

文章编号: 1004-4353(2019)03-0201-07

纵向数据下部分线性模型的二次光滑估计

李生彪

(兰州文理学院 教育学院, 甘肃 兰州 730000)

摘要: 利用二次光滑估计方法研究纵向数据下部分线性模型的估计问题, 给出了二次光滑估计的渐近性质. 进一步计算表明, 在渐近方差不变的前提下, 二次光滑估计的渐近偏差的阶 $o_p(h^4)$ 低于局部线性估计的渐近偏差的阶 $o_p(h^2)$, 即二次光滑估计的效果优于局部线性估计的效果. 利用 CD4 细胞数数据对二次光滑估计方法进行验证表明, 本文所得结果正确.

关键词: 纵向数据; 二次光滑; 部分线性模型; 渐近偏差

中图分类号: O212.1

文献标识码: A

Double smoothing estimation for partial linear model with longitudinal data

LI Shengbiao

(College of Education, Lanzhou University of Arts and Science, Lanzhou 730000, China)

Abstract: We used the double smoothing estimation method to discuss the estimation problem of partial linear model, the asymptotic properties of the double smoothing local linear estimators are given. Further calculation show that the order $o_p(h^4)$ of the asymptotic bias of the double smoothing estimation is lower than the order $o_p(h^2)$ of the asymptotic bias of the local linear estimation with the same asymptotic variance, that is to say, the double smoothing estimation is better than the local linear estimation. Using CD4 cell number data to verify the double smoothing local linear estimation method, the results obtained in this paper are correct.

Keywords: longitudinal data; double smoothing; partial linear model; asymptotic bias

0 引言

1994 年, Zeger 等^[1]首次提出了纵向数据下部分线性模型: $Y_{ij} = \beta^T X_{ij} + g(U_{ij}) + \epsilon_{ij}$, 其中 β 是未知参数向量, $g(\cdot)$ 是未知光滑函数. 因部分线性模型结合了线性模型和非参数模型的特点, 使得该模型具有很好的灵活性, 且具有削减建模偏差、避免“维数祸根”和解释性强等优点, 因而被广泛应用于计量经济学、生物医学等领域. 目前, 部分线性估计方法^[2]是处理独立数据下变系数模型估计问题的常用方法, 但其在部分线性模型的应用中时仍存在一些不足. 例如: 该方法只能在目标点的小区域内拟合直线段, 因而使得该直线段导数的估计值没有得到有效利用, 所得估计的渐近偏差的阶 $o_p(h^2)$ 偏大, 存在稀疏问题, 等等^[3]. 对此, 一些研究者对局部线性估计方法进行了一些改进, 如 HE 等^[4]提出了二次光滑局部线性估计. 该方法通过再次光滑处理, 整合目标点处的所有局部线性拟合值, 使其在不改变渐近方差的阶的前提下, 渐近偏差降低至 $o_p(h^4)$ 阶, 且整体估计效果与局部立方回归估计相当, 较好地克服了稀

疏问题. 此后, 一些学者对二次光滑局部线性估计进行了进一步研究, 但相关研究大多针对的是独立数据下的半参数回归模型估计^[5-6], 很少运用于纵向数据的分析中. 基于此, 本文尝试利用二次光滑局部线性估计研究纵向数据下部分线性模型的估计问题, 并对该方法的估计效果进行验证.

1 纵向数据下部分线性模型概述

纵向数据下部分线性模型有多种表达形式, 本文仅研究如下形式的纵向数据下部分线性模型:

$$\mathbf{Y}(t) = \mathbf{X}(t)^T \boldsymbol{\beta} + g(t) + \varepsilon(t), \quad (1)$$

其中 $\varepsilon(t)$ 是均值为 0 的随机过程. 假设观测 n 个个体, 第 i 个个体观测 m_i 次, $1 \leq i \leq n$. 记 t_{ij} ($1 \leq j \leq m_i$) 为第 i 个个体第 j 次观测时间, $Y_{ij} = Y_i(t_{ij})$ 为第 i 个个体在时间 t_{ij} 的响应变量的观测值, $X_i(t_{ij})$ 为第 i 个个体在时间 t_{ij} 的协变量的观测值.

2 纵向数据下部分线性模型的局部线性估计

假设 $\boldsymbol{\beta}$ 给定, 且令 $Y^*(t) = \mathbf{Y}(t) - \mathbf{X}(t)^T \boldsymbol{\beta}$, 则模型(1)可写成 $Y^*(t) = g(t) + \varepsilon(t)$. 由此, 半参数回归问题转化为非参数回归问题. 记未知函数 $g(\cdot)$ 在点 t_0 处的初始估计为 $\check{g}(t_0)$, 极小化式(2)即得 $\check{g}(t_0)$.

$$\sum_{i=1}^n \sum_{j=1}^{m_i} [Y_{ij}^* - a - b(t_{ij} - t_0)]^2 \omega(t_{ij}) K_h(t_{ij} - t_0). \quad (2)$$

记 $\mathbf{H}_0 = \begin{pmatrix} 1 & \cdots & 1 \\ h^{-1}(t_{11} - t_0) & \cdots & h^{-1}(t_{nm_n} - t_0) \end{pmatrix}$, $\mathbf{Y} = (\mathbf{Y}_1^T, \cdots, \mathbf{Y}_n^T)$, $\mathbf{Y}_i = (y_{i1}, \cdots, y_{im_i})^T$, $\mathbf{X} = (\mathbf{X}_1^T, \cdots, \mathbf{X}_n^T)$,

$\mathbf{X}_i = (X_i(t_{i1}), \cdots, X_i(t_{im_i}))^T$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^T, \cdots, \boldsymbol{\varepsilon}_n^T)$, $\boldsymbol{\varepsilon}_i = (\varepsilon_i(t_{i1}), \cdots, \varepsilon_i(t_{im_i}))^T$, $\mathbf{W} = \text{diag}(\omega(t_{11}), \cdots, \omega(t_{nm}))$, $\mathbf{W}_{t_0} = \text{diag}(\omega(t_{11})K_h(t_{11} - t_0), \cdots, \omega(t_{nm})K_h(t_{nm} - t_0))$. 其中 h 是带宽, $K(\cdot)$ 是核函数, $K_h(\cdot) = h^{-1}K(\cdot/h)$, $\omega(t_{ij})$ 是权函数. 由以上, 有

$$\check{g}(t_0) = (1, 0) [\mathbf{H}_{t_0}^T \mathbf{W}_{t_0} \mathbf{H}_{t_0}]^{-1} \mathbf{H}_{t_0}^T \mathbf{W}_{t_0} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \quad (3)$$

根据最小二乘理论可将模型(1)表示为 $(\mathbf{I} - \mathbf{F})\mathbf{Y} = (\mathbf{I} - \mathbf{F})\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, 其中 \mathbf{I} 是 n 阶单位矩阵, \mathbf{F} 是仅依赖于 t_{ij} 的光滑矩阵 (\mathbf{F} 可根据文献[7]指定). 对上式应用 Profile 最小二乘法, 可得到参数分量 $\boldsymbol{\beta}$ 的估计为

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}^T (\mathbf{I} - \mathbf{F})^T \mathbf{W} (\mathbf{I} - \mathbf{F}) \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{I} - \mathbf{F})^T \mathbf{W} \mathbf{Y}. \quad (4)$$

在式(3)中用 $\hat{\boldsymbol{\beta}}$ 代替 $\boldsymbol{\beta}$, 即可得到未知函数 $g(\cdot)$ 在点 t_0 处的最终估计:

$$\hat{g}(t_0) = (1, 0) [\mathbf{H}_{t_0}^T \mathbf{W}_{t_0} \mathbf{H}_{t_0}]^{-1} \mathbf{H}_{t_0}^T \mathbf{W}_{t_0} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \quad (5)$$

樊明智^[8] 在一系列的适当假设条件下, 得到了 $\hat{g}(t_0)$ 的渐近偏差和渐近方差:

$$\text{Bias}(\hat{g}(t_0)) = \frac{1}{2} \mu_2 g''(t_0) h^2 + o_p(h^2), \quad (6)$$

$$\text{Var}(\hat{g}(t_0)) = [nh\lambda(t_0)]^{-1} \sigma^2(t_0) \int K^2 du \{1 + o_p(1)\}. \quad (7)$$

3 纵向数据下部分线性模型的二次光滑估计

将局部拟合直线的截距值 $\hat{g}(t_0)$ 和未知函数的导数估计值 $\hat{g}^{(1)}(t_0)$ 相结合, 即可得到新的纵向数据下模型(1)在目标点 t_0 处的二次光滑局部线性估计的表达式^[9]:

$$\tilde{g}(t_0) = \int [\hat{g}(u) + \hat{g}^{(1)}(u)(t_0 - u)] L_{h'}(t_0 - u) du, \quad (8)$$

其中 h' 是第 2 步光滑的带宽, $L(\cdot)$ 是核函数, $L_{h'}(\cdot) = L(\cdot/h')h'$, $\omega(t_{ij})$ 是权函数. 在此, 本文取 $h' = h$, $L(\cdot) = K(\cdot)$, 这样不但能简化结论形式, 也能达到较好的估计效果. 由于二次光滑局部线性回归估计存在边界问题, 即在边界点处的渐近偏差不能像内点处的偏差那样减小, 因此需要对边界点处的渐近

性质进行单独考察. 设观测的时间区间为 $[0, T]$, 目标点 $t_0 \in [2h, T-2h]$, 则边界区间为 $[0, 2h)$ 和 $(T-2h, T]$. 再结合式(8) 可得到改进的二次光滑局部线性估计表达式:

$$\tilde{g}(t_0) = \int \left\{ \left(1, \frac{t_0 - u}{h} \right) [H_u^T W_u H_u]^{-1} H_u^T W_u [Y - X\hat{\beta}] \right\} K_h(t_0 - u) du. \quad (9)$$

4 二次光滑局部线性估计的渐近性质

首先给出渐近性质证明中常用的正则条件^[10]:

(C1) 核函数 $K(\cdot)$ 为具有紧支撑且有界的概率密度函数. 为简化计算, 在此假设 $K(\cdot)$ 具有对称性, 即 $K(-x) = K(x)$.

(C2) $\int_0^T \lambda(u) du < \infty$, 对于给定的时间 t , $\{N_i(t)\} (i=1, 2, \dots, n)$ 中的跳跃点是独立的, 且无相同的跳跃点.

(C3) 当 $n \rightarrow \infty$ 时, $nh^8 \rightarrow 0$, 且 $nh^2/(\lg n)^2 \rightarrow \infty$.

(C4) $g(\cdot)$ 在内点处存在有界的四阶连续导数.

(C5) 对于 $\forall t$, $\lambda(t)$ 是二阶连续可微函数, $X(t)$ 是连续函数.

(C6) 对于 $\forall t$, $\sigma^2(t) = \text{Var}\{\varepsilon(t)\}$ 有限, 且二次连续、可微.

(C7) $E \left[\int_0^\infty \{X(t) - EX(t)\}^{\otimes 2} dN(t) \right]$ 非奇异.

为了简化后续的证明, 令 $K_1(v) = \int K(v-u)K(u)du$, $K_2(v) = \int (v-u)uK(v-u)K(u)du$, $V(K) = \int [K_1(v) - K_2(v)/\mu_2]^2 dv$. 下面给出二次光滑局部线性估计的渐近偏差和渐近方差.

引理 1 在条件(C1)–(C3) 下, 有:

$$S_n(t_0) = n\lambda(t_0)\omega(t_0) \otimes \Gamma[1 + o_p(c_n)], \quad c_n = h^2 + \left(\frac{\lg n}{nh}\right)^{\frac{1}{2}}. \quad (10)$$

证明 由矩阵计算可得 $S_n(t_0) = \begin{pmatrix} S_{n,0} & S_{n,1} \\ S_{n,1} & S_{n,2} \end{pmatrix}$, 其中

$$S_{n,l} = \sum_{i=1}^n \sum_{j=1}^{m_i} \omega(t_{ij}) K_h(t_{ij} - t_0) \left(\frac{t_{ij} - t_0}{h}\right)^l = \sum_{i=1}^n \int_0^{+\infty} \omega(t) K_h(t - t_0) \left(\frac{t - t_0}{h}\right)^l dN_i(t).$$

当 $h \rightarrow 0$, $nh \rightarrow \infty$ 时, 有

$$E(S_{n,l}) = nE \left[\int_0^{+\infty} \omega(t) K_h(t - t_0) \left(\frac{t - t_0}{h}\right)^l \lambda(t) dt \right] = n\lambda(t_0)\omega(t_0) \int_{-\infty}^{+\infty} \mu^l K(u) d\mu (1 + o_p(c_n)).$$

将 $S_{n,l} = E(S_{n,l}) + o_p(\sqrt{\text{Var}(S_{n,l})})$ 代入上式即可证得式(10) 成立.

引理 2 在条件(C1)–(C4) 下, 有:

$$H_{t_0}^T W_{t_0} X = n\lambda(t_0)\omega(t_0) E[X^T(t) | t=t_0] \otimes (1, 0)^T (1 + o_p(c_n)),$$

$$H_{t_0}^T W_{t_0} G = n\lambda(t_0)\omega(t_0) g(t_0) \otimes (1, 0)^T (1 + o_p(c_n)),$$

$$H_{t_0}^T W_{t_0} \varepsilon = (1, 0)^T o_p(1).$$

证明 由于引理 2 中的 3 个式子的证明类似, 因此在此只给出第 1 式的证明. 由矩阵计算有

$$H_{t_0}^T W_{t_0} X = \begin{pmatrix} \sum_{i=1}^n \sum_{j=1}^{m_i} \omega(t_{ij}) K_h(t_{ij} - t_0) X_i^T(t_{ij}) \\ \sum_{i=1}^n \sum_{j=1}^{m_i} \omega(t_{ij}) K_h(t_{ij} - t_0) X_i^T(t_{ij}) \left(\frac{t_{ij} - t_0}{h}\right) \end{pmatrix} =$$

$$\left(\begin{array}{c} \sum_{i=1}^n \int_0^{+\infty} \omega(t) K_h(t-t_0) \mathbf{X}_i^T(t) dN_i(t) \\ \sum_{i=1}^n \int_0^{+\infty} \omega(t) K_h(t-t_0) \mathbf{X}_i^T(t) \left(\frac{t-t_0}{h} \right) dN_i(t) \end{array} \right) = \left(\begin{array}{c} n\lambda(t_0)\omega(t_0)E[\mathbf{X}^T(t) | t=t_0]\mu_0(1+o_p(c_n)) \\ n\lambda(t_0)\omega(t_0)E[\mathbf{X}^T(t) | t=t_0]\mu_1(1+o_p(c_n)) \end{array} \right).$$

又因为 $K(\cdot)$ 是对称核密度函数, 故有 $\mu_0 = 1, \mu_1 = 0$. 将 $\mu_0 = 1, \mu_1 = 0$ 代入上式, 第 1 式即可得证.

由于本文考察的是未知函数分量的估计效果, 因此只给出引理 3 和引理 4, 不给出其证明.

引理 3 在条件(C1)–(C5) 下, 有 $\frac{1}{n} \mathbf{X}^T (\mathbf{I} - \mathbf{F})^T \mathbf{W} (\mathbf{I} - \mathbf{F}) \mathbf{X} \xrightarrow{p} \mathbf{A}$.

引理 4 在条件(C1)–(C5) 下, 有 $\frac{1}{n} \mathbf{X}^T (\mathbf{I} - \mathbf{F})^T \mathbf{W} (\mathbf{I} - \mathbf{F}) \mathbf{G} = o_p(c_n^2)$.

由引理 3 和引理 4 可得到参数分量估计值 $\hat{\boldsymbol{\beta}}$ 的渐近正态性, 即有 $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{L} N(0, \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1})$.

定理 1 在条件(C1)–(C7) 下, 可得 $\hat{g}(t_0)$ 的渐近偏差为

$$\text{Bias}(\hat{g}(t_0)) = \frac{1}{4} g^{(4)}(t_0) (\mu_2^2 - \mu_4) h^4 + o_p(h^4). \quad (11)$$

证明 由上述引理有 $E(\hat{g}(t_0)) = (1, 0)^{-1} S_n^{-1}(t_0) \mathbf{H}_{t_0}^T \mathbf{W}_{t_0} [\mathbf{G} + \mathbf{X} E(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})]$. 再由 $E(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = o_p(1)$ 有

$$E(\hat{g}(t_0)) = (1, 0) S_n^{-1}(t_0) \mathbf{H}_{t_0}^T \mathbf{W}_{t_0} \mathbf{G} + o_p(1). \quad (12)$$

对向量 \mathbf{G} 的每一分量运用 Taylor 公式^[11], 有

$$\mathbf{G} = \begin{pmatrix} \sum_{l=0}^4 \frac{g^{(l)}(t_0)}{l!} (t_{11} - t_0)^l + o_p[(t_{11} - t_0)^l] \\ \vdots \\ \sum_{l=0}^4 \frac{g^{(l)}(t_0)}{l!} (t_{1m_1} - t_0)^l + o_p[(t_{1m_1} - t_0)^l] \\ \vdots \\ \sum_{l=0}^4 \frac{g^{(l)}(t_0)}{l!} (t_{nm_n} - t_0)^l + o_p[(t_{nm_n} - t_0)^l] \end{pmatrix},$$

其中 $g^{(0)}(t_0) = g(t_0)$, 进而有

$$\begin{aligned} \mathbf{H}_{t_0}^T \mathbf{W}_{t_0} \mathbf{G} &= \left(\begin{array}{c} \sum_{i=1}^n \sum_{j=1}^{m_i} \omega(t_{ij}) K_h(t_{ij} - t_0) \left\{ \sum_{l=0}^4 \frac{g^{(l)}(t_0)}{l!} (t_{ij} - t_0)^l + o_p[(t_{ij} - t_0)^l] \right\} \\ \vdots \\ \sum_{i=1}^n \sum_{j=1}^{m_i} \omega(t_{ij}) K_h(t_{ij} - t_0) \left(\frac{t_{ij} - t_0}{h} \right) \left\{ \sum_{l=0}^4 \frac{g^{(l)}(t_0)}{l!} (t_{ij} - t_0)^l + o_p[(t_{ij} - t_0)^l] \right\} \end{array} \right) = \\ &= \left(\begin{array}{c} \sum_{i=1}^n \int_0^{+\infty} \omega(t) K_h(t - t_0) \left\{ \sum_{l=0}^4 \frac{g^{(l)}(t_0)}{l!} (t - t_0)^l + o_p[(t - t_0)^l] \right\} dN_i(t) \\ \vdots \\ \sum_{i=1}^n \int_0^{+\infty} \omega(t) K_h(t - t_0) \left(\frac{t - t_0}{h} \right) \left\{ \sum_{l=0}^4 \frac{g^{(l)}(t_0)}{l!} (t - t_0)^l + o_p[(t - t_0)^l] \right\} dN_i(t) \end{array} \right) = \\ &= \left(\begin{array}{c} n\lambda(t_0)\omega(t_0) \sum_{l=0}^4 \frac{g^{(l)}(t_0)h^l}{l!} \mu_l [1 + o_p(1)] \\ \vdots \\ n\lambda(t_0)\omega(t_0) \sum_{l=0}^4 \frac{g^{(l)}(t_0)h^l}{l!} \mu_{l+1} [1 + o_p(1)] \end{array} \right). \end{aligned}$$

又因为 $K(\cdot)$ 是对称核密度函数, 故有 $\mu_0 = 1, \mu_1 = \mu_3 = \mu_5 = 0$. 则上式可化简为

$$\mathbf{H}_{t_0}^T \mathbf{W}_{t_0} \mathbf{G} = \left(\begin{array}{c} n\lambda(t_0)\omega(t_0) \left[g(t_0) + \frac{g^{(2)}(t_0)h^2}{2!} \mu_2 + \frac{g^{(4)}(t_0)h^4}{4!} \mu_4 \right] [1 + o_p(1)] \\ n\lambda(t_0)\omega(t_0) \left[g^{(1)}(t_0)h\mu_2 + \frac{g^{(3)}(t_0)h^3}{3!} \mu_4 + \frac{g^{(5)}(t_0)h^5}{5!} \mu_6 \right] [1 + o_p(1)] \end{array} \right). \quad (13)$$

对式(10) 进行计算可得

$$S_n^{-1}(t_0) = [n\lambda(t_0)\omega(t_0)]^{-1} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\mu_2} \end{pmatrix} [1 + o_p(c_n)]. \quad (14)$$

将式(13) 和(14) 代入式(12) 得

$$E(\tilde{g}(t_0)) = \left[g(t_0) + \frac{g^{(2)}(t_0)h^2}{2!}\mu_2 + \frac{g^{(4)}(t_0)h^4}{4!}\mu_4 \right] [1 + o_p(1)],$$

进而有

$$\text{Bias}(\tilde{g}(t_0)) = E(\tilde{g}(t_0)) - g(t_0) = \frac{g^{(2)}(t_0)h^2}{2!}\mu_2 + \frac{g^{(4)}(t_0)h^4}{4!}\mu_4 + o_p(h^4). \quad (15)$$

显然式(15) 的结果比式(6) 更为精确.

未知函数一阶导数的估计 $\hat{g}^{(1)}(t_0)$ 的表达式^[12] 为

$$\hat{g}^{(1)}(t_0) = (0, \frac{1}{h}) [\mathbf{H}_{t_0}^T \mathbf{W}_{t_0} \mathbf{H}_{t_0}]^{-1} \mathbf{H}_{t_0}^T \mathbf{W}_{t_0} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

类似于 $\text{Bias}(\tilde{g}(t_0))$ 的计算步骤, 对上式计算可得

$$E(\hat{g}^{(1)}(t_0)) = (0, \frac{1}{h}) S_n^{(-1)}(t_0) \mathbf{H}_{t_0}^T \mathbf{W}_{t_0} \mathbf{G} + o_p(1) = \left[g^{(1)}(t_0) + \frac{g^{(3)}(t_0)h^2}{3! \mu_2} \mu_4 + \frac{g^{(5)}(t_0)h^4}{5! \mu_2} \mu_6 \right] (1 + o_p(1)).$$

于是有

$$\text{Bias}(\hat{g}^{(1)}(t_0)) = \frac{g^{(3)}(t_0)h^2}{3! \mu_2} \mu_4 + \frac{g^{(5)}(t_0)h^4}{5! \mu_2} \mu_6 + o_p(h^4). \quad (16)$$

根据式(15) 和(16), 再结合 Taylor 公式, 二次光滑局部线性估计 $\tilde{g}(t_0)$ 的渐近偏差可写成

$$\begin{aligned} \text{Bias}(\tilde{g}(t_0)) &= \int E[\hat{g}(u) - g(u)] K_h(t_0 - u) du + \\ &\int E[\hat{g}^{(1)}(u) - g^{(1)}(u)] (t_0 - u) K_h(t_0 - u) du - \\ &\int \sum_{l=2}^4 \frac{g^{(l)}(u) (t_0 - u)^l}{l!} K_h(t_0 - u) du + o_p(h^4). \end{aligned} \quad (17)$$

上式的第 1 部分即为

$$\begin{aligned} \int \text{Bias}(\hat{g}(u)) K_h(t_0 - u) du &\stackrel{\frac{t_0 - u}{h} = \omega}{=} \int \text{Bias}(\hat{g}(u)) K(\omega) d\omega = \\ &\int \left\{ \frac{1}{2} \mu_2 h^2 [g^{(2)}(t_0) - \omega h g^{(3)}(t_0) + \frac{1}{2} \omega^2 h^2 g^{(4)}(t_0) + o_p(h^2)] + \right. \\ &\left. \frac{1}{4!} \mu_4 h^4 [g^{(4)}(t_0) - \omega h g^{(5)}(t_0) + o_p(h)] + o_p(h^4) \right\} K(\omega) d\omega = \\ &\frac{1}{2} \mu_2 h^2 g^{(2)}(t_0) + \frac{1}{4} \mu_2^2 h^4 g^{(4)}(t_0) + \frac{1}{4!} \mu_4 h^4 g^{(4)}(t_0) + o_p(h^4). \end{aligned}$$

通过类似方法, 可得式(17) 的第 2 部分和第 3 部分的结果, 分别为:

$$\begin{aligned} \int \text{Bias}(\hat{g}(u)) (t_0 - u) K_h(t_0 - u) du &= -\frac{1}{3!} \mu_4 h^4 g^{(4)}(t_0) + o_p(h^4), \\ -\int \sum_{l=2}^4 \frac{g^{(l)}(t_0 - \omega h) (\omega h)^l}{l!} K(\omega) d\omega &+ o_p(h^4) = \\ -\left[\frac{1}{2} \mu_2 h^2 g^{(2)}(t_0) + \frac{1}{4} \mu_4 h^4 g^{(4)}(t_0) - \frac{1}{3!} \mu_4 h^4 g^{(4)}(t_0) + \frac{1}{4!} \mu_4 h^4 g^{(4)}(t_0) \right] &+ o_p(h^4). \end{aligned}$$

将这 3 部分的结果相加即可得证式(11).

定理 2 在条件(C1)–(C7) 下, $\widetilde{g}(t_0)$ 的渐近方差为

$$\text{Var}(\widetilde{g}(t_0)) = \frac{1}{nh\lambda(t_0)}\sigma^2(t_0)V(K)[1 + o_p(1)].$$

(18)

证明 将式(9) 中的 u 换成 t , 有 $\widetilde{g}(t_0) = \int \left\{ \left(1, \frac{t_0 - t}{h} \right) [\mathbf{H}_t^T \mathbf{W}_t \mathbf{H}_t]^{-1} \mathbf{H}_t^T \mathbf{W}_t [\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}] \right\} K_h(t_0 - t) dt$.

令 $\mathbf{D} = \int \left(1, \frac{t_0 - t}{h} \right) [\mathbf{H}_t^T \mathbf{W}_t \mathbf{H}_t]^{-1} \mathbf{H}_t^T \mathbf{W}_t K_h(t_0 - t) dt$, $u = \frac{t - t_0}{h}$, $v = \frac{s - t_0}{h}$, $v_{ij} = \frac{t_{ij} - t_0}{h}$, 则类似于定理 1 的计算可得

$$\begin{aligned} \text{Var}(\widetilde{g}(t_0)) &= \mathbf{D} \cdot \text{Var}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \cdot \mathbf{D}^T = \mathbf{D} \cdot \text{diag}(\sigma^2(t_{11}), \cdots, \sigma^2(t_{nm_n})) \cdot \mathbf{D}^T = \\ &= \frac{1}{[nh\lambda(t_0)]^2} \sum_{i=1}^n \sum_{j=1}^{m_i} \sigma^2(t_{ij}) \left\{ \int [K(v_{ij} - u) - \frac{1}{\mu_2}(v_{ij} - u)K(v_{ij} - u)]K(u)du \right\}^2 = \\ &= \frac{nh\lambda(t_0)\sigma^2(t_0)}{[nh\lambda(t_0)]^2} \int \left\{ \int [K(v - u)K(u)du - \frac{1}{\mu_2} \int (v - u)K(v - u)uK(u)du] \right\}^2 dv [1 + o_p(1)] = \\ &= \frac{1}{nh\lambda(t_0)}\sigma^2(t_0)V(K)[1 + o_p(1)]. \end{aligned}$$

5 实例分析

实例分析的数据集(纵向数据)来自国际艾滋病研究中心记录的人体 CD4 细胞数的数据库, 本文选取其中 150 个患者的检查结果. 为了描述 CD4 细胞数损耗的平均时间趋势, 建立如下纵向数据下部分线性模型:

$$\mathbf{Y}(t) = \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + g(t) + \varepsilon(t),$$

其中 \mathbf{Y} 表示患者感染 HIV 之后在时刻 t 标准化后的 CD4 细胞浓度, \mathbf{X}_1 表示是否抽烟和年龄对患者感染 HIV 的交互效应, \mathbf{X}_2 表示患者感染之前标准化后的 CD4 细胞浓度, $g(t)$ 表示平均 CD4 细胞浓度随时间的变化趋势. 利用光滑交叉验证方法^[13] 求得模型的最优带宽 $h = 0.5912$, 核函数 $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2})$.

采用一步差分估计法求得参数分量的估计值为 $(\hat{\beta}_1, \hat{\beta}_2)^T = (-0.0296, 0.5408)^T$. 在区间 $[0, 6]$ 上取 100 个格子点, 采用局部线性估计方法和二次光滑局部线性估计方法对未知非参数分量 $g(t)$ 作拟合, 结果如图 1 所示.

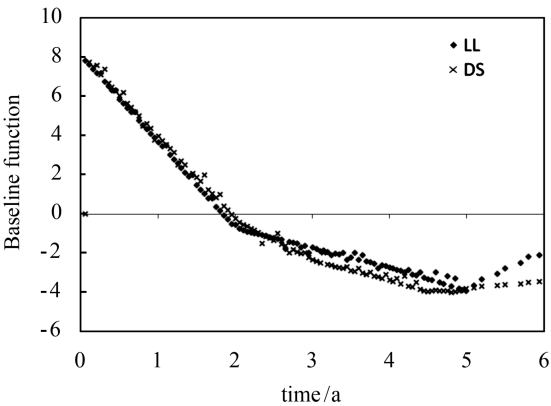


图 1 非参数分量 $g(t)$ 的估计曲线(LL 为局部线性估计, DS 为二次光滑估计)

由图 1 中 2 种估计曲线的走势可以看出, 患者在感染 HIV 的初期, 其平均 CD4 细胞浓度 $g(t)$ 下降

得很快,但在3年后下降趋势减缓.这两种方法的估计结果虽然在趋势上接近,但因二次光滑局部线性估计是在局部线性估计的基础上再次进行了光滑平均,所以其拟合曲线更为平滑.这表明二次光滑估计的整体效果优于局部线性估计,同时也证实了二次光滑估计可降低渐近偏差.

参考文献:

- [1] ZEGER S L, DIGGLE P J. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters[J]. *Biometrics*, 1994,50:689-699.
- [2] FAN J. Local linear regression smoothers and their minimax efficiency[J]. *The Annals of Statistics*, 1993,21:196-216.
- [3] LIN X H, CARROL R J. Semiparametric regression for clustered data using generalized estimating equations[J]. *Journal of the American Statistical Association*, 2001,96:1045-1056.
- [4] HE H, HUANG L S. Double-smoothing for bias reduction in local linear regression[J]. *Journal of Statistical Planning and Inference*, 2009,139(3):1056-1072.
- [5] HE H, TANG W. Statistical inference in the partial linear models with the double smoothing local linear regression method[J]. *Journal of Statistical Planning and Inference*, 2014,146:102-112.
- [6] HWANG R C, KEN M L. Double smoothing estimation of the multivariate regression function in nonparametric regression[J]. *Comm Statist Theory Methods*, 2002,31(3):419-434.
- [7] HE X M, ZHU Z Y. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure[J]. *Biometrika*, 2002,89:579-590.
- [8] 樊明智. 纵向数据下部分线性模型的渐近性质[J]. *数理统计与管理*, 2012,31(3):447-454.
- [9] HE H, TANG W, ZUO G X. Statistical inference in the partial linear models with the double smoothing local linear regression method[J]. *Journal of Statistical Planning and Inference*, 2014,146:102-112.
- [10] FAN J Q, LI R Z. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis[J]. *Journal of the American Statistical Association*, 2004,99(467):710-723.
- [11] 薛留根. 现代统计模型[M]. 北京:科学出版社,2012:24-32.
- [12] 罗美华,李元,周勇,等. 基于纵向数据的半参数变系数部分线性回归模型[J]. *应用数学学报*, 2007,30(3):540-554.
- [13] HALL P, MARRON J S. Smoothed cross validation Probab[J]. *Theory Related Field*, 1992,90:149-173.