应用概率统计　第二十五卷
第三期　2009年6月

Chinese Journal of Applied Probability
and Statistics Vol.25 No.3 Jun. 2009

纵向数据边际模型非参数光滑方法的比较 *

秦国友
(复旦大学公共卫生学院卫生统计教研室, 上海, 200032)

朱仲义
(复旦大学统计系, 上海, 200433)

摘　要

对于纵向数据边际模型的均值函数, 有很多非参数估计方法, 其中回归样条, 光滑样条, 似乎不相关(SUR)核估计等方法在工作协方差阵正确指定时具有最小的渐近方差. 回归样条的渐近偏差与工作协方差阵无关, 而SUR核估计和光滑样条估计的渐近偏差却依赖于工作协方差阵. 本文主要研究了回归样条, 光滑样条和SUR核估计的效率问题. 通过模拟比较发现回归样条估计的表现比较稳定, 在大多数情况下比光滑样条估计和SUR核估计的效率高.

关键词: 回归样条, 光滑样条, SUR核, 纵向数据, 效率.

学科分类号: O212.

§1. 引　　言

在纵向数据情形下, 假定数据有n个个体, 每个个体有m次观察. 记Y_{ij}和X_{ij}分别表示第i个个体第j次观察所对应的响应变量和自变量$(i = 1, \cdots, n;\ j = 1, \cdots, m)$. 本文我们讨论如下的边际非参数回归模型:

$$\mathsf{E}(Y_{ij}|X_i) = \mu_{ij}(X_{ij}), \qquad \mathsf{Cov}\,(Y_i|X_i) = \Sigma, \tag{1.1}$$

其中$\mu(\cdot)$是未知函数, Σ是真实的协方差阵, $X_i = (X_{i1}, \cdots, X_{in_i})^T$, $Y_i = (Y_{i1}, \cdots, Y_{in_i})^T$.

很多文献讨论了模型(1.1)中均值函数$\mu(\cdot)$的估计问题. 采用的非参数回归估计方法包括核, 光滑样条和回归样条等估计方法. Zeger and Diggle (1994), Wu, Chiang and Hoover (1998)采用传统的核估计方法研究了纵向数据情形下回归估计问题. 他们采用了所谓的"工作独立"方法, 完全忽略了纵向数据个体内部可能存在的相关性. Lin and Carroll (2000)从理论上证明了采用"工作独立"方法的有效性. 他们指出在一类传统的非参数函数的核估计中, 完全忽略个体内部相关性得到的核估计是渐近有效的. Chen and Jin (2005) 讨论了如何利用个体内部的相关性来提高局部多项式估计的效率问题. Wang (2003)提出了似乎不相关(SUR)核估计方法并指出在采用SUR核时考虑个体内部的相关性可以提高估计的效率. Welsh et al. (2002)指出在采用光滑样条方法估计非参数回归函数时, 考虑个体内部的相关性可以减少估计的方差, 并且在假定真实协方差阵时, 估计的方差达到最小. 该文同时

*国家自然科学基金项目(10801039, 10671038)、复旦大学青年科学基金项目(08FQ29)、上海市重点学科建设项目(B118)资助.

本文2008年2月18日收到, 2008年4月21日收到修改稿.

指出在纵向数据情形下, 样条估计比传统核估计具有更高的效率. Lin et al. (2004)证明了光滑样条估计和Wang (2003)提出的SUR核估计是渐近等价的. 虽然在工作协方差阵正确指定时, SUR核估计和光滑样条估计方法可以得到最小的渐近方差, 但是这些估计的渐近偏差是依赖于工作协方差阵的, 并且两者之间的关系非常复杂. 所以一般很难确定当估计的方差达到最小时, 估计的偏差是否会增加. 因而难以获得一个最有效的估计, 即具有最小均方误差的估计.

最近, Zhu, Fung and He (2008)研究了纵向数据情形下, 边际模型中回归样条的局部性质. 他们发现并且证明了回归样条估计一个有趣的性质: 回归样条的渐近偏差与工作协方差阵无关. 同时, 在工作协方差阵被正确指定时, 估计的渐近方差能够达到最小. 因此, 与光滑样条和SUR核估计不同, 在工作相关阵被正确指定时, 回归样条估计的均方误差(MSE)能够达到最小. 以上三种非参数光滑方法, 从理论的角度, 核方法比较方便; 从计算的有效性来看, 样条方法比较有效, 快捷. 但是对估计的效率, 由于光滑样条和SUR核方法中偏差项与相关结构的关系难以确定, 从理论层面难以比较三种光滑方法的估计效率. 本文通过大量的计算机模拟分析比较这三种非参数光滑方法.

本文主要研究回归样条, 光滑样条与SUR核估计的效率问题. 我们主要基于估计的MSE来比较三种估计方法的效率. 本文其他部分安排如下: 第二部分分别介绍三种非参数估计方法; 第三部分进行随机模拟研究三种估计方法的效率问题. 通过模拟研究发现, 回归样条估计的表现比较稳定, 在大多数情况下比光滑样条估计和SUR核估计的效率高. 这与回归样条估计的性质也是一致的, 即回归样条估计的渐近偏差与工作协方差阵无关, 只要正确指定工作协方差阵可以获得具有最小MSE的估计.

§2. 非参数估计方法

在这一部分将给出本文主要讨论的三种非参数估计方法: 回归样条, 光滑样条和SUR核估计方法.

2.1 回归样条方法

样条实际上是一个分段的多项式, 通过节点将各段多项式相连接. 记$x_0 = 0 < x_1 < x_2 < \cdots < x_k < 1 = x_{k+1}$表示区间$[0,1]$上的一个划分, 其中$x_1, \cdots, x_k$表示$k$个可区分的点. 利用这些点作为节点, 我们可以得到$k+r$个阶数为r的B-样条基函数, $N_i(x), i = 1, \cdots, k+r$. 关于$N_i(x)$的具体形式可以参见Schumaker (1981). 我们把这些基函数记成向量形式, 表示为$\pi(x) = (N_1(x), \cdots, N_p(x))^T$, 其中$p = k + r$. 回归样条方法就是利用$\pi(x)^T\alpha$逼近$\mu(x)$, 其中$\alpha$是某个未知的系数.

令$S_i = (\pi(X_{i1}), \cdots, \pi(X_{ij}))^T$, $V(\gamma)$是依赖于未知参数γ的可逆的工作协方差阵. 于是通过最小化

$$\sum_{i=1}^{n}(Y_i - S_i\alpha)^T V^{-1}(Y_i - S_i\alpha)$$

可以得到α的估计,

$$\widehat{\alpha}=\Big(\sum_{i=1}^{n}S_i^TV^{-1}S_i\Big)^{-1}\sum_{i=1}^{n}S_i^TV^{-1}Y_i=(S^TV_D^{-1}S)^{-1}S^TV_D^{-1}Y,$$

其中$S=(S_1^T,\cdots,S_n^T)^T$, $Y=(Y_1^T,\cdots,Y_n^T)^T$, $V_D=\mathrm{diag}(V,\cdots,V)$. 进一步有$\widehat{\mu}_B=\pi(x)\widehat{\alpha}$. $\widehat{\mu}_B(x)$就是本文讨论的回归样条估计. 根据Zhu, Fung and He (2008)可知, $\widehat{\mu}_B$的渐近偏差与工作协方差阵无关. 当工作协方差阵正确指定时, $\widehat{\mu}_B$渐近方差最小, 同时$\widehat{\mu}_B$的MSE也达到最小.

2.2 似乎不相关核方法

由于传统的核估计方程方法无法考虑纵向数据个体内部可能存在的相关性, 于是Wang (2003)提出了似乎不相关(SUR)核估计方法. 记$K_h(s)=h^{-1}K(s/h)$, 其中$K(\cdot)$是一个零均值的核函数, h是窗宽. 对任意的矩阵A, 记a^{jk}是矩阵A^{-1}的第j行k列的元素. 现在考虑一个q阶多项式核估计. SUR核估计通过下面的迭代过程实现.

(1) 记$l=0$; 给定初值$\widehat{\mu}_K^{(0)}$.

(2) 利用下面的估计方程解得$\widehat{\mu}_K^{(l+1)}(x)=\widehat{\alpha}_0$, 其中$\widehat{\alpha}=(\widehat{\alpha}_0,\widehat{\alpha}_1,\cdots,\widehat{\alpha}_q)^T$,

$$\sum_{i=1}^{n}\sum_{j=1}^{m}K_h(X_{ij}-x)B_{ij}(x)^TV^{-1}(Y_i-\mu_{i(j)}(x))=0,$$

其中$B_{ij}(x)$是一个除了第j行为$[1,(X_{ij}-x),\cdots,(X_{ij}-x)^q]^T$, 其它元素为零的$m\times(q+1)$矩阵,

$$\mu_{i(j)}(x)=\Big[\widehat{\mu}_K^{(l)}(X_{i,1}),\cdots,\widehat{\mu}_K^{(l)}(X_{i,j-1}),\sum_{k=0}^{q}\alpha_k(X_{ij}-x)^k,\widehat{\mu}_K^{(l)}(X_{i,j+1}),\cdots,\widehat{\mu}_K^{(l)}(X_{i,m})\Big]^T.$$

(3) 令$\widehat{\mu}_K^{(l)}(x)=\widehat{\mu}_K^{(l+1)}(x)$.

(4) 重复(2), (3)两步直至估计收敛. 最终得到的核估计称为SUR核估计, 记为$\widehat{\mu}_K(x)$.

Wang (2003)证明了在采用SUR核时考虑个体内部的相关性可以提高估计的效率. 并且在协方差阵正确指定时, $\widehat{\mu}_K(x)$的方差可以达到最小. 但是$\widehat{\mu}_K(x)$的渐近偏差和工作协方差阵有着非常复杂的关系.

2.3 光滑样条方法

光滑样条方法也是一种常用的非参数估计方法. 记$X_o=(X_{(1)},\cdots,X_{(N)})^T$, X_o是所有协变量X_{ij}的值从小到大排序后得到的向量. 矩阵U是与X关联的矩阵. 如果$X_{ij}=X_{(k)}$, 则它的第$((i,j),k)$个元素为1, 否则为0 ($i=1,\cdots,n;j=1,\cdots,m;k=1,\cdots,N$). 即$X=UX_o$. 这样定义的$U$满足$UU^T=I$, 其中$I$是单位阵. 记$\mu=\mu(X_o)$. 假定工作协方差阵为$V$, 于是通过最小化

$$(Y-U\mu)V^{-1}(Y-U\mu)+\lambda\int(\mu^{(p)}(x))^2\mathrm{d}x,$$

可以得到$\mu(\cdot)$的p阶光滑样条估计

$$\widehat{\mu}_S(X) = (U^T {V_D}^{-1} U + \lambda\Psi)^{-1} U^T V_D^{-1} Y,$$

其中λ是光滑参数用来平衡函数拟合和光滑程度, Ψ是光滑矩阵. 详细内容可参考Welsh, Lin and Carroll (2002).

Lin et al. (2004)证明了光滑样条估计和Wang (2003)提出的SUR核估计是渐近等价的.

2.4 效率的比较

前面介绍了本文主要讨论的三种非参数估计方法. 三种方法的共同之处在于当工作协方差阵被正确指定时, 所得到的估计具有最小的方差; 差异在于回归样条估计的渐近偏差和工作协方差阵是无关的, 而另两种方法则是有关的. 对于回归样条估计, 我们只要指定了正确的工作协方差阵就可以得到具有最小均方误差(MSE)的估计. 而对于光滑样条和SUR核估计方法, 由于渐近偏差和工作协方差阵有关, 即使正确指定工作协方差阵也未必能得到具有最小MSE的估计. 另外, 虽然Lin et al. (2004)证明了光滑样条估计和SUR核估计的渐近等价性, 但是在有限样本情形下的表现如何也是有趣的问题. 所以研究这三种估计的效率问题, 从中发现效率较高的估计方法是有意义的.

§3. 模拟研究

本节我们通过随机模拟来研究文中给出的三种非参数估计的效率问题. 我们主要通过计算估计的均方误差(MSE)来比较估计的效率. 为了简单和计算结果的可比性, 对于回归样条中的节点数, 光滑样条方法中的光滑参数λ以及SUR核估计方法中的窗宽, 我们选择统一的标准进行选择. 由于模拟中的均值函数$\mu(x)$是已知的, 类似于Lin et al. (2004),我们通过最小化估计的MSE, 即$\sum(\widehat{\mu}(x) - \mu(x))^2$, 来得到最优的节点数, 光滑参数和窗宽. 另外, 为了计算结果的可比性, 我们在区间$[-1.8, 1.8]$中等距的选取101个格子点. 利用这些点, 我们计算三种不同估计的MSE并作相应的比较.

我们从模型$y_{ij} = \mu(x_{ij}) + e_{ij}$, $i = 1, \cdots, n$, $j = 1, \cdots, m$中产生模拟观察值, 其中$m = 3$, $n = 100$. 自变量X_{ij}从$[-2, 2]$的均匀分布上随机抽取的. 误差项$e_i = (e_{i1}, \cdots, e_{in_i})^T$从正态分布$N(0, R)$中随机产生, R为相关矩阵. 类似于Welsh et al. (2002), 我们考虑如下三种相关结构:

1. EX: 可交换结构, 相关系数$\rho = 0.6$;

2. AR: 一阶自回归结构, 相关系数$\rho = 0.6$;

3. US: 无结构的相关结构, 其中$\rho_{12} = \rho_{21} = 0.8$, $\rho_{13} = \rho_{31} = 0.5$, ρ_{jk}是e_{ij}和e_{ik}之间的相关系数, $j \neq k$.

令$z = (x + 2)/4$. 我们分别选取$\mu(x)$为如下6种函数:

模型1: $\mu(x) = \sin(x)$;

模型2: $\mu(x)=\exp(x)$;
模型3: $\mu(x)=\sin(2x)$;
模型4: $\mu(x)=\sqrt{z(1-z)}\sin(z\pi(1+2^{-3/5})/(z+2^{-3/5}))$;
模型5: $\mu(x)=\sqrt{z(1-z)}\sin(z\pi(1+2^{-7/5})/(z+2^{-7/5}))$;
模型6: $\mu(x)=\sin(8z-4)+2\exp\{-256(z-0.5)^2\}$.
对于上述各种模型, 我们都进行了200次模拟.

表1 三种估计方法采用工作独立阵时的估计与采用正确协方差阵和估计的协方差阵时估计之间MSE之比R_1

Model	Corr	True			Estimated		
		R-spline	S-spline	SUR K	R-spline	S-spline	SUR K
1	1	1.34	1.25	1.26	1.28	1.21	1.22
	2	1.33	1.25	1.26	1.29	1.21	1.22
	3	1.89	1.69	1.80	1.86	1.64	1.75
2	1	1.30	1.25	0.99	1.23	1.19	0.97
	2	1.39	1.31	1.01	1.36	1.28	1.01
	3	1.76	1.65	1.32	1.73	1.60	1.31
3	1	1.36	1.30	1.31	1.31	1.24	1.26
	2	1.35	1.30	1.31	1.30	1.27	1.28
	3	1.94	1.79	1.92	1.84	1.71	1.84
4	1	1.32	1.27	1.27	1.29	1.25	1.25
	2	1.35	1.31	1.31	1.33	1.29	1.29
	3	1.93	1.76	1.79	1.87	1.73	1.78
5	1	1.36	1.31	1.31	1.34	1.29	1.30
	2	1.44	1.37	1.39	1.41	1.34	1.35
	3	2.41	2.07	2.11	2.32	2.03	2.08
6	1	1.51	1.46	1.42	1.48	1.43	1.39
	2	1.47	1.42	1.41	1.43	1.38	1.37
	3	2.67	2.43	2.27	2.57	2.33	2.28

True: 表示采用正确的协方差阵; Estimated: 表示采用估计的协方差阵;
Corr=1: 表示相关结构EX; Corr=2: 表示相关结构AR;
Corr=3: 表示相关结构US; R-spline: 表示回归样条估计;
S-spline: 表示光滑样条估计; SUR K: 表示看似不相关核估计.

表1给出了三种方法分别采用真实协方差阵和估计的协方差阵时得到的估计MSE之间比较的结果. 这里通过矩估计的方法得到工作相关阵的估计. 表1中比值的定义为

$$R_1=\frac{\mathrm{MSE}_W}{\mathrm{MSE}_D},$$

其中MSE_W表示采用工作独立协方差阵时得到的估计MSE, MSE_D表示采用真实协方差阵或估计的协方差阵时得到的估计MSE. 所以R_1值越大说明采用真实协方差阵时得到的估计效率越高. 由表1可以发现, 在采用正确的协方差阵和估计的协方差阵时, 三种估计的效率都有提高, 与理论的结果一致. 特别地, 回归样条估计效率提高的最多. 值得注意的是在模型2中, 相关结构为EX和AR时, SUR核估计的效率似乎没有明显的提高. 我们发现在此种模型下, 虽然在工作协方差阵被正确指定时SUR核估计的方差减少了, 但是其偏差却增大了, 故导致最终采用正确协方差阵的估计效率并没有显著的提高. 说明由于该估计的渐近偏差与工作协方差阵有关系, 所以即使正确指定了协方差阵也未必可以得到具有最小MSE的估计.

表2 采用正确协方差阵和估计的协方差阵时, 光滑样条和SUR核估计与回归样条估计MSE之比R_2

Model	Corr	True		Estimated	
		S-spline	SUR K	S-spline	SUR K
1	1	0.96	0.98	0.95	0.97
	2	0.95	1.01	1.81	1.01
	3	0.99	0.98	1.00	0.99
2	1	1.06	1.60	1.05	1.55
	2	1.12	1.79	1.13	1.76
	3	1.06	1.63	1.08	1.62
3	1	1.07	1.25	1.07	1.24
	2	1.08	1.25	1.07	1.24
	3	1.12	1.21	1.12	1.20
4	1	1.03	1.05	1.02	1.04
	2	1.03	1.05	1.03	1.05
	3	1.06	1.05	1.06	1.03
5	1	1.01	0.99	1.79	1.00
	2	1.02	0.99	1.01	0.99
	3	1.10	1.07	1.94	1.05
6	1	0.94	0.97	0.94	0.97
	2	0.96	0.98	0.96	0.98
	3	1.04	1.10	1.05	1.06

表2给出了采用正确的协方差阵和估计的协方差阵时, 三种估计之间MSE的比较. 表2中的比值定义为

$$R_2 = \frac{\mathrm{MSE}_C}{\mathrm{MSE}_B},$$

其中MSE_C表示由光滑样条估计或SUR核估计得到的MSE, 而MSE_B表示由回归样条得到的MSE. 当R_2大于1时, 表示回归样条估计的效率高. 由表2可以发现, 在大多数情况下, 采用正确的协方差阵和估计的协方差阵时, 回归样条估计的效率比其它两种方法的效率要高. 特别在模型2下, 与SUR核估计相比, 回归样条估计的效率有明显的提高.

参 考 文 献

[1] Chen, K. and Jin, Z.H., Local polynomial regression analysis of cluster data, *Biometrika*, **92**(2005), 59–74.

[2] Lin, X. and Carroll, R.J., Nonparametric function estimation for clustered data when the predictor is measured without/with error, *J. Am. Statist. Assoc.*, **95**(2000), 520–534.

[3] Lin, X., Wang, N., Welsh, A.H. and Carroll, R.J., Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data, *Biometrika*, **91**(2004), 177–193.

[4] Schumaker, L.L., *Spline Functions: Basic Theory*, New York: Wiley, 1981.

[5] Wang, N., Marginal nonparametric kernel regression accounting for within-subject correlation, *Biometrika*, **90**(2003), 43–52.

[6] Welsh, A.H., Lin, X. and Carroll, R.J., Marginal longitudinal nonparametric regression: locality and efficiency of spline and kernel methods, *J. Am. Statist. Assoc.*, **97**(2002), 482–493.

[7] Wu, C.O., Chiang, C.T. and Hoover, D.R., Asymptotic confidence regions for kernel smoothing of a varying coefficient model with longitudinal data, *J. Am. Statist. Assoc.*, **93**(1998), 1388–1402.

[8] Zeger, S.L. and Diggle, P.J., Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters, *Biometrics*, **50**(1994), 689–699.

[9] Zhu, Z.Y., Fung, W.K. and He, X., On the local asymptotics of marginal regression splines with longitudinal data, *Biometrika*, **95**(2008), 907–917.

The Comparison of Nonparamteric Smoothing Methods for Longitudinal Data

Qin Guoyou

(*Department of Biostatistics, School of Public Health, Fudan University, Shanghai, 200032*)

Zhu Zhongyi

(*Department of Statistics, Fudan University, Shanghai, 200433*)

There are many nonparametric estimation methods for the mean functions of marginal models for longitudinal data. Those estimators such as regression spline, smoothing spline and seemingly unrelated(SUR) kernel estimators can achieve the minimum asymptotic variance when the true covariance structure is specified. The asymptotic bias of the regression spline estimator does not depend on the working covariance matrix, but the asymptotic bias of smoothing spline and SUR kernel estimators depend on the working covariance matrix in a complicate manner. In this paper, we focus on the comparison of the estimation efficiency among the regression spline, smoothing spline and SUR kernel estimators. By simulation study, it is found that the regression spline estimator generally present higher efficiency than the other two estimators with smaller mean square errors.

Keywords: Regression spline, smoothing spline, SUR kernel, longitudinal data, efficiency.

AMS Subject Classification: 62G08.