应用概率统计　第二十九卷
第二期　2013年4月

Chinese Journal of Applied Probability
and Statistics Vol.29 No.2 Apr. 2013

纵向数据下变系数EV模型的光滑核估计 *

杨宜平　　李　佳

(重庆工商大学数学与统计学院, 重庆, 400067)

摘　要

考虑纵向数据下变系数EV模型, 提出了未知系数函数的局部纠偏光滑核估计, 在适当条件下, 证明了所提出的光滑核估计具有渐近正态性, 由此构造回归系数的逐点置信区间. 模拟研究了所提出方法的有限样本性质.

关键词: 变系数模型, 核估计, 置信区间, 渐近正态, 纵向数据.

学科分类号: O212.7.

§1. 引　　言

在生物医学和经济管理等领域中常常会收集到大量的纵向数据, 它是指对同一受试个体在不同时间点上的观察, 目前已有大量文献研究此类数据. 考虑来自n个个体的数据, 其第i个个体具有$n_i\,(i=1,\ldots,n)$次观察, 总的观测次数为$N=\sum\limits_{i=1}^{n}n_i$. 设$Y_i(t_{ij})$和$X_i(t_{ij})$分别是第$i$个个体在时间点$t_{ij}\,(j=1,\ldots,n_i)$上观测的响应变量和协变量, 其中$Y_i(t_{ij})$是实值变量, $X_i(t_{ij})$是$k\times 1$向量而t_{ij}是数量或时间. 响应变量和协变量的依赖关系由下式给出:

$$Y_i(t_{ij})=X_i^T(t_{ij})\beta(t_{ij})+\varepsilon_i(t_{ij}),\qquad i=1,\ldots,n,\ j=1,\ldots,n_i, \tag{1.1}$$

其中$X_i(t)=(X_{i1}(t),\ldots,X_{ik}(t))^T$是$t$的实值协变量, $\beta(t)=(\beta_1(t),\ldots,\beta_k(t))^T$是未知回归系数向量, 误差$\varepsilon_i(t)$是均值为0的随机过程且$\varepsilon_i(t)$是独立的.

变系数模型在纵向数据下的研究引起了许多统计学者的兴趣. Hoover等(1998)采用光滑样条和局部加权多项式方法研究了函数系数的估计; Wu等(1998)通过最下化局部加权最小二乘准则得到了函数系数的核估计及其渐近性质, 并讨论了函数系数的置信带; Huang等(2002)利用基函数逼近的思想把每一个系数函数转化成无限维的参数, 基于Bootstrap方法构造了函数系数的置信域; Xue和Zhu (2007)提出了均值校正和残差调整的经验似然, 证明了所构造的经验对数似然比渐近于卡方分布, 由此构造函数系数的置信带.

在实际应用中, 协变量$X_i(t_{ij})$的精确值往往不能直接观测到, 观测到的是带有测量误差的观测变量$W_i(t_{ij})$, 即

$$W_i(t_{ij})=X_i(t_{ij})+U_i(t_{ij}), \tag{1.2}$$

*国家自然科学基金数学天元青年基金(11126332)、国家社科基金(11CTJ004)、重庆市自然科学基金(cstc2011jjA00014)和广西自然科学基金(2010GXNSFB013051)资助.

本文2011年12月23日收到, 2012年8月16日收到修改稿.

其中$U(t)$为测量误差, 类似Li和Greene (2007), 假定$U(t)$与$Y(t)$以及$X(t)$相互独立, 均值为0, 协方差阵为Σ_u与t无关, 且Σ_u已知. 否则, 可以采用Carroll等(1995)提出的方法获得Σ_u相合的, 无偏的矩估计, 本文的结论仍成立. 如果$X(t)$中某个分量不含测量误差, 则U中相应的分量和Σ_u对应的部分令为0. 产生测量误差的原因有很多, 如仪器精度不够, 感兴趣变量无法直接测量等, 已有很多学者讨论了测量误差模型, 相关工作可以参考文献[7–12].

关于含测量误差的纵向数据的研究, 在医学实验中常常出现. Lin和Carroll (2000)通过构造估计方程, 研究了非参数的估计, 考虑了血液中CD4含测量误差. Yi等(2012)讨论了含测量误差的纵向数据的参数模型, 采用广义矩方法估计回归系数, 在对食品数据进行分析时, 考虑了维生素A和维生素C的测量含有误差. Pan等(2009), Xiao等(2010)对这类数据进行了系统研究. 基于上述原因, 促使本文考虑含测量误差的纵向数据的变系数模型的估计问题.

对于变系数EV模型, Li和Greene (2007), You等(2006)采用局部纠偏的方法讨论了函数系数的估计; 崔恒建(2007)利用调整的加权最小二乘法估计函数系数. 上述文献主要在独立数据下讨论了函数系数的估计并研究了估计的渐近性质. 本文则在纵向数据下研究变系数EV模型中函数系数的核估计, 在一些正则条件下, 证明了估计的渐近性质. 进一步, 给出估计的渐近偏差和方差的估计, 从而基于正态逼近方法构造函数系数的逐点置信区间.

§2. 方法与主要结果

2.1 局部纠偏光滑核估计

假设$(X(t),Y(t))$与$(X_i(t),Y_i(t))$同分布. 对给定的$t\in R$, 模型(1.1)的等价形式为

$$Y(t)=X^T(t)\beta(t)+\varepsilon(t), \tag{2.1}$$

其中$\varepsilon(t)$是均值为0的随机过程, 其方差为$\sigma^2(t)$且协方差为$\rho_\varepsilon(t_1,t_2)$, $\varepsilon(\cdot)$与$X(\cdot)$相互独立. 假定给定$t\in R$的条件期望$\mathsf{E}[X(t)X^T(t)]$和$\mathsf{E}[X(t)Y(t)]$存在, 并且$\mathsf{E}[X(t)X^T(t)]$可逆. 那么, 对任何给定的$t\in R$, Wu等(1998)通过局部最小二乘准则给出了$\beta(t)$的局部核估计, 即关于$\beta(t)$极小化

$$L_N(\beta(t))=\sum_{i=1}^{n}\sum_{j=1}^{n_i}[Y_i(t_{ij})-X_i^T(t_{ij})\beta(t)]^2K_h(t-t_{ij}),$$

其中h是带宽, $K_h(\cdot)=K(\cdot/h)$且K是核函数, 则

$$\widetilde{\beta}(t)=\Big\{\frac{1}{Nh}\sum_{i=1}^{n}\sum_{j=1}^{n_i}X_i(t_{ij})X_i^T(t_{ij})K_h(t-t_{ij})\Big\}^{-1}\frac{1}{Nh}\sum_{i=1}^{n}\sum_{j=1}^{n_i}X_i(t_{ij})Y_i(t_{ij})K_h(t-t_{ij}).$$

但是, $\widetilde{\beta}(t)$中的$X_i(t_{ij})$不能直接精确观测, $\widetilde{\beta}(t)$将不能直接应用于模型(2.1)中函数系数$\beta(t)$的估计. 下面类似Liang等(1999)提出的偏差校正方法, 给出一个关于$\beta(t)$的局部纠偏光滑核估计. 令

$$\widehat{V}(t)=\frac{1}{Nh}\sum_{i=1}^{n}\sum_{j=1}^{n_i}[W_i(t_{ij})W_i^T(t_{ij})-\Sigma_u]K_h(t-t_{ij}).$$

假定矩阵$\widehat{V}(t)$可逆, 则$\beta(t)$的局部纠偏光滑核估计为

$$\widehat{\beta}(t)=\widehat{V}^{-1}(t)(Nh)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}W_i(t_{ij})Y_i(t_{ij})K_h(t-t_{ij}). \tag{2.2}$$

注记 1 采用$\widehat{\beta}(t)$的主要优点是它的数学表达式简单, 在实际应用中容易实现, 且具有优良的渐近性质. 但$\widehat{\beta}(t)$中仅含一个带宽, 当$\beta_1(t),\ldots,\beta_k(t)$是不同光滑族时, 它不能对$\beta(t)$的所有分量提供适当的光滑. 因此, 有必要进一步研究其它最小二乘估计方法.

现考虑$\widehat{\beta}(t)$的渐近分布. 假定设计点列$\{t_{ij},1\le i\le n,1\le j\le n_i\}$ i.i.d., 且具有公共的密度f. 记$S(f)$为f的支撑. 设t_0是$S(f)$的内点, 并记

$$\sigma^2(t_0)=\mathsf{E}[\varepsilon^2(t_0)],\qquad \rho_\varepsilon(t_0)=\lim_{\delta\to0}\mathsf{E}[\varepsilon(t_0+\delta)\varepsilon(t_0)],$$

且

$$\gamma_{lr}(t_0)=\mathsf{E}[X_{il}(t_{ij})X_{ir}(t_{ij})|t_{ij}=t_0].$$

为了得到本文结果, 需给出一些正则条件:

(1) 对某一个常数$h_0>0$, $h=h_0N^{-1/5}$.

(2) 对某一$0\le\lambda<\infty$, $\lim\limits_{n\to\infty}N^{-6/5}\sum\limits_{i=1}^{n}n_i^2=\lambda$.

(3) 核函数$K(\cdot)$是具有紧支撑的对称概率密度函数且满足$\int u^4K(u)\mathrm{d}u<\infty$.

(4) 存在常数$\delta\in(2/5,2]$, 使得$\sup\limits_t\mathsf{E}[|\varepsilon_i(t_{ij})|^{2+\delta}|t_{ij}=t]<\infty$, $\sup\limits_t\mathsf{E}[X_{il}^4(t_{ij})|t_{ij}=t]<\infty$, 且$\sup\limits_t\mathsf{E}[U_{il}^4(t_{ij})|t_{ij}=t]<\infty$, $i=1,\ldots,n$, $j=1,\ldots,n_i$, $l=1,\ldots,k$.

(5) 对所有$l,r=1,\ldots,k$, $\gamma_{lr}(t),\beta_r(t)$和$f(t)$在$t_0$点具有连续的二阶导数.

(6) $\sigma^2(t)$和$\rho_\varepsilon(t)$在t_0连续.

(7) $\Gamma(t_0)=(\gamma_{lr}(t_0))$是$k\times k$正定矩阵.

定理 2.1 假设条件(1)-(7)成立, 则

$$\sqrt{Nh}[\widehat{\beta}(t_0)-\beta(t_0)]-B(t_0)\xrightarrow{\mathcal{L}}N(0,\Sigma(t_0)),$$

其中$\xrightarrow{\mathcal{L}}$表示依分布收敛,

$$B(t_0)=(f(t_0))^{-1}\Gamma^{-1}(t_0)b(t_0),\qquad \Sigma(t_0)=(f(t_0))^{-2}\Gamma^{-1}(t_0)\Sigma^*(t_0)\Gamma^{-1}(t_0),$$

$\Gamma(t_0)$在条件(7)中已定义, $b(t_0)=(b_1(t_0),\ldots,b_k(t_0))^T$, $b_l(t_0)$和$\Sigma^*(t_0)$为

$$\begin{aligned}b_l(t_0) &= h_0^{5/2}\sum_{r=0}^{k}[\beta_r'(t_0)\gamma_{lr}'(t_0)f(t_0)+\beta_r'(t_0)\gamma_{lr}'(t_0)f'(t_0)\\ &\quad+(1/2)\beta_r''(t_0)\gamma_{lr}'(t_0)f(t_0)]\int u^2K(u)\mathrm{d}u,\\ \Sigma^*(t_0) &= \{\sigma^2(t_0)(\Gamma(t_0)+\Sigma_u)+\mathsf{E}[\eta(t_0)\beta(t_0)\beta^T(t_0)\eta_1^T(t_0)]\}f(t_0)\int K^2(u)\mathrm{d}u\\ &\quad+\{\rho_\varepsilon(t_0)(\Gamma(t_0)+\Sigma_u)+\mathsf{E}[\eta(t_0)\beta(t_0)\beta^T(t_0)\eta^T(t_0)]\}\lambda h_0f^2(t_0),\end{aligned}$$

其中$\eta(t_0)=\Sigma_u-X(t_0)U^T(t_0)-U(t_0)U^T(t_0)$, h_0和λ在条件(1)和(2)中已定义.

注记 2 在定理2.1中, 如果将条件(1)改为$h=o(N^{-1/5})$, 但条件(2)-(7)仍满足, 则渐近偏差项消失且有

$$\sqrt{Nh}[\widehat{\beta}(t_0)-\beta(t_0)]\xrightarrow{\mathcal{L}}N(0,\Sigma(t_0)).$$

2.2 渐近置信区间

在2.1节中定理2.1给出了函数系数估计$\widehat{\beta}(t_0)$的渐近性质, 基于定理2.1可以构造$\beta(t_0)$的逐点置信区间. 但是, 需要估计$\widehat{\beta}(t_0)$的渐近偏差和渐近协方差. 显然, $\widehat{\beta}(t_0)$的渐近偏差和渐近方差依赖$f(t_0),\Gamma(t_0),b(t_0)$和$\Sigma^*(t_0)$. 密度函数$f(t_0)$和$\Gamma(t_0)$可以采用核光滑方法估计. 假定$K(\cdot)$是满足条件(3)的核函数, 带宽$h$满足$h\to0$, $nh\to\infty$, 带宽h可以与$\widehat{\beta}(t_0)$中的带宽不同, 则$f(t_0)$和$\gamma_{lr}(t_0)$的估计定义为

$$\widehat{f}(t_0)=\frac{1}{Nh}\sum_{i=1}^{n}\sum_{j=1}^{n_i}K_h(t_0-t_{ij}),$$
$$\widehat{\gamma}_{lr}(t_0)=\frac{1}{Nh\widehat{f}(t_0)}\sum_{i=1}^{n}\sum_{j=1}^{n_i}[W_{il}(t_{ij})W_{ir}(t_{ij})-\Sigma_{u,lr}]K_h(t_0-t_{ij}),$$

其中$\Sigma_{u,lr}$是矩阵Σ_u的(l,r)元素. $\widehat{\Gamma}(t_0)$是$k\times k$矩阵, 其(l,r)元素为$\widehat{\gamma}_{lr}(t_0)$. 类似文献[4]中(3.1)式的思想, $b_l(t_0)$和$\Sigma^*(t_0)$的相合估计定义为

$$\widehat{b}_l(t_0)=\frac{1}{\sqrt{Nh}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{r=1}^{k}[W_{il}(t_{ij})W_{ir}(t_{ij})-\Sigma_{u,lr}][\widehat{\beta}_r(t_{ij})-\widehat{\beta}_r(t_0)]K_h(t_0-t_{ij}),$$
$$\widehat{\Sigma}^*(t_0)=\frac{1}{Nh}\sum_{i=1}^{n}\widehat{Z}_i(\widehat{\beta}(t_0))\widehat{Z}_i^T(\widehat{\beta}(t_0)),$$

其中

$$\widehat{Z}_i(\widehat{\beta}(t_0))=\sum_{j=1}^{n_i}\big\{W_i(t_{ij})[Y_i(t_{ij})-W_i^T(t_{ij})\widehat{\beta}(t_0)]+\Sigma_u\widehat{\beta}(t_0)\big\}K_h(t_0-t_{ij}),\qquad i=1,\ldots,n.$$

因此, $\widehat{\beta}(t_0)$的渐近偏差和渐近协方差的估计分别为

$$\widehat{B}(t_0)=(\widehat{f}(t_0))^{-1}\widehat{\Gamma}^{-1}(t_0)\widehat{b}(t_0),\qquad\widehat{\Sigma}(t_0)=(\widehat{f}(t_0))^{-2}\widehat{\Gamma}^{-1}(t_0)\widehat{\Sigma}^*(t_0)\widehat{\Gamma}^{-1}(t_0).$$

基于定理2.1, $\beta_r(t_0)$的渐近$1-\alpha$逐点置信区间为

$$\widehat{\beta}_r(t_0)-(Nh)^{-1/2}\widehat{B}_r(t_0)\pm z_{\alpha/2}(Nh)^{-1/2}\widehat{\sigma}_r(t_0),$$

其中$\widehat{B}_r(t_0)$是$\widehat{B}(t_0)$的第r个分量, $\widehat{\sigma}_r^2(t_0)$是矩阵$\widehat{\Sigma}(t_0)$的$(r,r)$元素, 且$z_{\alpha/2}$为标准正态分布的$1-\alpha/2$分位数.

§3. 模拟研究

本节通过数据模拟来研究本文所提出方法的有限样本性质. 为实施模拟研究, 从如下模型产生数据

$$Y(t) = \sin(\pi t/6)X(t) + \varepsilon(t),$$

其中$X(t)$服从区间$[t/12, 2+t/12]$的均匀分布. 假定$W(t) = X(t)+U(t)$, 且$U(t) \sim N(0, \sigma_u^2)$. 在模拟过程中, σ_u^2分别取$0.2^2, 0.4^2$和0.6^2以代表数据不同的污染水平. 为了使得每个个体重复观测的次数n_i不同, 我们产生60个不同的时间点s_{i1}, 且$s_{i1} \sim U(0,1)$, 使得$s_{il} = s_{i1}+(l-1)$, $l = 1, 2, \ldots, 12$. 对时间点列s_{il}随机缺失30%, 剩下的时间点列记为t_{ij}, $i = 1, \ldots, 60$, $j = 1, \ldots, n_i$. $Y(t)$由模型产生, 其中$\varepsilon(t)$为零均值的Gaussian过程, 协方差函数为

$$\mathrm{Cov}\left(\varepsilon_{i_1}(t_{i_1j_1}), \varepsilon_{i_2}(t_{i_2j_2})\right) = \begin{cases} 0.2\exp(-|t_{i_1j_1} - t_{i_2j_2}|), & \text{if } i_1 = i_2; \\ 0, & \text{if } i_1 \neq i_2. \end{cases}$$

我们取核函数为$K(u) = 0.75(1-u^2)_+$, 并且用交叉核实法选取带宽h_{CV}使其满足

$$\mathrm{CV}(h) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\{[Y_i(t_{ij}) - W_i^T(t_{ij})\widehat{\beta}_{[i]}(t_{ij};h)]^2 - \widehat{\beta}_{[i]}(t_{ij};h)\Sigma_u\widehat{\beta}_{[i]}^T(t_{ij};h)\}$$

达到最小, 其中$\widehat{\beta}_{[i]}(t_{ij};h)$为去掉第$i$个个体后$\beta(t)$的估计. 模拟研究比较了纠偏的光滑核估计(CKE)和忽略测量误差的估计(NCKE). 重复500次的模拟结果见图1和图2. 由图1和2, 可以得到如下结论:

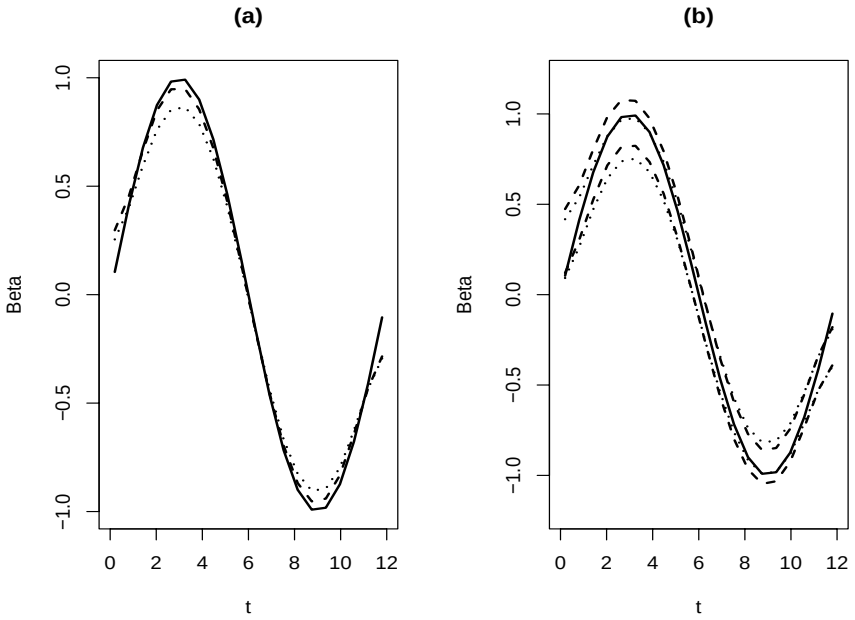

图1 基于CKE(虚线)和NCKE(点线)方法, $\beta(t)$的估计(图a)和置信水平为95%逐点置信区间(图b), $\sigma_u^2 = 0.4^2$

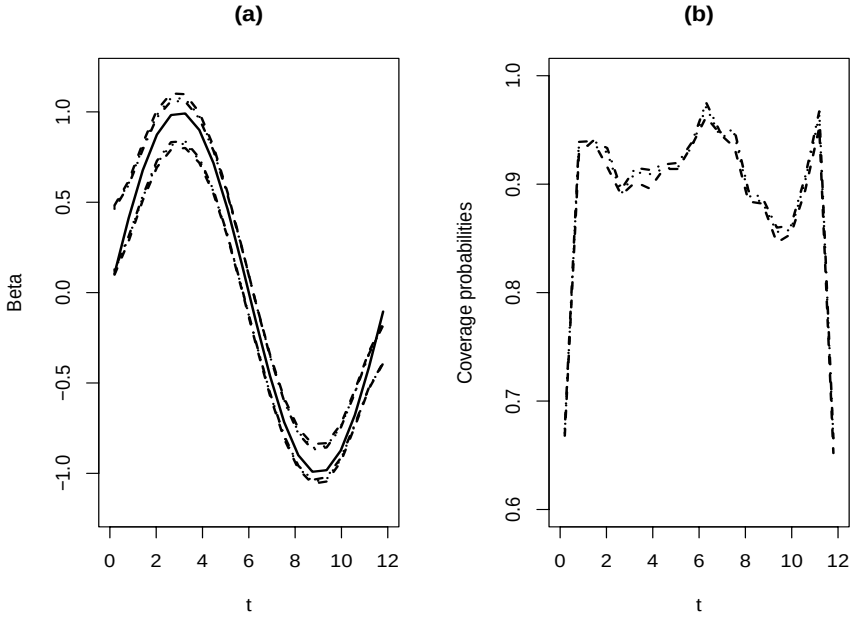

图2 $\beta(t)$置信水平为95%逐点置信区间(图a)及相应的覆盖概率(图b), $\sigma_u^2 = 0.2^2$(点虚线), 0.4^2(点线)和0.6^2(虚线)

(1) 图1表明基于NKE方法所得到的估计和逐点置信区间是有偏的. 本文提出的纠偏光滑核估计除了边界点外, 拟合效果较好.

(2) 从图2可以看出, 随着测量误差水平的增加, $\beta(t)$的逐点置信区间增加, 且除边界点外, 相应的覆盖概率都接近95%. 这表明测量误差水平影响估计.

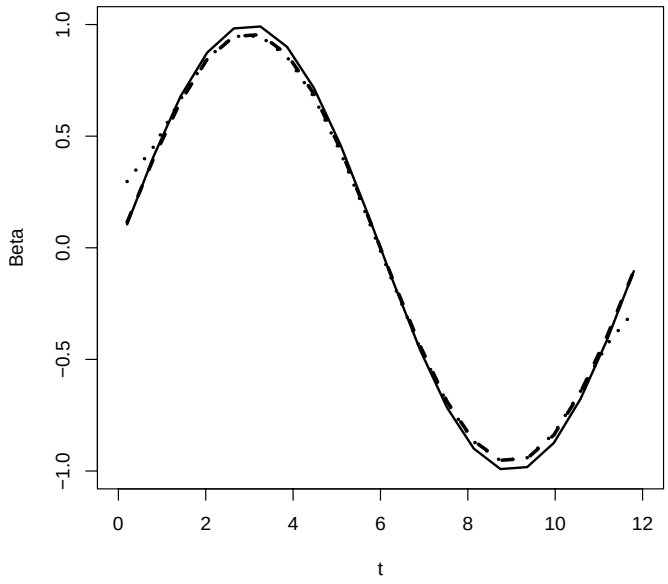

图3 基于CLPE(虚线)和CKE(点线)方法, 实线为真实曲线, $\sigma_u^2 = 0.4^2$

为了比较本文提出的纠偏的光滑核估计与纠偏的局部多项式估计(CLPE)方法, 模拟研究了这两种方法的有限样本性质. 文中仅给出了$\sigma_u^2 = 0.4^2$的结果, $\sigma_u^2 = 0.2^2$和$\sigma_u^2 = 0.6^2$有类似的结果. 由图3可以看出: 与纠偏的局部多项式方法相比, 本文提出的方法自变量在边界部分对回归函数的拟合较之于内点处存在较大的偏差, 该方法存在边界效应; 除边界点外, 两种方法拟合效果差不多.

§4. 实例分析

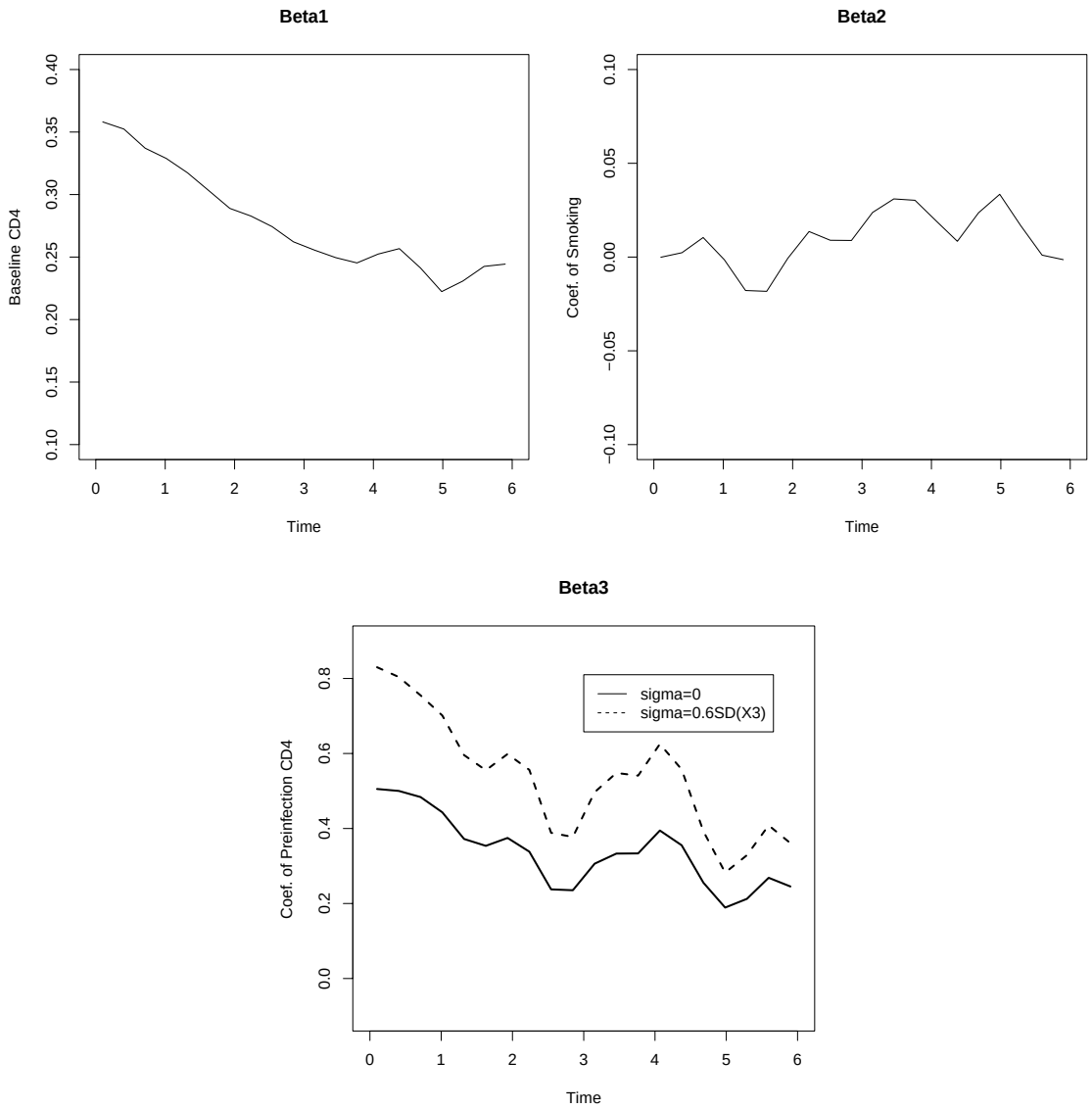

图4 AIDS数据分析. $\beta_1(t)$, $\beta_2(t)$以及在$\sigma_u^2 = 0$ (实线)和$\sigma_u^2 = 0.6\mathrm{SD}(X_3)$ (虚线)时$\beta_3(t)$的估计

本节分析一组来自对艾滋病(AIDS)群体研究的数据. 该数据可以从R程序包"timereg"获得. 该组数据采集了1984年到1991年期间感染HIV的283名男性同性恋者. Wu等(1998), Xue和Zhu (2007)采用变系数模型分析了该组数据. 数据包含患者感染HIV病毒时的年龄, 抽烟的状况, 感染前血液中的CD4细胞的百分比(Pre-CD4), 观察时间以及随着时间变化的感染HIV病毒后血液中CD4细胞的百分比. 用Y_{ij}表示CD4细胞的百分比, $X_{i1}\equiv 1$, X_{i2}表示是否吸烟, X_{i3}表示感染病毒前的CD4细胞百分比. 对该组数据建立如下模型

$$Y_{ij}=X_{i1}\beta_1(t_{ij})+X_{i2}\beta_2(t_{ij})+X_{i3}\beta_3(t_{ij})+\varepsilon_{ij}.$$

一些学者研究表明年龄对CD4细胞的百分比没有显著影响, 因此在建模中我们没有考虑年龄. Lin和Carroll (2000)考虑了CD4细胞的测量含有误差. 本节中我们假定Pre-CD4细胞的百分比含有测量误差

$$W_{i3}=X_{i3}+U_{i3}.$$

由于该数据缺少对含测量误差协变量的重复观察, 类似Lin和Carroll (2000), 我们取$\sigma_u^2=0$ (忽略测量误差)和$\sigma_u^2=0.6\mathrm{SD}(X_3)$. 由于在$\sigma_u^2=0$和$\sigma_u^2=0.6\mathrm{SD}(X_3)$时, $\beta_1(t)$和$\beta_2(t)$的估计类似, 我们仅给出了$\sigma_u^2=0$时$\beta_1(t)$和$\beta_2(t)$的估计. 从图4可以看出, 和忽略测量误差相比, 应用于$\sigma_u^2=0.6\mathrm{SD}(X_3)$的$\beta_3(t)$的估计在数值上有显著增大, 这表明, 当考虑测量误差时, Pre-CD4和CD4有更强的正相关, 这与Lin和Carroll (2000)所得的结果类似. 从$\beta_2(t)$的估计曲线可以看出吸烟对CD4细胞的百分比没有显著影响, 从$\beta_1(t)$的估计曲线可以看出CD4细胞百分比在感染HIV初期下降非常迅速, 而后下降速度开始减慢. 这与Xue和Zhu (2007)分析的结果一致.

§5. 定理的证明

引理 5.1 假设定理2.1中的条件(1)-(7)成立, 那么有

$$\mathsf{E}\Big[\frac{1}{\sqrt{Nh}}\sum_{i=1}^{n}Z_i(\beta(t_0))\Big]=b(t_0)+o(1),$$

且

$$\mathsf{Cov}\Big[\frac{1}{\sqrt{Nh}}\sum_{i=1}^{n}Z_i(\beta(t_0))\Big]=\Sigma^*(t_0)+o(1),$$

其中$Z_i(\beta(t_0))=\sum\limits_{j=1}^{n_i}\big\{W_i(t_{ij})[Y_i(t_{ij})-W_i^T(t_{ij})\beta(t_0)]+\Sigma_u\beta(t_0)\big\}K_h(t_0-t_{ij})$.

证明: 令$\widetilde{\xi}(t_0)=(Nh)^{-1/2}\sum\limits_{i=1}^{n}Z_i(\beta(t_0))$. 通过简单的计算, $\widetilde{\xi}(t_0)$的第l个分量为

$$\widetilde{\xi}_l(t_0)=\frac{1}{\sqrt{Nh}}\sum_{i=1}^{n}Z_{il}(\beta(t_0)),$$

其中

$$\begin{aligned} Z_{il}(\beta(t_0)) &= \sum_{j=1}^{n_i}[\xi_{il}(t_0,t_{ij})K_h(t_0-t_{ij})], \\ \xi_{il}(t_0,t_{ij}) &= \sum_{r=1}^{k}\{[W_{il}(t_{ij})W_{ir}(t_{ij})-\Sigma_{u,lr}][\beta_r(t_{ij})-\beta_r(t_0)]\} \\ &\quad + W_{il}(t_{ij})\varepsilon_i(t_{ij})+\eta_{il}(t_{ij})\beta(t_{ij}), \\ \eta_i(t_{ij}) &= \Sigma_u - X_i(t_{ij})U_i^T(t_{ij})-U_i(t_{ij})U_i^T(t_{ij}), \end{aligned}$$

$\eta_{il}(t_{ij})$表示$\eta_i(t_{ij})$的第l行. 我们可得

$$\begin{aligned} \mathsf{E}[Z_{il}(\beta(t_0))] &= \sum_{j=1}^{n_i}\int \mathsf{E}[\xi_{il}(t_0,t_{ij})|t_{ij}=u]K_h(t_0-u)f(u)\mathrm{d}u \\ &= n_ih\sum_{r=1}^{k}\int[\beta_r(t_0-hu)-\beta_r(t_0)]\gamma_{lr}(t_0-hu)f(t_0-hu)K(u)\mathrm{d}u. \end{aligned}$$

对上式右边进行Taylor展开, 结合条件(1), (3)和(6)可得

$$\mathsf{E}\Big[\frac{1}{\sqrt{Nh}}\sum_{i=1}^{n}Z_i(\beta(t_0))\Big]=b(t_0)+o(1).$$

现考虑$\widetilde{\xi}(t_0)$的协方差. 由于

$$\mathsf{Cov}\,[\widetilde{\xi}_l(t_0),\widetilde{\xi}_r(t_0)]=\mathsf{E}[\widetilde{\xi}_l(t_0)\widetilde{\xi}_r(t_0)]-\mathsf{E}[\widetilde{\xi}_l(t_0)]\mathsf{E}[\widetilde{\xi}_r(t_0)],$$

因此, 需要计算

$$\begin{aligned} &\mathsf{E}\Big\{\Big[\frac{1}{\sqrt{Nh}}\sum_{i=1}^{n}Z_{il}(\beta(t_0))\Big]\Big[\frac{1}{\sqrt{Nh}}\sum_{i=1}^{n}Z_{ir}(\beta(t_0))\Big]\Big\} \\ =\ &(Nh)^{-1}\Big\{\sum_{i=1}^{n}\mathsf{E}[Z_{il}(\beta(t_0))Z_{ir}(\beta(t_0))]+\sum_{i_1\neq i_2}\mathsf{E}[Z_{i_1l}(\beta(t_0))Z_{i_2r}(\beta(t_0))]\Big\}. \qquad (5.1) \end{aligned}$$

考虑(5.1)式中右边的第一项, 通过简单计算可得

$$\begin{aligned} Z_{il}(\beta(t_0))Z_{ir}(\beta(t_0)) &= \sum_{j=1}^{n_i}\xi_{il}(t_0,t_{ij})\xi_{ir}(t_0,t_{ij})K_h^2(t_0-t_{ij}) \\ &\quad + \sum_{j_1\neq j_2}\xi_{il}(t_0,t_{ij_1})\xi_{ir}(t_0,t_{ij_2})K_h(t_0-t_{ij_1})K_h(t_0-t_{ij_2}). \qquad (5.2) \end{aligned}$$

当$n\to\infty$, 可得

$$\begin{aligned} &\mathsf{E}[\xi_{il}(t_0,t_{ij})\xi_{ir}(t_0,t_{ij})|t_{ij}=u] \\ =\ &\sum_{c=1}^{k}\big\{[\beta_c(u)-\beta_c(t_0)]^2\mathsf{E}[(W_{il}(t_{ij})W_{ic}(t_{ij})-\Sigma_{u,lc})(W_{ir}(t_{ij})W_{ic}(t_{ij})-\Sigma_{u,rc})|t_{ij}=u]\big\} \\ &+\sigma^2(u)\mathsf{E}[W_{il}(t_{ij})W_{ir}(t_{ij})|t_{ij}=u]+\mathsf{E}[\eta_{il}(t_{ij})\beta(t_{ij})\beta^T(t_{ij})\eta_{ir}^T(t_{ij})|t_{ij}=u] \\ &+\sum_{c_1\neq c_2}\big\{[\beta_{c_1}(u)-\beta_{c_1}(t_0)][\beta_{c_2}(u)-\beta_{c_2}(t_0)] \\ &\times\mathsf{E}[(W_{il}(t_{ij})W_{ic_1}(t_{ij})-\Sigma_{u,lc_1})(W_{ir}(t_{ij})W_{ic_2}(t_{ij})-\Sigma_{u,rc_2})|t_{ij}=u]\big\} \\ \to\ &\sigma^2(t_0)[\gamma_{lr}(t_0)+\Sigma_{u,lr}]+\mathsf{E}[\eta_{1l}(t_0)\beta(t_0)\beta^T(t_0)\eta_{1r}^T(t_0)], \qquad \text{if } u\to t_0. \end{aligned}$$

结合上式, 我们有

$$\begin{aligned}
&\mathsf{E}\Big[\sum_{j=1}^{n_i}\xi_{il}(t_0,t_{ij})\xi_{ir}(t_0,t_{ij})K_h^2(t_0-t_{ij})\Big]\\
=\ &\sum_{j=1}^{n_i}\int \mathsf{E}[\xi_{il}(t_0,t_{ij})\xi_{ir}(t_0,t_{ij})|t_{ij}=u]K_h^2(t_0-u)f(u)\mathrm{d}u\\
=\ &n_ih\big\{\sigma^2(t_0)[\gamma_{lr}(t_0)+\Sigma_{u,lr}]+\mathsf{E}[\eta_{1l}(t_0)\beta(t_0)\beta^T(t_0)\eta_{1r}^T(t_0)]\big\}f(t_0)\int K^2(u)\mathrm{d}u\\
&+o(n_ih).
\end{aligned}\tag{5.3}$$

类似地, 当$n\to\infty$, $u_1\to t_0$且$u_2\to t_0$, 可以证明

$$\begin{aligned}
&\mathsf{E}[\xi_{il}(t_0,t_{ij_1})\xi_{ir}(t_0,t_{ij_2})|t_{ij_1}=u_1,t_{ij_2}=u_2]\\
\to\ &\rho_\varepsilon(t_0)[\gamma_{lr}(t_0)+\Sigma_{u,lr}]+\mathsf{E}[\eta_{1l}(t_0)\beta(t_0)\beta^T(t_0)\eta_{1r}^T(t_0)],
\end{aligned}$$

且(5.2)式中右边的第二项的期望为

$$\begin{aligned}
&\mathsf{E}\Big[\sum_{j_1\neq j_2}\xi_{il}(t_0,t_{ij_1})\xi_{ir}(t_0,t_{ij_2})K_h(t_0-t_{ij_1})K_h(t_0-t_{ij_2})\Big]\\
=\ &\sum_{j_1\neq j_2}\Big\{\int\int \mathsf{E}[\xi_{il}(t_0,t_{ij_1})\xi_{ir}(t_0,t_{ij_2})|t_{ij_1}=u_1,t_{ij_2}=u_2]\\
&\times K_h(t_0-u_1)K_h(t_0-u_2)f(u_1)f(u_2)\mathrm{d}u_1\mathrm{d}u_2\Big\}\\
=\ &h^2n_i(n_i-1)\big\{\rho_\varepsilon(t_0)[\gamma_{lr}(t_0)+\Sigma_{u,lr}]+\mathsf{E}[\eta_{1l}(t_0)\beta(t_0)\beta^T(t_0)\eta_{1r}^T(t_0)]\big\}f^2(t_0)\\
&+o(h^2n_i(n_i-1)).
\end{aligned}\tag{5.4}$$

结合(5.2), (5.3)和(5.4)式可得

$$\begin{aligned}
&(Nh)^{-1}\sum_{i=1}^{n}\mathsf{E}[Z_{il}(\beta(t_0))Z_{ir}(\beta(t_0))]\\
=\ &\big\{\sigma^2(t_0)[\gamma_{lr}(t_0)+\Sigma_{u,lr}]+\mathsf{E}[\eta_{1l}(t_0)\beta(t_0)\beta^T(t_0)\eta_{1r}^T(t_0)]\big\}f(t_0)\int K^2(u)\mathrm{d}u\\
&+hN^{-1}\Big(\sum_{i=1}^{n}n_i^2-N\Big)\big\{\rho_\varepsilon(t_0)[\gamma_{lr}(t_0)+\Sigma_{u,lr}]+\mathsf{E}[\eta_{1l}(t_0)\beta(t_0)\beta^T(t_0)\eta_{1r}^T(t_0)]\big\}\\
&\times f^2(t_0)+o\Big(hN^{-1}\Big(\sum_{i=1}^{n}n_i^2-N\Big)\Big).
\end{aligned}\tag{5.5}$$

由$h=N^{-1/5}h_0$和$\lim\limits_{n\to\infty}N^{-6/5}\sum\limits_{i=1}^{n}n_i^2=\lambda$, 很容易证明当$n\to\infty$, 有

$$hN^{-1}\Big(\sum_{i=1}^{n}n_i^2-N\Big)\to\lambda h_0.$$

类似文献[2]中(A.13)式的证明, 我们有

$$\begin{aligned}
&\Big|(Nh)^{-1}\sum_{i_1\neq i_2}\mathsf{E}[Z_{il}(\beta(t_0))Z_{ir}(\beta(t_0))]-\mathsf{E}\Big[(Nh)^{-1/2}\sum_{i=1}^{n}Z_{il}(\beta(t_0))\Big]\\
&\times\mathsf{E}\Big[(Nh)^{-1/2}\sum_{i=1}^{n}Z_{ir}(\beta(t_0))\Big]\Big|\to 0,\qquad \text{as}\ \ n\to\infty.
\end{aligned}\tag{5.6}$$

由条件(1), 结合(5.1), (5.5)和(5.6)式可得

$$\mathsf{Cov}\Big[\frac{1}{\sqrt{Nh}}\sum_{i=1}^{n}Z_i(\beta(t_0))\Big]=\Sigma^*(t_0)+o(1). \qquad \square$$

定理2.1的证明: 由(2.2)式, 经简单计算可得

$$\sqrt{Nh}[\widehat{\beta}(t_0)-\beta(t_0)]=\widehat{V}^{-1}\Big[\frac{1}{\sqrt{Nh}}\sum_{i=1}^{n}Z_i(\beta(t_0))\Big]. \tag{5.7}$$

结合条件(5)和引理5.1, 可得

$$\frac{1}{\sqrt{Nh}}\sum_{i=1}^{n}Z_i(\beta(t_0))\xrightarrow{\mathcal{L}}N(b(t_0),\Sigma^*(t_0)). \tag{5.8}$$

由大数定律可知

$$\widehat{V}\xrightarrow{\mathsf{P}}f(t_0)\Gamma(t_0). \tag{5.9}$$

因此, 由(5.7)-(5.9)以及Slutsky定理可得

$$\sqrt{Nh}[\widehat{\beta}(t_0)-\beta(t_0)]\xrightarrow{\mathcal{L}}N(B(t_0),\Sigma(t_0)).$$

由此可得定理2.1. $\square$

参 考 文 献

[1] Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.P., Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data, *Biometrika*, **85**(1998), 809–822.

[2] Wu, C.O., Chiang, C.T. and Hoover, D.R., Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data, *Journal of the American Statistical Association*, **93**(1998), 1388–1402.

[3] Huang, J.Z., Wu, C.O. and Zhou, L., Varying-coefficient models and basis function approximations for the analysis of repeated measurements, *Biometrika*, **89**(2002), 111–128.

[4] Xue, L.G. and Zhu, L.X., Empirical likelihood for a varying coefficient model with longitudinal data, *Journal of the American Statistical Association*, **102**(2007), 642–654.

[5] Li, L. and Greene, T., Varying coefficients model with measurement error, *Biometrika*, **84**(2007), 1–8.

[6] Carroll, R.J., Ruppert, D. and Stefansk, L.A., *Measurement Error in Nonlinear Models*, New York: Chapman and Hall, 1995.

[7] Zhou, Y. and Liang, H., Statistical inference for semiparametric varying-coefficient partially linear models with error-prone linear covariates, *The Annals of Statistics*, **37**(2009), 427–458.

[8] You, J.H., Zhou, Y. and Chen, G.M., Corrected local polynomial estimation in varying coefficient models with measurement errors, *The Canadian Journal of Statistics*, **34**(2006), 391–410.

[9] 崔恒建, 变系数EV模型参数的调整加权最小二乘估计及其渐近性质, 系统科学与数学, **27(1)**(2007), 82–92.

[10] Liang, H., Härdle, W. and Carroll, R.J., Estimation in a semiparametric partially linear errors-in-variables model, *The Annals of Statistics*, **27**(1999), 1519–1535.

[11] Wang, Q.H. and Zhang, R.Q., Statistical estimation in varying coefficient error-in-covariables models with surrogate data and validation sampling, *Journal of Multivariate Analysis*, **100**(2009), 2389–2405.

[12] 刘强, 薛留根, 纵向数据下部分线性EV模型的渐近性质, 应用数学学报, **32(1)**(2009), 178–189.

[13] Lin, X. and Carroll, R.J., Nonparametric function estimation for clustered data when the predictor is measured without/with error, *Journal of the American Statistical Association*, **95**(2000), 520–534.

[14] Yi, Y., Ma, Y. and Carroll, R.J., A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error, *Biometrika*, **99**(2012), 151–165.

[15] Pan, W., Zeng, D. and Lin, X., Estimation in semiparametric transition measurement error models for longitudinal data, *Biometrics*, **65**(2009), 728–736.

[16] Xiao, Z., Shao, J. and Palta, M., GMM in linear regression for longitudinal data with multiple covariates measured with error, *Journal of Applied Statistics*, **37**(2010), 791–805.

Kernel Smoothing Estimation for Varying Coefficient EV Models with Longitudinal Data

Yang Yiping Li Jia

(*College of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing, 400067*)

Varying coefficient EV models with longitudinal data are considered. The local bias-corrected kernel estimators for the unknown coefficient functions are proposed. It is shown that the proposed estimators are asymptotically normal under some suitable conditions, and hence it can be used to construct the pointwise confidence regions of the coefficient functions. The finite-sample properties of the proposed procedures are studied through a simulation study.

Keywords: Varying coefficient model, kernel estimator, confidence region, asymptotic normality, longitudinal data.

AMS Subject Classification: 62G05, 62G20.