

第5章 局部多项式回归

刘柏森

2014年10月11日

1 引言

给定一组观测数据 $(x_1, y_1), \dots, (x_n, y_n)$, 我们往往想研究**响应变量** Y 和**协变量** X 之间的关系, 这个关系可以用一个函数来刻画, 我们可以考虑如下模型:

$$y_i = g(x_i) + \epsilon_i, \quad E(\epsilon_i) = 0, \quad i = 1, \dots, n, \quad (1)$$

这里 $g(\cdot)$ 称为**回归函数**. 那么在一些假设条件下, 如何基于样本 $(X_1, Y_1), \dots, (X_n, Y_n)$ 去估计函数 $g(x)$ 呢? 这就是本章的主要研究内容. 首先我们不妨先看两个例子:

例5.1 (motorcycle data): 图1(左)展示的是一个关于摩托车碰撞试验的数据集的散点图, 该数据集来源于R程序包**MASS**. 此数据集由一系列在模拟摩托车事故中头部加速度的观测值构成, 目的主要是用来检测头盔的性能. 由散点图很容易看出: **加速度和时间之间存在显著的非线性关系**. 那么如何建模? 应该采用哪个回归函数?

例5.2 (Ethanol data): 图1(右)展示的是ethanol数据集的散点图, 该数据集来源于R程序包**locfit**. 此数据集是用来研究发动机排放气体中的nitric oxides的浓度和equivalence ratio(一种空气和ethanol混合程度的度量)的关

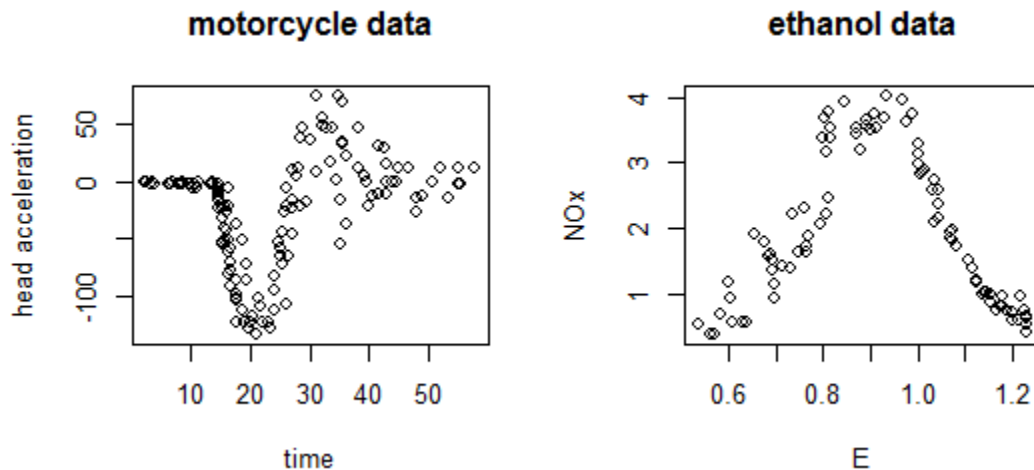


Figure 1: motorcycle 数据(左) 和ethanol 数据(右)的散点图.

系。响应变量是 NOx ，协变量是 E (发动机的equivalence ratio)和 C (压缩比例). 显然, NOx 和 E 之间存在显著的非线性关系。

传统上，我们往往采用多项式回归的trial-and-error方法.

2 线性模型和非线性模型的回顾

2.1 线性模型

线性模型是最经典、最常用的一种模型，其内容如下:

$$Y_i = \sum_{j=1}^p x_{ij}\beta_j + \epsilon_i, \quad i = 1, \dots, n, \quad (2)$$

这里 $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$. 线性模型里一般都包含截距，因而我们不妨设 $x_{i1} = 1, i = 1, \dots, n$.

定义一个设计矩阵

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

令 $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$. 模型(2)可改写为:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \quad (3)$$

则, 最小二乘估计 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T$ 使得如下残差平方和达到最小:

$$\text{RSS} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p x_{ij}\beta_j \right)^2.$$

若矩阵 $(\mathbf{X}^T\mathbf{X})^{-1}$ 存在, 则最小二乘估计为:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \quad (4)$$

从而函数 $g(x)$ 在点 $x = (x_1, \dots, x_p)^T$ 的估计值为:

$$\hat{g}(x) = \sum_{j=1}^p \hat{\beta}_j x_j = \mathbf{x}^T \hat{\boldsymbol{\beta}}.$$

此外, 拟合值 $\hat{\mathbf{g}} = (\hat{g}(\mathbf{x}_1), \dots, \hat{g}(\mathbf{x}_n))^T$ 为:

$$\hat{\mathbf{g}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{S}\mathbf{Y},$$

这里

$$\mathbf{S} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T,$$

称为**hat** 矩阵. 向量 $\hat{\epsilon} = \mathbf{Y} - \mathbf{g}$ 的元素称为**残差**. 需要指出的是: hat 矩阵是对称的(即 $\mathbf{S} = \mathbf{S}^T$) 且幂等的(即 $\mathbf{S}^2 = \mathbf{S}$). 参数 p 等于矩阵 \mathbf{S} 的迹, 即:

$$p = \text{tr}(\mathbf{S}).$$

σ^2 的无偏估计为

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{g}(x_i))^2}{n - p} = \frac{\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}}{n - p}.$$

2.2 非线性回归

非线性回归通常假定函数 $g(\cdot)$ 为一含有未知参数 $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ 的非线性函数,函数形式已知. 因而非线性模型为:

$$Y_i = g(\mathbf{x}_i; \boldsymbol{\beta}) + \epsilon_i, \quad i = 1, \dots, n,$$

其中 $E(\epsilon_i) = 0$, $\text{Var}(\epsilon_i) = \sigma^2$.

则**最小二乘估计** $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T$ 使得如下**残差平方和**达到最小:

$$\text{RSS} = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p g(x_{ij}; \beta_j) \right)^2.$$

一般地, 最小二乘估计 $\hat{\boldsymbol{\beta}}$ 没有显示表达. 然而我们可以用**加权最小二乘算法**来获得 $\hat{\boldsymbol{\beta}}$. 更多知识请参考Ritz and Streibig (2008).

例5.3: 我们用多项式回归来拟合motorcycle数据集, 结果见图2。显然, 所有拟合结果都存在较大偏差.

例5.4: 同样, 我们用多项式回归模型来拟合ethanol数据集, 结果见图3。显然, 所有拟合结果也都存在较大偏差。

可否考虑非线性回归模型?

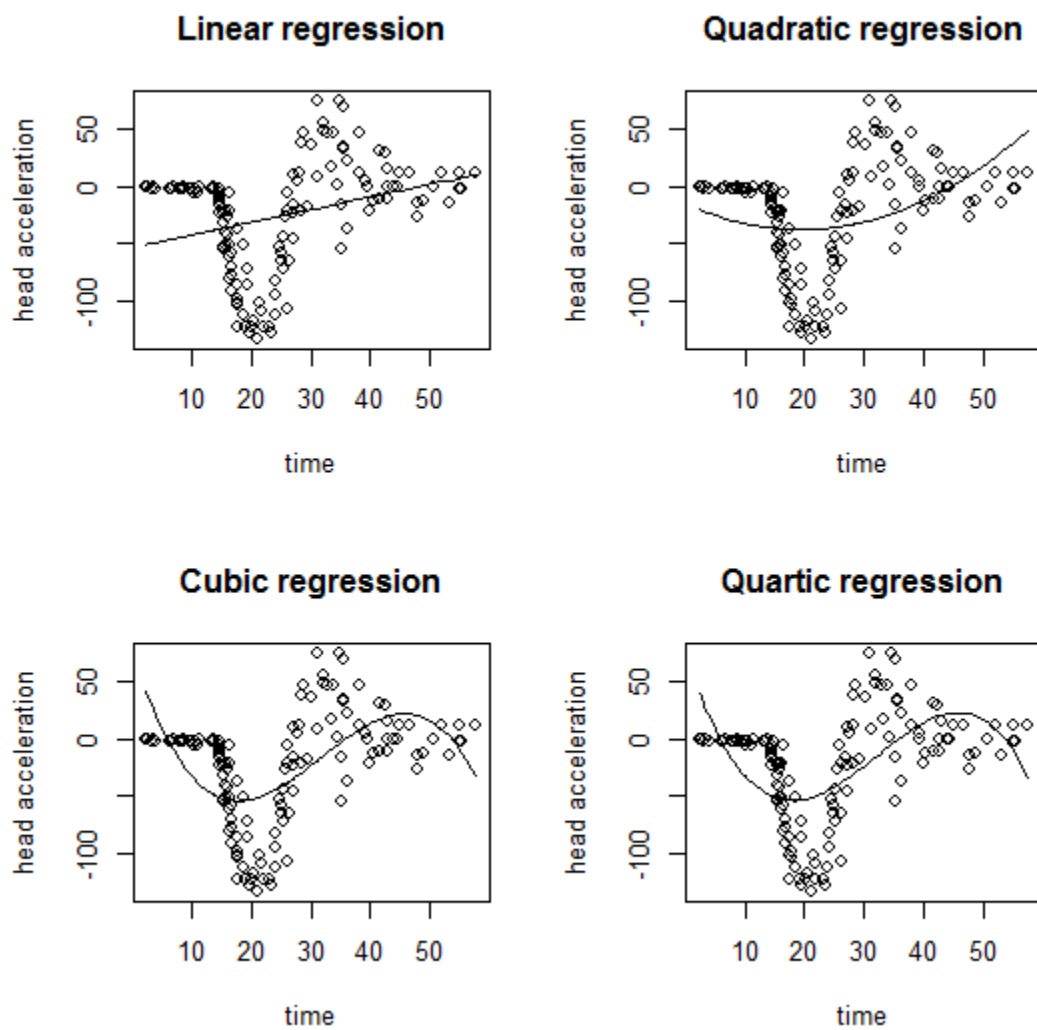


Figure 2: 多项式回归拟合motorcycle 数据.

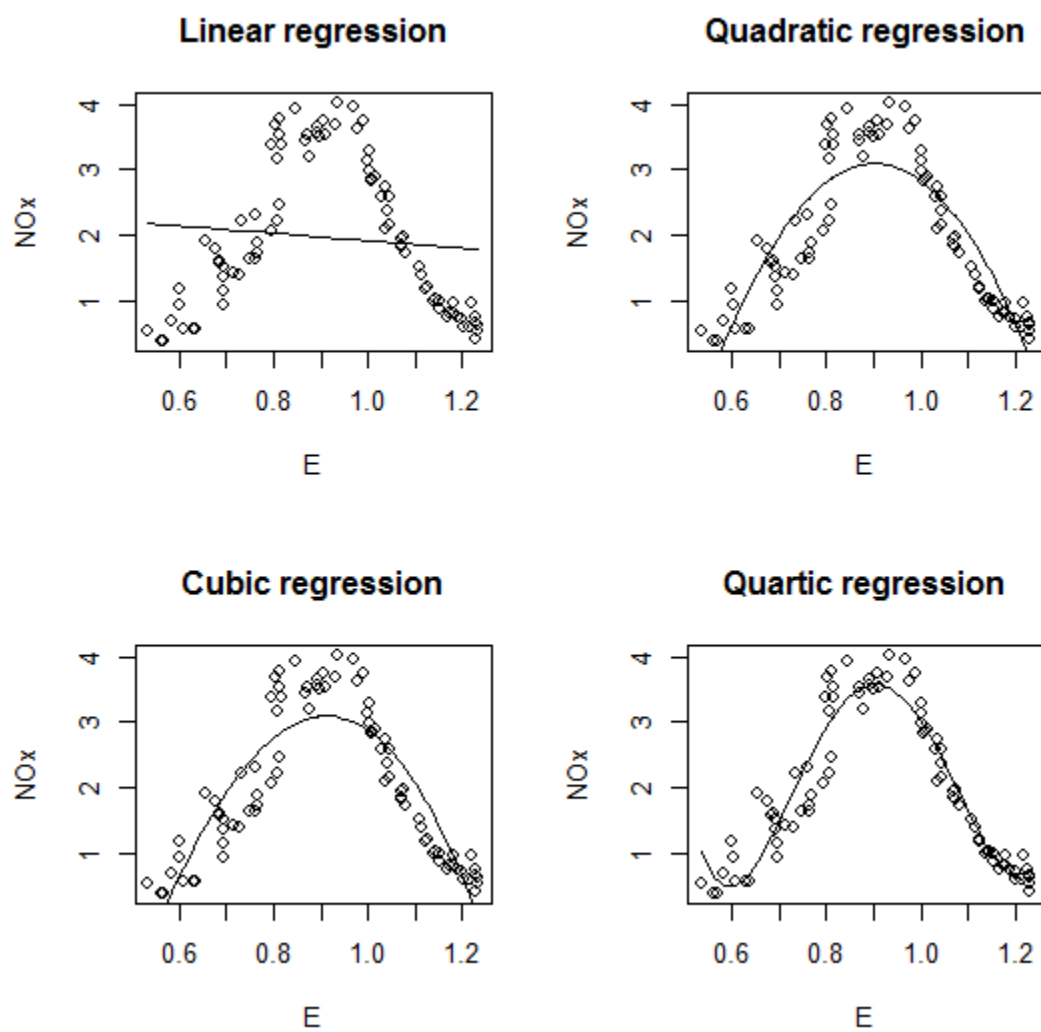


Figure 3: 多项式回归拟合ethanol 数据.

3 非参数回归

Li and Racine (2006, Page 59)证明了如下结论:

定理5.1 : 假定 $E(Y^2)$ 存在, 则对于任意连续函数 $g(x)$, 在平方损失准则下, 我们有:

$$E\{[Y - g(X)]^2\} \geq E\{[Y - E(Y|X)]^2\},$$

或等价地

$$\min_g E\{[Y - g(X)]^2\} = E\{[Y - E(Y|X)]^2\}.$$

证明: 对于任意连续函数 $g(x)$, 易证

$$\begin{aligned} E\{[Y - g(X)]^2\} &= E\{[(Y - E(Y|X)) + (E(Y|X) - g(X))]^2\} \\ &= E\{[Y - E(Y|X)]^2\} + E\{[E(Y|X) - g(X)]^2\} \\ &\quad + 2E\{[Y - E(Y|X)][E(Y|X) - g(X)]\} \end{aligned} \quad (5)$$

利用下面两个等式: 对于任意随机变量 X 和 Y , 我们有:

$$E\{E(Y|X)\} = E(Y), \quad E\{g(X)Y|X\} = g(X)E(Y|X).$$

则, 我们有

$$\begin{aligned} E\{[Y - E(Y|X)][E(Y|X) - g(X)]\} &= E\{E\{[Y - E(Y|X)][E(Y|X) - g(X)]|X\}\} \\ &= E\{[E(Y|X) - g(X)]E\{[Y - E(Y|X)]|X\}\}. \end{aligned}$$

由于

$$E\{[Y - E(Y|X)]|X\} = E(Y|X) - E(Y|X) = 0.$$

因而(5)中第三项为0, 从而我们证明了:

$$E\{[Y - g(X)]^2\} \geq E\{[(Y - E(Y|X))]^2\}.$$

定理5.1说明: 在平方损失意义下, 对于所有的连续函数中, $E(Y|X)$ 是Y的最优预测函数。因而在模型(1)中, 我们只研究 $g(x) = E(Y|x)$. 通常假定是一个未知的光滑函数(即3阶导数存在且连续).

假定二元数据 $(X_i, Y_i), i = 1, \dots, n$, 来自于下面的模型:

$$Y_i = g(X_i) + \sigma(X_i)\epsilon_i, i = 1, \dots, n, \quad (6)$$

其中 $E(\epsilon_i) = 0, Var(\epsilon_i) = 1$, 且 $X_i, i = 1, \dots, n$ 与 $\epsilon_i, i = 1, \dots, n$ 相互独立.

假设 $g(x)$ 在点 x_0 处存在 $(p+1)$ 阶导数. 我们要研究的内容是: 如何估计回归函数

$$g(x_0) = E(Y|X = x_0),$$

及其导数 $g'(x_0), g''(x_0), \dots, g^{(p)}(x_0)$.

3.1 Nadaraya-Watson核估计

由定理5.1知:

$$\begin{aligned} g(x) &= E(Y|X=x) \\ &= \int y f(y|x) dy \\ &= \int y \frac{f(x,y)}{f_X(x)} dy, \end{aligned} \tag{7}$$

这里 $f_X(x)$, $f(x,y)$, $f(y|x)$ 分别为 X 的边际密度, (X,Y) 的联合密度, 以及给定 $X=x$ 下, Y 的条件密度.

回忆一下密度函数 $f(x,y)$ 和 $f(x)$ 的光滑核估计分别由下式给出:

$$\begin{aligned} \hat{f}(x,y) &= \frac{1}{nh_x h_y} \sum_{i=1}^n K_x \left(\frac{x-X_i}{h_x} \right) K_y \left(\frac{y-Y_i}{h_y} \right), \\ \hat{f}(x) &= \frac{1}{nh_x} \sum_{i=1}^n K_x \left(\frac{x-X_i}{h_x} \right). \end{aligned}$$

将上述结果代入(7), 注意到核函数 $K(\cdot)$ 满足 $\int K_y(u) du = 1$ 且 $\int u K_y(u) du = 0$, 则Nadaraya-Watson(N-W)核估计定义为:

$$\hat{g}_{NW}(x) = \frac{\sum_{i=1}^n K \left(\frac{x-X_i}{h} \right) Y_i}{\sum_{i=1}^n K \left(\frac{x-X_i}{h} \right)} = \sum_{i=1}^n w_i Y_i, \tag{8}$$

其中

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

为权函数。因而, Nadaraya-Watson核估计实际上是因变量观察数据 Y_i 的加权平均。

3.2 最邻近(nearest-neighbor)估计

最邻近估计也叫 k -NN 估计, 也是因变量 Y 在 x 邻域内的一个加权平均, 只不过其邻域大小是变化的。更具体地说, 我们用离 x 最近的那 k 个点的 Y_i 的平均来估计 $g(x)$. 因而, k -NN 估计可写为

$$\hat{g}_k(x) = \frac{1}{n} \sum_{i=1}^n w_{ki}(x) Y_i, \quad (9)$$

这里的权函数 $\{w_{k1}(x), \dots, w_{kn}(x)\}$ 定义为

$$w_{ki}(x) = \begin{cases} n/k, & \text{如果 } i \in J_x, \\ 0 & \text{否则} \end{cases}$$

其中

$$J_x = \{i : X_i \text{属于距离 } x \text{ 最近的前 } k \text{ 个点}\}.$$

在具体操作时, 我们可以将 $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ 按照 X_i 到某个点 x 的距离从小到大的顺序进行排序, 不妨设排序后的 X_i 为:

$$|X_{(1)} - x| \leq |X_{(2)} - x| \leq \dots \leq |X_{(n)} - x|,$$

则 k -NN核估计为

$$\hat{g}_k(x) = \frac{1}{k} \sum_{j=1}^k Y_{(j)}.$$

事实上, k -NN 可以看作核函数为均匀密度函数 $K(u) = \frac{1}{2}I(|u| \leq 1)$ 的核估计, 其窗体宽度 $h = h(k)$ 为可变的, 即:

$$\hat{g}_k(x) = \frac{\sum_{i=1}^n K_h(x - X_i) Y_i}{\sum_{i=1}^n K_h(x - X_i)}.$$

3.3 局部多项式回归

3.3.1 局部线性回归

局部多项式回归的主要思想: 在某个点 x 附近, 我们用一个多项式函数(例如线性函数)来逼近未知的光滑函数 $g(u)$. 令 h 表示局部邻域的大小。对于任意给定一点 x_0 , 在其小邻域内, 我们做如下逼近

$$g(x) \approx a(x_0) + b(x_0)(x - x_0), \quad \forall x \in (x_0 - h, x_0 + h],$$

这里 $a(x_0)$ 和 $b(x_0)$ 是两个依赖于 x_0 的未知参数. 从而我们考虑如下模型:

$$Y_i = a(x_0) + b(x_0)X_i + \text{error}_i, \quad X_i \in (x_0 - h, x_0 + h].$$

于是我们就引出如下问题: 寻找参数值 $\hat{a}(x_0)$ 和 $\hat{b}(x_0)$, 使得如下准则达到最小:

$$\sum_{i=1}^n [Y_i - a(x_0) - b(x_0)X_i]^2 I\left(\frac{|X_i - x_0|}{h} \leq 1\right). \quad (10)$$

那么, 回归函数 $g(x)$ 在点 x_0 处的估计值为 $\hat{g}(x_0) = \hat{a}(x_0)$, 而 $\hat{g}'(x_0) = \hat{b}(x_0)$. 当 x_0 在一个感兴趣的区域内变动时, 我们就得到了一条回归曲线.

这里权函数 $I(|X_i - x_0|/h \leq 1)$ 是用来刻画观测值 X_i 对 x_0 的作用或者影响。显然, 上述准则中的权函数是不光滑的函数, 一种改进的方法是用一个光滑的权函数来替代。令 $K(\cdot)$ 为一非负单峰函数, 其支撑区间为 $[-1, 1]$ (我们这样做的目的是使得下面定义的准则(11)仅使用落在区间 $(x_0 - h, x_0 + h]$ 内的数据 X_i 及其对应的 Y_i)。则加权最小二乘问题转化为优化问题:

$$\sum_{i=1}^n [Y_i - a(x_0) - b(x_0)X_i]^2 K\left(\frac{X_i - x_0}{h}\right). \quad (11)$$

当 K 为 $[-1, 1]$ 上的均匀分布的密度函数时, (11) 转化为(10).

当(11) 中的 $b(x_0) = 0$ 时, 我们有

$$\min_{a(x_0)} \sum_{i=1}^n [Y_i - a(x_0)]^2 K \left(\frac{X_i - x_0}{h} \right). \quad (12)$$

令 $\hat{a}(x_0)$ 为(12) 的最优解, 易看出, $\hat{a}(x_0)$ 为我们刚刚讨论过的Nadaraya-Watson核估计, 因而

容易看出: (12)实际上是在点 x_0 的邻域内用一个常数 $a(x_0)$ 去逼近 $g(x)$ (或 Y), 或者说: (12) 只利用了 x_0 邻域内的那些 X_j 对应的 Y_j 的一个加权平均去估计 $g(x_0)$, 因而, N-W核估计也常常称为局部常数核估计。相比较而言, 局部线性回归方法不仅能提供 $g(x_0)$ 的估计, 还能提供其一阶导数 $g'(x_0) = \left. \frac{\partial g(x)}{\partial x} \right|_{x_0}$ 的估计。

令 $\hat{a}(x_0)$ 和 $\hat{b}(x_0)$ 为(11)的优化解, 可以证明: $\hat{a}(x_0)$ 为 $g(x_0)$ 的相合估计, $\hat{b}(x_0)$ 为 $g'(x_0)$ 的相合估计.

定义 $\beta = (a(x_0), b(x_0))^T$, $\mathbf{Y} = (Y_1, \dots, Y_n)^T$. 令 \mathbf{X} 为一具有如下形式的设计阵:

$$\mathbf{X} = \begin{pmatrix} 1 & x - X_1 \\ \vdots & \vdots \\ 1 & x - X_n \end{pmatrix}$$

此外, 定义一权矩阵:

$$\mathbf{W} = h^{-1} \text{diag} \left[K \left(\frac{x - X_1}{h} \right), \dots, K \left(\frac{x - X_n}{h} \right) \right]$$

则优化问题(11) 可改写为:

$$\min_{\beta} (\mathbf{Y} - \mathbf{X}\beta)^T \mathbf{W} (\mathbf{Y} - \mathbf{X}\beta). \quad (13)$$

若逆矩阵 $(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$ 存在, 则有

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y} \quad (14)$$

记拟合值向量为

$$\hat{\mathbf{g}} = (\hat{g}(\mathbf{x}_1), \dots, \hat{g}(\mathbf{x}_n))^T.$$

则

$$\hat{\mathbf{g}} = \mathbf{S} \mathbf{Y},$$

其中

$$\mathbf{S} = \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}, \quad (15)$$

被称为**smoothing matrix** 或者**hat matrix**, 其第 i 行元素为估计量 $\hat{g}(x_i)$ 中 Y_i 对应的权重.

定义

$$w_i = K_h(x - X_i) \{S_{n,2} - (x - X_i)S_{n,1}\},$$

其中 $S_{n,j} = \sum_{i=1}^n K_h(x - X_i)(x - X_i)^j, j = 1, 2$. 则 $g(x)$ 的局部线性核估计为:

$$\hat{g}(x) = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}. \quad (16)$$

图4为局部多项式回归方法的一个演示:

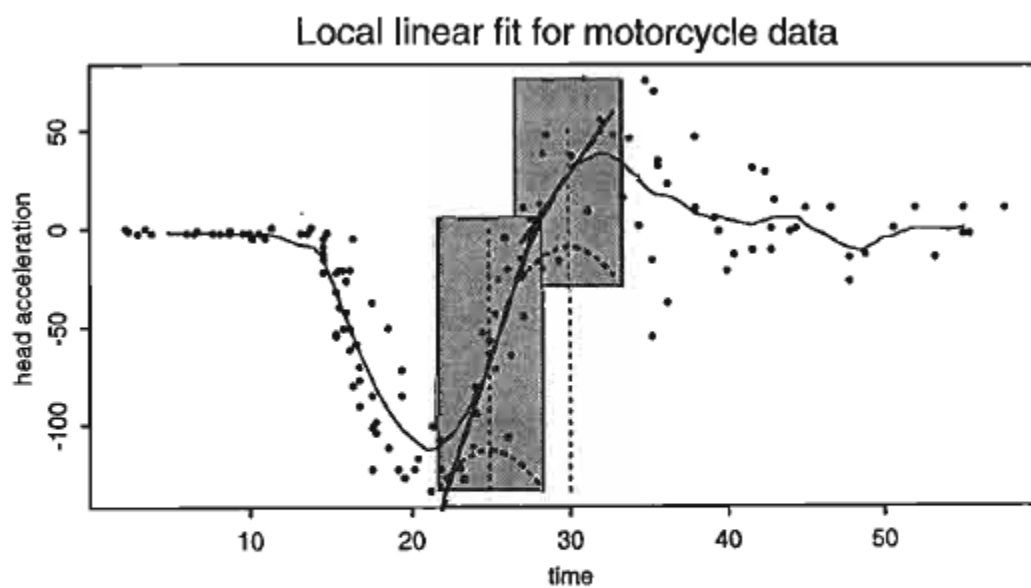


Figure 1.3. *Illustration of local modelling approach. For each given x , fit a linear model for the data points contained in the strip $x \pm 3.3$, using the weight function indicated at the bottom of the strip.*

Figure 4: 多项式回归拟合motorcycle 数据.

3.4 局部多项式回归

更一般地, 给定某个点 x_0 , 在其小邻域内, 我们可以用一个 p 阶多项式来局部逼近 $g(x)$. 由Taylor展开式, 我们有

$$\begin{aligned} g(x) &\approx g(x_0) + g'(x_0)(x-x_0) + \frac{g''(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{g^{(p)}(x_0)}{p!}(x-x_0)^p \\ &\equiv \beta_0 + \beta_1(x-x_0) + \cdots + \beta_p(x-x_0)^p, \end{aligned}$$

其中 $\beta_j = g^{(j)}(x_0)/j!$, $j = 0, 1, \dots, p$. 则局部多项式拟合问题转化为如下的加权最小二乘回归的优化问题:

$$\operatorname{argmin} \sum_{i=1}^n \left\{ Y_i - \sum_{j=0}^p \beta_j (x_0 - X_i)^j \right\}^2 K_h(x_0 - X_i), \quad (17)$$

定义设计矩阵

$$\mathbf{X} = \begin{pmatrix} 1 & (x_0 - X_1) & \cdots & (x_0 - X_1)^p \\ \vdots & \vdots & & \vdots \\ 1 & (x_0 - X_n) & \cdots & (x_0 - X_n)^p \end{pmatrix}$$

则 $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ 的最小二乘估计为:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}.$$

从而, $g^{(\nu)}(x_0)$, $\nu = 0, 1, \dots, p$ 的估计为

$$\hat{g}_\nu(x_0) = \nu! \hat{\beta}_\nu.$$

三个重要问题:

- 光滑参数(窗体宽度) h 的选取.

窗体宽度 h 对于局部回归起着至关重要的作用. 如果 h 太小, 只有少量数据落入 x 的小邻域内, 从而导致较大方差(undersmoothed). 相反地, 若 h 太大, 则有大量数据落入 x 的小邻域内, 导致回归函数 $g(x)$ 的某些局部特征被扭曲甚至丢失, 即产生较大偏差(oversmoothed). 因而在选择恰当的窗体宽度时, 必须要兼顾偏差和方差。

- 多项式的阶数 p 的选取.

对于给定窗体宽度 h , 较大的阶数 p 会减少建模的偏差, 但会导致较大的方差, 并会导致计算量增大. 一般地, 若估计 $g(x)$ 的 ν 阶导数 $g^{(\nu)}(x_0)$, 我们通常采用阶数为 $p = \nu + 1$, 或 $p = \nu + 3$.

- 核函数 K 的选取.

基于局部回归问题(17), 我们要求核函数 $K(u) \geq 0$. 事实上, 最优的核函数一般选用Epanechnikov 核函数

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

因为它能使得到的局部多项式回归估计的渐近均方误差达到最小。

4 局部多项式回归的理解

在研究线性回归时，我们主要关注于回归系数的估计问题。假如我们拟合的模型是正确的，我们常常会问所得到的回归系数的估计值与真实值“相差多远”；或者寻找其置信区间，检查哪些协变量是显著的；或者通过逐步回归等方法来决定保留哪些显著协变量和剔除哪些不显著的协变量，而对于所得到的回归曲线，倒是不很关心。

而在非参数回归中，我们的注意力是在所拟合的“回归曲线上”，一个很自然的问题就是：估计量 $\hat{g}(x)$ 与其真实值 $g(x)$ 相差多远？如何使得得到的函数估计更好，更贴近于观测数据。此外，我们依然可以研究其方差和置信区间，或者做变量选择问题。

5 局部多项式回归的统计性质

5.1 偏差和方差

设 \mathcal{X} 为 X 的一个基本空间，要注意到在回归模型中，协变量 X 可以为固定变量(例如时间 t)，也可以是一随机变量(例如身高)。

为方便，我们引入以下记号： $\mathbf{g} = \{g(X_1), \dots, g(X_n)\}^T$, $\boldsymbol{\beta} = \{g(x_0), \dots, g^{(p)}(x_0)/p!\}^T$, $\mathbf{r} = \mathbf{g} - \mathbf{X}\boldsymbol{\beta}$ 为局部多项式逼近的残差向量, $\boldsymbol{\Sigma} = \text{diag}\{K_h^2(x_0 - X_i)\sigma^2(X_i)\}$. 由公式(14)可得 $\boldsymbol{\beta}$ 的条件数学期望和条件方差分别为：

$$E(\hat{\boldsymbol{\beta}}|\mathcal{X}) = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{g} \quad (18)$$

$$= \boldsymbol{\beta} + (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{r},$$

$$\text{Var}(\hat{\boldsymbol{\beta}}|\mathcal{X}) = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} (\mathbf{X}^T \boldsymbol{\Sigma} \mathbf{X}) (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}. \quad (19)$$

虽然我们得到了 $\hat{\boldsymbol{\beta}}$ 的偏差和方差的表达式，但由于其含有未知的残差 \mathbf{r} 和

对角阵 Σ ，因而无法直接应用，下面我们给出它们的渐近表达式。为方便，我们先定义如下变量：

$$\mu_j = \int u^j K(u) du, \quad \nu_j = \int u^j K^2(u) du.$$

$$\mathbf{S} = (\mu_{j+l})_{0 \leq j, l \leq p}, \quad \tilde{\mathbf{S}} = (\mu_{j+l+1})_{0 \leq j, l \leq p}, \quad \mathbf{S}^* = (\nu_{j+l})_{0 \leq j, l \leq p},$$

$$\mathbf{c}_p = (\mu_{p+1}, \dots, \mu_{2p+1})^T, \quad \tilde{\mathbf{c}}_p = (\mu_{p+2}, \dots, \mu_{2p+2})^T,$$

此外，我们还定义向量 $\mathbf{e}_{\nu+1} = (0, \dots, 0, 1, 0, \dots, 0)^T$ ，该向量只有第 $(\nu+1)$ 个元素为1，其余元素都为0。

Fan and Gijbels (1996, Page 62) 证明了如下结论：

定理： 设 $f(x_0) > 0$ ，且在 x_0 的邻域内，函数 $f(\cdot)$ ， $g^{(p+1)}(\cdot)$ 和 $\sigma^2(\cdot)$ 均为连续函数。则当 $h \rightarrow 0$ ， $nh \rightarrow \infty$ ， $\hat{g}_\nu(x_0) = \nu! \hat{\beta}_\nu$ 的渐近条件方差为：

$$Var\{\hat{g}_\nu(x_0)|\mathcal{X}\} = \mathbf{e}_{\nu+1}^T \mathbf{S}^{-1} \mathbf{S}^* \mathbf{S}^{-1} \mathbf{e}_{\nu+1} \frac{\nu!^2 \sigma^2(x_0)}{f(x_0) n h^{1+2\nu}} + o_P(1/(n h^{1+2\nu})). \quad (20)$$

此外，当 $p - \nu$ 为奇数时，渐近偏差为：

$$Bias\{\hat{g}_\nu(x_0)|\mathcal{X}\} = \mathbf{e}_{\nu+1}^T \mathbf{S}^{-1} \mathbf{c}_p \frac{\nu!}{(p+1)!} g^{(p+1)}(x_0) h^{p+1-\nu} + o_P(h^{p+1-\nu}). \quad (21)$$

进而若函数 $f'(\cdot)$ 、 $g^{(p+2)}(\cdot)$ 在 x_0 的邻域内均连续，则当 $p - \nu$ 为偶数且 $nh^3 \rightarrow \infty$ 时，渐近偏差为：

$$\begin{aligned} Bias\{\hat{g}_\nu(x_0)|\mathcal{X}\} &= \mathbf{e}_{\nu+1}^T \mathbf{S}^{-1} \tilde{\mathbf{c}}_p \frac{\nu!}{(p+2)!} \{g^{(p+2)}(x_0) \\ &\quad + (p+2)g^{(p+1)}(x_0) \frac{f'(x_0)}{f(x_0)}\} h^{p+2-\nu} + o_P(h^{p+2-\nu}). \end{aligned} \quad (22)$$

由此定理可以看出： $p - \nu$ 为奇数和偶数时， $\hat{g}_\nu(x_0)$ 的渐近偏差是不同的，

而且事实证明 $p - \nu$ 为奇数时得到的估计量 $\hat{g}_\nu(x_0)$ 要优于 $p - \nu$ 为偶数时得到的估计量。

若假定在 x_0 的邻域内有 $\sigma^2(x) = \sigma^2(x_0)$,则条件方差(19)可以近似表示为:

$$(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{W}^2 \mathbf{X}) (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \sigma^2(x_0).$$

这里的 $\sigma^2(x_0)$ 可以由如下的加权残差平方和来估计:

$$\hat{\sigma}^2(x_0) = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 K_{h^*}(X_i - x_0)}{\text{trace}\{\mathbf{W}^* - \mathbf{W}^* \mathbf{X}^* (\mathbf{X}^{*T} \mathbf{W}^* \mathbf{X}^*)^{-1} \mathbf{X}^{*T} \mathbf{W}^*\}}. \quad (23)$$

其中 h^* , \mathbf{X}^* , \mathbf{W}^* 分别为 $(p + a)$ 阶局部多项式回归时的窗体宽度(bandwidth),设计阵和加权矩阵,通常我们可以取 $a = 1, 2$.

5.2 置信区间

在一些必要的正则条件下,可以证明:

$$\{V_\nu(x_0)\}^{1/2} \{\hat{g}_\nu(x_0) - g^{(\nu)}(x_0) - b_\nu(x_0)\} \xrightarrow{D} N(0, 1), \quad (24)$$

其中 $V_\nu(x_0)$, $b_\nu(x_0)$ 分别为 $\hat{g}_\nu(x_0)$ 的渐近方差和偏差[见((20)、21)式]。于是有:

定理: $\hat{g}^{(\nu)}(x_0)$ 的渐近 $(1 - \alpha)100\%$ 置信区间为:

$$\hat{g}_\nu(x_0) - \nu! \hat{b}_{\nu,p}(x_0) \pm z_{1-\alpha/2} \nu! \{\hat{V}_{\nu,p}(x_0)\}^{1/2}, \quad (25)$$

其中 $z_{1-\alpha/2}$ 为标准正态分布的 $(1 - \alpha/2)$ 分位数。

另外一种求置信区间的方法是基于Bootstrap的方法基础之上,方法如下: (1) 估计 $g(\cdot)$ 和 σ^2 .

(2) 计算残差 $\tilde{\epsilon}_i = Y_i - \hat{g}(X_i)$, 然后将残差中心化:

$$\hat{\epsilon}_i = \tilde{\epsilon}_i - \bar{\epsilon},$$

其中 $\bar{\epsilon} = n^{-1} \sum_{i=1}^n \tilde{\epsilon}_i$ 。

(3) 构建bootstrap重复抽样。从 $\hat{\epsilon}_1, \dots, \hat{\epsilon}_n$ 中随机重复抽样得到 $\epsilon_1^*, \dots, \epsilon_n^*$, 然后定义 $Y_i^* = \hat{g}(X_i) + \epsilon_i^*$ 。

(4) 基于 $\{(X_i, Y_i^*), i = 1, \dots, n\}$ 估计 $g(\cdot)$ 和 σ^2 , 不妨设为 $\hat{g}^*(x)$ 和 σ^{2*} 。重复做 B 次, 得到一系列估计 $\hat{g}_b(x), b = 1, \dots, B$, 于是对于任意一点 x , 得到 $g(x)$ 的 $(1 - \alpha)100\%$ 置信区间。

6 窗体宽度 h 的选取

窗体宽度 h 通常分为两类: (1) 若在函数 $g(x)$ 的整个定义域内为常数, 则称其为*constant bandwidth, or global bandwidth*. 当未知的回归函数在其定义域内表现为大致相同的光滑程度时, *global bandwidth*不失为一种理想选择。

而当回归函数在不同的区域内表现出不同的特点时, 比如, 在某个区域接近于线性, 而在另一个区域陡然升降、变化剧烈, 这时我们建议在不同的区域用不同的窗体宽度 h 。换句话说, 窗体宽度 h 跟 x 有关, 从而可表示为 $h(x)$ 。我们称之为*local variable bandwidth*。

6.1 Optimal bandwidth

首先我们要确定一个评价估计量优劣的准则, 此种准则通常都是建立在未知的回归函数与其估计量之间的某种距离基础之上。例如: L_1 , L_2 , 和 L_∞ 距离, 以及Hellinger距离。这里我们只考虑 L_2 损失函数, 即平方损失函数。

我们定义如下的均方误差(mean squared error, MSE)

$$\begin{aligned}\text{MSE}_{\nu,p}(x_0; h) &= E[\{\hat{g}_\nu(x_0) - g^{(\nu)}(x_0)\}^2 | \mathcal{X}] \\ &= [\text{Bias}\{\hat{g}_\nu(x_0) | \mathcal{X}\}^2 + \text{Var}\{\hat{g}_\nu(x_0) | \mathcal{X}\}],\end{aligned}\quad (26)$$

及其mean integrated squared error(MISE)

$$\begin{aligned}\text{MISE}_{\nu,p}(h) &= \int E[\{\hat{g}_\nu(x) - g^{(\nu)}(x)\}^2 | \mathcal{X}] w(x) dx \\ &= \int [\text{Bias}\{\hat{g}_\nu(x) | \mathcal{X}\}]^2 w(x) dx + \int \text{Var}\{\hat{g}_\nu(x) | \mathcal{X}\} w(x) dx\end{aligned}\quad (27)$$

其中 $w(\cdot)$ 为正的权函数，目的是为了保证MISE是渐近有限的。

最佳局部窗体宽度 $h(x)$ 则定义为：

$$h_{opt}(x_0) = \text{argmin}_h \text{MSE}_{\nu,p}(x_0; h).$$

等价地，

$$h_{opt}(x_0) = C_{\nu,p}(K) \left[\frac{\sigma^2(x_0)}{\{g^{(p+1)}(x_0)\}^2 f(x_0)} \right]^{1/(2p+3)} n^{-1/(2p+3)}, \quad (28)$$

其中

$$C_{\nu,p}(K) = \left[\frac{(p+1)!^2 (2\nu+1) \int K_\nu^{*2}(t) dt}{2(p+1-\nu) \{\int t^{p+1} K_\nu^*(t) dt\}^2} \right]^{1/(2p+3)}.$$

这里 K_ν^* 称为*equivalent kernel* [见Fan and Gijbels (2006), Page 64]. 表1 列出了一些常用的 $C_{\nu,p}(K)$.

此外，最佳全局窗体宽度 h 则定义为：

$$h_{opt} = \operatorname{argmin}_h \operatorname{MISE}_{\nu,p}(h).$$

即

$$h_{opt} = C_{\nu,p}(K) \left[\frac{\int \sigma^2(x)w(x)/f(x)dx}{\int \{g^{(p+1)}(x)\}^2 w(x)dx} \right]^{1/(2p+3)} n^{-1/(2p+3)}. \quad (29)$$

Table 1: 常数 $C_{\nu,p}(K)$ 的部分数值.

ν	p	Gaussian	Uniform	Epanechnikov	Biweight	Triweight
0	1	0.776	1.351	1.719	2.036	2.312
0	3	1.161	2.813	3.243	3.633	3.987
1	2	0.884	1.963	2.275	2.586	2.869
2	3	1.006	2.604	2.893	3.208	3.503

我们要指出的是，上面导出的最佳窗体宽度 h 都依赖于某些未知量，例如 X 的密度函数 $f(\cdot)$ ，方差 $\sigma^2(\cdot)$ 和导函数 $g^{(p+1)}(\cdot)$ 。在实际运用时，一种处理方法就是用其pilot estimators 去替代这些未知量，我们称这种方法为‘plug-in’型窗体选择法。我们还有另外一种选择方法：基于cross-validation准则上的选择。

6.2 Rule-of-Thumb Global Bandwidth Selector

主要思想就是寻找一个参数化模型作为参考模型来估计未知的回归函数 $g(\cdot)$ ，例如多项式回归。对于多项式的阶数，我们可以选取 $p+3$ (即 $g^{(p)}(\cdot)$ 为三次函数)。更具体地说，我们考虑如下参数模型：

$$g(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_{p+3} x^{p+3}, \quad (30)$$

并令(29) 中的 $w(\cdot)$ 为 $w(\cdot) = f_X(\cdot)w_0(\cdot)$, 这里 $w_0(\cdot) > 0$ 为一给定的正的权函数. 我们拟合线性模型(1)和(30), 从而得到回归曲线

$$\check{g}(x) = \check{\alpha}_0 + \check{\alpha}_1 x + \cdots + \check{\alpha}_{p+3} x^{p+3},$$

及方差的估计 $\check{\sigma}^2$. 然后把估计量

$$\check{g}^{(p+1)}(x) = (p+1)!\check{\alpha}_{p+1} + (p+2)!\check{\alpha}_{p+2}x + \frac{(p+3)!}{2}\check{\alpha}_{p+3}x^2,$$

和 $\check{\sigma}^2$ 带入到(29) 中, 就得到了估计值

$$C_{\nu,p}(K) \left[\frac{\check{\sigma}^2 \int w_0(x) dx}{n \int \{\check{g}^{(p+1)}(x)\}^2 w_0(x) f_X(x) dx} \right]^{1/(2p+3)}.$$

注意到分母可写为 $nE[\{\check{g}^{(p+1)}(X)\}^2 w_0(X)]$, 而该项可由如下量来估计

$$\sum_{i=1}^n \{\check{g}^{(p+1)}(X_i)\}^2 w_0(X_i).$$

于是, 我们就得到了(29) 中最优窗体宽度 $h_{\nu,p}^{opt}$ 的rule-of thumb (ROT) 估计

$$h_{\text{ROT}} = C_{\nu,p}(K) \left[\frac{\check{\sigma}^2 \int w_0(x) dx}{\sum_{i=1}^n \{\check{g}^{(p+1)}(X_i)\}^2 w_0(X_i)} \right]^{1/(2p+3)}. \quad (31)$$

本质上, 上述的rule-of thumb 选择方法与核密度估计中用正态分布密度做为参考密度的rule-of-thumb 窗体选择方法是一样的。虽然选出的窗体宽度不一定是最优的, 但在窗体宽度的选择上给我们提供了一个很好的参考值.

6.3 The leave-one-out cross-validation

令 $\hat{g}(x; h)$ 表示回归函数 $g(x)$ 的任意估计, 此估计量与窗体宽度 h 有关. 对于每一个 $i, i = 1, \dots, n$, 我们从样本 $\{(X_j, Y_j), i = 1, \dots, n\}$ 中剔除 (X_i, Y_i) , 用剩下的样本 $\{(X_j, Y_j), j \neq i\}$ 来估计回归函数 $g(\cdot)$, 得到其估计量 $\hat{g}_{-i}(x; h)$, 这里下角标 $-i$ 表示我们剔除了第 i 组观测数据, 然后我们检查预测偏差 $Y_i - \hat{g}_{-i}(X_i; h)$. 我们定义如下的cross-validation 准则:

$$CV(h) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{g}_{-i}(X_i; h))^2. \quad (32)$$

则我们选择使得cross-validation准则(32)达到最小的 h 为最优的窗体宽度 h_{opt} .

事实上, 此种方法可以进一步推广到 k -fold cross-validation, 即每次剔除 k 个观测值. 当 $k = 1$ 时, “one-fold cross-validation”即为“leave-one-out cross-validation”.

7 其它应用

7.1 极大似然估计

首先我们回忆一下参数模型的极大似然方法。

假定 $\{(X_i, Y_i), i = 1, \dots, n\}$ 是来自总体 (X, Y) 的一个样本, 我们想估计总体函数 $\theta(\cdot)$. 令 $\ell\{g(X_i), Y_i\}$ 为第 i 个观测数据 (X_i, Y_i) 的似然函数的对数(log-likelihood), 其中 g 为待估函数, 则有:

$$\ell(g; \mathbf{XY}) = \sum_{i=1}^n \ell\{g(X_i), Y_i\}.$$

若函数 g 可参数化为 $g(x) = g_\theta(x)$, 其中 θ 为未知参数, 则 θ 的极大似然估

计(MLE)为

$$\hat{\theta}_{\text{MLE}} = \operatorname{argmin}_{\theta} \sum_{i=1}^n \ell\{g_{\theta}(X_i), Y_i\}.$$

下面我们考虑 g 的非参数估计。给定点 x_0 ，在其小邻域内，我们可以将 $g(X_i)$ 做泰勒展开，即：

$$\begin{aligned} g(X_i) &\approx g(x_0) + g'(x_0)(X_i - x_0) + \cdots + \frac{g^{(p)}(x_0)}{p!}(X_i - x_0)^p \\ &\equiv \mathbf{B}_i^T \boldsymbol{\beta}, \end{aligned}$$

其中

$$\mathbf{B}_i = (1, X_i - x_0, \cdots, (X_i - x_0)^p)^T,$$

$$\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^T, \quad \beta_{\nu} = g^{(\nu)}(x_0)/\nu!, \quad \nu = 0, 1, \cdots, p.$$

于是我们得到如下的加权对数似然函数(local log-likelihood)

$$\ell_p(\boldsymbol{\beta}; h, x_0) = \sum_{i=1}^n \ell(\mathbf{B}_i^T \boldsymbol{\beta}, Y_i) K_h(X_i - x_0). \quad (33)$$

将上式关于 $\boldsymbol{\beta}$ 求最大值，即得 $\hat{g}^{(\nu)}(x_0) = \nu! \hat{\beta}_{\nu}, \nu = 0, 1, \dots, p$ 。

下面我们研究几个具体例子。

(1) 考虑如下模型：

$$Y = g(X) + \epsilon,$$

其中 $\epsilon \sim N(0, \sigma^2)$ ，且 X 与 ϵ 相互独立。则log-likelihood为：

$$-n \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n \{Y_i - g(X_i)\}^2.$$

在点 x_0 的邻域内对 $g(x)$ 做泰勒展开, 则相应的local log-likelihood为

$$-n \log(\sqrt{2\pi}\sigma) \sum_{i=1}^n K_h(X_i - x_0) - \frac{1}{2\sigma^2} \sum_{i=1}^n \left\{ Y_i - \sum_{j=0}^p \beta_j (X_i - x_0)^j \right\}^2 K_h(X_i - x_0).$$

然后解其关于 $\beta = (\beta_0, \dots, \beta_p)^T$ 的最大值问题, 所得结果恰为我们前面讲的局部多项式回归。

(2) 假设我们想研究吸烟跟肺癌的关系, X 表示烟龄, Y 表示是否有肺癌, 我们想估计吸烟得肺癌的概率, 这个概率用 $\theta \in (0, 1)$ 表示, 则我们的模型为:

$$\log(\theta/(1-\theta)) = g(x).$$

此时, 我们通常假定 Y 服从二项分布 $b(1, \theta)$, 于是我们可以写出log-likelihood:

$$\ell(g; X_i, Y_i) = \sum_{i=1}^n Y_i \log(\theta_i) + (1 - Y_i) \log(1 - \theta_i).$$

其中 $\theta_i = \frac{\exp\{g(X_i)\}}{1 + \exp\{g(X_i)\}}$. 然后利用局部多项式逼近的思想, 在点 x_0 的小邻域内用多项式逼近 $g(x)$, 于是有:

$$\ell^*(\beta; X_i, Y_i) = \sum_{i=1}^n \{Y_i \log(\theta_i^*) + (1 - Y_i) \log(1 - \theta_i^*)\} K_h(X_i - x_0),$$

这里 θ_i^* 满足 $\text{logit}(\theta_i^*) = \sum_{j=0}^p \beta_j (X_i - x_0)^j$, $\text{logit}(\pi) = \log\{\pi/(1-\pi)\}$ 称为logit函数。

(3) 普阿松分布:

设 $Y_i|X_i \sim \text{Poisson}(\lambda_i)$, $i = 1, \dots, n$, 其中 λ_i 满足 $\log(\lambda_i) = g(X_i)$. 则log-

likelihood函数为:

$$\ell(g(\cdot); X_i, Y_i) = C(\mathbf{Y}) + \sum_{i=1}^n \{Y_i \log(\lambda_i) - \lambda_i\}.$$

其中 $C(\mathbf{Y}) = C(Y_1, \dots, Y_n)$ 为常数, $\lambda_i = \exp(g(X_i))$. 在点 x_0 的邻域内对 $g(X_i)$ 做泰勒展开, 即:

$$g(X_i) \approx g(x_0) + g'(x_0)(X_i - x_0) + \dots + \frac{g^{(p)}(x_0)}{p!}(X_i - x_0)^p \equiv \mathbf{B}_i^T \boldsymbol{\beta}.$$

则local log-likelihood函数为:

$$\ell^*(\boldsymbol{\beta}) = C(\mathbf{Y}) + \sum_{i=1}^n \{Y_i(\mathbf{B}_i^T \boldsymbol{\beta}) - \exp(\mathbf{B}_i^T \boldsymbol{\beta})\} K_h(X_i - x_0).$$

7.2 条件分布函数的估计

给定 $X = x$, Y 的条件分布函数(CDF)为

$$\begin{aligned} F(y|x) &= \Pr(Y \leq y | X = x) \\ &= E\{I(Y \leq y) | X = x\}. \end{aligned}$$

注意到上式恰为随机变量 $I(Y \leq y)$ 的条件期望, 因而我们可以将其视为回归问题, 从而通过局部多项式回归来解决. 于是 $F(y|x)$ 的N-W核估计为:

$$\hat{F}(y|x) = \frac{\sum_{i=1}^n K_h(X_i - x) I(Y_i \leq y)}{\sum_{i=1}^n K_h(X_i - x)}$$

但此估计量不是smoothed的, 若想得到smoothed估计, 我们只需将 $I(Y_i \leq y)$ 替换为 $G\{(y - Y_i)/h_y\} = \int_{-\infty}^{\frac{y-Y_i}{h_y}} K(u) du$ 即可。

类似地，我们还可以得到local linear 估计。

更多内容请参考<http://www.ssc.wisc.edu/~bhansen/>

8 多元局部多项式回归

8.1 Nadaraya-Watson估计和局部线性一次估计

接下来我们研究 $\mathbf{X} \in R^d, d > 1$ 的情形，即：

$$g(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}).$$

其中 $\mathbf{x} = (x_1, \dots, x_d)^T$ ，因而 $g(\mathbf{x})$ 为一多元函数。利用条件期望定义，我们有：

$$E(Y|\mathbf{X} = \mathbf{x}) = \int y \frac{f(\mathbf{x}, y)}{f_{\mathbf{x}}(\mathbf{x})} dy.$$

分别代入 $f(\mathbf{x}, y)$ 和 $f_{\mathbf{x}}(\mathbf{x})$ 的核密度估计：

$$\hat{f}_{h, \mathbf{H}}(\mathbf{x}, y) = n^{-1} \sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}) K_h(Y_i - y), \quad \hat{f}_{\mathbf{x}}(\mathbf{x}) = n^{-1} \sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}).$$

于是我们得到多元的Nadaraya-Watson估计

$$\hat{g}_{\mathbf{H}}(\mathbf{x}) = \frac{\sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}) Y_i}{\sum_{i=1}^n \mathcal{K}_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})}. \quad (34)$$

事实上，上述Nadaraya-Watson估计(34)可视为下述优化问题(35)的解：

$$\min_{\beta_0} \sum_{i=1}^n \{Y_i - \beta_0\}^2 \mathcal{K}_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}), \quad (35)$$

相应地，局部线性一次估计为下述优化问题(36)的解：

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n \{Y_i - \beta_0 - (\mathbf{X}_i - \mathbf{x})^T \beta_1\}^2 \mathcal{K}_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}), \quad (36)$$

最优窗体宽度 \mathbf{H} 和 h 可以运用cross-validation方法来获得。

8.2 维数诅咒(curse-of-dimensionality)

9 应用

例9.1 考虑如下模型

$$Y_i = g(X_i) + \sigma \epsilon_i, i = 1, \dots, n,$$

其中 $g(x) = x + 2e^{-16x^2}$, $\sigma = 0.4$, $X_i \sim U(-2, 2)$, $\epsilon_i \sim N(0, 1)$. 随机产生数据 $\{(X_i, Y_i), i = 1, \dots, n\}$, 试估计 $g(x)$.

例9.2 考虑如下模型

$$Y_i = g(X_i) + \sigma \epsilon_i, i = 1, \dots, n,$$

其中 $g(x) = \sin(2x) + 2e^{-16x^2}$, $\sigma = 0.3$, $X_i \sim U(-2, 2)$, $\epsilon_i \sim N(0, 1)$. 随机产生数据 $\{(X_i, Y_i), i = 1, \dots, n\}$, 试估计 $g(x)$.

例9.3 考虑如下模型

$$Y_i = g(X_i) + \sigma \epsilon_i, i = 1, \dots, n,$$

其中 $g(x) = 0.3 \exp\{-4(x + 1)^2\} + 0.7 \exp\{-16(x - 1)^2\}$, $\sigma = 0.1$, $X_i \sim$

$U(-2, 2)$, $\epsilon_i \sim N(0, 1)$. 随机产生数据 $\{(X_i, Y_i), i = 1, \dots, n\}$, 试估计 $g(x)$.

例9.4 考虑如下模型

$$Y_i = g(X_i) + \sigma\epsilon_i, i = 1, \dots, n,$$

其中 $g(x) = 0.4x + 1$, $\sigma = 0.15$, $X_i \sim N(0, 1)$, $\epsilon_i \sim N(0, 1)$. 随机产生数据 $\{(X_i, Y_i), i = 1, \dots, n\}$, 试估计 $g(x)$.

例9.5 研究加拿大工人收入(age.income数据集)的年龄(age)和收入(income)的关系, 该数据来源于R程序包“SemiPar”。总共调查了 $n = 205$ 个加拿大工人的年龄(age)和收入(income), 所有工人的都是高中毕业, 取 $\log.income = \log(income)$, 试拟合age与log.income之间的函数关系。

例9.6 我们来研究摩托车碰撞试验的数据集mcycle, 该数据集来源于R程序包MASS. 试拟合time与acceleration之间的函数关系。

例9.7 我们来研究ethanol数据集, 该数据集来源于R程序包locfit. 试拟合E与NO_x之间的函数关系。

参考文献

- Fan, J. Q. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC, Boca Raton, USA.
- Hardle, W., Muller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Springer, New York, USA.
- Li, Q. and Racine, J. S. (2006). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, USA.
- Ritz, C. and Streibig, J. C. (2008). *Nonlinear Regression with R*. Springer, New York, USA.

Simonoff, J. S. (1998). *Smoothing Methods in Statistics* . Springer, New York, USA.

Wang, M. P. and Jones, M. C. (1994). *Kernel Smoothing*. Chapman & Hall/CRC, Boca Raton, USA.