文章编号:1007－6492(1999)01－0060－04

非参数分布自由的回归函数核估计的相合性

胡玉萍

(郑州工业大学数理力学系,河南 郑州 450002)

摘 要:设(X_i,Y_i),$i=1,\cdots,n$ 是从取值于$R^d\times R^1$的随机向量(X,Y)中抽取的iid样本,$E(|Y|)<\infty$,而以$m(x)=E(Y|X=x)$表示回归函数.在合适条件下获得了一类基于完全和截尾数据回归函数核估计的逐点相合性,所获的结果对于所有X的分布μ均成立,因此是分布自由的.

关键词:截尾数据;回归函数;核估计;分布自由

中图分类号:O 212.7 **文献标识码**:A

1 若干引理

设(X_1,Y_1),(X_2,Y_2),…是从取值于$R^d\times R$的随机向量(X,Y)中抽取的相互独立且同分布(简记为iid)的样本,$E|Y|<\infty$,文献中通常考虑$m(x):=E(Y|X=x)$的核估计为:

$$m_n^{(1)}(x)=\sum_{i=1}^n Y_iK\left(\frac{X_i-x}{h_n}\right)\Big\backslash\sum_{i=1}^n K\left(\frac{X_i-x}{h_n}\right),\quad(1)$$

$$m_n^{(2)}(x)=\sum_{i=1}^n Y_iK\left(\frac{X_i-x}{h_i}\right)\Big\backslash\sum_{j=1}^n K\left(\frac{X_j-x}{h_j}\right),\quad(2)$$

$$m_n(x)=\sum_{j=1}^n Y_ih_i^dK\left(\frac{X_i-x}{h_i\ us}\right)\Big\backslash\sum_{j=1}^n h_j^{-d}K\left(\frac{X_i-x}{h_j}\right),\quad(3)$$

这里$0<h_n\to 0$.

文献中,对这些估计的收敛性有不少的讨论,文献[1]在X为连续型分布(即X为分布μ关于Lebesgue测度具有密度函数$\mathrm{d}L(x)/\mathrm{d}\mu(x)=f(x)$,核函数$K$非负有界,$\int_{R^d}K(x)\mathrm{d}x<\infty$,$\int_{R^d}\psi(x)\mathrm{d}x<\infty$(其中:$\psi(x)=\sup_{\|y\|\geqslant\|x\|}K(y)$)等条件下,得到了基于完全与截尾样本时$m(x)$核估计的强相合.

文献[2]从应用的角度考虑,将上述结果在两个方面进行推广:

(a) 不要求X具有密度函数$f(x)$;

(b) 不要求核函数K可积.

所获得的结果对所有X的分布μ均成立,因此是分布自由的.核函数K满足下面的条件:

(c) $C_1(\|x\|)\leqslant K(x)\leqslant C_2H(\|x\|)$;

(d) $\lim_{t\to\infty}t^dH(t)=0$.

其中:C_1,C_2为两个正数,$H(t)$为$[0,\infty]$上的某个降函数.它最早由Greblicki,Krzyzak和Pawlak在文献[3]中提出,后来被人们广泛采用.文献[2]在一定条件下得到了回归函数核估计$m_n^{(1)}(x)$,$m_n^{(2)}(x)$的逐点强相合性.

本文在文献[2]的基础上进一步研究了Ahmed和Lin在文献[4]中提出的一类递归核估计$m_n(x)$的逐点强相合性,从而解决了文献[2]中提出的问题.

为叙述方便,用$C_1,C_2,C,C(x),C_1(x)$等表示与n无关的有限正数、正函数,且每次出现不要求相同,模$\|\cdot\|$均取为L_2模或均为L_∞模,$k:=\sup_x K(x)$,为获得主要结果,先介绍若干引理.

引理1[5] 设$X_1,X_2,\cdots$相互独立,且存在$b_n\uparrow\infty$,$\alpha_n\in[1,2]$,使

$$\sum_{n=1}^\infty E|X_n|^{\alpha_n}/b^{\alpha_n}<\infty,$$

那么

$$\lim_{n\to\infty}b_n^{-1}\sum_{i=1}^n(X_i-EX_i)=0\quad\text{a.s.}\quad(4)$$

引理2[6] 设$K(x)$满足条件(c),(d),则对任意μ可积函数f

$$\lim_{h\to 0}\int_{R^d}K\left(\frac{y-x}{h}\right)f(y)\,\mu(\mathrm{d}y)\Big/\int_{R^d}K\left(\frac{y-x}{h}\right)\mu(\mathrm{d}y)=f(x),\ \text{a.e.}(\mu)\ x\in R^d.\quad(5)$$

引理3[7] 记$\alpha_h(x)=h^d/\mu(S_h(x))$,则存

收稿日期:1998－09－23;**修订日期**:1998－11－23

作者简介:胡玉萍(1971—),女,河南省尉氏县人,郑州工业大学助教,硕士.

在非负有限函数 $g(x)$，使

$$\lim_{h\to 0}\alpha_h(x)=g(x),\ a.e.(\mu)\ x\in R^d,\quad (6)$$

其中：$S_h(x)$ 为 R^d 中以 x 为中心；h 为半径的球域.

2　回归函数估计的逐点强相合性

定理 1　设核函数 K 满足条件(c)，(d)，存在 $1<p\leqslant 2$，使 $E|Y|^p<\infty$.

$$\sum_{n=1}^{\infty}(nh_n^d)^{-p}<\infty \quad \text{时},$$

则有

$$\lim_{n\to\infty}m_n(x)\overset{a.s}{=}m(x),\ a.e.(\mu)\ x\in R^d.$$

证明　首先由式(c)，(d) 易推出(去掉无意义的 $K\equiv 0$ 情形)，存在正数 b 和 C 使

$$K(x)\geqslant C\mathrm{I}_{(\|x\|\leqslant b)}(x).\quad (7)$$

因为 $h_n\to 0$，所以存在正数 A，使 $h_n\leqslant A$，$n=1,2,\cdots$，由式(7)得 $EK(\frac{X-x}{h})\geqslant Ch^d/a_{bh}(x)$，从而 $(EK(\frac{X-x}{h}))^{-1}\leqslant a_{bh}(x)/ch^d$，由引理 3 推出，存在有限函数 $C(x)$，使

$$(EK(\frac{X-x}{h}))^{-1}\leqslant C(x)/h^d,$$
$$\text{任 } 0<h\leqslant A,\ a.e.(\mu)\ x\in R^d.\quad (8)$$

记

$$U_n(x)=\sum_{i=1}^{n}Y_ih_i^{-d}K(\frac{X_i-x}{h_i})\Big/\Big[\sum_{j=1}^{n}h_j^{-d}EK(\frac{X_j-x}{h_j})\Big],\quad (9)$$

$$V_n(x)=\sum_{i=1}^{n}h_i^{-d}K(\frac{X_i-x}{h_i})\Big/\Big[\sum_{j=1}^{n}h_j^{-d}EK(\frac{X_j-x}{h_j})\Big],\quad (10)$$

$$Z_n(x,X_n,Y_n)=Y_nh_n^{-d}K(\frac{X_n-x}{h_n}),\quad (11)$$

则

$$U_n(x)=\sum_{i=1}^{n}Z_i(x,X_i,Y_i)\Big/\Big[\sum_{j=1}^{n}h_j^{-d}EK(\frac{X_j-x}{h_j})\Big],$$

$$EZ_n(x,X_n,Y_n)/(h_n^{-d}EK(\frac{X_n-x}{h_n}))=E(K(\frac{X_n-x}{h_n})Y_n)/EK(\frac{X_n-x}{h_n}).$$

由引理 2 知

$$\lim_{n\to\infty}E(K(\frac{X_n-x}{h_n})Y_n)/EK(\frac{X_n-x}{h_n})=E(Y_n\mid X_1=x)=m(x),$$
$$a.e.(\mu)\ x\in R^d,\quad (12)$$

又

$$\sum_{j=1}^{n}h_j^{-d}EK(\frac{X_j-x}{h_n})\geqslant C^{-1}(x)n\uparrow\infty$$
$$a.e.(\mu)\ x\in R^d,\quad (13)$$

由 Toeplitz 引理得

么

$$\lim_{n\to\infty}\sum_{i=1}^{n}EZ_i(x,X_i,Y_i)\Big/\Big[\sum_{j=1}^{n}h_j^{-d}EK(\frac{X_j-x}{h_j})\Big]=m(x),\ a.e.(\mu)\ x\in R^d,\quad (14)$$

$$\sum_{n=1}^{\infty}E|Z_n(x,X_n,Y_n)\Big/\Big[\sum_{j=1}^{n}h_j^{-d}EK(\frac{X_j-x}{h_j})\Big]|^p=\sum_{n=1}^{n}E\Big|Y_nh_n^{-d}K(\frac{X_n-x}{h_n})\Big/\Big[\sum_{j=1}^{n}h_j^{-d}EK(\frac{X_j-x}{h_j})\Big]\Big|^p\leqslant$$

$$\sum_{n=1}^{\infty}C_1^{-p}(x)k^p(nh_n^d)^{-p}<\infty.\ a.e.(\mu)\ x\in R^d.$$

所以由引理 1 知

$$\lim_{n\to\infty}\sum_{i=1}^{n}(Z_i(x,X_i,Y_i)-EZ_i(x,X_i,Y_i))\Big/\Big[\sum_{j=1}^{n}h_j^{-d}EK(\frac{X_j-x}{h_j})\Big]\overset{a.s}{=}0,$$
$$a.e.(\mu)\ x\in R^d.\quad (15)$$

结合式(14)可知

$$\lim_{n\to\infty}U_n(x)\overset{a.s}{=}m(x)\ a.e.(\mu)\ x\in R^d.\quad (16)$$

由式(13)得

$$\sum_{n=1}^{\infty}E\Big|h_n^{-d}K(\frac{X_n-x}{h_n})\Big/\Big[\sum_{j=1}^{n}h_j^{-d}EK(\frac{X_j-x}{h_j})\Big]\Big|^p\leqslant\sum_{n=1}^{\infty}C^{-p}(x)k^p(nh_n^d)^{-p}<\infty.$$

所以由引理 1 知：

$$\lim_{n\to\infty}(V_n(x)-EV_n(x))\overset{a.s}{=}0,\ a.e.(\mu)\ x\in R^d,\quad (17)$$

从而

$$\lim_{n\to\infty}V_n(x)\overset{a.s}{=}1,\ a.e.(\mu)\ x\in R^d,\quad (18)$$

所以由式(16),(18)知

$$\lim_{n\to\infty} m_n(x) \overset{a.s.}{=} m(x), \text{a.e.}(P)\, x \in R^d.$$

证毕.

3 截尾样本时回归函数的核估计及改良核估计

当$(X_1,Y_1),\cdots,(X_n,Y_n)$不能完全观察到的情形,如有一独立于$Y_1,\cdots,Y_n$的$iid$随机变量$T_1,\cdots,T_n$干扰$Y_1,\cdots,Y_n$,即我们仅能观察到$Z_i=\min\{Y_i,T_i\}$以及$\delta_i=I(Y_i\leqslant T_i)$,也就是说,我们仅能看到$Y_i$和$T_i$中较小的一个,以及知道$Y_i$这个数据是否被截断.这类问题有广泛的实际意义,如在生存分析、人口统计、医药追踪试验、可靠性等方面.

设$(X_1,Y_1),\cdots,(X_n,Y_n)$为$(X,Y)$的$iid$样本,$T_i$是与$(X_i,Y_i)$独立的随机变量,且$\{T_i\}$相互独立,有共同的已知分布$G$,下面采用郑祖康提出的$K$类函数构造$m(x)$的截尾数据的核估计,先构造$Y_i^*$

$$Y_i^* = \delta_i\varphi_1(Z_i)+(1-\delta_i)\varphi_2(Z_i), i=1,2,\cdots, \tag{19}$$

其中φ_1,φ_2连续,与(X,Y)的分布无关,但可依赖于G,且满足

$$(1-G(y))\varphi_1(y)+\int_{-\infty}^{y}\varphi_2(t)G(dt)=y, \text{任}y \tag{20}$$

基于$(X_1,Y_1^*),\cdots,(X_n,Y_n^*)$,我们可构造$m(x)$的核估计

$$\hat{m}_n(x)=\sum_{i=1}^{n}Y_i^*h_i^{-d}K\left(\frac{X_i-x}{h_i}\right)\Big/\sum_{j=1}^{n}h_j^{-d}K\left(\frac{X_j-x}{h_j}\right). \tag{21}$$

定理2 设核函数K满足条件(c),(d),存在$1<p\leqslant 2$,使$E|Y^*|^p<\infty$,则当

$$\sum_{n=1}^{\infty}(nh_n^d)^{-p}<\infty$$

则有

$$\lim_{n\to\infty}\hat{m}_n(x)\overset{a.s.}{=}m(x), \text{a.e.}(P)\, x\in R^d.$$

证明 在满足式(19),(20)下,有

$$E(Y_i^*\mid X_i=x)=E(Y_i\mid X_i=x), i=1,2\cdots, \tag{22}$$

且$(X_1,Y_1^*),(X_2,Y_2^*),\cdots$仍为iid,于是由定理1知定理2结论成立.对于一般的回归函数的改良核估计首先由成平教授在文献[8]中提出,它的优点在于对Y的矩不需作其它要求,因此具有更广泛的这际意义.

当Y^*的P阶矩不存在时($P>1$),我们可采用下面的改良核估计:

$$\overline{m}_n(x)=\sum_{i=1}^{n}Y_i^*h_i^{-d}I(|Y_i^*|\leqslant b_i)\cdot K\left(\frac{X_i-x}{h_i}\right)\Big/\sum_{j=1}^{n}h_j^{-d}K\left(\frac{X_j-x}{h_j}\right). \tag{23}$$

定理3 设$E|Y|<\infty$,$0<b_n\to\infty$,核函数K满足条件(c),(d),若存在$1<p\leqslant 2$,使

$$\sum_{n=1}^{\infty}\left(\frac{b_n}{nh_n}\right)^p<\infty,$$

则有

$$\lim_{n\to\infty}\overline{m}_n(x)\overset{a.s.}{=}m(x), \text{a.e.}(P)\, x\in R^d.$$

证明 由$\sum_{n=1}^{\infty}\left(\frac{b_n}{nh_n}\right)^p<\infty$,故

$$\sum_{n=1}^{\infty}(nh_n)^{-p}<\infty,$$

由式(18)可知

$$\lim_{n\to\infty}\left(\sum_{i=1}^{n}h_i^{-d}K\left(\frac{X_i-x}{h_i}\right)\Big/\sum_{j=1}^{n}h_j^{-d}EK\left(\frac{X_j-x}{h_j}\right)\right)\overset{a.s.}{=}1, \text{a.e.}(P)\, x\in R^d. \tag{24}$$

记

$$Z_n(x,X_n,Y_n^*)=[h_n^{-d}(Y_n^*I(|Y_n^*|\leqslant b_n)-m(x))]K\left(\frac{X_n-x}{h_n}\right),$$

则

$$\overline{m}_n(x)-m(x)=\sum_{i=1}^{n}Z_i(x,X_i,Y_i^*)\Big/\sum_{j=1}^{n}h_j^{-d}K\left(\frac{X_j-x}{h_j}\right),$$

$$\sum_{n=1}^{\infty}E|Z_n(x,X_n,Y_n^*)\Big/\sum_{j=1}^{n}h_j^{-d}EK\left(\frac{X_j-x}{h_j}\right)|^p\leqslant C_2(x)\sum_{n=1}^{\infty}\left(\frac{b_n}{nh_n^d}\right)^p<\infty.$$

所以由引理1知

$$\lim_{n\to\infty}\sum_{i=1}^{n}(Z_i(x,X_i,Y_i^*)-EZ_i(x,X_i,Y_i^*))\Big/\sum_{j=1}^{n}h_j^{-d}EK\left(\frac{X_j-x}{h_j}\right)\overset{a.s.}{=}0 \text{a.e.}(P)\, x\in R^d. \tag{25}$$

记

$$g_{bi}(x)=E(|Y_i^*|I(|Y_i^*|>b_i)\mid X_i=x),$$

则

$$|EZ_n(x,X_n,Y_n^*)|/(h_n^{-d}EK(\frac{X_n-x}{h_n}))$$
$$\leqslant|E[h_n^{-d}K(\frac{X_n-x}{h_n})(Y_n^*-m(x))]|/$$
$$(h_n^{-d}EK(\frac{X_n-x}{h_n}))+|E(g_{bn}(x)h_n^{-d}\cdot$$
$$K(\frac{X_n-x}{h_n})|/(h_n^{-d}EK(\frac{X_n-x}{h_n})). \quad (26)$$

由引理 2 及式(22)可知

$$\lim_{n\to\infty}E(h_n^{-d}K(\frac{X_n-x}{h_n})(Y_n^*-m(x)))/$$
$$(h_n^{-d}EK(\frac{X_n-x}{h_n}))\overset{a.s}{=}0,\ a.e.(\mu)x\in R^d, \quad (27)$$

$$\lim_{n\to\infty}E[g_N(x)h_n^{-d}K(\frac{X-x}{h_n})]/(h_n^{-d}EK(\frac{X-x}{h_n}))$$
$$=g_N(x),\ a.e.(\mu)x\in R^d.$$

又 $b_n\to\infty$,故当 $b_n>N$ 时

$$E(g_{b_n}(x)h_n^{-d}K(\frac{X_n-x}{h_n}))\leqslant$$
$$E[g_N(x)K(\frac{X-x}{h_n})h_n^{-d}]. \quad (28)$$

另外,由 $E|Y|<\infty$ 及式(22)推出 $E|Y_n^*|<\infty$,故

$$\lim_{N\to\infty}Eg_N(x)=0,\ a.e.(\mu)x\in R^d. \quad (29)$$

由式(26)~(29)及 Lebesgue 控制收敛定理,我们先令 $n\to\infty$,再令 $N\to\infty$,可得

$$\lim_{n\to\infty}EZ_n(x,X_n,Y_n^*)/(h_n^{-d}EK(\frac{X_n-x}{h_n}))$$
$$\overset{a.s}{=}0,\ a.e.(\mu)x\in R^d. \quad (30)$$

再由式(13),结合 Toeplitz 引理得

$$\lim_{n\to\infty}\sum_{i=1}^{n}EZ_i(x,X_i,Y_i^*)/\sum_{j=1}^{n}h_j^{-d}EK(\frac{X_j-x}{h_j})$$
$$\overset{a.s}{=}0,\ a.e.(\mu)x\in R^d. \quad (31)$$

故由式(24),(26),(31)可得:

$$\lim_{n\to\infty}\overline{m}_n(x)\overset{a.s}{=}m(x),\ a.e.(\mu)x\in R^d.$$

定理 3 证毕.

参考文献

[1] 胡舒合.完全与截尾样本时回归函数的核估计[J].数学年刊,1995,16(3):269-274.

[2] 胡舒合.分布自由的回归函数核估计的相合性[J].应用数学学报,1997,20(3):448-455.

[3] GREBICKE W, KRZRZAK A, PAWLAK M. Distribution-free point wise consistency of kernel regression estimate[J]. Ann Statist, 1974, 12, 1570-1575.

[4] AHMADI A, LIN P. Nonparametric sequential estimation of a multiple regression function[J]. Bull Math Statist, 1976, 17: 63-75.

[5] 孙东初.回归函数核估计的强相合性[J].数学年刊,1985,6(4):481-486.

[6] 胡舒合.回归函数改良核估计的相合性[J].系统科学与数学,1993,13(2):141-151.

[7] DEVROYE L P. On the almost everywhere convergence of nonparametric regression function estimates [J]. Ann Statist, 1981, 9: 1310-1319.

[8] 成　平.回归函数改良核估计的强相合性及收敛速度[J].系统科学与数学,1983,3(4):304-315.

Distribution Free Convergence of Nonparametric Regression Function Kernel Estimates

HU Yu-ping

(Department of Mathematics, Physics & Mechanics, Zhengzhou University of Technology, Zhengzhou 450002, China)

Abstract: Let (X,Y) be a $R^d\times R^1$-Valued random vector with $E(|Y|)<\infty$, $m(x)=E(Y|X=x)$ be the regression function of Y with respect to X. Suppose that (X_i,Y_i), $i=1,\cdots,n$ are iid samples drawn from (X,Y). It is desired to estimate $m(x)$ based on these samples. Based on complete and censored data, we obtain point wise consistency of regression function kernel estimates under suitable conditions, the results are distribution-free in the sense that they are true for all distributions μ of X.

Key words: censored data; regression function; kernel estimate; distribution free