

含治愈个体的复发事件下半参数比率模型 *

曾小凤 陈传钟 李霓*

(海南师范大学数学与统计学院, 海口, 571158)

摘要

随着科技及医疗水平的不断提高, 对于一些反复发生且被认为是不可能被治愈的疾病, 近年来发现有疑似治愈个体的存在。针对这一现象, 本文在原来的复发事件数据的半参数比率模型基础之上, 利用Logistic模型回归治愈率部分, 提出一类含有治愈个体的半参数比率模型, 来刻画协变量对事件复发率的影响。同时给出该模型中未知参数的估计方法, 证明这些估计的相合性和渐近性正态性。并通过数值模拟验证了这些估计在有限样本下也是有效的, 并且把该模型及方法用于一组实际的膀胱癌数据分析中。

关键词: 复发事件, 半参数比率模型, 治愈率, Logistic模型, 估计方程.

学科分类号: O212.7.

§1. 引言

在生物医学、工程学、社会学和经济学等学科领域的纵向数据研究中, 生存分析(survival analysis)已经成为许多学者关注的热点。在早期生存分析的模型及方法研究中, 人们主要研究实验对象一次死亡或发病的情况(Cox, 1972)。近来, 人们开始研究同一个实验对象疾病多次复发的情况(Andersen和Gill, 1982)。如膀胱肿瘤患者在治疗过程中的多次复发; 冠心病患者在随访过程中的多次急性事件; 白血病患者在骨髓移植后会经历多次反复的感染等等。描述这类事件的数据称为复发事件数据(recurrent event data)。

复发事件数据虽属纵向数据, 但由于事件重复发生的时间有内在的次序和相依性, 个体之间又有异质性的存在, 因此又有其特殊结构。复发时间数据往往包含删失数据, 而且删失时间可能与事件重复发生的时间也具有相依性等等。介于以上原因, 对复发事件数据的建模和统计推断变得较为困难(Dai等, 2009)。尽管如此, 由于复发事件数据结构自身的特性和应用前景, 其统计分析已经引起了各学科门类尤其是医学界的高度重视, 对其的研究不仅具有重要的理论意义而且具有广阔及深远的实际应用价值(Liu等, 2011)。

*国家自然科学基金(11401146, 11361022, 11471135)、海南省自然科学基金(20151006, 20151010)、海南省教育厅高等学校科学研究项目(Hjkj2013-16)、海南师范大学博士启动基金项目和海南师范大学研究生创新科研项目(Hsyx2014-34)资助。

*通讯作者, E-mail: nl_hainnu@163.com.

本文2014年5月14日收到, 2014年11月21日收到修改稿。

doi: 10.3969/j.issn.1001-4268.2015.05.007

在复发事件数据的分析中, 人们常常关心协变量对事件复发率的影响. 记 $N^*(t)$ 为在区间 $[0, t]$ 上所发生的事件次数, 若 $E\{dN^*(t)\} = \mu(t)dt$, 则 $\mu(t)$ 称为 $N^*(t)$ 的比率函数. 其中 $dN^*(t) = N^*\{(t + dt)^-\} - N^*(t^-)$ 为 N^* 在小区间 $[t, t + dt)$ 上的增量(当 $dt \rightarrow 0$). 假设在一定时间范围内, 实验个体数目为 n , 且每个个体之间是相互独立的. 记 $N_i^*(t) = \int_0^t dN_i^*(s)$ 为个体 i ($i = 1, 2, \dots, n$) 在时刻 t 所经历目标事件的次数. 在实际研究及应用中, 考察个体不可能无限时间进行下去, 往往只能在有限删失时间内观察事件过程, 即 $N_i^*(\cdot)$ 不可能完全观测. 那么我们记第 i 个个体的删失时间为 C_i , 可观察的复发事件数据为 $N_i(t) = \int_0^t I(C_i \geq s) dN_i^*(s)$. 设 $X_i(t)$ 表示协变量, 并假设在给定 $X_i(\cdot)$ 条件下, 删失时间 C_i 与 $N_i^*(t)$ 相互独立, 即 $E\{dN_i^*(t)|X_i(t), C_i \geq t\} = E[dN_i^*(t)|X_i(t)]$, 对于乘性比率模型在给定 $X_i(\cdot)$ 的条件下复发事件的比率函数假定满足

$$E\{dN_i^*(t)|X_i(t)\} = \lambda_0(t) \exp\{\gamma_0' X_i(t)\} dt, \quad (1.1)$$

这里 γ_0 为未知的回归参数向量, $\lambda_0(\cdot)$ 未知的基本比率函数. 关于模型(1.1)的研究: Pepe和Cai (1993), 利用首次事件之后的复发率函数, 给出了 γ_0' 估计大样本理论, 但是证明方法较为粗糙; Lawless等(1995, 1997)研究 γ_0' 和 $\Lambda_0(t)$ 的估计, 且假设时间是离散时, 研究估计的渐近性; Lin等(2000), 改进了Lawless等(1995, 1997)的估计, 构造出均值函数的置信区间.

然而, 随着科技及医疗水平的不断提高, 对于一些反复发生且被认为是不能被治愈的疾病, 近年来发现有疑似治愈个体的存在, 这部分被治愈的病人或者被称为长期存活个体(long-term survivor) (Maller和Zhou, 1994). 这些治愈者在一定时刻后疾病就不复发了, 从整体上看, 很大程度地降低了疾病的复发率. 若对含有长期生存者的资料数据仍然采用传统分析方法, 就不太合适, 有可能得出与实际问题解释不符的结论(Lai, 2009).

在生存分析数据中, 已有学者对治愈个体做了研究. Boag (1949)最早提出了混合治愈模型的概念, 并对治愈率和未治愈病人的存活率给出了估计方法. Farewell (1982)推广了Cox比率模型, 提出在混合治愈模型中用Cox模型来刻画生存函数部分, 以此解释Cox模型中可能存在的治愈现象. Farewell对治愈个体和未治愈个体分别采用Logistic回归和Cox比率模型共同刻画所定义的风险函数. Kuk和Chen (1992)和Taylor (1995)利用半参数方法去估计混合模型中的回归参数和治愈概率. 因为在实际中不能观测到Kuk和Chen (1992)所定义的治愈变量, 所以Sy和Taylor (2000)提出用期望最大化(EM)算法将治愈变量看成潜在变量, 逐步迭代地估计参数. 但是, 在复发事件数据中, 研究个体多数被假设为一直反复发病, 没有治愈个体的存在. 虽有少数文献, 如Cook和Lawless (2007)提及用Logistic模型对治愈概率进行建模, 但并未详细研究. 因此在复发事件数据中探索治愈个体的新模型迫在眉睫.

1.1 治愈个体的识别

由于治愈个体的生存时间肯定大于试验观测时间(也会被删失掉), 所以怎样从复发事

件的删失数据中识别出这些治愈个体就成了一一个非常重要的任务. 类似的问题在一般的生存数据中被提到, 依据Maller等(1995, 1996)的研究成果, 当试验的观测时间(follow-up)足够长时, 在实际应用中, 通过Kaplan-Meier生存函数曲线去大体判断是否存在治愈个体就成为比较流行且有效的检验手段(Lai, 2009). 因此我们用累计发病次数曲线类似生存分析里的Kaplan-Meier函数曲线, 在复发数据中对识别治愈者做如下标记, 如图1示, 以横轴表示观察时间, 纵轴表示个体*i*的累积发病次数, 以累积发病次数趋于稳定来判定长期生存者的存在.

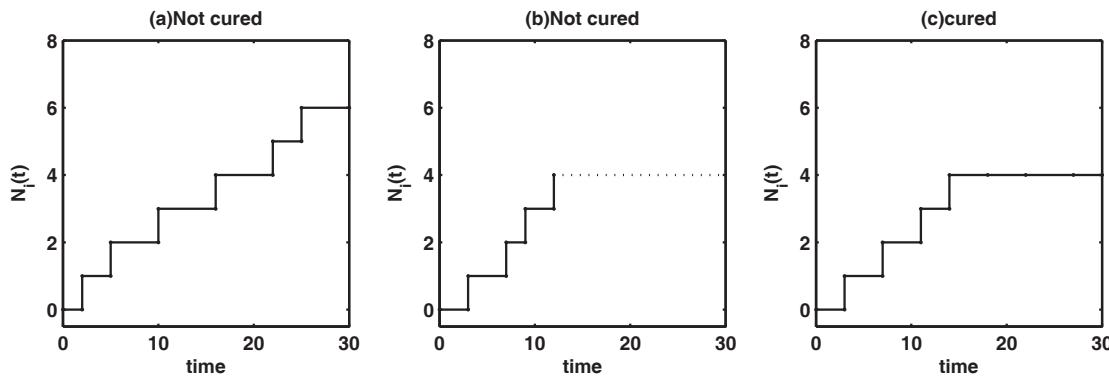


图1 判定治愈者是否存在

图1(a): 表明个体*i*为非治愈者. 因为, 曲线累积发病次数一直上升.

图1(b): 表明个体*i*为非治愈者. 因为, 数据中虽有一段趋于稳定的平台, 但这是由于个体*i*随机删失后, 此后数据收集不到, 才使得累积发病次数不变. 因此, 不认为它是治愈者.

图1(c): 表明个体*i*为治愈者. 因为, 在数据未随机删失的情况下累积发病次数有一段趋于稳定的平台.

本文结构如下: 第2节第一部分, 在原来半参数比例模型(1.1)回归事件的复发率基础上, 我们利用Logistic模型回归治愈率部分, 进而提出一类带有治愈率的半参数比率的新模型(2.1), 来刻画协变量对事件复发率的影响. 第2节第二部分, 利用估计方程的思想, 讨论我们模型中回归参数的估计. 第3节研究这些估计的渐近性. 第4节借助计算机Matlab软件, 通过一些数值模拟来验证所提出的方法是有效的. 第5节把所得的模型和方法应用到一组实际的纵向复发膀胱癌数据中. 本文涉及到的证明皆在附录中给出.

§2. 模型和估计方法

2.1 含治愈个体的半参数比率模型

假设在一定时间内, 我们的实验个体*i*总共个有*n*个($i = 1, 2, \dots, n$), 并且每个个体之间是相互独立的. $W(t)$ 和 $X(t)$ 都表示与复发率有关协变量, $W(t)$ 通过治愈率间接影响复发

率, $X(t)$ 通过经典的比率模型(1.1)直接影响复发率, 它们是不同的两部分协变量. $N^*(t)$ 为在时间 t 或之前拥有事件次数这样的计数过程. 但是在有限时间内, $N^*(t)$ 不能够完全观察, 记 C 为删失时间, 可设 $N(t) = N^*(t \wedge C)$, 其中 $a \wedge b = \min(a, b)$, $N(t)$ 则为可观测到的数据.

用 π 表示个体被治愈的概率, 而 $(1 - \pi)$ 表示个体未被治愈的概率. 当研究整体中含有是治愈者时, 我们考虑如下, 含治愈者的半参数比率回归模型

$$\mathbb{E}[dN_i^*(t)|W_i(t), X_i(t)] = [1 - \pi(W_i(t))] \exp\{\gamma_0' X_i(t)\} \lambda_0(t) dt, \quad (2.1)$$

$$\pi(W_i(t)) = \frac{\exp\{\beta_0' W_i(t)\}}{1 + \exp\{\beta_0' W_i(t)\}}. \quad (2.2)$$

我们将与复发率有关协变量分为两类, $W(t)$ 表示一类有可能改善患者疾病的因素, 如: 接受药物治疗, 接受化疗放疗, 手术仪器精密等. $X(t)$ 则表示一类有可能加重患者疾病复发的因素, 如患者的家族病史, 肿瘤扩散转移, 高龄, 酗酒等. $W(t)$ 这类协变量代入 Logistic 模型(2.2)得到治愈概率, 再结合(2.1)模型 $X(t)$ 这类协变量, 共同影响疾病的复发率. 其中 β_0, γ_0 为未知的回归参数向量, 分别表示协变量 $W_i(t)$ 与 $X_i(t)$ 与对复发事件比率的影响, $\lambda_0(\cdot)$ 是未知的基本比率函数.

2.2 模型的估计方法

在复发事件数据下, 可观测的数据是 $\{N_i(\cdot), X_i(\cdot), W_i(\cdot), C_i\}$ ($i = 1, 2, \dots, n$) 和指示变量 $Y_i(t) = I(C_i \geq t)$. 我们定义如下过程:

$$M_i(t; \theta_0) = N_i(t) - \int_0^t Y_i(s)[(1 + \exp\{\beta_0' W_i(s)\})^{-1} \exp\{\gamma_0' X_i(s)\}] d\Lambda_0(s).$$

我们可以知道 $M_i(t; \theta_0)$ 是均值为零的随机过程(附录给出证明). 为了方便书写, 记 $Z(t) = (W(t)', X(t)')'$, 对于给定的 $\theta = \{\beta', \gamma'\}'$, 因此, 我们很自然想到用下列下列式子来估计 $\Lambda_0(t)$:

$$\hat{\Lambda}_0(t; \theta) = \int_0^t \left[\sum_{i=1}^n dN_i(s) \right] / \left[\sum_{i=1}^n Y_i(s)(1 + \exp\{\beta' W_i(s)\})^{-1} \exp\{\gamma' X_i(s)\} \right]. \quad (2.3)$$

为了估计 $\theta_0 = \{\beta_0', \gamma_0'\}'$, 应用估计方程的思想, 我们可以利用下面方程:

$$\sum_{i=1}^n \int_0^\tau Z_i(s) \{dN_i(s) - Y_i(s)[(1 + \exp\{\beta' W_i(s)\})^{-1} \exp\{\gamma' X_i(s)\}] d\Lambda_0(s; \theta)\} = 0, \quad (2.4)$$

定义

$$S_z(t; \theta) = n^{-1} \sum_{i=1}^n Y_i(t)(1 + \exp\{\beta' W_i(t)\})^{-1} \exp\{\gamma' X_i(t)\} Z_i(t),$$

$$S_0(t; \theta) = n^{-1} \sum_{i=1}^n Y_i(t)(1 + \exp\{\beta' W_i(t)\})^{-1} \exp\{\gamma' X_i(t)\},$$

$$\bar{Z}(t; \theta) = \frac{S_z(t; \theta)}{S_0(t; \theta)},$$

且记 $s_z(t; \theta)$, $s_0(t; \theta)$, $\bar{z}(t; \theta)$ 分别为 $S_z(t; \theta)$, $S_0(t; \theta)$, $\bar{Z}(t; \theta)$ 的极限, 将式(2.3)代入估计方程(2.4), 有

$$\sum_{i=1}^n \int_0^\tau \{Z_i(s) - \bar{Z}(s; \theta)\} dN_i(s) = 0. \quad (2.5)$$

记方程(2.5)的解为 $\hat{\theta} = (\hat{\beta}', \hat{\gamma}')'$, 则 $\Lambda_0(t)$ 的估计为 $\hat{\Lambda}_0(t; \hat{\theta})$. 下面我们展示这些估计的大样本性质.

§3. 统计性质

在研究渐近性质之前, 我们先假设下列条件满足, 这些条件是复发数据所要求的一般结构形式:

- (Δ1) $\{N_i(\cdot), Y_i(\cdot), W_i(\cdot), X_i(\cdot)\}$ ($i = 1, 2, \dots, n$) 独立同分布.
- (Δ2) $P(Y_i(\tau) = 1) > 0$, 且几乎处处 $N_i(\tau) < \eta < \infty$, $i = 1, 2, \dots, n$, 其中 η 为常数.
- (Δ3) $W_i(\cdot)$ 和 $X_i(\cdot)$ 每一个分量函数的总变量分别以一个非随机常数为界.
- (Δ4) A 为非奇异矩阵, 其中

$$A = E \left\{ \int_0^\tau \{Z_i(s) - \bar{z}(s; \theta_0)\}^{\otimes 2} Y_i(s) [(1 + \exp\{\beta'_0 W_i(s)\})^{-1} \exp\{\gamma' X_i(s)\}] d\Lambda_0(s) \right\}, \quad (3.1)$$

其中, E 表示数学期望, 对于向量 $b^{\otimes 2} = bb'$, 记

$$U(t; \theta) = \sum_{i=1}^n \int_0^t \{Z_i(s) - \bar{Z}(s; \theta)\} dN_i(s).$$

接下来我们给出这些前面估计的渐近性质, 其证明皆在附录中给出.

性质 3.1 ($U(t; \theta)$ 的渐近性质) 如果条件(Δ1)–(Δ4)满足, 则 $n^{-1/2} U(t; \theta)$ 渐近服从均值为0, 协方差矩阵为 $\xi(s, t)$ 的正态分布, 其中

$$\xi(s, t) = E \left[\int_0^s \{Z_i(u) - \bar{z}(u; \theta_0)\} dM_i(u; \theta_0) \int_0^t \{Z_i(v) - \bar{z}(v; \theta_0)\}' dM_i(v; \theta_0) \right]. \quad (3.2)$$

由一致强大数定律和Lin等(2000)中的引理1可知, $\xi(s, t)$ 的相合估计为 $\hat{\xi}(s, t)$, 这里,

$$\begin{aligned} \hat{\xi}(s, t) &= n^{-1} \sum_{i=1}^n \left[\int_0^s \{Z_i(u) - \bar{Z}(u; \hat{\theta})\} d\hat{M}_i(u; \hat{\theta}) \int_0^t \{Z_i(v) - \bar{Z}(v; \hat{\theta})\}' d\hat{M}_i(v; \hat{\theta}) \right], \\ \hat{M}_i(t; \hat{\theta}_0) &= N_i(t) - \int_0^t Y_i(s) [(1 + \exp\{\hat{\beta}' W_i(s)\})^{-1} \exp\{\hat{\gamma}' X_i(s)\}] d\hat{\Lambda}_0(s; \hat{\theta}). \end{aligned}$$

性质 3.2 ($\hat{\theta}$ 的渐近性质) 如果条件(Δ1)–(Δ4)满足, 则 $\hat{\theta}$ 几乎处处一致收敛于 θ_0 , 且 $n^{1/2}\{\hat{\theta} - \theta_0\}$ 渐近服从均值为0, 协方差矩阵为 $A^{-1}\Sigma A^{-1}$ 的高斯过程, 其中 A 的定义在

$(\Delta 4)$ 中给出, 且 $\Sigma = \xi(\tau, \tau)$. 未知量被估计量替代, 便有 $n^{1/2}\{\hat{\theta} - \theta_0\}$ 渐近协方差的相合估计 $\hat{A}^{-1}\hat{\Sigma}\hat{A}^{-1}$, 这里, $\hat{\Sigma} = \hat{\xi}(\tau, \tau)$,

$$\hat{A} = n^{-1} \sum_{i=1}^n \left\{ \int_0^\tau \{Z_i(s) - \bar{Z}(s; \hat{\theta})\}^{\otimes 2} Y_i(s) [(1 + \exp\{\hat{\beta}' W_i(s)\})^{-1} \exp\{\hat{\gamma}' X_i(s)\} d\hat{\Lambda}_0(s)] \right\}.$$

性质 3.3 ($\hat{\Lambda}_0(t)$ 的渐近性质) 如果条件 $(\Delta 1)$ – $(\Delta 4)$ 满足, 则 $\hat{\Lambda}_0(t)$ 关于 $t \in [0, t]$ 几乎处处一致收敛于 $\Lambda_0(t)$, 且 $n^{1/2}\{\hat{\Lambda}_0(t; \hat{\theta}) - \Lambda_0(t)\}$ 渐近服从均值为0, 协方差矩阵为 $\Gamma(s, t) = E\{\Phi_i(s)\Phi_i(t)\}$ 的高斯过程, 其中,

$$\Phi_i(t) = \int_0^t \frac{dM_i(u; \theta_0)}{s_0(u; \theta_0)} - B(t; \theta_0)' A^{-1} \int_0^\tau Y_i(u) \{Z_i(u) - \bar{z}(u; \theta_0)\}' dM_i(u; \theta_0), \quad (3.3)$$

$$B(t; \theta) = \int_0^t \bar{z}(u; \theta) \lambda_0(u) du. \quad (3.4)$$

用估计量代替未知量, 我们可以得到协方差函数 $\Gamma(s, t)$ 的一个相合估计为

$$\hat{\Gamma}(s, t) = n^{-1} \sum_{i=1}^n \hat{\Phi}_i(s) \hat{\Phi}_i(t)', \quad (3.5)$$

其中

$$\hat{\Phi}_i(t) = \int_0^t \frac{d\hat{M}_i(u; \hat{\theta})}{S_0(u; \hat{\theta})} - \hat{B}(t; \hat{\theta})' \hat{A}^{-1} \int_0^\tau Y_i(u) \{Z_i(u) - \bar{Z}(u; \hat{\theta})\}' d\hat{M}_i(u; \hat{\theta}),$$

$$\hat{B}(t; \hat{\theta}) = \int_0^t \bar{Z}(u; \hat{\theta}) d\hat{\Lambda}_0(u; \hat{\theta}).$$

§4. 数值模拟

为了验证所提出的估计方法在有限样本容量下的表现, 本节进行一些数值模拟. 这里我们主要集中在 β_0 和 γ_0 的估计上. 在下面的模拟中, 受Li等(2010)的启发, 我们也类似地考虑: 协变量 X_i 和 W_i 都为Bernoulli随机变量, 其成功概率为0.5, 并且最长的跟踪时间为 τ , 删失时间 C_i 由均匀分布 $(\tau/2, \tau)$ 产生. 设 $\lambda_0(t) = c/\tau$, c 是一个常数, 且 $N_i(t)$ 是Poisson过程, 其均值函数为

$$\Lambda(C_i|Z_i) = \Lambda_0(C_i)(1 + \exp\{\beta W_i\})^{-1} \exp\{\gamma X_i\} = c C_i (1 + \exp\{\beta W_i\})^{-1} \exp\{\gamma X_i\}/\tau,$$

并且观测时间 $(t_{i1}, t_{i2}, \dots, t_{ik_i})$ 是长度为 k_i 的次序统计量, k_i 来自均匀分布 $U(0, C_i)$.

表1给出了 (β_0, γ_0) 取不同的 $(0, 0)$, $(0.1, 0)$, $(0, 0.1)$ 和 $(0.1, 0.1)$, τ 为1或2时, 样本量 n 变换100和200, 其中10%是治愈个体. 利用重复迭代1000次得到模拟结果. 表中包含有估计的Bias, SSE, SEE和CP, 其中Bias为估计量 $\hat{\theta}$ 的样本均值减去真实值, SSE为 $\hat{\theta}$ 的样本标准差, SEE为 $\hat{\theta}$ 的标准差估计的平均值, CP为 θ_0 的95%经验覆盖率. 从这些结果可以看: Bias接

近, 说明点估计是无偏的; SSE与SEE彼此接近, 则我们的方差估计起到了很好的效果; CP这一经验覆盖率也十分接近我们提出95%的置信水平. 综上, 我们提出的估计是渐近无偏的, 其方差估计和覆盖概率是合理的.

表1 含治愈个体复发事件模型的仿真结果

	$n = 100$		$n = 100$		$n = 200$		$n = 200$	
	$\tau = 1$		$\tau = 2$		$\tau = 1$		$\tau = 2$	
	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\beta}$
$\theta = (0, 0)$								
BIAS	0.0020	-0.0123	0.0016	-0.0053	-0.0034	-0.0063	-0.0054	-0.0063
SSE	0.1078	0.2359	0.1075	0.2324	0.0794	0.1573	0.0789	0.1593
SEE	0.1386	0.2428	0.1394	0.2438	0.0649	0.1665	0.0650	0.1665
CP	0.9140	0.9260	0.9240	0.9320	0.9410	0.9430	0.9260	0.9380
$\theta = (0, 0.1)$								
BIAS	-0.0044	-0.0532	-0.0048	-0.0604	-0.0041	-0.0460	-0.0057	-0.0456
SSE	0.1126	0.2181	0.1090	0.2187	0.0755	0.1442	0.0630	0.1725
SEE	0.1495	0.2548	0.1478	0.2534	0.0828	0.1745	0.0728	0.1747
CP	0.9280	0.9360	0.9280	0.9400	0.9330	0.9470	0.9420	0.9500
$\theta = (0.1, 0)$								
BIAS	-0.0399	-0.0154	-0.0372	-0.0134	-0.0358	-0.0042	-0.0380	-0.0132
SSE	0.1105	0.2283	0.1111	0.2320	0.0785	0.1639	0.0786	0.1618
SEE	0.1380	0.2428	0.1383	0.2427	0.0657	0.1672	0.0645	0.1660
CP	0.9310	0.9340	0.9230	0.9290	0.9260	0.9330	0.9270	0.9340
$\theta = (0.1, 0.1)$								
BIAS	-0.0408	-0.0540	-0.0417	-0.0613	-0.0415	-0.0565	-0.0433	-0.0575
SSE	0.1078	0.2220	0.1133	0.2218	0.0766	0.1559	0.0686	0.1579
SEE	0.1500	0.2549	0.1486	0.2532	0.0781	0.1739	0.0827	0.1579
CP	0.9270	0.9310	0.9170	0.9290	0.9150	0.9270	0.9240	0.9320

为了验证在有治愈个体存在的情况下, 我们所提出的模型和估计方法与原来的有所不同, 这里通过数值模拟来比较分析. 把我们所提出的模型和估计方法与Lin等(2000)中的模型与方法都用到相同的复发事件数据中, 计算出 γ_0 的偏差, 从而比较它们的优劣. 表2中两个模型的参数设置除了我们模型对样本量n中有10%的治愈个体及多考虑协变量 $W(t)$ 对复发率的影响, 而原来模型则忽略这些, 其它的二者均相同(都与表1的设置相同). 在表格中, BIAS和SSE分别表示所提出估计的偏差和估计的标准差, BIAS*和SSE*分别表示

由Lin等(2000)中给出估计的偏差和估计的标准差. 从数据对比的结果来看, 在相同设定的情况下, BIAS与SSE分别比BIAS*与SSE*的绝对值更小, 这说明在考虑协变量通过治愈率影响复发率时, 我们对未知参数所提出的估计比Lin等(2000)方法更为合理.

表2 所提模型与忽略治愈个体模型中 γ 的比较

	$n = 100$				$n = 200$			
	BIAS	BIAS*	SSE	SSE*	BIAS	BIAS*	SSE	SSE*
$\tau = 1$								
$\gamma = -0.1$	-0.0054	-0.1176	0.2238	0.2972	-0.0179	-0.2085	0.1749	0.3623
$\gamma = 0$	-0.0072	-0.1255	0.3152	0.3371	0.0019	-0.1575	0.2382	0.3722
$\gamma = 0.1$	0.0241	0.2184	0.2871	0.3865	0.0197	-0.1722	0.2247	0.4174
$\tau = 2$								
$\gamma = -0.1$	-0.0137	0.1452	0.2774	0.4012	-0.0109	-0.1528	0.2492	0.3458
$\gamma = 0$	0.0013	0.2015	0.1982	0.3208	0.0027	0.1346	0.2587	0.3635
$\gamma = 0.1$	0.0126	-0.1528	0.2498	0.4054	0.0265	0.2343	0.3481	0.4723

§5. 应用

本节把我们所提出的模型和估计方法应用到Byar (1980)提供的美国退伍军人管理局泌尿学研究组(The Veterans Administration Cooperation Urological Group)关于膀胱癌反复治疗的临床实验数据. 在Byar (1980)这部分研究中, 所有病人都是有膀胱癌特征的一组纵向膀胱癌数据. 该数据包括85个非治愈个体膀胱癌病人, 对每个病人, 可观察的信息包括入院诊断时间或观察时间以及每两次诊断之间新生长的肿瘤个数, 且每次发现的肿瘤都会被完全切除. 由于药品噻替派治疗对复发率的影响已经在Liu等(2011), Li等(2010), Sun等(2007)和He等(2009)等文献中分析论证了, 因此这里我们就研究初始肿瘤个数和累计肿瘤个数对复发率的影响. 定义两个协变量: 一个是病人进入研究之前的初始肿瘤个数; 另一个是累计肿瘤个数. 对所有病人而言, 最长的观察时间是53个月(Liu等, 2011), 我们这里的主要目的是研究肿瘤复发与我们记录的协变量的影响. 定义 $N_i(t)$ 为病人截止到时间 t 处, 去医院并确认有膀胱肿瘤的累计次数. 记 $X_i(t)$ 为病人的初始肿瘤个数, $W_i(t)$ 为病人在时间 t 处的累计肿瘤数目. 应用我们的方法, 得到表3.

表3 膀胱癌数据的回归参数估计

	回归系数估计值	标准误差	95%的置信区间	P值
初始个数	-0.0070	0.0450	(-0.0952, 0.0813)	0.8773
累计个数	-1.4487	0.1589	(-1.7602, -1.1372)	0.0000

表格结果显示, 在考虑整体的膀胱癌复发过程中存在治愈个体的情况下, 我们处理的结果说明, 初始个数与膀胱癌复发率影响不显著; 而累计肿瘤个数对肿瘤的复发率有显著影响, 即累计肿瘤个数越多, 治愈率越低, 肿瘤的复发率越高. 从数据上看, 这一结果有统计学意义; 从医学角度上看, 这一结果有与实际相符.

§6. 总结与展望

本文认为, 对复发事件进行回归分析时, 可能会涉及到个体在发病一定次数后不再发病的情况, 从而影响疾病的复发率. 针对该问题, 我们提出一类含治愈个体的半参数回归模型. 该模型假设有治愈个体的存在, 这样不仅可以提供更多的信息处理疾病复发率的影响因素; 而且, 随着医疗技术的不断提高, 复发事件中治愈者的存在与客观实际更吻合. 此外, 我们针对模型发展了估计方程, 并用数值仿真表明估计过程的可行性. 最后将模型运用到一组真实的数据中.

如前所述, 本文重点讨论协变量不仅通过比率函数, 还通过Logistic函数影响事件的复发率, 但由于我们没有给出检验模型的公式, 这使得忽略治愈率会对复发率产生偏差, 这样的结果难以察觉; 模型的运用中, 受到数据资料的限制, 我们对协变量的选取和分析不全面, 这可能导致我们提出的模型在矫正因素时有遗漏或矫枉过正之嫌, 但就模型结果而言, 累计肿瘤数对疾病的复发率有影响是符合实际的.

致谢 作者衷心感谢编委和审稿专家对本文提出的宝贵意见.

附录: 主要的理论证明

过程 $M_i(t; \theta_0)$ 的证明: 由条件期望知

$$\begin{aligned} E\{dN_i(t)\} &= E\{E[dN_i(t)|Z_i(t)]\} \\ &= E\{E[Y_i(t)dN_i^*(t)|Z_i(t)]\} \\ &= E\{E[Y_i(t)|Z_i(t)] E[dN_i^*(t)|Z_i(t)]\} \\ &= E\{[Y_i(t)][(1 - \pi(W_i(t))) \exp\{\gamma'_0 X_i(t)\} \lambda_0(t) dt]\} \\ &= E\{Y_i(t)(1 + \exp\{\beta'_0 W_i(t)\})^{-1} \exp\{\gamma'_0 X_i(t)\} \lambda_0(t) dt\}, \end{aligned}$$

所以 $E\{M_i(t; \theta_0)\} = 0$, 即 $M_i(t; \theta_0)$ 是均值为零的随机过程. \square

性质3.1的证明: 由于

$$\sum_{i=1}^n \{Z_i(s) - \bar{Z}(s; \theta)\} Y_i(s) (1 + \exp\{\beta' W_i(s)\})^{-1} \exp\{\gamma' X_i(s)\} = 0,$$

故 $U(t; \theta_0)$ 可以写成

$$U(t; \theta_0) = \sum_{i=1}^n \int_0^t \{Z_i(s) - \bar{Z}(s; \theta_0)\} dM_i(s; \theta_0) = \bar{M}_Z(t) - \int_0^t \bar{Z}(s; \theta_0) d\bar{M}(s),$$

其中

$$\bar{M}(t) = \sum_{i=1}^n M_i(t; \theta_0), \quad \bar{M}_Z(t) = \sum_{i=1}^n \int_0^t Z_i(s) dM_i(s; \theta_0).$$

类似于Lin等(2000)中附录A.2的证明, 我们可以知道 $n^{-1/2}U(t; \theta_0)$ 是弱收敛的, 并且它的极限分布的均值为0, 协方差函数在式(3.2)中给出. \square

性质3.2的证明: 这里我们先证明 $\hat{\theta}$ 的存在性, 唯一性及相合性.

记 $\hat{A}(\theta) = -n^{-1}\partial U(\tau; \theta)/\partial\theta'$. 由于

$$U(\tau; \theta) = \sum_{i=1}^n \int_0^\tau \{Z_i(s) - \bar{Z}(s; \theta)\} dM_i(s; \theta),$$

则有

$$\begin{aligned} A^*(\theta) &= n^{-1} \sum_{i=1}^n \int_0^\tau \{Z_i(s) - \bar{Z}(s; \theta)\}^{\otimes 2} Y_i(s) [(1 + \exp\{\beta' W_i(s)\})^{-1} \exp\{\gamma' X_i(s)\}] d\Lambda_0(s) \\ &\quad + n^{-1} \sum_{i=1}^n \int_0^\tau \frac{\partial \bar{Z}(s; \theta)}{\partial\theta'} dM_i(s; \theta). \end{aligned} \quad (7.1)$$

由一致强大数定理我们知道, $A^*(\theta)$ 关于 θ 几乎处处一致收敛到一个非随机的 $A(\theta)$ 函数, 且 $A(\theta_0) = A$, 其中 A 在($\Delta 4$)给出, 且由条件($\Delta 4$)知 $A(\theta_0)$ 是非奇异的, $A(\theta)$ 关于 θ 是连续的. 那么由 $A^*(\theta)$ 的一致收敛及连续性, $A(\theta_0)$ 的非奇异性知, 有 θ_0 的一个小范围, 在这个范围内, 任意大的 n , $A^*(\theta)$ 的特征根有界且非零. 类似于Lin和Ying(1995)中定理2.2的证明, 可知 $\hat{\theta}$ 存在唯一, 且为 θ_0 的强相合估计.

现在我们把目光放到 $\hat{\theta}$ 的渐近正态性上.

用 $U(\tau; \hat{\theta})$ 的级数展开, $A^*(\theta_0)$ 一致收敛性, $\hat{\theta}$ 的相合性以及 A 的非奇异性, 我们可以知道

$$n^{1/2}(\hat{\theta} - \theta_0) = A^{-1}n^{-1/2}U(\tau; \theta_0) + o_p(1). \quad (7.2)$$

故由性质3.1有, $n^{1/2}(\hat{\theta} - \theta_0)$ 依分布收敛到一个均值为0, 协方差矩阵是 $A^{-1}\Sigma A^{-1}$ 的正态随机向量, 并且 $\hat{A}^{-1}\hat{\Sigma}\hat{A}^{-1}$ 是其协方差矩阵的相合估计. \square

性质3.3的证明: 我们首先证明 $\hat{\Lambda}_0(t)$ 的相合性. 由微分中值定理有

$$\hat{\Lambda}_0(t) - \Lambda_0(t) = n^{-1} \sum_{i=1}^n \int_0^t \frac{dM_i(u; \theta_0)}{S_0(u; \hat{\theta})} - \int_0^t \frac{S_Z(u; \theta^*)'}{S_0(u; \hat{\theta})} du_i(\hat{\theta}, \theta_0), \quad (7.3)$$

其中 θ^* 在 $\hat{\theta}$ 和 θ_0 之间. 由一致强大数定理和Lin等(2000)中的引理1知, 式(7.3)等号右边第一部分一致收敛到0. 由条件($\Delta 1$)–($\Delta 4$)知, 式(7.3)最后一部分中积分几乎处处一致有界. 因

此由 $\hat{\theta}$ 的相合性得, 式(7.3)最后一部分一致收敛到0. 因此 $\hat{\Lambda}_0(t)$ 在 $t \in [0, \tau]$ 内几乎处处一致收敛于 $\Lambda_0(t)$.

接着证明 $\hat{\Lambda}_0(t)$ 的弱收敛性. 我们发现

$$n^{1/2}\{\hat{\Lambda}_0(t) - \Lambda_0(t)\} = n^{1/2}\{\hat{\Lambda}_0(t; \hat{\theta}) - \hat{\Lambda}_0(t; \theta_0)\} + n^{1/2}\{\hat{\Lambda}_0(t; \theta) - \Lambda_0(t)\}. \quad (7.4)$$

利用 $\hat{\Lambda}_0(t)$ 在 θ_0 处级数展开, 式(7.4)等号右边第一部分可以写为

$$n^{1/2}\{\hat{\Lambda}_0(t; \hat{\theta}) - \hat{\Lambda}_0(t; \theta_0)\} = \frac{\partial \hat{\Lambda}_0(t; \theta_0)}{\partial \theta'} n^{1/2}(\hat{\theta} - \theta_0) + o_p(n^{1/2}\|\hat{\theta} - \theta_0\|).$$

由Lin等(2000)中的引理1以及一致强大数定理知, 几乎处处有

$$\sup_{0 \leq t \leq \tau} \left\| \frac{\partial \hat{\Lambda}_0(t; \theta_0)}{\partial \theta} + B(t; \theta_0) \right\| \rightarrow 0,$$

其中 $B(t; \theta_0)$ 由式(3.4)给出. 由性质3.1和式(7.2)有

$$n^{1/2}(\hat{\theta} - \theta_0) = A^{-1}n^{-1/2} \sum_{i=1}^n \int_0^t Y_i(u)\{Z_i(u) - \bar{z}(u; \theta_0)\}dM_i(u; \theta_0) + o_p(1).$$

则对于 t 一致有

$$\begin{aligned} & n^{1/2}\{\hat{\Lambda}_0(t; \hat{\theta}) - \hat{\Lambda}_0(t; \theta_0)\} \\ &= -B(t; \theta_0)' A^{-1} n^{-1/2} \sum_{i=1}^n \int_0^t Y_i(u)\{Z_i(u) - \bar{z}(u; \theta_0)\}dM_i(u; \theta_0) + o_p(1). \end{aligned} \quad (7.5)$$

另一方面, 对于 $0 \leq t \leq \tau$ 可以得到

$$n^{1/2}\{\hat{\Lambda}_0(t; \theta_0) - \Lambda_0(t)\} = n^{-1/2} \sum_{i=1}^n \int_0^t \frac{dM_i(u; \theta_0)}{S_0(u; \theta_0)}.$$

于是利用Lin等(2000)中的引理1知, 对 t 一致有

$$n^{1/2}\{\hat{\Lambda}_0(t; \theta_0) - \Lambda_0(t)\} = n^{-1/2} \sum_{i=1}^n \int_0^t \frac{dM_i(u; \theta_0)}{s_0(u; \theta_0)} + o_p(1). \quad (7.6)$$

最后由式(7.4)–(7.6)知, 对 t 一致有

$$n^{1/2}\{\hat{\Lambda}_0(t; \hat{\theta}) - \Lambda_0(t)\} = n^{-1/2} \sum_{i=1}^n \Phi_i(t) + o_p(1). \quad (7.7)$$

由式(7.4)知, 对于固定 t , $n^{1/2}\{\hat{\Lambda}_0(t; \hat{\theta}) - \Lambda_0(t)\}$ 渐近于均值为0的独立同分布随机变量之和. 接着从多元中心极限我们发现, $n^{1/2}\{\hat{\Lambda}_0(t; \hat{\theta}) - \Lambda_0(t)\}$ 以有限维分布收敛到均值为零的高斯过程. 注意到 $\Phi_i(t)$ 中 $B(t; \theta_0)$ 是非随机的函数, 且 $\int_0^\tau Y_i(u)\{Z_i(u) - \bar{z}(u; \theta_0)\}dM_i(u; \theta_0)$ 与 t 无关, 由此可知 $\Phi_i(t)$ 中第二部分是胎紧的. 另一方面, 与性质3.1的证明相似, $\Phi_i(t)$ 中第一部分也是胎紧的. 那么 $\Phi_i(t)$ 也是胎紧的. 且 $n^{1/2}\{\hat{\Lambda}_0(t; \hat{\theta}) - \Lambda_0(t)\}$ 弱收敛到一个均值为0的高斯过程, 且在 (s, t) 处的协方差函数为 $\Gamma(s, t) = E\{\Phi_i(s)\Phi_i(t)\}$. 同时 $\Gamma(s, t)$ 的相合估计在式(3.5)给出. \square

参考文献

- [1] Cox, D.R., Regression models and life-tables, *Journal of the Royal Statistical Society: Series B (Methodological)*, **34(2)**(1972), 187–220.
- [2] Andersen, P.K. and Gill, R.D., Cox's regression model for counting processes: a large sample study, *The Annals of Statistics*, **10(4)**(1982), 1100–1120.
- [3] Dai, J.J., Sun, L.Q. and Yang, Z.H., A general additive-multiplicative rates model for recurrent event data, *Science in China Series A: Mathematics*, **52(10)**(2009), 2257–2265.
- [4] Liu, H.B., Miao, R. and Sun, L.Q., Analysis of panel data with informative observation and censoring times under biased sampling, *Scientia Sinica Mathematica*, **41(4)**(2011), 365–376. (in Chinese)
- [5] Pepe, M.S. and Cai, J., Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates, *Journal of the American Statistical Association*, **88(423)**(1993), 811–820.
- [6] Lawless, J.F. and Nadeau, C., Some simple robust methods for the analysis of recurrent events, *Technometrics*, **37(2)**(1995), 158–168.
- [7] Lawless, J.F., Nadeau, C. and Cook, R.J., Analysis of mean and rate functions for recurrent events, In: *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, Volume 123 of the series *Lecture Notes in Statistics*, Springer, New York, 1997, 37–49.
- [8] Lin, D.Y., Wei, L.J., Yang, I. and Ying, Z., Semiparametric regression for the mean and rate functions of recurrent events, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62(4)**(2000), 711–730.
- [9] Maller, R.A. and Zhou, S., Testing for sufficient follow-up and outliers in survival data, *Journal of the American Statistical Association*, **89(428)**(1994), 1499–1506.
- [10] Lai, X., Extensions on long-term survivor model with random effects, Doctoral Dissertation of University of Science and Technology of China, 2009. (in Chinese)
- [11] Boag, J.W., Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *Journal of the Royal Statistical Society: Series B (Methodological)*, **11(1)**(1949), 15–53.
- [12] Farewell, V.T., The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics*, **38(4)**(1982), 1041–1046.
- [13] Kuk, A.Y.C. and Chen, C.H., A mixture model combining logistic regression with proportional hazards regression, *Biometrika*, **79(3)**(1992), 531–541.
- [14] Taylor, J.M.G., Semi-parametric estimation in failure time mixture models, *Biometrics*, **51(3)**(1995), 899–907.
- [15] Sy, J.P. and Taylor, J.M.G., Estimation in a Cox proportional hazards cure model, *Biometrics*, **56(1)**(2000), 227–236.
- [16] Cook, R.J. and Lawless, J.F., *The Statistical Analysis of Recurrent Events*, Springer, New York, 2007.
- [17] Maller, R.A. and Zhou, S., Testing for the presence of immune or cured individuals in censored survival data, *Biometrics*, **51(4)**(1995), 1197–1205.
- [18] Maller, R.A. and Zhou, X., *Survival Analysis with Long-Term Survivors*, John Wiley and Sons, New York, 1996.
- [19] Li, N., Sun, L.Q. and Sun, J.G., Semiparametric transformation models for panel count data with dependent observation process, *Statistics in Biosciences*, **2(2)**(2010), 191–210.

- [20] Byar, D.P., The veterans administration study of chemoprophylaxis for recurrent stage I bladder tumours: comparisons of placebo, pyridoxine and topical thiotepa, In: *Bladder Tumors and other Topics in Urological Oncology*, Volume 1 of the series *Ettore Majorana International Science Series*, Springer, New York, 1980, 363–370.
- [21] Sun, J.G., Sun, L.Q. and Liu, D.D., Regression analysis of longitudinal data in the presence of informative observation and censoring times, *Journal of the American Statistical Association*, **102**(480) (2007), 1397–1406.
- [22] He, X., Tong, X.W. and Sun, J.G., Semiparametric analysis of panel count data with correlated observation and follow-up times, *Lifetime Data Analysis*, **15**(2)(2009), 177–196.
- [23] Lin, D.Y. and Ying, Z.L., Semiparametric analysis of general additive-multiplicative hazard models for counting processes, *The Annals of Statistics*, **23**(5)(1995), 1712–1734.

Semiparametric Rate Model for Recurrent Event Data with Cure Rate

ZENG XIAOFENG CHEN CHUANZHONG LI NI

(School of Mathematics and Statistics, Hainan Normal University, Haikou, 571158)

Recurrent event data usually occur in long-term studies which concern recurrence rates of the disease. In studies of medical sciences, patients who have infected with the disease, like cancer, were conventionally regarded as impossible to be cured. However, with the development of medical sciences, recently those patients were found to be possibly recovered from the disease. The recurrence rate of the events, which is of primary interest, may be affected by the cure rate that may exist. Therefore, we proposed semiparametric statistical analysis for recurrent event data with subjects possibly being cured. In our approach, we present a proportional rate model for recurrence rate with the cure rate adjusted through a Logistic regression model, and develop some estimating equations for estimation of the regression parameters, with their large sample properties, including consistency and asymptotic normality established. Numerical studies under different settings were conducted for assessing the proposed methodology and the results suggest that they work well for practical situations. The approach is applied to a bladder cancer dataset which motivated our study.

Keywords: Recurrent event, proportional rate model, cure rate, Logistic model, estimating equation.

AMS Subject Classification: 62G05.