An Introduction to Statistical Learning
Ch7: Moving beyond linearity

Authors: Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani
Presenter: Fei Yang
2022/01/13

Content目录

非参数分析方法的发展简史

1946-1948

Spectral Density Estimation

1956, 1962

Rosenblatt (1956) -Parzen (1962): Kernel Probability Density Estimation

1964

Nadaraya (1964)-Watson (1964): Regression Estimation

1970-1990

- Series
- Splines
- Local Polynomial
- Artificial Neural Network

$$h(w) = \int_{-\pi}^{\pi} \hat{I}(w) W_T(\lambda - w) d\lambda, w \in [-\pi, \pi]$$

平滑（Smoothing）：非参数分析的基本思想

全局平滑法

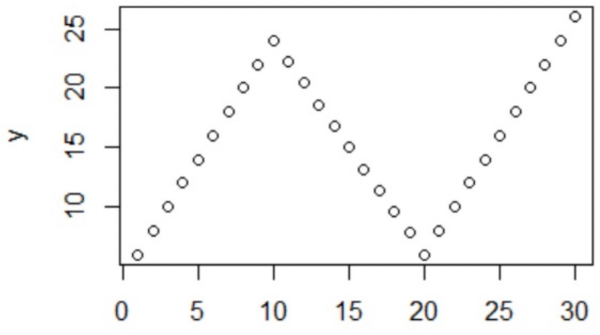
局部平滑法

PART 1

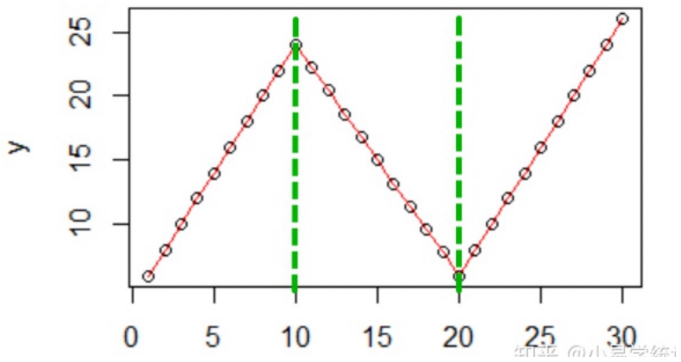
全局平滑法
Global Smoothing

7.1 Polynomial Regression多项式回归

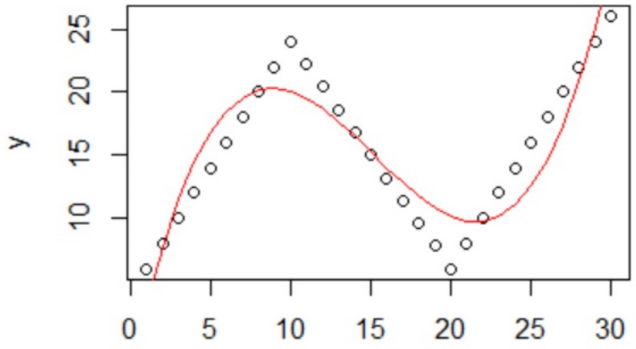
- 现实中，线性假设很可能不能满足实际需求，甚至直接违背实际情况。在第6、7章会介绍一些方法来弥补线性模型的不足，通过降低模型的复杂度和估计量的方差来改善模型。但我们还可以通过另一种方式来改善模型，那就是改变”线性假设”。



dataset



分段处理



全局多项式回归

7.1 Polynomial Regression多项式回归

- 现实中，线性假设很可能不能满足实际需求，甚至直接违背实际情况。在第6、7章会介绍一些方法来弥补线性模型的不足，通过降低模型的复杂度和估计量的方差来改善模型。但我们还可以通过另一种方式来改善模型，那就是改变"线性假设"。接下来，我们介绍一些由线性假设下，扩展得到的其他模型。

多项式回归是一种通过增加自变量上的次数，而将数据映射到高维空间的方法，从而提高模型拟合复杂数据的效果。

	线性模型	非线性模型-多项式回归
优势	• 易于描述和实现 • 解释性能和推断理论更有优势	• 放松线性假设，且尽可能保证可解释性 • 随着变量的增加，可以拟合出异常极端变化的曲线
劣势	• 预测效力较弱	• 降低可解释性，随着数据维度和多项式次数的上升，方程也变得异常复杂 • 多重共线性

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

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i,$$

$$y_i = \beta_0 + \sum_{j=1}^{d} \beta_j x_i^j + \epsilon_i$$

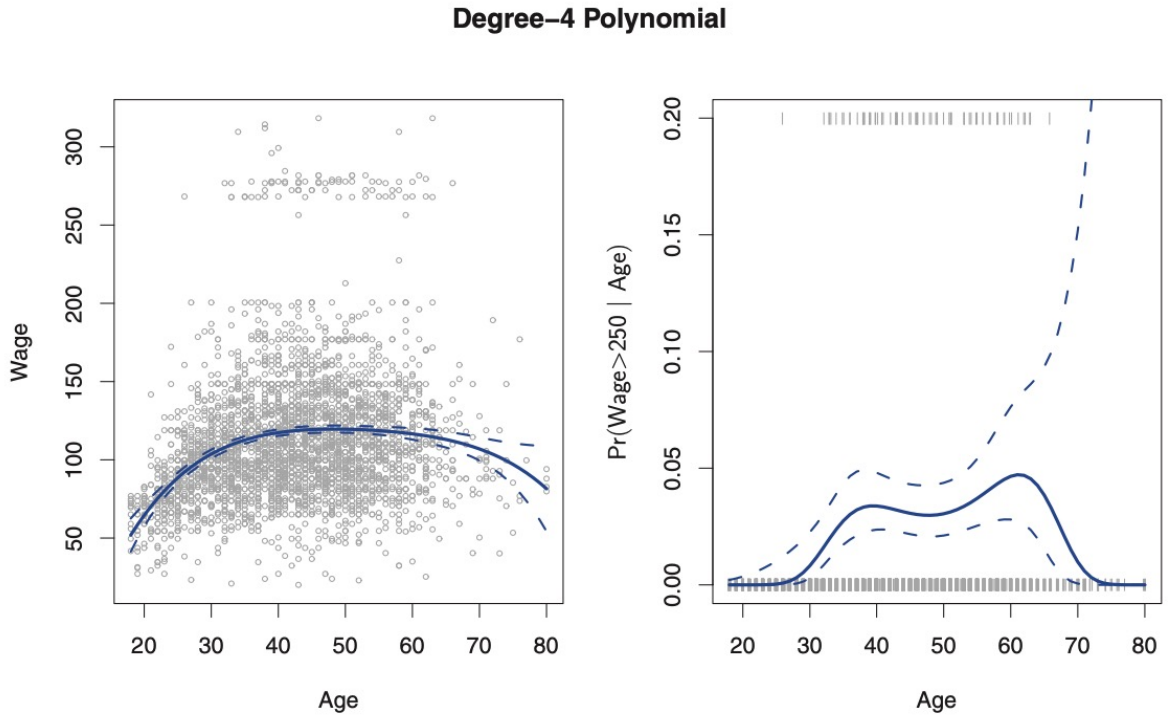
7.1 Polynomial Regression 多项式回归

Wage & Age Non-Linear Relation

回归问题

Let x_0 be the value of age , to predict wage:

$$\hat{f}(x_0) = \hat{\beta}_0 + \sum_{d=1}^{4} \hat{\beta}_d x_0^d$$



分类问题

通过分成[high/ low earners] 处理为二分类变量

$$P(y_i > 250|x_i) = \frac{e^{(\beta_0+\beta_1 x_i^1+\cdots+\beta_d x_i^d)}}{1+e^{(\beta_0+\beta_1 x_i^1+\cdots+\beta_d x_i^d)}}$$

「Variance」

Compute Variance of the fit, $Var(\hat{f}(x_0))$, we need:

- Variance Estimates for each of the fitted coefficients $\hat{\beta}_j$ from Least Squares
- The Covariances between pairs of coefficient estimates,
  - Let $\hat{C}$ be the 5x5 covariance matrix of the $\hat{\beta}_{j=0,1,2,3,4,}$
- Let $X_0^T = (1, x_0, x_0^2, x_0^3, x_0^4)$:

$$Var[(\hat{f}(x_0))] = X_0^T \hat{C} X_0$$

$\sqrt{Var(\hat{f}(x_0))}$ is the *estimated pointwise standard error* of $\hat{f}(x_0)$

- As *EACH reference point* x_0 , this computation is repeated and get the fitted curve and twice the standard error

7.2 Step Functions 阶跃函数/piecewise linear regression 逐段线性回归

- 不论简单线性回归、多项式回归等都是具有全局性的结构。如不考虑全局性回归时，可以用到逐段线性回归。
- 最简单的分段回归应该就形如阶跃函数了，阶跃函数是使得连续变量离散化成有序分类变量的方法。

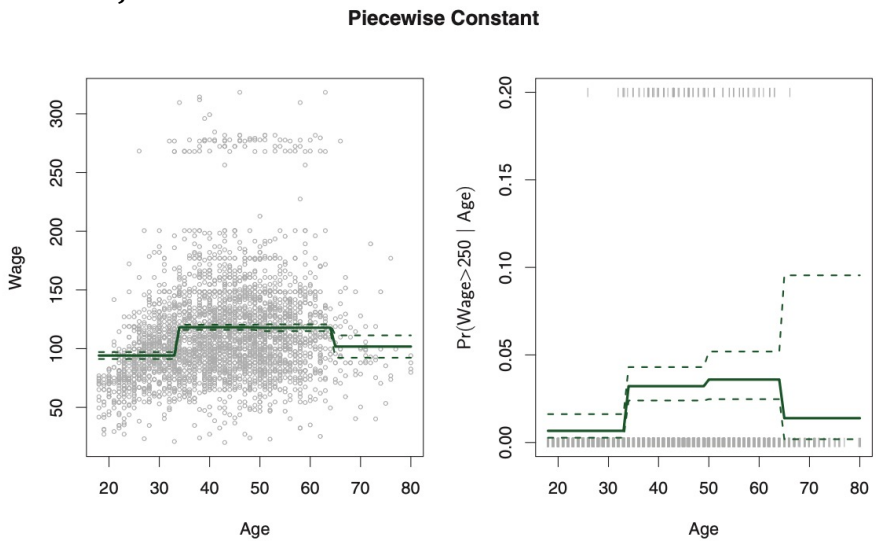
- 过程：
  - 将自变量X分成多区间，各组

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

- where I(·) is an indicator function that returns a 1 if the condition is true, and returns a 0 otherwise.
- 用K个分割点 生成K+1个新变量，且X必出现在其中某一区间,则有 $C_0(X) + C_1(X) + \ldots + C_K(X) = 1$
- 因此可用 $C_1(X), \ldots, C_K(X)$ 作为预测变量建模

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

- β_0 can be interpreted as the mean value of Y for X $< c_1$
- β_j can represent the average increase in the response for X in $c_j \le X \le c_{j+1}$ relative to $X < c_i$



$$P(y_i > 250|x_i) = \frac{e^{\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i)}}{1 + e^{\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i)}}$$

7.2 Step Functions 阶跃函数/piecewise linear regression 逐段线性回归

- 不论简单线性回归、多项式回归等都是具有全局性的结构。如不考虑全局性回归时，可以用到逐段线性回归。
- 最简单的分段回归应该就形如阶跃函数了，阶跃函数是提供连续变量离散化成有序分类变量的方法。

- 过程：
  - 将自变量X分成多区间，各组

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

- where I(·) is an indicator function that returns a 1 if the condition is true, and returns a 0 otherwise.
- 用K个分割点 生成K+1个新变量，且X必出现在其中某一区间,则有 $C_0(X) + C_1(X) + \ldots + C_K(X) = 1$
- 因此可用 $C_1(X), \ldots, C_K(X)$ 作为预测变量建模

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

- 关于分割点的选取类似**分箱法**：
  - (1)等距：将X的取值范围分成等宽的箱
  - (2)等量：将各区间包含相同数量的观测
  - 注意：此处各箱互不重合

Precaution: unless there are natural breakpoints in the predictors, piecewise-constant functions can miss the action.

7.3 Basis Functions 基函数

- 前两种回归方式，**都可以总结为是基底函数的特例。**所谓的基底函数是不直接对x进行回归，而是用函数变化后的 $b_i(x_i)$ 值进行回归，形如：

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

- 基函数：$b_1(\cdot), b_2(\cdot), \ldots, b_K(\cdot)$ 确定已知的：
  - 多项式回归： $b_j(x_i) = x_i^j$
  - 分段回归：$b_j(x_i) = \mathrm{I}\left(c_j \leq x_i < c_{j+1}\right)$
  - 其他：傅里叶序列、小波基
- 可以视为是以$b_1(x_i), b_2(x_i), \ldots, b_k(x_i)$ 为预测变量的标准线性模型
- 估计方法：least squares

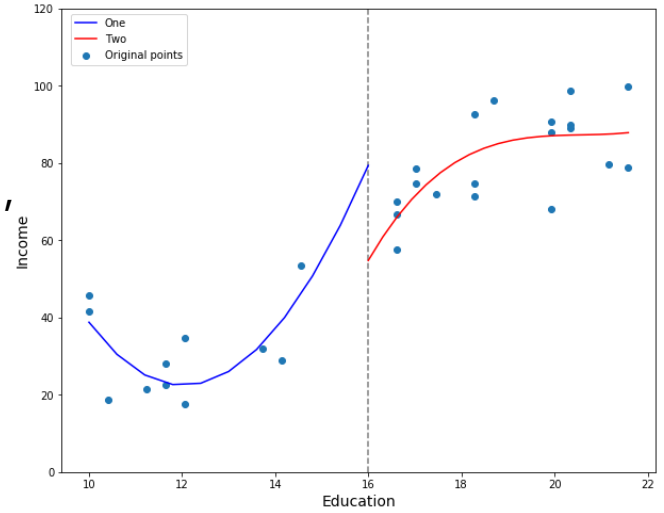
7.4 Regression Splines 样条回归

7.4.1 Piecewise Polynomials 分段多项式回归

- 由基底函数可知，基函数可以是任意的组合形式，接下来介绍一种基函数是**结合了多项式和逐段线性回归**的形式，即分段多项式回归

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i,$$

- **动机：**
  - Fitting separate low-degree polynomials over different region of X to avoid fitting a high-degree polynomial over entire range of
- **结点：**
  - 改变点称为knots，K个结点, 即K+1个方程
  - 自由度$(d+1)\times(K+1)$：1条d阶多项式曲线的自由度是(d+1):$\beta_0 + \beta_1 x + \cdots + \beta_d x^d$ ，K个节点将空间分成K+1份,总自由度是$(d+1)\times(K+1)$



7.4 Regression Splines 样条回归

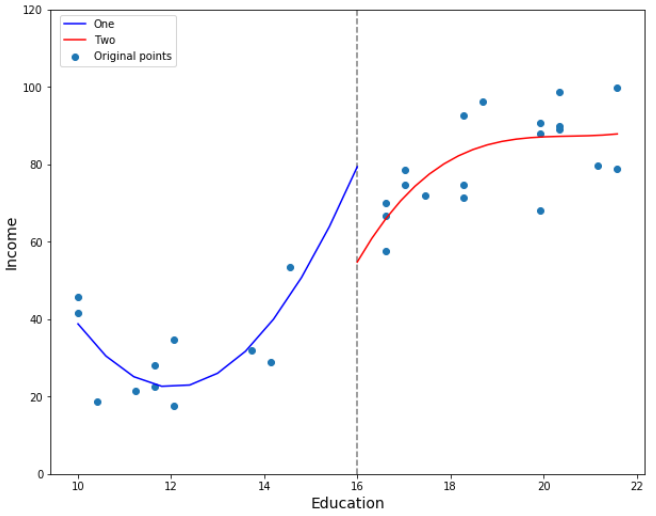
7.4.1 Piecewise Polynomials 分段多项式回归

- 由基底函数可知，基函数可以是任意的组合形式，接下来介绍一种基函数是**结合了多项式和逐段线性回归**的形式，即分段多项式回归

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i,$$

- **动机：**
  - Fitting separate low-degree polynomials over different region of X to avoid fitting a high-degree polynomial over entire range of X 可理解为具有变系数的多项式回归
- **结点：**
  - 改变点称为knots，K个结点, 即K+1个方程
  - 自由度: $(d+1)\times(K+1)$
  - 实际举例：税法改革使相关支出的回归线会在法案生效时点发生变化
- **缺点：**受异常点影响较大，需要加上额外的边界约束。
  - 举例：一个三次多项式分段回归如下，但该回归不满足“平滑特性”,在结点处可能出现断裂或跳跃（如右图）。

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



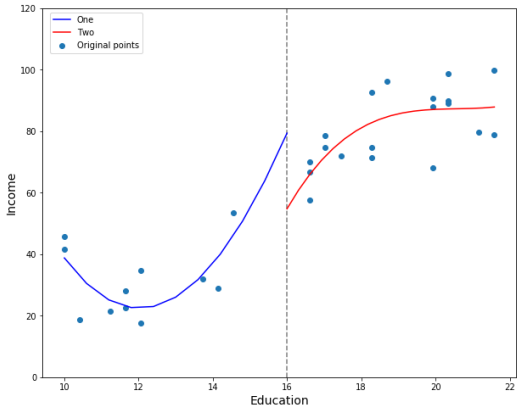
7.4 Regression Splines 样条回归

7.4.2 Constraints and Splines带约束的分段多项式回归=样条回归

- 样条是一个函数，由多项式构造的分段函数，并且在分段节点处要具有高度平滑的特性，即在分段结点处连续的导数。

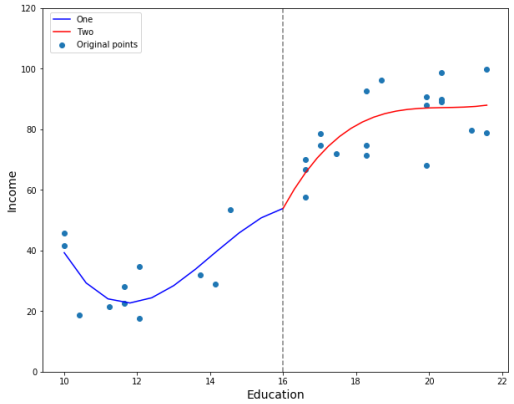
- **性质：**
  - **样条函数具有连续性和光滑性；（**在结点处的函数值、斜率上相同**）**

三次多项式分段回归



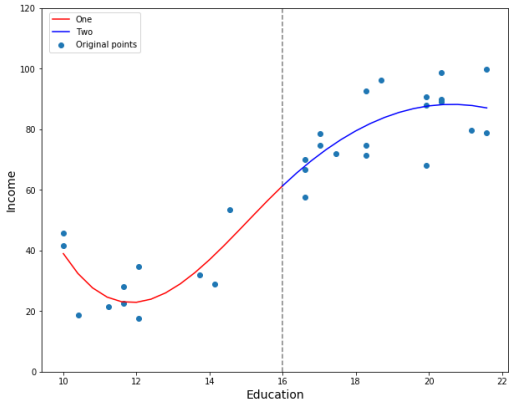
- 无约束

连续的三次多项式分段回归



- 结点处连续
- $f(C_-) = f(C_+)$

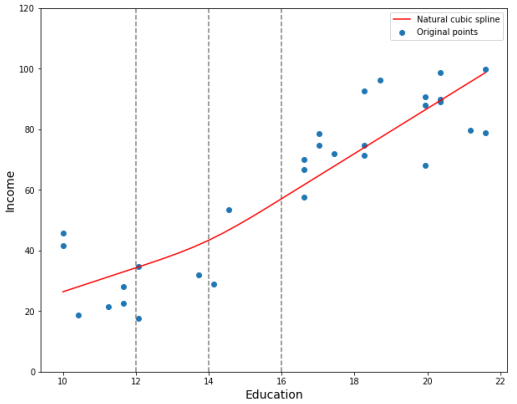
三次样条回归



- 结点处连续且光滑

(一阶导与二阶导在节点处连续)

自然三次样条回归



- 结点处连续且光滑
- 边界线性-超出末端节点之外的拟合是线性的

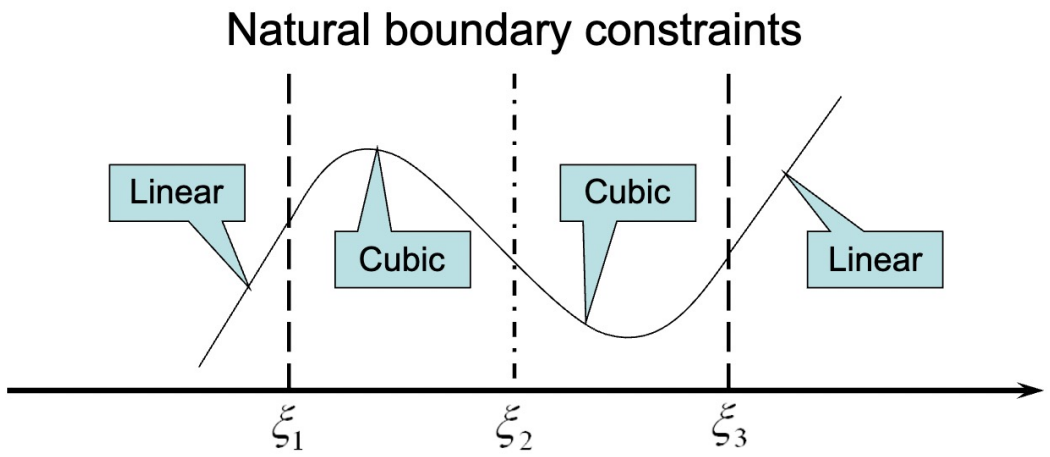
自由度$(d+1)\times(K+1)$

由于每个节点上有d个约束（从0到$d-1$阶导数相等），最终自由度是总自由度减去总约束度：$(d+1)\times(K+1)-K\times d = K+d+1$

7.4 Regression Splines 样条回归
关于自然三次样条回归的补充说明

Natural cubic spline

- Natural cubic spline adds additional constraints, namely that the function is linear beyond the boundary knots.



0、定义

已知函数$f(x)$在区间$[a,b]$上$n+1$个互异节点，$a=x_0<x_1<\cdots<x_n=b$处的函数值为$y_i=f(x_i)$,若构造函数$s(x)$,满足:

1. $s(x_i)=y_i(i=0,1,\cdots,n)$
2. 在每个小区间$[x_i,x_{i+1}]$上是一个不超过三次的多项式
3. $s(x),s^{'}(x),s^{''}(x)$在$[a,b]$上连续

则称$s(x)$为$f(x)$的三次样条插值函数。

根据定义知道规律为:

已知:

- n+1个数据点[xi, yi], i = 0, 1, …, n
- 每一分段都是三次多项式函数曲线
- 节点达到二阶连续
- 左右两端点处特性（自然边界，固定边界，非节点边界）

根据定点，求出每段样条曲线方程中的系数，即可得到每段曲线的具体表达式。

7.4 Regression Splines 样条回归
关于自然三次样条回归的补充说明

则这个三次方程可以构造成如下形式：

$y = a_i + b_i x + c_i x^2 + d_i x^3$ 这种形式,我们称这个方程为三次样条函数 $S_i(x)$ 。

从 $S_i(x)$ 可以看出每个小区间有四个未知数 (a_i, b_i, c_i, d_i) ，有n个小区间，则有4n个未知数，要解出这些未知数，则我们需要4n个方程来求解。

求解

我们要找出4n个方程来求解4n个未知数

首先，由于所有点必须满足插值条件， $S(x_i) = y_i \quad (i = 0, 1, \ldots, n)$ ，除了两个端点，所有n-1个内部点的每个点都满足 $S_i(x_{i+1}) = y_{i+1} \quad S_{i+1}(x_{i+1}) = y_{i+1}$ 前后两个分段三次方程，则有2(n-1)个方程，再加上两个端点分别满足第一个和最后一个三次方程，则总共有2n个方程；

其次，n-1个内部点的一阶导数应该是连续的，即在第 i 区间的末点和第 i+1 区间的起点是同一个点，它们的一阶导数应该也相等，即 $S_i'(x_{i+1}) = S_{i+1}'(x_{i+1})$ 则有n-1个方程

另外，内部点的二阶导数也要连续，即 $S_i''(x_{i+1}) = S_{i+1}''(x_{i+1})$,也有n-1个方程

现在总共有4n-2个方程了，还差两个方程就可以解出所有未知数了，这两个方程我们通过边界条件得到。

有三种边界条件：自然边界，固定边界，非节点边界

1，自然边界 (Natural Spline)：指定端点二阶导数为0， $S''(x_0) = 0 = S''(x_n)$

2, 固定边界 (Clamped Spline): 指定端点一阶导数，这里分别定为A和B。即
$S_0'(x_0) = A, \quad S_{n-1}'(x_n) = B$

3, 非扭结边界(Not-A-Knot Spline): 强制第一个插值点的三阶导数值等于第二个点的三阶导数值，最后第一个点的三阶导数值等于倒数第二个点的三阶导数值. 即
$S_0'''(x_0) = S_1'''(x_1) \quad and \quad S_{n-2}'''(x_{n-1}) = S_{n-1}'''(x_n)$

7.4 Regression Splines 样条回归
关于自然三次样条回归的补充说明

Natural Cubic Splines

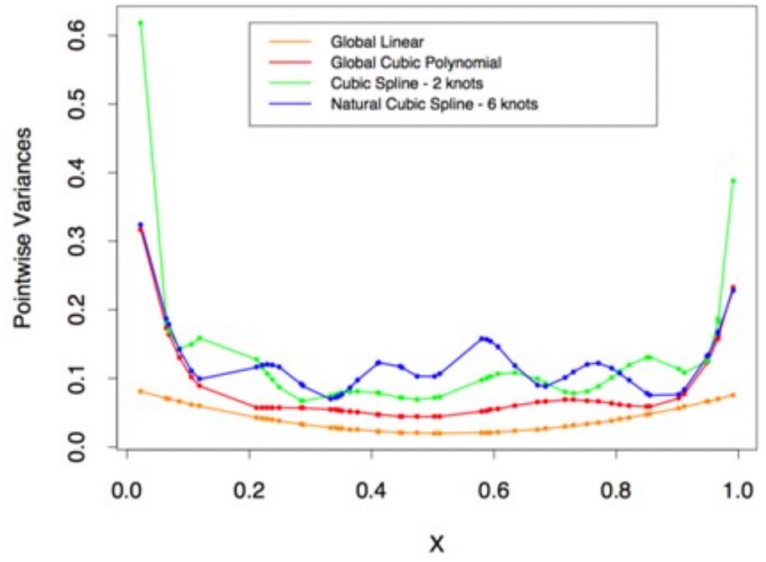
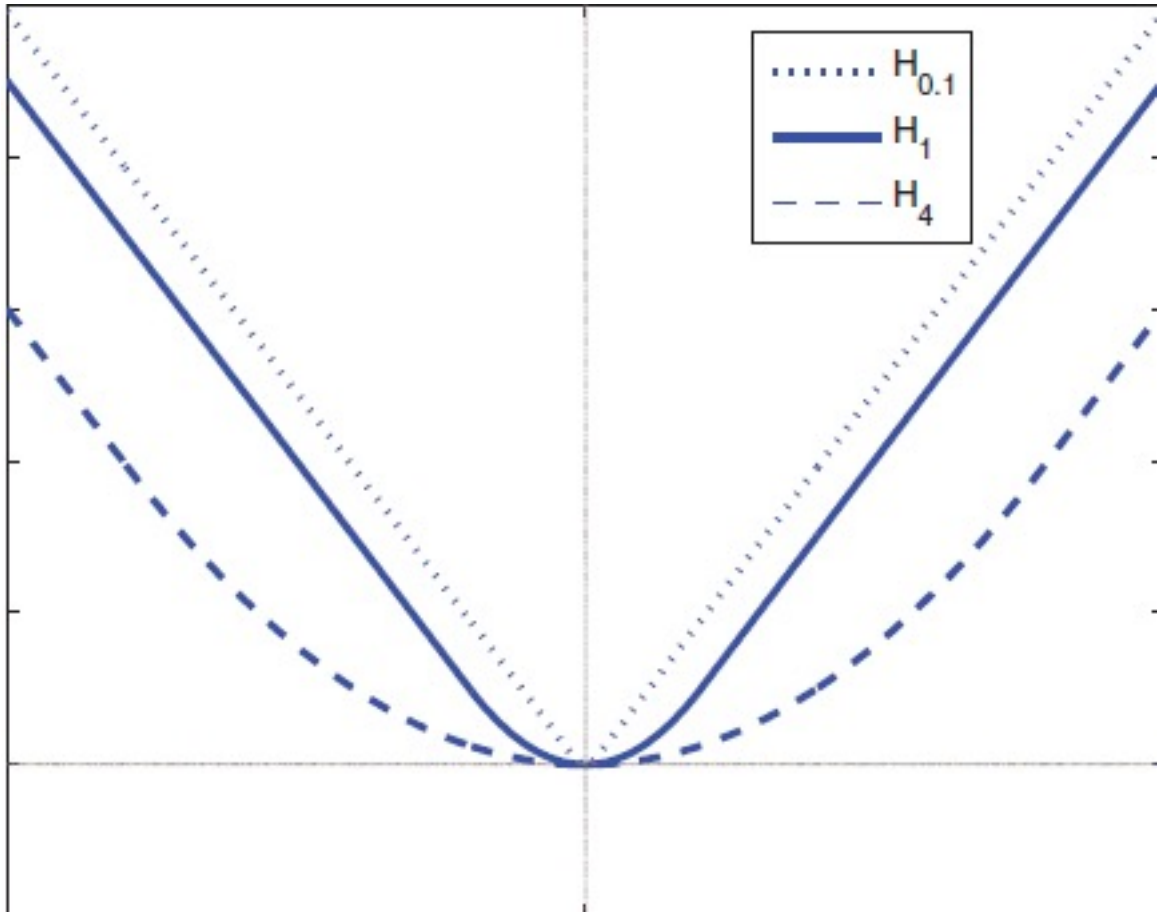


FIGURE 5.3. *Pointwise variance curves for four different models, with X consisting of 50 points drawn at random from U[0, 1], and an assumed error model with constant variance. The linear and cubic polynomial fits have two and four degrees of freedom, respectively, while the cubic spline and natural cubic spline each have six degrees of freedom. The cubic spline has two knots at 0.33 and 0.66, while the natural spline has boundary knots at 0.1 and 0.9, and four interior knots uniformly spaced between them.*

How do you calculate a cubic spline?

- Given (x_1, y_1), (x_2, y_2), and (x_3, y_3)
- $ax_2^3 + bx_2^2 + cx_2 + d = y_2$
- $ex_2^3 + fx_2^2 + gx_2 + h = y_2$
- $3ax_2^2 + 2bx_2 + c = 3ex_2^2 + 2fx_2 + g$
- $6ax_2 + 2b = 6ex_2 + 2f$
- $6ax_1 + 2b = 0$
- $6ex_3 + 2f = 0$

通过增加约束条件实现：连续函数→光滑函数



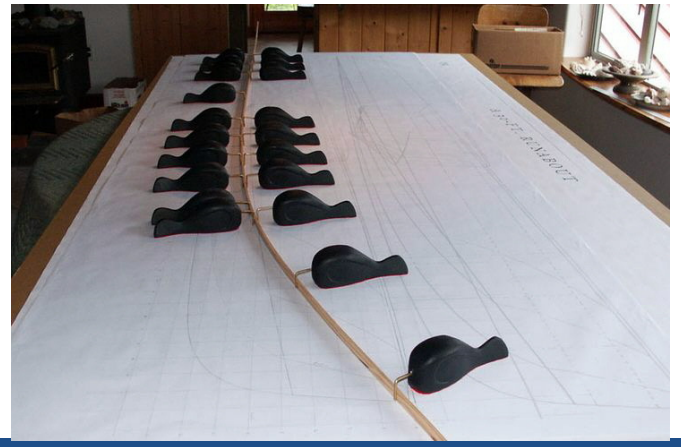
7.4 Regression Splines 样条回归

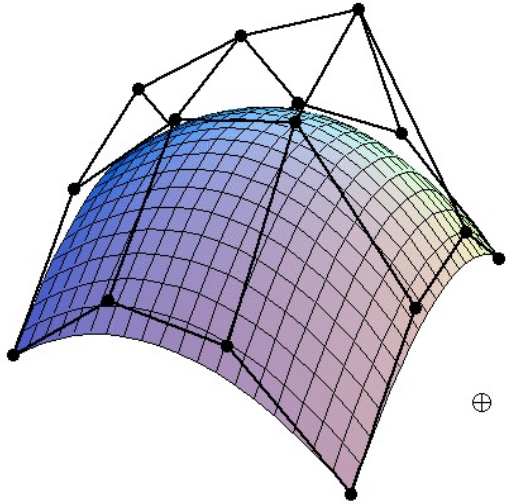
7.4.2 Constraints and Splines带约束的分段多项式回归=样条回归

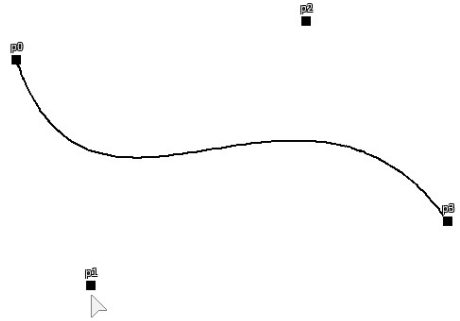
- 样条是一个函数，由多项式构造的分段函数，并且在分段节点处要具有高度平滑的特性，即在分段结点处连续的导数。

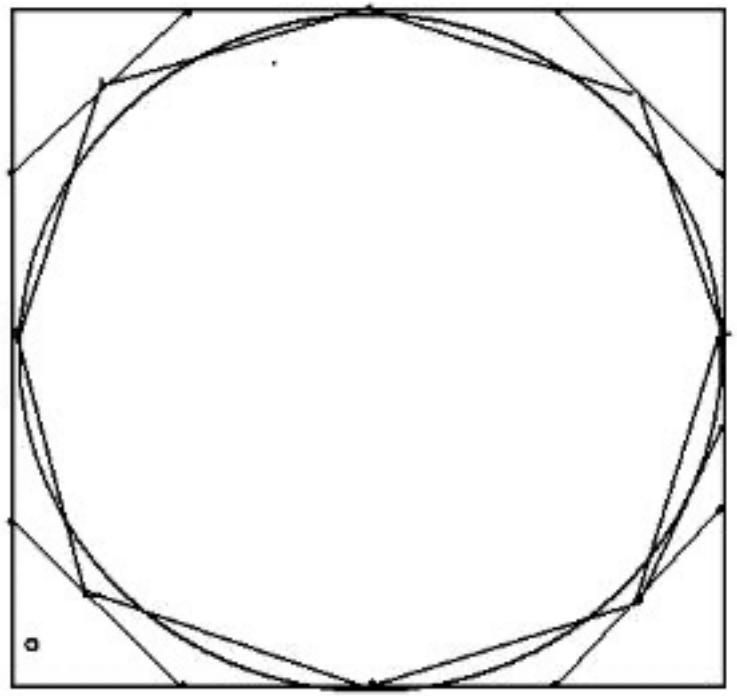
- **历史来源：**
  - 在船体数学放样或飞机外形设计中，人们借助细长的木质样条和压铁，绘制所需曲线，这就是样条曲线。
  - **材料力学视角：**
    - 木质样条=弹性梁，压铁作用点=集中载荷作用点，样条曲线=集中载荷作用下的细梁挠度曲线或弹性曲线。
    - 小挠度的弹性曲线，在数学上是分段三次多项式，集中载荷作用点即为节点。
  - **计算机图形学视角：**
    - Bezier 曲线, Bezier 曲面,

- **种类举例：**
  - **三次样条**
  - **B-样条**
  - **自然样条**





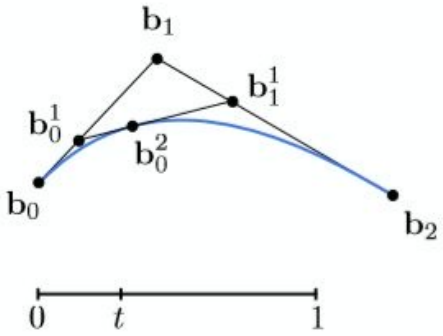




通过正方形内切画圆

B样条基函数

Example: quadratic Bézier curve from three points



$$\mathbf{b}_0^1(t) = (1-t)\mathbf{b}_0 + t\mathbf{b}_1$$

$$\mathbf{b}_1^1(t) = (1-t)\mathbf{b}_1 + t\mathbf{b}_2$$

$$\mathbf{b}_0^2(t) = (1-t)\mathbf{b}_0^1 + t\mathbf{b}_1^1$$

$$\mathbf{b}_0^2(t) = (1-t)^2\mathbf{b}_0 + 2t(1-t)\mathbf{b}_1 + t^2\mathbf{b}_2$$

7.4 Regression Splines 样条回归

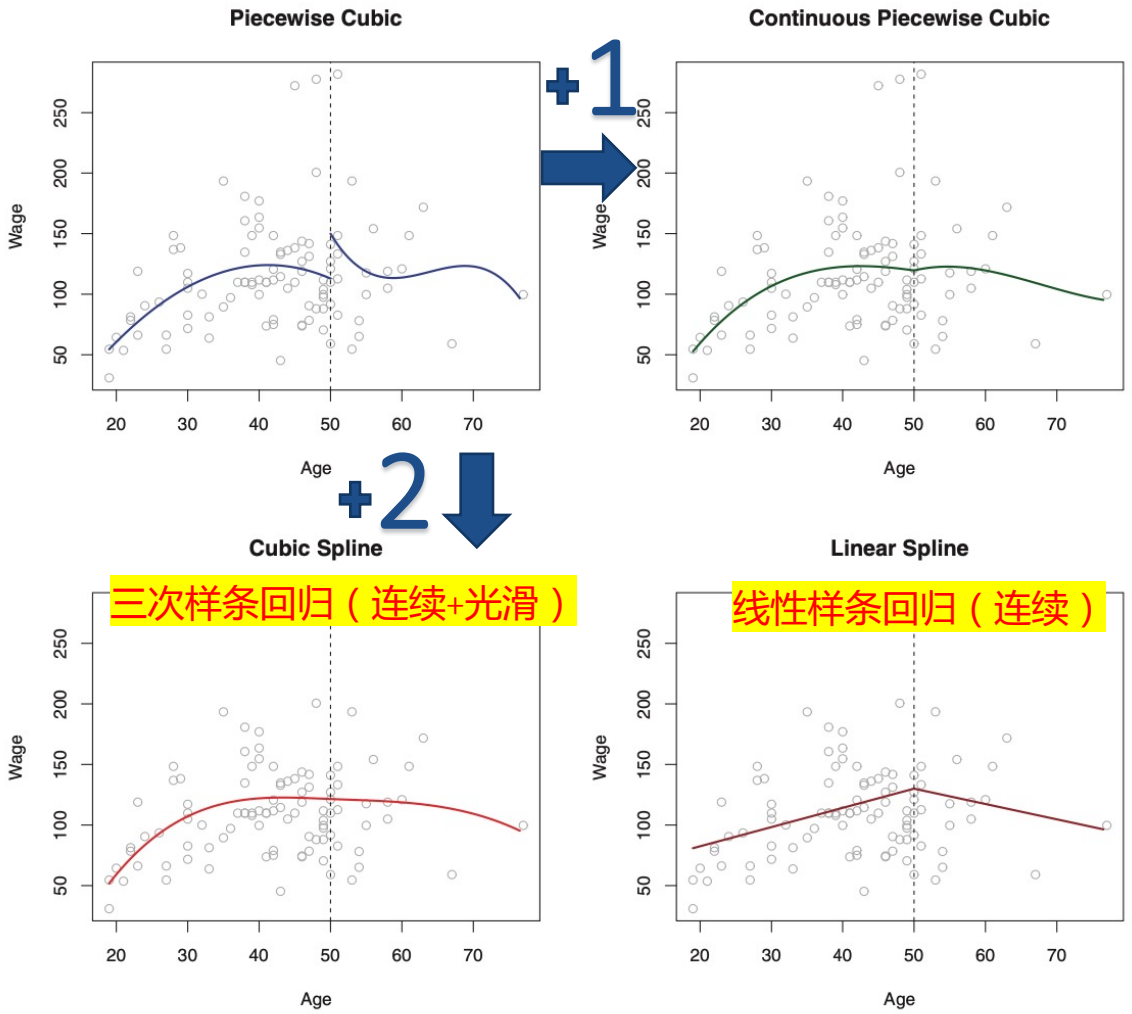
7.4.2 Constraints and Splines从分段多项式到样条回归

1）节点已知的情况

- 左上：
  - 三次多项式分段回归
  - 结点处出现跳跃，自由度为(3+1)(1+1)=8
- 右上：
  - [连续的]三次多项式分段回归
  - 在函数值、斜率上相同
  - 左上基础上增加“结点处连续”的条件；
  - 但仍呈现不自然的V-shape
- 左下：
  - 三次样条回归（cubic spline）
  - 在函数值、斜率上相同；最高为三次项
  - 使用K 个结点, 4+K个自由度
  - 在左上基础上增加两个限制条件，使其变成样条回归：
    - 结点处连续（add一阶可导条件）
    - 结点处光滑（add二阶可导条件）
  - 当前自由度为8-3=5，(放松自由度：方程连续性、一阶连续、二阶连续)
- 右下：
  - 逐段线性样条回归（linear spline）
  - 在函数值上相同；最高为一次项
  - 保证各个结点在d-1阶导数都具有连续性

三次多项式分段回归

三次多项式分段回归（连续）



7.4 Regression Splines 样条回归
7.4.3 The Spline Basis Representation

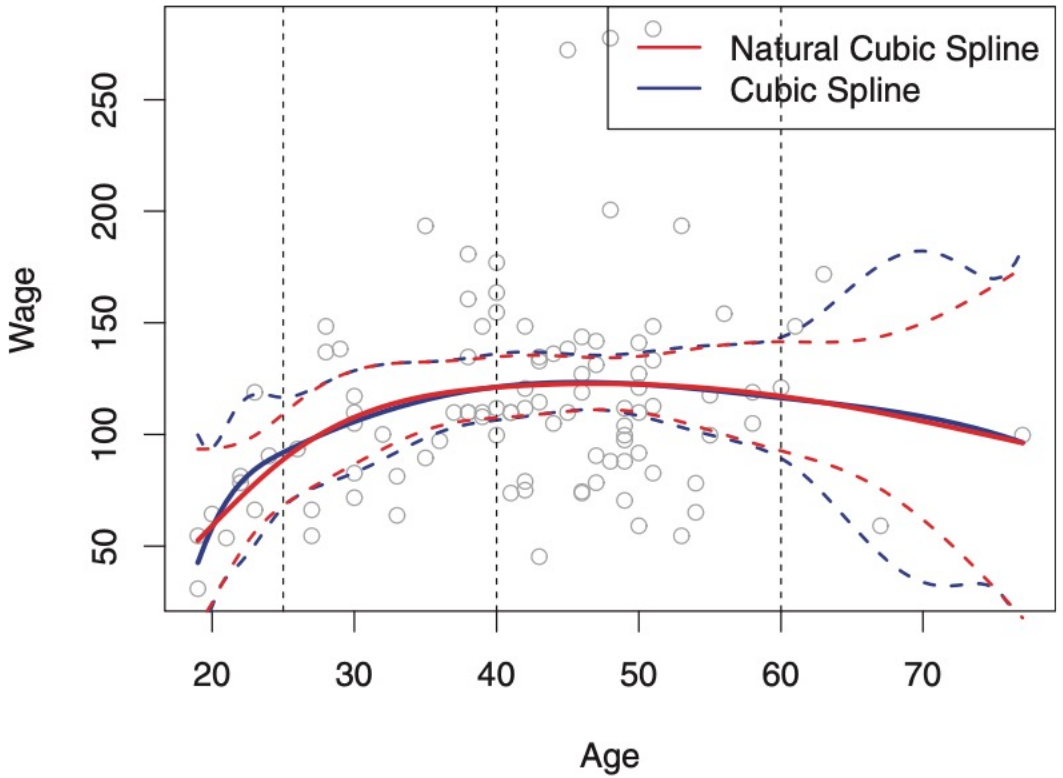
- **三次样条回归模型**Cubic Splines

$$y_i = \beta_0 + \sum_{d=1}^{K+3} \beta_d b_d(x_i) + \epsilon_i$$

- 三次样条下，基函数b()也有非常多的选择，最直接的方法就是对$X, X^2, X^3, h(X, \xi_1), \ldots h(X, \xi_K)$进行拟合，其中$\xi_K$是结点

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

- 对公式的理解：类似虚拟变量，在结点后发生突变



7.4 Regression Splines 样条回归
7.4.3 The Spline Basis Representation

- **三次样条回归模型**Cubic Splines

A cubic spline with knots at ξ_k, $k = 1, \ldots, K$ is a piecewise cubic polynomial with continuous derivatives up to order 2 at each knot.

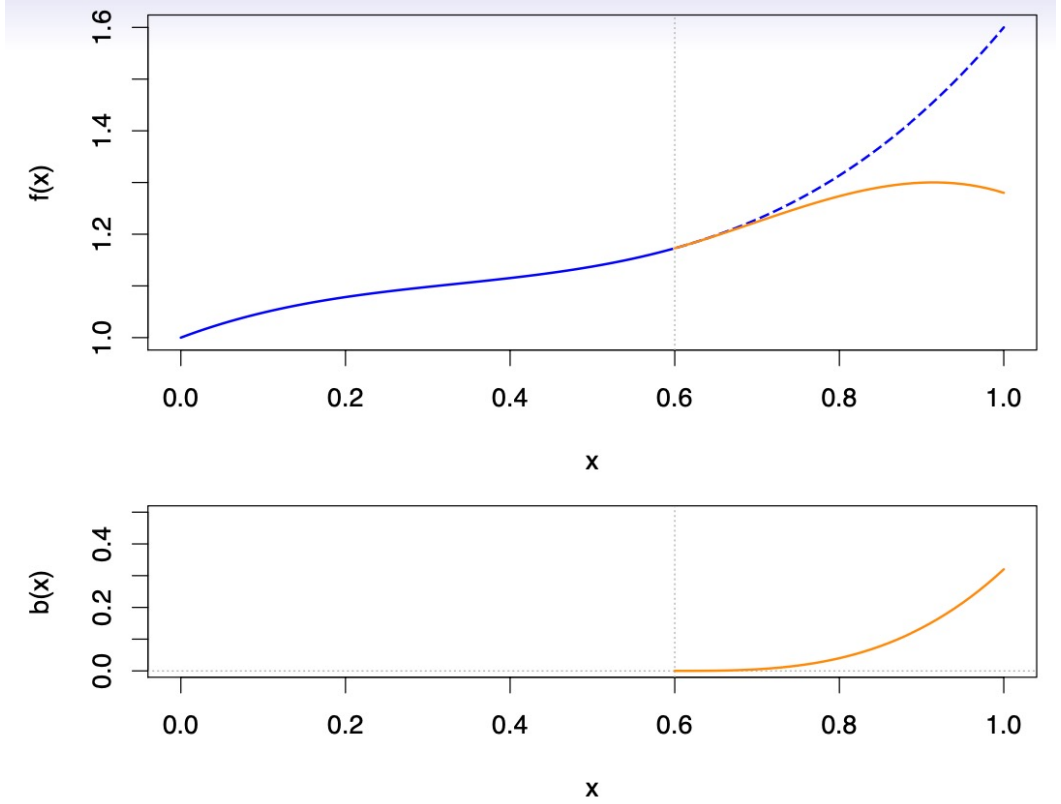
Again we can represent this model with truncated power basis functions

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

$$\begin{aligned} b_1(x_i) &= x_i \\ b_2(x_i) &= x_i^2 \\ b_3(x_i) &= x_i^3 \\ b_{k+3}(x_i) &= (x_i - \xi_k)_+^3, \quad k = 1, \ldots, K \end{aligned}$$

where

$$(x_i - \xi_k)_+^3 = \begin{cases} (x_i - \xi_k)^3 & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$$



A cubic spline with K knots has K+4 degrees of freedom.

7.4 Regression Splines 样条回归
7.4.3 The Spline Basis Representation

- **线性样条回归模型**Linear Splines

A linear spline with knots at ξ_k, $k = 1, \ldots, K$ is a piecewise linear polynomial continuous at each knot.

We can represent this model as

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+1} b_{K+1}(x_i) + \epsilon_i,$$

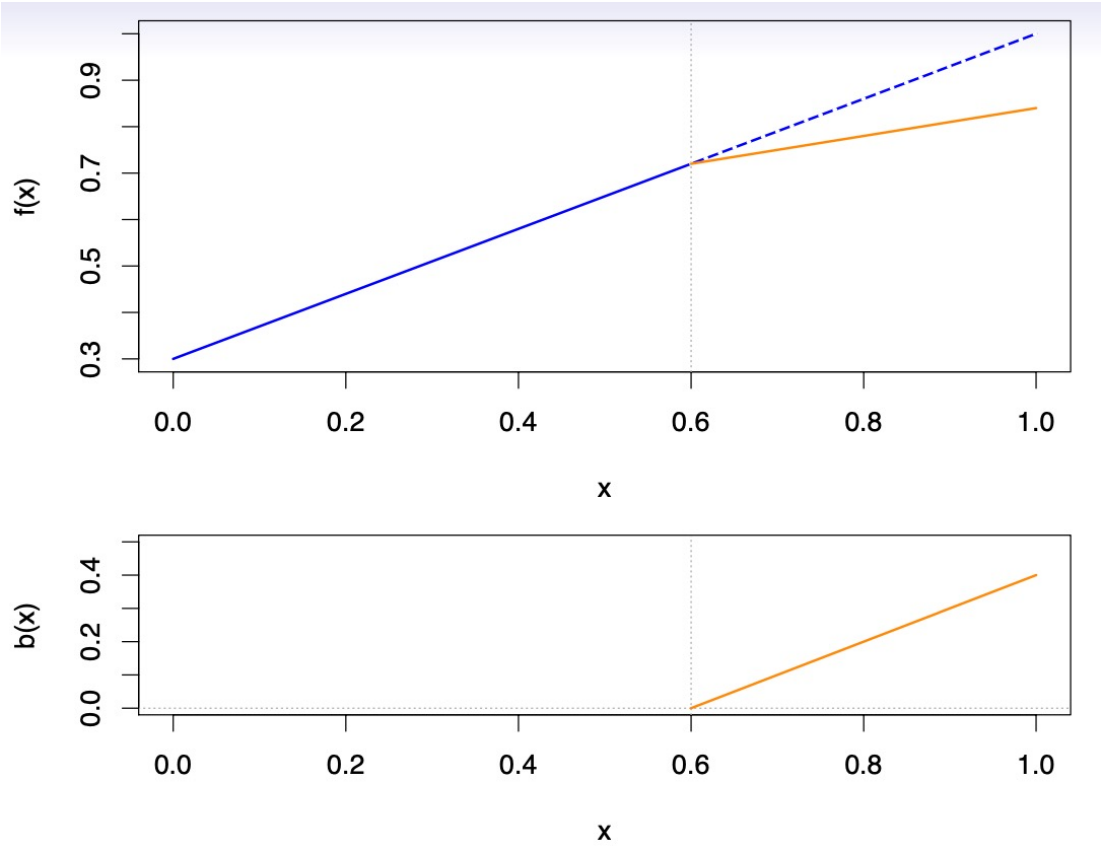
where the b_k are *basis functions*.

$$\begin{aligned} b_1(x_i) &= x_i \\ b_{k+1}(x_i) &= (x_i - \xi_k)_+, \quad k = 1, \ldots, K \end{aligned}$$

Here the $()_+$ means *positive part*; i.e.

$$(x_i - \xi_k)_+ = \begin{cases} x_i - \xi_k & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$$

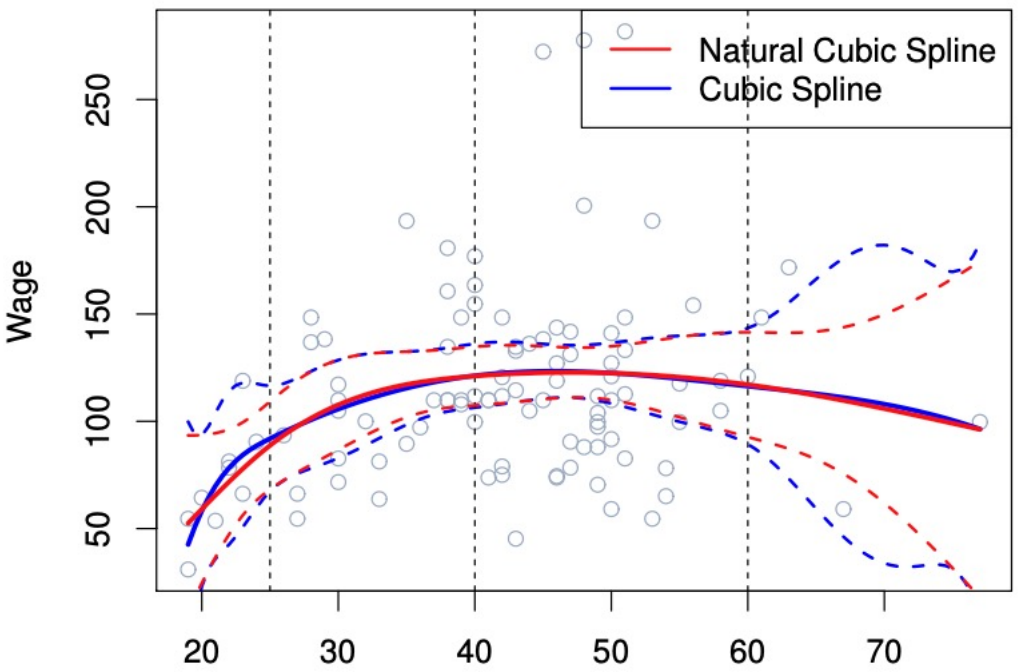
A natural spline with K knots has K degrees of freedom.



7.4 Regression Splines 样条回归
7.4.3 The Spline Basis Representation

- **自然三次样条回归**Natural Cubic Splines

- 为了解决B样条回归的边界预测误差大的问题，统计学家们又在B样条回归增加约束，这种回归成为自然样条回归，R中对应函数是ns()；
- 通过上图的对比，红色为自然样条回归，蓝色为B样条回归，红色的虚线间距比蓝色的虚线间距窄，尤其是在age的两端，表明自然回归在age的边界处得到的结果更加稳健。



7.4 Regression Splines 样条回归

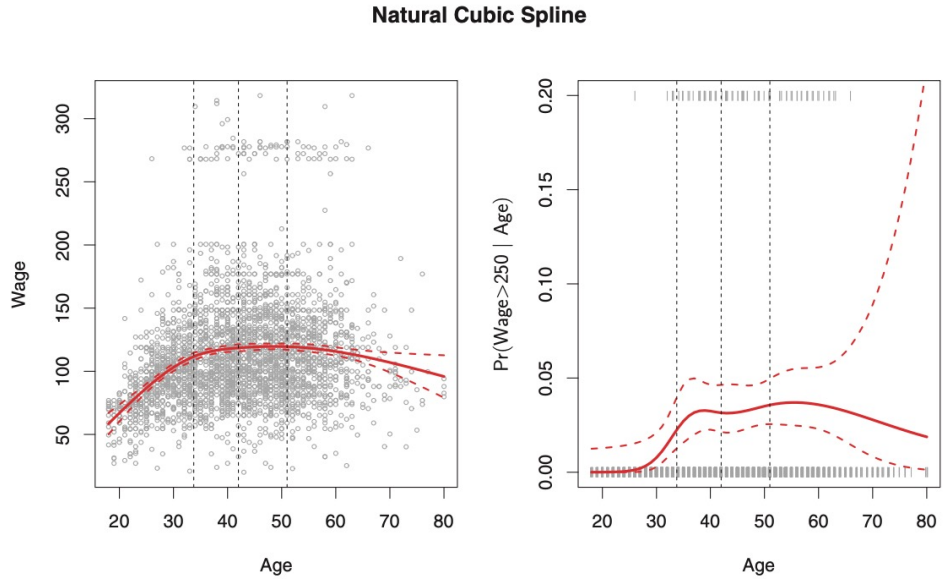
7.4.4 Choosing the Number and Locations of the Knots

2）节点数量位置未知的情况

确定结点的个数与位置：

- （1）一种方法是构建连续型等距测量：在变化相对稳定的区间设置尽量少的结点，在变化相对快速的地方设置尽量多的结点。尽量让每个结点区间内的变量趋于均匀分布。
- （2）另一种方法是设置自由度，根据算法自动跑出最优的结点位置。自由度的个数可以结合交叉验证来验证。

- As in Figure 7.4, we have fit a natural cubic spline with three knots, except this time the knot locations were chosen automatically as the 25th, 50th, and 75th percentiles natural spline of age.



7.4 Regression Splines 样条回归

7.4.4 Choosing the Number and Locations of the Knots

2）节点数量位置未知的情况

- Decide the number of knots:
  - Cross-validation: find the K which gives the smallest RSS.

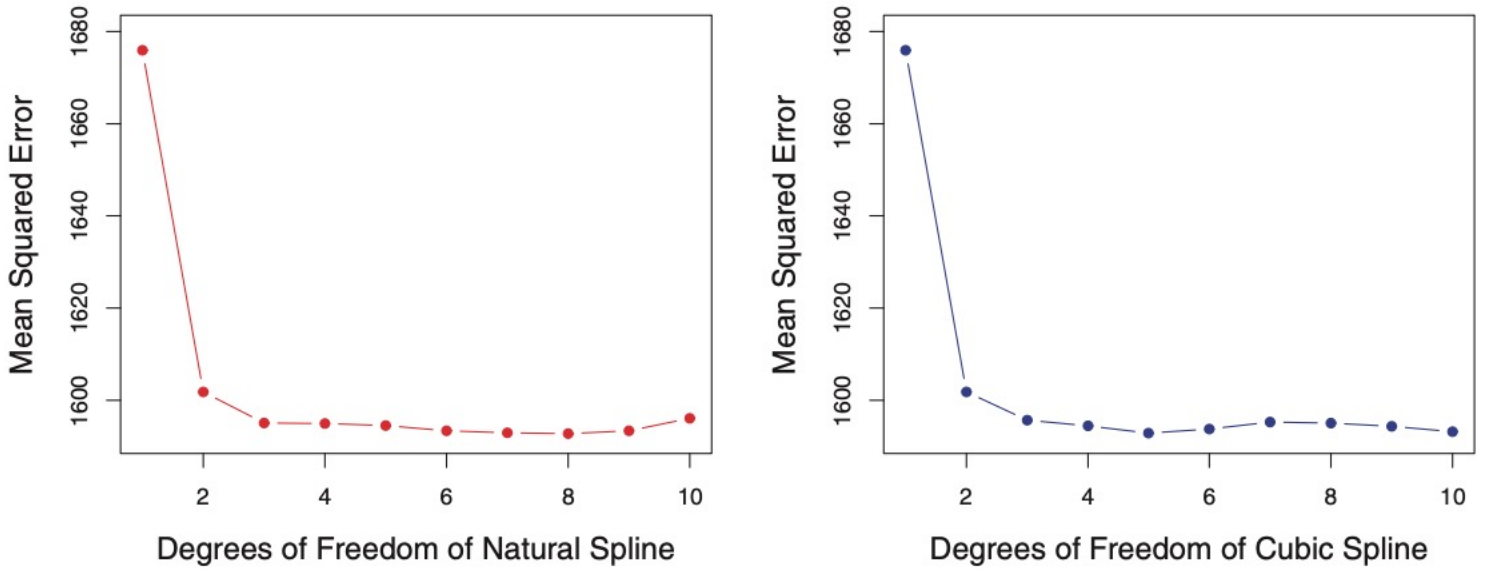


FIGURE 7.6. *Ten-fold cross-validated mean squared errors for selecting the degrees of freedom when fitting splines to the* `Wage` *data. The response is* `wage` *and the predictor* `age`. Left: *A natural cubic spline.* Right: *A cubic spline.*

7.4 Regression Splines 样条回归

7.4.5 Comparison to Polynomial Regression样条回归与多项式回归的比较

- **多项式回归**Polynomials :
  - 多项式回归则可能需要更多的次数(e.g. x^{15})
- **样条回归**Splines :
  - 由于样条有结点的帮助，所以在变动很大的数据背景下，仍然可以保证多项式的次数较小。
  - Keep the degree fixed, introduce flexibility by increasing the number of knots

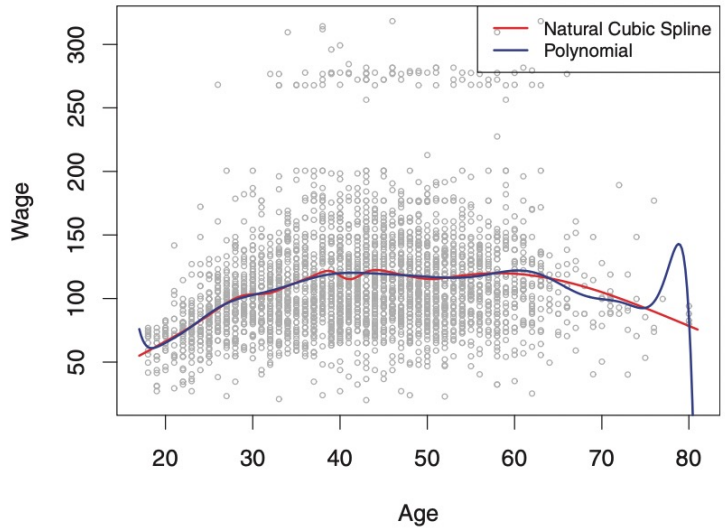


FIGURE 7.7. *On the* `Wage` *data set, a natural cubic spline with 15 degrees of freedom is compared to a degree-15 polynomial. Polynomials can show wild behavior, especially near the tails.*

7.5 Smoothing Splines 光滑样条

7.5.1 An Overview of Smoothing Splines

- 避免多项式样条估计的节点选择问题对光滑程度造成过多主观性影响，我们采用Ch6正则化手段，在自然三次样条中，引入光滑参数λ对拟合的粗糙度变化进行惩罚。即假设我们样条函数为g(x), 寻找一个光滑函数使得残差平方和最小

- Goal:
  - Find some function g(x) that makes RSS small: $\boldsymbol{RSS = \sum_{i=1}^{n}(y_i - g(x_i))^2}$
  - Guarantee g(x) is also smooth.
    - Find the function g that minimizes: (加入光滑因子后的均方误差)

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

how g fit the data well

penalizes the variability in g

7.5 Smoothing Splines 平滑样条

7.5.1 An Overview of Smoothing Splines

- 避免多项式样条估计的节点选择问题对光滑程度造成过多主观性影响，我们采用Ch6正则化手段，在自然三次样条中，引入光滑参数λ对拟合的粗糙度变化进行惩罚。即假设我们样条函数为g(x), 寻找一个光滑函数使得残差平方和最小

- Goal:
  - Find some function g(x) that makes RSS small: $\boldsymbol{RSS} = \sum_{i=1}^{n}(\boldsymbol{y_i} - \boldsymbol{g(x_i)})^2$
  - Guarantee g(x) is also smooth.
    - Find the function g that minimizes: (加入光滑因子后的均方误差)

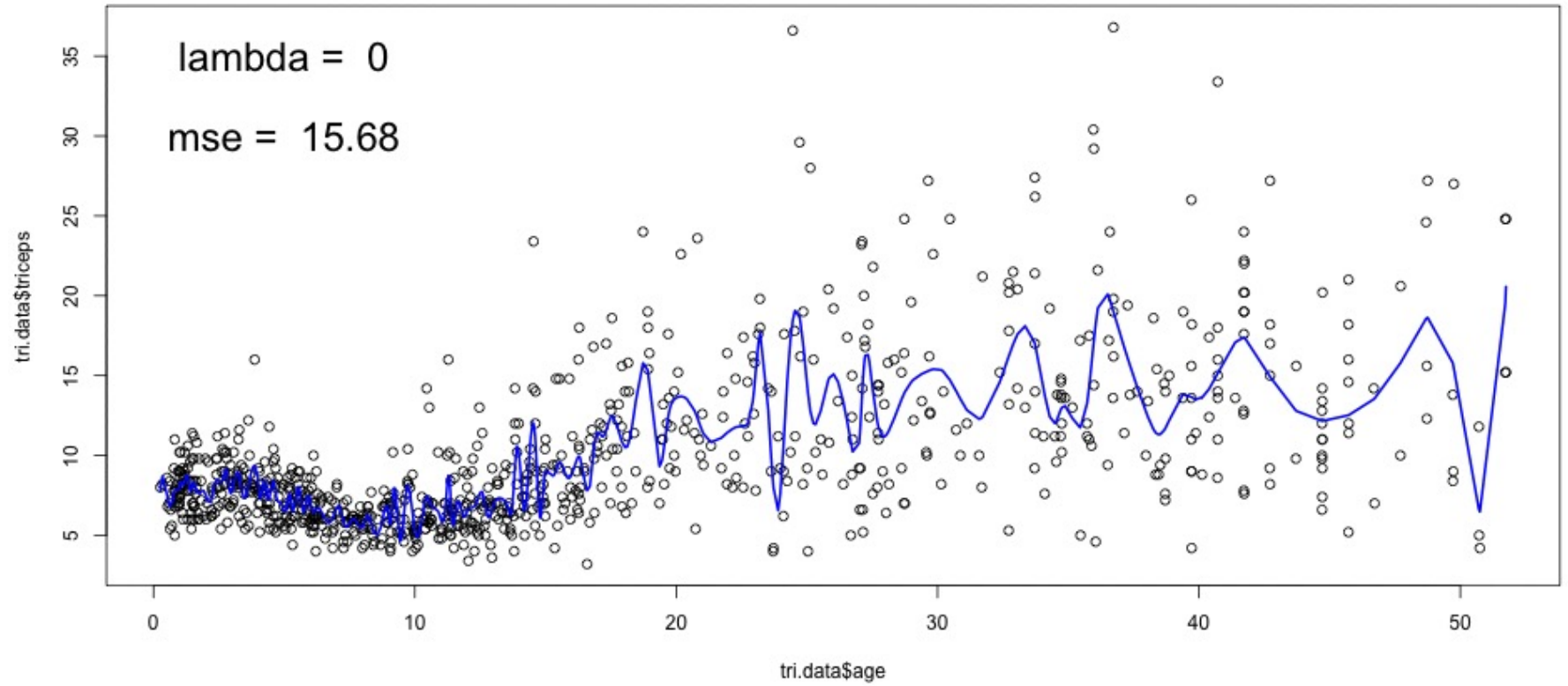
$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

how g fit the data well　　penalizes the variability in g

    - Where λ is a nonnegative tuning parameter.(bias-variance tradeoff)
      - λ =0, penalty term no effect, function g will be jumpy,
      - $\lambda \to \infty$ g will be smooth,
      - 超参λ ，用来衡量惩罚项的重要性占比，一般用n折交叉验证或留一交叉验证法来确定。
- ”Effective Degrees of Freedom”
  - The number of free parameters is an inappropriate measure of model complexity due to λ

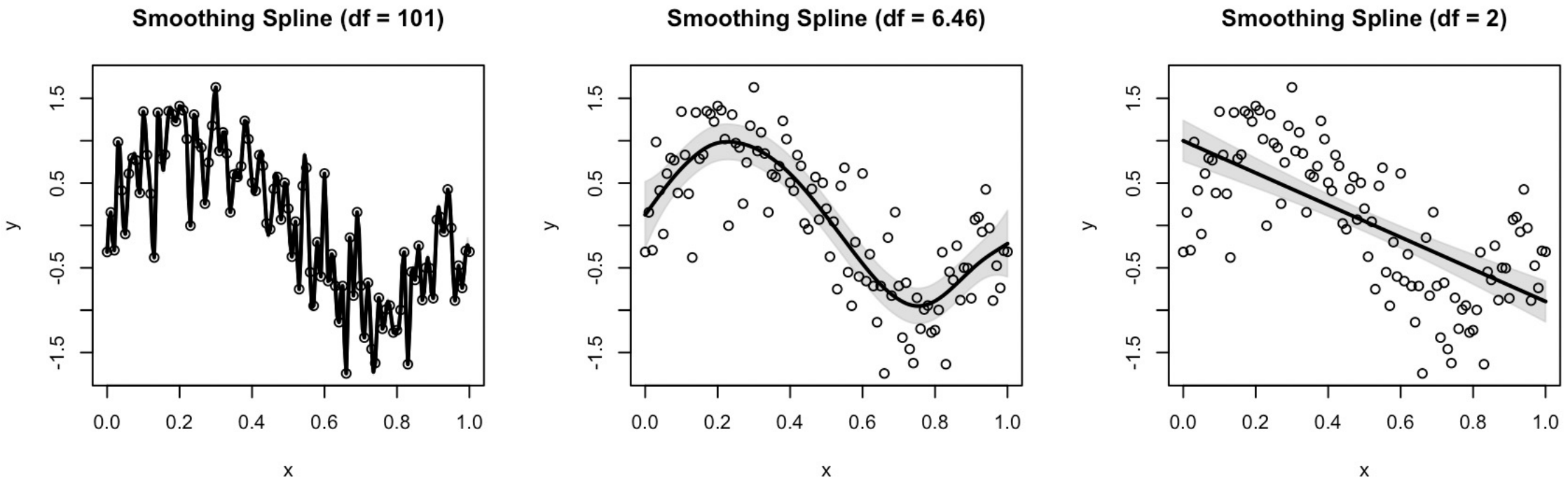
- 一阶导：该点曲线的斜率；
- 二阶导 ：斜率的变化率，即，惩罚项即表示为函数曲线在该点的平滑性。
- 二阶导的积分为：对区间t内，二阶导数累积的变化情况，因此可以用来衡量该段区间整体的平滑性。

7.5 Smoothing Splines 平滑样条
7.5.1 An Overview of Smoothing Splines



- Lambda与MSE的关系

7.5 Smoothing Splines 平滑样条
7.5.1 An Overview of Smoothing Splines



- ”Effective Degrees of Freedom”
  - The number of free parameters is an inappropriate measure of model complexity due to λ

7.5 Smoothing Splines 平滑样条

7.5.2 Choosing the Smoothing Parameter λ

- We can specify *df* rather than λ!

  In R: `smooth.spline(age, wage, df = 10)`

- The leave-one-out (LOO) cross-validated error is given by

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

  In R: `smooth.spline(age, wage)`

平滑矩阵$S_{\lambda} = \begin{bmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nn} \end{bmatrix}$

$$\hat{y} = Sy$$

7.5 Smoothing Splines 平滑样条
7.5.2 Choosing the Smoothing Parameter λ

Delete the i-th row and column,
renormalize the rows to sum to one

$$\hat{y}_{-i} = \frac{1}{1 - s_{ii}} \sum_{\substack{j=1 \\ j \neq i}}^{n} s_{ij} y_j$$

$$\hat{y}_i = \sum_{j=1}^{n} s_{ij} y_j$$

$$\hat{y}_{-i} = \sum_{\substack{j=1 \\ j \neq i}}^{n} s_{ij} y_j + s_{ii} \hat{y}_{-i}$$

$$= \sum_{j=1}^{n} s_{ij} y_j + s_{ii} \hat{y}_{-i} - s_{ii} y_i$$

$$= \hat{y}_i + s_{ii} \hat{y}_{-i} - s_{ii} y_i$$

$$y_i - \hat{y}_{-i} = y_i - \hat{y}_i + s_{ii}(y_i - \hat{y}_{-i})$$

$$y_i - \hat{y}_{-i} = \frac{y_i - \hat{y}_i}{1 - s_{ii}}$$

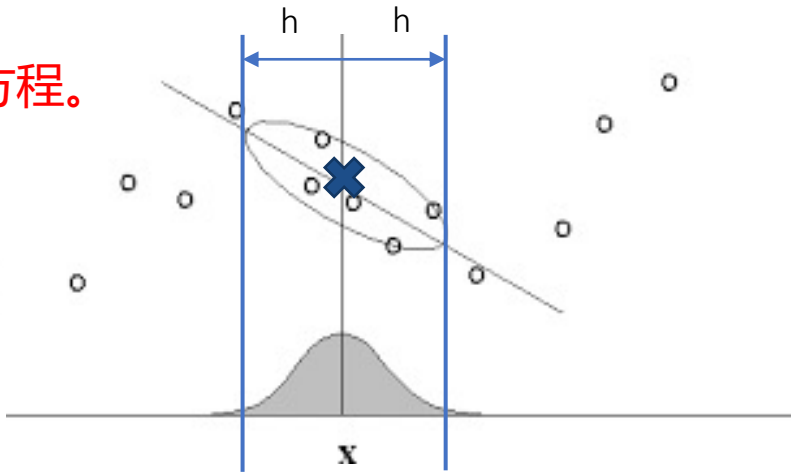
$$\mathrm{cv}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_{\lambda,i}}{1 - s_{\lambda,ii}} \right)^2$$

PART 2

局部平滑法
Local Smoothing

7.6 Local Regresssion 局部回归

- 对x邻域内的样本做加权平均得到y的sample mean，用来估计未知的回归方程。
- 带宽的选择对估计的影响非常大
  - Rule of thumb、 Pulg-in、最小二乘交叉验证法



Kernel Probability Density Estimation (Rosenblatt, 1956; Parzen, 1960)

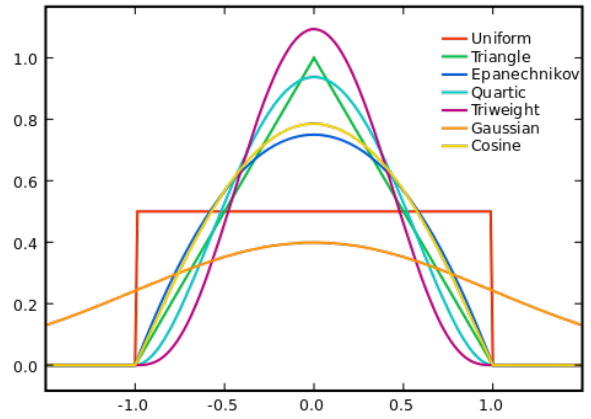
- Suppose $\{X_i\}_{i=1}^n$ is an IID random sample from an unknown **probability density function** $f(x)$ with support $[a, b]$, where $a < b$.
- Then a kernel estimator for $f(x)$ at a given point $x \in [a, b]$ is

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x),$$

where

$$K_h(X_i - x) = \frac{1}{h}K\left(\frac{X_i - x}{h}\right),$$

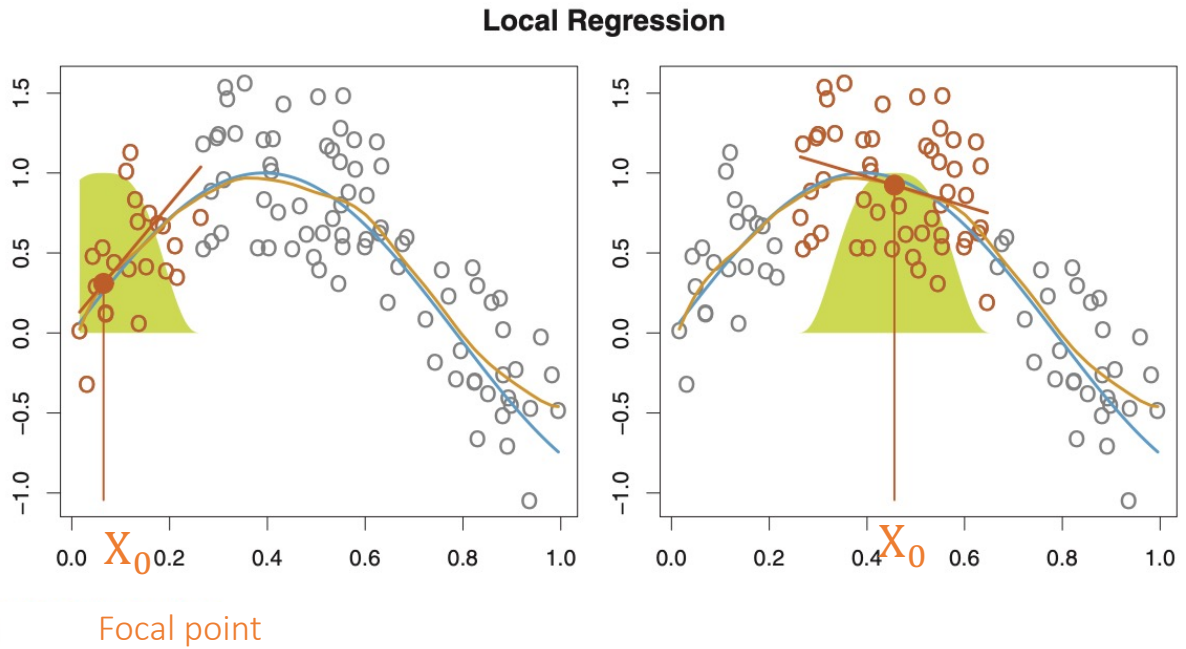
$K: [-1, 1] \to \mathbb{R}^+$ is a kernel function, and $h = h(n) \to 0$ as $n \to \infty$ is a bandwidth.



7.6 Local Regresssion 局部回归

- 局部回归是另一种拟合非线性模型的方法，它对于窗内的每个点都予以不同的权重 ，有效降低异常值的影响。

- 为了实现局部回归，有许多重要步骤需要确定：
  - - 加权函数K
  - - 局部回归的模型：线性 / 二次 / 还是常数等，模型估计方法：加权最小二乘法(WLS)
  - 对“局部”的定义-最近邻bin，如何确定x_0点邻近范围s。范围s的值越小，每次拟合的数据区间越小，并且到下一个拟合点到距离也越小。
    - 交叉验证的方法来确定s
    - 或者人为设定一个值
  - Goal: “总残差平方和”最小



Focal point

总结：几乎所有非参数平滑估计量均可以表示为一个 Local Weighted Sample Mean of $\{Y_i\}_{i=1}^n$

- 不同的非参数方法体现在确定不同邻域（neighbors）和不同权重（weights）

7.6 Local Regresssion 局部回归

- Nadaraya-Watson **核估计/局部常数核回归方法**
  - 用常数r来对回归函数$\hat{r}(x)$进行估计

Kernel Regression Estimation (Nadaraya, 1964; Watson, 1964)

- The Nadaraya-Watson estimator is a local weighted sample mean

$$\hat{r}(x) = \arg\min_{r} \sum_{i=1}^{n} (Y_i - r)^2 K_h(x - X_i)$$

- FOC

$$\hat{r}(x) = \frac{\hat{m}(x)}{\hat{f}(x)} = \sum_{i=1}^{n} \hat{w}_i(x) Y_i$$

其中权重

$$\hat{w}_i(x) = \frac{K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)},$$

7.6 Local Regresssion 局部回归

- **局部多项式回归方法**
  - 用多项式来对回归函数$\hat{r}(x)$进行估计

  - The idea of using a **local polynomial**: Katkovnik (1979, 1983, 1985) and Lejeune (1985):

$$\min_{\alpha} \sum_{i=1}^{n} (Y_i - Z_i'\alpha)^2 K_h(x - X_i)$$

where $Z_i = (1, X_i - x, \dots, (X_i - x)^p)' = (p+1) \times 1$

$$\alpha = (\alpha_0, \alpha_1, \dots, \alpha_p)' = (p+1) \times 1$$

  - when $p = 0$, it becomes a local constant smoothing.
  - when $p = 1$, it becomes a local linear smoothing.

7.6 Local Regresssion 局部回归

- **局部多项式回归方法**
  - 用多项式来对回归函数$\hat{r}(x)$进行估计
  - Locally WLS: $\hat{\alpha} = (Z'WZ)^{-1}Z'WY$
    - $p \geq 1,\ Bias = O(h^2)\ \forall\ x \in [a,b]$
    - $\hat{\alpha}$ has an equivalent kernel representation:

$$\hat{\alpha}_v = \sum_{i=1}^{n} \hat{w}_v\left(\frac{X_i - x}{n}\right) Y_i,\ \hat{w}_v(u) = \frac{\tilde{K}_v(u)}{Thf(x)}[1 + o_p(1)]$$

其中 $\tilde{K}_v(u)$ 是一个核函数，而 $f(x)$ 是 X_i 的概率密度函数。

7.6 Local Regresssion 局部回归

- 局部回归是另一种拟合非线性模型的方法，它对于窗内的每个点都予以不同的权重 ，有效降低异常值的影响。

1）最简单的局部模型

$$\hat{f}(x_0) = \frac{\sum_i y_i I(|x_i - x_0| < h)}{\sum_i I(|x_i - x_0| < h)}$$

- 缺点：估计结果非连续

2）Nadaraya-Watson 核估计(NW)

$$\hat{f}(x_0) = \frac{\sum_i y_i K_h(x_i, x_0)}{\sum_i K_h(x_i, x_0)},$$

- 引用核估计方法，将local average模型进行改造，当K是连续时，f即为连续的。
- 其局部性参数h可以通过交叉验证的方式计算得到。
- 不过当x不是均匀分布的时候，该方法将产生偏倚

3）LOWESS

lowess 算法：

1）给定数据集$D = (x_i, y_i)$， 核函数K(),

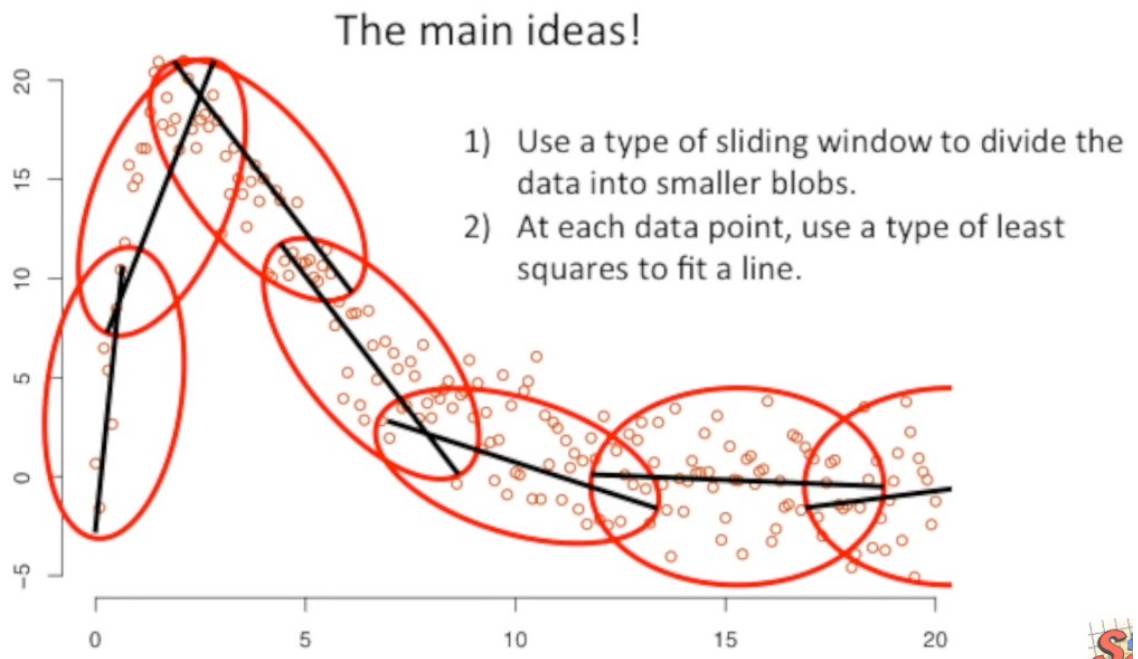
2）拟合局部回归$\hat{w}(x) = argmin \sum_{i=1}^{n} K(x, x_i) * (w^T x_i - y_i)^2$

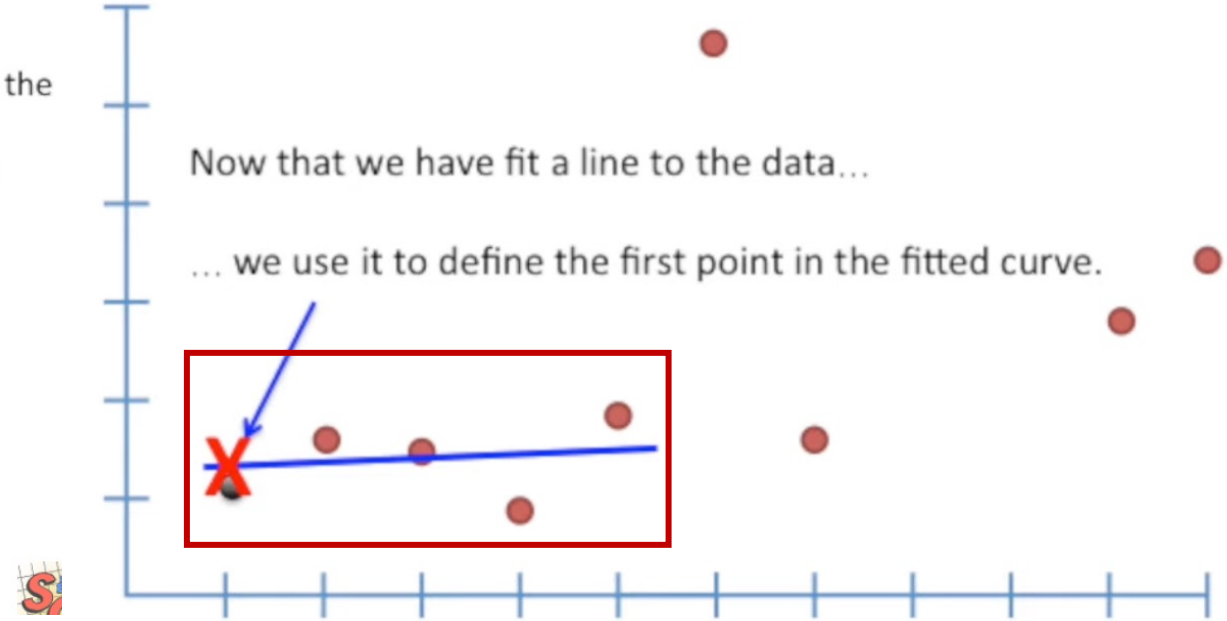
3）得到拟合结果$\hat{w}(x)^T x$

- 局部加权回归散点平滑法（ locally weighted scatterplot smoothing ）
- 结合参数和非参数模型的方法，通过对局部进行线性/二次拟合，来解决NW的缺点
- New-new

7.6 Local Regresssion 局部回归

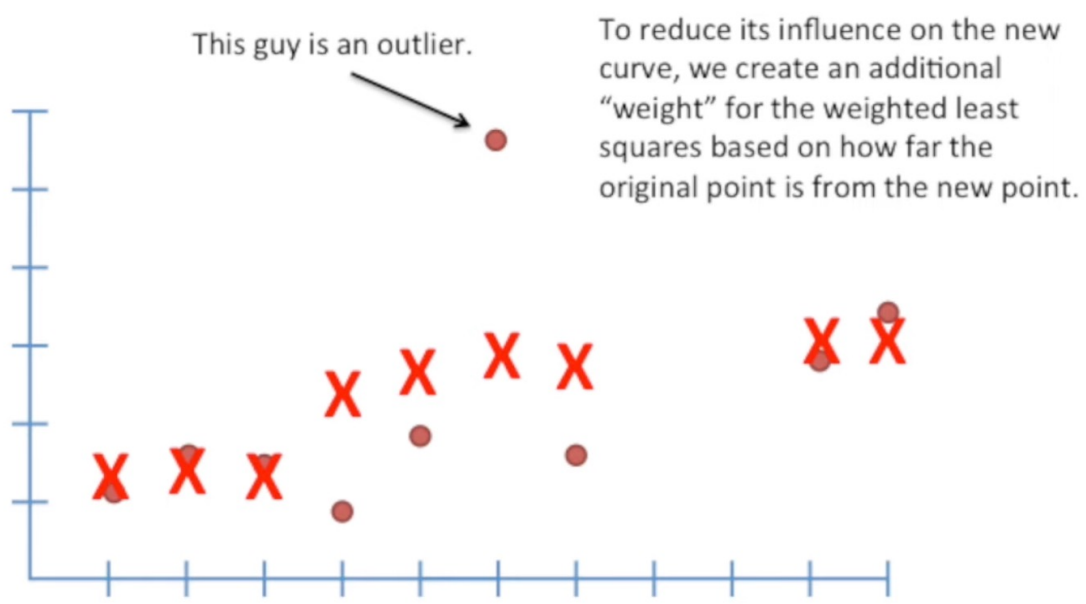
- LOWESS

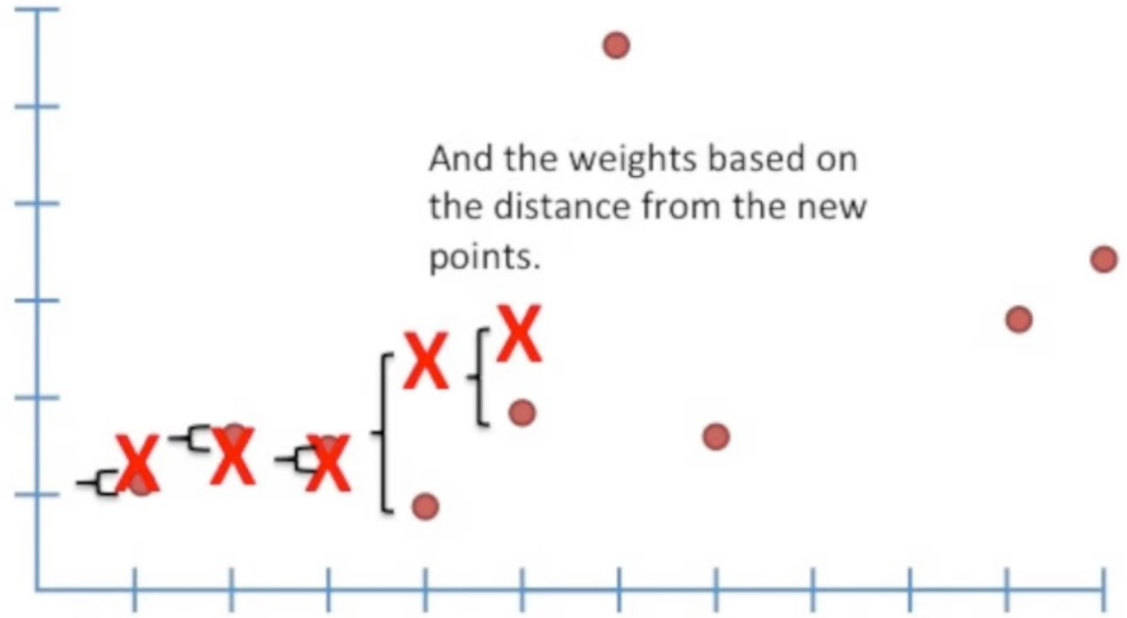




7.6 Local Regresssion 局部回归

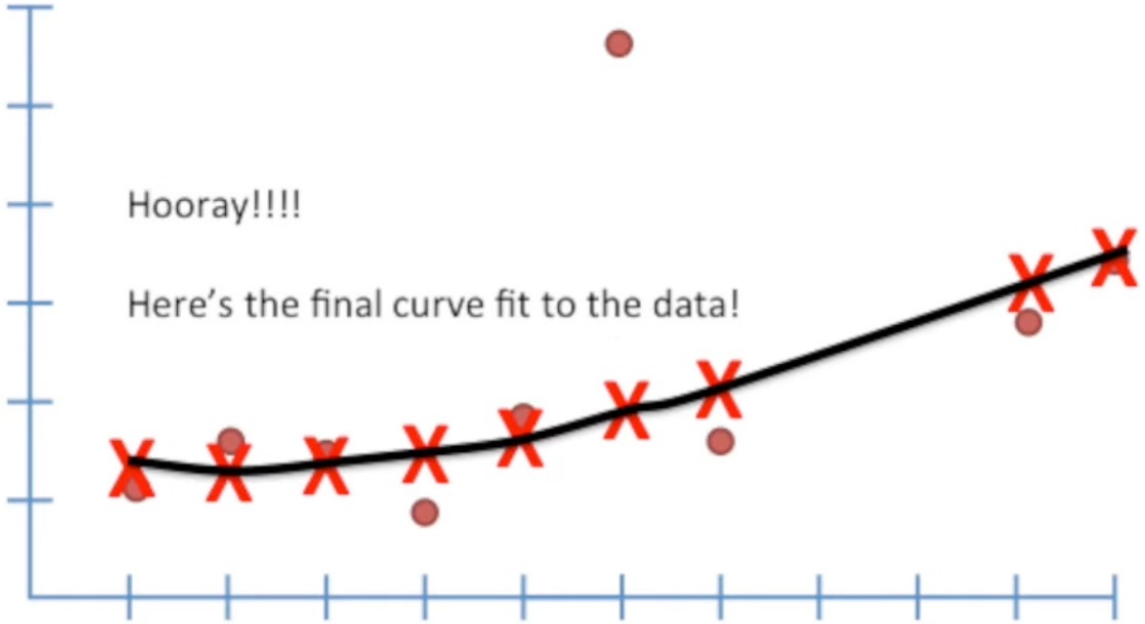
- LOWESS





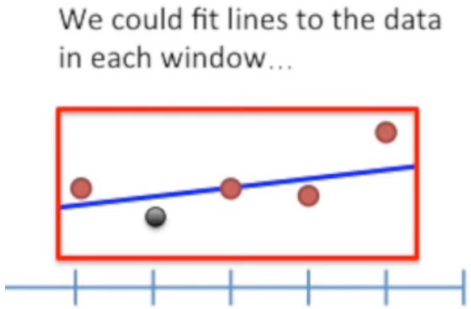
7.6 Local Regresssion 局部回归

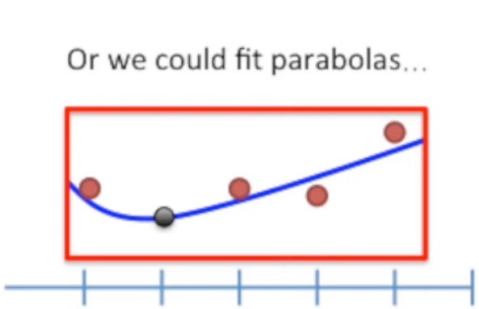
- LOWESS

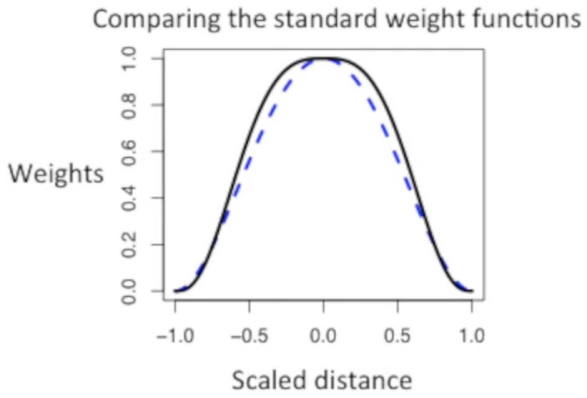


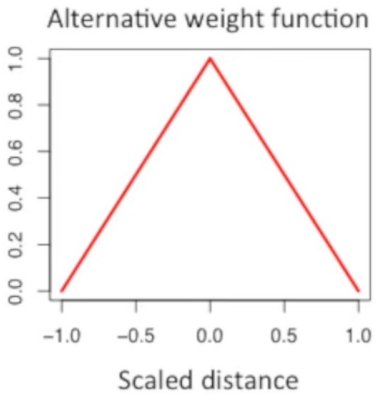
7.6 Local Regresssion 局部回归

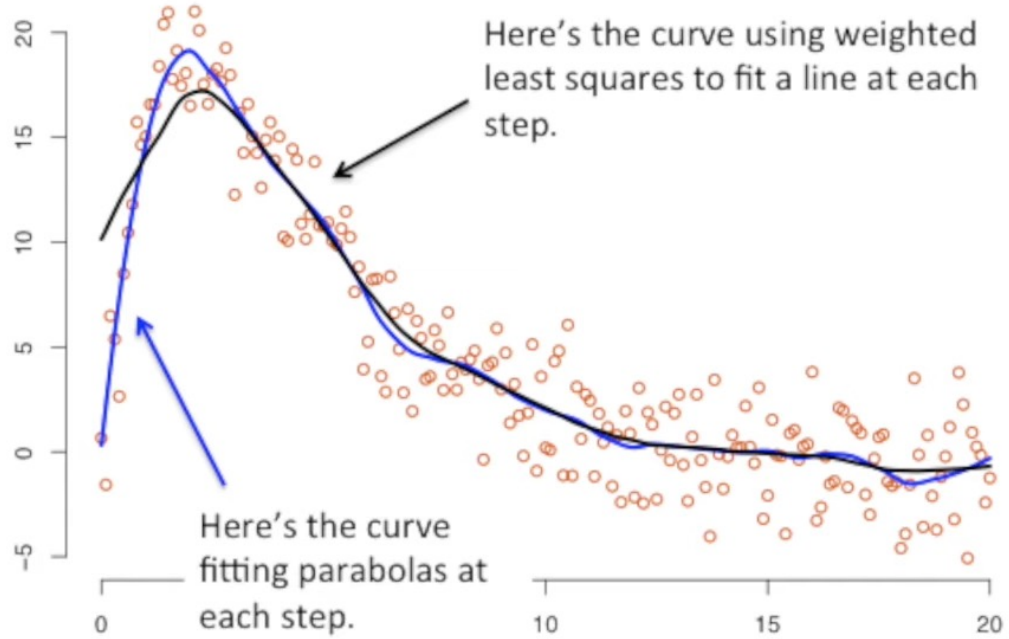
- LOWESS











7.7 Generalized Additive Model 广义可加模型
7.7.1 GAMs for Regression Problems

- 前面，我们已经介绍了一些方法，作为简单线性模型的扩展。这些方法可以归纳为广义相加模型(GAMs)的框架里，形如

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

- 之所以是“相加性”的，是因为我们对于每个变量X_j 都单独计算f_j, f_j,可以是任意形式的函数，最后统一加起来用来预测Y
- GAMs的优缺点：
  - 可以引入非线性函数f_j,(可用样条/局部多项式回归)
  - 非线性可能使得对Y预测地更准确
  - 因为是“相加性”的，所以线性模型的假设检验方法仍然可以使用
  - 因为是“相加性的”，GAMs中可能会缺失重要的交互作用$X_j \times X_k$，只能通过手动添加交互项来补充

$$\texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \epsilon$$

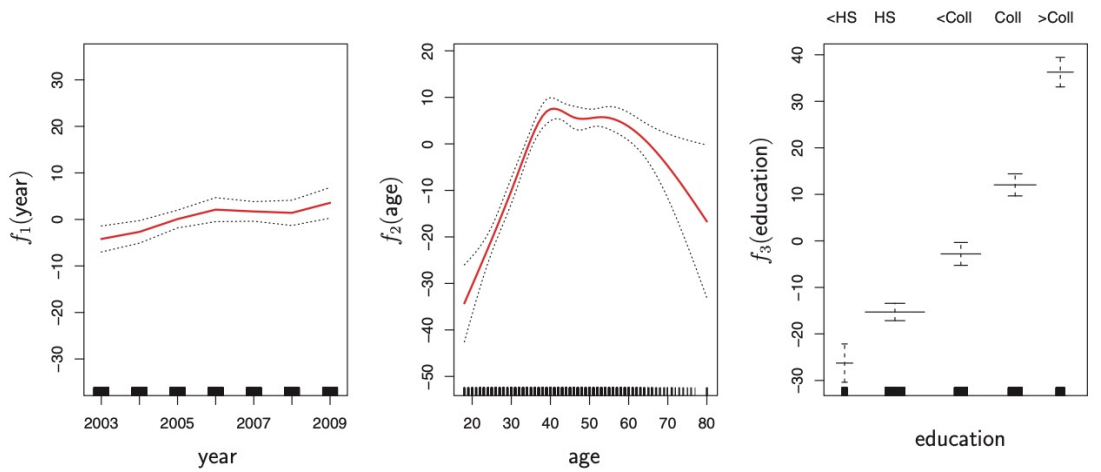


FIGURE 7.11. *For the* Wage *data, plots of the relationship between each feature and the response,* wage*, in the fitted model (7.16). Each plot displays the fitted function and pointwise standard errors. The first two functions are natural splines in* year *and* age*, with four and five degrees of freedom, respectively. The third function is a step function, fit to the qualitative variable* education.

7.7 Generalized Additive Model 广义可加模型
7.7.2 GAMs for Classification Problems

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p).$$

GAMs can also be used in situations where Y is qualitative

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 \times \texttt{year} + f_2(\texttt{age}) + f_3(\texttt{education}),$$

$$p(X) = \Pr(\texttt{wage} > 250 | \texttt{year}, \texttt{age}, \texttt{education}).$$

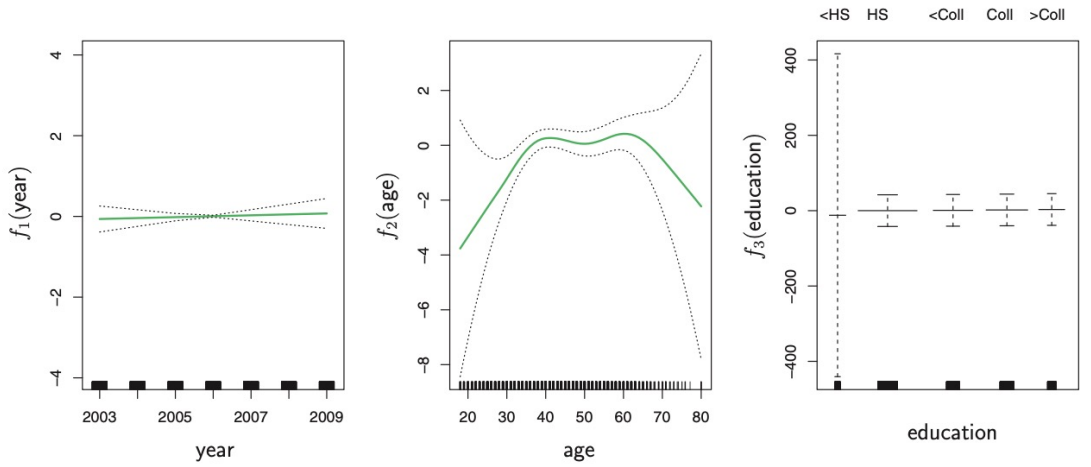


FIGURE 7.13. *For the* `Wage` *data, the logistic regression GAM given in (7.19) is fit to the binary response* `I(wage>250)`*. Each plot displays the fitted function and pointwise standard errors. The first function is linear in* `year`*, the second function a smoothing spline with five degrees of freedom in* `age`*, and the third a step function for* `education`*. There are very wide standard errors for the first level* `<HS` *of* `education`*.*

Summary-非参数方法

- Motivation: 如何表征复杂数据的特征？
- Idea:不预先设定模型的具体形式和误差分布，让数据告诉真实的函数形式。
  - (1)**分段处理**
    - 等距/等量/实际意义/
  - (2)**局部平滑**-(无权/带权)
  - (3)**全局平滑**
- Realization:**基函数**
  - 从一般到特殊
    - 多项式回归
    - 阶跃函数
  - 分段多项式样条估计法
    - 约束条件
    - 确定元素：1、多项式阶次；2、分段结点数量位置
  - 光滑样条估计法 min”Loss + Penalty”
  - 局部多项式估计法
  - 核回归估计法
- Model: GAM

Summary-非参数方法

总结：非参数分析的基本假设、思想与方法

（1）Key Assumption: Data generating process (DGP) is an **unknown stochastic process**.

（2）**Model-free**：用于未知函数形式、非线性情景

（3）Criterion: Mean Squared Error (MSE)

（4）关键是如何控制平滑程度（即如何选取平滑参数值）？

Trade-off between variance and squared bias

Summary-非参数方法

总结：非参数分析的基本假设、思想与方法

(5) 收敛速度（Convergence Rate）比较慢，需要比较大的样本（n）

✓ For kernel multivariate density estimation, the optimal convergence rate

$$MSE\left[\hat{f}(x), f(x)\right] \propto n^{-\frac{4}{4d+1}},$$

where d is the dimension of X_i.

(6) 解释变量 X_i 事先给定；存在**维度灾难**（Curse of Dimensionality）

(7) 非参数模型的可解释性：系数一般没有经济意义的解释

Summary-非参数方法

	优点	缺点
多项式样条估计	• 全局估计方法； • 多项式技巧的有用推广； • 适应于样本较多的异常值情况	• 节点选择，有较大主观性，若过多会造成过拟合； • 全局则需考虑边界修正问题
光滑样条估计	• 对多项式样条的修正方法，使得节点选择没有光滑参数λ的选取重要； • 计算有效 • Kernel变体	
局部多项式估计	• 局部估计，无需考虑边界修正； • 可用于异常值数据 • Kernel变体	• 受到异常值影响大-修正见LOWESE/Robust Regression

感谢聆听！请大家批评指正!

THANK YOU FOR YOUR CRITICISM

presenter: Fei Yang 2022/01/13