系 统 科 学 与 数 学
J. Sys. Sci. & Math. Scis.
27(6) (2007, 12), 847–857

纵向数据下半参数回归模型的统计分析*

田 萍

(许昌学院数学系，许昌 461000; 北京工业大学应用数理学院，北京 100022)

薛 留 根

(北京工业大学应用数理学院，北京 100022)

摘要 对于纵向数据下半参数回归模型，基于广义估计方程和一般权函数方法构造了模型中参数分量和非参数分量的估计．在适当的条件下证明了参数估计量具有渐近正态性，并得到了非参数回归函数估计量的最优收敛速度．通过模拟研究说明了所提出的估计量在有限样本下的精确性．

关键词 纵向数据，半参数回归模型，广义估计方程，渐近正态性，收敛速度．

MR(2000) 主题分类号 62G05, 62G20

1 引 言

考虑纵向数据下半参数回归模型

$$y_{ij} = x'_{ij}\beta + g(t_{ij}) + e_{ij}, \quad i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, m_i, \tag{1.1}$$

其中 $(x_{ij}, t_{ij}) \in R^p \times R$ 是已知的设计点列， β 是 p 维未知参数， $g(\cdot)$ 是定义在闭区间 $[0,1]$ 上的未知回归函数， e_{ij} 是随机误差．记 $e_i = (e_{i1}, e_{i2}, \cdots, e_{im_i})'$, $\{e_i, i = 1, 2, \cdots, n\}$ 相互独立， $E(e_i) = 0$, $\mathrm{Var}(e_i) = \Sigma_i$(正定). 总的观测个数为 $N = \sum_{i=1}^{n} m_i$. 本文假定 n 可以充分大，而 m_i 为有界正整数序列．

近年来，该模型已经引起了许多统计学者的兴趣．他们分别使用不同的方法构造了 β 和 $g(\cdot)$ 的估计量，并研究了估计量的渐近性质．譬如，后移算法，核广义估计方程， M 估计和最小二乘估计等，参见文献 [1–4]. 但是，这些研究都是在 (x_{ij}, t_{ij}) 为随机设计点列情形下进行的．当 (x_{ij}, t_{ij}) 为固定设计点列时的研究所见不多．钱伟民 [5] 在 x_{ij} 为随机设计点列而 t_{ij} 为固定设计点列下采用二阶段估计方法构造了 β 和 $g(\cdot)$ 的估计量，证明了它们具有强相合性．孙孝前和尤进红 [6] 在 (x_{ij}, t_{ij}) 为固定设计点列下提出了参数分量的一个迭代加权偏样条最小二乘估计方法，并证明了所提出估计量具有渐近正态性．

* 国家自然科学基金 (10571008)，河南省自然科学基金 (0511013300) 和河南省教育厅自然科学基金 (2007110033) 资助课题．

收稿日期：2005-09-02, 收到修改稿日期：2006-10-16.

自 Liang 和 Zeger[7] 在纵向数据下研究广义线性模型参数估计问题中提出广义估计方程 (GEE) 以来， GEE 方法在参数回归模型下得到了广泛的使用． Lin 和 Carroll[2,8] 在随机设计下，结合非参数核估计方法把 GEE 推广到纵向数据下非参数和半参数回归模型．本文在固定设计情形下考虑模型 (1.1); 基于参数分量的 GEE 和非参数分量的一般权函数方法分别构造了未知参数 β 和未知函数 $g(\cdot)$ 的估计量；在适当条件下证明了估计量 $\widehat{\beta}$ 的渐近正态性，并给出了 $\widehat{g}(\cdot)$ 的最优收敛速度．本文还做了数据模拟计算，利用模拟结果说明了所提出的估计量在有限样本下的精度．

2 主要结论

为了表述方便起见，我们引入如下记号

$$\widetilde{x}_{ij} = x_{ij} - \sum_{k=1}^{n}\sum_{l=1}^{m_k} W_{kl}(t_{ij})x_{kl}, \quad \widetilde{y}_{ij} = y_{ij} - \sum_{k=1}^{n}\sum_{l=1}^{m_k} W_{kl}(t_{ij})y_{kl},$$

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{im_i})', \qquad y_i = (y_{i1}, y_{i2}, \cdots, y_{im_i})',$$

$$\widetilde{x}_i = (\widetilde{x}_{i1}, \widetilde{x}_{i2}, \cdots, \widetilde{x}_{im_i})', \qquad \widetilde{y}_i = (\widetilde{y}_{i1}, \widetilde{y}_{i2}, \cdots, \widetilde{y}_{im_i})',$$

其中 $W_{kl}(t) = W_{kl}(t; t_{11}, t_{12}, \cdots, t_{nm_n})$ 为概率权函数．

由 (1.1) 式可得， $y_{ij} - x'_{ij}\beta = g(t_{ij}) + e_{ij}$, 由非参数估计中的一般权函数方法可以定义非参数分量 $g(\cdot)$ 的初始估计为

$$\check{g}(t) = \sum_{k=1}^{n}\sum_{l=1}^{m_k} W_{kl}(t)(y_{kl} - x'_{kl}\beta).$$

记 $\check{g}(t_i) = \big(\check{g}(t_{i1}), \check{g}(t_{i2}), \cdots, \check{g}(t_{im_i})\big)'$, V_i 为任意指定的 e_i 的作业协方差矩阵．定义 β 的估计是下式的解

$$\sum_{i=1}^{n}\big(y_i - x_i\beta - \check{g}(t_i)\big)'V_i^{-1}\big(y_i - x_i\beta - \check{g}(t_i)\big) = \min!. \tag{2.1}$$

该极小化问题可转化为解估计方程

$$\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}(\widetilde{y}_i - \widetilde{x}_i\beta) = 0.$$

可解得 β 的估计量为

$$\widehat{\beta} = \left(\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\widetilde{x}_i\right)^{-1}\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\widetilde{y}_i. \tag{2.2}$$

将 $\widehat{\beta}$ 代入 $\check{g}(t)$, 得 $g(t)$ 的最终估计为

$$\widehat{g}(t) = \sum_{k=1}^{n}\sum_{l=1}^{m_k} W_{kl}(t)(y_{kl} - x'_{kl}\widehat{\beta}). \tag{2.3}$$

本文中我们作如下假定．

条件 2.1　存在定义在 [0,1] 上的函数 $h_s(\cdot)$, 使得对 $1 \le s \le p$, 有

$$x_{ijs} = h_s(t_{ij}) + u_{ijs}, \quad i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, m_i,$$

其中 u_{ijs} 满足

i) $\lim\limits_{n\to\infty} \frac{1}{n} \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{m_i} u_{ij}u'_{ij} = A$;

ii) $\limsup\limits_{N\to\infty} \frac{1}{\sqrt{N \log N}} \max\limits_{1\le k\le N} \left\| \sum\limits_{i=1}^{k} v_{j_i} \right\| < \infty$,

这里 A 是正定阵，$u_{ij} = (u_{ij1}, u_{ij2}, \cdots, u_{ijp})'$, $v_1 = u_{11}, \cdots, v_{m_1} = u_{1m_1}, \cdots, v_N = u_{nm_n}$. $(j_1, j_2, \cdots, j_N)$ 是 $(1, 2, \cdots, N)$ 的任意排列. $\|\cdot\|$ 表示 Euclid 模.

条件 2.2　$g(\cdot)$ 和 $h_s(\cdot)$ 在 [0,1] 上均满足一阶 Lipschitz 条件.

条件 2.3　当 n 充分大时，$W_{ij}(\cdot)$ 满足

i) $\max\limits_{1\le i\le n, 1\le j\le m_i} \sum\limits_{k=1}^{n} \sum\limits_{l=1}^{m_k} W_{ij}(t_{kl}) = O(1)$;

ii) $\sup\limits_{0\le t\le 1} \max\limits_{1\le i\le n, 1\le j\le m_i} W_{ij}(t) = O\big(n^{-\frac{2}{3}}\big)$;

iii) $\sup\limits_{0\le t\le 1} \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{m_i} W_{ij}(t) I\big(|t_{ij} - t| > \delta n^{-\frac{1}{3}}\big) = O\big(n^{-\frac{1}{3}}\big)$, 对任给的$\delta > 0$.

条件 2.4　存在常数 c_1, c_2, c_3, c_4, 使得对所有的 $i = 1, 2, \cdots, n$, 有

$$0 < c_1 \le \min_{1\le i\le n} \lambda_{i1} \le \max_{1\le i\le n} \lambda_{im_i} \le c_2 < \infty,$$

$$0 < c_3 \le \min_{1\le i\le n} \lambda'_{i1} \le \max_{1\le i\le n} \lambda'_{im_i} \le c_4 < \infty,$$

其中 λ_{i1} 和 λ_{im_i} 表示 Σ_i 的最小和最大特征根，λ'_{i1} 和 λ'_{im_i} 表示 V_i 的最小和最大特征根.

条件 2.5　$\sup\limits_i E\|e_i\|^4 < \infty$.

注 1　上述条件都是很普通的假定. 条件 2.1 的合理性可参看文献 [6,9–11] 等. 满足上述条件 2.3 的权函数是存在的, 例如常见的核权和近邻权, 可分别仿文献 [12,13] 进行验证, 这里不再详细给出.

在上述假定下，我们有以下结论.

定理 2.1　设条件 2.1–2.4 成立, 则当 $n \to \infty$ 时，有

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{L} N(0, B^{-1}CB^{-1}),$$

其中 $B = \lim\limits_{n\to\infty} \frac{1}{n} \sum\limits_{i=1}^{n} \widetilde{x}'_i V_i^{-1} \widetilde{x}_i$, $C = \lim\limits_{n\to\infty} \frac{1}{n} \sum\limits_{i=1}^{n} \widetilde{x}'_i V_i^{-1} \Sigma_i V_i^{-1} \widetilde{x}_i$.

注 2　当作业协方差阵 $V_i = I$ 时，即假定作业独立 (参见文献 [2,8]), 则 β 的估计为

$$\widehat{\beta}_I = \left(\sum_{i=1}^{n} \widetilde{x}'_i \widetilde{x}_i \right)^{-1} \sum_{i=1}^{n} \widetilde{x}'_i \widetilde{y}_i.$$

它是通常的最小二乘极小化问题 $\sum\limits_{i=1}^{n} \big(y_i - x_i\beta - \check{g}(t_i)\big)' \big(y_i - x_i\beta - \check{g}(t_i)\big) = \min!$ 的解.

注 3 当作业协方差阵 $V_i = \Sigma_i$ 时，β 的估计为

$$\widehat{\beta}^* = \Big(\sum_{i=1}^n \widetilde{x}_i'\Sigma_i^{-1}\widetilde{x}_i\Big)^{-1}\sum_{i=1}^n \widetilde{x}_i'\Sigma_i^{-1}\widetilde{y}_i.$$

记 $D \widehat{=} \lim\limits_{n\to\infty}\frac{1}{n}\sum\limits_{i=1}^n \widetilde{x}_i'\Sigma_i^{-1}\widetilde{x}_i$, 则 $\widehat{\beta}^*$ 的渐近方差为 D^{-1}. 由于 $((C,B)',(B,D)')$ 为对称且半正定矩阵，由广义的 Cauchy-Schwarz 不等式可得 $B^{-1}CB^{-1} \geq D^{-1}$, 即 $\widehat{\beta}^*$ 的渐近方差达到最小，估计 $\widehat{\beta}^*$ 在所有的估计 $\widehat{\beta}$ 中最有效.

当 Σ_i 未知时，$\widehat{\beta}$ 的渐近方差在实际中不可用，为此我们还需要给出 C 的相合估计 $\widehat{C} = \frac{1}{n}\sum\limits_{i=1}^n \widetilde{x}_i'V_i^{-1}\widehat{e}_i\widehat{e}_i'V_i^{-1}\widetilde{x}_i$, 其中 $\widehat{e}_i = y_i - x_i\widehat{\beta}_I - \widehat{g}(t_i)$. 下述定理表明 $\widehat{C}$ 是 C 的相合估计.

定理 2.2 设条件 2.1–2.5 成立，则当 $n\to\infty$ 时，有

$$\widehat{C} \xrightarrow{P} C.$$

定理 2.3 设条件 2.1–2.4 成立，则当 $n\to\infty$ 时，有

$$\widehat{g}(t) - g(t) = O_p\big(n^{-\frac{1}{3}}\big).$$

注 4 本文的估计方法是参数回归模型中 GEE 方法在半参数回归模型中的推广. 正如参数 GEE 一样，本文的半参数模型 (1.1) 也属于边缘模型，其中感兴趣的是回归参数 β 的估计，而不是组内协方差阵. 由定理 2.1 和 2.2 可见，无论如何指定 V_i, 都不影响 $\widehat{\beta}$ 及其渐近方差估计量的相合性. 一般说来，V_i 的选择应与观测到的协方差阵相一致. 然而，即使错误指定组内协方差矩阵，基于个体间的独立性仍然可以获得 $\widehat{\beta}$ 及其渐近方差之估计量的相合性，这是 GEE 方法的一个非常吸引人的特点. 尽管错误指定 V_i 可能会降低 $\widehat{\beta}$ 的效，但当个体数目 n 增大时，效的损失会逐渐减少. 实际应用中，作业协方差矩阵 V_i 可指定为依赖某个相同的参数 α, 即 $V_i = V_i(\alpha)$. 具体有独立、可交换、 AR(1) 和 m 相依等几种形式. α 可以是已知的，也可以是未知的. 当 α 未知时，可以采用矩法来估计 (参见文献 [2, 7] 等). 例如，当 $V_i \equiv V$ 时，V 的估计可以采用 $\widehat{V} = n^{-1}\sum\limits_{i=1}^n r_i r_i'$, 其中 $r_i = y_i - x_i'\widehat{\beta} - \widehat{g}(t_i)$.

3 定理的证明

以下用 c 表示任一不依赖于 n 的正常数，在不同地方可取不同的值.

引理 3.1 设条件 2.2 和条件 2.3 iii) 成立，则当 n 充分大时，有

$$\max_{1\leq i\leq n, 1\leq j\leq m_i}\Big|\widetilde{f}(t_{ij})\Big| \widehat{=} \max_{1\leq i\leq n, 1\leq j\leq m_i}\Big|f(t_{ij}) - \sum_{k=1}^n\sum_{l=1}^{m_k} W_{kl}(t_{ij})f(t_{kl})\Big| = O\big(n^{-\frac{1}{3}}\big),$$

其中 $f(\cdot) = g(\cdot), h_s(\cdot), s = 1, 2, \cdots, p$.

证 由条件 2.2 和条件 2.3 iii) 即证得. 这里证明省略.

引理 3.2 设条件 2.1 和条件 2.3 成立，且 $h_s(\cdot)$ 满足条件 2.2, 则有

$$A = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \widetilde{x}_i'\widetilde{x}_i \tag{3.1}$$

存在且正定．若再满足条件 2.4, 则有

$$B=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\widetilde{x}_i, \tag{3.2}$$

$$C=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\Sigma_iV_i^{-1}\widetilde{x}_i \tag{3.3}$$

存在且正定.

证　仿文 [14] 中引理 4 的证明即可证得 (3.1) 式．(3.2) 与 (3.3) 式的证明方法类同，下面仅给出 (3.2) 式的证明.

由于 V_i^{-1} 为实对称阵, 则存在标准正交化特征向量构成的矩阵 $\Phi_i=(\varphi_{i1},\varphi_{i2},\cdots,\varphi_{im_i})$, 使得 $V_i^{-1}=\Phi_i\mathrm{diag}(\lambda_{i1},\lambda_{i2},\cdots,\lambda_{im_i})\Phi_i'$, 其中 φ_{ij} 是对应 V_i^{-1} 的特征根 λ_{ij} 的特征向量，并且存在 $m_i\times p$ 的矩阵

$$b_i=\begin{pmatrix} b_{i11} & b_{i12} & \cdots & b_{i1p}\\ b_{i21} & b_{i22} & \cdots & b_{i2p}\\ \vdots & \vdots & \ddots & \vdots\\ b_{im_i1} & b_{im_i2} & \cdots & b_{im_ip}\end{pmatrix}$$

使得 $\widetilde{x}_i=\Phi_ib_i$. 记 $b_{ij}=(b_{ij1},b_{ij2},\cdots,b_{ijp})'$, 由 (3.1) 式知

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i'\widetilde{x}_i=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}b_i'\Phi_i'\Phi_ib_i=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}b_{ij}b_{ij}'=A.$$

故由条件 2.4 可得

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\widetilde{x}_i=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}b_i'\Phi_i'V_i^{-1}\Phi_ib_i=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\lambda_{ij}b_{ij}b_{ij}'=B.$$

引理 3.3　设条件 2.1 ii), 条件 2.3 ii) 和条件 2.4 成立，则对 $1\le s\le p$, 有

$$\max_{1\le k\le n,1\le l\le m_k}\left|\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}W_{kl}(t_{iv})u_{ijs}\right|=o(1),$$

$$\max_{1\le k\le n,1\le l\le m_k}\left|\sum_{r=1}^{n}\sum_{q=1}^{m_r}\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}W_{kl}(t_{iv})W_{rq}(t_{ij})u_{rqs}\right)\right|=o(1),$$

其中 σ_i^{jv} 为 V_i^{-1} 的第 (j,v) 元.

证　应用 Abel 不等式，仿文献 [10] 中定理 2 的证明可证得

$$\begin{aligned}&\left|\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}W_{kl}(t_{iv})u_{ijs}\right|\\ \le\ & c\max_{1\le i\le n,1\le j,v\le m_i}|\sigma_i^{jv}|\cdot\max_{1\le i\le n,1\le v\le m_i}W_{kl}(t_{iv})\cdot\max_{1\le r\le N}\left|\sum_{i=1}^{r}v_{j_is}\right|.\end{aligned}$$

由条件 2.1 ii), 条件 2.3 ii) 和条件 2.4 易见第一式成立. 同理可证明第二式.

定理 2.1 的证明 由 (2.2) 式, 经过简单运算可得

$$\begin{aligned}\widehat{\beta}-\beta &= \Big(\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\widetilde{x}_i\Big)^{-1}\Big[\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}e_i-\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\overline{e}_i+\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\widetilde{g}(t_i)\Big]\\ &\widehat{=}\Big(\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\widetilde{x}_i\Big)^{-1}(B_1-B_2+B_3),\end{aligned}\tag{3.4}$$

其中 $\overline{e}_i=\sum\limits_{k=1}^{n}\sum\limits_{l=1}^{m_k}W_{kl}(t_i)e_{kl}$, 且 $\widetilde{g}(t_i)$ 在引理 1 中定义. 记

$$B_{2,s}=\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}\widetilde{x}_{ijs}\overline{e}_{iv},\qquad B_{3,s}=\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}\widetilde{x}_{ijs}\widetilde{g}(t_{iv}),$$

$$B_{21,s}=\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}\widetilde{h}_s(t_{ij})\overline{e}_{iv},\ B_{31,s}=\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}\widetilde{h}_s(t_{ij})\widetilde{g}(t_{iv}),$$

$$B_{22,s}=\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}\widetilde{u}_{ijs}\overline{e}_{iv},\qquad B_{32,s}=\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}\widetilde{u}_{ijs}\widetilde{g}(t_{iv}),$$

其中 $\widetilde{u}_{ijs}=u_{ijs}-\sum\limits_{k=1}^{n}\sum\limits_{l=1}^{m_k}W_{kl}(t_{ij})u_{ijs},\ s=1,2,\cdots,p$. 经过计算可得

$$\begin{aligned}&E(B_{21,s}^2)\\ &=\sum_{k=1}^{n}\sum_{l=1}^{m_k}\Big(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}\widetilde{h}_s(t_{ij})W_{kl}(t_{iv})\Big)^2Ee_{kl}^2\\ &\quad+\sum_{k=1}^{n}\sum_{l\neq q}\Big[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}\widetilde{h}_s(t_{ij})W_{kl}(t_{iv})\Big]\Big[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}\widetilde{h}_s(t_{ij})W_{kq}(t_{iv})\Big]E(e_{kl}e_{kq})\\ &\widehat{=}E_{11}+E_{12}.\end{aligned}$$

由引理 3.1, 条件 2.3 i), ii) 和条件 2.4 可知

$$\begin{aligned}E_{11}&\le c\sum_{k=1}^{n}\sum_{l=1}^{m_k}\Big(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}\widetilde{h}_s(t_{ij})W_{kl}(t_{iv})\Big)^2=o(n),\\ E_{12}&\le\sum_{k=1}^{n}\sum_{l\neq q}\Big[\Big(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}\widetilde{h}_s(t_{ij})W_{kl}(t_{iv})\Big)^2+\Big(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}\widetilde{h}_s(t_{ij})W_{kq}(t_{iv})\Big)^2\Big]\\ &=o(n).\end{aligned}$$

故 $E(B_{21,s}^2)=o(n)$. 对 $B_{22,s}^2$ 求期望得

$$\begin{aligned}E(B_{22,s}^2)=&\sum_{k=1}^{n}\sum_{l=1}^{m_k}\Big(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}\widetilde{u}_{ijs}W_{kl}(t_{iv})\Big)^2Ee_{kl}^2\\&+\sum_{k=1}^{n}\sum_{l\neq q}\Big[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}\widetilde{u}_{ijs}W_{kl}(t_{iv})\Big]\Big[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}\widetilde{u}_{ijs}W_{kq}(t_{iv})\Big]E(e_{kl}e_{kq})\\\widehat{=}\;&E_{21}+E_{22}.\end{aligned}$$

由引理 3.3 可知

$$\begin{aligned}E_{21}\le cn\Big\{&\max_{1\le k\le n,1\le l\le m_k}\Big|\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}W_{kl}(t_{iv})u_{ijs}\Big|^2\\&+\max_{1\le k\le n,1\le l\le m_k}\Big|\sum_{r=1}^{n}\sum_{q=1}^{m_r}\Big(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}\sigma_i^{jv}W_{kl}(t_{iv})W_{rq}(t_{ij})\Big)u_{rqs}\Big|^2\Big\}\\=&\,o(n).\end{aligned}$$

类似地，可以得到 $E_{22}=o(n)$. 由此可得 $E(B_{22,s}^2)=o(n)$. 故

$$B_{2,s}=o_p(n^{\frac12}). \tag{3.5}$$

由引理 3.1 知

$$B_{31,s}\le cn\max_{1\le i\le n,1\le j,v\le m_i}|\sigma_i^{jv}|\cdot\max_{1\le i\le n,1\le j\le m_i}\Big|\widetilde{h}_s(t_{ij})\Big|\cdot\max_{1\le i\le n,1\le v\le m_i}\Big|\widetilde{g}(t_{iv})\Big|=o(n^{\frac12}).$$

由引理 3.1 且利用 Abel 不等式，类似于引理 3.3 的证明，可证得 $B_{32,s}=o(n^{\frac12})$. 故

$$B_{3,s}=o(n^{\frac12}). \tag{3.6}$$

由引理 3.2 知

$$\sqrt{n}\Big(\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\widetilde{x}_i\Big)^{-1}B_2=o_p(1),\quad \sqrt{n}\Big(\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\widetilde{x}_i\Big)^{-1}B_3=o(1).$$

由 (3.4),(3.5) 和 (3.6) 式可见，为证定理 3.1, 只需证明

$$\sqrt{n}\Big(\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\widetilde{x}_i\Big)^{-1}\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}e_i\xrightarrow{L}N(0,B^{-1}CB^{-1}). \tag{3.7}$$

记 $\xi=\sum\limits_{i=1}^{n}\widetilde{x}_i'V_i^{-1}e_i$. 对于任意 $c\in R^p$, 且满足 $\|c\|=1$, $c'\xi=\sum\limits_{i=1}^{n}c'\widetilde{x}_i'V_i^{-1}e_i$ 为独立随机变量之和，且 $E(c'\xi)=0$, $\mathrm{Var}(c'\xi)=\sum\limits_{i=1}^{n}c'\widetilde{x}_i'V_i^{-1}\Sigma_iV_i^{-1}\widetilde{x}_ic$. 因此可以记 $c'\xi\widehat{=}\sum\limits_{i=1}^{n}a_i\varepsilon_i$, 其中 $a_i^2=c'(\widetilde{x}_i'V_i^{-1}\Sigma_iV_i^{-1}\widetilde{x}_i)c$, ε_i 相互独立，且 $E\varepsilon_i=0$, $E\varepsilon_i^2=1$. 这样，为证 (3.7) 式，需要证明

$$\frac{\sum\limits_{i=1}^{n}a_i\varepsilon_i}{\sqrt{\sum\limits_{i=1}^{n}a_i^2}}\xrightarrow{L}N(0,1). \tag{3.8}$$

由引理 3.2 知

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}a_i^2=\lim_{n\to\infty}\frac{1}{n}c'\Big(\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\Sigma_iV_i^{-1}\widetilde{x}_i\Big)c>0.$$

则由文 [15] 中引理 3 可知 $\max\limits_{1\le i\le n}a_i^2\Big(\sum\limits_{i=1}^{n}a_i^2\Big)^{-1}\longrightarrow 0$. 再由文 [16] 中命题 2.2 知 (3.8) 式成立. 故

$$\Big(\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\Sigma_iV_i^{-1}\widetilde{x}_i\Big)^{-\frac{1}{2}}\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}e_i\xrightarrow{L}N(0,I_p).$$

由此式和引理 3.2 即证得 (3.7) 式成立. 故定理 3.1 得证.

定理 2.2 的证明 记

$$I_1=\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\big(\widehat{e}_i\widehat{e}_i'-e_ie_i'\big)V_i^{-1}\widetilde{x}_i,\quad I_2=\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\big(e_ie_i'-\Sigma_i\big)V_i^{-1}\widetilde{x}_i,$$

其中 $\widehat{e}_i=(\widehat{e}_{i1},\widehat{e}_{i2},\cdots,\widehat{e}_{im_i})$, $\widehat{e}_{ij}=\widetilde{y}_{ij}-\widetilde{x}_{ij}'\widehat{\beta}_I$. 则有

$$\widehat{C}=\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}\Sigma_iV_i^{-1}\widetilde{x}_i+I_1+I_2.\tag{3.9}$$

用 $d_k=(0,0,\cdots,1,\cdots,0)'(k=1,2,\cdots,p)$ 表示第 k 个分量为 1，其余分量均为 0 的 p 维单位向量. 对于任意的 k,l, 由引理 3.1, 3.2 及条件 2.3, 2.4 可证得

$$\begin{aligned}E\big|d_k'I_1d_l\big|&\le c\bigg(d_k'\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}V_i^{-1}\widetilde{x}_id_k+d_l'\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i'V_i^{-1}V_i^{-1}\widetilde{x}_id_l\bigg)\max_{i,j,q}E\big|\widehat{e}_{ij}\widehat{e}_{iq}-e_{ij}e_{iq}\big|\\&=o(1).\end{aligned}$$

因此， $I_1=o_p(1)$. 记 $W_i\widehat{=}V_i^{-1}\widetilde{x}_i$, W_{ij} 表示 W_i 的第 j 行，则对于任意的 k,l, 由引理 3.2 和条件 2.5 可知

$$\begin{aligned}E\big[d_k'I_2d_l\big]^2&\le\frac{c}{n^2}\sum_{i=1}^{n}\bigg[\sum_{j=1}^{m_i}\sum_{q=1}^{m_i}(W_{ij}'d_k)^2\cdot(W_{iq}'d_l)^2\bigg]\\&\le\frac{c}{n}\max_{i,q}(W_{iq}'d_l)^2\cdot\frac{1}{n}\sum_{i=1}^{n}d_k'\widetilde{x}_i'V_i^{-1}V_i^{-1}\widetilde{x}_id_k\\&=o(1).\end{aligned}$$

则 $I_2=o_p(1)$. 故由 (3.9) 式和引理 3.2 的 (3.3) 式即证得定理 3.2.

定理 2.3 的证明 经过计算得

$$\begin{aligned}\widehat{g}(t)-g(t)&=\sum_{k=1}^{n}\sum_{l=1}^{m_k}W_{kl}(t)x_{kl}'\big(\beta-\widehat{\beta}\big)+\sum_{k=1}^{n}\sum_{l=1}^{m_k}W_{kl}(t)e_{kl}\\&\quad+\bigg(\sum_{k=1}^{n}\sum_{l=1}^{m_k}W_{kl}(t)g(t_{kl})-g(t)\bigg)\\&\widehat{=}G_1+G_2+G_3.\end{aligned}\tag{3.10}$$

由条件 2.2, 存在正数 M, 使得 $\sup\limits_{0\le t\le 1}|h_s(t)|\le M$, 则

$$\left|\sum_{k=1}^{n}\sum_{l=1}^{m_k}W_{kl}(t)h_s(t_{kl})\right|\le M.$$

又由 $\left|\sum\limits_{k=1}^{n}\sum\limits_{l=1}^{m_k}W_{kl}(t)u_{kls}\right|=o(1)$ 和定理 3.1 知 $G_1=O_p(n^{-\frac{1}{2}})$. 由于

$$E\left[\sum_{k=1}^{n}\sum_{l=1}^{m_k}W_{kl}(t)e_{kl}\right]^2=O_p(n^{-\frac{2}{3}}),$$

则 $G_2=O_p(n^{-\frac{1}{3}})$. 类似于引理 3.1 的证明可知 $G_3=O_p(n^{-\frac{1}{3}})$. 故由 (3.10) 式可证得定理 3.3.

4　模拟结果

模拟中我们考虑重复观测次数相等的情况. 取 $p=1$, $m=3$, $\beta=2$, 假定组内协方差阵相等. (x_{ij},t_{ij}) 可分别看作来自于 $N(0,1)$ 和 $U(0,1)$ 的样本观测值. e_{ij} 来自于 $N(0,0.8^2)$, 组内相关系数 $\rho_{e_{ij}e_{ik}}\equiv 0.6$. 取权函数 $W(\cdot)$ 为核权, 其中核函数取为 Epanechnikov 核: $K(u)=0.75(1-u^2)_+$. 此时窗宽应满足 $h=O(n^{-\frac{1}{3}})$. 因为这个条件没有达到非参数回归函数核估计的最优窗宽的速度, 即 $O\left(n^{-\frac{1}{5}}\right)$, 那么要得到一个合理的窗宽选择规则是困难的. 因此我们需要采用 "undersmoothing" 方法选取窗宽 (见文 [17]), 其做法如下: 首先, 用删除一组的交叉核实方法选择一个最优窗宽 h_{opt}. 然后用 h_{opt} 乘以 $n^{-\frac{2}{15}}$ 即可得到一个近似的窗宽 $\widehat{h}$, 即 $\widehat{h}=h_{opt}\cdot n^{-\frac{2}{15}}$. 由于 h_{opt} 的收敛速度为 $O(n^{-\frac{1}{5}})$, 因此 $\widehat{h}$ 的收敛速度为 $O(n^{-\frac{1}{3}})$, 即可满足本文对窗宽的要求.

对于不同的个体数目和不同的函数 $g(t)$, 分别就 $V=I$ 和 $V=\Sigma_i$ 对 β 的估计精度进行了模拟计算, 重复模拟次数为 5000 次. 就估计量的估计值、偏差 (Bias) 、标准差 (SD) 和均方误差 (MSE) 进行模拟计算, 其结果如表 1 所示.

表 1　参数 β 的估计值及其精度

$g(\cdot)$	n	估计量	估计值	Bias	SD	MSE
	30	$\widehat{\beta}_I$	2.0080	0.0080	0.0889	0.0079
		$\widehat{\beta}^*$	2.0045	0.0045	0.0663	0.0044
$\sin(2\pi t)$	50	$\widehat{\beta}_I$	1.9972	-0.0028	0.0728	0.0054
		$\widehat{\beta}^*$	1.9974	-0.0026	0.0539	0.0029
	30	$\widehat{\beta}_I$	1.9915	-0.0085	0.0954	0.0091
		$\widehat{\beta}^*$	1.9915	-0.0085	0.0721	0.0053
e^{2t}	50	$\widehat{\beta}_I$	2.0168	0.0168	0.0663	0.0047
		$\widehat{\beta}^*$	2.0042	0.0042	0.0458	0.0021

当个体数目为 30 时, 对于函数 $g(t)=\sin(2\pi t)$ 和 $g(t)=e^{2t}$, 其估计量 $\widehat{g}^*(t)$ 的计算结果分别如图 1 和图 2 所示.

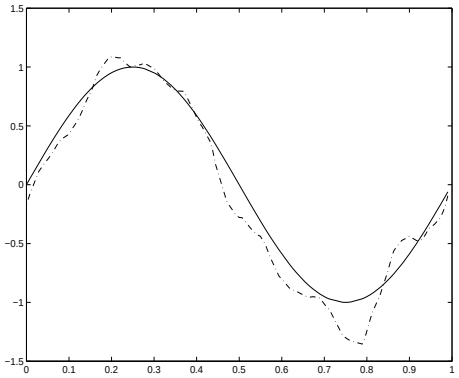

图 1 $g(t)=\sin(2\pi t)$ 的真实曲线 (实线) 和估计曲线 (虚线)

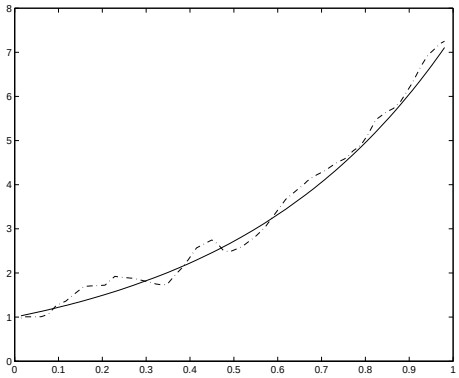

图 2 $g(t)=\mathrm{e}^{2t}$ 的真实曲线 (实线) 和估计曲线 (虚线)

从表 1 我们容易看到两个明显的特点：一是随着观测个体数目的增加, 对于不同的函数 $g(t)$, β 的两种估计 $\widehat{\beta}_I$ 和 $\widehat{\beta}^*$ 都越来越接近于 β 的真值 2, 估计的偏差, 标准差和均方差也相应减小, 这说明样本容量增大时 β 拟合的效果越来越好; 二是对于相同的函数和观测样本来说, $\widehat{\beta}^*$ 的拟合效果明显优于 $\widehat{\beta}_I$, 这是由于估计 $\widehat{\beta}^*$ 考虑了组内的协方差结构, 而 $\widehat{\beta}_I$ 则完全忽视了组内的协方差结构, 其结果必然造成估计的效的降低. 图 1 表明函数 $g(\cdot)$ 的估计也有较好的效果.

参 考 文 献

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

[2] Lin X H and Carrol R J. Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, 2001, **96**: 1045–1056.

[3] He X M, Zhu Z Y and Fung W K. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, 2002, **89**(3): 579–590.

[4] Fan J Q and Li R Z. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, 2004, **99**: 710–723.

[5] 钱伟民. 纵向数据统计模型中的估计方法. 同济大学博士学位论文， 2003.

[6] 孙孝前, 尤进红. 纵向数据半参数建模中的迭代加权偏样条最小二乘估计. 中国科学， 2003, **33**(5): 470–480.

[7] Liang K Y and Zeger S L. Longitudinal data analysis using generalized linear models. *Biometrika*, 1986, **73**: 13–22.

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

[9] 尤进红，陈歌迈. 带有异方差的部分线性回归模型的 B 样条估计. 数学年刊， 2004, **25**(5): 661–676.

[10] 高集体，洪圣岩，梁华. 部分线性模型中估计的收敛速度. 数学学报， 1995, **38**(5): 658–669.

[11] Härdle W, Liang H and Gao J T. Partially Linear Models. Heidelberg: Physica-Verlag, 2000.

[12] 高集体，赵林城. 部分线性模型中自适应估计. 中国科学 (A 辑), 1992, **22**(8): 791–803.

[13] 石坚. 部分线性模型中的 Edgeworth 展开. 数学学报， 1998, **41**(4): 683–686.

[14] 高集体，陈希孺，赵林城. 部分线性模型中估计的渐近正态性. 数学学报， 1994, **37**(2): 256–268.

[15] Wu C F. Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics*, 1981, **9**: 501–513.
[16] Huber P. Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1973, **1**: 799–821.
[17] Carroll R J, Fan J, Gijbels I and Wand M P. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 1997, **92**: 477–489.

STATISTICAL ANALYSIS OF THE SEMIPARAMETRIC REGRESSION MODEL FOR LONGITUDINAL DATA

Tian Ping

(*Department of Mathematics, Xuchang University, Xuchang* 461000;

College of Applied Sciences, Beijing University of Technology, Beijing 100022)

Xue Liugen

(*College of Applied Sciences, Beijing University of Technology, Beijing* 100022)

Abstract For the semiparametric regression model with longitudinal data, the estimators of paranetric component and nonparametric component are obtained by using generalized estimating equations and usual nonparametric weight function method. Under some suitable conditions, the asymptotic normality of the parametric estimator is shown, and the optimal convergence rate of the estimator of nonparametric regression function is obtained. A simulation result illustrates the finite sample performance of the proposed estimators.

Key words Longitudinal data, semiparametric regression model, generalized estimating equation, asymptotic normality, convergence rate.