分布函数的非参数局部多项式估计

马鸿杰,胡汉辉,吴 崇
(东南大学 经济管理学院,南京 210096)

摘 要:文章讨论了当分布函数 F(x)经过- log(- log(F(x)))变换后,可以采用局部多项式逼近;使用非参数分布的泛函估计,建立了分布函数的强相合估计。模拟结果表明,估计拟合情况要优于经验分布,较为理想。

关键词:局部多项式;非参数估计;变换;相合性
中图分类号:O212. 文献标识码:A 文章编号:1002- 6487(2008)11- 0010- 03

0 引言

设 $X_1,X_2,\cdots,X_n$ 是来自 F(x)的独立样本,若已知 F(x)的泛函形式,但其中含有未知参数 $\theta_1,\theta_2,\cdots,\theta_r$ 时,就可以使用矩估计法,但这种方法并不渐进有效,因而有 Soong[1]的混合矩估计法。若完全未知时,应用最广泛,最直接的是经验分布函数估计,这种方法已有许多著名的大样本结果。在 F(x)的泛函形式未知时,但可用[0,1]上的多项式逼近,郑祖康已提出一种混合矩估计法,并建立了估计量的强相合性。而郑的方法假定其分布函数的区间为[0,1],限制了它的应用范围。柴根象[3]提到的 logsit 变换多项式逼近,柴根象[4]使用了核函数估计,但其效果还有待进一步提高。

本文则讨论 F(x)的泛函形式未知时,使用局部多项式方

基金项目:国家自然科学基金资助项目(704730103)

$$n=\frac{S^2[1+(K-1)r_2]}{V_0}=1000=\sqrt{\frac{100\times(1-0.3)}{2+20\times0.7}}[1+(2.092-1)0.3]$$
$$\approx1328$$

$$m=\frac{n_2}{K}=\frac{1328\times0.3}{3.092}\approx190$$

即需要事先寄出 n=1328 份问卷,预计会有 $n_2=1328\times0.3=398$ 名被访者无应答,需要从中抽出 190 名,派出调查员进行当面访问。邮寄问卷和当面访问总的调查费用为:

$$C=c_0n+c_1n+c_2m=c_0n+c_1(n-n_2)+c_2m$$
$$=2\times1328+20\times(1328-398)+100\times190=40256(元)$$

4 结语

通过以上分析不难看出,在对无应答单位进行替换的调查中,计算应答率时必须考虑无应答替换的影响,而实现这一点的条件是全面保留无应答替换的相关资料。从目前国内由研究者自行主持的调查实践看,一些调查并未做到全面保留无应答替换的相关资料,以至于无法准确计算出调查应答率。希望此番有关应答率计算的讨论,能使众人对无应答替换现象引起足够的重视,在今后的调查中,制定出保留调查执行资料的程序,严格执行之。

其次,有关无应答替换的讨论,还提示我们关注无应答现象产生偏差的可能性。在目前大多数调查中,‘明显替换”是经常被采用的替换方法。由于替换无应答单位的替换样本是从总体中独立抽取来的,因此,在每一轮无应答替换中,总是用那些‘应答单位’来代替那些‘无应答单位”。此时,如果调查中的‘应答单位’和‘无应答单位’在样本特征上存在着实质性的差异,那我们极有可能漏掉那些特征不同的‘无应答单位’的答案,使参数估计出现偏差,特别是当无应答单位比例较高时更是如此。如何走出用‘应答单位’替换‘无应答单位’的困境,降低偏差出现的可能性,则是“二重抽样’和其他缺失值处理方案的努力方向。

最后,任何讨论抽样的文章都应该强调:社会调查应答率是由多种因素决定的,其中有宏观社会结构方面的因素(比如人口流动、犯罪率、社会信任等),也有地域/社区层面的影响(碰到高层公寓,封闭式宿舍,等等),还有调查者研究方案设计和实施方面的因素。提高应答率的唯一办法就是尽可能完善设计与实施方案,想方设法克服那些能够被克服的困难(通过与政府相关部门合作,提供适当的激励机制,想办法增加调查员与被访人之间的互信,等等),坦然面对那部分由于结构性因素产生的无应答。

参考文献:
[1]杜子芳.抽样技术及其应用[M].北京:清华大学出版社,2005.
[2]樊鸿康.抽样调查[M].北京:高等教育出版社,2000.
[3]福勒(Floyd J. Fowler. Jr).调查研究方法(第 3 版)[M].重庆:重庆大学出版社,2004.
[4]金勇进,蒋研,李序颖.抽样技术[M].北京:中国人民大学出版社,2002.
[5]扎如(Ronald Czaja),布莱尔(Johnny Blair).抽样调查设计导论[M].重庆:重庆大学出版社,2007.

(责任编辑/亦 民)

法估计，去掉了关于支持集为[0,1]的限制。提供的方法可解决许多保险精算模型的推断问题。

1 估计方法及其一些基本假定

假设(a):$g(x)=-\log(-\log(F(x)))$在 R 上有 m+1 阶连续导数，在 x_0 点泰勒展开

$$g(x)=-\log(-\log(F(x)))\approx g(x_0)+g'(x_0)(x-x_0)+\cdots+\frac{g^{(m)}(x_0)(x-x_0)^m}{m!}$$

令$=\beta_i(x_0)=\frac{g^{(i)}(x_0)}{i!}$ $(i=0,1,\cdots,m$ $\beta(x_0)=(\beta(x_0),\beta_1(x_0),\cdots,\beta_m(x_0)))$

(b):F(x)有密度函数 r(x)

记 $X_{(1)},X_{(2)},\cdots,X_{(3)}$是样本 $X_1,X_2,\cdots,X_n$ 的顺序统计量，由假设(b)可得

$$E(F(X_{(i)}))=\frac{i}{1+n}\quad (i=1,2,\cdots,n)$$

记 $Y_i=-\log(-\log(\frac{i}{1+n}))(i=1,2,\cdots,n)$

则可以用 Y_i 作为$-\log(-\log(F(x)))$的近似，考虑到局部多项式法则，极小以下目标函数：

$$L(\beta_0(x_0),\cdots,\beta_m(x_0))$$

$$\hat{=}\sum_{i=1}^{n}(Y_i-(\beta_0(x_0)+\beta_1(x_0)(X_{(i)}-x_0)+\cdots+\beta_m(x_0)(X_{(i)}-x_0)^m))^2K_h(X_{(i)}-x_0) \quad (1)$$

$$X(x_0)=\begin{bmatrix}1 & X_{(1)}-x_0 & \cdots & (X_{(1)}-x_0)^m\\ \vdots & \vdots & \cdots & \vdots\\ 1 & X_{(n)}-x_0 & \cdots & (X_{(n)}-x_0)^m\end{bmatrix}$$

$$Y=(Y_1,\cdots,Y_n)'$$

$$\hat{\beta}(x_0)=(\hat{\beta}_0(x_0),\hat{\beta}_1(x_0),\cdots,\hat{\beta}_m(x_0))'$$

$$W(x_0)=\mathrm{diag}(K_h(X_{(i)}-x_0))\quad K_h(x)=h^{-1}K(\frac{x}{h})$$

则可表示极小问题解为

$$\hat{\beta}(x_0)=(X(x_0)',W(x_0)X(x_0)^{-1}X(x_0)'W(x_0)Y \quad (2)$$

记 $a_0=\sup\{x:F(x)=0\},b_0=\inf\{x:F(x)=1\}$, $g(x)=-\log(-\log(F(x)))$

则 g(x)仅仅在(a_0,b_0)上有定义，今固定$(a,b)\subset(a_0,b_0)$，于是 g(x)在[a,b]上处处连续，且在(a,b)上可微。

2 主要结果及其证明

令 $A(x_0)=X(x_0)'W(x_0)X(x_0)$, $B(x_0)=X(x_0)'W^2(x_0)X(x_0)$

$S_\nu'(x_0)=e_{\nu+1}'(X(x_0)'W(x_0)X(x_0))^{-1}X(x_0)'W(x_0)$, $H=\mathrm{diag}[1,h,\cdots,h^m]$

$u_j=\int u^jK(u)du$, $\nu_j=\int u^jK^2(u)du$

$A_1=(u_{i+j})$ $0\le j,i\le m$, $A_2=(\nu_{i+j})$ $0\le j,i\le m$

此外本文需要如下假设：

①核函数 K(t)连续，有界变差，有紧支撑[-1,1], $A_1=(u_{i+j})$ $0\le j,i\le m$非奇异。

②$nh^2/\log n\to\infty, nh^{2m+4}\to 0 (n\to\infty)$

③X_1 有连续密度 r(t)，且 $0<\inf_{0\le t\le 1} r(t)\le \sup_{0\le t\le 1} r(t)<\infty$

④g(x)有 m+1 阶连续有界导数

本文证明采用的矩阵范数为 F-范数，即若 $A=(a_{ij}),1\le i,j\le n$,则$\|A\|_F^2=\sum_i\sum_i a_{ij}^2$

引理 1

$$\sup_{-\infty<t<\infty}|Fn(u)-F(u)|=O((n^{-1}\log\log n)^{1/2})\quad a.s \quad (3)$$

其中 $F_n(u)$为经验分布函数，F(u)为分布函数

证明：见文献[5]。

引理 2

设 L(x)是实函数，令 $L_h(x)=h^{-1}L(x/h)$。若存在正常数 C 使对所有 $x,y\in R$,有 $|L(x)-L(y)|\le C|x-y|$, $\sup_{x\in R}|L(x)|<\infty$, $\int|L(x)|dx<\infty$及条件(②)(③)成立，则

$$\sup_{a<x<b}\left|n^{-1}\sum_{i=1}^{n}L_h(X_i-x)-r(x)\int_{a<x<b}L(u)du\right|=o(1)\quad a.s \quad (4)$$

证明：因为 $\sup_{a<x<b}\left|n^{-1}\sum_{i=1}^{n}L_h(X_i-x)-r(x)\int L(u)du\right|\le\sup_{a<x<b}$

$$\left|\int L_h(u-x)dF_n(u)-\int L_h(u-x)r(u)du\right|+$$

$$\sup_{a<x<b}\left|\int r(u)L_h(u-x)du-r(x)\int L(u)du\right|\hat{=}I_{11}+I_{12}$$

现只需证 $I_{11}=o(1),I_{12}=o(1)$ a.s

对 I_{11}，利用引理 1 和条件②

$$I_{11}=\sup_{a<x<b}\left|\frac{1}{h}\int L(u)d(F_n(x+hu)-F(x+hu))\right|\le\frac{C_2}{h}\sup_{a<x<b}$$

$|F_n(u)-F(u)|=O((n^{-1}h^{-2}\log\log n)^{1/2})=o(1)$,a.s($C_2$ 为常数)

现在考虑 I_{12} 。

由于 r(t)在(a,b)上连续，所以对 $\forall\varepsilon>0$ 存在 $a_1=a_1(\varepsilon)>0$，使得对$|x-y|<a_1$,有$|r(x)-r(y)|<\varepsilon$,又有 $\int|L(x)|dx<\infty$可知，存在 $a_2=a_2(\varepsilon)>0$

使得 $\int_{|u|\ge a_2}|L(u)||du\le\varepsilon$ |对于 a_1,a_2 存在 $n_1=n_1(t)>0$ 使得对所有 $n>n_1$,有 $ha_2\le a_1$,这样对所有 $n>n_1$ 和 $|u|<a_2$ 有 $x+hu-x=hu\le a_1$ 因而当 $n>n_1$ 时

$$\sup_{a<x<b}\int_{|u|\ge a_2}|r(x+hu)-r(x)|\left|L(u)du\le\varepsilon\int\right|L(u)|du=\varepsilon C_3,C_3,\text{为常数}$$

所以 $\sup_{a<x<b}\left|\int r(u)L_h(u-x)du-r(x)\int L(u)du\right|\le\sup_{a<x<b}\int_{|u|\ge a_2}$

$$|r(x+hu)-r(x)||L(u)|du+\sup_{a<x<b}\int_{|u|\ge a_2}|r(x+hu)-r(x)||L(u)|du\le\varepsilon 2$$

$$\max_{a<x<b}|r(x)|+C_3\varepsilon$$

令 $\varepsilon\to 0$,r(x)的有界性可知 $I_{12}=o(1)$

故引理 2 得证。

引理 3

若条件①、②、③满足，则

$$\sup_{a<x_0<b}\|n^{-1}H^{-1}A(x_0)H^{-1}-r(x_0)A_1\|_F=o(1) \quad a.s \quad (5)$$

$$\sup_{a<x_0<b}\|nHA^{-1}(x_0)H-r(x_0)^{-1}A_1^{-1}\|_F=o(1) \quad a.s \quad (6)$$

$$\sup_{a<x_0<b}\|n^{-1}hH^{-1}B(x_0)H^{-1}-r(x_0)A_2\|_F=o(1) \quad a.s \quad (7)$$

$$\sup_{a<x_0<b}\sum_{k=1}^{n}S^2_{\nu k}(x_0)=O(n^{-1}h^{2\nu-1}) \quad a.s \quad (8)$$

$$\max_i \sup_{a<x_0<b}|S_{\nu i}(x_0)|=O(n^{-1}h^{\nu-1}) \quad a.s \quad (9)$$

证明：（略）。

引理 4

$$\sup_{a<x_0<b}\left|\beta_\nu(x_0)-\sum_{i=1}^{n}S_{\nu i}(x_0)g(x_{i1})\right|=0\ (h^{m+1-\nu}),\quad a.s \quad (10)$$

证明：（略）。

定理 1

在条件①、③下，若 $n^{-1}h^{2\nu}\log\log n\to 0$

则 $\sup_{a<x_0<b}\left|\hat\beta_\nu(x_0)-\beta_\nu(x_0)\right|\to 0$，a.s $\nu=0,1,\cdots,m$

证明：（略）。

定理 2 若 $\overline{\lim}_{n\to\infty}\frac{(\log\log n)^{1/2}}{n^{1/2}h^{m+1}}<\infty$，有

$$\sup_{a<x_0<b}\left|\hat\beta_\nu(x_0)-\beta_\nu(x_0)\right|=O(h^{m+1-\nu}) \ a.s$$

证明：

知若 $(n^{-1}h^{2\nu}\log\log n)^{1/2}=O(h^{m+1-\nu})$，定理 2 的结论等价于：

$(n^{-1}\log\log n)^{1/2}=O(h^{m+1})$

由条件可推出 $\overline{\lim}_{n\to\infty}\frac{(\log\log n)^{1/2}}{n^{1/2}h^{m+1}}<\infty$，因而定理 2 得证。

推论 1 若定理 2 的条件成立，则有

$$\sup_{a<x_0<b}\left|\hat F_\nu(x_0)-F_\nu(x_0)\right|=O(h^{m+1-\nu}) \quad a.s$$

证明：由(*)式知 $\beta(x_0)=-\log(-\log(F(x_0))$

故 $\hat\beta(x_0)=-\log(-\log(\hat F(x_0)))$

$F(x_0)=\exp(-\exp(-\beta(x_0)))$

$\hat F(x_0)=\exp(-\exp(-\hat\beta(x_0))$

又因为 $\exp(-\exp x)$在(a,b)上是连续可微的，由定理 2 可得

$$\sup_{a<x_0<b}\left|\hat F(x_0)-F(x_0)\right|=O(h^{m+1-\nu}) \quad a.s$$

3 多项式逼近法模拟结果

局部多项式方法各次计算采用的核函数均为：

$$K(t)=\begin{cases}\Phi(t)/(2\Phi(2.5h)-1) & |t|\le 2.5h\\ 0 & \text{其他}\end{cases}\quad(\Phi\text{为标准正态分布})$$

记号含义：

$F_2(x)$ 表示局部多项式法，$\overline{F}_2(x)$ 表示 $F_2(x)$ 估计的平均，$\hat\sigma_2$ 表示 $F_2(x)$ 估计的的标准偏差，h 分别取 $h_1=100^{-\frac{1}{8}}$，$h_2=200^{-\frac{1}{8}}$；

$F_n(x)$表示经验分布，$\overline{F}_n(x)$表示 $F_n(x)$估计的平均，$\hat\sigma_n$ 表示 $F_n(x)$估计的的标准偏差，F(x)表示理论值。

例 1. Φ(x)，Φ 为标准正态分布，样本容量 n=100，重复试验次数 s=10 组，h 分别取 h_1，h_2，局部多项式阶数 m=5 计算结果见表 1。

例 2. F(x)=exp(-exp(-x))，样本容量 n=100，重复试验次数 s=10 组 h 分别取 h_1，h_2，局部多项式阶数 m=5 计算结果见表 2。

表 1

x	-1.1379	-0.8892	-0.3571	-0.0513	0.1871	0.4307	0.7441	1.0538	1.3083	1.6742
F(x)	0.1276	0.1869	0.3605	0.4795	0.5742	0.6667	0.7716	0.8540	0.9046	0.9529
$\overline{F}n(x)$	0.1120	0.1810	0.3480	0.4900	0.5910	0.6830	0.7830	0.8630	0.9060	0.9440
$\hat\sigma_n$	0.0956	0.0752	0.0458	0.0088	0.0255	0.0327	0.0381	0.0438	0.0911	0.1004
$\hat\sigma_1(h_1)$	0.0899	0.0522	0.0355	0.0012	0.0122	0.0299	0.0301	0.0325	0.0897	0.0986
$\overline{F}_2(x)(h_2)$	0.1223	0.1820	0.3616	0.4801	0.5830	0.6701	0.7814	0.8572	0.9087	0.9423
$\hat\sigma 2(h_2)$	0.0956	0.0752	0.0458	0.0088	0.0255	0.0327	0.0381	0.0438	0.0911	0.1004

表 2

x	-0.7134	-0.5493	-0.2533	-0.0061	0.2603	0.5735	0.8572	1.2614	1.8396	3.9629
F(x)	0.1289	0.1769	0.2758	0.3656	0.4626	0.5692	0.6542	0.7533	0.8531	0.9812
$\overline{F}_n(x)$	0.1430	0.1880	0.2570	0.3540	0.4540	0.5680	0.6650	0.7720	0.8610	0.9880
$\overline{F}_2(x)(h_1)$	0.1440	0.1860	0.2775	0.3665	0.4619	0.5688	0.6595	0.7777	0.8626	0.9899
$\hat\sigma_n$	0.0293	0.0328	0.0360	0.0496	0.0522	0.0481	0.0488	0.0451	0.0421	0.0097
$\hat\sigma_2(h_1)$	0.0390	0.0399	0.0340	0.0355	0.0385	0.0376	0.0499	0.0500	0.0520	0.0120
$\overline{F}_2(x)(h_2)$	0.1428	0.1820	0.2764	0.3659	0.4620	0.5690	0.6556	0.7667	0.8600	0.9877
$\hat\sigma_2(h_2)$	0.0410	0.0421	0.0399	0.0455	0.0390	0.0432	0.0555	0.0544	0.0531	0.0189

注：①从计算机模拟结果看，本文所采用的局部多项式变换法与经验函数法相比，在平均值方面靠近点 x_0 好些，离 x_0 点远的点差些。在标准差方面也是如此。

②从计算机模拟结果看，h 的选择对结果也有一定的影响。h 小，在平均值方面会更好些，即与真值的误差会小，但是标准差更大。

③从计算机模拟结果看，多项式逼近法与经验函数相比，平均值方面相差无几，但是在标准差方面多项式逼近法要优于经验函数。

参考文献：

[1]Soong, T.T. An Extension of the Moment Method in Statistical Estimation, SIAM J.Appl.Math, 1969, (17).

[2]Zheng Zukang. Method of Moments with Unknown Distribution form[J].应用概率统, (1995),(3).

[3] 柴根象， 花虹.Nonparametric Least Square Estimation of Distribution Function,高校应用数学学报,(2001), 17(4).

[4]柴根象.相依样本分布函数，回归函数的非参数估计的强相合性[J].系统数学与科学, 1988, 8(3).

[5]Kiefer.J.On large Deviation of the Empiric Distribution Function Function of Vector Chance Varia and a law of the Iterated logarithm[J].Pacific.I.Math, 1961, (11).

（责任编辑/亦 民）