Vol. 41 (2021)
No. 3

数 学 杂 志
J. of Math. (PRC)

病例队列设计下加性风险回归分析及其在乳腺癌数据中的应用

杨 杰, 丁洁丽
(武汉大学数学与统计学院, 湖北 武汉, 430072)

摘 要: 本文研究了如何应用加性风险模型拟合由病例队列设计获取的生存数据的问题. 利用参数的一种加权估计方法并综述其渐近性质. 通过模拟研究获得了这种方法在有限样本量下的优良表现, 并评估了病例队列设计相较于简单随机抽样设计的有效性. 本文推广简单随机抽样至病例队列设计, 并将此种方法应用于一个实际的乳腺癌数据, 展示其在实际中的应用价值和前景.

关键词: 病例队列设计; 加性风险模型; 估计方程; 逆概率加权
MR(2010) **主题分类号:** 62N01; 62N02; 62N86 **中图分类号:** O212.1
文献标识码: A **文章编号:** 0255-7797(2021)03-0270-13

1 引言

生存数据应用在许多学科领域, 如生物医学、流行病学、公共卫生学、可靠性工程学、经济学和保险精算学等领域. 生存分析主要研究某特定事件的发生时间 (例如: 死亡时间、疾病发生时间、系统失效时间、索赔时间等等) 与重要影响因素和协变量之间的关联. 对于生存数据的研究, 比例风险模型是应用最为广泛的统计半参数模型之一 (Cox, 1972[1]; Andersen & Gill, 1982[2]; Lin, 1994[3]; Chen & Little, 1999[4]; Kalbfleisch & Prentice, 2002[5] 等等). 作为比例风险模型的一个重要替代, 加性风险模型假设基准风险函数与协变量效应之间具有一个加性结构. 在很多实际应用中, 加性风险模型往往能更好地拟合数据 (Lin & Ying, 1994[3]; Mckeague & Sasieni, 1994[6]). 而当加性风险模型和比例风险模型均能较好地拟合数据时, 加性风险模型的回归参数更容易解释其实际意义 (Lin & Ying, 1994[3]; Zeng & Cai, 2010[7]).

在许多大型队列研究中, 往往涉及对大量研究个体的长期追溯和随访. 当重要影响因素或协变量的采集非常昂贵时, 采用传统的简单随机抽样可能会导致实验过于昂贵而超过预算. 因此, 发展和研究能节约成本和提高效率的抽样机制具有非常重要的意义. 对于带有删失的生存数据, 病例队列设计 (Case-Cohort design) 是应用最为广泛的有偏抽样机制之一. 其主要机制是: 首先, 从全队列中随机地抽取一个子队列 (subcohort), 全队列中所有发生了感兴趣事件的个体称为病例 (case). 然后, 子队列和子队列之外所有的病例组成病例队列样本. 最后, 仅对病例队列样本中的个体进行昂贵协变量的采集和观测. 对病例队列设计相关统计方法的研究已有大量和广泛的工作, 比例风险模型 (Prentice, 1986[8]; Self & Prentice, 1988[9]; Chen & Lo, 1999[10]; Lin & Ying, 1993[11]; Kulich & Lin, 2004[12]), 加性风险模型 (Kulich & Lin, 2000[13]; Sun et al, 2004[14]), 加速失效模型 (Kong et al, 2004[15]; Lu & Tsiatis, 2006[16]).

*收稿日期: 2020-12-14 接收日期: 2021-01-14
基金项目: 国家自然科学基金资助 (11671310) .
作者简介: 杨杰 (1997–), 女, 吉林敦化, 硕士, 主要研究方向: 生存分析.
通讯作者: 丁洁丽.

近来, 基于多元失效时间的病例队列设计的研究越来越广泛 (Kong & Cai, 2009[17]; Kang et al, 2013[18]; Kim et al, 2018[19]; Maitra et al, 2020[20]).

本文主要探讨病例队列设计中加性风险模型下参数的统计推断方法和应用. 首先, 我们探讨如何应用加性风险模型拟合由病例队列设计获取的生存数据, 考虑参数的一种加权估计方法并综述其渐近理论. 然后, 重点研究这种病例队列设计下的分析方法在实际中的应用问题. 一方面, 我们编写了可实现这种统计分析方法的 R 软件应用程序. 通过模拟研究展示了这种方法在有限样本量下的优良表现, 并评估了病例队列设计相较于简单随机抽样设计的有效性. 另一方面, 我们应用该方法分析了一个实际的乳腺癌数据, 展示了其成本效益与应用价值. 乳腺癌是女性最常见的恶性肿瘤之一, 全世界每年约有 120 万女性患上乳腺癌, 约有 50 万女性患者死亡. 中国癌症年发病数为 160 万, 现症病人 200 多万. 乳腺癌占女性恶性肿瘤发病率第一位, 每年约有 4 万女性死于乳腺癌 [21]. 本文将基于来源于美国国家癌症研究所 (National Cancer Institute)[22] 的乳腺癌数据, 探索影响乳腺癌患者生存时间的影响因素. 首先, 我们基于全队列数据应用加性风险回归方法分析数据, 探索出了一些对乳腺癌患者生存时间有显著性影响的因素. 进一步, 应用病例队列设计抽取样本, 并基于病例队列样本进行加性风险回归分析. 结果表明, 病例队列设计仅用了较小的样本量就达到了与全队列研究几乎一致的结果. 当协变量的测量非常昂贵时, 病例队列设计可提高研究效率和节约成本.

本文的主要结构为: 第 2 节介绍病例队列设计下加性风险模型参数的加权估计方法并综述其渐近理论, 第 3 节为模拟计算, 第 4 节为乳腺癌数据的生存分析.

2 加性风险模型与病例队列抽样

假设一个大型队列研究包含 N 个独立的研究个体. 我们用 $\widetilde{T}_i$ 表示第 i 个个体的潜在失效时间, C_i 表示第 i 个个体的删失时间或随访时间 ($i = 1, \cdots, N$). 记观测时间为 $T_i = \min(\widetilde{T}_i, C_i)$, 右删失示性变量为 $\Delta_i = I(\widetilde{T}_i \le C_i)$. 用 $Y_i(t) = I(T_i \ge t)$ 表示风险过程, $N_i(t) = \Delta_i I(T_i \le t)$ 表示计数过程, 其中 $I(\cdot)$ 表示示性函数. 记 $Z_i(t)$ 为第 i 个个体在 t 时刻处的 p 维协变量. 记 τ 为事件终止时间.

我们考虑如下加性风险模型: 给定协变量 $Z_i(t)$ 时, 失效时间 $\widetilde{T}_i$ 的风险函数有如下形式:

$$\lambda(t|Z) = \lambda_0(t) + \beta' Z_i(t), \tag{2.1}$$

其中 $\lambda_0(t)$ 为形式未知的基准风险函数, β 为感兴趣的 p 维待估参数. 记基准累积风险函数为 $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$. 在队列研究中, 当对每个个体的协变量进行观测时, 对 β 和 $\Lambda_0(t)$ 的推断广泛地使用如下估计方程方法 (Lin & Ying, 1994)[3]:

$$\begin{aligned}
&\sum_{i=1}^{N} \int_0^{\tau} \left[dN_i(t) - Y_i(t) \left\{ d\Lambda_0(t) + \beta' Z_i(t)dt \right\} \right] = 0, \\
&\sum_{i=1}^{N} \int_0^{\tau} Z_i(t) \left[dN_i(t) - Y_i(t) \left\{ d\Lambda_0(t) + \beta' Z_i(t)dt \right\} \right] = 0.
\end{aligned} \tag{2.2}$$

记 $\widehat{\beta}$ 和 $\widehat{\Lambda}_0(t)$ 为上述估计方程组的解, 有如下形式:

$$\widehat{\beta} = \left[\sum_{i=1}^{N} \int_0^{\tau} Y_i(t) \left\{ Z_i(t) - \overline{Z}(t) \right\}^{\otimes 2} dt \right]^{-1} \left[\sum_{i=1}^{N} \int_0^{\tau} \left\{ Z_i(t) - \overline{Z}(t) \right\} dN_i(t) \right], \tag{2.3}$$

和

$$\widehat{\Lambda}_0(t)=\int_0^t\frac{\sum_{i=1}^N\left\{dN_i(u)-Y_i(u)\widehat{\beta}^{'}Z_i(u)du\right\}}{\sum_{i=1}^N Y_i(u)},\tag{2.4}$$

其中 $a^{\otimes 2}=aa'$, 以及

$$\overline{Z}(t)=\frac{\sum_{i=1}^N Y_i(t)Z_i(t)}{\sum_{i=1}^N Y_i(t)}.$$

在研究的病例队列设计下, 首先, 通过简单随机抽样方式从全队列中抽取一个样本容量为 n_0 的子队列. 子队列中的个体和子队列之外的所有病例个体组成病例队列样本, 记其样本量为 n. 然后, 仅对病例队列样本中的个体观测其协变量. 具体来说, 记 ξ_i 为子队列示性变量, 即: $\xi_i=1$ 表示第 i 个个体被选入子队列, $\xi_i=0$ 表示第 i 个个体未被选入子队列. 假设每个个体入选子队列的概率 $P(\xi_i=1)=\widetilde{\alpha}=n_0/N$. 因此, 病例队列设计下, 当 $\xi_i=1$ 或 $\Delta_i=1$ 时, 观测数据为 (T_i,Δ_i,Z_i); 当 $\xi_i=0$ 且 $\Delta_i=0$ 时, 观测数据为 (T_i,Δ_i).

由于病例队列设计中, 协变量不是对每一个个体都进行了观测, 因此 (2.2) 中的估计方程方法不再适用, 需要提出新的推断方法. 受 Horvitz & Thompson(1951)[23] 逆概率加权思想和 Liang & Ziger(1986)[24] 广义估计方程思想的启发. 在病例队列的设计下, 对加性风险模型 (2.1) 中的 β 和 $\Lambda_0(t)$ 的推断可建立如下加权估计方程:

$$\begin{aligned}&\sum_{i=1}^N\int_0^\tau\omega_i\left[dN_i(t)-Y_i(s)\left\{d\Lambda_0(t)+\beta^{'}Z_i(t)dt\right\}\right]=0,\\&\sum_{i=1}^N\int_0^\tau\omega_iZ_i(t)\left[dN_i(t)-Y_i(t)\left\{d\Lambda_0(t)+\beta^{'}Z_i(t)dt\right\}\right]=0,\end{aligned}\tag{2.5}$$

其中

$$\omega_i=\Delta_i+\frac{\xi_i(1-\Delta_i)}{\widetilde{\alpha}}.$$

上述加权估计方程有如下显式解:

$$\widehat{\beta}_w=\left[\sum_{i=1}^N\int_0^\tau\omega_iY_i(t)\{Z_i(t)-\overline{Z}_w(t)\}^{\otimes 2}dt\right]^{-1}\left[\sum_{i=1}^N\int_0^\tau\omega_i\{Z_i(t)-\overline{Z}_w(t)\}dN_i(t)\right],\tag{2.6}$$

和

$$\widehat{\Lambda}_0^w(t)=\int_0^\tau\frac{\sum_{i=1}^N\omega_i\left\{dN_i(u)-Y_i(u)\widehat{\beta}_w^{'}Z_i(u)du\right\}}{\sum_{i=1}^N\omega_iY_i(u)},\tag{2.7}$$

其中

$$\overline{Z}_w(t)=\frac{\sum_{i=1}^N\omega_iY_i(t)Z_i(t)}{\sum_{i=1}^N\omega_iY_i(t)}.$$

这种逆概率加权的思想由 Kalbfleisch & Lawless(1988)[25] 首次应用于生存分析数据. 基于多类型疾病的病例队列设计, Kang et al(2013)[18] 在加性风险模型下为多元失效时间发展

了逆概率加权推断方法. 以下我们讨论和综述了上述 $\widehat{\beta}_w$ 的渐近性质. 设 β_0 为 β 的真值. 假设如下正则条件成立:

(C1) β 的参数空间 $\mathbb{B}$ 是紧集, Z 的取值空间 $\mathbb{Z}$ 是紧集.

(C2) 给定 Z_i 时, $\widetilde{T}_i$ 与 C_i 相互独立, ξ_i 与 (T_i, Δ_i, Z_i) 相互独立.

(C3) $P(T_i \geq \tau) > 0$ 且 $\Lambda_0(t) < \infty$.

(C4) 存在某个 $\alpha \in (0,1)$, 使得当 $N \to \infty$ 时, $\widetilde{\alpha} = n_0/N \to \alpha$.

(C5) 矩阵

$$A = E\left[\int_0^\tau Y_i(t)\left\{Z_i(t) - \bar{z}(t)\right\}^{\otimes 2} dt\right]$$

是非奇异矩阵, 其中

$$\bar{z}(t) = \frac{E\left[Y_i(t)Z_i(t)\right]}{E\left[Y_i(t)\right]}.$$

定理 1 ($\widehat{\beta}_w$ 的渐近性质) 在正则条件 (C1)-(C5) 下, $\widehat{\beta}_w$ 是 β_0 的相合估计, 即:

$$\widehat{\beta}_w \xrightarrow{\mathrm{P}} \beta_0.$$

进一步, 有

$$\sqrt{N}\{\widehat{\beta}_w - \beta_0\} \xrightarrow{d} N(0, A^{-1}\left\{\Sigma_1 + \Sigma_2\right\}A^{-1}),$$

其中

$$\Sigma_1 = E\left[\int_0^\tau \left\{Z_i(t) - \overline{z}(t)\right\} dM_i^{(0)}(t)\right]^{\otimes 2},$$

$$\Sigma_2 = \frac{1-\alpha}{\alpha} E\left[(1-\Delta_i)\int_0^\tau \left\{Z_i(t) - \overline{z}(t)\right\} dM_i^{(0)}(t)\right]^{\otimes 2},$$

这里

$$M_i^{(0)}(t) = N_i(t) - \int_0^t Y_i(s)\left\{d\Lambda_0(s) + \beta_0' Z_i(s)ds\right\}.$$

进一步, 为了实际应用中的计算问题, 我们为渐近方差 $A^{-1}\left\{\Sigma_1 + \Sigma_2\right\}A^{-1}$ 提出了如下一种相合估计. 定义:

$$\widehat{A} = \frac{1}{N}\sum_{i=1}^N \int_0^\tau Y_i(t)\left\{Z_i(t) - \overline{Z}_w(t)\right\}^{\otimes 2} dt,$$

$$\widehat{\Sigma}_1 = \frac{1}{N}\sum_{i=1}^N \omega_i \widehat{\varphi}_i^{\otimes 2},$$

以及

$$\widehat{\Sigma}_2 = \left\{\frac{1-\widetilde{\alpha}}{\widetilde{\alpha}}\right\}\left[\frac{1}{N}\sum_{i=1}^N \omega_i\left(1-\Delta_i\right)\widehat{\varphi}_i^{\otimes 2}\right],$$

其中

$$\widehat{\varphi}_i = \int_0^\tau \left\{Z_i(t) - \overline{Z}_w(t)\right\}\left\{dN_i(t) - Y_i(t)d\widehat{\Lambda}_0(t) - Y_i(t)\widehat{\beta}_w' Z_i(t)dt\right\}.$$

易得

$$\widehat{A}^{-1}\left\{\widehat{\Sigma}_1+\widehat{\Sigma}_2\right\}\widehat{A}^{-1}$$

是 $A^{-1}\{\Sigma_1+\Sigma_2\}A^{-1}$ 的一个相合估计.

3 模拟研究

本节我们通过一系列模拟研究来展示上节中讨论的加权估计方法在有限样本量下的优良表现, 展示病例队列设计下加性风险回归方法的应用价值. 考虑如下加性风险模型:

$$\lambda(t|Z)=\lambda_0(t)+\beta^{'}Z(t),$$

设定参数真值 $\beta_1=0$ 或 0.5, $\beta_2=0$ 或 0.5. 协变量 Z_1 分别从均匀分布 $U(0,1)$ 和正态分布 $N(0,1)$ 中生成, Z_2 从成功率为 0.5 的 Bernoulli 分布中生成. 设定基准风险函数 $\lambda_0(t)=1$, 则失效时间 $\widetilde{T}$ 可以从风险率为 $\lambda_0(t)+\beta_1Z_1+\beta_2Z_2$ 的指数分布中生成. 删失时间 C 从均匀分布 $U[0,c]$ 中生成, 通过挑选 c 的不同取值从而产生相应的删失率, 分别为 $\rho=80\%$, 85% 和 90%. 对于病例队列设计, 设定全队列样本总量 $N=1000$, 子队列样本量为 $n_0=200$.

为了阐明问题, 我们比较以下几种方法:

Full: 基于全队列的估计方程方法 ($\widehat{\beta}_F$);

SRS: 基于子队列的估计方程方法 ($\widehat{\beta}_S$);

Naive: 基于与病例队列样本相同样本量的简单随机样本的估计方程方法 ($\widehat{\beta}_N$);

CC: 病例队列设计下的逆概率加权估计法 ($\widehat{\beta}_P$).

在每种参数设定下, 比较上述四种方法的参数估计值的均值 (Mean), 估计值的样本标准差 (SD), 标准差估计值的均值 (SE), 95% 的正态区间覆盖率 (CP) 以及估计的相对效率 (RE), 结果均基于 1000 次的模拟结果计算获得. 模拟结果请见表 1 和表 2.

在所有考虑的情况下, 关于 β_1 和 β_2 的四种估计都是无偏的, 标准误差估计的均值 (SEs) 很好地估计了估计值的样本标准差 (SDs), 置信区间覆盖率 (CPs) 均约为 95%. 模拟结果表明, 较之传统简单随机抽样设计, 病例队列设计能有效地提高估计的效率. 在所有情况下, $\widehat{\beta}_P$ 均比 $\widehat{\beta}_S$ 和 $\widehat{\beta}_N$ 更有效. 例如: 关于 β_1 的估计, 当 $\rho=0.9$, $\beta_1=\beta_2=0$, $Z_1\sim U(0,1)$ 时, 相较于 $\widehat{\beta}_F$, $\widehat{\beta}_S$, $\widehat{\beta}_N$ 和 $\widehat{\beta}_P$ 的相对效率分别为 0.20, 0.28 和 0.62. 这一方面说明, 在相同样本量下, 病例队列设计的效率约为简单随机抽样设计的 2.2 倍. 另一方面说明, 病例队列设计仅用了全队列 28% 的样本量可达到约 62% 的效率. 再例如, 关于 β_2 的估计, 当 $\rho=0.85$, $\beta_1=0$, $\beta_2=0.5$, $Z_1\sim N(0,1)$ 时, 相较于 $\widehat{\beta}_F$, $\widehat{\beta}_S$, $\widehat{\beta}_N$ 和 $\widehat{\beta}_P$ 的相对效率分别为 0.20, 0.32 和 0.54. 病例队列设计仅用了全队列 32% 的样本量却达到了约 54% 的效率, 且其效率为相同样本量下简单随机抽样设计的 1.7 倍. 当删失率提高时, $\widehat{\beta}_P$ 的效率更高.

总体来说, 在有限样本量下, 本文研究的病例队列设计下加性风险模型中的加权估计方法表现优异. 病例队列设计作为一种基于生存数据的有偏抽样机制, 通过将资源集中到认为包含有更多信息的群体上, 能够提高估计的效率并节约研究成本. 因此, 相比简单随机抽样设计有显著的成本效益. 尤其是感兴趣的事件是稀发事件时, 此时抽样设计非常高效.

表 1: 参数 β_1 和 β_2 的模拟结果, 其中 $Z_1 \sim U(0,1)$.

			$\widehat{\beta}_1$					$\widehat{\beta}_2$				
ρ	(β_1,β_2)	Method	Bias	SD	SE	CP	RE	Bias	SD	SE	CP	RE
0.8	(0,0)	Full	0.005	0.236	0.231	0.945	1.00	0.006	0.135	0.133	0.944	1.00
		SRS	−0.006	0.546	0.519	0.944	0.20	0.016	0.299	0.299	0.957	0.20
		Naive	0.002	0.397	0.386	0.955	0.36	0.006	0.228	0.222	0.940	0.36
		CC	−0.009	0.349	0.335	0.946	0.47	0.010	0.198	0.193	0.958	0.47
	(0,0.5)	Full	0.014	0.285	0.286	0.961	1.00	−0.027	0.166	0.166	0.945	1.00
		SRS	0.018	0.645	0.644	0.955	0.20	−0.036	0.384	0.373	0.938	0.20
		Naive	0.015	0.474	0.476	0.946	0.36	−0.017	0.271	0.276	0.955	0.36
		CC	0.016	0.424	0.419	0.955	0.47	−0.019	0.242	0.243	0.960	0.47
	(0.5,0)	Full	−0.029	0.280	0.286	0.952	1.00	−0.001	0.168	0.165	0.958	1.00
		SRS	−0.015	0.655	0.640	0.940	0.20	0.004	0.363	0.369	0.963	0.20
		Naive	−0.036	0.470	0.476	0.946	0.36	0.009	0.285	0.275	0.939	0.36
		CC	−0.025	0.410	0.416	0.951	0.47	−0.004	0.238	0.240	0.960	0.47
	(0.5,0.5)	Full	−0.012	0.349	0.346	0.947	1.00	−0.022	0.202	0.200	0.957	1.00
		SRS	−0.016	0.777	0.781	0.956	0.20	−0.009	0.466	0.450	0.946	0.20
		Naive	−0.003	0.583	0.577	0.942	0.36	−0.030	0.336	0.334	0.950	0.36
		CC	0.006	0.510	0.506	0.955	0.47	−0.010	0.293	0.292	0.952	0.47
0.85	(0,0)	Full	0.029	0.277	0.265	0.949	1.00	0.000	0.154	0.153	0.957	1.00
		SRS	0.027	0.608	0.596	0.941	0.20	−0.006	0.357	0.344	0.961	0.20
		Naive	0.042	0.480	0.470	0.938	0.32	0.001	0.272	0.271	0.949	0.32
		CC	0.028	0.361	0.362	0.948	0.54	−0.009	0.214	0.209	0.953	0.54
	(0,0.5)	Full	0.001	0.323	0.332	0.960	1.00	−0.029	0.196	0.192	0.942	1.00
		SRS	−0.012	0.787	0.745	0.942	0.20	−0.020	0.452	0.432	0.939	0.20
		Naive	0.019	0.591	0.586	0.955	0.32	−0.054	0.339	0.340	0.944	0.32
		CC	0.004	0.454	0.454	0.958	0.53	−0.026	0.269	0.264	0.961	0.53
	(0.5,0)	Full	−0.033	0.340	0.339	0.948	1.00	−0.001	0.190	0.196	0.955	1.00
		SRS	−0.017	0.793	0.763	0.942	0.20	0.004	0.444	0.440	0.950	0.20
		Naive	−0.046	0.618	0.602	0.954	0.32	0.014	0.338	0.347	0.959	0.32
		CC	−0.040	0.465	0.461	0.950	0.54	0.000	0.270	0.266	0.954	0.54
	(0.5,0.5)	Full	−0.032	0.412	0.403	0.942	1.00	−0.038	0.230	0.233	0.957	1.00
		SRS	−0.038	0.927	0.906	0.946	0.20	−0.011	0.533	0.521	0.950	0.20
		Naive	−0.026	0.720	0.715	0.957	0.32	−0.029	0.410	0.414	0.952	0.32
		CC	−0.028	0.546	0.550	0.954	0.54	−0.022	0.322	0.318	0.952	0.54
0.9	(0,0)	Full	0.007	0.325	0.325	0.955	1.00	0.002	0.190	0.187	0.946	1.00
		SRS	−0.025	0.759	0.729	0.946	0.20	−0.006	0.437	0.421	0.952	0.20
		Naive	0.015	0.634	0.613	0.953	0.28	0.003	0.353	0.353	0.949	0.28
		CC	−0.013	0.417	0.413	0.961	0.62	−0.002	0.248	0.238	0.946	0.62
	(0,0.5)	Full	0.004	0.424	0.411	0.948	1.00	−0.034	0.242	0.238	0.938	1.00
		SRS	0.041	0.989	0.924	0.941	0.20	−0.016	0.550	0.536	0.952	0.20
		Naive	0.015	0.788	0.782	0.946	0.28	−0.027	0.437	0.453	0.958	0.28
		CC	0.010	0.539	0.523	0.950	0.62	−0.024	0.306	0.303	0.951	0.62
	(0.5,0)	Full	−0.022	0.408	0.412	0.956	1.00	−0.001	0.242	0.237	0.950	1.00
		SRS	−0.046	0.974	0.924	0.951	0.20	−0.016	0.552	0.532	0.948	0.20
		Naive	−0.023	0.804	0.783	0.947	0.28	0.005	0.468	0.452	0.954	0.28
		CC	−0.020	0.528	0.522	0.957	0.62	−0.010	0.314	0.300	0.957	0.63
	(0.5,0.5)	Full	−0.013	0.487	0.480	0.951	1.00	−0.022	0.283	0.278	0.949	1.00
		SRS	0.029	1.075	1.076	0.954	0.20	0.001	0.638	0.623	0.952	0.20
		Naive	−0.002	0.927	0.907	0.949	0.28	−0.031	0.514	0.524	0.953	0.28
		CC	−0.009	0.624	0.616	0.957	0.61	−0.015	0.353	0.356	0.957	0.61

表 2: 参数 β_1 和 β_2 的模拟结果, 其中 $Z_1 \sim N(0,1)$.

			$\widehat{\beta}_1$					$\widehat{\beta}_2$				
ρ	(β_1,β_2)	Method	Bias	SD	SE	CP	RE	Bias	SD	SE	CP	RE
0.8	(0,0)	Full	−0.001	0.138	0.134	0.940	1.00	0.004	0.258	0.268	0.957	1.00
		SRS	0.002	0.320	0.302	0.946	0.20	−0.016	0.611	0.603	0.956	0.20
		Naive	0.005	0.230	0.224	0.940	0.36	−0.001	0.435	0.448	0.964	0.36
		CC	0.000	0.207	0.193	0.940	0.48	0.000	0.395	0.387	0.946	0.48
	(0,0.5)	Full	0.002	0.149	0.149	0.951	1.00	−0.017	0.308	0.298	0.937	1.00
		SRS	−0.003	0.351	0.337	0.944	0.19	0.007	0.664	0.673	0.959	0.20
		Naive	0.016	0.260	0.249	0.943	0.36	0.001	0.487	0.497	0.956	0.36
		CC	−0.001	0.220	0.217	0.961	0.47	0.004	0.432	0.436	0.958	0.47
	(0.5,0)	Full	−0.026	0.136	0.133	0.938	1.00	0.003	0.270	0.267	0.946	1.00
		SRS	0.000	0.315	0.303	0.939	0.19	0.010	0.613	0.603	0.952	0.20
		Naive	−0.020	0.225	0.225	0.947	0.35	0.009	0.444	0.448	0.952	0.35
		CC	−0.008	0.200	0.198	0.952	0.45	−0.002	0.389	0.390	0.961	0.47
	(0.5,0.5)	Full	−0.023	0.149	0.148	0.946	1.00	−0.037	0.296	0.296	0.948	1.00
		SRS	−0.031	0.343	0.333	0.933	0.20	−0.031	0.689	0.666	0.948	0.20
		Naive	−0.025	0.248	0.247	0.945	0.36	−0.033	0.488	0.494	0.955	0.36
		CC	−0.021	0.225	0.219	0.932	0.46	−0.035	0.438	0.434	0.953	0.47
0.85	(0,0)	Full	0.006	0.176	0.173	0.950	1.00	−0.022	0.332	0.346	0.959	1.00
		SRS	0.001	0.404	0.387	0.944	0.20	−0.019	0.783	0.776	0.946	0.20
		Naive	−0.009	0.304	0.308	0.954	0.32	−0.039	0.610	0.611	0.947	0.32
		CC	0.003	0.241	0.236	0.956	0.54	−0.001	0.472	0.473	0.953	0.53
	(0,0.5)	Full	−0.005	0.155	0.154	0.948	1.00	0.000	0.301	0.306	0.959	1.00
		SRS	0.002	0.356	0.345	0.954	0.20	0.008	0.705	0.688	0.945	0.20
		Naive	−0.010	0.270	0.272	0.956	0.32	0.015	0.547	0.540	0.947	0.32
		CC	−0.007	0.211	0.210	0.959	0.54	−0.004	0.412	0.418	0.955	0.54
	(0.5,0)	Full	−0.029	0.155	0.153	0.939	1.00	−0.002	0.306	0.305	0.952	1.00
		SRS	−0.031	0.358	0.344	0.947	0.20	0.005	0.701	0.685	0.944	0.20
		Naive	−0.027	0.288	0.270	0.932	0.32	0.023	0.533	0.539	0.954	0.32
		CC	−0.018	0.223	0.214	0.934	0.51	−0.003	0.421	0.420	0.956	0.53
	(0.5,0.5)	Full	−0.021	0.170	0.172	0.951	1.00	0.002	0.337	0.345	0.953	1.00
		SRS	−0.014	0.371	0.383	0.960	0.20	0.024	0.751	0.771	0.956	0.20
		Naive	−0.009	0.319	0.305	0.934	0.32	−0.023	0.607	0.611	0.952	0.32
		CC	−0.013	0.232	0.238	0.954	0.52	0.007	0.479	0.473	0.954	0.53
0.9	(0,0)	Full	0.002	0.188	0.188	0.955	1.00	−0.007	0.377	0.375	0.939	1.00
		SRS	0.009	0.421	0.421	0.965	0.20	−0.022	0.854	0.839	0.959	0.20
		Naive	0.006	0.350	0.352	0.955	0.28	0.019	0.737	0.707	0.945	0.28
		CC	−0.001	0.235	0.239	0.962	0.62	−0.024	0.478	0.476	0.954	0.62
	(0,0.5)	Full	−0.004	0.206	0.209	0.955	1.00	−0.026	0.428	0.417	0.944	1.00
		SRS	0.004	0.472	0.472	0.947	0.20	−0.016	0.984	0.938	0.951	0.20
		Naive	0.002	0.403	0.395	0.950	0.28	−0.006	0.806	0.787	0.936	0.28
		CC	−0.011	0.275	0.269	0.944	0.61	−0.024	0.548	0.532	0.954	0.62
	(0.5,0)	Full	−0.023	0.187	0.187	0.942	1.00	−0.008	0.375	0.374	0.946	1.00
		SRS	−0.014	0.424	0.419	0.939	0.20	−0.007	0.856	0.840	0.953	0.20
		Naive	−0.033	0.353	0.354	0.943	0.28	−0.024	0.733	0.709	0.954	0.28
		CC	−0.010	0.243	0.244	0.958	0.59	−0.027	0.492	0.479	0.956	0.61
	(0.5,0.5)	Full	−0.022	0.205	0.209	0.949	1.00	−0.042	0.407	0.417	0.958	1.00
		SRS	−0.017	0.469	0.467	0.954	0.20	−0.037	0.953	0.932	0.938	0.20
		Naive	−0.008	0.407	0.395	0.945	0.28	−0.038	0.762	0.783	0.952	0.28
		CC	−0.014	0.270	0.272	0.952	0.59	−0.035	0.526	0.534	0.956	0.61

4 乳腺癌数据的生存分析

本节我们研究乳腺癌相关数据. 数据来源于美国国家癌症研究所 (National Cancer Institute)[22]. 我们选取了 1975-2017 年期间 40 岁以上的女性乳腺癌患者的共 118763 条数

据. 我们基于加性风险模型对此数据集进行生存回归分析, 探索影响乳腺癌患者生存时间的主要影响因素. 进一步, 我们应用病例队列设计分析数据集, 展示此种有偏抽样机制的实际应用价值.

我们感兴趣的因变量是乳腺癌患者的生存时间, 而观测到的生存时间存在删失, 其删失率为 84.4%. 我们考虑如下 6 个潜在影响因素. 患者年龄 (Age) 分为 5 组: 40-49 岁 (Age=1), 50-59 岁 (Age=2), 60-69 岁 (Age=3), 70-79 岁 (Age=4), 80 岁以上 (Age=5). 种族 (Race) 分为 3 种: 其他种族 (Race=1), 白种人 (Race=2), 黑种人 (Race=3). 癌细胞分化程度 (Grade) 分为 4 种类型: 肿瘤高度分化 (Grade=1), 肿瘤中度分化 (Grade=2), 肿瘤低分化 (Grade=3), 肿瘤未分化 (Grade=4). 肿瘤直径大小的 T 分期 (T) 分为 5 种类型: 肿瘤直径 $\leq 2cm$(T=1), $2cm <$ 肿瘤直径 $\leq 5cm$(T=2), 肿瘤直径 $> 5cm$(T=3), 肿瘤直接侵犯胸壁或皮肤 (T=4), 肿瘤无法评估 (T=5). 是否并发淋巴癌的 N 分期分为 5 种类型: 同侧腋窝无肿大淋巴结 (N=1), 同侧腋窝有尚可推动的肿大淋巴结 (N=2), 同侧腋窝肿大淋巴结融合或粘连 (N=3), 有同侧胸骨旁淋巴结转移 (N=4), 区域淋巴结无法评估 (N=5). 肿瘤是否转移的 M 分期 (M) 分为 2 种类型: 肿瘤未转移 (M=1), 肿瘤转移 (M=2). 我们对数据中考虑的上述影响因素进行了描述性统计分析, 画出了其条形图 (图 1) 及生存曲线图 (图 2).

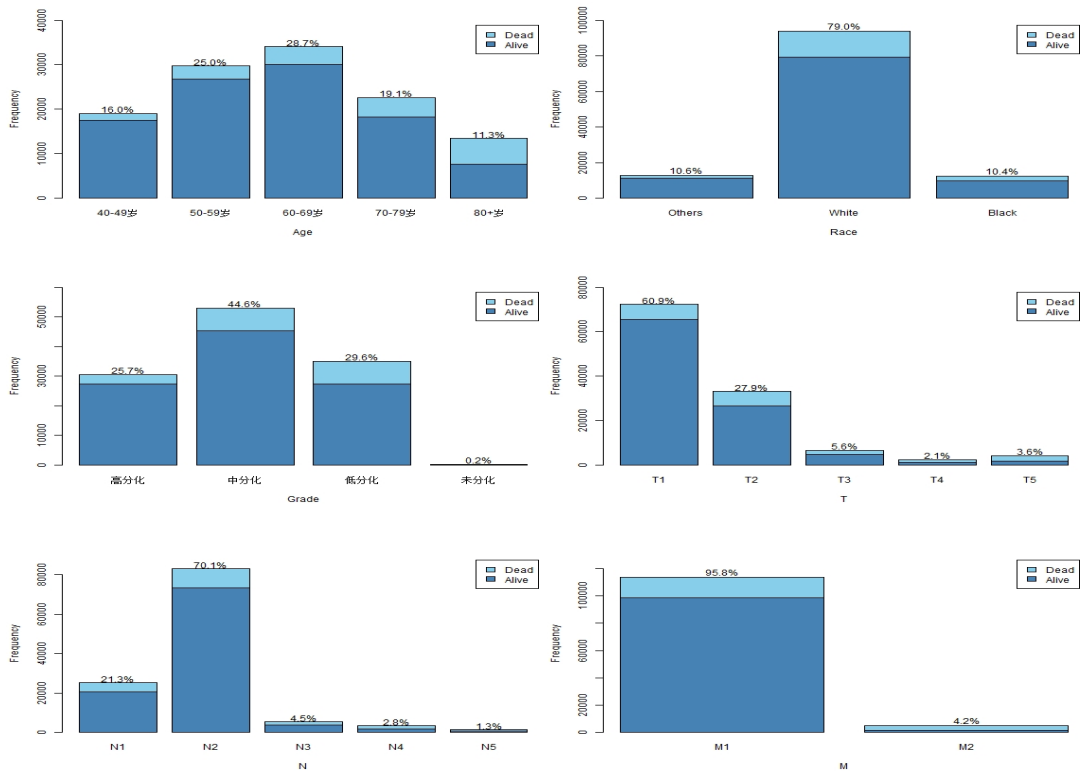

图 1 影响因素条形图

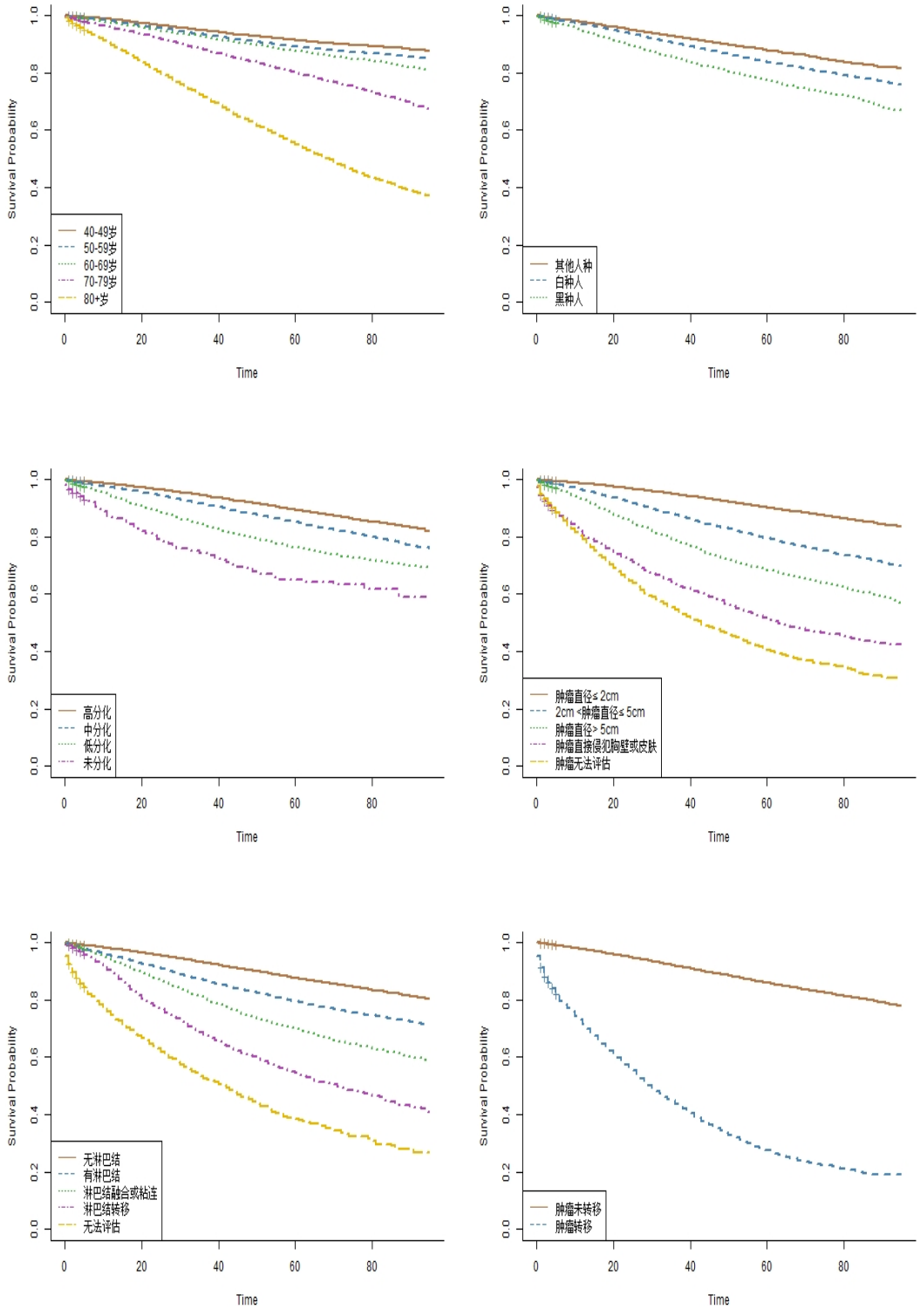
Survival Probability
Time
40-49岁
50-59岁
60-69岁
70-79岁
80+岁
其他人种
白种人
黑种人
高分化
中分化
低分化
未分化
肿瘤直径≤2cm
2cm <肿瘤直径≤5cm
肿瘤直径> 5cm
肿瘤直接侵犯胸壁或皮肤
肿瘤无法评估
无淋巴结
有淋巴结
淋巴结融合或粘连
淋巴结转移
无法评估
肿瘤未转移
肿瘤转移

图 2 生存曲线条形图

为了探究乳腺癌患者生存时间的影响因素. 我们建立如下加性风险模型:

$$\lambda(t|Z) = \lambda_0(t) + \beta_1 Age + \beta_2 Race + \beta_3 Grade + \beta_4 T + \beta_5 N + \beta_6 M.$$

首先, 基于全队列 118763 条数据进行加性风险回归分析, 结果总结在表 3 中 (见 Full-Cohort). 结果表明, 考虑的 6 个协变量对乳腺癌患者的生存时间均有显著的影响. 具体来说, 年龄越大的患者死亡率越高. 黑种人死亡风险率最高, 白种人次之, 其他人种最低. 癌细胞分化程度越高, 肿瘤大小的 T 分期等级越高, 区域淋巴癌的 N 分期等级越高, 患者的生存率越低. 癌细胞有远处转移的患者比无远处转移的患者的死亡风险率高出 1.476%.

为了评估病例队列设计并展现其在实际应用中的可行性与有效性, 我们首先从全队列中随机抽取了一个容量为 $n_0 = 30000$ 的子队列, 然后将子队列和子队列之外的病例 ($n_c = 13903$) 组成病例队列样本, 其样本容量为 43903. 我们应用本文研究的病例队列设计下加性风险模型下的加权估计方法对数据进行分析, 其结果总结在表 3 中 (见 Case-Cohort). 结果表明, 病例队列设计采用了较小的样本量, 但估计结果与基于全队列数据分析得到的估计结果十分接近. 病例队列设计仅用了全队列约 37% 的样本量, 对于 Grade, T, N 的估计分别达到了约 52.6%, 41.2% 和 45.8% 的效率. 这说明在实际应用中, 当影响因素 Grade, T 和 N 的测量非常昂贵或耗时时, 采用病例队列设计会提高研究效率和节约成本.

表 3: 乳腺癌细胞瘤研究数据分析结果

	Full-Cohort (N=118763)			Case-Cohort ($n_0 = 30000, n_c = 13903$)		
	Est$\times 10^{-3}$	SE$\times 10^{-3}$	p-value	Est$\times 10^{-3}$	SE$\times 10^{-3}$	p-value
Age	1.65	0.02	$< 1 \times 10^{-16}$	1.65	0.03	$< 1 \times 10^{-16}$
Race	0.71	0.05	$< 1 \times 10^{-16}$	0.70	0.07	$< 1 \times 10^{-16}$
Grade	0.75	0.03	$< 1 \times 10^{-16}$	0.74	0.04	$< 1 \times 10^{-16}$
T	1.60	0.04	$< 1 \times 10^{-16}$	1.71	0.07	$< 1 \times 10^{-16}$
N	0.98	0.05	$< 1 \times 10^{-16}$	0.98	0.07	$< 1 \times 10^{-16}$
M	14.76	0.36	$< 1 \times 10^{-16}$	14.64	0.60	$< 1 \times 10^{-16}$

参 考 文 献

[1] Cox D R. Regression models and life-tables[J]. Journal of the Royal Statistical Society (SeriesB), Methodological, 1972, 34(2): 187–220.

[2] Andersen P K, Gill R D. Cox' s regression model for counting processes: a large sample study[J]. The Annals of Statistics, 1982, 10(4): 1100–1120.

[3] Lin D Y, Ying Z. Semiparametric analysis of the additive risk model[J]. Biometrika, 1994, 81(1): 61–71.

[4] Chen H Y, Little R J. Proportional hazards regression with missing covariates[J]. Journal of the American Statistical Association, 1999, 94(447): 896–908.

[5] John P Klein. The statistical analysis of failure time data[J]. Technometrics, 2012, 24(3): 251–251.

[6] Mckeague I W, Sasieni P D. A partly parametric additive risk model[J]. Biometrika, 1994, 81(3): 501–514.

[7] Zeng Donglin, Cai Jianwen. Additive transformation models for clustered failure time data[J]. Lifetime Data Analysis, 2010, 16(3): 333–352.

[8] Prentice R L. A case-cohort design for epidemiologic cohort studies and disease prevention trials[J]. Biometrika, 1986, 73(1): 1–11.

[9] Self S G, Prentice R. Asymptotic distribution theory and efficiency results for case-cohort studies[J]. The Annals of Statistics, 1988, 16(1): 64–81.

[10] Chen K, Lo S. Case-cohort and case-control analysis with Cox' s model[J]. Biometrika, 1999, 86(4), 755–764.

[11] Lin D Y, Ying Z. Cox regression with incomplete covariate measurements[J]. Journal of the American Statistical Association, 2012, 88(424): 1341–1349.

[12] Kulich M, Lin D Y. Improving the efficiency of relative-risk estimation in case-cohort studies[J]. Journal of the American Statistical Association, 2004, 99(467): 832-844.

[13] Kulich M, Lin D Y. Additive hazards regression for case-cohort studies[J]. Biometrika, 2000, 87(1): 15.

[14] Sun J, Sun L Q, Flournoy N. Additive hazards model for competing risks analysis of the case-cohort design[J]. Communications in Statistics - Theory and Methods, 2004, 33(2): 351–366.

[15] Kong L, Cai J, Sen P K. Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design[J]. Biometrika, 91(2), 305–319.

[16] Lu W, Tsiatis A A. Semiparametric transformation models for the case-cohort study[J]. Biometrika, 2006, 93(1): 8.

[17] Kong L, Cai J. Case-cohort analysis with accelerated failure time model[J]. Biometrics., 2009, 65(1) : 135–142.

[18] Kang S, Cai J, Chambless L. Marginal additive hazards model for case-cohort studies with multiple disease outcomes: an application to the Atherosclerosis Risk in Communities (ARIC) study[J]. Biostatistics, 2013, 14(1): 28–41.

[19] Kim S, Zheng D, Cai J. Analysis of multiple survival events in generalized case-cohort designs[J]. Biometrics, 2018, 74(4): 1250–1260.

[20] Maitra Poulami, Amorim Leila D A F, Cai Jianwen. Multiplicative rates model for recurrent events in case-cohort studies[J]. Lifetime data analysis., 2020, 26(1): 134–157.

[21] 陈津利, 莫邦豪, 冯可立. 乳腺癌患者在自助组织中的情感支持: 一个探索性的研究 [J]. The Hong Kong Journal of Social Work, 2003, 37(2): 173–190.

[22] https://seer.cancer.,gov/data/

[23] Horvitz D G, Thompson D J. A generalization of sampling without replacement from a finite universe[J]. Journal of the American Statistical Association., 1952, 47(260): 663–685.

[24] Zeger Scott L, Liang Kung-Yee. Longitudinal data analysis for discrete and continuous outcomes[J]. Biometrics., 1986, 42(1): 121–130.

[25] Kalbfleisch J D, Lawless J F. Likelihood analysis of multi-state models for disease incidence and mortality[J]. Statistics in medicine., 1988, 7(1-2): 149–160.

[26] Pollard D. Empirical processes: theory and applications[J]. NSF-CBMS Regional Conference Series in Probability and Statistics, 1990,2(1), 1–86

[27] van der Vaart AW, Wellner JA. Weak convergence and empirical processes[M]. NewYork: Springer, 1996.

[28] Kulich Michal, Lin D Y. Additive hazards regression with covariate measurement Error[J]. Journal of the American Statistical Association., 2012, 95(449): 238–248.

ADDITIVE HAZARDS REGRESSION UNDER CASE-COHORT DESIGN AND ITS APPLICATION IN BREAST CANCER DATA

YANG Jie, DING Jie-li

(*School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China*)

Abstract: In this paper, we discuss how to fit the additive hazards model to case-cohort data. We consider a weighted estimating estimation approach, and review asymptotic properties of the proposed estimator. We conduct a series of simulation studies to assess the good finite-sample performance of such a method, and evaluate the efficiency of the case-cohort design to the simple random sampling scheme. We promote the traditional simple random sampling design to case-cohort design. We apply the proposed method to analyze a real data for breast cancer and illustrate the practical applications.

Keywords: case-cohort design; additive hazards model; estimating equations; inversed probability weighted

2010 MR Subject Classification: 62N01; 62N02; 62N86

附录

基于前人的研究工作 (Lin, 1994[3], Kang et al, 2013[18]), 我们在此给出定理 1 的证明.

定理 1 的证明 根据一致强大数定律 (Pollard, 1990[26]), 对于 $t \in [0,\ \tau]$ 一致地有:

$$\frac{1}{N}\sum_{i=1}^{N}\omega_i Y_i(t) \xrightarrow{\text{a.s.}} E\left[Y_i(t)\right],$$

$$\frac{1}{N}\sum_{i=1}^{N}\omega_i Y_i(t)Z_i(t) \xrightarrow{\text{a.s.}} E\left[Y_i(t)Z_i(t)\right].$$

由此可得, 对于 $t \in [0,\ \tau]$ 一致地成立:

$$\overline{Z}_w(t) \xrightarrow{\text{a.s.}} \bar{z}(t), \qquad \widehat{A} \xrightarrow{\text{a.s.}} A. \tag{A.1}$$

注意到: 在模型 (2.1) 下, $M_i^{(0)}(t)$ 是均值为 0 的随机过程. 因此, 我们有

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\omega_i\int_0^{\tau}\left\{Z_i(t)-\overline{Z}_w(t)\right\}dN_i(t) =& \frac{1}{N}\sum_{i=1}^{N}\omega_i\int_0^{\tau}\left\{Z_i(t)-\overline{Z}_w(t)\right\}dM_i^{(0)}(t)\\ &+\frac{1}{N}\sum_{i=1}^{N}\omega_i\int_0^{\tau}\left\{Z_i(t)-\overline{Z}_w(t)\right\}Y_i(t)d\Lambda_0(t)\\ &+\left[\frac{1}{N}\sum_{i=1}^{N}\omega_i\int_0^{\tau}Y_i(t)\left\{Z_i(t)-\overline{Z}_w(t)\right\}^{\otimes 2}Y_i(t)dt\right]\beta_0\\ \xrightarrow{\text{a.s.}}& A\beta_0.\end{aligned} \tag{A.2}$$

由 (A.1) 和 (A.2) 式, 可得:

$$\widehat{\beta}_w = \widehat{A}^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\omega_i\int_0^\tau \left\{Z_i(t)-\overline{Z}_w(t)\right\}dN_i(t)\right] \xrightarrow{\text{a.s.}} \beta_0.$$

由此可得 $\widehat{\beta}_w$ 的相合性.

进一步, 我们有

$$\begin{aligned}\sqrt{N}(\widehat{\beta}_w-\beta_0) =&\widehat{A}^{-1}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega_i\int_0^\tau \left\{Z_i(t)-\overline{Z}_w(t)\right\}dM_i^{(0)}(t)\right]\\ =&\widehat{A}^{-1}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega_i\int_0^\tau \left\{Z_i(t)-\bar{z}(t)\right\}dM_i^{(0)}(t)\right]\\ &-\widehat{A}^{-1}\left[\int_0^\tau \left\{\overline{Z}_w(t)-\bar{z}(t)\right\}d\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega_i dM_i^{(0)}(t)\right\}\right].\end{aligned} \tag{A.3}$$

对于 (A.3) 式右侧的第二项, 由于零均值过程 $\omega_i M_i^{(0)}(t)$ 可以表示成两个单调过程的和, 因此, 根据单调随机过程的中心极限定理 (van der Vaart & Wellner, 1996)[27], 我们有 $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega_i M_i^{(0)}(t)$ 收敛到一个在 $[0,\tau]$ 上具有连续通道的紧高斯过程. 故而, 由 Kulich & Lin (2000)[28] 中的引理 A.1, 可得: (A.3) 式右边第二项收敛到零.

因此, 我们有

$$\begin{aligned}\sqrt{N}(\widehat{\beta}_w-\beta_0) &= A^{-1}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega_i\int_0^\tau \left\{Z_i(t)-\bar{z}(t)\right\}dM_i^{(0)}(t)\right]+o_P(1)\\ &= A^{-1}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega_i\varphi_i\right]+o_P(1).\end{aligned} \tag{A.4}$$

其中,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega_i\varphi_i = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\varphi_i+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\omega_i-1)\varphi_i.$$

由中心极限定理, 易知,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\varphi_i \xrightarrow{d} N(0,\,\Sigma_1), \tag{A.5}$$

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\omega_i-1)\varphi_i \xrightarrow{d} N(0,\,\Sigma_2). \tag{A.6}$$

由此可得,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega_i\varphi_i \xrightarrow{d} N(0,\,\Sigma_1+\Sigma_2).$$

基于 (A.4) 式以及 Slutsky 定理, 我们可得如下渐近正态性:

$$\sqrt{N}(\widehat{\beta}_w-\beta_0) \xrightarrow{d} N(0,\,A^{-1}(\Sigma_1+\Sigma_2)A^{-1}).$$