非参数密度估计与非参数回归

1 非参数密度估计

密度估计的实质是用给定的样本$X_1, \cdots, X_n \sim f(x)$,估计$\hat{f}(x)$, 本质上是未知参数个数为无穷情况下的估计.

密度估计方法

- 直方图
- 核方法
- K近邻

核方法的带宽选择

- 大拇指法则(假定$f(x)$为正态分布, 选取Gauss核)
- 交叉验证

1.1 直方图

构造直方图估计,

步骤一:

选取起点x_0, 正数h, 把直线分为一些形如$[x_0 + mh, x_0 + (m+1)h)$的区间, 记

$$B_j = [x_0 + (j-1)h, x_0 + jh)$$

假设格子$B_1, B_2, \cdots, B_k$覆盖了所有的数据点$x_1, \cdots, x_n$, 即

$$x_0 \le x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)} < x_0 + kh$$

步骤二:

计数n_j个数据点落入B_j的频数. 即

$$n_j = \sharp\{i : X_i \in B_i, i = 1, \cdots, n\}$$

步骤三:

取

$$\hat{f}_j(x) = \frac{n_j}{nh}$$

为格子B_j的直方图估计. 直方图的格子B_j, 高度为$\hat{f}_j$, 宽度为h.

正式的, 直方图的密度估计为,

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j=1}^{k} I(X_i \in B_j) I(x \in B_j), \forall x \in \mathbb{R}$$

证明. 令m_j为格子B_j的中心, 即

$$m_j = x_0 + (j - \frac{1}{2})h, \quad B_j = [m_j - \frac{h}{2}, m_j + \frac{h}{2})$$

故

$$\begin{aligned} P(X \in B_j) &= P(X \in [m_j - \frac{h}{2}, m_j + \frac{h}{2})) \\ &= \int_{m_j - \frac{h}{2}}^{m_j + \frac{h}{2}} dF(x) \\ &\approx f(m_j)h \end{aligned}$$

另外基于数据$X_1, \cdots, X_n$的大数定律

$$P(X \in B_j) = \frac{\sharp\{X_i \in B_j; i = 1, \cdots, n\}}{n} = \frac{n_j}{n}$$

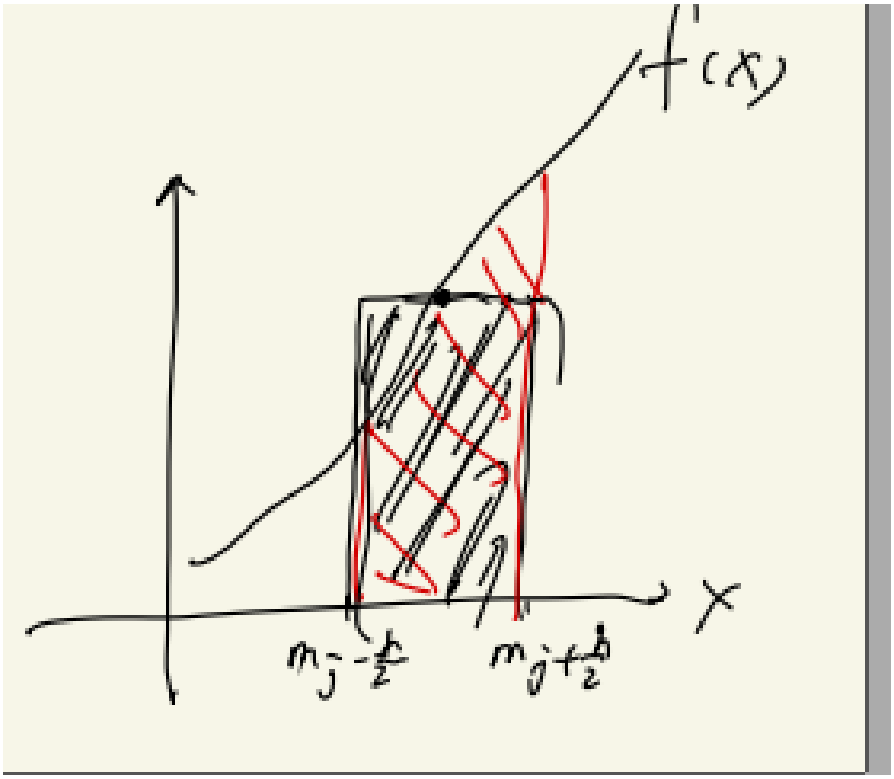
综合上述两式即可得出密度估计的表达式. 即,

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j=1}^{k} I(X_i \in B_j) I(x \in B_j), \forall x \in \mathbb{R}$$

□

1.1.1 对偏倚的估计

$Bias(\hat{f}(x)) = E(\hat{f}(x)) - f(x), \forall x \in \mathbb{R}.$



由于$\mathbb{E}(I(X_i \in B_j)) = P(X_i \in B_j) = \int_{B_j} f(x)\,dx$, 故对于任给的$x \in B_j$, $\mathbb{E}(\hat{f}(x)) = \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}(I(X_i \in B_j)) = \frac{1}{h}\int_{B_j} f(x)\,dx$. 因此

$$\begin{aligned} Bias &= \mathbb{E}_{\hat{f}}(\hat{f}(x)) - f(x) \\ &= \frac{1}{h}\int_{B_j} f(u)\,du - f(x) \\ &= \frac{1}{h}\int_{B_j} f(u) - f(x)\,du \end{aligned}$$

由Taylor展式,

$$f(u) - f(x) \approx f'(m_j)(u - x) \approx f'(m_j)(m_j - x)$$

可以将积分项化为与n无关, 因此

$$\begin{aligned} Bias &= \frac{1}{h}\int_{B_j} f(u) - f(x)\,du \\ &= \frac{1}{h}\int_{B_j} f'(m_j)(m_j - x)\,du \\ &= \frac{h}{h}f'(m_j)(m_j - x) \end{aligned}$$

即, $Bias = f'(m_j)(m_j - x)$, 因此当$h \searrow 0$, $(m_j - x) \searrow 0$时, $Bias \searrow 0$, 也就是说每一个区间中心的密度估计较为精准, 而边缘部分结果较差.

1.1.2 对方差的估计

对任给的$x \in B_j$

$$Var(\hat{f}(x)) = Var\left(\frac{1}{nh}\sum_{i=1}^{n} I(X_i \in B_j)\right)$$

由于X_i个体独立同分布, 因此, $Var(\hat{f}(x)) = \frac{1}{n^2h^2}\sum_{i=1}^{n} Var(I(X_i \in B_j))$.

注意到

$$I(X_i \in B_j) = \begin{cases} 1, x_i \in B_j \\ 0, x_i \notin B_j \end{cases} \sim Bernoulli(p)$$

其中$p = \mathbb{E}(I(X_i \in B_j)) = P(X_i \in B_j) = \int_{B_j} f(u)du \approx f(x)h$, 其中$x \in B_j$. 因此

$$\begin{aligned} Var(\hat{f}(x)) &= \frac{1}{nh^2} Var\left(I(X_i \in B_j)\right) \\ &= \frac{1}{nh^2} pq \\ &\approx \frac{1}{nh^2} f(x)h\left[1 - hf(x)\right] \\ &\approx \frac{1}{nh} f(x)[1-0] \\ &\approx \frac{f(x)}{nh} \end{aligned}$$

事实上, 若记$p_j = \int_{B_j} f(u)du$, 有$Var(\hat{f}(x)) = \frac{p_j(1-p_j)}{nh^2}$, 因此可以看出, 当$h \searrow 0$时, $Var(\hat{f}(x)) \nearrow$, 这表明了偏倚方差分解中的偏倚方差权衡.

1.2 均方误差

为评价统计学习方法对某一数据集的效果, 需要用一些方法评价模型的预测结果与实际观测数据的一致性, 在回归中, 最常用的评价标准是均方误差(Mean Square Error), 其表达式如下所示,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

上式的计算数值使用训练数据计算出来的, 由于用训练数据计算出的MSE往往较高, 我们形象地称之为**训练均方误差(training MSE)**, 一般而言我们不关注在训练集上怎样, 而更关注测试集上的均方误差, 测试均方误差可以写作

$$AVE(\hat{f}(x_0) - y_0)^2$$

从数学形式上看, 假设对训练集中的观测值$\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$拟合统计学习模型, 输出估计函数$\hat{f}$. 于是可以计算出, $\hat{f}(x_1), \cdots, \hat{f}(x_n)$, 这些值可能与$y_i$非常接近, 但是我们在应用中注重未参与建模的点$(x_0, y_0)$.

上课介绍的期望形式的MSE是指, 对于一个拟合的函数$\hat{f}(x)$

$$\begin{aligned}
MSE\left\{\hat{f}(x)\right\} &= E_f\left[\left(\hat{f}(x)-f(x)\right)^2\right] \\
&= E\left[\left(\hat{f}(x)-E\hat{f}(x)+E\hat{f}(x)-f(x)\right)^2\right] \\
&= E\left[\left(\hat{f}(x)-E\hat{f}(x)\right)^2\right]+E\left[\left(E\hat{f}(x)-f(x)\right)^2\right]+2E\left[\underbrace{\left(\hat{f}(x)-E\hat{f}(x)\right)\left(E\hat{f}(x)-f(x)\right)}_{\text{外面期望求进来为零 与}\hat{f}\text{无关, 相当于常数}}\right] \\
&= E\left[\left(\hat{f}(x)-E\hat{f}(x)\right)^2\right]+\left(E\hat{f}(x)-f(x)\right)^2 \\
&= Var(\hat{f}(x))+Bias(\hat{f}(x))^2
\end{aligned}$$

这里期望E_f是对大量训练数据重复估计f后, 又在x处代入不同的值所得的平均测试估计误差, 表示在真分布为f时的计算.

例 (直方图的MSE)**.** 对$x \in B_j$, 可以证明

$$\begin{aligned}
MSE\{\hat{f}(x)\} &\approx \frac{1}{nh}f(x)+\left[f'(m_j)(m_j-x)\right]^2 \\
&= \frac{1}{nh}f(x)+\left[f'\left(\left(j-\frac{1}{2}\right)h\right)\cdot\left(\left(j-\frac{1}{2}\right)h-x\right)\right]^2+o(h)+o(\frac{1}{nh})
\end{aligned}$$

上式中的两个无穷小量反映了直方图的偏倚方差权衡问题.

1.3 积分均方误差(MISE)

对密度拟合精度的估计更有实际意义应该是整体的度量而非逐点的度量. 我们因此引入积分均方误差的概念(Mean Integrated Squared Error). 形式上可以写成

$$MISE(T_n) = \int MSE(T_n(x))dx$$

具体的,

$$\begin{aligned}
MISE(\hat{f}) &= E\left[\int_{-\infty}^{\infty}\left(\hat{f}(x)-f(x)\right)^2 dx\right] \\
&= \int_{-\infty}^{\infty} E\left(\hat{f}(x)-f(x)\right)^2 dx \\
&= \int_{-\infty}^{\infty} MSE(\hat{f}(x))dx
\end{aligned}$$

例 (直方图的MISE)**.** 我们要选取最优的带宽h, 使得$MISE(\hat{f})$最小. 可以证明,

$$MISE(\hat{f}) \approx \underbrace{\int_{-\infty}^{\infty}\frac{1}{nh}f(x)dx}_{:=\mathcal{A}} + \underbrace{\sum_{j=1}^{k}\int_{B_j}\left[(j-\frac{1}{2})h-x\right]^2\cdot\left[f'\left((j-\frac{1}{2})h\right)^2\right]dx}_{:=\mathcal{B}}$$

对$\mathcal{A}$, 由密度函数的性质

$$\mathcal{A} = \frac{1}{nh}\int_{\mathbb{R}} f(x)dx = \frac{1}{nh}$$

对$\mathcal{B}$, 注意到$\left[f'\left((j-\frac{1}{2})h\right)^2\right]$与积分变元$x$无关, 因此提出来先对前面一项$\left[(j-\frac{1}{2})h-x\right]^2$积分, 我们有

$$\int_{B_j}\left[(j-\frac{1}{2})h-x\right]^2dx=\frac{1}{3}\left[x-\left(j-\frac{1}{2}\right)h\right]^3|_{(j-1)h}^{jh}=\frac{h^3}{12}$$

且还有

$$\left[f'\left((j-\frac{1}{2})h\right)\right]^2=\left[f'(m_j)\right]^2\approx\frac{1}{h}\int_{(j-1)h}^{jh}\left[f'(x)\right]^2\,dx$$

因此

$$\begin{aligned}\mathcal{B}&=\sum_{j=1}^{K}\int_{B_j}f'\left((j-\frac{1}{2})h\right)^2\cdot\frac{h^3}{12}\,dx\\&\approx\sum_{j=1}^{k}\frac{h^3}{12}\int_{B_j}\frac{f'(x)^2}{h}\,dx\\&=\frac{h^2}{12}\sum_{j=1}^{k}\int_{B_j}f'(x)^2\,dx\\&\approx\frac{h^2}{12}\int_{\mathbb{R}}f'(x)^2dx\end{aligned}$$

因此对于直方图而言, $MISE(\hat{f})=\frac{1}{nh}+\frac{h^2}{12}\|f'\|^2_{L^2(\mathbb{R})}$, 其中$\|f'\|^2_{L^2(\mathbb{R})}=\int_{\mathbb{R}}f'(x)^2$.

1.4 直方图的最优带宽

极小化MISE

$$\frac{\partial MISE(\hat{f})}{\partial h}=0\Rightarrow-\frac{1}{nh^2}+\frac{h}{6}\|f'\|^2_{L^2(\mathbb{R})}=0$$

因此最优带宽

$$h_o=\left(\frac{6}{n\|f'\|^2_{L^2(\mathbb{R})}}\right)^{\frac{1}{3}}\sim O(n^{-\frac{1}{3}})$$

使用大拇指法则

经常假设$f(x)$为正态分布的密度函数, 即$X\sim N(0,1)$且有$f(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$. 由此可以得出$f'(x)=-xf(x)$, $\|f'\|^2_{L^2(\mathbb{R})}=\frac{1}{4\sqrt{\pi}}$, 因此最优带宽

$$h_o=\left(\frac{6}{n\frac{1}{4\sqrt{\pi}}}\right)^{\frac{1}{3}}\approx3.5n^{-\frac{1}{3}}$$

1.5 核方法(Kernel density estimation)

设随机变量X来源于分布$f(\cdot)$. 由于对任给的$x\in\mathbb{R}$,

$$P(X\in(x-h,x+h))=\int_{x-h}^{x+h}f(x)dx=f(x)\cdot2h$$

又因为大数定律

$$P(X \in (x-h, x+h)) \approx \frac{\sharp\{i : X_i \in (x-h, x+h)\}}{n}$$

因此

$$\hat{f}(x) = \frac{\sharp\{i : X_i \in (x-h, x+h)\}}{2nh} = \frac{1}{2nh}\sum_{i=1}^{n} I(x-h \le x_i \le x+h)$$

注意核方法与直方图不同之处在于没有事先对坐标轴进行分割.

将上式换种方法写出, 即可获得

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2} I\left(\left|\frac{x_i - x}{h}\right| \le 1\right) = \frac{1}{nh}\sum_{i=1}^{n} W_i$$

其中

$$I\left(\left|\frac{x_i - x}{h}\right| \le 1\right)$$

往往被称为均匀核函数.

核函数需要满足的三个要求

- (对称性) $K(x) = K(-x)$
- (半正定) $K(u) \ge 0$
- (规范性) $\int_{\mathbb{R}} K(x)dx = 1$

因此对核密度估计而言, 其估计函数满足

- (半正定) $\hat{f}(x) \ge 0$
- (规范性) $\int_{\mathbb{R}} \hat{f}(x)dx = 1$
- $\hat{f}(x)$的连续性,可导性可以由$K(\cdot)$的连续性, 可导性决定.

Epanechnikov核

利用对$\hat{f}(x)$的均方误差极小化的方法可以得到最优核函数为Epanechnikov核.

$$K(u) = \frac{3}{4}(1-u^2)I(|u| \le 1)$$

此时x_i的权重为

$$w_i = \begin{cases} \frac{3}{4}\left(1 - \left(\frac{x-x_i}{h}\right)^2\right), & \text{if} \left|\frac{x-x_i}{h}\right| \le 1 \\ 0, & \text{if} \left|\frac{x-x_i}{h}\right| > 1 \end{cases}$$

带宽h往往被称为光滑参数, h越大, 参加平均的Y_i就越多, 会提高回归估计的精度, 但可能会增大估计的偏差, 反之h越小, 参加平均的Y_i越少, 会降低估计的精度, 但可以减小偏差.

对最优带宽的选取, 应用大拇指法则, 可以确定

$$h_o = 1.06\hat{\sigma}n^{-\frac{1}{5}}$$

其中

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

注记. 满足核函数条件下的均匀核, *Gauss*核, *Epanechnikov*核的最优性几乎一致, 因此核函数的不同选择在核密度估计中并不敏感.

1.6 K近邻(KNN)

近邻估计法也是常用的估计方法, 此法适用于密度的局部估计, 其基本思想如下: 设 $X_1, \cdots, X_n$ 是来自未知密度 $f(x)$ 样本, 先选定一个与n有关的整数$k = k_n$满足$1 \le k \le n$. 对于固定的$x \in \mathbb{R}$, 记$a_n(x)$为最小的正数a, 使得$[x-a, x+a]$中至少包含$X_1, \cdots, X_n$中的k个. 注意到对于每一个$a > 0$, 可以期望在$X_1, \cdots X_n$中大约有 $2af_n(x)$个观测值落入区间$[x-a, x+a]$中, 即$k \approx 2a_n(x)n\hat{f}(x)$, 于是定义

$$\hat{f}(x) = \frac{k}{2a_n(x)n}$$

为$f(x)$的最近邻(K近邻)估计.

与直方图估计不同, 此处区间长度$2a_n(x)$是随机的而区间内所含的观察数是固定的.

此外,虽然$\hat{f}(x)$是连续的, 但它的导数不一定是连续的, 这是因为$a_n(x)$在$\frac{X_{(j)} + X_{(j+1)}}{2}$的每一点处其导数是不连续的.

1.7 多元密度估计: 用核密度估计

$$\tilde{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$$

度量: 欧氏距离, 即$||\mathbf{x} - \mathbf{x}_i||_2^2$. 度量体现在核函数$K$的定义上, 当为一维时, 均匀核

$$\frac{1}{2} I(|u| \le 1)$$

而高维时就需要写成

$$\frac{1}{2} I(\|u\| \le 1)$$

1.8 维数灾祸(curse of dimensionality)

当维数增大时, 数据越来越稀疏(sparse), 衰减速度: 指数级.

例 (均匀核)**.**

$$K(u) = \frac{1}{2} I(\|u\|_2 \le 1)$$

例 (Epanechikov核)**.**

$$K(u) = \frac{3}{4}(1 - \|u\|_2^2) I(\|u\|_2 \le 1)$$

考虑数据$x_1, \cdots, x_n \overset{i.i.d.}{\sim} U[-1, 1], n = 10^5$, 大概有$\frac{10^5}{10} = 10^4$的数据点落入$[-0.1, 0.1]$(即$x = 0$近邻). 当维数增大时, $d = 10$, 考虑数据$x_1, \cdots, x_n \overset{i.i.d.}{\sim} U[-1, 1]^d, n = 10^5$, 则仅有$\frac{10^5}{10^{10}} = 10^{-5}$的数据点落入$[-0.1, 0.1]^{10}$($x = 0$近邻).

因此核函数方法往往仅适用于较小的维数$d = 1, 2, 3, 4$, 是一种局部(local)的方法.

2 非参数回归

对比参数回归模型, 非参数回归模型对回归函数提供了大量额外信息, 当假定的模型成立时往往具有较高的精度. 线性回归是参数方法的特例.

设响应变量Y是随机变量, 解释变量X是随机变量或非随机变量. 给定样本观察值$\{(X_i, Y_i)\}_{i=1}^n$, X与Y之间的回归关系可由下式确定:

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, 2, \cdots, n$$

其中$m(\cdot)$是未知的回归函数, ε_i是随机误差. 假定ε_i独立同分布, 满足

1. 当X非随机时, $\mathbb{E}\varepsilon_i = 0, Var(\varepsilon_i) = \sigma^2 < \infty$.
2. 当X为随机变量时, $(E\varepsilon_i|X_i) = 0, Var(\varepsilon_i|X_i = x) = \sigma^2(x) < \infty$.

我们主要讨论X也为随机变量的情况, 此时$m(x) = \mathbb{E}(Y|X = x)$.

注记. *按吴喜之教材上的记号*

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, \cdots, n$$

注记. *在线性回归中$m(x_i) = \beta_0 + \beta_1 x_i$, 线性回归是一种全局算法, 而以下三种方法只是**局部**的方法.*

2.1 K近邻(KNN)

近邻估计的直观想法:对给定的样本$X_1, \cdots, X_n$和$x \in \mathbb{R}^d$, 虽然可能没有一个X_i对给定的x恰好相等,但可以将“等于x”的要求降低为“与x接近”. 依每个X_i对给定x之间的距离重新排序, 与x距离越近的点其重要程度越大.

将$X_1, \cdots, X_n$依$||\cdot||$与x接近的程度排序

$$||X_{R_1} - x|| \le ||X_{R_2} - x|| \le \cdots \le ||X_{R_n} - x||$$

当有等好出现时, 采用下标靠前的原则, 然后选定n个常数$\{C_{ni}\}_{i=1}^n$, 满足

$$C_{n1} \ge C_{n2} \ge \cdots \ge C_{nn} \ge 0, \quad \sum_{i=1}^n C_{ni} = 1$$

定义. *令J_x^k为与x距离(欧氏)*

相关的推导可以写为

$$\hat{m}_k(x) = \sum_{i=1}^n W_k(x, x_i) y_i$$

其中

$$W_k(x, x_i) = \begin{cases} \dfrac{1}{k} & i \in J_x^k \\ 0 & i \notin J_x^k \end{cases}$$

J_x^k表示离x最近的k个数据点.

$W_k(x)$为权重(weight), 权重的要求有一下两个条件

1. $W_k(x) \geq 0$
2. $\sum_{i=1}^{n} W_k(x, x_i) = 1$

本质: 响应变量的加权平均, 近邻权也是一种概率权.

注记. 引进$\mathbb{R}^d$中的距离$\|\cdot\|$可以是欧氏距离, 也可以是$\max_{1\leq i\leq d}|x_i|$.

2.2 NW核回归

由于$m(x) = E[Y|X = x]$, 为估计$\hat{m}(x)$, 注意到

$$\begin{aligned} E[Y|X = x] &= \int y f_{Y|X}(y \mid x)\, dy \\ &= \int y \frac{f_{X,Y}(x, y)}{f_X(x)}\, dy \\ &= \frac{\int y f_{X,Y}(x, y)\, dy}{f_X(x)} \end{aligned}$$

对核密度估计

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

为估计$E[Y|X = x]$, 因此估计二元联合密度函数,

$$\hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right) \frac{1}{g} K\left(\frac{y - y_i}{g}\right)$$

注意到

$$\begin{aligned} &\int_{\mathbb{R}} \frac{y}{g} K\left(\frac{y - y_i}{g}\right) dy \\ &= \underbrace{\int_{\mathbb{R}} xK(x)dx}_{=0} + \underbrace{\frac{y_i}{g} \int_{\mathbb{R}} K(x)g\,dx}_{Y_i} \end{aligned}$$

则

$$\begin{aligned} \int y \hat{f}(x, y)\, dy &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right) \int \frac{y}{g} K\left(\frac{y - y_i}{g}\right) dy \\ &= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) Y_i \end{aligned}$$

除以$\hat{f}$得到

$$\hat{m}(x) = \frac{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) y_i}{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)} = \frac{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}$$

Nadaraya-Watson形式的核回归. 也可以写成

$$\hat{m}(x) = \sum_{i=1}^{n} W(x, x_i) y_i$$

也是Y_i的加权平均, 其中$W(x, x_i) = \dfrac{K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}$

$$\hat{m}_n(x) = \arg\min_{m(x)} \sum_{i=1}^{n} (W_i - m(x_i))^2 K\left(\frac{x - x_i}{h_n}\right)$$

上式x是固定的. 该式的含义是, 当离得近时, 预测应更加准确, 当离得远时, 有些误差不重要. 若对上式$m(x)$求导并令导数为0, 则有

$$2m \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) - 2 \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) Y_i = 0$$

即

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}$$

正是NW核回归.

例. *核回归,*

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}$$

之前的带宽选择方法

1. *理论最优带宽, 选取h使得$MISE = \dfrac{1}{nh} + \dfrac{h^2}{12}\|f'\|_2^2$最小$h_O = \left(\dfrac{6}{n\|f'\|_2^2}\right)^{\frac{1}{3}}$, 大拇指法则, 假设$X \sim N(0,1)$, $h_O \sim 3.5n^{-\frac{1}{3}}$.*
2. *交叉验证方法*

注记. *NW核回归相当于对模型$y_i = \beta_0 + \epsilon_i$做加权最小二乘回归.*

书中的例子

```r
library(MASS)
par(mfrow=c(2,2))
X=mcycle[,1]
Y=mcycle[,2]
bw=list("h=1", "h=2", "h=3", "h=5")
plot(X,Y,main=bw[[1]])
lines(ksmooth(X,Y,"normal",bandwidth=1), col = "blue")
plot(X,Y,main=bw[[2]])
lines(ksmooth(X,Y,"normal",bandwidth=2), col = "blue")
plot(X,Y,main=bw[[3]])
lines(ksmooth(X,Y,"normal",bandwidth=3), col = "blue")
plot(X,Y,main=bw[[4]])
lines(ksmooth(X,Y,"normal",bandwidth=5), col = "blue")
```

Listing 1: **code.R**

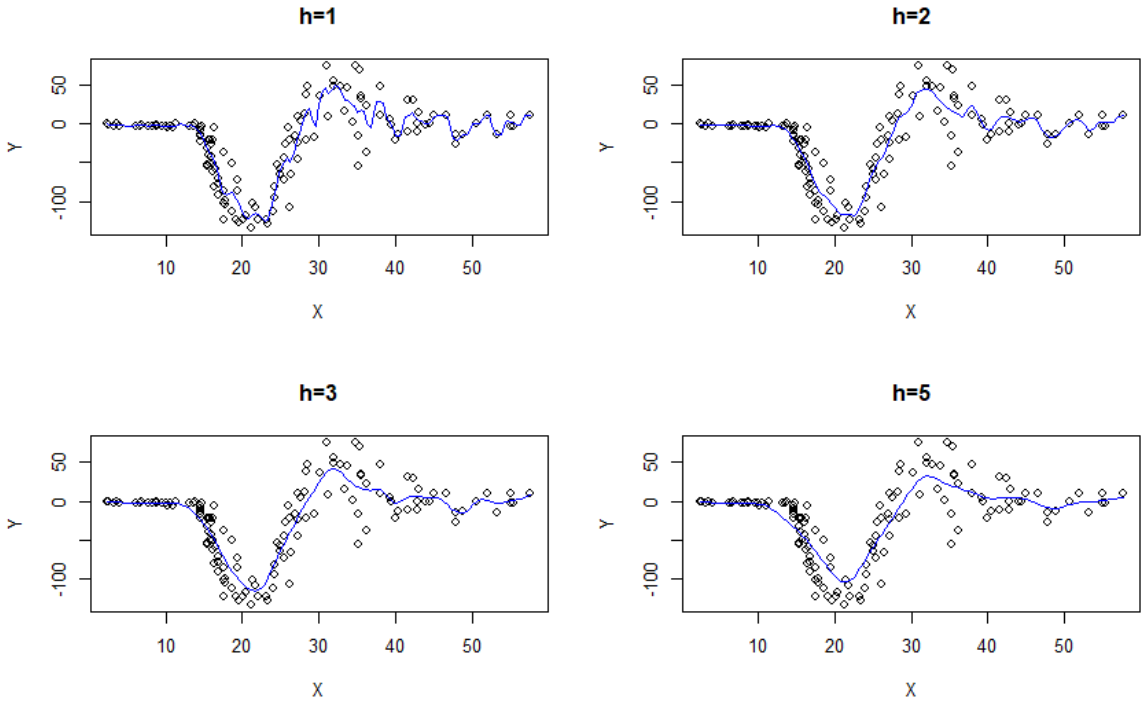


图 1: 使用核函数方法拟合的回归曲线

2.3 局部多项式(Local Polynomial)

回归函数的核估计存在边界效应, 即: 它在边界处收敛于真实函数的速度慢于在内点出的收敛速度. 核估计的本质是局部加权最小二乘得到的局部常数估计, 我们将其推广为局部p阶多项式估计.

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \cdots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

回顾

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, \cdots, n$$

其中$E(\varepsilon) = 0, Var(\varepsilon|X) = \sigma^2(X)$, 采用加权最小二乘法(WLS)

设 $m(x)$ 有 $p+1$ 阶导数, 对于任给的 $x_0 \in \mathscr{D}$, 由 Taylor 公式, 在 x_0 的邻域内有

$$\begin{aligned} m(x) &\approx m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p \\ &= \sum_{j=0}^{p} \beta_j (x - x_0)^j, \end{aligned}$$

其中$\beta_j = \dfrac{m^{(j)}(x_0)}{j!}$ 因此我们要找的就是$\beta_0, \cdots, \beta_p$, 即

$$\underset{\beta_0, \cdots, \beta_p}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j (x_i - x)^j \right)^2 K\left(\frac{x - x_i}{h} \right)$$

当$p=0$时,退化为NW方法.

显然$\hat{m}(x)=\hat{\beta}_0$, 实际计算中取遍x在定义域中, 即得到$m(x)$的估计曲线.

写成矩阵的形式,

$$\mathbf{X}(x)=\begin{pmatrix} 1 & x_1-x_0 & \cdots & (x_1-x_0)^p \\ 1 & x_2-x_0 & \cdots & (x_2-x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & x_n-x_0 & \cdots & (x_n-x_0)^p \end{pmatrix}$$

$$\mathbf{W}(x)=\begin{pmatrix} K\left(\frac{x-x_1}{h}\right) & 0 & \cdots & 0 \\ 0 & K\left(\frac{x-x_2}{h}\right) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & K\left(\frac{x-x_n}{h}\right) \end{pmatrix}$$

$$\mathbf{Y}=\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

$$\boldsymbol{\beta}(x)=\begin{pmatrix} \beta_0(x) \\ \beta_1(x) \\ \vdots \\ \beta_p(x) \end{pmatrix}$$

则估计值为 $\underset{\beta_0,\cdots,\beta_p}{\arg\min}(Y-X\beta)^\top W(Y-X\beta)=(X^\top WX)^{-1}X^\top WY$

证明.

$$\begin{aligned}&(Y-X\beta)^\top W(Y-X\beta)\\&=Y^\top WY-Y^\top WX\beta-\beta^\top X^\top WY+\beta^\top X^\top WX\beta\end{aligned}$$

因此

$$\frac{\partial}{\partial\beta}(Y-X\beta)^\top W(Y-X\beta)=-2Y^\top WX+\beta^\top(X^\top WX+X^\top WX)$$

故最小二乘估计量为

$$(X^\top WX)^{-1}X^\top WY$$

□

注记. *常常使用$p=1$, 即$y_i=\beta_0+\beta_1x_i+\epsilon_i$, 在这种情况下被称为局部回归(local regression).*

为什么考虑局部?

> 原本的想法是相邻的两个点X, X^*应当有相似的 $X \sim Y$ 关系, 从而$m(X)$与$m(X^*)$相似.($m(\cdot)$是一个光滑的函数, 即变化不快.)

反例: **信号处理**,出现信号加噪声, 不适用局部的方法.

2.4 局部回归的算法

以下算法给出在$X = x_0$处的局部回归模型,

1. 选取占所有数据$\frac{k}{n}$比例的最靠近x_0的数据$x_i, i = n_1, n_k$.
2. 对选出

3 交叉验证(Cross Validation)

应该有什么准则选取h?

应该对于任意的新数据(test data), (X^{new}, Y^{new}), 使得

$$\mathbb{E}(\underbrace{Y^{new}}_{\text{观测}} - \underbrace{\hat{m}_h(X^{new})}_{\text{预测}})^2 \tag{1}$$

尽可能小.

残差平方和, residual sum of square,

$$RSS(h) = \sum_{i=1}^{n}(Y_i - \hat{m}_h(X_i))^2 \tag{2}$$

而使得1和2式最小的h往往不是一个h, 1式使用了新的数据而2式(RSS(h))仅仅使用已有数据. 仅仅使用2式, 由于我们使用现有数据(training data)构造$\hat{m}(x)$, 会出现过拟合现有数据的情况.

$$\hat{h}_{CV} = \arg\min_h RSS_{CV}(h) = \arg\min_h \sum_{i=1}^{n}(Y_i - \hat{m}_h^{(-i)}(X_i))^2 \tag{3}$$

其中

$$\hat{m}_h^{(-i)}(x) = \frac{\sum_{j\neq i} K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{j\neq i} K\left(\frac{x-x_i}{h}\right)}$$

即只用到了$1, 2, \cdots, i-1, i+1, \cdots, n$共$n-1$个数据点, 没有用到$X_i, Y_i$, (X_i, Y_i)对$\hat{m}_h^{(-i)}(x)$而言是新的数据点. 3式与1式更为接近.

注记. 可证明,

$$RSS_{CV}(h) = RSS_{GCV}(h) = \sum_{i=1}^{n}(Y_i - \hat{m}_h(X_i))^2 \cdot H_i$$

其中H_i是一个对过小h的惩罚项. 上述方法大大降低了计算成本.

4 基函数(Basic Functions)方法

基本原理是对变量x的函数或变换进行建模.

$$m(x) = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \ldots + \beta_K b_K(x_i)$$

其中$b_1(\cdot), \cdots, b_K(\cdot)$的值是给定的, 已知的, 即在建模以前就确定了基函数的形式.

注记. 对比多项式回归,

$$m(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_K x^K$$

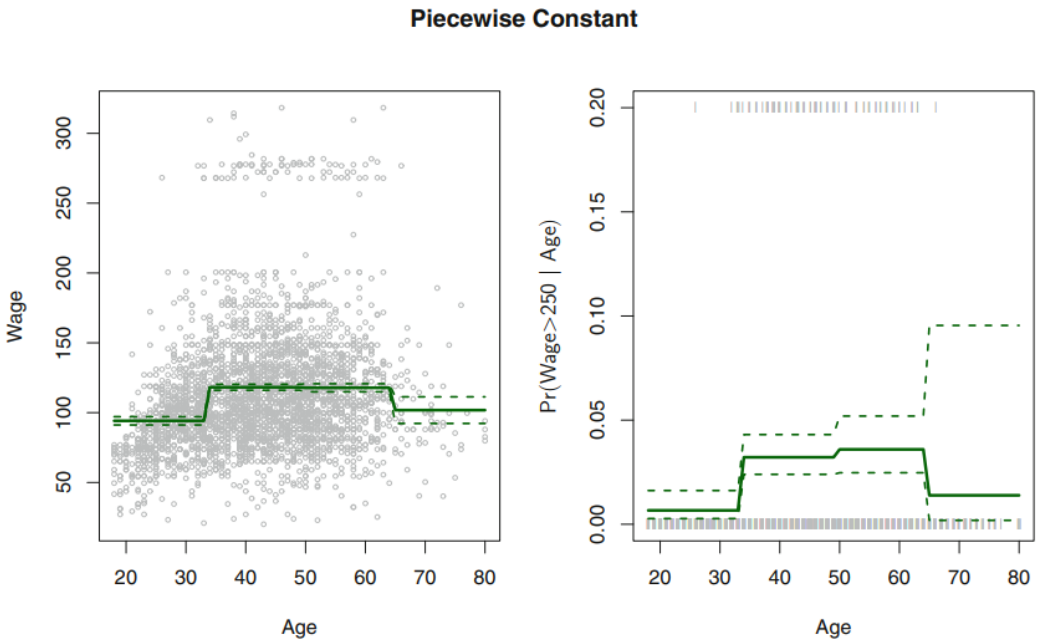
常见的基函数方法如下

1. 多项式回归(Polynomial Regression): 全局的方法, 用到了所有的数据点
2. 阶梯函数(Step Functions):在X的取值范围内选取k个固定的节点(knots):$c_1, \cdots, c_k$, 令

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

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

如此拟合: 非连续, 非光滑(局部的方法).

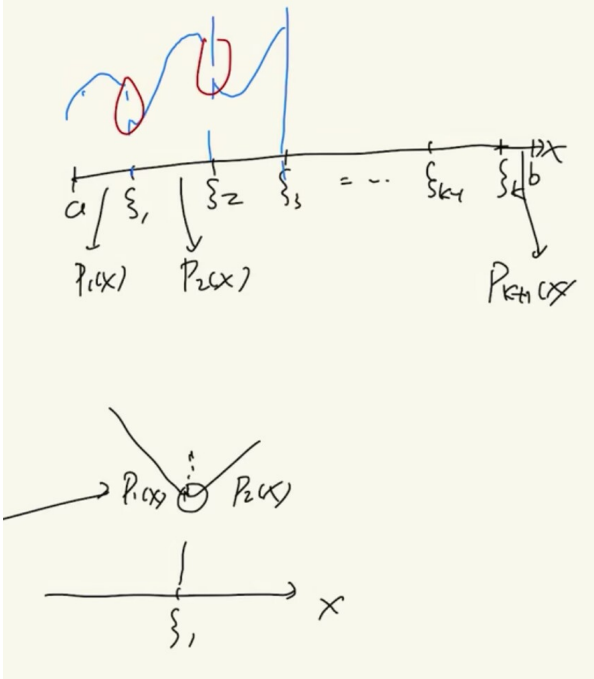


4.1 回归样条(spline)

- 三次样条(cubic spline)
- 自然三次样条(natural cubic spline)
- 光滑样条(smoothing spline)

4.1.1 三次样条

在X范围内选取k个节点, $\xi_1, \cdots, \xi_k$ 记$a = x_{(1)}, b = x_{(n)}$, 共$k+1$个区间,



$$(a, \xi_1), (\xi_1, \xi_2), \cdots, (\xi_{k-1}, \xi_k), (\xi_k, b)$$

每个区间均拟合一个**三次**多项式.

$$P_1(x) = \alpha_1 + \lambda_1 x + \gamma_1 x^2 + \delta_1 x^3$$
$$\vdots$$
$$P_{k+1}(x) = \alpha_{k+1} + \lambda_{k+1} x + \gamma_{k+1} x^2 + \delta_{k+1} x^3$$

为满足光滑性条件, 应满足如下的约束条件

$$P_1(\xi_1) = P_2(\xi_2) \tag{4}$$

式4保证了$m(x)$在ξ_1处的连续性.

$$P_1'(\xi_1) = P_2'(\xi_2) \tag{5}$$

式5保证了$m(x)$在ξ_1处的光滑性.

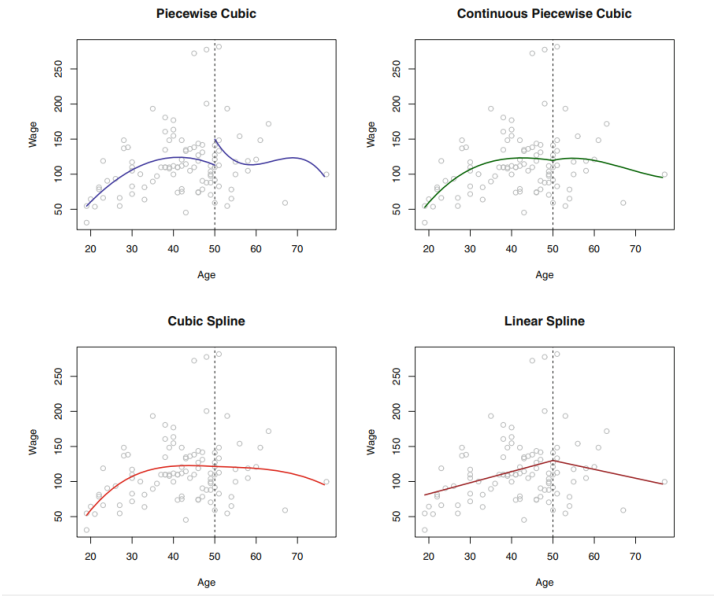
$$P_1''(\xi_1) = P_2''(\xi_2) \tag{6}$$

式6保证了$m(x)$在ξ_1处导数的连续性.

注记. 若仅有4式和5式, ξ_1左端一阶导变化快, 右边一阶导变化慢. 需要保证曲线的弯曲程度相同.

在边界上还要进行一些约束:$P_1''(a) = P_{k+1}'' = 0$, 即具有线性性质.

注记. 样条方法结合了多项式回归和阶梯函数的优点, 既保持了局部的信息, 又保持了回归函数的光滑性.



例. $\xi_1 = 0, P_1(x) = x^2, P_2(x) = X^3$, 在0附近, $P_2(x)$比$P_1(x)$更弯曲因为我们观察到

$$\begin{cases} P_1(0) = P_2(0) = 0 \\ P_1'(0) = P_2'(0) = 0 \\ P_1''(0) \neq P_2''(0) \end{cases}$$

最后, 三次样条可以表示为

$$Y_i = \beta_0 + \beta_1 b_1(x_i) + \cdots + \beta_{k+3} b_{k+3}(x_i) + \epsilon_i$$

其中$b_1(x) = x, b_2(x) = x^2, b_3(x) = x^3, b_l(x) = h(x, \xi_{l-3}), 4 \leq l \leq k+4$ 其中

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

k表示节点数, 被称为截断幂基(truncated power basis)函数. 因此有k个结点需要估计$k+4$个系数, 因此拟合三次样条总共需要$k+4$个自由度.

4.1.2 光滑样条(Smoothing spline)

回顾: 残差平方和(RSS), 式2给出

$$RSS = \sum_{i=1}^{n}(Y_i - \hat{m}(X_i))^2$$

在线性回归中, 对上述RSS应用最小二乘法, 即得到最佳的关于参数的估计. 而若在不对$m(\cdot)$做任何约束的情况下是否能对$m(\cdot)$做出一个好的估计?

> **思路.** 构造$\hat{m}(\cdot)$, 使得RSS越小越好.
>
> **注记.** 往往会造成过拟合*(overfitting)*.
>
> **例.** 任意的$\hat{m}(\cdot)$, 满足$\hat{m}(X_i) = Y_i$, 即内插法*(interpolate)*所有的数据点, 则有$RSS = 0$, 但是不光滑, 方差较大, 相应偏倚方差分解中, 方差较大.
>
> 解决方法是加惩罚项(penalize), 使得$\hat{m}(\cdot)$不能太不光滑.

$$\widehat{m}_\lambda(\cdot) = \arg\min_{m(\cdot)} \sum_{i=1}^{n}\left(\underbrace{y_i}_{\text{观测}} - \underbrace{m(x_i)}_{\text{预测}}\right)^2 + \lambda \int \underbrace{m''(x)^2}_{\text{曲线的光滑程度}} dx \tag{7}$$

其中$\lambda > 0$为超参数. 当λ较大时, 二阶导$m''(x)$ 需要很小, 即曲线很光滑, 才能使损失函数(loss)最小化, 换句话说λ越大, $m(\cdot)$越光滑.

注记. 式7有显性表达式, 且属于自然三次样条的特殊情况,

λ的值允许从0增加到∞, 这样实际的自由度df_λ就从n下降到2.

若记对于某一特定的λ, $\hat{g}$是相应的解, 即对应光滑样条的拟合值, 由推导可知

$$\hat{g}_\lambda = S_\lambda y$$

由此可以定义df_λ

$$df_\lambda = \sum_{i=1}^{n}\{S_\lambda\}_{ii} = tr(S_\lambda)$$

由此

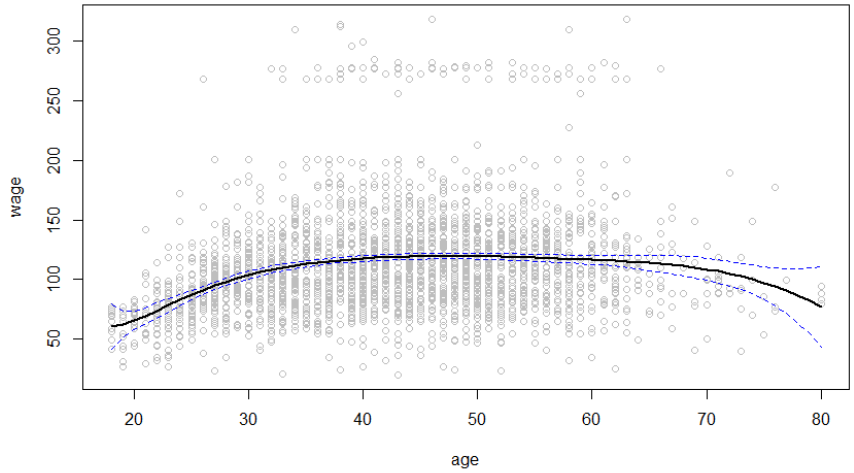
$$\text{RSS}_{cv}(\lambda) = \sum_{i=1}^{n}\left(y_i - \hat{g}_\lambda^{(-i)}(x_i)\right)^2 = \sum_{i=1}^{n}\left[\frac{y_i - \hat{g}_\lambda(x_i)}{1 - \{\mathbf{S}_\lambda\}_{ii}}\right]^2$$

```r
# 中包中提供拟合回归样条的函数Rspline
library(splines)
library(ISLR)
attach(Wage)
age.grid=seq(from=range(age)[1], to=range(age)[2])

# bs()默认生成三次样条
```

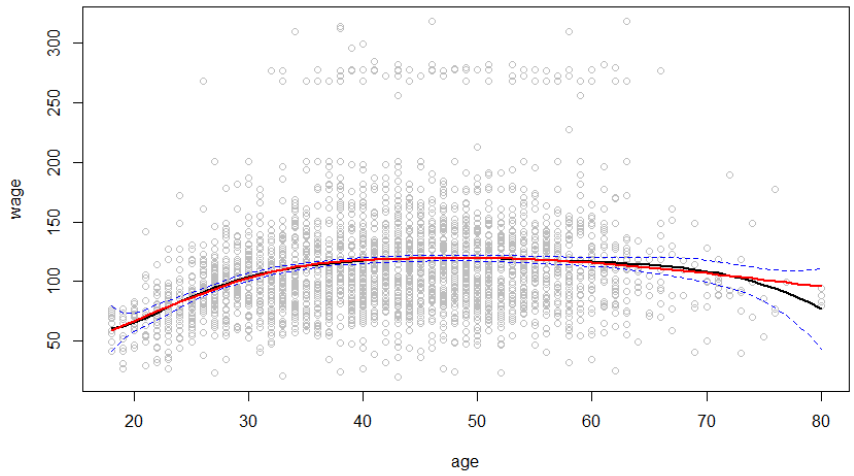
```
fit=lm(wage~bs(age, knots=c(25, 40, 60)), data = Wage)
pred=predict(fit, newdata = list(age = age.grid), se=T)
plot(age, wage, col="gray")
lines(age.grid, pred$fit, lwd = 2)
lines(age.grid, pred$fit + 2*pred$se.fit, lty = "dashed", col="blue")
lines(age.grid, pred$fit - 2*pred$se.fit, lty = "dashed", col="blue")
```

上述代码节点设置为25, 40, 60. 我们用$k+3$个基函数进行建模, 这里自由度$df = k+3$, 因此电脑也可以自动通过选取分位点设置数据节点.



自然样条也很容易拟合,

```
fit2=lm(wage~ns(age, df=4), data = Wage)
pred2=predict(fit2, newdata = list(age = age.grid), se=T)
lines(age.grid, pred2$fit, col = "red", lwd = 2)
```



光滑样条也很好拟合, 用`smooth.splines()`即可, 选择λ的两种方法: 自由度方法, 交叉验证方法.

5 广义可加模型(GAM, generalized additive model)

数据:$(Y_i, X_{i1}, \cdots, X_{ip})^\top, i = 1, \cdots, n$, 我们这回考虑多元线性回归的推广.

$$Y_i = m(X_{i1}, \cdots, X_{ip}) + \epsilon_i$$

多元线性回归

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

例. $p = 3$, Y为工资, X_1为年份, X_2为年龄, X_3为受教育程度. 若使用线性回归,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

从上述图中可以看出工资的增长与年龄的关系并非线性.

一种自然的推广方法是用一个光滑的**非线性**函数$f_j(x_{ij})$替代$\beta_j x_{ij}$, 于是模型可以重新写为

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} f_j\left(x_{ij}\right) + \epsilon_i \\ &= \beta_0 + f_1\left(x_{i1}\right) + f_2\left(x_{i2}\right) + \cdots + f_p\left(x_{ip}\right) + \epsilon_i \end{aligned}$$

上述的f可以用比如spline去拟合.

注记. 上述的f不能分开去拟合, 而应该同时拟合p个函数f.

> 拟合的算法: 向后拟合方法(backfitting). R语言采用这种方法, 循环依次对每个变量更新系数并保持其它系数不变的情况下拟合模型, 这种方法的好处是在每次更新函数时, 只需要将拟合的方法应用到变量的部分残差上.
>
> 部分残差(partial residual)是指, 比如要拟合X_3, 当前模型已经有残差$r_i = y_i - f_1(x_{i1}) - f_2(x_{i2})$, 如果$f_1, f_2$已知, 可以应用这个残差来对$X_3$建立非线性回归, 从而拟合$f_3$.

GAM的模型复杂度低于单纯的非参数方法.

注记. 无论是加性模型还是*GAM*, 均不考虑交互作用项. 若要考虑交互作用*(interaction)*, 对应$r_i x_{i1} x_{i2}$, 或在*GAM*中$g(x_i, x_j)$

```
library(gam)
gam.m3=gam(wage~s(year, 4)+s(age, 5)+education, data = Wage)
par(mfrow=c(1,3))
plot.Gam(gam.m3, se=T, col="blue")
gam.m1=gam(wage~s(age, 5)+education, data = Wage)
gam.m2=gam(wage~year + s(age, 5)+education, data = Wage)
anova(gam.m1,gam.m2,gam.m3,test = "F")
```

上述的ANOVA, 第一个p值,

$$H_0 : \text{no year}, \quad vs. \quad H_1 : \text{year linear}$$

第二个p值,

$$H_0 : \text{year linear}, \quad vs. \quad H_1 : \text{year spline}$$

```
> anova(gam.m1, gam.m3, test = "F")
Analysis of Deviance Table

Model 1: wage ~ s(age, 5) + education
Model 2: wage ~ s(year, 4) + s(age, 5) + education
  Resid. Df Resid. Dev Df Deviance     F    Pr(>F)
1      2990    3711731
2      2986    3689770  4    21960 4.443 0.001397 **
---
Signif. codes:  0 ‘’ *** 0.001 ‘’ ** 0.01 ‘’ * 0.05 ‘’ . 0.1 ‘ ’ 1
```

上述数据表明, 带有year线性函数项的模型比不含year的模型更好, 没有理由支持关于year的非线性项是必须的.

局部线性回归(local regression)也能作为GAM的一部分, 用`lo()`即可实现.

产生交互作用项: 应用`lo()`产生交互作用, 再用`gam()`函数来拟合模型.

$$\text{wage}_i = \beta_i + f_i(\text{year}, \text{age}) + \beta_1 z_1 + \cdots + \beta_4 z_4 + \epsilon_i$$

其中

$$z_1 = \begin{cases} 1, \text{if HS} \\ 0, \text{OW(otherwise)} \end{cases} \quad \cdots \quad z_4 = \begin{cases} 1, \text{if >Coll} \\ 0, \text{OW(otherwise)} \end{cases}$$

```
gam.lo.i=gam(wage~lo(year,age, span = 0.5) + education, data = Wage)
```