori knowledge
- Very unstable in extreme ranges
- Different problems require different structure

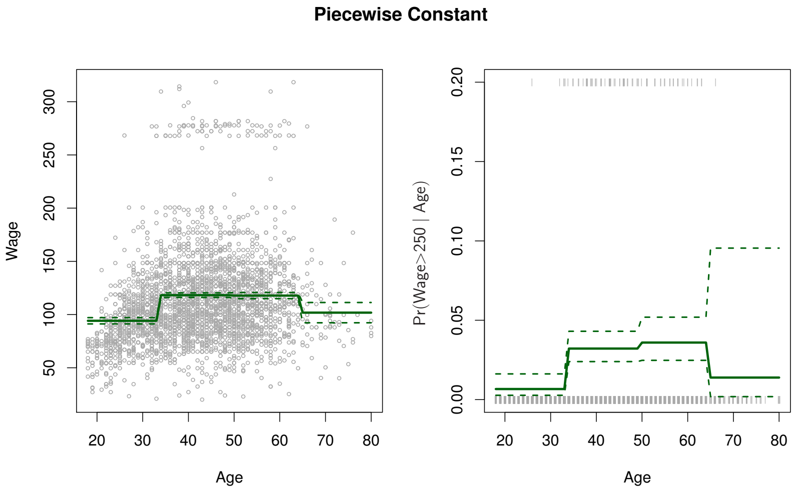
Step Functions

- Similar to dummy variables, but for quantitative features
- Create cut points $c_1.c_2.\ldots, c_K$
- Construct $K+1$ new features:

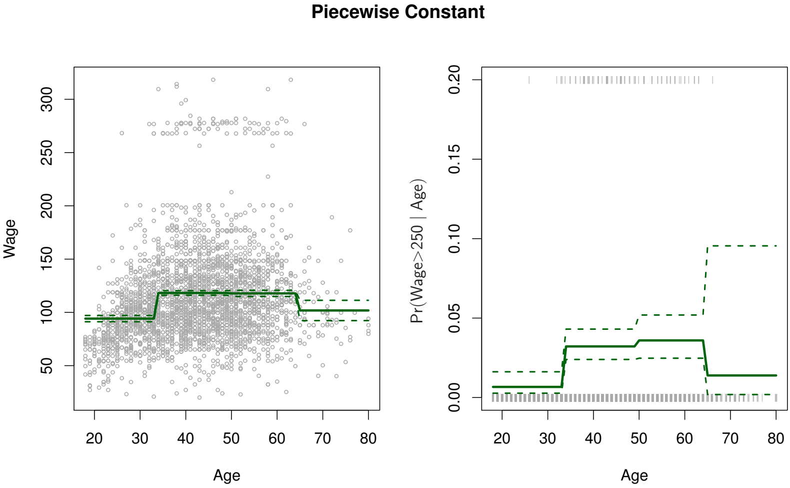
$$
\begin{aligned}
C_0(X) &= I(X < c_1)\\
C_1(X) &= I(c_1 \le X < c_2)\\
&\vdots
\end{aligned}
$$

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

Step Functions Example



Step Functions Example



Step functions are not continuous!

Basis Functions

- Polynomial functions are new **basis functions**
- Step functions are new **basis functions**
- **Basis Functions**: Span linear space
- Linear algebra detour

Basis of Vector Space

- Vectors $X_1, X_2, \ldots, X_K$
- **Span** of vectors (space):

$$\alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_K X_K$$

- **Basis**: smallest set of vectors that spans a space

Column View of Linear Regression

- Linear regression:

$$\min_{\beta} \|y - X\beta\|_2^2$$

- Treat vectors as columns:

$$\min_{\beta} \|y - X_1\beta_1 - \ldots - X_K\beta_K\|_2^2$$

- **Interpretation**: closest point to y in space spanned by $X_1, \ldots, X_K$

Column View of Linear Regression

- Linear regression:
$$\min_{\beta} \|y - X\beta\|_2^2$$
- Treat vectors as columns:
$$\min_{\beta} \|y - X_1\beta_1 - \ldots - X_K\beta_K\|_2^2$$
- **Interpretation**: closest point to y in space spanned by $X_1, \ldots, X_K$
- **Features are the basis!**

Regression Splines

- **Polynomials** are not local
- **Step functions** are not continuous or smooth

Regression Splines

- **Polynomials** are not local
- **Step functions** are not continuous or smooth

- Regression splines are local and smooth

Regression Splines

- **Polynomials** are not local
- **Step functions** are not continuous or smooth

- Regression splines are local and smooth
- Derivation in several steps

Step 1: Step Function as Piecewise Polynomials

- Step functions (minor change in $\leq$):

$$
\begin{aligned}
C_0(X) &= I(X < c_1) \\
C_1(X) &= I(c_1 < X \leq c_2) \\
&\vdots
\end{aligned}
$$

- Step-functions are piece-wise polynomials of degree 0

$$
C_i(X) = \begin{cases} 1 & \text{if } X > c_i \text{ and } X \leq c_{i+1} \\ 0 & \text{otherwise} \end{cases}
$$

- Different representation (**basis spans the same space!**):

$$
C_i(X) = \begin{cases} 1 & \text{if } X > c_{i-1} \\ 0 & \text{otherwise} \end{cases}
$$

Step 2: Piecewise Polynomials

- Piecewise polynomials of degree 1:

$$P_i(X) = \begin{cases} X & \text{if } X > c_i \text{ and } X \leq c_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

- Piecewise polynomials of degree 2:

$$P_i(X) = \begin{cases} X^2 & \text{if } X > c_i \text{ and } X \leq c_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

- Piecewise polynomials of degree 3:

$$P_i(X) = \begin{cases} X^3 & \text{if } X > c_i \text{ and } X \leq c_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

Step 2: Piecewise Polynomials

- Piecewise polynomials of degree 1:

$$P_i(X) = \begin{cases} X & \text{if } X > c_i \text{ and } X \leq c_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

- Piecewise polynomials of degree 2:

$$P_i(X) = \begin{cases} X^2 & \text{if } X > c_i \text{ and } X \leq c_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

- Piecewise polynomials of degree 3:

$$P_i(X) = \begin{cases} X^3 & \text{if } X > c_i \text{ and } X \leq c_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

- Local but **not continuous!**

Step 3: Continuity and Regression Splines

- Piecewise polynomials of degree 1:

$$P_i(X) = \begin{cases} X & \text{if } X > c_i \text{ and } X \leq c_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

Step 3: Continuity and Regression Splines

- Piecewise polynomials of degree 1:

$$P_i(X) = \begin{cases} X & \text{if } X > c_i \text{ and } X \leq c_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

- **Must prevent discontinuity in knots**

Step 3: Continuity and Regression Splines

- Piecewise polynomials of degree 1:

$$P_i(X) = \begin{cases} X & \text{if } X > c_i \text{ and } X \leq c_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

- **Must prevent discontinuity in knots**
- Different representation:

$$H_i(X) = \begin{cases} X - c_{i-1} & \text{if } X > c_{i-1} \\ 0 & \text{otherwise} \end{cases}$$

Step 3: Continuity and Regression Splines

- Piecewise polynomials of degree 1:

$$P_i(X) = \begin{cases} X & \text{if } X > c_i \text{ and } X \leq c_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

- **Must prevent discontinuity in knots**
- Different representation:

$$H_i(X) = \begin{cases} X - c_{i-1} & \text{if } X > c_{i-1} \\ 0 & \text{otherwise} \end{cases}$$

- Each feature is 0 in its knot

General Regression Splines

- Regression splines of degree d:

$$H_i(X) = \begin{cases} (X - c_{i-1})^d & \text{if } X > c_{i-1} \\ 0 & \text{otherwise} \end{cases}$$

General Regression Splines

- Regression splines of degree d:

$$H_i(X) = \begin{cases} (X - c_{i-1})^d & \text{if } X > c_{i-1} \\ 0 & \text{otherwise} \end{cases}$$

- Compact representation:

$$h(x, \xi) = ([x - \xi]_+)^d = (\max\{x - \xi, 0\})^d$$

General Regression Splines

- Regression splines of degree d:

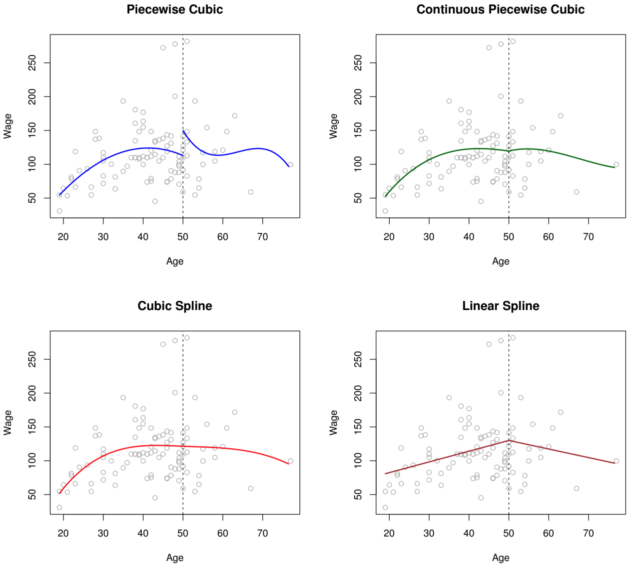
$$H_i(X) = \begin{cases} (X - c_{i-1})^d & \text{if } X > c_{i-1} \\ 0 & \text{otherwise} \end{cases}$$

- Compact representation:

$$h(x, \xi) = ([x - \xi]_+)^d = (\max\{x - \xi, 0\})^d$$

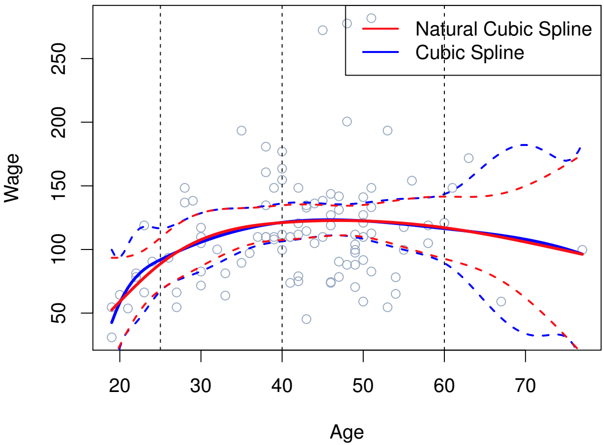
- Most common are **cubic splines**: continuous and continuously differentiable

Example Splines

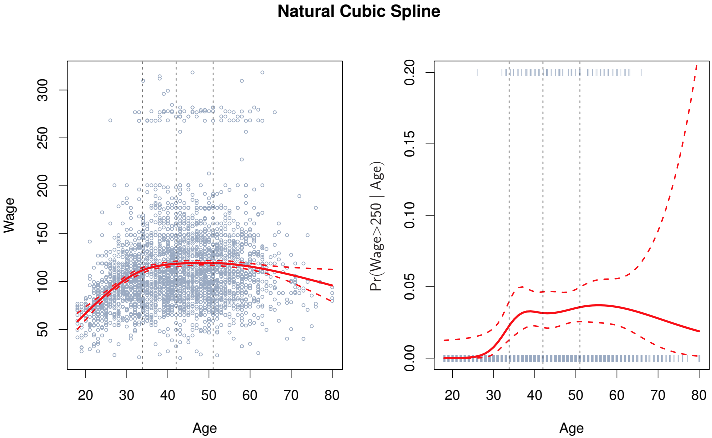


Natural Splines

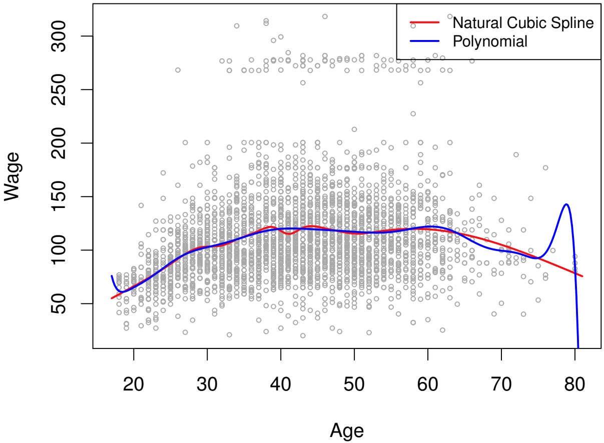
Boundary segments are linear



Natural Splines and Logistic Regression



Natural Splines vs Polynomials



Choosing Knots

- Domain dependent
- Change of mode (retirement?)
- Quantiles of data is generally a good choice
- Number of knots = degrees of freedom

Smoothing Splines

- Extreme version of regression splines
- Knot in every data point

Smoothing Splines

- Extreme version of regression splines
- Knot in every data point
- Must have regularization to generalize

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

- Smoothing parameter λ chosen by LOOCV
- Effective degrees of freedom: technical, but not very important

Finishing the Book

Read also 7.6 and 7.7:

- Local regression
- General Additive Models