第八章　非参数回归

在实际中, 我们经常要研究两个变量X与Y的函数关系，如图8.1(见chap8 数据motor.txt)所示为两幅二元函数的散点图,左图由230个成对样本点构成, 其中$Y_i = \sin(4X_i) + \varepsilon_i, X_i \sim U(0,1)$,$\varepsilon_i \sim N(0, 1/3), i = 1, 2, \cdots, 230$. X和Y看似存在某种非线性函数关系, 可以尝试非线性回归。最常见的一种做法是用一个多项式回归来刻画二者的关系，如下所示：

$$y(x, \beta) = \sum_{j=0}^{p} \beta_j x^j . x^0 = 1$$

如果线性关系成立, 那么$p = 1$,如果关系不是线性的，$p > 1$. 选用高阶回归可以在一定程度上改善线性模型的拟合优度. 但是, 多项式回归的不足在于对其阶数的选择。单从拟合优度来看，一般更倾向于取较高的阶数，这时模型会非常强烈地依赖于几个关键点,对这些点的变化非常敏感, 如果这些点出现小的扰动,则可能会波及远离这些点的一些点的估计以及它们附近的曲线走向. 多项式回归需要调整参数p的大小，当关系复杂时，p 也倾向于取更高阶，选择高阶的p的代价是高阶的系数不仅不容易估得准确，常常具有较大的方差，而且还会出现系数膨胀现象，这样很容易产生错误的回归估计模型。本章将讨论复杂数据关系的非参数回归模型的解决方案，这些方案具有两个共同的特点：一是模型不是事先设定的，二是模型中引入了灵活可调解的参数，从而尽可能用低阶的回归模型去解决复杂的数据关系问题。

在图8.1中，右图是很多统计学家都研究过的摩托车碰撞模拟数据的散点图, 由133个成对数据构成. X为模拟的摩托车发生相撞事故后的某一短暂时刻(单位为百万分之一秒), Y是该时刻驾驶员头部的加速度(单位为重力加速度g). X和Y之间直觉上是有某种函数关系的, 但是很难用参数方法进行回归, 也很难用普通的多项式回归拟合. 因此考虑更如下一般的模型：

给定一组样本观测值$(Y_1, X_1), (Y_2, X_2), \cdots, (Y_n, X_n)$, X_i和Y_i之间的任意函数模型表示为

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, 2, \cdots, n. \tag{8.1}$$

其中$m(\cdot) = E(Y|X)$, ε为随机误差项. 一般假定$E(\varepsilon|X = x) = 0$, $\mathrm{var}(\varepsilon|X = x) = \sigma^2$, 不必是常数.

§8.1　核回归光滑模型

回顾上一章刚刚介绍过的核密度估计法, 它相当于求x附近的平均点数. 平均点

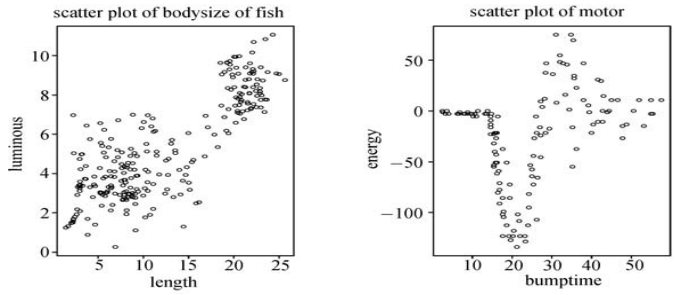

图 8.1 复杂的二元关系数据散点图

数的求法是对可能影响到x 的样本点, 按照距离x的远近作距离加权平均. 核回归光滑的基本思路与之类似, 这里不是求平均点数, 而是估计点x处y的取值. 仍然按照距离x的远近对样本观测值y_i加权即可. 这就是纳达拉亚(Nadaraya)及沃森(Watson)(1964)提出的Nadaraya-Watson核回归的基本思想.

定义8.1 选定原点对称的概率密度函数$K(\cdot)$为核函数及带宽$h_n > 0$,

$$\int K(u)\mathrm{d}u = 1. \tag{8.2}$$

定义加权平均核为

$$\omega_i(x) = \frac{K_{h_n}(X_i - x)}{\sum_{j=1}^{n} K_{h_n}(X_j - x)}. \tag{8.3}$$

其中$K_{h_n}(u) = h_n^{-1}K(uh_n^{-1})$也是一个概率密度函数. Nadaraya-Watson 核估计定义为

$$\hat{m}_n(x) = \sum_{i=1}^{n} \omega_i(x) Y_i. \tag{8.4}$$

注意到

$$\hat{\theta} = \min_{\theta} \sum_{i=1}^{n} \omega_i(x)(Y_i - \theta)^2 = \sum_{i=1}^{n} \frac{\omega_i Y_i}{\sum_{i=1}^{n} \omega_i}, \tag{8.5}$$

因此, 核估计等价于局部加权最小二乘估计. 权重$\omega_i = K(X_i - x)$. 常用的核函数与上一章的表7.1类似.

若$K(\cdot)$是$[-1, 1]$上的均匀概率密度函数, 则$m(x)$的Nadaraya-Watson 核估计就是落在$[x - h_n,\ x + h_n]$上的X_i对应的Y_i的简单算术平均值. 称参数h_n为带宽, h_n越小, 参与平均的Y_i就越少; h_n越大, 参数平均的Y_i就越多.

若$K(\cdot)$是$[-1,1]$上的概率密度函数, 则$m(x)$的Nadaraya-Watson 核估计就是落在$[x-h_n,\ x+h_n]$上的X_i对应的Y_i的加权算术平均值.

若$K(\cdot)$是$(-\infty,\ +\infty)$上关于原点对称的标准正态密度函数, 则$m(x)$的Nadaraya-Watson 核估计就是Y_i的加权算术平均值. 当X_i离x越近时, 权数就越大; 离x越远时, 权数就越小; 当X_i落在$[x-3h_n,\ x+3h_n]$之外时, 权数为零.

Nadaraya-Watson核估计直接使用密度加权, 但是在实际估计参数和计算带宽的时候, 可能需要对权重取导数运算, 这时将核表达为密度积分的形式是比较方便的, 这就导致了另一种核估计——Gasser-Müller核估计：

$$\hat{m}(x)=\sum_{i=1}^{n}\int_{s_{i-1}}^{s_i}K\left(\frac{u-x}{h}\right)\,\mathrm{d}u\ y_i.$$

式中, $s_i=(x_i+x_{i+1})/2, x_0=-\infty, x_{n+1}=+\infty$.显然它是用面积而不是密度本身作为权重.

例8.1(核回归的例子)　图8.2所示为鲑鱼和鲈鱼体长与光泽度之间的Nadaraya-Watson核回归光滑. 为了说明带宽h的作用, 这里的h分别取3, 1.5, 0.5 和0.1.

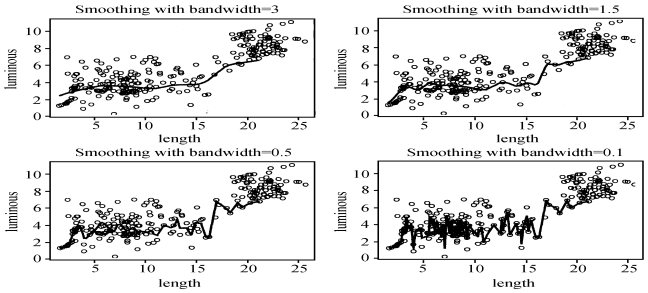

图 8.2　不同带宽的鲑鱼和鲈鱼体长(length)和光泽度(luminous)核回归

§8.2　局部多项式回归

§8.2.1　局部线性回归

核估计虽然实现了局部加权, 但是这个权重在局部邻域内是常量, 由于加权是基于整个样本点的, 因此在边界的估计往往不理想. 如图8.3所示, 真实的曲线用虚线表示, Nadaraya-Watson核回归拟合曲线用虚线表示. 在左边和右边的边界点处, 曲线真实的走向有很大的线性斜率, 但是在拟合曲线上, 显然边界的估计有高估的现象. 这是因为核函数是对称的, 因而在边界点处, 起决定作用的是内点, 比如影响左边界点走势的主要是右边的点. 同样, 影响到右边界点走势的主要是左边的点.

越到边界这种情况越突出. 显然问题并非仅对外点而言, 如果内部数据分布不均匀, 则那些恰好位于高密度附近的内点的核估计也会存在较大偏差.

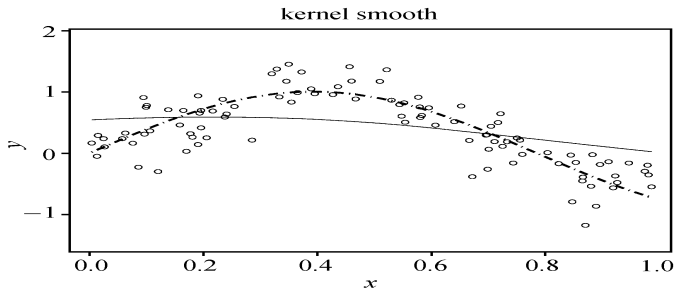

图 8.3 核回归和真实函数曲线比较

解决的方法是用一个变动的函数取代局部固定的权, 这样就可能避免这种边界效应. 最直接的做法就是在待估计点x 的邻域内用一个线性函数$Y_i = a(x) + b(x)X_i,\ X_i \in [x-h, x+h]$取代$Y_i$的平均, 其中$a(x)$和$b(x)$是两个局部参数. 因而就得到了局部线性估计.

具体而言, 局部线性估计为最小化

$$\sum_{i=1}^{n}\{Y_i - a(x) - b(x)X_i)\}^2 K_{h_n}(X_i - x), \tag{8.6}$$

其中$K_{h_n}(u) = h_n^{-1}K(h_n^{-1}u)$,$K(\cdot)$为概率密度函数. 若$K(\cdot)$是$[-1,1]$上的均匀概率密度函数$K_0(\cdot)$, 则$m(x)$的局部线性估计就落在$[x-h_n,\ x+h_n]$的$X_i$与其对应的$Y_i$关于局部模型

$$\hat{m}(x) = \hat{a}(x) + \hat{b}(x)X_i \tag{8.7}$$

的最小二乘估计.

若$K(\cdot)$是$[-1,1]$上的概率密度函数$K_2(\cdot)$, 则$m(x)$的局部线性估计就落在$[x-h_n,\ x+h_n]$的X_i与其对应的Y_i关于局部模型(8.6)的加权最小二乘估计. 当X_i越接近x时, 对应Y_i的权数就越大; 反之, 则越小.

若$K(\cdot)$是$(-\infty,+\infty)$上关于原点对称的标准正态密度函数$K_2(\cdot)$, 则$m(x)$的局部线性估计就是局部模型(8.6)的加权最小二乘估计. 当X_i离x越近时, 权数就越大; 反之, 就越小. 当X_i落在$[x-3h_n,\ x+3h_n]$之外时, 权数基本上为零.

$m(x)$的局部线性估计的矩阵表示为

$$\begin{aligned}\hat{m}_n(x,h_n)&=\boldsymbol{e}_1^{\mathrm{T}}(\boldsymbol{X}_x^{\mathrm{T}}\boldsymbol{W}_x\boldsymbol{X}_x)^{-1}\boldsymbol{X}_x^{\mathrm{T}}\boldsymbol{W}_x\boldsymbol{Y}\\&=\sum_{i=1}^{n}l_i(x)y_i.\end{aligned}\tag{8.8}$$

其中

$$\boldsymbol{e}_1=(1,0)^{\mathrm{T}},\ \boldsymbol{X}_x=(X_{x,1},\cdots,X_{x,n})^{\mathrm{T}},\ \boldsymbol{X}_{x,i}=(1,(X_i-x))^{\mathrm{T}},$$

$$\boldsymbol{W}_x=\mathrm{diag}[K_{h_n}(X_1-x),\cdots,K_{h_n}(X_n-x)],\ \boldsymbol{Y}=[Y_1,\cdots,Y_n]^{\mathrm{T}}.$$

当解释变量为随机变量时, 局部线性估计$\hat{m}_n(x,h_n)$在内点处的逐点渐近偏差和方差如表8.1所示.

表8.1 局部线性估计内点渐近偏差和方差

	渐近偏差	渐近方差
总变异	$h_n^2\dfrac{m''(x)}{2}\mu_2(K)$	$\dfrac{\sigma^2(x)}{nh_nf(x)}R(K)$

使得$\hat{m}_n(x,h_n)$的均方误差达最小的最佳窗宽为

$$h_n=cn^{-1/5}.\tag{8.9}$$

其中c与n无关, 只与回归函数、解释变量的密度函数和核函数有关. 在内点, 使得$\hat{m}_n(x,h_n)$的均方误差达到最小的最优的核函数为$K(z)=0.75(1-z^2)_+$, 此时, 局部线性估计可达到收敛速度$O(n^{-2/5})$.

例8.2 如图8.4显示了用局部线性回归对图8.3关系的重新拟合, 可见边界效应问题有所缓解, 即其在边界点的收敛速度与内点几乎一样, 且等于核估计在内点处的收敛速度, 它的偏差比核估计小, 而且其偏差与解释变量的密度函数无关. 此外, 局部线性估计在估计出回归函数$m(x)$的同时也估计出回归函数的导函数$m'(x)$, 导数在实际中可用于分析边际变化率.

§8.2.2 局部多项式回归的基本原理

如图8.4所示, 与真实函数比较起来, 局部线性回归虽然较好地克服了边界的偏差, 但在曲线导数符号改变的附近, 仍然产生偏差, 又由于导数改变的点通常为极值点, 因而呈现出“山头被削, 谷底添满”的光滑效果, 这时就需要考虑高阶局部多项式的情况. 局部线性回归很容易扩展到一般的局部多项式回归.

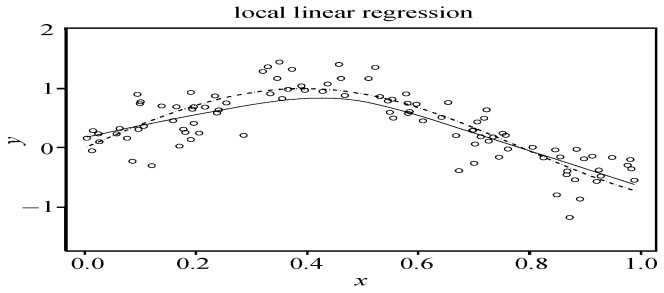

图 8.4 局部线性回归和真实曲线的比较图

考虑二元数据对$\{(X_1,Y_1),\cdots,(X_n,Y_n)\}$, 它们独立同分布取自总体$(X,Y)$, 待估的回归函数是: $m(x)=E(Y|X=x)$, 它的各阶导数记为$m'(x),m''(x),\cdots,m^{(p)}(x)$.

定义8.2 局部p阶多项式估计为最小化p阶多项式

$$\sum_{i=1}^{n}[Y_i-\beta_0-\cdots-\beta_p(X_i-x)^p]^2K\left(\frac{X_i-x}{h}\right), \tag{8.10}$$

这里的记号与前面类同. h是带宽, K是核函数.

令

$$\boldsymbol{X}=\begin{pmatrix}1 & X_1-x & \cdots & (X_1-x)^p\\ \vdots & \vdots & & \vdots\\ 1 & X_n-x & \cdots & (X_n-x)^p\end{pmatrix},$$

$$\boldsymbol{\beta}=\begin{bmatrix}\hat\beta_0\\ \hat\beta_1\\ \vdots\\ \hat\beta_p\end{bmatrix}_{(p+1)\times 1},\quad \boldsymbol{y}=\begin{bmatrix}Y_1\\ Y_2\\ \vdots\\ Y_n\end{bmatrix}_{n\times 1},$$

$$\boldsymbol{W}=h^{-1}\mathrm{diag}\left[K\left(\frac{X_1-x}{h}\right),\cdots,K\left(\frac{X_n-x}{h}\right)\right].$$

因此有加权最小二乘问题的估计$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}$.

例8.3 如图8.5所示, 图中实线表示真实曲线走向, 虚线显示了用局部二项回归对图8.4关系的重新拟合, 可见极值点的问题有所缓解.

§8.3 LOWESS稳健回归

异常点可能造成线性回归模型最小二乘估计发生偏差, 因而有必要改进局部线

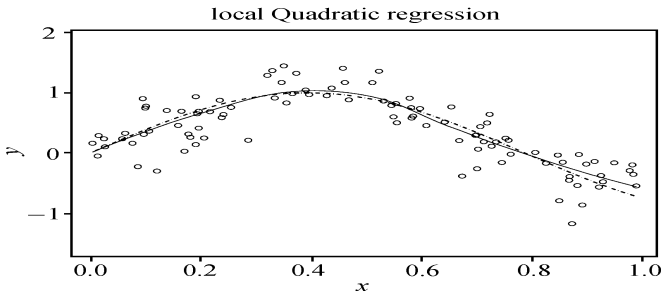

图 8.5 局部线性回归和真实曲线的比较图

性拟合方法来降低异常点对估计结果的影响. LOWESS(locally weighted scatter plot smoothing) 稳健估计方法就是在这样的背景条件下产生的, 它是由威廉姆.克里维兰德(Williams.Cleveland,1979)提出的, 目前已在国际上得到了广泛的应用. LOWESS的基本思想是先用局部线性估计进行拟合, 然后定义稳健的权数并进行平滑, 重复运算几次后就可消除异常值的影响, 从而得到稳健的估计. LOWESS稳健估计的计算步骤如下.

第一步：对模型(8.6)进行局部线性估计, 得到$m(X_i)$的估计$\hat{m}(X_i)$, 进而得到残差$r_i = Y_i - \hat{m}(X_i)$.

第二步：计算稳健权数$\delta_i = B(r_i/(6 \cdot \text{median}(|r_1|, |r_2|, \cdots, |r_n|)))$, 其中$B(t) = (1 - |t|^2)^2 I_{[-1,1]}(t)$. 式中

$$I_{[-1,1]}(t) = \begin{cases} 1, & \text{当} |t| \leqslant 1\text{时}, \\ 0, & \text{当} |t| > 1\text{时}. \end{cases}$$

第三步：使用权δ_i $K(h_n^{-1}(X_i - x))$对模型(8.1)进行局部加权最小二乘估计, 就可得到新的r_i.

第四步：重复第二步和第三步s次后就可得到稳健估计.

由于稳健权数δ_i可将异常值排除在外, 并且初始残差大(小)的观测值在下一次局部线性回归中的权数就小(大), 因而, 重复几次后就可将异常值不断地排除在外, 并最终得到稳健的估计. 克里维兰德(Cleveland,1979)推荐$s = 3$.

例8.4(见教学资源数据fish.txt) 本例仍然是关于鲑鱼和鲈鱼两种鱼类长度和光泽度之间关系的进一步研究, 假设现在有3个异常点被加入, 3个异常点分别为：$\boldsymbol{x}_1 = (22.03784, -18.22867)$, $\boldsymbol{x}_2 = (24.21510, -20.62153)$, $\boldsymbol{x}_3 = (22.70523, -20.90481)$. 这些异常点可能是由于仪器损坏、人为疏漏或黑客侵犯等原因造成的. 图8.6 左图为局部线性核最小二乘估计的拟合值与鲑鱼和鲈鱼两种鱼类长度和光泽度之间散点图的比较, 右图为LOWESS稳健估计的拟合值和实际值散点图的比

较. 左图曲线的右端显然有向下的偏差, 这是异常值造成的, 而右边图形中向下的偏差并不明显. 由此可见, LOWESS稳健估计方法通过三次对异常点权重的减少, 基本上消除了异常点对非参数回归模型估计的影响. 而且该方法不需要知道异常点的位置, 简单易行, 因而在国际上得到广泛的应用.

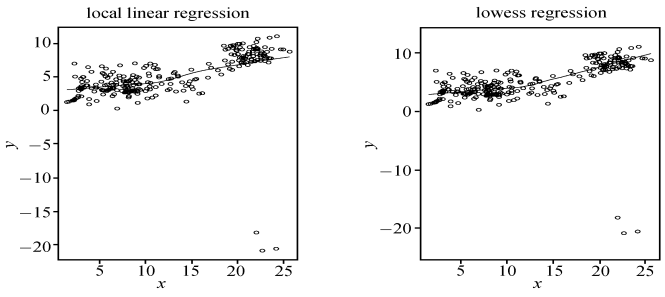

图 8.6 局部线性回归和LOWESS稳健回归拟合效果比较图

§8.4 k近邻回归

与k近邻密度估计类似, 这里也有k近邻回归, 它的基本原理是用距离待估计点最近的k个样本点处y_i的值来估计当前点的取值. 按照是否对这些点按距离加权, k近邻回归又分为普通k近邻估计和k近邻核加权回归两类, 下面分别介绍两者的应用.

k近邻估计的主要优点在于该方法可以自动地适应局部信息. 也就是说, 局部的点越多, 所选的带宽越小, 这是k近邻和核估计的主要不同之处. 但另一方面, k近邻估计过于局限于局部信息, 而失去了一些全局的信息, 对于有些数据, 这种方法有一定的缺陷.

1. k近邻估计

令$1 < k < n$, 记

$$I_{x,k} = \{i:\ X_i\text{是离}x\text{最近的}k\text{个观测值之一}\}. \tag{8.11}$$

非参数回归模型(8.1)的k近邻估计为

$$\hat{m}_n(x,k) = \sum_{i=1}^{n} w_i(x,k) Y_i. \tag{8.12}$$

其中

$$w_i(x,k) = \begin{cases} 1/k, & i \in I_{x,k}, \\ 0, & i \notin I_{x,k}. \end{cases}$$

当解释变量为随机变量时, 如果当$n \to \infty$时, $k \to \infty$, $k/n \to 0$, 则$\hat{m}_n(x,k)$在内点处逐点渐近偏差和方差如表8.2. 此外, 在适当的条件下, $\hat{m}_n(x,k)$还具有一致性和渐近正态性.

表8.2 k近邻估计内点逐点渐近偏差和方差

	渐近偏差	渐近方差
总变异	$\frac{1}{24f(x)^3}[(m''f+2m'f')(x)](k/n)^2$	$\frac{\sigma^2(x)}{k}$

k近邻估计既适合于解释变量是确定性的模型, 也适合于解释变量是随机变量的模型.

2. k近邻核估计

非参数回归模型(8.1)的近邻核估计为

$$\hat{m}_n(x,k)=\frac{\sum_{i=1}^{n}K((X_i-x)/R(x,k))Y_i}{\sum_{i=1}^{n}K((X_i-x)/R(x,k))}. \tag{8.13}$$

其中$R(x,k)=\max\{|X_i-x|:\ i\in I_{x,k}\}$.

由上式可见, k近邻估计是k近邻核估计的特例. 由式(8.12)可知, k近邻估计就是用最靠近x的k个观测值进行加权平均. 它的基本原理与核估计相似, 性质也相似. 当解释变量为随机变量时, 当$n \to \infty$时, $k \to \infty$, $k/n \to 0$, 则$\hat{m}_n(x,k)$在内点处的逐点渐近偏差和方差如表8.3所示. 此外, 在适当的条件下, $\hat{m}_n(x,k)$还具有一致性和渐近正态性. 易见, k近邻估计在内点处的收敛速度可达到$O(n^{-2/5})$.

表8.3 k近邻核估计的内点逐点渐近偏差和方差

	渐近偏差	渐近方差
总变异	$\frac{\mu(K)}{8f(x)^3}[(m''f+2m'f')(x)](k/n)^2$	$2\frac{\sigma^2(x)}{k}R(K)$

例8.5 本例是关于鲑鱼和鲈鱼两种鱼类长度和光泽度之间关系的k近邻回归, 图8.7左图表示$k=3$时的近邻估计, 右图表示$k=6$时的近邻估计. 我们发现随着k的增加, 曲线的光滑度也在增加, 但是与核回归比较, k近邻回归显然在k较小的时候不够光滑.

k近邻回归的主要优点在于该方法可以自动地对数据进行局部估计. 也就是说, 当一个点的附近有许多观测点时, 所选的带宽越小, 这是k近邻回归与8.1节的核回

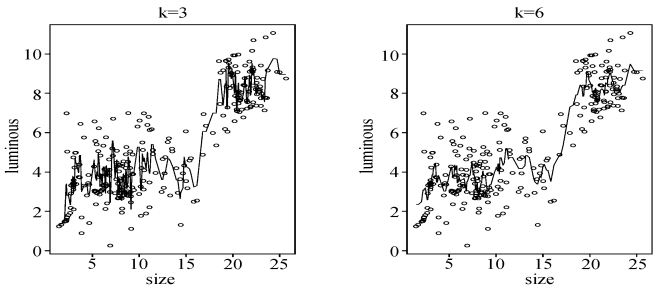

图 8.7 $k-$近邻回归

归的主要不同之处.但另一方面, k近邻回归过于强调局部估计,这样就有可能忽视较远观测值对局部模式的影响，选择合适的k是k近邻回归有效的必要条件。

§8.5 正交序列回归

前面介绍的非参数回归模型的核估计、局部线性估计和近邻估计属局部估计方法, 局部估计方法用于预测时只能预测数据区域内的回归函数值,对于附近没有观察点的回归函数值则无法预测, 因而全局估计法仍然需要. 正交序列回归的一个优势在于正交基函数比较容易构造, 比如数学上常用的Fourier序列. 因此, 整个方法在结构上比较简单, 而且在数学上比较容易分析其性质.本节将简单介绍正交序列估计的基本原理.

设回归函数$m(x) \in C[a,b]$, 假设$\{\varphi_i\}_{j=0}^{\infty}$构成$[a,b]$上的一组正交基, 即

$$\int_a^b \varphi_i(x)\varphi_j(x)\mathrm{d}x = \delta_{ij} = \begin{cases} 0, & i \neq j, \\ c_i, & i = j, \end{cases}$$

则$m(x)$有正交序列展开$m(x) = \sum\limits_{i=1}^{\infty} \theta_i \varphi_i(x)$. 可将非参数回归模型(8.1)近似为

$$Y_i = \sum_{j=1}^{m} \theta_j \varphi_j(X_i) + \nu_i. \tag{8.14}$$

对模型(8.14)进行最小二乘估计, 得到

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y}, \tag{8.15}$$

其中$\boldsymbol{Z} = (\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_m)$, $\boldsymbol{Z}_i = (\varphi_i(X_1), \cdots, \varphi_i(X_n))^{\mathrm{T}}$. 于是, $m(x)$有正交序列估计：

$$\hat{m}_n(x) = \boldsymbol{z}(x)^{\mathrm{T}}\hat{\boldsymbol{\theta}}, \tag{8.16}$$

其中$\boldsymbol{z}(x)=(\varphi_1(x),\cdots,\varphi_m(x))^{\mathrm{T}}$.

设解释变量为确定性变量. 记$\nu(x)=\sigma_u^2(\boldsymbol{z}(x)^{\mathrm{T}}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z})^{-1}\boldsymbol{z}(x))$, 则当$n\to\infty$, $m\to\infty$时, 正交序列估计有如下性质：

① $\nu(x)^{-1/2}(\hat{m}_n(x)-E\hat{m}_n(x))\xrightarrow{\mathcal{L}}N(0,1)$;

② $\nu(x)^{-1/2}(E\hat{m}_n(x)-m)\to 0$;

③ $\hat{\sigma}_u^2=n^{-1}\sum\limits_{i=1}^{n}(Y_i-\hat{m}_n(X_i))^2$是$\sigma_u^2$的一个一致估计.

区间$[-1,1]$上Legendre 多项式正交基为

$$\begin{aligned}
\mathrm{P}_0(x)&=1/\sqrt{2},\\
\mathrm{P}_1(x)&=x/\sqrt{2/3},\\
\mathrm{P}_2(x)&=\frac{1}{2}(3x^2-1)/\sqrt{2/5},\\
\mathrm{P}_3(x)&=\frac{1}{2}(5x^3-3x)/\sqrt{2/7},\\
\mathrm{P}_4(x)&=\frac{1}{8}(35x^4-30x^2+3)/\sqrt{2/9},\\
\mathrm{P}_5(x)&=\frac{1}{8}(63x^5-70x^3+15x)/\sqrt{2/11}.
\end{aligned}$$

其他高阶Legendre 多项式可由下式递推地推出：

$$(m+1)\mathrm{P}_{m+1}(x)=(2m+1)x\mathrm{P}_m(x)-m\mathrm{P}_{m-1}(x). \tag{8.17}$$

Legendre 多项式正交基$\{\mathrm{P}_j(x)\}_{j=0}^{\infty}$满足

$$\int_{-1}^{1}\mathrm{P}_i(x)\mathrm{P}_j(x)\mathrm{d}x=\begin{cases}0, & i\neq j,\\ 1, & i=j.\end{cases}$$

例8.6 图8.8给出了前六个Legendre 多项式的图像.

例8.7 图8.9是对摩托车数据采用Legendre多项式正交基进行正交序列估计拟合效果图. 若解释变量X在区间$[a,b]$上取值, 则必须作变量替换$Z=\dfrac{2X-a-b}{b-a}$, 使得变量Z的取值区间为[−1,1].

§8.6 罚最小二乘法

考虑在普通最小二乘问题中, 求函数m使得

$$\sum_{i=1}^{n}[Y_i-m(X_i)]^2 \tag{8.18}$$

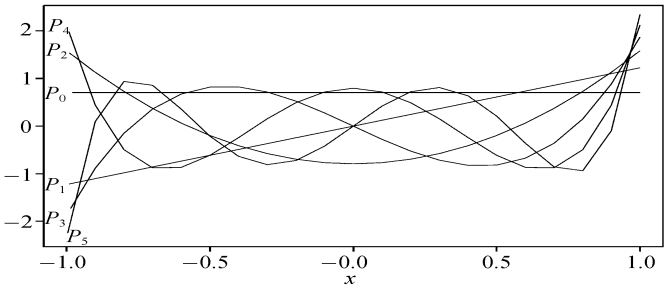

图 8.8 Legendre 多项式的函数图

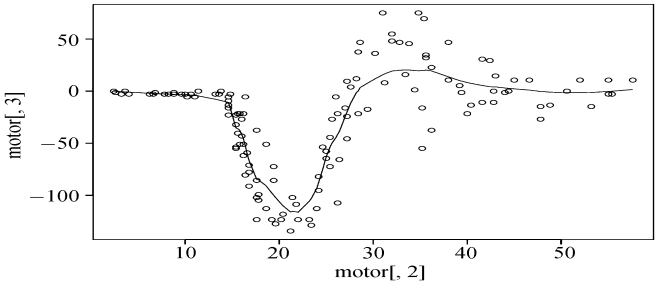

图 8.9 Legendre 多项式正交函数拟合摩托车数据效果图

达到最小, 该问题有无穷多解. 比如：通过所有观察点的折线和通过所有观察点的任意阶多项式光滑曲线都是解. 但这些解没有应用价值, 它们的残差全为0, 虽然完整地拟合了数据, 但是模型的泛化能力和预测效果都很差, 随机误差项产生的噪声没有在模型中得到体现, 这样的问题称为“过度拟合”现象. 因而这些解并非我们真正需要的. 为了寻求既可排除随机误差项产生的噪声, 又使得解具有一定的光滑性(二阶导数连续), 罚方法是控制墨西哥不至过于复杂的一种选择, 其中较为有代表性的是二次罚,叫做罚最小二乘法，它的原理是使

$$\sum_{i=1}^{n}(Y_i - m(X_i))^2 + \lambda \int_0^1 (m''(x))^2 \mathrm{d}x \tag{8.19}$$

达到最小的解$\hat{m}_{n,\lambda}(\cdot)$，其中$\lambda > 0$. 式中, λ称为罚（penalty）参数.

该问题有唯一解, 它的解可以表达为Y_i的的线性组合，其中$i = 1, 2, \cdots, n$. 由于求解过程复杂且解也没有显示表达式, 因而这里省略其求解过程.

而通过所有观察点的折线所对应的模型, 虽然使得式(8.19)第一项平方和为零, 但它不满足光滑性; 对于直线模型, 式(8.19)第二项为零,但却会使得式(8.19)的第一项平方和过大. 因而, 罚最小二乘法实际上是在最小二乘法和解的光滑性之间的平衡. 式(8.19)的第二项实际上就是对第一项平方和过小的一个罚系数, 也称为光滑系数. 罚最小二乘法的光滑系数λ可以人为确定, 并不是对每一个λ, 罚最小二乘法的解都能够充分排除随机误差项产生的噪声. 当$\lambda = 0$时, 通过所有观察点的任意高方差的曲线没有意义; 当$\lambda = +\infty$时, 直线解也没有意义. 最优的光滑系数应该界于0和$+\infty$ 之间. 应该说, 非参数回归模型的罚最小二乘估计的估计效果完全取决于λ的选择. 最佳的平滑参数一般采用如下的广义交叉验证法确定. 在实际应用中, 需要不断调整λ, 直到找到满意解为止.

例8.8 本例是对摩托车数据进行的罚最小二乘估计的效果图, 如图8.10所示. 左图显示的是$\lambda = 10$时的拟合效果, 从图上看出, 采用较大的λ, 拟合效果不好; 右图显示的是$\lambda = 3$时的拟合效果, 从图上看出, 采用较小的λ, 拟合效果较好.

值得一提的是, 罚方法不仅用于直接对函数部分进行惩罚, 更多的则是表现在系数求罚上, 从而也使得罚方法成为模型选择的重要组成部分.

§8.7 样条回归

§8.7.1 模型

在正交序列回归中, 我们假设$\varphi_j(t)$是正交的. 在样条回归中, 我们不做这样的要求, 因此我们可以选择更多可能的基函数. 我们希望通过减少要求的条件, 得到

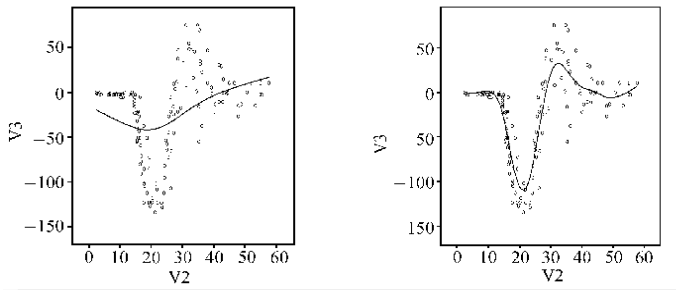

图 8.10 罚最小二乘拟合的摩托车数据拟合结果

更好的拟合效果.

假设我们观测到如下n组数据$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, 其中$x_i \in [a, b]$。在很多情况下，我们并不知道$(x_i, y_i)$满足什么关系，在这种情况下，我们假设$(x_i, y_i)$满足如下关系

$$y_i = f(x_i) + \varepsilon_i, i = 1, 2, \ldots, n,$$

其中$f(x)$是关于x的未知函数，ε_i是独立同分布的正态分布$N(0, \sigma^2)$. 在上述假设下，我们有$E(y) = f(x)$。

对于未知的函数$f(x)$，我们采用样条基函数去估计，这里我们以线性样条基函数来介绍样条回归模型。首先介绍线性样条基函数。对于$x \in [a, b]$, x的线性样条基函数定义为

$$1, x, (x - \kappa_1)_+, (x - \kappa_2)_+, \ldots, (x - \kappa_K)_+$$

这里$\kappa_j \in [a, b]$称为结点。我们可以采用上述样条基函数去逼近$f(x)$，即

$$f(x) \approx \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} b_k (x_i - \kappa_k)_+$$

在本节后面的部分，我们假设存在一组基函数，使得$f(x) = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} b_k (x_i - \kappa_k)_+$。当然，事实上，等号一般是不能取到的，但如果差别足够小，我们可以认为上述的假设是合理的。

定义8.3 一个样条模型(Spline Model)可以写成

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} b_k (x_i - \kappa_k)_+ + \varepsilon_i, i = 1, \ldots, n. \tag{8.20}$$

我们引入以下的记号，$\boldsymbol{y}=(y_1,\ldots,y_n)^T$代表观测到的因变量，设计矩阵为

$$\boldsymbol{X}=\begin{pmatrix} 1 & x_1 & (x_1-\kappa_1)_+ & (x_1-\kappa_2)_+ & \ldots & (x_1-\kappa_K)_+ \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & (x_n-\kappa_1)_+ & (x_n-\kappa_2)_+ & \ldots & (x_n-\kappa_K)_+ \end{pmatrix}$$

和多元线性回归类似，参数$(\beta_0,\beta_1,b_1,b_2,\ldots,b_K)$的估计值为

$$\widehat{\boldsymbol{\beta}}=(\widehat{\beta}_0,\widehat{\beta}_1,\widehat{b}_1,\widehat{b}_2,\ldots,\widehat{b}_K)^T=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

$f(x)$的估计值为$\widehat{f}(x)=\widehat{\beta}_0+\widehat{\beta}_1x_i+\sum_{k=1}^K\widehat{b}_k(x_i-\kappa_k)_+$. (Ruppert D. et.al 2003)

§8.7.2 样条回归模型的节点

对于样条回归模型，一个重要的问题是如何选择节点(knot)。节点的选择有如下两个方法，第一个方法是根据点的疏密程度人为地选择。基本原则是如果x_i比较均匀的分布在区间$[a,b]$ 上，我们可以取等距的节点。如果x_i 在有些区域比较密，我们可以在该区域上多取一些节点。上述的方法比较的主观，另一个方法则是把样条基函数看成多元线性模型中的自变量，然后通过常用的模型选择的方法，例如AIC规则。

除了对节点进行选择外，我们还可以通过控制这些节点的影响。即在$\boldsymbol{\beta}^T\boldsymbol{D}\boldsymbol{\beta}\leqslant C$ 条件下，最小化如下公式

$$\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2, \tag{8.21}$$

这里$\boldsymbol{\beta}=(\beta_0,\beta_1,b_1,\ldots,b_K)$, $D=\begin{pmatrix}\mathcal{O}_{2\times2} & \mathcal{O}_{2\times K}\\ \mathcal{O}_{K\times2} & \mathcal{I}_K\end{pmatrix}$。其中$\mathcal{O}_{m\times n}$是$m\times n$阶的零矩阵，$\mathcal{I}_K$是$K$阶单位矩阵。

类似于岭回归，上述问题可以等价地转化为如下的最小化问题

$$\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2+\lambda\boldsymbol{\beta}^T\boldsymbol{D}\boldsymbol{\beta}$$

观察上述公式，容易看出来$\boldsymbol{\beta}^T\boldsymbol{D}\boldsymbol{\beta}=\sum_{i=1}^K b_i^2$, 因此可以看到我们只是对带有节点的基函数$(x-\kappa_1)_+,(x-\kappa_2)_+,\ldots,(x-\kappa_K)_+$进行了限制，对没有节点的基函数$1,x$没有限制。

对于上述问题，参数$(\beta_0,\beta_1,b_1,b_2,\ldots,b_K)$的估计值为

$$\widehat{\boldsymbol{\beta}}=(\widehat{\beta}_0,\widehat{\beta}_1,\widehat{b}_1,\widehat{b}_2,\ldots,\widehat{b}_K)^T=(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{D})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$f(x)$的估计值为

$$\widehat{f}(x)=\widehat{\beta}_0+\widehat{\beta}_1x_i+\sum_{k=1}^K\widehat{b}_k(x_i-\kappa_k)_+.$$

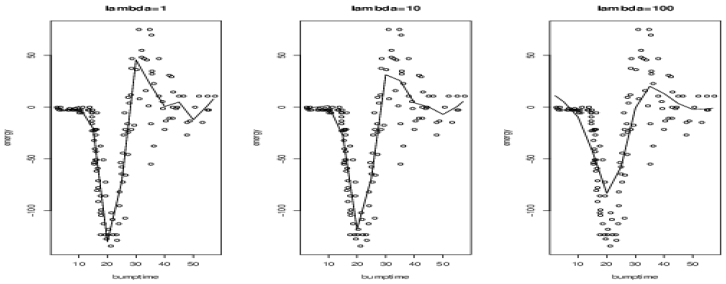

图 8.11 不同λ对估计的影响

如图8.11所示，我们把样条回归模型对motor数据进行分析，我们三个不同的λ值，即$\lambda = 1, 10, 100$. 我们可以看到，λ比较小时，估计值波动比较多，随着λ的增大，估计值逐渐光滑，但是当λ过大时，估计值会出现较大偏差。

§8.7.3 常用的样条基函数

上面的线性样条基函数在节点处不光滑(不可导)。为了克服这个缺点，我们可以采用二次样条基函数(quadratic spline basis functions)，

$$1, x, x^2, (x-\kappa_1)^2, (x-\kappa_2)^2, \ldots, (x-\kappa_K)^2,$$

我们可以看到，二次样条基函数在节点处是可导的。

我们也可以扩张线性样条基函数, 引入p阶截断样条基函数(Truncated power basis of degree p)，即

$$1, x, \ldots, x^p, (x-\kappa_1)_+^p, (x-\kappa_2)_+^p, \ldots, (x-\kappa_K)_+^p$$

容易看到，当$p = 1$时，截断样条基函数即为线性样条基函数。当$p \geqslant 2$时, 截断样条基函数在节点处是可导的。

另一类常用的样条基函数称为B-样条基函数(B-spline basis functions)。B-样条基函数是通过递推公式来定义，0 阶B- 样条基函数定义为

$$B_{j,0}(x) = I(\kappa_j \leqslant x < \kappa_{j+1})$$

这里的$I(\cdot)$是示性函数。p阶B-样条基函数通过如下递推公式定义，

$$B_{i,p} = \frac{x-\kappa_i}{\kappa_{i+p-1}-\kappa_i} B_{i,p-1}(x) + \frac{\kappa_{i+p}-x}{\kappa_{i+p}-\kappa_{i+}} B_{i+1,p-1}(x)$$

一阶B-样条基函数

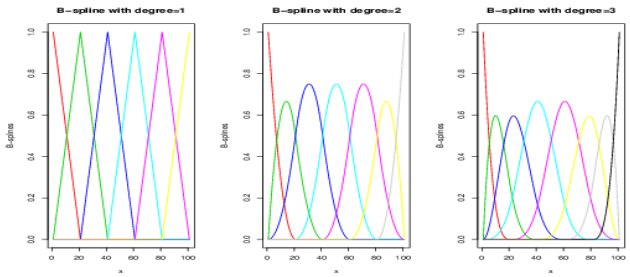

图 8.12 1,2,3阶的B-样条基函数.

§8.7.4 样条模型自由度

这里我们来看误差的自由度. 对于样本模型

$$y_i = f(x_i) + \varepsilon_i, i = 1, 2, \ldots, n,$$

通过最小罚二乘法, 我们知道参数的估计值为$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{D})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, $f(x)$的估计值为

$$\widehat{\boldsymbol{y}} = \boldsymbol{S}_\lambda\boldsymbol{y},$$

其中$\boldsymbol{S}_\lambda = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{D})^{-1}\boldsymbol{X}^T$.

这里误差的自由度定义为

$$df_{res} = n - 2\mathrm{tr}(\boldsymbol{S}_\lambda) + \mathrm{tr}(\boldsymbol{S}_\lambda\boldsymbol{S}_\lambda^T).$$

令残差平方和$SSE = (\widehat{\boldsymbol{y}} - \boldsymbol{y})^T(\widehat{\boldsymbol{y}} - \boldsymbol{y})$, 通过计算我们可以知道

$$\begin{aligned} E(SSE) &= E\{(\widehat{\boldsymbol{y}} - \boldsymbol{y})^T(\widehat{\boldsymbol{y}} - \boldsymbol{y})\} \\ &= E\{\boldsymbol{y}^T(\boldsymbol{S}_\lambda - \boldsymbol{I})^T(\boldsymbol{S}_\lambda - \boldsymbol{I})\boldsymbol{y}\} \\ &= \boldsymbol{y}^T(\boldsymbol{S}_\lambda - \boldsymbol{I})^T(\boldsymbol{S}_\lambda - \boldsymbol{I})\boldsymbol{y} + \sigma^2\mathrm{tr}\{(\boldsymbol{S}_\lambda - \boldsymbol{I})^T(\boldsymbol{S}_\lambda - \boldsymbol{I})\} \\ &= \boldsymbol{y}^T(\boldsymbol{S}_\lambda - \boldsymbol{I})^T(\boldsymbol{S}_\lambda - \boldsymbol{I})\boldsymbol{y} + \sigma^2 df_{res} \end{aligned}$$

上面我们用到如下性质: 对于任意随机向量$\boldsymbol{v}$和对称矩阵$\boldsymbol{A}$, 我们有$E(\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v}) = E(\boldsymbol{v})^T\boldsymbol{A}E(\boldsymbol{v}) + \mathrm{tr}\{\boldsymbol{A}\mathrm{Cov}(\boldsymbol{v})\}$.

如果$\boldsymbol{y}^T(\boldsymbol{S}_\lambda - \boldsymbol{I})^T(\boldsymbol{S}_\lambda - \boldsymbol{I})\boldsymbol{y}$比较小, 那么SSE/$df_{res}$是对$\sigma^2$的一个估计. 我们可以把上面的结果和参数线性模型进行比较, 在线性模型中, $\boldsymbol{S}_\lambda$对应的是$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, 并且$\boldsymbol{H}\boldsymbol{H}^T = \boldsymbol{H}$. 在$df_{res}$的定义中, 用$\boldsymbol{H}$ 代替$\boldsymbol{S}_\lambda$, 我们有

$$df_{res} = n - 2\mathrm{tr}(H) + \mathrm{tr}(\boldsymbol{H}\boldsymbol{H}^T) = n - \mathrm{tr}(H) = n - p.$$

因此df_{res}可以看成是对线性模型中误差自由度的推广.

习题

8.1 令$u_i \sim N(0, 0.025), i = 1, 2, \cdots, 300$, 令$X_i = i/300$, 则$X_i \in [0, 1]$. 模拟产生如下数据：$Y_i = \sin(2\exp(X_i + 1)) + u_i$, 尝试R中所有可能的核函数估计X与Y的函数曲线.

8.2 对于Nadaraya-Watson核估计$\widehat{m}_n(x) = \sum_{i=1}^{n} w_i(x)Y_i$. 在给定的点$x$, 假设$Y_1, \ldots, Y_n$满足独立同分布$N(m(x), \sigma^2)$, 计算$E(\widehat{m}_n(x))$和$\text{var}(\widehat{m}_n(x))$.

8.3 令$X_i \sim N(0, 1), u_i \sim N(0, 0.025X_i^2), i = 1, 2, \cdots, 300$为相互独立的变量. 模拟产生如下数据：$Y_i = \exp(|X_i|) + u_i$, 用局部线性和局部二项式估计$X$与$Y$的函数曲线.

8.4 用求导的方法最小化(8.6), 写出具体步骤并给出$a(x)$和$b(x)$的估计公式.

8.5 数据见教学资源文件Indchina.txt, 记Y_t=居民消费价格指数, X_t=商品进出口额(亿美元). 本文采用1993年4月到1998年11月68个月的月度资料, 应用LOWESS稳健估计方法对通货膨胀与进出口的关系进行非参数回归模型估计.

8.6 产生B-样条基函数, 定义域为$[0, 100]$,节点为$0, 20, 50, 90, 100$. 写出B-样条基函数在$d = 0, 1, 2$的形式.

8.7 用B-样条基函数拟合摩托车数据(见教学资源数据motor.txt). 注明所用的节点,d和λ.

8.8 本题中使用波士顿（Boston）数据中变量到波士顿五个就业中心的加权平均距离(dis)和每十万分之一的氮氧化物颗粒浓度(nox)。将加权平均距离(dis)作为预测变量，氮氧化物颗粒浓度(nox)作为响应变量。

a. 用poly()函数对加权平均距离(dis)和氮氧化物颗粒浓度(nox)拟合三次多项式回归模型，并输出回归结果并画出数据点及拟合曲线。

b. 选择阶数从从1到10的多项式模型的拟合结果，绘制相应的残差平方和曲线。

c. 运用交叉验证或者其他方法选择合适的多项式模型的阶数并解释结果。

d. 用bs()函数对加权平均距离(dis)和氮氧化物颗粒浓度(nox)拟合回归样条，输出自由度为4时的拟合结果，说明选择结点时使用了什么准则，最后绘制出拟合曲线.

e. 尝试不同的自由度拟合回归样条，绘制拟合曲线图和相应的RSS，并解释结果。

f. 运用交叉验证或者其他方法选择合适的回归样条模型的自由度并解释你的结果。

案例与讨论:排放物成分与燃料-空气当量比（EquivRatio）和发动机的压缩比（CompRatio）

案例背景

随着城市汽车保有量的增加，汽车尾气排放对环境的影响越来越大。节能降耗、降低汽车尾气排放，减少对大气的污染，已成为当今社会亟待解决的问题。通过改进发动机使用清洁燃料，能够有效控制汽车尾气的排放量。尾气排放受发动机压缩比技术性能和燃料空气当量比影响。发动机压缩比指混合气提压缩程度，高压缩比发动机可输出较大的动能，但较大压缩比发动机高温时，在中高负荷中出现高温轻微爆燃现象，就会导致NOx排放的增加。另一方面，发动机的燃料空气当量比

也影响发动机的动力性能和尾气排放。燃料空气当量比是发动机空燃比的重要组成部分，用于测量汽油与空气混合燃烧时，发动机进气冲程中吸入气缸的燃料（汽油）重量与空气的重量之比，燃料与混合气中的空气的比例在1附近，对应着空气量多或者少时空气都不能完全燃烧，造成燃烧效率低下，从而产生较多的尾气，污染环境。因此，研究发动机尾气排放量与发动机压缩比和燃料空气当量比之间的关系对于检测车辆尾气超标情况，推动清洁能源使用，设计环保尾气过滤装置以及倡导绿色出行都有积极意义。

数据描述

例中ethanol数据集所用的排放物数据来自于一项以纯乙醇作为单缸发动机的燃料的调查研究（Brinkman,1981）。1. ethanol数据集共有88个样本；2. 2个连续数值型自变量，CompRatio表示发动机压缩比，EquivRatio表示燃料空气当量比，NOx表示氮氧化物，主要成分有一氧化碳(CO)、碳氢化合物(HC)等以及微粒污染物(或称颗粒污染物)在大城市的许多空气质量监测点NOx已成为左右空气污染指数的首要污染物。3. 无缺失值。

研讨题目

1.请查阅文献解释发动机压缩比对NOx排放量的边际影响;

2.请查阅文献解释燃料空气当量比对NOx排放量的边际影响;

3.请结合下列提示的数据分析提示分析燃料空气当量比与发动机压缩比对NOx排放量的影响;

4.请调整局部多项式覆盖邻域的数据比例（在R中用span控制）分析.

数据分析提示

排放物成分取决于两个预测变量，燃料-空气当量比（EquivRatio）和发动机的压缩比（CompRatio10）。首先给出排放物的密度直方图：

```
## first we read in the data
  ethanol=read.csv("e:\\data\\ethanol.csv",header=T,sep=",")
  #density histogram and add density curve
  hist(ethanol$NOx,freq=FALSE,breaks=15)
  lines(density(ethanol$NOx))

  library(locfit)
  plot(NOx~CompRatio,data=ethanol)

  ## local polynomial regression of NOx on the equivalence ratio
```

```
## fit with a 50% nearest neighbor bandwidth.
fit <- locfit(NOx~lp(EquivRatio,nn=0.5),data=ethanol)
plot(EquivRatio,NOx,data,data=ethanol)
lines(fit)
```