Lecture 13 Nonparametric Regression I

STAT 441/505: Applied Statistical Methods in Data Mining

Linglong Kong

Department of Mathematical and Statistical Sciences
University of Alberta

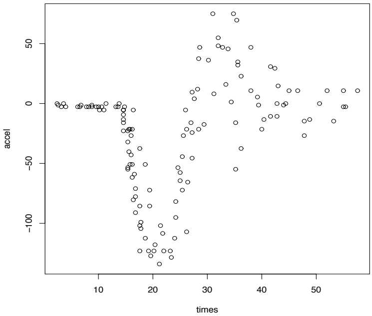
Winter, 2016

Outline

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Motorcycle Data

- The truth is never linear! Or almost never!
- We observe $y_i = f(\mathbf{x}_i) + \varepsilon_i$ but no knowledge of $f(\cdot)$; determine $\hat{f}(\mathbf{x})$ from the data alone - no model.
- Output from these methods is typically graphical and used for prediction and interpolation.



- The motorcycle data gives measurements on head acceleration vs. milliseconds after impact in a simulated motorcycle accident; it is used to test crash helmets.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Polynomial Regression

- One might try to fit a linear combination of certain basis functions, say orthogonal polynomials (see `help(poly)`)

$$y_i = \beta_0 + x_i\beta_1 + \cdots + x_i^p\beta_p + \epsilon_i,$$

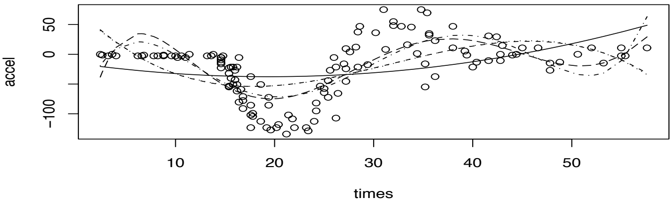
essentially, a multiple linear regression.

- Not really interested in the coefficients; more interested in the fitted function values at any values x_0

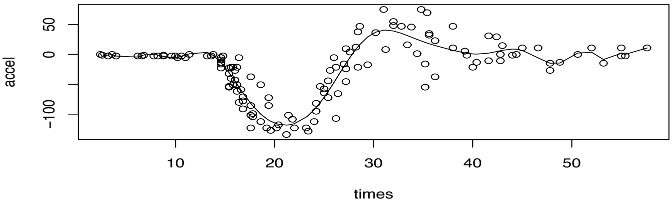
$$\hat{f}(x_0) = \beta_0 + x_0\beta_1 + \cdots + x_0^p\beta_p.$$

- Since $\hat{f}(x_0)$ is a linear function of $\hat{\beta}_j$, its variance can be easily obtained for pointwise-variances $\text{Var}[\hat{f}(x_0)]$. Therefore the confidence interval of the prediction can be calculated.
- Polynomials can be very unstable to fit, and behave erratically away from the region where there are data.

Motorcycle Data



Polynomial fits of degrees 2,...,6 to motorcycle data



Polynomial fit of degree 20 (!) to motorcycle data

Linear Splines

- A linear spline with knots at ε_k, $k = 1, \cdots, K$ is a piecewise linear polynomial continuous at each knot.
- We can represent this model as

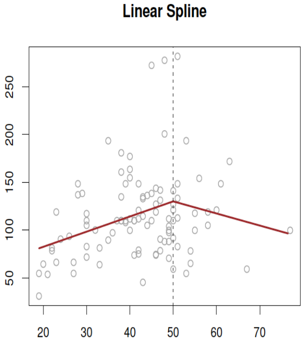
$$y_i = \beta_0 + \beta_1 b_1(x_i) + \cdots + \beta_{K+1} b_{K+1}(x_i) + \epsilon_i,$$

where the b_k are basis functions

$$b_1(x_i) = x_i,\ b_{k+1}(x_i) = (x_i - \varepsilon_i)_+,\ k = 1, \cdots, K,$$

where the $()_+$ means positive part, i.e.

$$(x_i - \varepsilon_i)_+ = \max(0, x_i - \varepsilon_i).$$



UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Cubic Splines

- A cubic spline with knots at ε_k, $k = 1, \cdots, K$ is a piecewise cubic polynomial with continuous derivatives up to order 2 at each knot.
- We can represent this model as

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \cdots + \beta_{K+1} b_{K+3}(x_i) + \epsilon_i,$$

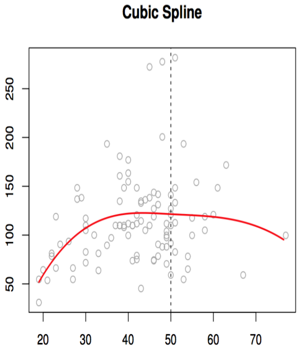
where the b_k are basis functions

$$b_1(x_i) = x_i, \; b_2(x_i) = x_i^2, \; b_3(x_i) = x_i^3,$$

$$b_{k+3}(x_i) = (x_i - \varepsilon_i)_+^3, \; k = 1, \cdots, K,$$

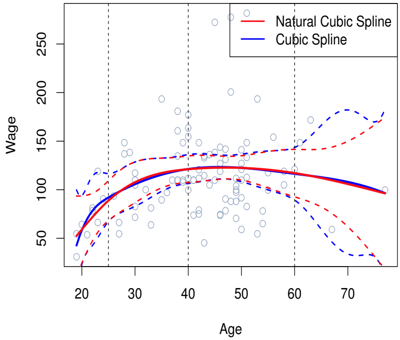
where the $()_+^3$ means positive part, i.e.

$$(x_i - \varepsilon_i)_+^3 = \max(0, (x_i - \varepsilon_i)^3).$$



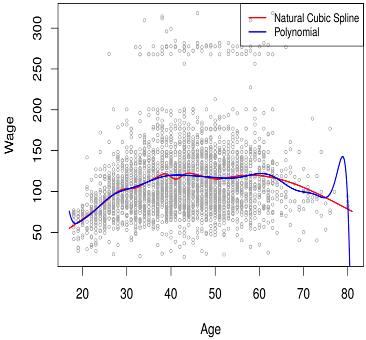
Natural Cubic Splines

- A natural cubic spline extrapolates linearly beyond the boundary knots.
- This adds $4 = 2 \times 2$ extra constraints, and allows us to put more internal knots for the same degrees of freedom as a regular cubic spline.



Knot Placement

- One strategy is to decide K, the number of knots, and then place them at appropriate quantiles of the observed X.
- A cubic spline with K knots has $K+4$ parameters or degrees of freedom.
- A natural spline with K knots has K degrees of freedom.



UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Smoothing Splines

- Suppose now that we have data and to fit a smooth function $y = g(x)$ and we can achieve so by solving the penalized regression problem

$$\min_{g \in \mathcal{S}} \left\{ \sum_{i=1}^{N} (y_i - g(x_i))^2 + \lambda \int_{x_1}^{x_N} \left[g''(x) \right]^2 dx \right\}, \tag{3.1}$$

  for a *smoothing parameter* $\lambda > 0$, where $\mathcal{S}$ is certain function space and can be $C^2[x_1, x_N]$.
- The first term is a fidelity term and tries to make $g(x)$ match the data at each x.
- The smoothing parameter λ is a roughness penalty and controls how wiggly $g(x)$ is.
- The smaller the λ, the more wiggly the function, eventually interpolating y_i when $\lambda = 0$.
- As $\lambda = \infty$, the function $g(x)$ becomes linear.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Smoothing Splines

- The solution is a natural cubic spline with a knot at every unique value of x_i.
- The roughness penalty term still controls the roughness via λ.
- Once the basis functions are determined, we have a parametric problem:

$$\min_{\theta}\left\{\|\mathbf{y}-\mathbf{L}\theta\|^2+\lambda\theta^T\mathbf{G}\theta\right\}$$

where $\mathbf{L}$: $n \times N$ has rows $\mathbf{b}^T(x_i)$ and $\boldsymbol{G}_{jk}=\int_{x_1}^{x_N} b_j''(x)\, b_k''(x)\, dx$.

- This gives

$$\hat{\boldsymbol{\theta}}=\left[\mathbf{L}^T\mathbf{L}+\lambda\mathbf{G}\right]^{-1}\mathbf{L}^T\mathbf{y},\ \ \hat{\boldsymbol{y}}=L\hat{\boldsymbol{\theta}}=\mathbf{S}_\lambda\mathbf{y},$$

where the smoother matrix $\mathbf{S}_\lambda$ plays the same role as the hat matrix.

- The *equivalent degrees of freedom* are (thusly)

$$df_\lambda=tr\left[\mathbf{S}_\lambda\right].$$

Smoothing Splines

- The smoothness is controlled by λ; it and df_λ can be determined from each other.
- The R function will determine an optimal λ_0 by generalized cross-validation:

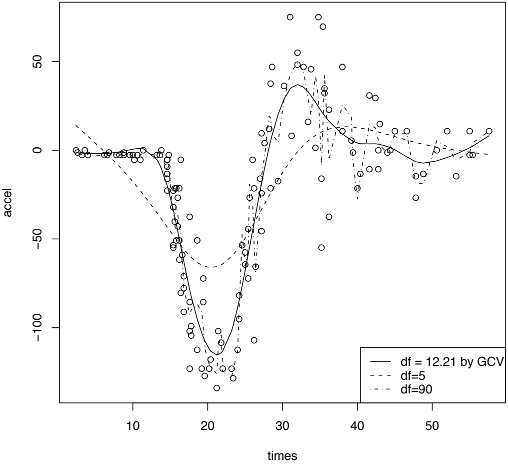
$$\lambda_0 = \arg\min_{\lambda} \frac{\left\|\left(\mathbf{I} - \mathbf{S}_\lambda\right)\mathbf{y}\right\|^2}{n - df_\lambda}.$$

(What is the RHS as $\lambda \to \infty$?)

- This is the default; an option is ordinary cross-validation:

$$\lambda_0 = \arg\min_{\lambda} \sum_{i=1}^{n} \left(\frac{e_i(\lambda)}{1 - \left[\mathbf{S}_\lambda\right]_{ii}}\right)^2.$$

Motorcycle Data



Summary and Remark

- Polynomial regression
- Splines
- Read textbook Chapter 5 and `R` code
- Do `R` lab