带终止事件纵向数据半参数部分线性单指标模型

周洁¹, 周晓华^{2*}, 孙六全^{3,4}

1. 首都师范大学数学科学学院, 北京 100048;

2. 北京大学北京国际数学研究中心, 北京 100871;

3. 中国科学院数学与系统科学研究院, 北京 100190;

4. 中国科学院大学数学科学学院, 北京 100049

E-mail: zhoujie@amss.ac.cn, azhou@uw.edu, slq@amt.ac.cn

收稿日期: 2018-02-03; 接受日期: 2018-05-15; 网络出版日期: 2018-12-10; * 通信作者

国家自然科学基金 (批准号: 11671275, 11471223, 11301355, 11690015 和 11771431)、科技创新服务能力建设 - 基本科研业务费 (科研类) (批准号: 025185305000/204)、首都师范大学青年科研创新团队和 NIH/NIA (批准号: U01AG016976) 资助项目

摘要 在纵向研究中, 我们常常会对一些非终止事件感兴趣, 例如, 与健康相关的生存质量. 然而, 死亡事件可能会在我们观测到这些感兴趣的变量之前发生, 这种情况称之为截断数据. 此外, 每个个体的观察时间可能是不规律的, 而且协变量对于感兴趣事件的影响也可能很复杂. 本文提出了一个新的用于拟合被死亡截断的数据的半参数部分线性单指标模型, 并用局部线性核方法近似非参数的连接函数, 构造了参数部分的估计方程. 本文证明了连接函数的局部线性核估计具有一致相合性, 但收敛速度低于 $n^{1/2}$. 为了消除这一低速收敛的影响, 本文提出了一种数据分割的方法, 证明了参数部分估计量的渐近正态性. 本文通过一些数值模拟检验所提方法在有限样本下的表现, 并分析了一组老年痴呆症的数据.

关键词 纵向数据 部分线性单指标模型 终止事件

MSC (2010) 主题分类 62G05, 62N02

1 引言

纵向数据常见于对个体进行持续跟踪观测研究中, 已有各种各样的方法用于分析这种数据 (参见文献 [1–6]). Diggle 等^[7] 及 Lin 和 Ying^[8] 对常用的方法做了非常好的总结, 包括很多常用的参数、半参数和非参数回归模型, 然而, 所有这些模型都没有包含死亡等终止事件. 在很多应用场合, 终止事件是客观存在的, 如死亡, 它的发生将终止对个体的继续观察. 例如, 患者重复就诊这一事件, 在患者死亡后将不复存在.

近年来, 很多研究者提出了一些带有终止事件纵向数据的联合建模方法^[9–16]. 在大多数文献中, 终止事件被当作带信息的删失事件来处理, 然而, 将它们看作阻止纵向响应变量后续存在性的截断数

英文引用格式: Zhou J, Zhou X H, Sun L Q. A partially linear single index model for longitudinal data with terminal event (in Chinese). Sci Sin Math, 2019, 49: 73–88, doi: 10.1360/SCM-2018-0091

据可能更合适. 当有截断数据存在时, 条件建模方法比较通行, 如果关注点是治疗效果的因果推断, 则往往需要用到分析潜在观测结果的相关方法^[12, 17–19]; 如果关注点只是风险因素与死亡的相互关系, 通常会对生存群体建立模型^[16]. Wen 等^[20]对被死亡截断的纵向数据的分析方法做了一个非常好的总结, 绝大多数方法都简单地假定响应变量服从可加模型或者可乘模型, 这在实际应用中有一定的局限性, 我们需要一个更灵活的模型. 部分线性单指标模型通过单指标部分来控制不感兴趣的协变量的效应, 而把感兴趣的协变量放在线性部分使其具有易解释性^[21–24]. 对于不含有终止事件的纵向数据, Li 等^[25]基于经验似然方法考虑了一个部分线性单指标模型, 但在他们的模型中, 时间只能被当作一个协变量作用于线性部分或者单指标部分, 这很可能与实际不相符, 一种更加自然的方式是设定一个非参数的基准时间效应函数^[6, 8]. 本文将研究一个用于分析带有终止事件的纵向数据的一个含有非参数基准时间效应的半参数部分线性单指标模型. 该模型非常灵活, 感兴趣的协变量可以具有易解释的线性效应, 而其他混杂因素则一起被包含在单指标部分共同对响应变量起作用. 此外, 随着时间变化的基准效应也没有作参数化假定.

本文余下部分的内容如下: 第 2 节介绍一些记号以及我们所提出的模型; 第 3 节给出计算估计量的一个迭代算法, 并推导了所得估计量的大样本性质; 第 4 节提供了一些数值模拟结果; 第 5 节将所提出的方法应用于分析老年痴呆症的数据; 第 6 节对我们的方法进行了总结.

2 模型建立

考虑一个样本量为 n 的纵向数据研究, 对于个体 i 和时刻 t , 令 $Y_i(t)$ 表示感兴趣的纵向响应过程, $X_i(t)$ 和 $Z_i(t)$ 分别为 $p \times 1$ 维和 $q \times 1$ 维的具有不同效应的协变量向量. 进一步地, 令 D_i 和 C_i 分别表示终止事件的发生时间和删失时间, 并记 $T_i = D_i \wedge C_i$ 为真实观测到的时间, 其中 $D_i \wedge C_i = \min(D_i, C_i)$. 令 $\delta_i = I(D_i \leq C_i)$, 其中 $I(\cdot)$ 为示性函数. 以 $N_i^*(t)$ 表示 t 时刻前的观测次数所对应的计数过程, 注意到终止事件将终止观测过程和响应过程, 即有 $N_i^*(t) = N_i^*(t \wedge D_i)$, 且 $Y_i(t)$ 在 $t > D_i$ 时不再有定义. 事实上, 观测过程 $N_i^*(t)$ 通常会被 C_i 所删失, 从而仅有 $N_i(t) = N_i^*(t \wedge C_i)$ 被观测到, 而纵向响应过程 $Y_i(t)$ 也仅仅是在 $N_i(t)$ 有跳跃的那些时刻被观测到, 其中 $t \leq T_i$. 对于一组容量为 n 的随机样本, 我们观测到的数据为

$$\{Y_i(t)dN_i(t), T_i, \delta_i, X_i(t), Z_i(t), N_i(t), 0 \leq t \leq T_i, i = 1, \dots, n\}.$$

对于纵向响应过程, 考虑生存条件下的半参数部分线性单指标模型:

$$E\{Y_i(t) \mid X_i(t), Z_i(t), D_i \geq t\} = \alpha_0(t) + g(\beta_0^T X_i(t)) + \gamma_0^T Z_i(t), \quad (2.1)$$

其中 α_0 和 g 均为未知函数, β_0 和 γ_0 为未知参数. 为了参数的可识别性, 假定 $\|\beta_0\| = 1$, $g(v_0) = 0$, 其中 v_0 为函数 g 的定义域中任意点. 此外, 我们还假定函数 g 二阶可导. 该模型中的协变量, 有的具有线性效应, 有的具有非线性效应, 具有很大的灵活性. 通常, 包含在线性部分的协变量 $Z_i(t)$ 是我们主要感兴趣的因素, 其对应系数 γ_0 可看作一种加性效应, 而包含在单指标部分的协变量 $X_i(t)$ 则不是我们感兴趣的因素, 但也不是可忽略因素. 如果使用非参数模型来刻画单指标部分协变量的效应, 我们将面临著名的维数祸根问题, 因而, 单指标模型实际上是一种降维的做法. 事实上, 单指标建模可以看作是将协变量向方向 β_0 的投影然后再对投影结果建立非参数模型^[21]. 近年来, Li 等^[25]研究了一个不带终止事件的部分线性单指标模型, 但他们没有对随时间变化的基准效应做特殊对待, 然而, 在纵向

数据研究中, 对随时间变化的基准效应不做参数化假定会更加合理, 我们通过引入非参数项 $\alpha_0(t)$ 使得该效应更加广泛和稳健. 接下来的研究中, 我们感兴趣的主要是对参数 γ_0 的估计和推断.

对于观测过程, 假定生存条件下的比例比率模型为

$$E\{dN_i^*(t) \mid X_i(t), Z_i(t), D \geq t\} = \exp\{\eta_0^T W_i(t)\} d\Lambda_0(t), \quad (2.2)$$

其中 $W_i(t)$ 由 $X_i(t)$ 和 $Z_i(t)$ 的部分组成, η_0 是一个未知的回归参数, $\Lambda_0(t)$ 是一个未知的基准累加比率函数. 该比例比率模型已被广泛应用于不带终止事件的复发事件的建模分析中^[26]. 当有终止事件存在时, 我们选择对生存条件下的个体进行建模^[27]. 对于联合建模, 假定在给定协变量 $X_i(\cdot)$ 和 $Z_i(\cdot)$ 以及个体生存的条件下, 响应过程 $Y_i(\cdot)$ 与观测过程 $N_i^*(\cdot)$ 相互独立, 即生存个体的响应变量取值与其观测过程的联系仅通过协变量的作用来间接实现. 特别地, 当观测时间是固定的或者仅仅依赖于个体的协变量特征时, 这一假定是满足的. 此外, 我们还假定, 在给定协变量 $X_i(\cdot)$ 和 $Z_i(\cdot)$ 条件下, 删失时间 C_i 与终止事件 D_i 、观测时间 $N_i(\cdot)$ 和响应过程 $Y_i(\cdot)$ 皆独立.

3 参数估计

首先, 考虑观测过程模型 (2.2) 相关参数的估计, 给定协变量 $X_i(\cdot)$ 和 $Z_i(\cdot)$, 在观测过程 $N_i^*(\cdot)$ 与删失时间 C_i 相互独立的假定条件下, 我们有

$$\begin{aligned} E_{XZ}\{dN_i(t)\} &= E_{XZ}\{I(D_i \geq t)I(C_i \geq t)dN_i^*(t)\} \\ &= E_{XZ}\{I(T_i \geq t)\}E_{XZ}\{dN_i^*(t) \mid D_i \geq t\} \\ &= E_{XZ}\{I(T_i \geq t) \exp\{\eta_0^T W_i(t)\}\}d\Lambda_0(t), \end{aligned} \quad (3.1)$$

其中 $E_{XZ}(\cdot) = E(\cdot \mid X, Z)$ 表示给定协变量下的条件期望. 从而, 在给定协变量 $X_i(\cdot)$ 和 $Z_i(\cdot)$ 的条件下, $dN_i(t) - I(T_i \geq t) \exp\{\eta_0^T W_i(t)\}d\Lambda_0(t)$ 的均值为 0, 因而, 我们无需使用逆概率加权即可构造估计方程. Λ_0 和 η_0 的估计量可由如下估计方程得到:

$$\begin{aligned} \sum_{i=1}^n [dN_i(t) - I(T_i \geq t) \exp\{\eta_0^T W_i(t)\}d\Lambda_0(t)] &= 0, \quad 0 \leq t \leq \tau, \\ \sum_{i=1}^n \int_0^\tau W_i(t) \{dN_i(t) - I(T_i \geq t) \exp\{\eta_0^T W_i(t)\}d\Lambda_0(t)\} &= 0. \end{aligned}$$

简单推导可知, η_0 可由如下估计方程的解 $\hat{\eta}$ 估计:

$$\sum_{i=1}^n \int_0^\tau \{W_i(t) - \bar{W}(t; \eta)\} dN_i(t) = 0,$$

其中

$$\bar{W}(t; \eta) = \frac{\sum_{i=1}^n I(T_i \geq t) \exp\{\eta^T W_i(t)\} W_i(t)}{\sum_{i=1}^n I(T_i \geq t) \exp\{\eta^T W_i(t)\}}.$$

注意到我们考虑的是条件建模, 所以无需使用逆概率加权方法, 因而所得到的估计方程跟文献 [26] 是非常相似的.

为了估计纵向响应过程的模型 (2.1), 类似于 (3.1) 的推导, 我们可以证明 $M_i(t)$ 也是一个零均值过程, 其中

$$dM_i(t) = \{Y_i(t) - g(\beta_0^T X_i(t)) - \gamma_0^T Z_i(t)\} dN_i(t) - I(T_i \geq t) \exp\{\eta_0^T W_i(t)\} \alpha_0(t) d\Lambda_0(t).$$

我们可以基于此零均值过程构造估计方程. 由于 $g(\cdot)$ 是未知函数, 我们使用局部线性核平滑方法, 对函数 g 定义域中任意给定的点 v , 当 u 在 v 的邻域时, 由 Taylor 展开得到近似表达

$$g(u) \approx g(v) + \dot{g}(v)(u - v),$$

其中 $\dot{g}$ 是 g 的导数. 将 η_0 用其估计量 $\hat{\eta}$ 替代, 定义

$$M_i(s; v, \beta, \gamma, \dot{g}, \mathcal{A}, h) = \int_0^s K_h(\beta^T X_i(t) - v) [\{Y_i(t) - \dot{g}(v)(\beta^T X_i(t) - v) - \gamma^T Z_i(t)\} dN_i(t) - I(T_i \geq t) \exp\{\hat{\eta}^T W_i(t)\} d\mathcal{A}(t; v)],$$

其中

$$\mathcal{A}_0(t; v) \equiv \int_0^t \{\alpha_0(s) + g(v)\} d\Lambda_0(s), \quad K_h(x) = h^{-1} K\left(\frac{x}{h}\right),$$

而 K 和 h 则分别为核函数和窗宽.

给定 β 和 γ 时, 对函数 g 定义域中的任意点 v , $\mathcal{A}_0(t; v)$ 和 $\dot{g}(v)$ 可通过求解如下估计方程得到估计量:

$$\sum_{i=1}^n M_i(t; v, \beta, \gamma, \dot{g}, \mathcal{A}, h) = 0, \quad 0 < t < \tau, \quad (3.2)$$

$$\sum_{i=1}^n \int_0^\tau (\beta^T X_i(t) - v) dM_i(t; v, \beta, \gamma, \dot{g}, \mathcal{A}, h) = 0. \quad (3.3)$$

简单计算可得估计量具有如下显式表达:

$$\begin{aligned} \hat{g}(v; \beta, \gamma, h) &= \left[\sum_{i=1}^n \int_0^\tau \beta^T \{X_i(t) - \bar{X}(t; v, h)\} K_h(\beta^T X_i(t) - v) (\beta^T X_i(t) - v) dN_i(t) \right]^{-1} \\ &\quad \times \sum_{i=1}^n \int_0^\tau \beta^T \{X_i(t) - \bar{X}(t; v, h)\} K_h(\beta^T X_i(t) - v) \{Y_i(t) - \gamma^T Z_i(t)\} dN_i(t), \\ \hat{\mathcal{A}}(t; v, \beta, \gamma, h) &= \int_0^t S_x^{(0)}(t; v, h)^{-1} n^{-1} \sum_{i=1}^n K_h(\beta^T X_i(t) - v) \\ &\quad \times \{Y_i(t) - \hat{g}(v; \beta, \gamma, h)(\beta^T X_i(t) - v) - \gamma^T Z_i(t)\} dN_i(t), \end{aligned}$$

其中

$$\bar{X}(t; v, h) = \frac{\sum_{i=1}^n I(T_i \geq t) K_h(\beta^T X_i(t) - v) \exp\{\hat{\eta}^T W_i(t)\} X_i(t)}{\sum_{i=1}^n I(T_i \geq t) K_h(\beta^T X_i(t) - v) \exp\{\hat{\eta}^T W_i(t)\}}.$$

为了估计纵向响应过程模型 (2.1) 中的参数, 我们先对 β 进行重新参数化, 令

$$\beta_\phi = \left(\left(1 - \sum_{k=1}^{p-1} \phi_k^2 \right)^{1/2}, \phi_1, \dots, \phi_{p-1} \right)^T,$$

从而, β_ϕ 自动满足模长为 1 的要求. 易知 β_ϕ 关于 ϕ 求导的 Jacobi 矩阵为

$$J_\phi = \begin{pmatrix} -(1 - \sum_{k=1}^{p-1} \phi_k^2)^{-1/2} \phi^T \\ \mathbf{I}_{p-1} \end{pmatrix}.$$

由于核估计不是 $\sqrt{n}$ 相合的, 为了消除其在参数部分的估计中的影响, 从而得到 $\sqrt{n}$ 相合的参数估计, 我们借鉴文献中关于简单数据的单指标模型的渐近理论结果 (参见文献 [22, 23]), 直接引入一种中心化方法, 其效果和动机也可从附录 A 中的引理 A.3 窥见一斑. 具体地, 我们采用如下的中心化的估计方程:

$$U_1(\phi, \gamma) = \sum_{i=1}^n \int_0^\tau J_\phi^T \{X_i(t) - \bar{X}_i(t; h)\} d\hat{M}_i(t; \phi, \gamma, h) = 0, \quad (3.4)$$

$$U_2(\phi, \gamma) = \sum_{i=1}^n \int_0^\tau \{Z_i(t) - \bar{Z}_i(t; h)\} d\hat{M}_i(t; \phi, \gamma, h) = 0, \quad (3.5)$$

其中 $\bar{X}_i(t; h) = \bar{X}(t; \beta_\phi^T X_i(t), h)$, $\bar{Z}_i(t; h) = \bar{Z}(t; \beta_\phi^T X_i(t), h)$, 及

$$d\hat{M}_i(t; \phi, \gamma, h) = \{Y_i(t) - \gamma^T Z_i(t)\} dN_i(t) - I(T_i \geq t) \exp\{\hat{\eta}^T W_i(t)\} d\hat{\mathcal{A}}(t; \beta_\phi^T X_i(t), \beta_\phi, \gamma, h).$$

令 $\hat{\phi}$ 和 $\hat{\gamma}$ 为上述估计方程 (3.4) 和 (3.5) 的解, 并记 $\hat{\beta} = \beta_{\hat{\phi}}$. 于是, 函数 g 可通过如下积分估计:

$$\hat{g}(v) = \int_{v_0}^v \hat{g}(u; \hat{\beta}, \hat{\gamma}, h) du.$$

在给出参数估计量的计算方法之前, 我们先考虑估计量 $\hat{\mathcal{A}}_v(t; v, \beta, \gamma, h)$ 和 $\hat{g}(v; \beta, \gamma, h)$ 关于 β, γ 和 v 的导数, 事实上, 对等式 (3.2) 和 (3.3) 两边同时求导即可得到. 具体地, 令

$$\begin{aligned} S^{*(k)}(t; v, h) &= n^{-1} \sum_{i=1}^n \int_0^t K_h(\beta^T X_i(s) - v)(\beta^T X_i(s) - v)^k dN_i(s), \\ S_x^{*(k)}(t; v, h) &= n^{-1} \sum_{i=1}^n \int_0^t K_h(\beta^T X_i(s) - v) X_i(s) (\beta^T X_i(s) - v)^k dN_i(s), \\ S^{(k)}(t; v, h) &= n^{-1} \sum_{i=1}^n I(T_i \geq t) K_h(\beta^T X_i(t) - v) \exp\{\hat{\eta}^T W_i(t)\} (\beta^T X_i(t) - v)^k, \\ \hat{D}_0(v; h) &= S^{*(2)}(\tau; v, h) - \int_0^\tau S^{(0)}(t; v, h)^{-1} S^{(1)}(t; v, h) dS^{*(1)}(t; v, h), \end{aligned}$$

则近似地, 有

$$\begin{aligned} \frac{\partial}{\partial \beta^T} \hat{g}(v; \beta, \gamma, h) &= -\hat{D}_0(v)^{-1} \left[S_x^{*(1)}(\tau; v, h) - \int_0^\tau S^{(0)}(t; v, h)^{-1} S^{(1)}(t; v, h) dS_x^{*(0)}(t; v, h) \right] \hat{g}(v; \beta, \gamma, h), \\ \frac{\partial}{\partial \gamma^T} \hat{g}(v; \beta, \gamma, h) &= -\hat{D}_0(v; h)^{-1} \left[S_z^{*(1)}(\tau; v, h) - \int_0^\tau S^{(0)}(t; v, h)^{-1} S^{(1)}(t; v, h) dS_z^{*(0)}(t; v, h) \right], \\ \frac{\partial}{\partial v} \hat{g}(v; \beta, \gamma, h) &= \hat{D}_0(v; h)^{-1} \left[S^{*(1)}(\tau; v, h) - \int_0^\tau S^{(0)}(t; v, h)^{-1} S^{(1)}(t; v, h) dS^{*(0)}(t; v, h) \right] \hat{g}(v; \beta, \gamma, h), \\ d\hat{\mathcal{A}}_\beta(t; v, \beta, \gamma, h) &= -S^{(0)}(t; v, h)^{-1} \left[\hat{g}(v; \beta, \gamma, h) dS_x^{*(0)}(t; v, h) + \frac{\partial}{\partial \beta^T} \hat{g}(v; h) dS^{*(1)}(t; v, h) \right], \\ d\hat{\mathcal{A}}_\gamma(t; v, \beta, \gamma, h) &= -S^{(0)}(t; v, h)^{-1} \left[dS_z^{*(0)}(t; v, h) + \frac{\partial}{\partial \gamma^T} \hat{g}(v; \beta, \gamma, h) dS^{*(1)}(t; v, h) \right], \end{aligned}$$

以及

$$d\hat{\mathcal{A}}_v(t; v, \beta, \gamma, h) = -S^{(0)}(t; v, h)^{-1} \left[\frac{\partial}{\partial v} \hat{g}(v; \beta, \gamma, h) dS^{*(1)}(t; v, h) - \hat{g}(v; \beta, \gamma, h) dS^{*(0)}(t; v, h) \right],$$

其中 $\hat{\mathcal{A}}_x = \partial \hat{\mathcal{A}} / \partial x^T$, $x = \gamma, \beta, v$. 有了这些导数的表达式, 定义

$$\begin{aligned} \Phi_{1i}(t; h, \phi, \gamma) &= \int_0^t [\{Z_i(s) - \bar{Z}_i(s; h)\} dN_i(s) + I(T_i \geq s) \exp\{\hat{\eta}^T W_i(s)\} d\hat{\mathcal{A}}_\gamma(s; \beta^T X_i(s), \beta, \gamma, h)], \\ \Phi_{2i}(t; v, h, \phi, \gamma) &= \int_0^t I(T_i \geq s) \exp\{\hat{\eta}^T W_i(s)\} J_\phi^T \{d\hat{\mathcal{A}}_\beta(s; v, \beta, h, \gamma, h) + X_i(s) d\hat{\mathcal{A}}_v(s; v, \beta, \gamma, h)\}, \end{aligned}$$

以及

$$\begin{aligned} \hat{\Gamma}_1(\phi, \gamma) &= \sum_{i=1}^n \int_0^\tau J_\phi^T \{X_i(t) - \bar{X}_i(t; h)\} d\Phi_{1i}(t, h)^T, \\ \hat{\Gamma}_2(\phi, \gamma) &= \sum_{i=1}^n \int_0^\tau \{Z_i(t) - \bar{Z}_i(t; h)\} d\Phi_{2i}(t, h)^T, \\ \hat{D}_1(\phi, \gamma) &= \sum_{i=1}^n \int_0^\tau [J_\phi^T \{X_i(t) - \bar{X}_i(t; h)\} - \hat{\Gamma}_1^{-1} \{Z_i(t) - \bar{Z}_i(t; h)\}] d\Phi_{2i}(t; \beta^T X_i(t), h), \\ \hat{D}_2(\phi, \gamma) &= \sum_{i=1}^n \int_0^\tau \{Z_i(t) - \bar{Z}_i(t; h)\} d\Phi_{1i}(t, h)^T. \end{aligned}$$

有了以上结果, 我们给出估计参数 β 和 γ 的迭代算法如下:

步骤一 为了构造参数的初始值, 令函数 g 具有线性形式 $g(v) = b(v - v_0)$, 从而有 $g(\beta^T X_i(t)) = b\beta^T X_i(t) + bv_0$, 然后应用文献 [8] 的估计方法, 可以得到 $b\beta$ 和 γ 的估计, 分别记作 $\widehat{b\beta}$ 和 $\gamma_{(0)}$. 令 $\beta_{(0)} = \text{sign}_1(\widehat{b\beta})\widehat{b\beta}/\|\widehat{b\beta}\|$, 其中 $\text{sign}_1(\cdot)$ 为求向量第一个维度的符号函数.

步骤二 在第 k 步, 给定 $\beta_{(k)}$ 和 $\gamma_{(k)}$ 时, 计算 $\hat{g}_{(k)}(v, h) \equiv \hat{g}(v; \beta_{(k)}, \gamma_{(k)}, h)$ 和 $\hat{\mathcal{A}}_{(k)}(t; v, h) \equiv \hat{\mathcal{A}}(t; v, \beta_{(k)}, \gamma_{(k)}, h)$. 设 $\hat{D}_{1(k)}$ 、 $\hat{D}_{2(k)}$ 、 $\hat{\Gamma}_{1(k)}$ 和 $\hat{\Gamma}_{2(k)}$ 分别为 $\hat{D}_1(\phi, \gamma)$ 、 $\hat{D}_2(\phi, \gamma)$ 、 $\hat{\Gamma}_1(\phi, \gamma)$ 和 $\hat{\Gamma}_2(\phi, \gamma)$ 在参数 $\phi_{(k)}$ 和 $\gamma_{(k)}$ 处的取值. 对等式 (3.4) 和 (3.5) 在当前参数 $\phi_{(k)}$ 和 $\gamma_{(k)}$ 处进行 Taylor 展开, 我们得到 ϕ 和 γ 的更新公式为

$$\begin{aligned} \phi_{(k+1)} &= \phi_{(k)} + \delta^{j_0} \hat{D}_{1(k)}^{-1} \sum_{i=1}^n \int_0^\tau \{J_{\phi_{(k)}}^T \{X_i(t) - \bar{X}_i(t; h)\} d\hat{M}_i(t; \phi_{(k)}, \gamma_{(k)}, h) \\ &\quad - \hat{\Gamma}_{1(k)} \hat{D}_{2(k)}^{-1} \{Z_i(t) - \bar{Z}_i(t; h)\} d\hat{M}_i(t; \phi_{(k)}, \gamma_{(k)}, h)\}, \\ \gamma_{(k+1)} &= \gamma_{(k)} + \hat{D}_{2(k)}^{-1} \left[\sum_{i=1}^n \int_0^\tau \{Z_i(t) - \bar{Z}_i(t; h)\} d\hat{M}_i(t; \phi_{(k)}, \gamma_{(k)}, h) - \hat{\Gamma}_{2(k)}(\phi_{(k+1)} - \phi_{(k)}) \right], \end{aligned}$$

其中 $\phi_{(k+1)}$ 的表达式中 j_0 为使得不等式 $\|\phi_{(k+1)}\| < 1$ 成立的最小非负整数, δ 为取值于区间 $(0, 1)$ 的任意常数, 实际应用中, 我们建议取黄金分割数 [28] $\delta = 0.618$ 以提高收敛效率.

步骤三 重复进行步骤二, 直到收敛.

接下来将推导所得估计量 $\hat{\phi}$ 和 $\hat{\gamma}$ 的渐近性质, 结果见下面的定理, 其证明见附录 A. 相较于文献中关于部分线性单指标模型的常见证明思路, 我们采用了一种全新的证明方法. 为了消除非参数部分估计量收敛速度的影响而得到参数部分估计量的渐近正态性, 文献中常用的方法是直接计算残余项的方差并通过对方差的控制来证明其可忽略 [22]. 然而, 在我们的模型中, 随机过程的复杂性使得方差的

直接计算不可行. 为此, 我们提出一种数据分割的证明方法, 详见附录 A 中的引理 A.3. 为了介绍定理中用到的一些记号, 定义

$$\hat{\Omega} = n^{-1} \sum_{i=1}^n \int_0^\tau \{W_i(t) - \bar{W}(t)\}^\otimes dN_i(t),$$

$$\hat{\zeta}_i = \int_0^\tau \{W_i(t) - \bar{W}(t)\} [dN_i(t) - I(T_i \geq t) \exp\{\hat{\eta}^T W_i(t)\} d\hat{\Lambda}_0(t)],$$

其中 $\bar{W}(t) = \sum_{i=1}^n I(T_i \geq t) \exp\{\hat{\eta} W_i(t)\} W_i(t) / \sum_{i=1}^n I(T_i \geq t) \exp\{\hat{\eta} W_i(t)\}$. 令 $\hat{D}_1$ 、 $\hat{D}_2$ 、 $\hat{\Gamma}_1$ 和 $\hat{\Gamma}_2$ 表示 $\hat{D}_1(\phi, \gamma)$ 、 $\hat{D}_2(\phi, \gamma)$ 、 $\hat{\Gamma}_1(\phi, \gamma)$ 和 $\hat{\Gamma}_2(\phi, \gamma)$ 在 $\hat{\phi}$ 和 $\hat{\gamma}$ 处的取值, 定义

$$\hat{A} = \begin{pmatrix} \hat{D}_1 & \hat{\Gamma}_1 \\ \hat{\Gamma}_2 & \hat{D}_2 \end{pmatrix},$$

$\hat{\xi}_i = (\hat{\xi}_{1i}^T, \hat{\xi}_{2i}^T)^T$, 其中

$$\hat{\xi}_{1i} = \int_0^\tau J_\phi^T \{X_i(t) - \bar{X}_i(t; h)\} d\hat{M}_i(t; \phi, \gamma, h) - \hat{B}_1 \hat{\Omega}^{-1} \hat{\zeta}_i,$$

$$\hat{\xi}_{2i} = \int_0^\tau \{Z_i(t) - \bar{Z}_i(t; h)\} d\hat{M}_i(t; \phi, \gamma, h) - \hat{B}_2 \hat{\Omega}^{-1} \hat{\zeta}_i,$$

以及

$$\hat{B}_1 = n^{-1} \sum_{i=1}^n \int_0^\tau J_\phi^T \{X_i(t) - \bar{X}_i(t; h)\} I(T_i \geq t) \exp\{\hat{\eta}^T W_i(t) W_i(t) d\hat{A}(t; \hat{\beta}^T X_i(t), \hat{\beta}, \hat{\gamma}, h)\},$$

$$\hat{B}_2 = n^{-1} \sum_{i=1}^n \int_0^\tau \{Z_i(t) - \bar{Z}_i(t; h)\} I(T_i \geq t) \exp\{\hat{\eta}^T W_i(t) W_i(t) d\hat{A}(t; \hat{\beta}^T X_i(t), \hat{\beta}, \hat{\gamma}, h)\}.$$

定理 3.1 在附录 A 所列的正则条件 (C1)–(C7) 下, $\sqrt{n}[(\hat{\phi}^T, \hat{\gamma}^T)^T - (\phi_0^T, \gamma_0^T)^T]$ 依分布收敛到一个零均值的正态随机向量, 其协方差矩阵可由 $\hat{A}^{-1} \hat{\Sigma} (\hat{A}^T)^{-1}$ 进行相合估计, 其中 $\hat{\Sigma} = n^{-1} \sum_{i=1}^n \hat{\xi}_i \hat{\xi}_i^T$.

注意到, $\hat{\beta}$ 的渐近性质并没有在定理 3.1 中直接给出, 但我们可以通过 delta 方法得到. 具体来说, 令 $\hat{D}_\beta = (\hat{J}_\phi \mathbf{0}_{p \times q})$, 则 $\sqrt{n}(\hat{\beta} - \beta_0)$ 依分布收敛于一个零均值正态分布, 其方差可由 $\hat{D}_\beta \hat{A}^{-1} \hat{\Sigma} (\hat{A}^T)^{-1} \hat{D}_\beta^T$ 进行相合估计.

核估计的窗宽选择在理论层面是一个很大的挑战, 我们提供一个 K 层交叉核实方法来选择最优窗宽. 记 $\mathcal{I}_k$ 为第 k 层个体的下标所构成的集合, 对于给定窗宽 h , 定义预测误差^[29] 为

$$PE(h) = \sum_{k=1}^K PE_k(h),$$

其中 $PE_k(h) = \sum_{i \in \mathcal{I}_k} \int_0^\tau \hat{M}_i^{(k)}(t)^2 d\bar{N}_k(t)$, $\bar{N}_k(t) = \sum_{i \in \mathcal{I}_k} N_i(t)$, 而 $\hat{M}_i^{(k)}(t)$ 则为 (3.4) 所定义零均值过程在不使用 $\mathcal{I}_k$ 中的数据所得到的估计量 $\hat{\gamma}^{(k)}$ 、 $\hat{\beta}^{(k)}$ 和 $\hat{\mathcal{A}}^{(k)}$ 下的取值.

4 数值模拟研究

我们将在本节提供一些数值模拟试验来检验前述估计量在有限样本下的表现, 试验中, 我们取时间相依协变量为 $X_{ij}(t) = X_{ij}^* \sqrt{t}$, $j = 1, 2, 3$, 其中 X_{ij}^* 独立产生于均匀分布 $U(0, 1)$, 而协变量 Z_i 则定

义为成功概率为 0.5 的 Bernoulli 随机变量. 终止事件 D_i 由均值为 $0.2 \exp(0.5Z_i)$ 的指数分布产生, 而删失时间则取为 $C_i = C_i^* \wedge \tau$, 其中 $\tau = 3$, C_i^* 服从均匀分布 $U(1, 8)$.

给定协变量以及 $T_i = D_i \wedge C_i$ 条件下, 我们通过强度函数如下的非齐次 Poisson 过程来产生观测时间: $\lambda_i(t) = 0.6(1+t) \exp(0.5Z_i) I(T_i \geq t)$, 平均每个个体有大约 3 次观测. 对于纵向响应变量 $Y_i(t)$, 我们通过如下模型来产生:

$$Y_i(t) = 1 + 0.5t + g(\beta_0^T X_i(t)) + Z_i + \epsilon_i(t),$$

其中

$$\beta_0 = (\beta_{10}, \beta_{20}, \beta_{30})^T = (2 + c^2)^{-1/2} (1, 1, c)^T,$$

取 c 为 1、-1 或者 0 三种情形, $g(x) = x + 1/2x^2$, $\epsilon_i(t)$ 是均值为 ψ_i 方差为 0.25 的正态分布随机变量, 而 ψ_i 则是产生于标准正态分布的随机扰动.

在模拟试验中, 考虑三种窗宽 h 取值 0.05、0.10 或者 0.15, 以及不同的样本量 n 取值 300、500、1,000 和 2,000. 基于 1,000 次重复模拟, β_0 和 γ_0 的估计量的模拟结果分别展示在表 1 和 2 中. 表中, Bias 为样本均值减去参数真实值, SE 为估计量的样本标准差, SEE 为估计量的标准差估计的样本均值, 而 CP 则为 95% 经验覆盖率. 为了节省空间, 我们仅在表 1 中展示 h 取值 0.05 和 0.10 以及 $c = 1$ 时 β_0 的估计结果, 其余设置下的结果类似. 模拟结果表明, 所提出的关于 β_0 的估计量可以很好地估计 β_0 , 但当样本量较小时, 估计量的稳定性稍差, 而当样本量增加时, 估计将变得更好. 表 2 中展示了样本量 n 取值为 300 和 500 时 γ_0 的估计结果, 易见, γ_0 的估计量都是无偏的, 且标准差的估计 (SEE) 与估计量的标准差 (SE) 吻合很好, 而 95% 经验覆盖率 (CP) 也都合理地取值于 0.95 左右. 此外, 模拟表明, 我们的估计量的对于窗宽变化并不敏感, 而且, 想要得到 β_0 的更好的估计, 将比得到 γ_0 的更好的估计需要更多的样本量. 这一现象在部分线性单指标模型中很常见, 主要是因为, 相较于 γ_0 , β_0 的估计更多地依赖于对非参数连接函数 g 的估计效果.

表 1 β_0 的估计结果

	n	$h = 0.05$				$h = 0.10$			
		Bias	SE	SEE	CP	Bias	SE	SEE	CP
β_{10} 的估计结果	300	-0.0259	0.1746	0.1569	0.883	-0.0133	0.1606	0.1550	0.907
	500	-0.0257	0.1316	0.1222	0.907	-0.0207	0.1241	0.1240	0.935
	1,000	-0.0093	0.0902	0.0872	0.939	-0.0080	0.0862	0.0880	0.945
	2,000	-0.0019	0.0612	0.0624	0.948	-0.0042	0.0606	0.0623	0.951
β_{20} 的估计结果	300	-0.0218	0.1712	0.1619	0.902	-0.0226	0.1597	0.1569	0.928
	500	-0.0100	0.1329	0.1206	0.913	-0.0136	0.1275	0.1233	0.929
	1,000	-0.0017	0.0903	0.0866	0.931	-0.0046	0.0900	0.0870	0.936
	2,000	0.0004	0.0621	0.0619	0.943	-0.0031	0.0621	0.0618	0.943
β_{30} 的估计结果	300	-0.0347	0.1816	0.1598	0.888	-0.0330	0.1628	0.1573	0.925
	500	-0.0099	0.1298	0.1208	0.905	-0.0074	0.1256	0.1223	0.909
	1,000	-0.0103	0.0904	0.0872	0.940	-0.0073	0.0861	0.0871	0.943
	2,000	-0.0087	0.0636	0.0624	0.949	-0.0025	0.0616	0.0617	0.946

表 2 γ_0 的估计结果

c	h	$n = 300$				$n = 500$			
		Bias	SE	SEE	CP	Bias	SE	SEE	CP
1	0.05	-0.0003	0.2537	0.2360	0.940	0.0052	0.1522	0.1543	0.949
	0.10	-0.0024	0.1939	0.1938	0.946	0.0031	0.1446	0.1445	0.953
	0.15	-0.0041	0.1873	0.1850	0.946	0.0037	0.1429	0.1427	0.949
-1	0.05	0.0011	0.1919	0.1959	0.953	-0.0034	0.1417	0.1737	0.932
	0.10	-0.0004	0.1700	0.1714	0.945	0.0110	0.1318	0.1303	0.948
	0.15	-0.0011	0.1668	0.1681	0.941	0.0113	0.1310	0.1282	0.949
0	0.05	-0.0012	0.2389	0.2145	0.933	0.0057	0.1510	0.1510	0.939
	0.10	0.0049	0.1828	0.1846	0.943	-0.0039	0.1404	0.1391	0.947
	0.15	0.0048	0.1798	0.1788	0.953	-0.0037	0.1381	0.1366	0.956

5 老年痴呆症数据分析

我们在本节将把所提出的估计方法应用于分析美国国家老年痴呆症协调中心统一数据集 (NACC UDS)^[30]. 该数据总共有 10,211 个患者, 都注册于 2005 年至 2007 年之间, 年龄皆大于 65 岁, 且分布于 31 个不同的老年痴呆症疾病中心. 在老年痴呆症的研究中, 为了发现是否有认知障碍, 患者需要参加一个简易精神状态检查 (MMSE), 所得分数取值于 0 到 30, 得分越低说明认知障碍越严重. 我们主要感兴趣的是以下因素对患者 MMSE 得分的影响: 性别 (二值变量, 1 表示男性)、年龄 (在患者注册当年)、抑郁 (二值变量, 1 表示存在抑郁), 以及是否高血压. 其他一些可能会对 MMSE 得分有影响的因素为种族 (二值变量, 1 表示白种人)、受教育年限 (除以 10)、母亲是否患有痴呆症、婚姻状态 (二值变量, 1 表示已婚), 以及糖尿病. 这些因素中, 糖尿病、是否高血压以及婚姻状态为随时间变化的时间相依变量.

研究中, 患者在删失或死亡之前有 1 到 9 次观测, 平均每位患者有 3 次观测, 观测时间设置为患者每次到访时间距离其注册年份年初的时间. 为了建立模型, 我们将感兴趣的因素 (性别、年龄、抑郁和高血压) 分配为模型的线性部分的协变量, 而将其余因素分配为单指标部分的协变量. 为了避免数值过大使得系数过小, 我们对年龄作变换 $(\text{age} - 70)/10$. 对于高血压, 我们引入两个哑变量, 一个表示当前时间是否有高血压, 一个表示过去是否患过高血压, 而参照变量则是没有患过高血压. 通过最小化 $PE(h)$, 我们得到最优窗宽为 $h_{\text{opt}} = 0.09$, 在此窗宽下的估计结果见表 3. 这些估计结果表明, 性

表 3 老年痴呆症数据的分析结果

线性部分 ($\hat{\gamma}$)			单指标部分 ($\hat{\beta}$)		
协变量	EST	SE	协变量	EST	SE
性别	-1.2051	0.1319	种族	0.3026	0.0406
年龄	-0.9581	0.0861	教育年限	0.9438	0.0161
抑郁	-0.6459	0.1726	母亲患老年痴呆症	-0.0606	0.0473
患高血压 (当前)	0.7335	0.1736	婚姻状态	-0.1155	0.0522
患高血压 (曾经)	-0.5016	0.4165	糖尿病	-0.0247	0.0617

别、年龄、抑郁和高血压全都具有显著效应. 具体来讲, 男性的 MMSE 得分明显比女性更低, 痴呆症状更加严重; 随着年龄的增加, 得分也会明显降低, 痴呆症状加重; 抑郁症会明显地加重老年痴呆症患者的病情; 而患有高血压对痴呆症的影响则具有滞后性, 即当前患有高血压不会使得痴呆症病情加重, 而曾经患有高血压则会使得痴呆症病情加重, 这一结果很有意义, 还未被文献所揭示.

6 结论

本文提出了一个分析带有终止事件纵向数据的半参数部分线性单指标模型, 该模型既包含了线性和单指标的协变量效应, 也包含了一个非参数的随时间变化的基准效应, 具有很好的广泛性和灵活性. 我们给出了基于核平滑的估计方程方法, 并将主要精力放在对参数部分估计量的渐近正态性的推导. 在引入了一些关于窗宽的假定后, 我们通过一种数据分割的方法证明了参数部分的估计量的 $\sqrt{n}$ 相合性, 得到其渐近正态分布. 然而, 对于非参数部分, 我们仅能得到估计量的相合性, 无法做统计推断, 由于随机过程的复杂性, 想要得到其渐近分布, 还需要更多努力. 数值模拟研究表明, 我们的估计方法可以很好地工作, 表现可靠. 我们还将所提估计方法用于分析一组老年痴呆症的数据, 得到了一些新的结果. 对于窗宽选择, 我们建议使用基于数据的分层交叉核实法, 但是如何寻找一个理论上的最优窗宽还有待进一步的研究.

参考文献

- 1 Laird N M, Ware J H. Random-effects models for longitudinal data. *Biometrics*, 1982, 38: 963–974
- 2 Liang K Y, Zeger S L. Longitudinal data analysis using generalized linear models. *Biometrika*, 1986, 73: 13–22
- 3 Welsh A H, Lin X, Carroll R J. Marginal longitudinal nonparametric regression. *J Amer Statist Assoc*, 2002, 97: 482–493
- 4 Davis C S. *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer-Verlag, 2002
- 5 Fitzmaurice G M, Laird N M, Ware J H. *Applied Longitudinal Analysis*. New York: John Wiley and Sons, 2004
- 6 Fan J, Li R. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J Amer Statist Assoc*, 2004, 99: 710–723
- 7 Diggle P J, Heagerty P, Liang K Y, et al. *Analysis of Longitudinal Data*, 2nd ed. Oxford: Oxford University Press, 2002
- 8 Lin D Y, Ying Z. Semiparametric and nonparametric regression analysis of longitudinal data. *J Amer Statist Assoc*, 2001, 96: 103–126
- 9 Wu M C, Carroll R J. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 1988, 44: 175–188
- 10 Roy J, Lin X. Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates. *J Amer Statist Assoc*, 2002, 97: 40–52
- 11 Lin D Y, Ying Z L. Semiparametric regression analysis of longitudinal data with informative drop-outs. *Biostatistics*, 2003, 4: 385–398
- 12 Jin Z, Liu M L, Albert S, et al. Analysis of longitudinal health-related quality of life data with terminal events. *Lifetime Data Anal*, 2006, 12: 169–190
- 13 Liu M, Ying Z. Joint analysis of longitudinal data with informative right censoring. *Biometrics*, 2007, 63: 363–371
- 14 Li L, Hu B, Greene T. A semiparametric joint model for longitudinal and survival data with application to hemodialysis study. *Biometrics*, 2009, 65: 737–745
- 15 Liu L, Huang X, O’Quigley J. Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics*, 2008, 64: 950–958
- 16 Sun L, Song X, Zhou J, et al. Joint analysis of longitudinal data with informative observation times and a dependent terminal event. *J Amer Statist Assoc*, 2012, 107: 688–700

- 17 Zhang J L, Rubin D B. Estimation of causal effects via principal stratification when some outcomes are truncated by “death”. *J Educ Behav Stat*, 2003, 28: 353–368
- 18 Ding P, Geng Z, Yan W, et al. Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *J Amer Statist Assoc*, 2011, 106: 1578–1591
- 19 Wang L, Richardson T S, Zhou X H. Causal analysis of ordinal treatments and binary outcomes under truncation by death. *J R Stat Soc Ser B Stat Methodol*, 2017, 79: 719–735
- 20 Wen L, Terrera G M, Seaman S R. Methods for handling longitudinal outcome processes truncated by dropout and death. *Biostatistics*, 2018, 19: 407–425
- 21 Carroll R J, Fan J, Gijbels I, et al. Generalized partially linear single-index models. *J Amer Statist Assoc*, 1997, 92: 477–489
- 22 Xia Y, Härdle W. Semi-parametric estimation of partially linear single-index models. *J Multivariate Anal*, 2006, 97: 1162–1184
- 23 Wang J L, Xue L, Zhu L, et al. Estimation for a partial-linear single-index model. *Ann Statist*, 2010, 38: 246–274
- 24 Dong C, Gao J, Tjøstheim D. Estimation for single-index and partially linear single-index integrated models. *Ann Statist*, 2016, 44: 425–453
- 25 Li G, Zhu L, Xue L, et al. Empirical likelihood inference in partially linear single-index models for longitudinal data. *J Multivariate Anal*, 2010, 101: 718–732
- 26 Lin D Y, Wei L J, Yang I, et al. Semiparametric regression for the mean and rate functions of recurrent events. *J R Stat Soc Ser B Stat Methodol*, 2000, 62: 711–730
- 27 Ye Y, Kalbfleisch J D, Schaebel D E. Semiparametric analysis of correlated recurrent and terminal events. *Biometrics*, 2007, 63: 78–87
- 28 Kiefer J. Sequential minimax search for a maximum. *Proc Amer Math Soc*, 1953, 4: 502–506
- 29 Tian L, Zucker D, Wei L J. On the cox model with time-varying regression coefficients. *J Amer Statist Assoc*, 2005, 100: 172–183
- 30 Chen B, Zhou X H. Generalized partially linear models for incomplete longitudinal data in the presence of population-level information. *Biometrics*, 2013, 69: 386–395
- 31 Sun Y, Sun L, Zhou J. Profile local linear estimation of generalized semiparametric regression model for longitudinal data. *Lifetime Data Anal*, 2013, 19: 317–349
- 32 Yin G, Li H, Zeng D. Partially linear additive hazards regression with varying coefficients. *J Amer Statist Assoc*, 2008, 103: 1200–1213
- 33 van der Vaart A W, Wellner J A. *Weak Convergence and Empirical Processes*. New York: Springer, 1996
- 34 van der Vaart A W. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 1998

附录 A 定理证明

我们列出如下正则性条件:

- (C1) 核函数 $K(\cdot)$ 是有界对称且定义于紧支撑上的密度函数.
- (C2) $P(T \geq \tau) > 0$, 且存在 $C > 0$ 和 $\rho \in (0, 2)$ 使得 $E\{N(t) - N(s)\}^2 \leq C|t - s|^\rho$ 成立.
- (C3) 函数 $g(\cdot)$ 二阶连续可导.
- (C4) $h \rightarrow 0$, 且存在 $\epsilon > 0$ 使得 $\log(h)/\sqrt{nh^{4+\epsilon}} \rightarrow 0$ 成立.
- (C5) $X(t)$ 、 $Z(t)$ 和 $Y(t)$ 的样本路径在 $[0, \tau]$ 上一致有界且具有有界变差.
- (C6) 给定 $X(t)^T \beta = v_0$ 时 $(X(t), Z(t))$ 的条件密度函数关于 v_0 在 $\mathcal{V}$ 上连续可微. $X(t)^T \beta$ 的边缘密度函数, 记作 $f_{\beta, t}(v_0)$, 关于 v_0 二阶连续可微, 且满足条件 $\inf_{v_0 \in \mathcal{V}, t \in [0, \tau], \beta \in \mathcal{B}} f_{\beta, t}(v_0) > 0$.
- (C7) 矩阵 A 是非奇异的, 其中 A 是 $\hat{A}$ 的极限.

上述条件 (C1)–(C5) 是核平滑方法中的常见条件 (参见文献 [31]). 条件 (C6) 是关于密度的正则

性条件的随机过程版本^[32]. 在证明定理 3.1 之前, 我们先介绍一些记号及几个引理. 令

$$\begin{aligned}\mu_k &= \int u^k K(u) du, \quad P(v; t, \beta) = E\{I(T_i \geq t \exp(W_i(t)^T \eta_0) \mid \beta^T X_i(t) = v)\}, \\ s^{(k)}(t; \beta, v, h) &= P(v; t, \beta) f_{\beta, t}(v) \mu_k + \{P(v; t, \beta) \dot{f}_{\beta, t}(v) + \dot{P}(v; t, \beta) f_{\beta, t}(v)\} \mu_{k+1} h, \\ s^{*(k)}(t; \beta, v, h) &= \int_0^t s^{(k)}(u; \beta, v, h) du,\end{aligned}$$

其中 $\dot{f}_{\beta, t}(v)$ 和 $\dot{P}(v; t, \beta)$ 分别为 $f_{\beta, t}(v)$ 和 $P(v; t, \beta)$ 关于 v 的导数. 记 $\bar{x}_i(t)$ 、 $\bar{z}_i(t)$ 和 $\bar{w}(t)$ 分别为 $\bar{X}_i(t)$ 、 $\bar{Z}_i(t)$ 和 $\bar{W}(t)$ 的极限.

引理 A.1 在条件 (C1)–(C6) 下, 对于 $k = 0, 1, 2$, 我们有

$$\sup_{t \in [0, \tau], v \in \mathcal{V}, \beta \in \mathcal{B}} |S^{(k)}(t; \beta, v, h) h^{-k} - s^{(k)}(t; \beta, v, h)| = O_p\left(\frac{\log(h)}{\sqrt{nh}}\right) + O(h^2) + O_p(n^{-1/2}), \quad (\text{A.1})$$

$$\sup_{t \in [0, \tau], v \in \mathcal{V}, \beta \in \mathcal{B}} |S^{*(k)}(t; \beta, v, h) h^{-k} - s^{*(k)}(t; \beta, v, h)| = O_p\left(\frac{\log(h)}{\sqrt{nh}}\right) + O(h^2), \quad (\text{A.2})$$

以及

$$\sup_{t \in [0, \tau], v \in \mathcal{V}, \beta \in \mathcal{B}} |S_a^{*(k)}(t; \beta, v, h) h^{-k} - s_a^{*(k)}(t; \beta, v, h)| = O_p\left(\frac{\log(h)}{\sqrt{nh}}\right) + O(h^2), \quad a = x, z. \quad (\text{A.3})$$

证明 对于 (A.1), 注意到

$$\begin{aligned}S^{(k)}(t; \beta, v, h) h^{-k} &= n^{-1} \sum_{i=1}^n I(T_i \geq t) K_h(\beta^T X_i(t) - v) \exp\{\eta_0 W_i(t)\} \left\{ \frac{\beta^T X_i(t) - v}{h} \right\}^k \\ &\quad + n^{-1} \sum_{i=1}^n I(T_i \geq t) K_h(\beta^T X_i(t) - v) \exp\{\eta^* W_i(t)\} \left\{ \frac{\beta^T X_i(t) - v}{h} \right\}^k W_i(t)^T (\hat{\eta} - \eta_0),\end{aligned}$$

其中 η^* 处于 $\hat{\eta}$ 与 η_0 之间的连线上, 而 $\hat{\eta} - \eta_0 = O_p(\frac{1}{\sqrt{n}})$, 因而, 我们可以很容易地证明上述表达式中的第二项是 $O_p(\frac{1}{\sqrt{n}})$, 余下的证明与文献 [32, 引理 A.1] 的证明类似.

(A.2) 与 (A.3) 的证明相似, 我们仅在此给出 (A.2) 的证明. 设 $\mathbf{P}_n$ 和 $\mathbf{G}_n$ 分别表示 n 个样本对应的经验测度与经验过程^[33], $H_k(s; \beta, v) \equiv K_h(\beta^T X(s) - v) \{(\beta^T X_i(s) - v)/h\}^k$, 并定义函数类

$$\mathcal{F}_k = \left\{ \int_0^t H_k(s; \beta, v) dN(s) : \beta \in \mathcal{B}, v \in \mathcal{V}, t \in [0, \tau] \right\}.$$

在条件 (C1) 和 (C5) 下, 易得

$$\sup_{s \in [0, \tau]} |H_k(s; \beta_1, v_1) - H_k(s; \beta_2, v_2)| \lesssim (\|\beta_1 - \beta_2\| + |v_1 - v_2|) h^{-2}$$

对 $k = 0, 1, 2$ 成立, 其中 $a \lesssim b$ 表示存在 $c > 0$ 使得 $a \leq cb$. 从而, 对任意两个点 (β_1, v_1, t_1) 和 (β_2, v_2, t_2) , 如果它们之间的 Euclid 距离小于 δ 且 $t_1 < t_2$, 则在条件 (C1)、(C2) 和 (C5) 下, 我们有如下结果:

$$\begin{aligned}E \left\{ \int_0^{t_1} H_k(s; \beta_1, v_1) dN(s) - \int_0^{t_2} H_k(s; \beta_2, v_2) dN(s) \right\}^2 \\ \leq 2E \left\{ \int_{t_1}^{t_2} H_k(s; \beta_1, v_1) dN(s) \right\}^2 + 2E \left\{ \int_0^{t_2} (H_k(s; \beta_1, v_1) - H_k(s; \beta_2, v_2)) dN(s) \right\}^2\end{aligned}$$

$$\begin{aligned} &\lesssim \mathbb{E}\{N(t_2) - N(t_1)h^{-1}\}^2 + \sup_{s \in [0, \tau]} |H_k(s; \beta_1, v_1) - H_k(s; \beta_2, v_2)|^2 \{\mathbb{E}N(t_2)\}^2 \\ &\lesssim h^{-2}|t_2 - t_1|^\rho + \delta^2 h^{-4} = O(\delta^\rho h^{-4}). \end{aligned}$$

我们知道, 集合 $\{(\beta^\top, v, t)^\top : \beta \in \mathcal{B}, v \in \mathcal{V}, t \in [0, \tau]\}$ 在 Euclid 距离下直径为 δ 的覆盖数^[33] 为 $m = O(\delta^{-(p+2)})$, 因而, 结合上述不等式结果, 在 L_2 距离下, $\mathcal{F}_k$ 的覆盖数为

$$N(\delta, \mathcal{F}_k, L_2(P)) = O((\delta^2 h^4)^{-(p+2)/\rho}).$$

由于函数类 $\mathcal{F}_k$ 的覆盖函数是 $O(h^{-1})$ 的, 应用文献 [34, 推论 19.38], 我们得到

$$\mathbb{E}^* \|\mathbf{G}_n\|_{\mathcal{F}_k} \leq \int_0^1 \sqrt{1 + \log O((\delta^2 h^4)^{-(p+2)/\rho})} d\delta h^{-1} = O(\log(h)h^{-1}),$$

其中 $\mathbb{E}^*$ 表示外期望. 故有

$$\sup_{\beta \in \mathcal{B}, v \in \mathcal{V}, t \in [0, \tau]} \left| \mathbf{P}_n \int_0^t H_k(s; \beta, v) dN(s) - \mathbb{E} \left(\int_0^t H_k(s; \beta, v) dN(s) \right) \right| = O_p \left(\frac{\log(h)}{\sqrt{nh}} \right). \quad (\text{A.4})$$

简单计算并应用 Taylor 展开, 有

$$\begin{aligned} \mathbb{E} \left(\int_0^t H_k(s; \beta, v) dN(s) \right) &= \mathbb{E} \int_0^t K_h(\beta^\top X(s) - v) \left\{ \frac{\beta^\top X_i(s) - v}{h} \right\}^k I(T_i \geq s) \exp(W(s)^\top \eta_0) d\Lambda_0(s) \\ &= \mathbb{E} \int_0^t K_h(\beta^\top X(s) - v) \left\{ \frac{\beta^\top X_i(s) - v}{h} \right\}^k P(u; s, \beta) d\Lambda_0(s) \\ &= \int_0^t \int K(u) u^k P(v + uh; s, \beta) f_\beta(v + uh, s) du d\Lambda_0(s) \\ &= s^{(k)}(t; \beta, v, h) + O(h^{-2}). \end{aligned}$$

综合以上两个结果, 我们证得 (A.2) 成立. $\square$

引理 A.2 在条件 (C1)–(C6) 下,

$$\sup_{v \in \mathcal{V}} |\hat{g}(v; \beta_0, \gamma_0, h) - \dot{g}(v)|h = O_p \left(\frac{\log(h)}{\sqrt{nh}} \right) + O(h^2), \quad (\text{A.5})$$

$$\sup_{t \in [0, \tau], v \in \mathcal{V}} |\hat{\mathcal{A}}(t; v, \beta_0, \gamma_0, h) - \mathcal{A}_0(t; v)| = O_p \left(\frac{\log(h)}{\sqrt{nh}} \right) + O(h^2). \quad (\text{A.6})$$

证明 (A.5) 和 (A.6) 可由引理 A.1 的结论以及类似的方法得到证明. $\square$

引理 A.3 设 $O_i(t)$ 的路径一致有界且具有有界变差, 且 $\mathbb{E}\{O_i(t) \mid \beta_0^\top X_i(t)\} = 0$, 则在条件 (C1)–(C6) 下,

$$n^{-1/2} \sum_{i=1}^n \int_0^\tau O_i(t) d\{\hat{\mathcal{A}}(t; \beta_0^\top X_i(t)) - \mathcal{A}_0(t; \beta_0^\top X_i(t))\} = o_p(1), \quad (\text{A.7})$$

其中 $\hat{\mathcal{A}}(t; \beta_0^\top X_i(t)) = \hat{\mathcal{A}}(t; \beta_0^\top X_i(t), \beta_0, \gamma_0, h)$.

证明 由于核估计 $\hat{\mathcal{A}}(t; v)$ 不是 $n^{1/2}$ 相合的, 为了消除该非参数估计低收敛速度的影响, 我们采用一种数据分割的方法来进行证明, 其主要想法来自于借助外部信息进行估计时数据的独立性结构大大降低了理论难度. 首先, 我们将所有数据近似等分为 L 组, 并以 $\mathcal{T}_l$ 表示第 l 组的下标所构成的集合.

设 $\hat{g}^{(l)}(\cdot)$ 和 $\hat{\mathcal{A}}^{(l)}(\cdot)$ 为基于去掉第 l 组的所有其他组的数据得到的估计量, 则它们与第 l 组数据相互独立, 例如, $\hat{\mathcal{A}}^{(l)}(\cdot)$ 的定义与第 l 组数据没有关系, 因而与 $\{Y_i(\cdot), i \in \mathcal{T}_l\}$ 独立. 我们断言:

$$\sup_{t,v} |\hat{\mathcal{A}}(t;v) - \hat{\mathcal{A}}^{(l)}(t;v)| = O_p\left(\frac{\log(h)}{\sqrt{nLh^2}}\right), \quad l = 1, \dots, L. \quad (\text{A.8})$$

为了证明此论断, 令 $S^{(kl)}(t; \beta, v, h)$ 为第 l 个组定义的相对应于 $S^{(k)}(t; \beta, v, h)$ 的统计量, 注意到 $\hat{\mathcal{A}}(t; v)$ 的定义实质上是通过若干类似于 $S^{(k)}(t; \beta, v, h)$ 这样的求和项的组合得到, 我们仅举例严格证明如下式子:

$$\sup_{t,v} |S^{(kL)}(t; \beta_0, v, h) - S^{(k)}(t; \beta_0, v, h)| h^{-k} = O_p\left(\frac{\log(h)}{\sqrt{nLh^2}}\right), \quad (\text{A.9})$$

而 (A.8) 的证明可通过类似的方法以及泛函版的 Taylor 展开得到. 下面来证明 (A.9), 令

$$Q_k(t, v, h) = I(T \geq t) K_h(\beta_0^T X(t) - v) \exp\{\eta_0 W(t)\} \frac{\{\beta_0^T X(s) - v\}^k}{h^k},$$

则有 $S^{(k)}(t; \beta_0, v, h) = \mathbf{P}_n Q_k$ 和 $S^{(kL)}(t; \beta_0, v, h) = \mathbf{P}_{n-n/L} Q_k$. 简单计算可得如下不等式:

$$|\mathbf{P}_n Q_k - \mathbf{P}_{n-n/L} Q_k| = L^{-1} |\mathbf{P}_{n-n/L} Q_k - \mathbf{P}_{n/L} Q_k| \leq L^{-1} \{|\mathbf{P}_{n-n/L} Q_k - \mathbf{E} Q_k| + |\mathbf{P}_{n/L} Q_k - \mathbf{E} Q_k|\}.$$

应用类似于 (A.4) 的结论, 我们有

$$L^{-1} \sup |\mathbf{P}_{n-n/L} Q_k - \mathbf{E} Q_k| = O_p\left(\frac{\log(h)}{\sqrt{nL^2h^2}}\right) \quad \text{和} \quad L^{-1} \sup |\mathbf{P}_{n/L} Q_k - \mathbf{E} Q_k| = O_p\left(\frac{\log(h)}{\sqrt{nLh^2}}\right),$$

从而 (A.9) 得证.

我们回到 (A.7) 的证明, 注意到以下结果:

$$\begin{aligned} & n^{-1/2} \sum_{i=1}^n \int_0^\tau O_i(t) d\{\hat{\mathcal{A}}(t; \beta_0^T X_i(t), \beta_0, \gamma_0, h) - \mathcal{A}_0(t; \beta_0^T X_i(t))\} \\ &= n^{-1/2} \sum_{l=1}^L \sum_{i \in \mathcal{T}_l} \int_0^\tau O_i(t) d\{\hat{\mathcal{A}}(t; \beta_0^T X_i(t)) - \hat{\mathcal{A}}^{(l)}(t; \beta_0^T X_i(t))\} \\ & \quad + n^{-1/2} \sum_{l=1}^L \sum_{i \in \mathcal{T}_l} \int_0^\tau O_i(t) d\{\hat{\mathcal{A}}^{(l)}(t; \beta_0^T X_i(t)) - \mathcal{A}_0(t; \beta_0^T X_i(t))\} \\ & \equiv \Pi_1 + \Pi_2. \end{aligned}$$

由于 $O_i(t)$ 的路径是一致有界且有界变差, 应用分部积分公式以及 (A.8) 的结论, 我们有

$$|\Pi_1| \lesssim n^{-1/2} \sum_{i=1}^n \|O_i\|_{BV} \sup_{t,v,l} |\hat{\mathcal{A}}(t;v) - \hat{\mathcal{A}}^{(l)}(t;v)| = O_p\left(\frac{\log(h)}{\sqrt{Lh^2}}\right), \quad (\text{A.10})$$

其中 $\|O_i\|_{BV} = |O_i(0)| + |O_i(\tau)| + \int_0^\tau |dO_i(t)|$ 是一致有界的. 对于第二项 Π_2 , 注意到 $\mathbf{E}(O_i(t) | \beta_0^T X_i(t)) = 0$ 以及 $\hat{\mathcal{A}}^{(l)}$ 的定义不依赖于第 l 组中的个体, 从而, Π_2 实际上是一组独立同分布的零均值随机变量的和. 应用 Cauchy-Schwarz 不等式和分部积分公式, 以及引理 A.1 的结论, 我们得到如下结果:

$$\mathbf{E} \Pi_2^2 \leq Ln^{-1} \sum_{l=1}^L \mathbf{E} \left[\sum_{i \in \mathcal{T}_l} \int_0^\tau O_i(t) d\{\hat{\mathcal{A}}^{(l)}(t; \beta_0^T X_i(t)) - \mathcal{A}_0(t; \beta_0^T X_i(t))\} \right]^2$$

$$\begin{aligned}
&= LE \left[\int_0^\tau O_i(t) d\{\hat{\mathcal{A}}^{(1)}(t; \beta_0^T X_i(t)) - \mathcal{A}_0(t; \beta_0^T X_i(t))\} \right]^2 \\
&= O\left(\frac{L \log^2(h)}{nh^2} + Lh^4\right). \tag{A.11}
\end{aligned}$$

取 $L = h^{-2-\epsilon}$, 则由正则条件 (C3) 有 $\frac{\log(h)}{\sqrt{Lh^2}} = o(1)$ 和 $\frac{L \log^2(h)}{nh^2} + Lh^4 = o(1)$, 从而, 由 (A.10) 和 (A.11) 分别可知, $\Pi_1 = o_p(1)$ 和 $\Pi_2 = o_p(1)$. 此外, 每组的样本量 $n/L \rightarrow \infty$ 可以保证组内的大样本性质. $\square$

定理 3.1 的证明 为简单起见, 我们仅给出证明的大致思路. 首先, 我们不加证明地引入下述经典结论^[26]:

$$n^{1/2}(\hat{\eta} - \eta_0) = \Omega^{-1} n^{-1/2} \sum_{i=1}^n \zeta_i + o_p(1),$$

其中

$$\begin{aligned}
\Omega &= E \int_0^\tau \{W_i(t) - \bar{w}(t)\}^{\otimes 2} dN_i(t), \\
\zeta_i &= \int_0^\tau \{W_i(t) - \bar{w}(t)\} [dN_i(t) - I(T_i \geq t) \exp\{\eta_0^T W_i(t)\} d\Lambda_0(t)].
\end{aligned}$$

下面推导 $n^{-1/2}U_1(\phi_0, \gamma_0)$ 的渐近正态性, 首先, 注意到

$$n^{-1/2}U_1(\phi_0, \gamma_0) = n^{-1/2} \sum_{i=1}^n \int_0^\tau J_{\phi_0}^T \{X_i(t) - \bar{X}_i(t; h)\} d\hat{M}_i(t; \phi_0, \gamma_0, h) \equiv I + II + III + IV,$$

其中

$$\begin{aligned}
I &= n^{-1/2} \sum_{i=1}^n \int_0^\tau J_{\phi_0}^T \{X_i(t) - \bar{x}_i(t)\} dM_i(t), \\
II &= n^{-1/2} \sum_{i=1}^n \int_0^\tau J_{\phi_0}^T \{X_i(t) - \bar{x}_i(t)\} d\{\hat{M}_i(t; \phi_0, \gamma_0, h) - M_i(t)\}, \\
III &= n^{-1/2} \sum_{i=1}^n \int_0^\tau J_{\phi_0}^T \{\bar{X}_i(t; h) - \bar{x}_i(t)\} dM_i(t), \\
IV &= n^{-1/2} \sum_{i=1}^n \int_0^\tau J_{\phi_0}^T \{\bar{X}_i(t; h) - \bar{x}_i(t)\} d\{\hat{M}_i(t; \phi_0, \gamma_0, h) - M_i(t)\}.
\end{aligned}$$

第一项可由中心极限定理证明是渐近正态的. 对于第二项, 我们有

$$\begin{aligned}
II &= n^{-1/2} \sum_{i=1}^n \int_0^\tau J_{\phi_0}^T \{X_i(t) - \bar{x}_i(t)\} I(T_i \geq t) \{\exp\{\hat{\eta}^T W_i(t)\} - \exp\{\eta_0^T W_i(t)\}\} d\hat{\mathcal{A}}(t; \beta_0^T X_i(t), \beta_0, \gamma_0, h) \\
&\quad + n^{-1/2} \sum_{i=1}^n \int_0^\tau J_{\phi_0}^T \{X_i(t) - \bar{x}_i(t)\} I(T_i \geq t) \exp\{\eta_0^T W_i(t)\} \\
&\quad \times d\{\hat{\mathcal{A}}(t; \beta_0^T X_i(t), \beta_0, \gamma_0, h) - \mathcal{A}_0(t; \beta_0^T X_i(t))\} \\
&\equiv II_1 + II_2.
\end{aligned}$$

易知 $II_1 = B_1 \Omega^{-1} n^{-1/2} \sum_{i=1}^n \zeta_i + o_p(1)$, 其中 B_1 是下式的极限:

$$n^{-1} \sum_{i=1}^n \int_0^\tau J_{\phi_0}^T \{X_i(t) - \bar{x}_i(t)\} I(T_i \geq t) \exp\{\eta_0^T W_i(t)\} W_i(t) d\hat{\mathcal{A}}(t; \beta_0^T X_i(t), \beta_0, \gamma_0, h).$$

由于 $E(J_{\phi_0}^T \{X_i(t) - \bar{x}_i(t)\} I(T_i \geq t) \exp\{\eta_0^T W_i(t)\} W_i(t) \mid \beta_0^T X_i(t)) = 0$, 故由正则性条件 (C5) 和引理 A.3 的结论可得 $II_2 = o_p(1)$. $\square$

类似于文献 [26] 对其引理 A.1 的证明过程, 我们可以证得第三项也是可忽略的, 即 $III = o_p(1)$. 类似于前面对于引理 A.3 的证明, 我们可以证得 $IV = o_p(1)$. 从而, 合并针对这四项的结论, 可得

$$n^{-1/2}U_1(\phi_0, \gamma_0) = n^{-1/2} \sum_{i=1}^n \xi_{1i} + o_p(1),$$

其中

$$\xi_{1i} = \int_0^\tau J_{\phi_0}^T \{X_i(t) - \bar{x}_i(t; h)\} dM_i(t) + B_1 \Omega^{-1} \zeta_i.$$

类似的方法可用于证明

$$n^{-1/2}U_2(\phi_0, \gamma_0) = n^{-1/2} \sum_{i=1}^n \xi_{2i} + o_p(1),$$

其中

$$\xi_{2i} = \int_0^\tau \{Z_i(t) - \bar{z}_i(t)\} dM_i(t) + B_2 \Omega^{-1} \zeta_i + o_p(1).$$

令 $U(\phi, \gamma) = (U_1(\phi, \gamma)^T, U_2(\phi, \gamma)^T)^T$ 和 $\xi_i = (\xi_{1i}^T, \xi_{2i}^T)^T$, 则由中心极限定理, $n^{-1/2}U(\phi_0, \gamma_0)$ 依分布收敛于一个均值为 0 协方差为 $\Sigma = E\xi_i \xi_i^T$ 的正态分布随机向量. 最后, 类似于引理 A.1 的讨论, 我们可以证明, $\frac{\partial}{\partial(\phi^T, \gamma^T)^T} n^{-1}U(\phi, \gamma)$ 依概率一致收敛到它的极限, 一些简单推导可得, $\sqrt{n}[(\hat{\phi}^T, \hat{\gamma})^T - (\phi_0^T, \gamma_0)^T]$ 依分布收敛于一个均值为 0 协方差为 $A^{-1}\Sigma A^{-1}$ 的正态分布随机向量.

A partially linear single index model for longitudinal data with terminal event

Jie Zhou, Xiaohua Zhou & Liuquan Sun

Abstract In longitudinal studies, we may be interested in some non-terminal event, such as health related quality of life. However, before the outcome of interest can be observed, a terminal event, such as death, can occur. If this situation occurs, we call this problem as truncation by death. In addition, observed times for subjects may be irregular, and the effect of covariates on the outcome of interest may be complicated. In this paper, we propose a new semiparametric partially linear single index model in the presence of truncation by death. A local linear approximation is used for the nonparametric link function, and estimating equation approaches are developed. The obtained estimator for the link function is proved to be uniformly consistent with convergence rate under root n . A data-splitting technique is developed to eliminate the side effect of this low convergence rate to establish the asymptotic normality of the parametric part. The finite sample behavior is examined by simulation studies, and an application to Alzheimer's disease data is illustrated.

Keywords longitudinal data, partially linear single index, terminal event

MSC(2010) 62G05, 62N02

doi: 10.1360/SCM-2018-0091