DOI:10.13656/j.cnki.gxkx.2012.03.021

广西科学 Guangxi Sciences 2012,19(3):230～233

固定设计下回归函数局部线性核估计的渐近性质*

Asymptotic Properties of Local Linear Kernel Estimator for Fixed Design

吴果林[1],陈雄伟[2]

WU Guo-lin[1] CHEN Xiong-wei[2]

(1.桂林航天工业学院信息工程系,广西桂林 541004;2.桂林电子科技大学,广西桂林 541004)

(1. Department of Information Engineering, Guilin College of Aerospace Techology, Guilin, Guangxi, 541004, China; 2. Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China)

摘要:讨论固定设计下,局部线性核估计内部和外界地区的渐近性质.固定设计下回归函数局部线性核估计具有自适应性,在内部地区和外界地区具有相同的收敛速度.

关键词:固定设计 非参数回归 核估计 边界偏倚

中图法分类号:O212.7 **文献标识码**:A **文章编号**:1005-9164(2012)03-0230-04

Abstract: Asymptotic properties of the local linear kernel estimator for fixed design were discussed in their interior and boundary regions. The properties show that this estimator is adaptive in the boundary regions and has the same convergence speed as that in internal areas.

Key words: fixed design, nonparametric regression, kernel estimator, boundary bias

近年来,回归问题的非参数估计已成为数理统计中的热门话题,受到了许多学者的关注.回归问题的非参数估计方法有很多种,常见的有核函数、样条函数和小波,其中核函数回归估计(简称核估计)易于理解、结构简单、便于计算,已成为非参数回归估计的主要方法.核估计最早由 Nadaraya[1] 和 Watson[2] 给出,称为 Nadaraya-Watson 估计.在此基础上加以改进的核估计是局部多项式核估计[3,4],它是局部加权最小二乘的多项式估计.易见,当多项式为常数时,局部多项式核估计就是 Nadaraya-Watson 估计[5].在局部多项式核估计类中,最重要的一个估计是局部线性核估计,它不但拥有 Nadaraya-Watson 估计一些良好的性质,而且改进了 Nadaraya-Watson 估计设计偏倚与边界偏倚的不足[6,7].文献[6]和[7]讨论了自变量为随机变量的情形,对于固定点设计的情形还未见有报道.事实上,在许多科学研究[8]和试验设计过程中,常常碰到固定点设计的情形,因此本文研究固定设计下回归函数的局部线性核估计的渐近性质,特别是它在边界地区的渐近性质.

收稿日期:2011-11-22

修回日期:2012-04-22

作者简介:吴果林(1977-),男,讲师,硕士,主要从事非参数统计、数值计算研究。

*广西教育厅科研项目(201106LX717)资助。

1 回归函数固定设计下局部线性核估计

固定设计下的非参数回归模型为

$$Y_i = r(x_i) + v^{1/2}(x_i)\varepsilon_i, i = 1,2,\cdots,n, \quad (1.1)$$

其中 $x_i, i = 1,2,\cdots,n$ 为已知的固定设计点列,$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 是一组相互独立的随机变量,且满足 $E(\varepsilon_i) = 0, \mathrm{Var}(\varepsilon_i) = 1$. 显然 $E(Y_i) = r(x_i), \mathrm{Var}(Y_i) = v(x_i)$,故称 r 为均值回归函数,v 为方差函数,常常假设对于任意 i,有 $v(x_i) = \sigma^2$.

假设 $r(x)$ 具有二阶导数,在点 x 的某邻域,有 $r(t) \approx r(x) + r'(x)(t-x) \equiv \beta_0 + \beta_1(t-x)$,则回归函数 $r(x)$ 的估计等价于估计截距 β_0. 考虑一个权重局部线性回归问题,求解 β_0 和 β_1,使得它们最小化表达式:

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1(x_i - x))^2 K_h(x - x_i). \quad (1.2)$$

这里 K 为对称概率密度函数,称为权函数或核函数,并记 $K_h(u) = h^{-1}K(u/h)$;h 为正实数,常称为带宽

或窗宽.

这是一个最小二乘问题,(1.2) 式对 β_0,β_1 求偏导数,并令其为 0,得

$$\begin{cases}-2\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1(x_i-x))K_h(x_i-x)=0,\\ -2\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1(x_i-x))K_h(x_i-x)(x_i-x)=0.\end{cases} \tag{1.3}$$

方程组(1.3) 写成矩阵形式为

$$(X^{\mathrm{T}}WX)\beta=X^{\mathrm{T}}WY, \tag{1.4}$$

其中

$$X=\begin{pmatrix}1 & x_1-x\\ 1 & x_2-x\\ \vdots & \vdots\\ 1 & x_n-x\end{pmatrix},W=\mathrm{diag}\{K_h(x_i-x)\},Y=\begin{pmatrix}Y_1\\ Y_2\\ \vdots\\ Y_n\end{pmatrix},\beta=\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}.$$

记 $e_1=(1,0)^{\mathrm{T}}$,若 $X^{\mathrm{T}}WX$ 可逆,由式(1.4) 得回归函数 $r(x)$ 的局部线性核估计为

$$\hat{r}(x)=e_1^{\mathrm{T}}(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}WY. \tag{1.5}$$

若记 $\hat{s}_l(x)=n^{-1}\sum_{i=1}^{n}(x_i-x)^lK_h(x_i-x),l=0,1,\cdots,$式(1.5) 可以写为

$$\hat{r}(x)=n^{-1}\sum_{i=1}^{n}\frac{\{\hat{s}_2(x)-\hat{s}_1(x)(x_i-x)\}K_h(x_i-x)Y_i}{\hat{s}_2(x)\hat{s}_0(x)-\hat{s}_1^2(x)}. \tag{1.6}$$

2 局部线性核估计的渐近性质

用均方误差(MSE) 来考查估计的一些渐近性质:

$$\mathrm{MSE}\{\hat{f}(x)\}=E^2\{\hat{f}(x)-f(x)\}=\mathrm{Var}\{\hat{f}(x)\}+\mathrm{Bias}^2\{\hat{f}(x)\}, \tag{2.1}$$

其中 $\hat{f}(x)$ 为 $f(x)$ 的估计,$\mathrm{Bias}\{\hat{f}(x)\}=E\hat{f}(x)-f(x)$.

由文献[1,2] 可知,非参数核估计的内部和边界地区有不同阶的渐近偏差,事实上局部线性核估计也同样存在这个问题. 由式(1.6) 可得 $\hat{r}(x)=n^{-1}\sum_{i=1}^{n}\omega_iY_i$,其中

$$\omega_i=\frac{\{\hat{s}_2(x)-\hat{s}_1(x)(x_i-x)\}K_h(x_i-x)}{\hat{s}_2(x)\hat{s}_0(x)-\hat{s}_1^2(x)},$$

故局部线性核估计也是线性估计器,$\hat{r}(x)$ 为 $Y_i(i=1,2,\cdots,n)$ 的加权平均,权 ω_i 是关于 $K_h(x_i-x)$ 的函数,而 $K_h(x_i-x)$ 是关于点 x 对称的,因此当点 x 位于边界地区,以不完全的观测数据作加权平均,相对于内部点 x,$\hat{r}(x)$ 的偏差和肯定要受影响. 图 1 是观测数据和在点 $t=0.08,0.5,0.9$ 处核函数 $K_h(x-t)$ 的曲线,核函数取 Epanechnikov 核,窗宽 $h=0.2$. 从图 1 可以看出,当 $0\leqslant x<h$ 或 $1-x<h\leqslant 1$ 时,此时核函数部分曲线位于区间的外面,则在该点处(如 $x=0.08$) 权值 ω_i 的数量有所减少,而在内部,地区 $(h\leqslant x\leqslant 1-h)$ 将不会出现这种情况. 综合上述分析后,分两种情形推导固定设计下局部线性核估计的渐近均方误差.

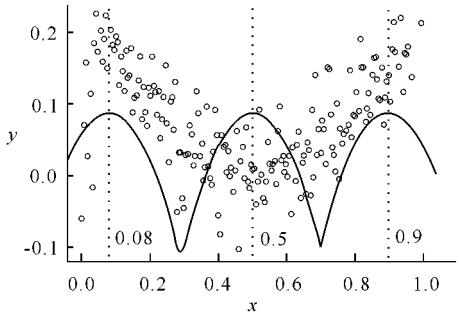

图 1 观测数据以及核函数 $k_h(x-t)$

Fig. 1 Observed data and the kernel function $k_h(x-t)$

2.1 内部地区的渐近性质

考虑固定设计下的回归模型(1.1),不失一般性,设自变量 x 的取值范围为$[0,1]$,为方便分析,作如下假定:

(1) 回归函数 $r(x)$ 的二阶导数 $r''(x)$ 和方差函数 $v(x)$ 是区间$[0,1]$ 上的连续函数.

(2) 核函数 $K(x)$ 在区间$[0,1]$ 上是关于原点对称的密度函数,即 $\int_{-1}^{1}xK(x)\mathrm{d}x=0,\int_{-1}^{1}K(x)\mathrm{d}x=1.$

(3) 窗宽 $h=h(n)$,且当 $n\to\infty$ 时,$h\to 0$,$n^{\frac{1}{2}}h\to\infty$,$n^{-1}=o(h^2)$.

(4) 对于估计点 x,满足 $h<x<1-h$.

由式(2.1) 可知,欲计算 $\hat{r}(x)$ 的渐近均方误差,只需计算它的渐近方差和渐近偏差即可. 从式(1.5) 计算可得

$$E\hat{r}(x)=e_1^{\mathrm{T}}(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}WR, \tag{2.2}$$

其中 $R=(r(x_1),r(x_2),\cdots,r(x_n))^{\mathrm{T}}$. 由泰勒公式,对于任意 $x\in[0,1]$,

$$r(x_i)=r(x)+(x_i-x)r'(x)+\frac{1}{2}(x_i-x)^2r''(x)+\cdots$$

则 R 可以改写为

$$R = X\begin{bmatrix} r(x) \\ r'(x) \end{bmatrix} + \frac{1}{2}r''(x)\begin{bmatrix} (x_1-x)^2 \\ \vdots \\ (x_n-x)^2 \end{bmatrix} + \cdots$$

因此，式(2.2)的第一项可以化简为

$$e_1^{\mathrm{T}}(X^{\mathrm{T}}WX)^{-1}(X^{\mathrm{T}}WX)\begin{bmatrix} r(x) \\ r'(x) \end{bmatrix} = e_1^{\mathrm{T}}\begin{bmatrix} r(x) \\ r'(x) \end{bmatrix} = r(x).$$

于是，估计函数 $\hat{r}(x)$ 的偏差为

$$\mathrm{Bias}\{\hat{r}(x)\} = E\hat{r}(x) - r(x) = \frac{1}{2}r''(x)e_1^{\mathrm{T}}(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W\begin{bmatrix} (x_1-x)^2 \\ \vdots \\ (x_n-x)^2 \end{bmatrix} + \cdots \quad (2.3)$$

为了推导渐近偏差，应用前面的记号 $\hat{s}_l(x) = n^{-1}\sum_{i=1}^{n}(x_i-x)^l K_h(x_i-x)$，那么

$$n^{-1}X^{\mathrm{T}}WX = \begin{bmatrix} \hat{s}_0(x) & \hat{s}_1(x) \\ \hat{s}_1(x) & \hat{s}_2(x) \end{bmatrix},$$

$$n^{-1}X^{\mathrm{T}}W\begin{bmatrix} (x_1-x)^2 \\ \vdots \\ (x_n-x)^2 \end{bmatrix} = \begin{bmatrix} \hat{s}_2(x) \\ \hat{s}_3(x) \end{bmatrix}.$$

利用定积分的定义，及假设条件(1)～(4)，当 $n\to\infty$，有

$$\hat{s}_l(x) = n^{-1}\sum_{i=1}^{n}(x_i-x)^l K_h(x_i-x) = \int_0^1 (t-x)^l K_h(t-x)\mathrm{d}t + O(n^{-1}) = h^l\int_{-x/h}^{(1-x)/h} u^l K(u)\mathrm{d}u + O(n^{-1}) = h^l\int_{-1}^{1} u^l K(u)\mathrm{d}u + O(n^{-1}).$$

记 $\mu_l(K) = \int_{-1}^{1} u^l K(u)\mathrm{d}u$，并注意到条件(2)，则上面的矩阵可以化简为

$$n^{-1}X^{\mathrm{T}}WX = \begin{bmatrix} 1+O(n^{-1}) & O(n^{-1}) \\ O(n^{-1}) & h^2\mu_2(K)+O(n^{-1}) \end{bmatrix}, \quad (2.4)$$

$$n^{-1}X^{\mathrm{T}}W\begin{bmatrix} (x_1-x)^2 \\ \vdots \\ (x_n-x)^2 \end{bmatrix} = \begin{bmatrix} h^2\mu_2(K)+O(n^{-1}) \\ O(n^{-1}) \end{bmatrix}. \quad (2.5)$$

由矩阵知识可知

$$(n^{-1}X^{\mathrm{T}}WX)^{-1} = (h^2\mu_2(K))^{-1}\begin{bmatrix} h^2\mu_2(K)+O(n^{-1}) & O(n^{-1}) \\ O(n^{-1}) & 1+O(n^{-1}) \end{bmatrix}, \quad (2.6)$$

将式(2.5)、(2.6)代入式(2.3)，得局部线性核估计 $\hat{r}(x)$ 的渐近偏差为

$$\mathrm{Bias}\{\hat{r}(x)\} = E\hat{r}(x) - r(x) = \frac{1}{2}h^2 r''(x)\mu_2(K) + O(n^{-1}) + O((nh)^{-2}) = \frac{1}{2}h^2 r''(x)\mu_2(K) + o(h^2). \quad (2.7)$$

对于渐近方差，由式(1.5)计算可得

$$\mathrm{Var}\{\hat{r}(x)\} = e_1^{\mathrm{T}}(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}WVWX(X^{\mathrm{T}}WX)^{-1}e_1, \quad (2.8)$$

其中 $V = \mathrm{diag}\{v(x_i)\}$. 类似上面的计算，得

$$n^{-1}X^{\mathrm{T}}WVWX = n^{-1}\sum_{i=1}^{n}K_h^2(x_i-x)v(x_i)\begin{bmatrix} 1 & (x_i-x) \\ (x_i-x) & (x_i-x)^2 \end{bmatrix} = \begin{bmatrix} h^{-1}\mu_0(K^2)v(x)+O(n^{-1}) & O(n^{-1}) \\ O(n^{-1}) & h^2\mu_2(K^2)+O(n^{-1}) \end{bmatrix}. \quad (2.9)$$

将式(2.9)代入式(2.8)，并结合前面的式子，得局部线性核估计 $\hat{r}(x)$ 的渐近方差为

$$\mathrm{Var}\{\hat{r}(x)\} = e_1^{\mathrm{T}}(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}WVWX(X^{\mathrm{T}}WX)^{-1}e_1 = (nh)^{-1}\mu_0(K^2)v(x) + O(n^{-2}h^{-3}) = (nh)^{-1}\mu_0(K^2)v(x) + o\{(nh)^{-1}\}. \quad (2.10)$$

再联合式(2.7)和(2.10)，得局部线性核估计 $\hat{r}(x)$ 的渐近均方误差为

$$\mathrm{MSE}\{\hat{r}(x)\} = \frac{1}{4}h^4(r''(x)\mu_2(K))^2 + (nh)^{-1}\mu_0(K^2)v(x) + o\{(nh)^{-1}+h^4\}. \quad (2.11)$$

定理 1 假设有一组关于两变量 x 和 Y 的数据 $\{(x_i,Y_i), i=1,2,\cdots,n\}$ 来自模型(1.1)，且满足条件(1)～(4)，则对于任意 $x\in[h,1-h]$，由式(1.6)所得的估计 $\hat{r}(x)$ 满足式(2.11).

2.2 边界地区的渐近性质

不失一般性，只考虑区间[0,1]左端的边界地区的点. 若回归模型(1.1)满足以下条件：

① 回归函数 $r(x)$ 的二阶导数 $r''(x)$ 和方差函数 $v(x)$ 是区间[0,1]上的连续函数.

② 核函数 $K(x)$ 在区间[0,1]上是关于原点对称的密度函数，即 $\int_{-1}^{1}xK(x)\mathrm{d}x = 0$，$\int_{-1}^{1}K(x)\mathrm{d}x = 1$.

③ 窗宽 $h = h(n)$，且当 $n\to\infty$ 时，$h\to 0$，$n^{\frac{1}{3}}h\to\infty$，$n^{-1} = o(h^3)$.

④ 对于估计点 $x = \alpha h$，其中 $0\leqslant\alpha<1$.

定义 $\mu_l(K,\alpha) = \int_{-\alpha}^{1}u^l K(u)\mathrm{d}u$，重新计算式(2.4)和式(2.5)，得

$$n^{-1}X^{\mathrm{T}}WX=\begin{bmatrix}\mu_0(K,\alpha)+O(n^{-1}) & h\mu_1(K,\alpha)+O(n^{-1})\\ h\mu_1(K,\alpha)+O(n^{-1}) & h^2\mu_2(K,\alpha)+O(n^{-1})\end{bmatrix},\tag{2.12}$$

$$n^{-1}X^{\mathrm{T}}W\begin{bmatrix}(x_1-x)^2\\ \vdots\\ (x_n-x)^2\end{bmatrix}=\begin{bmatrix}h^2\mu_2(K,\alpha)+O(n^{-1})\\ h^3\mu_3(K,\alpha)+O(n^{-1})\end{bmatrix}.\tag{2.13}$$

同样，将式(2.12)、(2.13) **代入式**(2.3)**，得局部线性核估计** $\hat{r}(x)$ **的边界渐近偏差为**

$$\mathrm{Bias}\{\hat{r}(x)\}=E\hat{r}(x)-r(x)=\frac{1}{2}h^2r''(x)A(\alpha)+O(n^{-1}h),$$

其中

$$A(\alpha)=\frac{\mu_2^2(K,\alpha)-\mu_1(K,\alpha)\mu_3(K,\alpha)}{\mu_0(K,\alpha)\mu_2(K,\alpha)-\mu_1^2(K,\alpha)}.$$

利用条件 ③**，上式还可以化为**

$$\mathrm{Bias}\{\hat{r}(x)\}=\frac{1}{2}h^2r''(x)A(\alpha)+o(h^2).\tag{2.14}$$

类似于上面的计算，得局部线性核估计 $\hat{r}(x)$ **的边界渐近方差为**

$$\mathrm{Var}\{\hat{r}(x)\}=(nh)^{-1}B(\alpha)v(x)+\mathrm{O}(n^{-2}h^{-4})=(nh)^{-1}B(\alpha)v(x)+O\{(nh)^{-1}\},\tag{2.15}$$

其中

$$B(\alpha)=(\mu_2^2(K,\alpha)\mu_0(K^2,\alpha)-2\mu_1(K,\alpha)\mu_1(K^2,\alpha)\mu_2(K,\alpha)+\mu_1^2(K,\alpha)\mu_2(K^2,\alpha))/(\{\mu_0(K,\alpha)\mu_2(K,\alpha)-\mu_1^2(K,\alpha)\}^2).$$

由此，联合式(2.14) **和式**(2.15)**，得局部线性核估计** $\hat{r}(x)$ **的边界渐近均方误差为**

$$\mathrm{MSE}\{\hat{r}(x)\}=\frac{1}{4}h^4(r''(x)A(\alpha))^2+(nh)^{-1}B(\alpha)v(x)+o\{(nh)^{-1}+h^4\}.\tag{2.16}$$

定理 2　假定有一组关于两变量 x **和** Y **的数据** $\{(x_i,Y_i),i=1,2,\cdots,n\}$ **来自模型**(1.1)**，且满足条件** ①～④**，则对于任意** $x\in[0,h)$**，由式**(1.6) **所得的估计** $\hat{r}(x)$ **满足式**(2.16).

对比式(2.11)**和**(2.16)**不难发现，固定设计下局部线性核估计在内部和边界地区具有相同的收敛阶，这与随机设计下局部线性核估计的情形相同**[7]**. 也就是说，固定设计下局部线性核估计具有自适应性，它不需要改变边界地区的估计方法就与内部具有相同的收敛速度，这对固定设计下** Nadaraya-Watson **估计**[9]**的边界收敛速度有所改进. 因此，固定设计下局部线性核估计是一个实用且性质优良的非参数回归估计.**

参考文献：

[1] Nadaraya E A. On estimating regression[J]. Theory of Probability and Its Applications, 1964, 10: 186-190.

[2] Watson G S. Smooth regression analysis[J]. Sankhya Series A, 1964, 26: 359-372.

[3] Ruppert D, Wand M P. Multivariate weighted least squares regression[J]. The Annals of Statistics, 1994, 22: 1346-1370.

[4] Fan J, Gijbels I. Local polynomial modelling and its applications[M]. London: CHAPMAN & HALL, 1996.

[5] Wasserma L. All of nonparametric statistic[M]. New York: Spring-Verlag, 2005.

[6] Fan J. Design-adaptive nonparametric regression[J]. Journal of the American Statistical Association, 1992, 20: 2008-2036.

[7] Fan J, Gijbels I. Variable bandwidth and local linear regression smoothers[J]. The Annals of Statistics, 1992, 87: 998-1004.

[8] Brown L, Cai T, Zhou H, et al. The root-unroot algorithm for density estimation as implemented via wavelet block thresholding[J]. Probab Theory Relat Fields, 2010, 146: 401-433.

[9] Gasser T, Müller H G. Estimating regression functions and their derivatives by the kernel method[J]. Scand J of Statist, 1984, 11: 171-185.

（责任编辑：尹　闯）