Lecture 3 – Linear Regression and Extensions

Martin Spindler

2016-04-26

Extensions

- Polynomial Regression
- Step Functions
- Basis Functions
- Regression Splines
- Smoothing Splines

Regression Splines

- Polynomial regressions often leads to rough and instable estimates.
- Solution: Fitting separate low-degree polynomials over different regions (and make them smooth)
- Construction of splines:
- Partition x-axis into different smaller sub-intervals and estimate a separate polynomial for each interval.
- Additionally, it is required that the combined function is smooth (e.g. continuously differentiable) at the boundary points (knots).

Regression Splines | Defintion

Let $a = c_1 < c_2 < \ldots < c_K = b$ be a partition of the interval $[a, b]$. A function $s : [a, b] \to \mathbb{R}$ is called a polynomial spline of degree l if

1. $s(z)$ is a polynomial of degree l for $z \in [c_j, c_{j+1}), 1 \leq j < m$.
2. $s(z)$ is $(l-1)$-times continuously differentiable.

$c_1, \ldots, c_K$ are called knots of the splines and $\Omega = \{c_1, \ldots, c_k\}$ knot set.

Regression Splines

It can be shown that regression splines form a vector space of dimension of dimension $K + l - 1$.
Hence every regression spline can be represented as the sum of $K + l - 1$ basis functions:

$$s(z) = \beta_0 B_0(z) + \ldots + \beta_{k+l-2} B_{K+l-2}(z).$$

Basis functions: truncated power series basis, B-spline basis (numerical more stable)
Truncated power series:

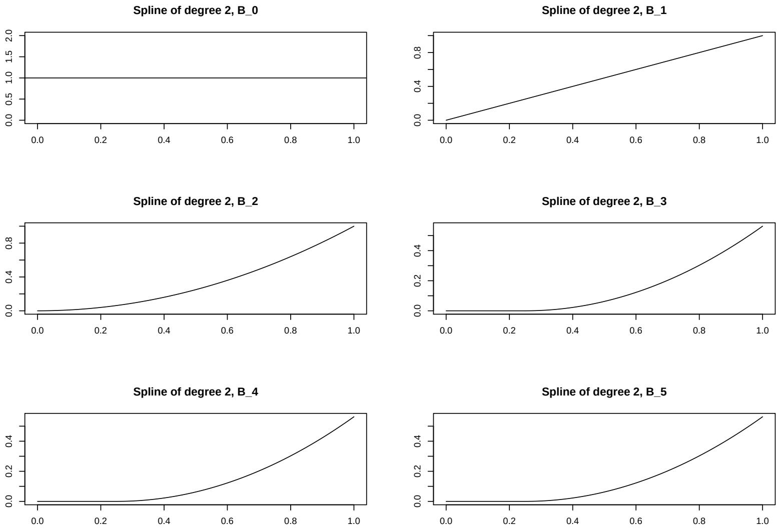
$$s(z) = \sum_{j=0}^{l} \beta_j z^j + \sum_{j=2}^{K-1} \beta_{l+j-1} (z - c_j)_+^l$$

with $(z - c_j)_+^l = \max(0, (z - c_j))^l$.

Regression Splines | Example Basis Functions

We consider the interval $[0, 1]$ and knots $0 < 0.25 < 0.5 < 0.75 < 1$. For a quadratic spline and 5 knots, the number of basis functions is given by $2 + 5 - 1 = 6$.

Regression Splines | Example Basis Functions

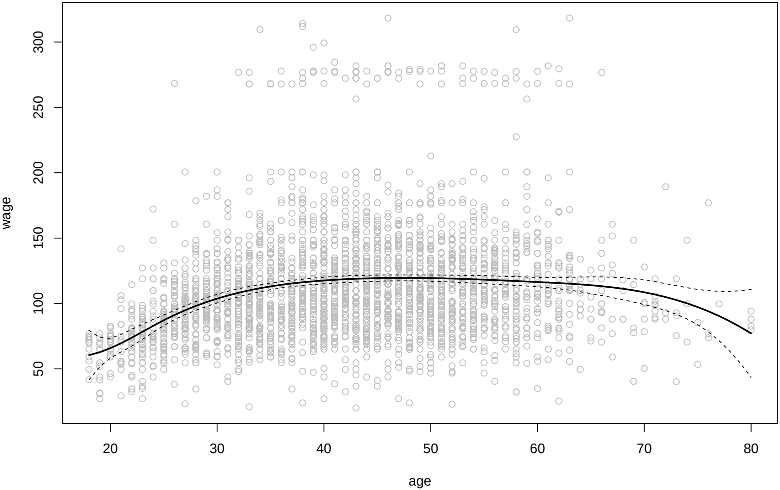


Regression Splines

How to choose the number and location of knots?

- Equi-distant knots
- Choice according to quantiles of the x-variable

Regression Splines | Example



Natural Splines

- Problem: Regression splines tend to display erratic behavior at the boundaries of the domain leading to high variance.
- Solution: additional constraints at the boundary (left of the leftmost knot and right of the most rightmost knot)
- Definition **Natural Spline**
  A natural spline of degree l is a regression spline of degree l with the additional constraint that
  it is a polynomial of degree $(k-1)/2$ on $(-\infty, c_0]$ and $[c_K, +\infty)$.
- Most popular natural splines are cubic which are linear beyond the boundaries.
- Modifications of the truncated power basis and B-spline basis for natural splines (here dimension K!)

Smoothing Splines

- Optimization problem: Among all functions $f(x)$ with two continuous derivatives, minimize:

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

- λ is called *smoothing parameter* (interpretation?)
- It can be shown that the solution of the optimization problem is unique and a natural cubic spline with knots at the unique values of the $x_i, i = 1, \ldots, n$.
- Here: no problem how to choose the knots (as in the regression spline case)
- Intuition: Overparametrization (because of n knots), but penalization

Snoothing Splines

Since the solution is a natural spline, we can write it as

$$f(x) = \sum_{j=1}^{n} b_j(x)\beta_j$$

with $b_1(\cdot), \ldots, b_n(\cdot)$ an n dimensional set of basis functions for representing the family of natural splines.

Snoothing Splines

Then the criterion reduces to

$$RSS(\beta, \lambda) = (y - B\beta)^T (y - B\beta) + \lambda \beta^T \Omega_n \beta$$

where $B_{ij} = b_j(x_i)$ and $(\Omega_n)_{jk} = \int b_j''(d) b_k''(t) dt$.
The solution is given by

$$\hat{\beta} = (B^T B + \lambda \Omega_n)^{-1} B^T y.$$

(generalized Ridge regression).

Snoothing Splines

The fitted smoothing spline is given by

$$\hat{f}(x) = \sum_{j=1}^{n} b_j(x)\hat{\beta}_j$$