

增长曲线模型的非参数估计 *

高采文

甘华来

(山西大同大学数学与计算机科学学院, 大同, 037009) (华东师范大学金融与统计学院, 上海, 200241)

摘要

增长曲线在研究中通常假定为时间的多项式形式, 大多数研究者都是通过选取高阶多项式的方式来提高估计的精度. 但这种方法存在很多缺陷, 如模型易受异常点的影响, 多项式假设要求过高等. 本文首次将局部多项式这种非参数估计方法应用到增长曲线模型中, 提出了非参数增长曲线模型, 给出了它的局部多项式估计, 并讨论了估计的渐近性质和理论带宽的选择. 最后对参数估计和非参数估计进行了模拟比较, 从拟合图和平均均方误差箱形图得到的结论是非参数估计效果较好.

关键词: 增长曲线模型, 非参数估计, 局部多项式, 最优窗宽.

学科分类号: O212.7.

§1. 引言

增长曲线模型由Wishart于1938年在研究不同组间动植物的生长情况时首次引入, 是一种广义多元方差分析模型, 在现代医学、农业及生物等领域中有着广泛的应用.

一般的增长曲线模型(GCM)为

$$Y_{p \times n} = X_{p \times m} B_{m \times r} Z_{r \times n} + \epsilon_{p \times n}, \quad \epsilon_{p \times n} \sim N_{p \times n}(\mathbf{0}, I_n \otimes \Sigma), \quad (1.1)$$

其中 Y 是观测矩阵, X 和 Z 是设计矩阵, $\text{Rank}(X) = m < p$, $\text{Rank}(Z) = r < n$, B 是未知参数矩阵, ϵ 是随机误差矩阵, Σ 是 $p \times p$ 阶正定矩阵, \otimes 表示Kronecker乘积.

几十年来, 许多统计学家对此模型做了大量研究, 给出了该模型在许多不同条件下对未知参数的极大似然估计(MLE). 国内学者潘建新(1988)讨论了增长曲线模型的回归参数在“trace”意义下的广义最小二乘估计(GLSE). 我国学者张日权等(1988)研究了在实际应用中, 若 Σ 不清楚或 Σ^{-1} 不易计算时, 用最小二乘估计代替最佳线性无偏估计所蒙受的损失问题.

在实际应用中, 研究者们大多数是采用Rao(1965)所提出的协方差调整的方法即通过选取高阶多项式的方式来获得比较精确的对未知参数矩阵 B 的估计. 虽然这种方法得到了广泛应用和推广, 但仍存在很多缺陷: (1)要求多项式具有任意阶导数, 且要求各阶导数处处存在, 然而大多数实际问题很难满足这一要求; (2)如何合理的确定多项式的阶数比较困

*国家自然科学基金项目(11171112)和国家统计局重点科研项目(2011LZ051)资助.

本文2013年8月28日收到, 2013年10月22日收到修改稿.

难, 高的阶数将带来参数的增加和模型的不稳定, 而低的阶数带来的是模型误差的增加; (3)异常点的影响, 异常点会对多项式的形式有较大的影响, 导致模型的不稳定; (4)多项式的假设有时不太合理. 如俞启泰(2003)提出的油田开发指标的增长曲线:

$$N_p = N_{\max} \left\{ \sin \left[\frac{\pi}{2} (1 - a^{-t}) \right] \right\}^b,$$

其中 N_p 为累积产油量, N_{\max} 为最大可采储量. 油田累计产量、储采比等开采指标与时间之间满足一种三角函数关系.

鉴于以上缺陷, 本文将非参数回归方法引入到增长曲线模型中, 提出非参数增长曲线模型(nonparametric growth curve model)为

$$Y_{p \times n} = \eta_{p \times r} Z_{r \times n} + \epsilon_{p \times n}, \quad \epsilon_{p \times n} \sim N_{p \times n}(\mathbf{0}, I_n \otimes \Sigma), \quad (1.2)$$

其中 Y 是观测矩阵, Z 是秩为 r ($r < n$) 的设计矩阵, $n = \sum_{i=1}^r n_i$, $\eta(t) = [\eta_1(t), \eta_2(t), \dots, \eta_r(t)]$, $\eta_i(t)$ ($i = 1, 2, \dots, r$) 是第 i 组的光滑增长曲线.

我们的目的是求出增长曲线函数 $\eta(t)$ 以及它的导函数 $\eta^{(v)}(t)$ 的估计值. 第二节给出了非参数增长曲线模型的局部多项式估计. 第三节借鉴Fan和Gijbels(1996)的研究方法, 对非参数增长曲线模型的局部多项式估计的渐近性质进行了讨论. 在非参数回归中, 一个重要的问题是窗宽的选择. 第四节中讨论了非参数增长曲线模型估计的窗宽选择问题. 由于大量的真实数据比较难获得, 在第五节采用模拟的方法对参数估计和非参数估计进行了比较. 从模拟的结果可以看出本文所采用的非参数回归方法比传统的参数估计方法效果要好, 从而验证了新方法的优势所在.

§2. 非参数增长曲线模型的局部多项式估计

令 t_0 为任意给定的时间点, 假设 $\eta(t)$ 在 t_0 处具有 $(m+1)$ 阶连续导数, 其中 m 为非负整数. 由 Taylor 展开式, $\eta(t)$ 可以在 t_0 的局部近似展开为 m 阶多项式, 即

$$\begin{aligned} \eta(t) &\approx \eta(t_0) + \eta'(t_0)(t - t_0) + \frac{\eta''(t_0)}{2!}(t - t_0)^2 + \cdots + \frac{\eta^{(m)}(t_0)}{m!}(t - t_0)^m \\ &\equiv \sum_{j=0}^m \beta_j^T (t - t_0)^j, \end{aligned}$$

其中 $\beta_j^T = \eta^{(j)}(t_0)/j!$.

令 $\beta = [\beta_0, \beta_1, \dots, \beta_m]^T$, 由模型残差最小性原则:

$$\text{tr}(Q(\beta)) = \text{tr}\{(Y - X\beta Z)^T W(Y - X\beta Z)\}, \quad (2.1)$$

其中 $X = (x_{ij})_{p \times (m+1)}$, $x_{ij} = (t_i - t)^{j-1}$, $W = \text{diag}\{K_h(t_i - t)\}_{p \times p}$, $K_h(\cdot) = K(\cdot/h)/h$, $i = 1, 2, \dots, p$, $j = 1, 2, \dots, m+1$, $K(\cdot)$ 为核函数, h 为带宽.

$\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m]^T$ 是为极小化(2.1)得到的 β 的估计, 即 $\hat{\beta} = \arg \min\{\text{tr}(Q(\beta))\}$.

定理 2.1 在非参数增长曲线模型(2.1)中, $\eta(t)$ 及 $\eta^{(v)}(t)$ 的估计值分别为

$$\hat{\eta}(t) = e_1^T \hat{\beta}, \quad \hat{\eta}^{(v)}(t) = v! e_{v+1}^T \hat{\beta}, \quad v = 0, 1, \dots, m.$$

证明: 显然 $\text{tr}(Q(\beta))$ 可以分解为

$$\text{tr}(Q(\beta)) = \text{tr}(Y^T W Y) - 2\text{tr}(Y^T W X \beta Z) + \text{tr}(Z^T \beta^T X^T W X \beta Z).$$

又

$$\begin{aligned} \text{tr}(Y^T W X \beta Z) &= \text{vec}(\beta)^T \text{vec}(X^T W Y Z^T), \\ \text{tr}(Z^T \beta^T X^T W X \beta Z) &= \text{vec}(\beta)^T [Z Z^T \otimes X^T W X] \text{vec}(\beta), \end{aligned}$$

有

$$\partial \text{tr}(Y^T W X \beta Z) / \partial \beta = \text{vec}(X^T W Y Z^T),$$

以及

$$\partial \text{tr}(Z^T \beta^T X^T W X \beta Z) / \partial \beta = 2\text{vec}(X^T W X \beta Z Z^T).$$

又 $\hat{\beta}$ 是下面方程的解:

$$\partial Q(\beta) / \partial \beta = 0,$$

则易得 $\text{vec}(X^T W Y Z^T) = \text{vec}(X^T W X \beta Z Z^T)$, 即有

$$X^T W Y Z^T = X^T W X \beta Z Z^T.$$

因此

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y Z^T (Z Z^T)^{-1}. \quad (2.2)$$

进而易得

$$\hat{\eta}^{(v)}(t) = v! e_{v+1}^T \hat{\beta}, \quad v = 0, 1, \dots, m. \quad (2.3)$$

特别地, 令 $v = 0$, 我们有 $\hat{\eta}(t) = e_1^T \hat{\beta}$. \square

令 $D = \{t_i, i = 1, 2, \dots, p\}$ 为所有观测点组成的集合, 则易有

$$\begin{aligned} \mathbb{E}(\hat{\eta}^{(v)}(t)|D) &= v! e_{v+1}^T (X^T W X)^{-1} X^T W \eta \\ &= \eta^{(v)}(t) + v! e_{v+1}^T (X^T W X)^{-1} X^T W R, \end{aligned}$$

以及

$$\text{Cov}(\hat{\eta}^{(v)}(t)|D) = v!^2 (Z Z^T)^{-1} \otimes [e_{v+1}^T (X^T W X)^{-1} X^T W \Sigma W X (X^T W X)^{-1} e_{v+1}],$$

其中 $R = \eta - X\beta$.

§3. 非参数增长曲线模型估计的渐近性质

在第二节中虽然得到了估计 $\hat{\eta}^{(v)}(t)$ 的精确期望和方差, 但是实际应用中不能直接使用它们, 因为其中含有未知元素: 残差 $v!e_{v+1}^T(X^TWX)^{-1}X^TR$ 和协方差 Σ . 在本节我们讨论估计 $\hat{\eta}^{(v)}(t)$ 的一些渐近性质.

下面描述几个假设:

A1: 观测时间点 t_i ($i = 1, 2, \dots, p$) 独立同分布, 密度函数为 $f(\cdot)$, 组内协方差

$$\Sigma = \text{diag}(\sigma^2(t_1), \sigma^2(t_2), \dots, \sigma^2(t_p));$$

A2: 给定观测时间点 t 在 $f(\cdot)$ 的紧支撑里, 且 $f(t) > 0$, $f'(t)$ 存在;

A3: $f(\cdot)$, $\eta^{(m+1)}(\cdot)$, $f'(\cdot)$, $\eta^{(m+2)}(\cdot)$ 以及 $\sigma^2(\cdot)$ 在 t 的某邻域中都是连续的;

A4: $h \rightarrow 0$ 且 $ph \rightarrow +\infty$.

我们记

$$\begin{aligned} \mu_i &= \int u^i K(u) du, & v_i &= \int u^i K^2(u) du; \\ S &= (\mu_{j+l})_{0 \leq j, l \leq m}, & \tilde{S} &= (\mu_{j+l+1})_{0 \leq j, l \leq m}, & S^* &= (v_{j+l})_{0 \leq j, l \leq m}; \\ S_{p,j} &= \sum_{i=1}^p K_h(t_i - t)(t_i - t)^j, & S_{p,j}^* &= \sum_{i=1}^p (t_i - t)^j K_h^2(t_i - t) \sigma^2(t_i); \\ c_m &= (\mu_{m+1}, \dots, \mu_{2m+1})^T, & \tilde{c}_m &= (\mu_{m+1}, \dots, \mu_{2m+1}); \\ c_p &= (S_{p,m+1}, \dots, S_{p,2m+1})^T, & \tilde{c}_p &= (S_{p,m+1}, \dots, S_{p,2m+1}); \\ \sigma^2 &= W\Sigma W = \text{diag}\{\sigma^2(t_1)K_h^2(t_1 - t), \dots, \sigma^2(t_p)K_h^2(t_p - t)\}; \\ S_p &= X^TWX, & S_p^* &= X^T\sigma^2 X; \\ \beta_{m+1} &= \eta^{(m+1)}(t)/(m+1)! = [\eta_1^{(m+1)}(t)/(m+1)!, \dots, \eta_r^{(m+1)}(t)/(m+1)!]; \\ H &= \text{diag}(1, h, \dots, h^m). \end{aligned}$$

定理 3.1 在假设A1-A4成立的条件下, 有

$$\text{Cov}(\hat{\eta}^{(v)}(t)|D) = (ZZ^T)^{-1} \otimes \left[v!^2 e_{v+1}^T S^{-1} S^* S^{-1} e_{v+1} \frac{\sigma^2(t)}{f(t)ph^{2v+1}} + op\left(\frac{1}{ph^{2v+1}}\right) \right]. \quad (3.1)$$

当 $m-v$ 为奇数时, 有

$$\text{Bias}(\hat{\eta}^{(v)}(t)|D) = v!e_{v+1}^T S^{-1} c_m \beta_{m+1} h^{m+1-v} + op(h^{m+1-v}); \quad (3.2)$$

当 $m-v$ 为偶数时, 有

$$\text{Bias}(\hat{\eta}^{(v)}(t)|D) = v!e_{v+1}^T S^{-1} \tilde{c}_m \left\{ \beta_{m+2} + \frac{f'(t)}{f(t)} \beta_{m+1} \right\} h^{m+2-v} + op(h^{m+2-v}). \quad (3.3)$$

证明: 易知

$$(X^T W X)^{-1} X^T \sigma^2 X (X^T W X)^{-1} = S_p^{-1} S_p^* S_p^{-1}.$$

对于任意随机变量 ξ , 若其一二阶矩有限, 则有

$$\xi = E(\xi) + Op(\text{Var}^{1/2}(\xi)).$$

根据上式, 有

$$\begin{aligned} S_{p,j} &= E(S_{p,j}) + Op\{\text{Var}^{1/2}(S_{p,j})\} \\ &= ph^j \int u^j K(u) f(t+hu) du + Op\{\sqrt{pE[(t_1-t)^2 K_h^2(t_1-t)]}\} \\ &= ph^j \{f(t)\mu_j + o(1) + Op(1/\sqrt{ph})\} \\ &= ph^j f(t)\mu_j \{1 + op(1)\}, \quad h \rightarrow 0, ph \rightarrow +\infty. \end{aligned}$$

易得

$$S_p = pf(t)HSH\{1 + op(1)\}.$$

同理, 有

$$\begin{aligned} S_{p,j}^* &= ph^{j-1} f(t) \sigma^2(t) v_j \{1 + op(1)\}, \\ S_p^* &= ph^{-1} f(t) \sigma^2(t) HS^* H \{1 + op(1)\}. \end{aligned}$$

由以上三式, 我们有

$$S_p^{-1} S_p^* S_p^{-1} = \frac{\sigma^2(t)}{phf(t)} H^{-1} S^{-1} S^* S^{-1} H^{-1} \{1 + op(1)\}.$$

代入 $\text{Cov}(\hat{\beta}|D)$ 和 $\text{Cov}(\hat{\eta}^{(v)}(t)|D)$, 有

$$\text{Cov}(\hat{\beta}|D) = (ZZ^T)^{-1} \otimes \left[\frac{\sigma^2(t)}{phf(t)} H^{-1} S^{-1} S^* S^{-1} H^{-1} \{1 + op(1)\} \right],$$

及

$$\begin{aligned} \text{Cov}(\hat{\eta}^{(v)}(t)|D) &= v!^2 (ZZ^T)^{-1} \otimes \left[e_{v+1}^T \frac{\sigma^2(t)}{phf(t)} H^{-1} S^{-1} S^* S^{-1} H^{-1} \{1 + op(1)\} e_{v+1} \right] \\ &= (ZZ^T)^{-1} \otimes \left[v!^2 e_{v+1}^T S^{-1} S^* S^{-1} e_{v+1} \frac{\sigma^2(t)}{f(t)ph^{2v+1}} + op\left(\frac{1}{ph^{2v+1}}\right) \right]. \end{aligned}$$

对于偏差, 分 $m - v$ 为奇数还是偶数两种情况讨论.

(1) 当 $m - v$ 为奇数时:

由Taylor展开式, 易知 $\hat{\beta}$ 的条件偏差 $S_p^{-1}X^TWR$ 可以表示为如下形式:

$$\begin{aligned} S_p^{-1}X^TWR &= S_p^{-1}X^TW\{(t_i-t)^{m+1}\}_{1 \leq i \leq p}\beta_{m+1} + [op((t_i-t)^{m+1})]_{1 \leq i \leq p}\mathbf{1}\} \\ &= S_p^{-1}\{c_p\beta_{m+1} + op(ph^{m+1})\} \\ &= H^{-1}S^{-1}c_m\beta_{m+1}h^{m+1}\{1 + op(1)\}. \end{aligned}$$

所以

$$\text{Bias}(\hat{\beta}|D) = H^{-1}S^{-1}c_m\beta_{m+1}h^{m+1}\{1 + op(1)\}.$$

进而, 有

$$\text{Bias}(\hat{\eta}^{(v)}(t)|D) = e_{v+1}^TS^{-1}c_m\beta_{m+1}h^{m+1-v} + op(h^{m+1-v}).$$

(2) 当 $m - v$ 为偶数时:

$$S_{p,j} = ph^j\{f(t)\mu_j + hf'(t)\mu_{j+1} + Op(a_n)\},$$

其中 $a_n = h^2 + 1/\sqrt{ph}$. 则有

$$S_p = ph\{f(t)S + hf'(t)\tilde{S} + Op(a_n)\}H.$$

使用高一阶的Taylor展开式, $\hat{\beta}$ 的条件偏差 $S_p^{-1}X^TWR$ 可以表示为

$$\begin{aligned} S_p^{-1}X^TWR &= S_p^{-1}X^TW\{(t_i-t)^{m+1}\}_{1 \leq i \leq p}\beta_{m+1} + [(t_i-t)^{m+2}]_{1 \leq i \leq p}\beta_{m+2} \\ &\quad + [op((t_i-t)^{m+2})]_{1 \leq i \leq p}\mathbf{1}\} \\ &= S_p^{-1}\{c_p\beta_{m+1} + \tilde{c}_p\beta_{m+2} + op(ph^{m+2})\}. \end{aligned}$$

所以

$$\begin{aligned} \text{Bias}(\hat{\beta}|D) &= H^{-1}\{f(t)S + hf'(t)\tilde{S} + Op(a_n)\}h^{m+1} \\ &\quad \cdot \{f(t)c_m\beta_{m+1} + h\tilde{c}_m[f'(t)\beta_{m+1} + f(t)\beta_{m+2}] + Op(a_n)\} \\ &= h^{m+1}H^{-1}\{S^{-1}c_m\beta_{m+1} + hb(t) + Op(a_n)\}, \end{aligned}$$

其中

$$b(t) = S^{-1}\tilde{c}_m \frac{f'(t)\beta_{m+1} + f(t)\beta_{m+2}}{f(t)} - \frac{f'(t)}{f(t)}S^{-1}\tilde{S}S^{-1}c_m\beta_{m+1}.$$

又 $\hat{\eta}(t) = e_{v+1}^T\hat{\beta}$ 且 $S^{-1}c_m$ 和 $S^{-1}\tilde{S}S^{-1}c_m$ 的第 $v + 1$ 个元素为0, 因此易得

$$\begin{aligned} \text{Bias}(\hat{\eta}^{(v)}(t)|D) &= e_{v+1}^Th^{m+2}H^{-1}S^{-1}\tilde{c}_m\left\{\beta_{m+2} + \frac{f'(t)}{f(t)}\beta_{m+1}\right\} + op(h^{m+2-v}) \\ &= e_{v+1}^TS^{-1}\tilde{c}_m\{\beta_{m+2} + \frac{f'(t)}{f(t)}\beta_{m+1}\}h^{m+2-v} + op(h^{m+2-v}). \quad \square \end{aligned}$$

《应用概率统计》

定理 3.2 在假设A1-A4成立的条件下, 当 $m - v$ 为奇数, $h \rightarrow 0$, $ph \rightarrow +\infty$ 时, 有

$$\hat{\eta}^{(v)}(t) \rightarrow N_{1 \times r}(\eta^{(v)} + b_v(t), (ZZ^T)^{-1} \otimes \sigma_v^2(t)), \quad (3.4)$$

其中

$$\begin{aligned} b_v(t) &= \left\{ \int t^{m+1} K_v^*(t) dt \right\} \frac{v! \eta^{m+1}(t)}{(m+1)!} h^{m+1-v} + op(h^{m+1-v}), \\ \sigma_v^2(t) &= \int K_v^{*2}(t) dt \frac{v!^2 \sigma^2(t)}{f(t) ph^{2v+1}} + op\left(\frac{1}{ph^{2v+1}}\right), \\ K_v^*(t) &= e_{v+1}^T S^{-1}(1, t, \dots, t^m)^T K(t). \end{aligned}$$

证明: 由定理3.1知, 当 $m - v$ 为奇数时,

$$\begin{aligned} \text{Bias}(\hat{\eta}^{(v)}(t)|D) &= v! e_{v+1}^T S^{-1} c_m \beta_{m+1} h^{m+1-v} + op(h^{m+1-v}), \\ \text{Cov}(\hat{\eta}^{(v)}(t)|D) &= (ZZ^T)^{-1} \otimes \left[v!^2 e_{v+1}^T S^{-1} S^* S^{-1} e_{v+1} \frac{\sigma^2(t)}{f(t) ph^{2v+1}} + op\left(\frac{1}{ph^{2v+1}}\right) \right]. \end{aligned}$$

记

$$K_v^*(t) = e_{v+1}^T S^{-1}(1, t, \dots, t^m)^T K(t),$$

则易得

$$\begin{aligned} \text{Bias}(\hat{\eta}^{(v)}(t)|D) &= \left\{ \int t^{m+1} K_v^*(t) dt \right\} \frac{v! \eta^{m+1}(t)}{(m+1)!} h^{m+1-v} + op(h^{m+1-v}), \\ \text{Cov}(\hat{\eta}^{(v)}(t)|D) &= (ZZ^T)^{-1} \otimes \int K_v^{*2}(t) dt \frac{v!^2 \sigma^2(t)}{f(t) ph^{2v+1}} + op\left(\frac{1}{ph^{2v+1}}\right). \end{aligned}$$

由(2.3)知非参数估计 $\hat{\eta}^{(v)}(t)$ 是观测矩阵 Y 的线性函数, 所以易得

$$\hat{\eta}^{(v)}(t) \rightarrow N_{1 \times r}(\eta^{(v)} + b_v(t), (ZZ^T)^{-1} \otimes \sigma_v^2(t)),$$

其中

$$\begin{aligned} b_v(t) &= \left\{ \int t^{m+1} K_v^*(t) dt \right\} \frac{v! \eta^{m+1}(t)}{(m+1)!} h^{m+1-v} + op(h^{m+1-v}), \\ \sigma_v^2(t) &= \int K_v^{*2}(t) dt \frac{v!^2 \sigma^2(t)}{f(t) ph^{2v+1}} + op\left(\frac{1}{ph^{2v+1}}\right). \quad \square \end{aligned}$$

§4. 窗宽的选择

在实际应用中, 窗宽 h 的选择很重要: 若窗宽选的过小, 会使最好的估计结果含有很多的噪音, 引起大的估计方差; 若窗宽选的过大, 则会过于光滑数据导致遗失原始数据中很多重要信息, 引起大的估计偏差. 对局部多项式估计, Fan和Gijbels(1992)讨论了关于理论常

数窗宽、局部最优窗宽以及全局最优窗宽的选择. 我们讨论当 $m - v$ 为奇数时局部最优窗宽的选择问题, 对于其它窗宽的选择可以参照Fan和Gijbels (1992)使用类似的方法获得.

根据Fan和Gijbels (1996), 在定理3.1中 $\hat{\eta}_v(t)$ 的渐近偏差和渐近协方差可以表示为如下形式:

$$\text{Bias}(\hat{\eta}^{(v)}(t)|D) = \left\{ \int x^{m+1} K_v^*(x) dx \right\} v! h^{m+1-v} \beta_{m+1} + o_p(h^{m+1-v}),$$

以及

$$\text{Cov}(\hat{\eta}^{(v)}(t)|D) = (ZZ^T)^{-1} \left[\int K_v^{*2}(x) dx \frac{v!^2 \sigma^2(t)}{f(t)p h^{2v+1}} + o_p\left(\frac{1}{ph^{2v+1}}\right) \right].$$

因此MSE可以近似的表示为

$$\begin{aligned} & \left\{ \int x^{m+1} K_v^*(x) dx \right\}^2 v!^2 h^{2(m+1-v)} \beta_{m+1}^T \beta_{m+1} + \left[\int K_v^{*2}(x) dx \frac{v!^2 \sigma^2(t)}{f(t)p h^{2v+1}} \right] (ZZ^T)^{-1} \\ &= \left\{ \int x^{m+1} K_v^*(x) dx \right\}^2 v!^2 \beta_{m+1}^T \beta_{m+1} h^{2(m+1-v)} + \left[\int K_v^{*2}(x) dx (ZZ^T)^{-1} \frac{v!^2 \sigma^2(t)}{f(t)p} \right] h^{2v+1}. \end{aligned}$$

在“trace”意义下极小化上面的渐近MSE, 可以得到局部最优窗宽为

$$h_{\text{opt}}(t) = \left(\frac{2v+1}{2(m+1-v)} \right)^{1/(2m+3)} (\text{tr}(a^{-1}b))^{1/(2m+3)},$$

其中

$$a = \left\{ \int x^{m+1} K_v^*(x) dx \right\}^2 v!^2 \beta_{m+1}^T \beta_{m+1}, \quad b = \int K_v^{*2}(x) dx (ZZ^T)^{-1} \frac{v!^2 \sigma^2(t)}{f(t)p}.$$

但这个局部最优窗宽依赖于一些未知量, 如设计密度函数 $f(\cdot)$ 、条件方差 $\sigma^2(\cdot)$ 以及 $\hat{\eta}^{(m+1)}(t)$, 因此不能直接使用. 在实际应用中, 窗宽的选择一般都采用交叉核实法(cross validation, 简称CV)或广义交叉核实法(generalized cross validation, 简称GCV).

本文中我们采用CV法获得最优局部窗宽, 定义CV得分为

$$\text{CV}(h) = \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p [y_{ij} - \hat{\eta}_i^{(-i)}(t_j)]^2, \quad (4.1)$$

其中 $\hat{\eta}_i^{(-i)}(t)$ 为去掉第*i*个数据点后, 由剩下的数据估计出的 $\eta_i(t)$ 的拟合值. 极小化上面定义的 $\text{CV}(h)$, 即可获得所需的局部最优窗宽 h_{opt} .

§5. 模拟与分析

这节我们采用模拟的方法对参数估计和本文提出的非参数估计进行比较. 参考Wu和Zhang (2006)中的模拟模型为

$$y_i(t) = \beta_{i0} + \beta_{i1} \cos(2\pi t) + \beta_{i2} \sin(2\pi t) + \epsilon_i(t), \quad i = 1, 2, \dots, r, \quad (5.1)$$

其中 r 是总组数, β_i 和 $\epsilon_i(t)$ 相互独立, 且 $\beta_i = [\beta_{i0}, \beta_{i1}, \beta_{i2}]^T \sim N((a_{i0}, a_{i1}, a_{i2}), \text{diag}(\sigma_0^2, \sigma_1^2, \sigma_2^2))$, $\epsilon_i(t) \sim N(0, \sigma_\epsilon^2 \cdot t)$, $E(\beta_i, \beta_j) = 0, i \neq j$, $E(\epsilon_i(t), \epsilon_j(t)) = 0, i \neq j, t = i/(m+1)$, $i = 1, 2, \dots, m$, 其中 m 为事先给定的一正整数.

易得模拟的均值函数和协方差函数, 以及噪音部分的方差函数分别为

$$\eta_i(t) = a_{i0} + a_{i1} \cos(2\pi t) + a_{i2} \sin(2\pi t); \quad (5.2)$$

$$\gamma(s, t) = \sigma_0^2 + \sigma_1^2 \cos(2\pi s) \cos(2\pi t) + \sigma_2^2 \sin(2\pi s) \sin(2\pi t), \quad s \neq t; \quad (5.3)$$

$$\gamma_\epsilon(t, t) = \sigma_\epsilon^2 \cdot t. \quad (5.4)$$

不难发现, 当 $\sigma_1^2 \neq \sigma_2^2$ 时, 协方差函数 $\gamma(s, t)$ 不是常值, 且当 $\sigma_\epsilon \neq 0$ 时, 噪音部分的方差 γ_ϵ 不是齐次的, 因此模型能代表一大类实际的情形. 易得 $y_i(s)$ 与 $y_i(t)$ 的相关系数.

特别地, 当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 时, 有

$$\text{Cor}\{y_i(s), y_i(t)\} = \frac{\sigma_0^2 + \sigma^2 \cos\{2\pi(s-t)\}}{[(\sigma_0^2 + \sigma^2 + \sigma_\epsilon^2 s)(\sigma_0^2 + \sigma^2 + \sigma_\epsilon^2 t)]^{1/2}}, \quad s \neq t. \quad (5.5)$$

在模拟中, 令 $(a_{10}, a_{11}, a_{12}) = (4, 2, 4)$, $(a_{20}, a_{21}, a_{22}) = (3, 2, 5)$, $\sigma_0^2 = 0.4$, $\sigma_1^2 = \sigma_2^2 = 0.1$, $n_1 = 50$, $n_2 = 50$ 以及 $m = 200$. 定义平均均方误差(average square error, 简称ASE):

$$\text{ASE} = \sqrt{\sum_{j=1}^n \sum_{i=1}^m (y_{ij} - \hat{\eta}(t_i))^2 / nm}. \quad (5.6)$$

图1是参数估计和非参数估计的拟合图, 图2是参数估计和非参数估计的ASE箱形图. 从这两个图都能得到结论, 对于这个模拟模型, 本文所采用的非参数回归方法比传统的参数估计方法效果要好.

§6. 结 论

迄今为止, 对增长曲线的研究很多, 但大多数都是用参数方法研究, 用非参数方法研究的几乎没有. 本文首次将非参数估计方法应用到增长曲线模型中. 从模拟的结果可以很清楚的看出本文所采用的非参数回归方法比传统的参数估计方法效果要好.

当多项式的假设对数据的拟合不再合适的时候, 非参数方法是个很好的选择. 在非参数回归模型中, 回归函数的形式自由, 约束较少, 同时对数据的分布一般不做任何要求, 且这种模型的精度和稳健性都较高. 局部多项式估计具有相对小的偏差和方差、不受边界效应的影响、适用于各种设计等优点. 因此, 本论文在对增长曲线进行估计时使用局部多项式方法. 在今后, 可以进一步研究非参数增长曲线模型的方差估计和偏差估计, 也可考虑用其它方法继续研究这个模型.

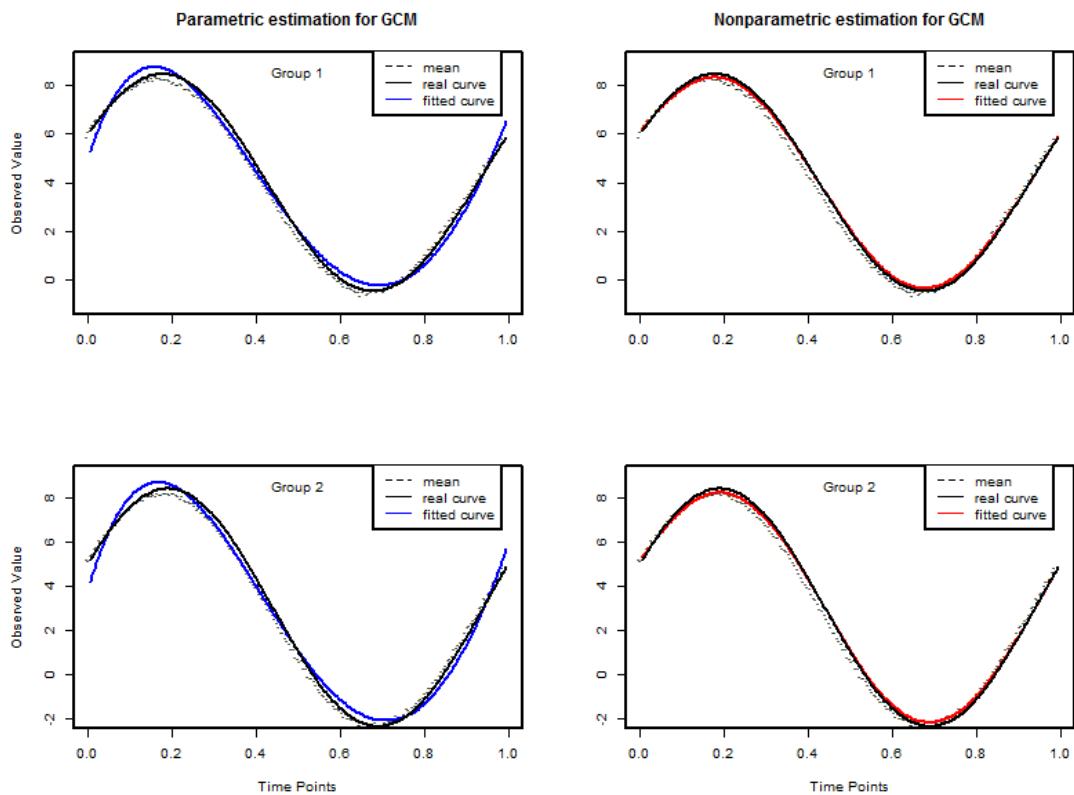


图1 增长曲线模型的参数估计和非参数估计拟合图, 其中蓝色和红色实线分别表示参数估计和非参数估计的拟合曲线, 黑色实线表示真实曲线, 虚线表示均值曲线, 核函数选取Epanechnikov

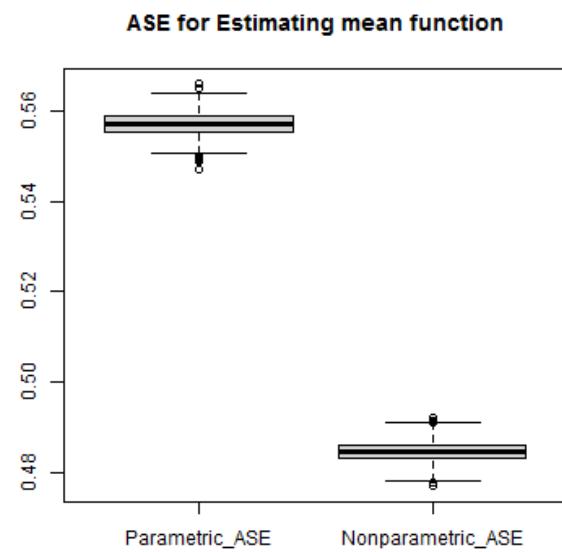


图2 参数估计和非参数估计的ASE箱形图

《应用概率统计》版权所有

参 考 文 献

- [1] 潘建新, 增长曲线模型中回归参数的最小二乘估计及Gauss-Markov定理, *数理统计与应用概率*, **3(2)**(1988), 169–185.
- [2] 张日权, 生长曲线模型中LSE的一个新的相对效率, *应用概率统计*, **14(3)**(1998), 297–300.
- [3] Rao, C.R., The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves, *Biometrika*, **52(3/4)**(1965), 447–458.
- [4] 俞启泰, 三角函数型增长曲线预测油田开发指标, *新疆石油地质*, **24(1)**(2003), 55–59.
- [5] Fan, J. and Gijbels, I., *Local Polynomial Modeling and its Applications*, Chapman and Hall, London, 1996.
- [6] Fan, J. and Gijbels, I., Variable bandwidth and local linear regression smoothers, *The Annals of Statistics*, **20(4)**(1992), 2008–2036.
- [7] Wu, H.L. and Zhang, J.T., *Nonparametric Regression Methods for Longitudinal Data Analysis*, John Wiley and Sons, Canada, 2006.

Nonparametric Regression Method for Growth Curve Model

GAO CAIWEN

(School of Mathematics and Computer Science, Shanxi Datong University, Datong, 037009)

GAN HUALAI

(School of Finance and Statistics, East China Normal University, Shanghai, 200241)

In the research it is frequently assumed that the growth curve is a polynomial in time. In practice, researchers mainly use higher-order polynomials to obtain more precise estimates. But this method has many defects, such as the model can be easily affected by outliers and the polynomial hypothesis may be much strong in practice. So in this paper we first proposed nonparametric approach, local polynomial, instead of parametric method for estimation in growth curve model. We give the nonparametric growth curve model, and its nonparametric estimation. Then discuss the large sample character of local polynomial estimate. The ideal theoretical choice of a local bandwidth is also discussed in detail in this paper. Finally, through the simulation study, from the fitting curve and average square error box plot we can clearly see that the performance of nonparametric approach is much better than parametric technique.

Keywords: Growth curve model, nonparametric estimation, local polynomial smoother, ideal choice of bandwidth.

AMS Subject Classification: 62G08.