Vol. 42 (2022)
No. 5

数 学 杂 志
J. of Math. (PRC)

比例风险模型下病例队列设计中两种加权估计方法及其应用

张佳倩 [1], 邓立凤 [2], 丁洁丽 [1]

(1. 武汉大学数学与统计学院, 湖北 武汉 430072)

(2. 山东科技大学数学与系统科学学院, 山东 青岛 266590)

摘 要: 对于大型队列研究或观察型研究, 基于生存数据的病例队列设计是一种能有效节约成本和提高效率的抽样机制. 这种抽样设计仅对一个随机抽取的子队列以及子队列之外所有经历了感兴趣事件的病例个体进行关键协变量的测量, 具有显著的成本效益. 本文研究如何应用比例风险模型拟合病例队列研究数据. 探讨逆概率加权和与时间相关加权这两种基于加权估计方程的统计推断方法和其渐近性质等理论结果. 通过一系列的统计模拟研究展示了病例队列设计的优良性以及相较于传统简单随机抽样设计的高效性. 进一步, 应用这两种推断方法分析了两个实际数据, 展示了其在实际中的应用价值和前景.

关键词: 病例队列设计; 比例风险模型; 逆概率加权法; 与时间相关权法

MR(2010) **主题分类号:** 62D05; 62N01; 62N02 **中图分类号:** O212.1

文献标识码: A **文章编号:** 0255-7797(2022)05-0445-16

1 引言

流行病学、生物医学和遗传学等领域的观察型研究对了解人类疾病的影响因素至关重要. 在许多这样的研究中, 往往涉及对大量研究个体的长期追踪和随访. 因而, 这些研究的主要的预算和成本通常来自于昂贵协变量的采集与观测. 对于预算有限的大型观察型研究, 若采用传统的简单随机抽样可能会导致试验过于昂贵而超出预算. 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理论意义和应用价值.

对于带有删失的失效时间数据, 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. 其关键思想是: 首先, 从全队列中随机地抽取一个子队列 (subcohort). 全队列中所有发生或者经历了感兴趣的事件 (例如: 疾病, 死亡等) 的个体称为病例 (case). 然后, 子队列和子队列之外所有的病例一起组成病例队列样本 (case-cohort sample). 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. 这种病例队列设计对于需要观测昂贵协变量的大型队列研究或观察型研究是具有成本效益的. 尤其是感兴趣的事件是稀发事件时, 此种抽样设计非常高效.

在 Prentice (1986)[1] 首次提出病例队列设计之后, 对于病例队列设计和相关统计方法的研究有了大量和广泛的工作, 这些工作主要沿着两条研究思路, 一条思路是基于似然函数的方法 (例如: Prentice, 1986[1]; SelfPrentice, 1988[2]; ChenLo, 1999[3]; Kang et al, 2016[4]; Liu et al, 2018[5] 等等), 一条思路是估计方程的方法 (例如: LinYing, 1993[6]; KulichLin, 2000[7]; Sun

*收稿日期: 2021-04-06 接收日期: 2021-05-16

基金项目: 国家自然科学基金资助 (11671310).

作者简介: 张佳倩 (1996–), 女, 湖北武汉, 研究生, 研究方向: 统计学, 生物统计, 生存分析.

通讯作者: 丁洁丽 (1979–), 女, 湖北武汉, 博士, 副教授, 研究方向: 统计学, 生物统计, 生存分析.

et al, 2004[8]; CaiZeng, 2004[9], 2007[10]; KulichLin, 2004[11]; Kong et al, 2004[12]; LuTsiatis, 2006[13]; BreslowWellner, 2007[14]; Kang et al, 2013[15]; SteingrimssonStrawderman, 2017[16] 等等). 在估计方程方法的研究中, 对基于不完全观测数据的逆概率加权思想的探讨尤其广泛, 例如: 加性风险模型 (KulichLin, 2000[7]); 加速失效模型 (KongCai, 2009[17]); 半参数转移模型 (Kong et al, 2004[12]; LuTsiatis, 2006[13]). 在对带有多元失效时间的病例队列研究中, 一些加权估计方程应用了一种与时间相关的权函数并发展了相应的统计推断方法 (KangCai, 2009[18]; Kang et al, 2013[15]; Yan et al, 2017[19]).

病例队列设计作为一种具有成本效益的有偏抽样机制, 其在各个领域的应用越来越广泛. 本文主要探讨病例队列研究中的逆概率加权和与时间相关加权这两种思想下的估计方程方法及其应用. 首先, 我们探讨如何用比例风险模型拟合病例队列数据, 考虑基于逆概率加权和与时间相关加权两种思想下的统计推断方法, 综述其渐近性质等理论结果. 然后, 我们重点研究上述两种方法在实际中的应用问题, 编写可实现这两种方法的操作性强的统计软件应用程序. 通过一系列的模拟研究来评估上述两种方法在有限样本下的表现, 展示了病例队列设计相较于简单随机抽样方法的优良性和有效性. 最后，我们将这两种方法应用于分析两个实际数据, 展示了该方法在实际中的应用价值与前景. 本文主要结构为: 第 2 节介绍比例风险模型和病例队列设计的抽样机制, 第 3 节探讨逆概率加权法和与时间相关的加权法这两种估计方法, 第 4 节为研究模拟计算, 第 5 节研究实际数据分析与应用.

2 模型与抽样设计

用 $\widetilde{T}$ 表示潜在失效时间, C 表示删失时间. 记 $T=\min(\widetilde{T},C)$ 为观测时间, $\Delta=I(\widetilde{T}\leq C)$ 为右删失示性变量, 其中 $I(\cdot)$ 为示性函数. 记 p 维协变量为 $\boldsymbol{Z}=(Z_1,Z_2,...,Z_p)^T$. 假设研究队列包含 N 个独立个体, $\{T_i,\Delta_i,\boldsymbol{Z}_i:i=1,...,N\}$ 为其观测数据. 记 τ 为实验终止时间. 我们考虑如下比例风险模型 (Cox, 1972[20]), 即: 给定协变量 $\boldsymbol{Z}$ 时, 失效时间 $\widetilde{T}$ 的风险率函数为:

$$\lambda(t|\boldsymbol{Z})=\lambda_0(t)\exp(\boldsymbol{Z}^T\boldsymbol{\beta}), \tag{2.1}$$

其中 $\lambda_0(t)$ 是未知的基准风险函数, $\boldsymbol{\beta}=(\beta_1,...,\beta_p)^T$ 为待估的 p 维回归系数.

在队列研究中, 当协变量对每个个体均进行观测时, 对 $\boldsymbol{\beta}$ 的推断问题广泛地应用如下的偏似然函数法 (Cox, 1972[20]; AndersenGill, 1982[21]):

$$\ell_F(\boldsymbol{\beta})=\sum_{i=1}^{n}\Delta_i\left[\boldsymbol{Z}_i^T\boldsymbol{\beta}-\log\left\{\sum_{l=1}^{N}Y_l(T_i)\exp(\boldsymbol{Z}_l^T\boldsymbol{\beta})\right\}\right]. \tag{2.2}$$

$\boldsymbol{\beta}$ 的估计可定义为: $\widehat{\boldsymbol{\beta}}_F=\arg\max_{\boldsymbol{\beta}}\ell_F(\boldsymbol{\beta})$.

在病例队列设计下, 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 $\widetilde{n}$ 的子队列. 子队列中的个体和所有子队列之外的所有病例组成病例队列样本, 记其容量为 n. 然后, 仅对病例队列样本进行协变量的测量. 特别地, 记 ξ_i 为子队列示性变量, 即: 如果第 i 个个体被选入子队列, 则 $\xi_i=1$; 否则, $\xi_i=0$. 假设每个个体入选子队列的概率为 $\mathrm{P}(\xi_i=1)=\widetilde{\alpha}=\widetilde{n}/N$. 因此, 病例队列设计下的观测数据结构可总结为:

$$\begin{cases}\text{当 } \xi_i=1 \text{ 或 } \Delta_i=1 \text{ 时}: & (T_i,\,\Delta_i,\,Z_i);\\ \text{当 } \xi_i=0 \text{ 且 } \Delta_i=0 \text{ 时}: & (T_i,\,\Delta_i).\end{cases} \tag{2.3}$$

由于病例队列设计中, 协变量不是对每个个体进行观测, 因此, 传统推断方法不再适用, 需要发展新的分析方法. 以下我们探讨两种基于加权估计方程的推断方法.

3 估计方法

3.1 逆概率加权法

受 HorvitzThompson (1951)[22] 思想的启发, 即: 对于不完全观测的数据进行逆概率加权 (inverse-probability weight, 以下简称 "IPW"), 在病例队列设计下, 对模型 (2.1) 中 $\boldsymbol{\beta}$ 的推断可建立如下加权估计方程:

$$U_w(\boldsymbol{\beta}) = \sum_{i=1}^{N} \Delta_i \left[Z_i - \frac{S_w^{(1)}(\boldsymbol{\beta}, T_i)}{S_w^{(0)}(\boldsymbol{\beta}, T_i)} \right], \tag{3.1}$$

其中 $S_w^{(d)}(\boldsymbol{\beta}, t) = \frac{1}{N} \sum\limits_{i=1}^{N} w_l Y_l(t) e^{\boldsymbol{Z}_i'\boldsymbol{\beta}} \boldsymbol{Z}_l^{\otimes d}$, 这里 $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, $a^{\otimes 2} = aa'$, 其中 a 是一个向量. 权重 w_i 定义为:

$$w_i = \Delta_i + \frac{(1 - \Delta_i)\xi_i}{\widetilde{\alpha}}, \qquad i = 1, ..., N. \tag{3.2}$$

注意到, 对于病例组中的个体, 无论是否选入子队列, 其权重均为 1; 而对于子队列中的对照组个体, 其权重为 $\widetilde{\alpha}^{-1}$. 这种逆概率加权的思想由 KalbfleischLawless (1988)[23] 首次应用于分析生存数据. Borgan et al (2000)[24] 在分层病例队列研究中引入了相似的逆概率权. 对于两阶段抽样设计, KulichLin (2004)[11] 也提出了类似的加权估计方法. 在比例风险模型下, KangCai (2009)[18] 为多类型疾病的病例队列设计发展了基于多元失效时间的逆概率加权推断方法.

估计方程 (3.1) 的解定义为逆概率加权估计, 记为 $\widehat{\boldsymbol{\beta}}_{IPW}$. 以下, 我们综述 $\widehat{\boldsymbol{\beta}}_{IPW}$ 的渐近性质. 记 $\boldsymbol{\beta}_0$ 为 $\boldsymbol{\beta}$ 的真值. 定义 $N_i(t) = \Delta_i I(T_i \le t)$ 和 $Y_i(t) = I(T_i \ge t)$ 分别为计数过程和风险过程. 定义 $M_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda_0(s)e^{\boldsymbol{Z}_i'\boldsymbol{\beta}} ds$. 假设如下正则条件成立.

(A1) $\boldsymbol{\beta}$ 的参数空间 $\mathcal{B}$ 是紧的.

(A2) 协变量 $\boldsymbol{Z}$ 的取值空间 $\mathcal{Z}$ 是紧的.

(A3) 给定 Z_i 时, T_i 与 C_i 相互独立. ξ_i 与 (T_i, δ_i, Z_i) 相互独立.

(A4) 存在某个 $\alpha \in (0, 1)$, 使得 $\widetilde{\alpha} = \widetilde{n}/N \to \alpha$.

(A5) $\int_0^\tau \lambda_0(t)dt < \infty$

(A6) $\mathrm{P}(I(T_1 \ge t) = 1, \text{对任意 } t \in [0, \tau]) > 0$.

(A7) 对 $d = 0, 1, 2$, $\sup\limits_{\boldsymbol{\beta}\in\mathcal{B},\ t\in[0,\tau]} \left\| S_w^{(d)}(\boldsymbol{\beta}, t) - s^{(d)}(\boldsymbol{\beta}, t) \right\| \xrightarrow{P} 0$. 其中 $s^{(d)}(\boldsymbol{\beta}, t) = E\left[Y_1(t) e^{\boldsymbol{Z}_1'\boldsymbol{\beta}} \boldsymbol{Z}_1^{\otimes d} \right]$. $s^{(d)}(\boldsymbol{\beta}, t)$ 在 $t \in [0, \tau]$ 上一致的关于 $\boldsymbol{\beta}$ 连续, 并且 $s^{(0)}(\boldsymbol{\beta}, t)$ 下方有界且大于零.

(A8) 矩阵

$$\Sigma(\boldsymbol{\beta}_0) = \int_0^\tau \left[\frac{s^{(2)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}_0, t)} - \left\{ \frac{s^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}_0, t)} \right\}^{\otimes 2} \right] s^{(0)}(\boldsymbol{\beta}_0, t)\lambda_0(t)dt,$$

是正定的.

下面定理给出了 $\widehat{\boldsymbol{\beta}}_{IPW}$ 的渐近性质.

定理 1 ($\widehat{\boldsymbol{\beta}}_{IPW}$ 的渐近性质) 在正则条件 (A1)-(A8) 下, $\widehat{\boldsymbol{\beta}}_{IPW}$ 依概率收敛于 $\boldsymbol{\beta}_0$, 即:

$$\widehat{\boldsymbol{\beta}} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}_0.$$

进一步,

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{IPW} - \boldsymbol{\beta}_0) \xrightarrow{d} N\left(0,\ \Omega(\boldsymbol{\beta}_0)\right), \tag{3.3}$$

其中: 渐近方差矩阵 $\Omega(\boldsymbol{\beta}_0) = \Sigma^{-1}(\boldsymbol{\beta}_0)\left\{\Sigma_1(\boldsymbol{\beta}_0) + \Sigma_2(\boldsymbol{\beta}_0)\right\}\Sigma^{-1}(\boldsymbol{\beta}_0)$, $\Sigma_1(\boldsymbol{\beta}_0) = E\left(G_1(\boldsymbol{\beta}_0)^{\otimes 2}\right)$, $\Sigma_2(\boldsymbol{\beta}_0) = \frac{1-\alpha}{\alpha}E\left[(1-\Delta_1)G_1(\boldsymbol{\beta}_0)^{\otimes 2}\right]$, 这里

$$G_1(\boldsymbol{\beta}) = \int_0^\tau \left[Z_1 - \frac{s^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)}\right] dM_1(t).$$

此定理基于与 KangCai (2009)[18] 相似的讨论和证明思路即可证明.

进一步, 为了实际应用中的计算问题, 我们为渐近方差 $\Omega(\boldsymbol{\beta}_0)$ 提出如下一种相合估计. 定义:

$$\widehat{\Sigma}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\Delta_i\left[\frac{S_w^{(2)}(\boldsymbol{\beta}, T_i)}{S_w^{(0)}(\boldsymbol{\beta}, T_i)} - \left(\frac{S_w^{(1)}(\boldsymbol{\beta}, T_i)}{S_w^{(0)}(\boldsymbol{\beta}, T_i)}\right)^{\otimes 2}\right],$$

$$\widehat{G}_i(\boldsymbol{\beta}) = \Delta_i\left[Z_i - \frac{S_w^{(1)}(\boldsymbol{\beta}, T_i)}{S_w^{(0)}(\boldsymbol{\beta}, T_i)}\right] - \frac{1}{N}\sum_{l=1}^{N}\frac{w_l\Delta_l I(T_l \le T_i)e^{\boldsymbol{Z}_i'\boldsymbol{\beta}}}{S_w^{(0)}(\boldsymbol{\beta}, T_l)}\left[\boldsymbol{Z}_i - \frac{S_w^{(1)}(\boldsymbol{\beta}, T_l)}{S_w^{(0)}(\boldsymbol{\beta}, T_l)}\right],$$

$$\widehat{\Sigma}_1(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} w_i\widehat{G}_i(\boldsymbol{\beta})^{\otimes 2},$$

$$\widehat{\Sigma}_2(\boldsymbol{\beta}) = \frac{1-\widetilde{\alpha}}{\widetilde{\alpha}}\frac{1}{N}\sum_{i=1}^{N} w_i(1-\Delta_i)\widehat{G}_i(\boldsymbol{\beta})^{\otimes 2}.$$

则 $\Omega(\boldsymbol{\beta}_0)$ 可由 $\widehat{\Omega}(\widehat{\boldsymbol{\beta}}_{IPW}) = \widehat{\Sigma}^{-1}(\widehat{\boldsymbol{\beta}}_{IPW})\left(\widehat{\Sigma}_1(\widehat{\boldsymbol{\beta}}_{IPW}) + \widehat{\Sigma}_2(\widehat{\boldsymbol{\beta}}_{IPW}\right)\widehat{\Sigma}^{-1}(\widehat{\boldsymbol{\beta}}_{IPW})$ 估计.

3.2 时间相关权法

在逆概率加权思想基础上, Barlow (1994)[25] 考虑了一种随时间变化的加权思想 (time-varying weight, 以下简称 "TVW "). Borgan et al (2000)[24] 和 KulichLin (2004)[11] 将这种与时间相关的权分别应用于分层病例队列设计和两阶段抽样设计. 还有一些研究工作发展和建立了分析多类型疾病的病例队列数据的与时间相关加权 (TVW) 法 (KangCai, 2009[18]; KangCai, 2013[15]; Yan et al, 2017[19]). 受到上述研究的启发, 我们探讨如下与时间相关的权函数:

$$w_i(t) = \Delta_i + \frac{(1-\Delta_i)\xi_i}{\widehat{\alpha}(t)}, \tag{3.4}$$

其中

$$\widehat{\alpha}(t) = \frac{\sum_{i=1}^{N}(1-\Delta_i)\xi_i Y_i(t)}{\sum_{i=1}^{N}(1-\Delta_i)Y_i(t)}.$$

在 TVW 的加权思想下, 所有病例组的个体其权重均为 1; 而被选入子队列中的对照组个体权重为 $\widehat{\alpha}(t)$. 注意到 $\widehat{\alpha}(t)$ 的定义为: 在时刻 t, 入选子队列并仍在风险集中的对照组个体数与所有仍在风险集中的对照组个体个数之比. 故而, $\widehat{\alpha}(t)$ 为真实的子队列入选概率 $\widetilde{\alpha}$ 的一个估计. 基于上述 TVW 权, 在病例队列下, 对模型 (2.1) 中 $\boldsymbol{\beta}$ 的估计可建立如下加权得分方程,

$$\widetilde{U}_W(\boldsymbol{\beta}) = \sum_{i=1}^{N} \Delta_i \left[Z_i - \frac{\widetilde{S}_w^{(1)}(\boldsymbol{\beta}, T_i)}{\widetilde{S}_w^{(0)}(\boldsymbol{\beta}, T_i)} \right], \tag{3.5}$$

其中 $\widetilde{S}_w^{(d)}(\boldsymbol{\beta}, t) = \frac{1}{N} \sum\limits_{i=1}^{N} w_l(t) Y_l(t) e^{\boldsymbol{Z}_l'\boldsymbol{\beta}} \boldsymbol{Z}_l^{\otimes d}$. 由方程 (3.5) 求解得到的 $\boldsymbol{\beta}$ 的估计, 我们记为 $\widehat{\boldsymbol{\beta}}_{TVW}$. 以下, 我们研究和综述 $\widehat{\boldsymbol{\beta}}_{TVW}$ 的渐近理论. 我们修改条件 (A7) 为如下的条件 (A7'):

(A7') 对 $d = 0, 1, 2,$

$$\sup_{\boldsymbol{\beta} \in \mathcal{B},\ t \in [0,\tau]} \left\| \widetilde{S}_w^{(d)}(\boldsymbol{\beta}, t) - s^{(d)}(\boldsymbol{\beta}, t) \right\| \xrightarrow{P} 0.$$

其中 $s^{(d)}(\boldsymbol{\beta}, t) = E\left[Y_l(t) e^{\boldsymbol{Z}_l'\boldsymbol{\beta}} \boldsymbol{Z}_l^{\otimes d} \right]$. $s^{(d)}(\boldsymbol{\beta}, t)$ 在 $t \in [0, \tau]$ 上一致的关于 $\boldsymbol{\beta} \in \mathcal{B}$ 连续, 并且 $s^{(0)}(\boldsymbol{\beta}, t)$ 下方有界且大于零.

下面定理给出了 $\widehat{\boldsymbol{\beta}}_{TVW}$ 的渐近性质.

定理 2 ($\widehat{\boldsymbol{\beta}}_{TVW}$ 渐近性质) 在正则条件 (A1)-(A6),(A7') 和 (A8) 下, $\widehat{\boldsymbol{\beta}}_{TVW}$ 依概率收敛于 $\boldsymbol{\beta}_0$. 进一步,

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{TVW} - \boldsymbol{\beta}_0) \xrightarrow{d} N\left(0,\ \widetilde{\Omega}(\boldsymbol{\beta}_0)\right), \tag{3.6}$$

其中渐近方差矩阵 $\widetilde{\Omega}(\boldsymbol{\beta}_0) = \Sigma^{-1}(\boldsymbol{\beta}_0) \left\{ \Sigma_1(\boldsymbol{\beta}_0) + \widetilde{\Sigma}_2(\boldsymbol{\beta}_0) \right\} \Sigma^{-1}(\boldsymbol{\beta}_0)$, $\widetilde{\Sigma}_2(\boldsymbol{\beta}) = \frac{1-\alpha}{\alpha} V_1(\boldsymbol{\beta}_0)$, 此处

$$V_1(\boldsymbol{\beta}) = Var\left((1 - \Delta_1) \int_0^{\tau} \left[\widetilde{R}_1(\boldsymbol{\beta}, t) - \frac{Y_1(t) E\left\{ (1 - \Delta_1) \widetilde{R}_1(\boldsymbol{\beta}, t) \right\}}{E\left\{ (1 - \Delta_1) Y_1(t) \right\}} \right] d\Lambda_0(t) \right),$$

其中 $\widetilde{R}_i(\boldsymbol{\beta}, t) = Y_i(t) \left[Z_i(t) - \frac{s^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \right] e^{\boldsymbol{\beta}' Z_i(t)}$.

此定理证明可参考 KangCai (2013)[15].

进一步, 我们提出渐近方差矩阵 $\widetilde{\Omega}(\boldsymbol{\beta}_0)$ 的一种估计方法. 我们采用非参数自助法 (Hjort, 1985[26]; Efron & Tibshirani, 1993[27]; Burr, 1994[28]) 来估计 $\widehat{\boldsymbol{\beta}}_{TVW}$ 的渐近方差. 其基本思想是通过对观测数据的重复抽样建立经验分布函数. 当相关分布未知时, 非参数自助法是一种应用广泛的计算估计值方差或标准差的数值计算方法. 具体地, 记 $Y_{obs} = \{\boldsymbol{X}_1, ..., \boldsymbol{X}_N\}$ 为一组观测数据, 其中 $\boldsymbol{X}_i = (T_i, \Delta_i, \boldsymbol{Z}_i)$. 从观测数据 $Y_{obs} = \{\boldsymbol{X}_1, ..., \boldsymbol{X}_N\}$ 中有放回地随机抽取一组新样本 $Y_{obs}^* = \{\boldsymbol{X}_1^*, ..., \boldsymbol{X}_N^*\}$, 其中每个 $\boldsymbol{X}_i^*$ 的入样概率均为 $1/N$. 基于抽取的 Bootstrap 样本 Y_{obs}^*, 我们计算加权估计 $\widehat{\boldsymbol{\beta}}_{TVW}^*$. 独立地重复上述过程 B 次, 可得到 B 个 Bootstrap 估计 $\{\widehat{\boldsymbol{\beta}}_{TVW}^*(b)\}_{b=1}^B$, 其中 $\widehat{\boldsymbol{\beta}}_{TVW}^*(b) = (\widehat{\beta}_1^*(b), ..., \widehat{\beta}_p^*(b))^T$. 因此, $\widehat{\boldsymbol{\beta}}_{TVW}$ 的方差的第 k 个分量可由如下的样本方差估计:

$$\widehat{\mathrm{Var}}(\widehat{\beta}_k) = \frac{1}{B-1} \sum_{b=1}^{B} \left[\widehat{\beta}_k^*(b) - \frac{1}{B} \sum_{b=1}^{B} \widehat{\beta}_k^*(b) \right]^2, \quad k = 1, ..., p.$$

进一步地, 可以得到 $\widehat{\beta}_k$ 的 Wald 型 Bootstrap 置信区间. 当 $\{\widehat{\beta}_k^*(b)\}_{b=1}^B$ 近似正态分布时, $\widehat{\beta}_k$ 的 $(1-\alpha)100\%$ 置信区间为 $[\widehat{\beta}_k - z_{\alpha/2}\cdot\widehat{\mathrm{se}}(\widehat{\beta}_k),\ \widehat{\beta}_k + z_{\alpha/2}\cdot\widehat{\mathrm{se}}(\widehat{\beta}_k)]$, 其中 z_α 表示标准正态分布的上 α 分位数, $\widehat{\mathrm{se}}(\widehat{\beta}_k) = \sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_k)}$.

4 模拟研究

本节通过一系列模拟研究来评估这两种方法在有限样本量下的表现, 展示其在实际中的可行性和应用性. 我们考虑如下比例风险模型:

$$\lambda(t|Z_1, Z_2) = \lambda_0(t)\exp\{\beta_1 Z_1 + \beta_2 Z_2\}.$$

我们设定参数真值为 $\beta_1 = 0,\ 0.693$ 和 $\beta_2 = 0,\ -0.5$. 协变量 Z_1 从成功概率为 0.5 的 Bernoulli 分布中生成, Z_2 从标准正态分布中生成. 设定基准风险函数 $\lambda_0(t)$ 为 1 和 $2t$. 由此, 失效时间 $\widetilde{T}$ 可以分别由风险率为 $\exp\{\beta_1 Z_1 + \beta_2 Z_2\}$ 的指数分布和形状参数为 2、尺度参数为 $[\exp\{\beta_1 Z_1 + \beta_2 Z_2\}^{-1/2}$ 的 Weibull 分布中随机生成. 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 $\rho = 80\%$ 和 90%. 对于病例队列设计, 设定子队列的样本量分别为 $\widetilde{n} = 200$ 和 300, 从样本总量为 $N = 1000$ 的全队列中随机地抽取获得.

我们关注以下几个问题: 第一, 使用病例队列设计替代简单随机抽样设计能提高多少效率? 第二, 逆概率加权法和与时间相关权法这两种推断方法在有限样本量下表现如何? 第三, 两种推断方法的估计效率是否不同? 因此, 我们比较了以下几种方法:

Full: 全队列下基于比例风险模型的偏似然估计法 ($\widehat{\boldsymbol{\beta}}_{Full}$);

Naive: 基于与病例队列样本相同样本量的简单随机抽样下的偏似然估计法 ($\widehat{\boldsymbol{\beta}}_{Naive}$);

IPW: 病例队列设计下的逆概率加权法 ($\widehat{\boldsymbol{\beta}}_{IPW}$);

TVW: 病例队列设计下与时间相关权法 ($\widehat{\boldsymbol{\beta}}_{TVW}$).

在每种参数设定下, 比较上述四种方法参数估计值相较于真值的偏差 (Bias), 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). 所有结果均基于 1000 次的模拟获得. 模拟结果请见表 1-4.

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%. 从结果来看, $\widehat{\boldsymbol{\beta}}_{IPW}$ 和 $\widehat{\boldsymbol{\beta}}_{TVW}$ 的表现相似, $\widehat{\boldsymbol{\beta}}_{IPW}$ 比 $\widehat{\boldsymbol{\beta}}_{TVW}$ 的效率稍高一点. 这说明 IPW 方法由于依据的是真实抽样概率 $\widetilde{\alpha}$, 其估计要比依据真实抽样概率估计 $\widehat{\alpha}(t)$ 的 TVW 方法更有效一些, 因为 $\widehat{\alpha}(t)$ 仅通过删失个体信息得到. 特别地, 对于高删失率的情况, 两者的差距要更小一些. 从相对效率 (REs) 来看, $\widehat{\boldsymbol{\beta}}_{IPW}$ 和 $\widehat{\boldsymbol{\beta}}_{TVW}$ 均比 $\widehat{\boldsymbol{\beta}}_{Naive}$ 更有效. 这说明相较于简单随机抽样设计, 病例队列设计能有效地提高估计的效率. 例如: 关于 β_1 的估计, 当 $\widetilde{n} = 200,\ \lambda_0(t) = 1,\ \beta_1 = 0,\ \beta_2 = -0.5,\ \rho = 90\%$ 时, 相较于 $\widehat{\boldsymbol{\beta}}_{Full}$, $\widehat{\boldsymbol{\beta}}_{Naive}$, $\widehat{\boldsymbol{\beta}}_{IPW}$ 和 $\widehat{\boldsymbol{\beta}}_{TVW}$ 的相对效率分别为 0.30, 0.61 和 0.70, 即: 一方面, 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下的 2 倍; 另一方面, 病例队列设计仅用了约全队列 30% 的样本量, 却达到了 60% 多的效率. 再比如, 关于 β_2 的估计, 当 $\widetilde{n} = 300,\ \lambda_0(t) = 2t,\ \beta_1 = 0.693,\ \beta_2 = -0.5,\ \rho = 90\%$ 时, 相较于 $\widehat{\boldsymbol{\beta}}_{Full}$, $\widehat{\boldsymbol{\beta}}_{Naive}$, $\widehat{\boldsymbol{\beta}}_{IPW}$ 和 $\widehat{\boldsymbol{\beta}}_{TVW}$ 的相对效率分别为 0.35, 0.68 和 0.67, 即病例队列设计仅用约全队列 35% 的样本量实现了相较于简单随机抽样设计的两倍效率.

总体来说, 删失率较高 ($\rho = 90\%\ v.s.\ 80\%$) 或子队列样本量更大 ($\widetilde{n} = 200\%\ v.s.\ 300$) 时, $\widehat{\boldsymbol{\beta}}_{IPW}$ 和 $\widehat{\boldsymbol{\beta}}_{TVW}$ 的效率更高. 另外, 当参数真值更接近于 0 时 ($\beta_1 = 0.693\ v.s.\ \beta_1 = 0$, $\beta_2 = -0.5\ v.s.\ \beta_2 = 0$), 两种估计方法的效率也有小幅度的提高, 其中 $\widehat{\boldsymbol{\beta}}_{TVW}$ 相较于 $\widehat{\boldsymbol{\beta}}_{IPW}$ 提高得尤为明显. 例如: 当 $\widetilde{n} = 200$, $\rho = 90\%$, $\lambda_0(t) = 1$ 时, 相较于全样本估计, $\widehat{\boldsymbol{\beta}}_{TVW}$ 在 $\beta_1 = 0$ 和 $\beta_2 = 0$ 时的效率比在 $\beta_1 = 0.693$ 和 $\beta_2 = -0.5$ 时的效率提高了约 20%.

表 1 : 参数 β_1 和 β_2 的模拟结果, 其中子队列样本量为 $\widetilde{n} = 200$, 基准风险函数 $\lambda_0(t) = 1$.

			$\widehat{\beta}_1$					$\widehat{\beta}_2$				
ρ	(β_1,β_2)	Method	Bias	SD	SE	CP	RE	Bias	SD	SE	CP	RE
80%	$(0.693, -0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0073	0.0749	0.0744	0.940	1.00	0.0026	0.1461	0.1452	0.958	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0022	0.1248	0.1263	0.954	0.35	−0.0047	0.2511	0.2445	0.949	0.34
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0183	0.1120	0.1072	0.934	0.44	−0.0013	0.2159	0.2132	0.946	0.46
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0068	0.1350	0.1251	0.923	0.30	−0.0014	0.2206	0.2232	0.949	0.44
	$(0.693, 0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0064	0.0758	0.0747	0.945	1.00	0.0046	0.1473	0.1432	0.948	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	−0.0028	0.1260	0.1267	0.949	0.36	−0.0271	0.2443	0.2414	0.944	0.36
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0068	0.1107	0.1080	0.931	0.47	−0.0039	0.2263	0.2120	0.934	0.42
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0011	0.1391	0.1253	0.929	0.30	−0.0005	0.2146	0.2159	0.948	0.47
	$(0, -0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0015	0.0728	0.0713	0.946	1.00	0.0010	0.1457	0.1457	0.948	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	−0.0001	0.1187	0.1190	0.944	0.38	−0.0053	0.2485	0.2457	0.950	0.34
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0007	0.1050	0.1034	0.950	0.48	−0.0058	0.2043	0.2057	0.954	0.51
		$\widehat{\boldsymbol{\beta}}_{TVW}$	−0.0035	0.1055	0.1069	0.938	0.48	0.0020	0.2192	0.2167	0.937	0.44
	$(0, 0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0017	0.0727	0.0714	0.950	1.00	0.0026	0.1463	0.1426	0.947	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0044	0.1265	0.1197	0.937	0.33	0.0063	0.2477	0.2403	0.944	0.35
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0007	0.1027	0.1023	0.959	0.50	0.0018	0.2123	0.2031	0.946	0.47
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0009	0.1030	0.1061	0.946	0.50	−0.0024	0.2091	0.2080	0.942	0.49
90%	$(0.693, -0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0034	0.1043	0.1032	0.946	1.00	0.0022	0.2118	0.2075	0.947	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0086	0.2164	0.1998	0.937	0.23	−0.0370	0.4114	0.4055	0.953	0.27
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0177	0.1398	0.1386	0.941	0.57	−0.0151	0.2760	0.2693	0.942	0.59
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0202	0.1558	0.1543	0.925	0.45	−0.0109	0.2790	0.2798	0.936	0.58
	$(0.693, 0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0067	0.1049	0.1020	0.951	1.00	0.0089	0.2046	0.2001	0.951	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0115	0.1979	0.1949	0.940	0.28	−0.0110	0.4192	0.3876	0.941	0.24
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0241	0.1446	0.1370	0.938	0.53	−0.0008	0.2707	0.2629	0.944	0.57
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0219	0.1512	0.1480	0.932	0.48	−0.0030	0.2709	0.2708	0.938	0.57
	$(0, -0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0022	0.1047	0.1004	0.938	1.00	0.0040	0.2101	0.2070	0.951	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0018	0.1915	0.1884	0.941	0.30	−0.0302	0.3966	0.4004	0.961	0.28
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0022	0.1340	0.1275	0.945	0.61	−0.0057	0.2569	0.2554	0.948	0.67
		$\widehat{\boldsymbol{\beta}}_{TVW}$	−0.0038	0.1250	0.1321	0.949	0.70	−0.0181	0.2598	0.2644	0.941	0.65
	$(0, 0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0045	0.1063	0.1015	0.934	1.00	0.0083	0.2117	0.2042	0.947	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0009	0.1996	0.1926	0.931	0.28	0.0140	0.4229	0.3997	0.945	0.25
		$\widehat{\boldsymbol{\beta}}_{IPW}$	−0.0060	0.1305	0.1280	0.940	0.66	−0.0027	0.2580	0.2527	0.953	0.67
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0038	0.1254	0.1320	0.949	0.72	0.0116	0.2485	0.2582	0.953	0.73

表 2：参数 β_1 和 β_2 的模拟结果, 其中子队列样本量为 $\tilde{n} = 300$, 基准风险函数 $\lambda_0(t) = 1$.

			$\hat{\beta}_1$					$\hat{\beta}_2$				
ρ	(β_1,β_2)	Method	Bias	SD	SE	CP	RE	Bias	SD	SE	CP	RE
80%	$(0.693,-0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	−0.0000	0.0758	0.0747	0.947	1.00	−0.0000	0.1436	0.1457	0.953	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0015	0.1149	0.1136	0.947	0.44	0.0012	0.2257	0.2209	0.948	0.40
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0093	0.1017	0.0962	0.926	0.56	−0.0088	0.1986	0.1886	0.935	0.52
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0050	0.1070	0.1071	0.939	0.50	0.0075	0.1903	0.1946	0.949	0.57
	$(0.693,0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0000	0.0757	0.0751	0.952	1.00	−0.0000	0.1423	0.1436	0.956	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0047	0.1156	0.1148	0.947	0.43	−0.0019	0.2239	0.2193	0.960	0.40
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0072	0.0981	0.0973	0.953	0.60	0.0003	0.1902	0.1872	0.954	0.56
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0056	0.1092	0.1053	0.940	0.48	0.0041	0.1847	0.1896	0.951	0.59
	$(0,-0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	−0.0011	0.0726	0.0713	0.947	1.00	−0.0028	0.1411	0.1461	0.956	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0010	0.1122	0.1082	0.942	0.42	−0.0030	0.2210	0.2220	0.952	0.41
		$\widehat{\boldsymbol{\beta}}_{IPW}$	−0.0005	0.0921	0.0916	0.955	0.62	−0.0055	0.1839	0.1830	0.954	0.59
		$\widehat{\boldsymbol{\beta}}_{TVW}$	−0.0036	0.0935	0.0932	0.935	0.60	−0.0014	0.1931	0.1901	0.932	0.53
	$(0,0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	−0.0016	0.0720	0.0714	0.948	1.00	−0.0005	0.1392	0.1429	0.953	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0028	0.1112	0.1083	0.939	0.42	0.0043	0.2145	0.2177	0.955	0.42
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0034	0.0905	0.0909	0.953	0.63	0.0055	0.1824	0.1805	0.952	0.58
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0062	0.0914	0.0931	0.938	0.62	−0.0094	0.1804	0.1824	0.942	0.60
90%	$(0.693,-0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0000	0.1057	0.1032	0.940	1.00	0.0000	0.2064	0.2074	0.948	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0134	0.1721	0.1720	0.948	0.38	−0.0163	0.3549	0.3483	0.954	0.34
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0189	0.1265	0.1263	0.941	0.70	−0.0090	0.2479	0.2464	0.954	0.69
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0099	0.1323	0.1326	0.929	0.64	−0.0127	0.2511	0.2521	0.937	0.68
	$(0.693,0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	−0.0000	0.1048	0.1021	0.946	1.00	−0.0000	0.2035	0.2004	0.951	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0050	0.1683	0.1704	0.955	0.39	0.0119	0.3429	0.3355	0.948	0.35
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0101	0.1251	0.1251	0.944	0.70	0.0019	0.2422	0.2391	0.947	0.71
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0044	0.1282	0.1314	0.934	0.67	0.0030	0.2342	0.2437	0.942	0.76
	$(0,-0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	−0.0000	0.1027	0.1001	0.937	1.00	−0.0000	0.2032	0.2072	0.958	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	−0.0011	0.1750	0.1639	0.934	0.34	−0.0100	0.3503	0.3468	0.957	0.34
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0051	0.1171	0.1167	0.957	0.77	−0.0074	0.2314	0.2361	0.960	0.77
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0040	0.1209	0.1193	0.933	0.72	−0.0024	0.2405	0.2414	0.932	0.71
	$(0,0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0017	0.1021	0.1014	0.945	1.00	0.0018	0.2078	0.2040	0.953	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0011	0.1744	0.1668	0.941	0.34	−0.0089	0.3483	0.3424	0.960	0.36
		$\widehat{\boldsymbol{\beta}}_{IPW}$	−0.0051	0.1196	0.1168	0.941	0.73	0.0022	0.2452	0.2339	0.949	0.72
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0002	0.1154	0.1200	0.947	0.78	−0.0019	0.2333	0.2377	0.944	0.79

表 3：参数 β_1 和 β_2 的模拟结果, 其中子队列样本量为 $\widetilde{n}=200$, 基准风险函数 $\lambda_0(t)=2t$.

			$\widehat{\beta}_1$					$\widehat{\beta}_2$				
ρ	(β_1,β_2)	Method	Bias	SD	SE	CP	RE	Bias	SD	SE	CP	RE
80%	$(0.693,-0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0060	0.0759	0.0757	0.950	1.00	−0.0012	0.1412	0.1458	0.964	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0055	0.1307	0.1285	0.942	0.34	−0.0039	0.2426	0.2468	0.961	0.34
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0194	0.1153	0.1097	0.919	0.43	−0.0048	0.2270	0.2214	0.934	0.39
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0001	0.1394	0.1474	0.930	0.30	0.0120	0.2126	0.2396	0.990	0.44
	$(0.693,0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0038	0.0779	0.0753	0.945	1.00	0.0023	0.1423	0.1416	0.954	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0154	0.1311	0.1272	0.938	0.35	0.0040	0.2519	0.2391	0.932	0.32
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0184	0.1166	0.1090	0.920	0.45	−0.0020	0.2359	0.2180	0.929	0.36
		$\widehat{\boldsymbol{\beta}}_{TVW}$	−0.0123	0.1930	0.1615	0.910	0.16	0.0081	0.2057	0.2236	0.950	0.48
	$(0,-0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	−0.0020	0.0709	0.0714	0.950	1.00	−0.0045	0.1474	0.1457	0.949	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	−0.0064	0.1229	0.1195	0.950	0.33	−0.0036	0.2538	0.2459	0.948	0.34
		$\widehat{\boldsymbol{\beta}}_{IPW}$	−0.0010	0.1135	0.1091	0.944	0.39	−0.0173	0.2252	0.2187	0.941	0.43
		$\widehat{\boldsymbol{\beta}}_{TVW}$	−0.0052	0.1089	0.1127	0.970	0.42	0.0242	0.2611	0.2448	0.940	0.32
	$(0,0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	−0.0026	0.0712	0.0710	0.956	1.00	−0.0016	0.1402	0.1419	0.952	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	−0.0016	0.1197	0.1192	0.945	0.35	0.0000	0.2456	0.2388	0.946	0.33
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0014	0.1169	0.1086	0.927	0.37	0.0088	0.2139	0.2160	0.952	0.43
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0036	0.1174	0.1139	0.970	0.37	0.0549	0.2284	0.2222	0.940	0.38
90%	$(0.693,-0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0073	0.1062	0.1035	0.947	1.00	−0.0057	0.2090	0.2063	0.956	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0253	0.2087	0.1994	0.945	0.26	−0.0417	0.3959	0.4024	0.969	0.28
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0282	0.1564	0.1417	0.923	0.46	−0.0235	0.2841	0.2801	0.945	0.54
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0377	0.1775	0.1607	0.900	0.36	0.0101	0.3293	0.2931	0.910	0.40
	$(0.693,0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0073	0.1090	0.1051	0.941	1.00	0.0035	0.2029	0.2035	0.956	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0105	0.2095	0.2051	0.939	0.27	−0.0032	0.4171	0.4003	0.947	0.24
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0133	0.1513	0.1433	0.932	0.52	−0.0050	0.2817	0.2776	0.947	0.52
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0150	0.1920	0.1680	0.890	0.32	−0.0316	0.2938	0.2902	0.970	0.48
	$(0,-0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	−0.0048	0.1030	0.1007	0.939	1.00	−0.0134	0.2055	0.2080	0.948	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0045	0.1977	0.1919	0.946	0.27	−0.0349	0.4233	0.4061	0.952	0.24
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0004	0.1372	0.1356	0.951	0.56	−0.0139	0.2763	0.2700	0.948	0.55
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0041	0.1443	0.1440	0.960	0.51	0.0445	0.2825	0.2897	0.960	0.53
	$(0,0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	−0.0018	0.1015	0.1010	0.939	1.00	−0.0077	0.2015	0.2030	0.946	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0013	0.1995	0.1929	0.941	0.26	−0.0050	0.3916	0.3962	0.966	0.26
		$\widehat{\boldsymbol{\beta}}_{IPW}$	−0.0009	0.1373	0.1340	0.932	0.55	0.0066	0.2590	0.2661	0.956	0.61
		$\widehat{\boldsymbol{\beta}}_{TVW}$	−0.0334	0.1560	0.1440	0.900	0.42	−0.0259	0.2701	0.2785	0.970	0.56

表 4：参数 β_1 和 β_2 的模拟结果, 其中子队列样本量为 $\widetilde{n}=300$, 基准风险函数 $\lambda_0(t)=2t$.

			$\widehat{\beta}_1$					$\widehat{\beta}_2$				
ρ	(β_1,β_2)	Method	Bias	SD	SE	CP	RE	Bias	SD	SE	CP	RE
80%	$(0.693,-0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0037	0.0776	0.0759	0.945	1.00	−0.0017	0.1495	0.1458	0.946	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0142	0.1217	0.1164	0.929	0.41	−0.0076	0.2270	0.2225	0.956	0.43
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0099	0.1038	0.0981	0.936	0.56	−0.0044	0.2010	0.1947	0.941	0.55
		$\widehat{\boldsymbol{\beta}}_{TVW}$	−0.0035	0.1182	0.1233	0.950	0.43	0.0153	0.2000	0.2051	0.940	0.56
	$(0.693,0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0043	0.0779	0.0751	0.938	1.00	−0.0012	0.1446	0.1417	0.946	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0000	0.1171	0.1149	0.954	0.44	0.0000	0.2073	0.2155	0.965	0.49
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0120	0.1030	0.0978	0.925	0.57	0.0037	0.1934	0.1911	0.953	0.56
		$\widehat{\boldsymbol{\beta}}_{TVW}$	−0.0054	0.1502	0.1265	0.930	0.27	−0.0269	0.1793	0.1932	0.960	0.65
	$(0,-0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	−0.0000	0.0725	0.0714	0.936	1.00	0.0000	0.1414	0.1460	0.956	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0072	0.1105	0.1083	0.941	0.43	−0.0065	0.2198	0.2222	0.952	0.41
		$\widehat{\boldsymbol{\beta}}_{IPW}$	−0.0012	0.0954	0.0953	0.952	0.58	−0.0074	0.1899	0.1919	0.952	0.55
		$\widehat{\boldsymbol{\beta}}_{TVW}$	−0.0168	0.1019	0.0989	0.940	0.51	0.0284	0.1983	0.2047	0.950	0.51
	$(0,0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	−0.0000	0.0714	0.0711	0.948	1.00	0.0000	0.1415	0.1423	0.946	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	−0.0016	0.1055	0.1076	0.954	0.45	0.0004	0.2145	0.2160	0.955	0.44
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0036	0.0938	0.0947	0.952	0.57	−0.0009	0.1932	0.1888	0.937	0.54
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0067	0.1003	0.0967	0.940	0.50	0.0176	0.1919	0.1922	0.960	0.54
90%	$(0.693,-0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	0.0023	0.1064	0.1039	0.945	1.00	−0.0114	0.2104	0.2064	0.952	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0070	0.1752	0.1750	0.942	0.37	−0.0135	0.3535	0.3478	0.953	0.35
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0167	0.1364	0.1286	0.936	0.61	−0.0073	0.2560	0.2526	0.952	0.68
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0019	0.1538	0.1365	0.900	0.48	−0.0249	0.2573	0.2613	0.900	0.67
	$(0.693,0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	−0.0003	0.1073	0.1050	0.949	1.00	−0.0046	0.2094	0.2037	0.945	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	0.0098	0.1717	0.1757	0.959	0.39	−0.0075	0.3533	0.3423	0.949	0.35
		$\widehat{\boldsymbol{\beta}}_{IPW}$	0.0090	0.1345	0.1292	0.943	0.64	0.0057	0.2625	0.2501	0.946	0.64
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0020	0.1443	0.1456	0.930	0.55	0.0254	0.2231	0.2561	0.950	0.88
	$(0,-0.5)$	$\widehat{\boldsymbol{\beta}}_{Full}$	−0.0029	0.1020	0.1005	0.940	1.00	−0.0088	0.2090	0.2083	0.955	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	−0.0010	0.1767	0.1673	0.936	0.33	−0.0163	0.3589	0.3511	0.949	0.34
		$\widehat{\boldsymbol{\beta}}_{IPW}$	−0.0062	0.1224	0.1224	0.951	0.69	−0.0029	0.2431	0.2461	0.952	0.74
		$\widehat{\boldsymbol{\beta}}_{TVW}$	0.0054	0.1273	0.1258	0.950	0.64	−0.0031	0.2644	0.2612	0.930	0.62
	$(0,0)$	$\widehat{\boldsymbol{\beta}}_{Full}$	−0.0027	0.1009	0.1009	0.946	1.00	−0.0045	0.2060	0.2035	0.949	1.00
		$\widehat{\boldsymbol{\beta}}_{Naive}$	−0.0046	0.1650	0.1662	0.940	0.37	0.0057	0.3680	0.3417	0.941	0.31
		$\widehat{\boldsymbol{\beta}}_{IPW}$	−0.0024	0.1235	0.1219	0.948	0.67	−0.0064	0.2496	0.2420	0.940	0.68
		$\widehat{\boldsymbol{\beta}}_{TVW}$	−0.0060	0.1218	0.1263	0.990	0.69	0.0069	0.2333	0.2488	0.960	0.78

5 实际数据分析

本节通过分析一个肾母细胞瘤的病例队列数据和一个员工离职管理的队列数据, 来展示病例队列设计以及其所探讨的两种加权估计推断方法在实际中的应用.

5.1 肾母细胞瘤

肾母细胞瘤是一种多发于幼儿的罕见肾癌. 美国国家肾母细胞瘤研究组织 (NWTSG) 进

行了一系列研究来探索患儿的肿瘤组织学类型与发病时间之间的联系. 我们研究的数据来自于 NWTSG 的第三次、第四次临床试验的数据 (BreslowChatterjee, 1999[29]; D'Angio et al, 1989[30]; Green et al, 1998[31]), 其中包含 4028 个幼儿的记录.

患儿的肿瘤组织学类型是最重要的暴露因素. 地方医疗机构的病理学家在患儿最初接受治疗时, 对其肿瘤组织学类型进行初步评估. 然后, 来自 NWTSG 病理中心的有经验的病理学家对其组织学类型进行确定性的评估. 后者的评估结果更加的准确, 但其费用也更加的昂贵且其过程更加的耗时. 因此, BreslowChatterjee (1999)[29] 为此研究提出了一个病例队列设计方案. 具体地, 从全队列 4028 个患儿中随机地抽取了 668 个患儿作为子队列. 病例队列样本由子队列以及子队列之外所有经历了发病或者死亡的患儿组成. 在 NWTSG 病理中心中仅对病例队列样本中的患儿肿瘤确定其组织类型. 我们应用本文研究的两种加权估计方程方法来分析这样一个病例队列数据.

我们感兴趣的因变量是患儿的发病时间, 而观测到的发病时间带有右删失, 其删失率约为 85.8%. 我们考虑三个潜在影响因素. 肿瘤组织学类型 (Histype) 分为两种类型: 一种是肿瘤由被称为"组织结构不良型" (Histype = 1) 的罕见细胞类型组成, 另一种是肿瘤由"组织结构良好型" (Histype = 0) 的细胞组成. 疾病分期 (Stage) 分为以下四种类型: 肿瘤广泛分布于肾脏并被完全切除 (Stage = 1), 肿瘤超出肾脏但完全切除 (Stage = 2), 腹腔中有残余肿瘤或淋巴结中有肿瘤 (Stage = 3), 癌细胞转移到肺或者肝脏 (Stage = 4). 确诊年龄 (Age) 以月为单位, 我们对其进行了中心标准化处理. 我们对数据中所有考虑的影响因素进行了描述性统计分析, 画出了组织学类型条形图, 疾病分期饼图以及确诊年龄直方图, 见图 1, 2 和 3.

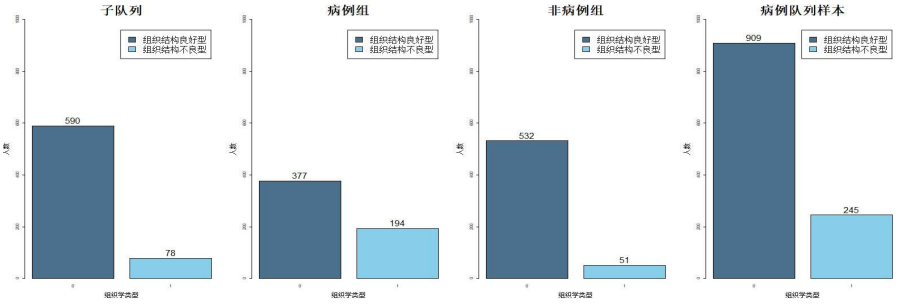

图 1: 组织学类型条形图

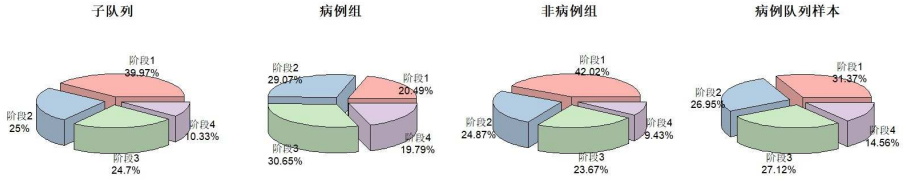

图 2: 疾病分期条形图

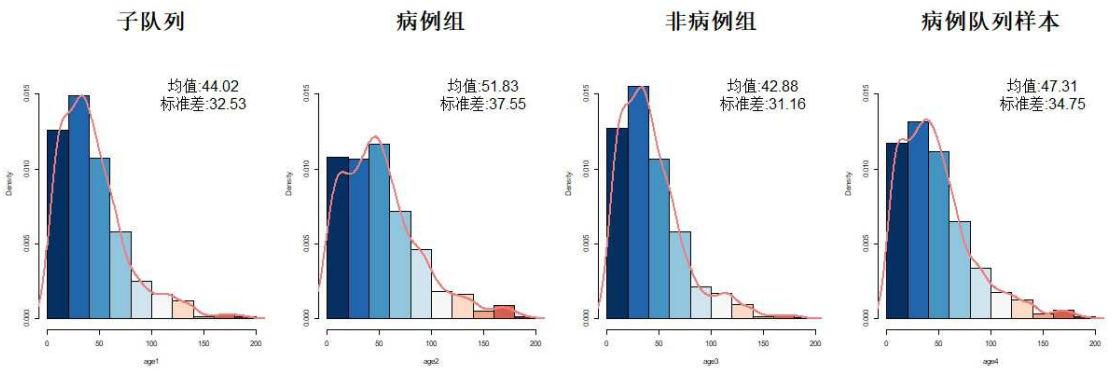

图 3: 确诊年龄直方图

我们的目的是评估肿瘤组织学类型, 疾病分期和确诊年龄对患儿发病时间 (T) 的影响. 我们考虑如下比例风险模型:

$$\lambda(t|Z) = \lambda_0(t)\exp\{\beta_1 \text{Histype} + \beta_2 \text{Stage} + \beta_3 \text{Age}\}.$$

在此模型框架下, 我们应用逆概率加权法 (IPW), 与时间相关权法 (TVW), 以及基于子队列样本的偏似然估计法 (SRS) 分析了肾母细胞瘤研究数据. 分析结果总结在表 5 中.

总体来说, 三种方法分析得到的估计结果是一致的. 结果表明, 肿瘤组织学类型对发病时间的影响是显著的. 组织结构不良型的患儿较之组织结构良好型的患儿发病或者死亡的风险更高, $e^{1.3489} = 3.9$ 倍 (SRS), $e^{1.3336} = 3.8$ 倍 (IPW) 或 $e^{1.3428} = 3.8$ 倍 (TVW). 结果也显示, 疾病分期程度越高的患儿发病和死亡的风险越高, 确诊年龄越小的患儿发病和死亡的风险越高. 在三种估计中, 病例队列设计下的 IPW 法和 TVW 法比简单随机抽样下的 SRS 法更有效. 考虑到肿瘤组织学类型的评估十分昂贵且耗时, 采用病例队列抽样设计能提高研究效率和节约成本.

表 5: 肾母细胞瘤研究数据分析结果

	SRS			IPW			TVW		
	Est.	SE	p-value	Est.	SE	p-value	Est.	SE	p-value
Histype	1.3489	0.2310	$< 0.0001^*$	1.3336	0.1301	$< 0.0001^*$	1.3428	0.1423	$< 0.0001^*$
Stage	0.2716	0.1076	0.0116^*	0.3438	0.0537	$< 0.0001^*$	0.3647	0.0590	$< 0.0001^*$
Age	0.1292	0.1078	0.2305	0.1093	0.0541	0.0435^*	0.1092	0.0643	0.0895

5.2 员工离职管理

在现代商业社会, 研究员工自身的特点与企业文化是否相适应对于稳定企业员工团队至关重要. 我们研究的数据来源于国内某大型商业银行人力资源部门 [32], 共有 1300 个样本, 每

个样本对应于该银行当年的一位销售人员. 我们关注的因变量是这些员工的工作年限 T (单位: 月), 在观测时间内有的员工已离职而有的员工仍在继续工作, 因此观测数据带有右删失, 删失率约为 22.69%. 为了展示病例队列抽样设计, 我们从全队列 1300 个样本中随机地抽取了 300 个作为子队列, 子队列以及子队列之外所有已离职的员工数据组成病例队列样本. 我们考虑比例风险模型, 研究可能与员工工作年限相关的三个潜在影响因素: 一个是员工户籍 X_1, $X_1 = 0$ 表示本地员工, $X_1 = 1$ 表示异地员工; 一个是员工性别 X_2, $X_2 = 0$ 表示性别为男, $X_2 = 1$ 表示性别为女; 还有一个是员工年龄 X_3, 我们对其作了中心标准化处理. 对数据中上述所有协变量进行了描述性统计分析, 结果见表 6.

我们应用逆概率加权法 (IPW), 与时间相关权法 (TVW) 以及基于相同样本量的简单随机抽样下的偏似然估计法 (Naive) 分析了员工离职数据. 分析结果总结在表 7 中.

表 6: 员工离职数据描述性统计分析表.

	子队列	病例组	非病例组	病例队列样本
户籍 (%)				
0 = 本地	40.67 (122/300)	39.70 (399/1005)	47.76 (32/67)	40.21 (431/1072)
1 = 异地	59.33 (178/300)	60.30 (606/1005)	52.24 (35/67)	59.79 (641/1072)
性别 (%)				
0 = 男	67.67 (203/300)	67.96 (683/1005)	46.27 (31/67)	66.60 (714/1072)
1 = 女	32.33 (97/300)	32.04 (322/1005)	53.73 (36/67)	33.40 (358/1072)
确诊年龄 (均值 ± 标准差)				
	28.12 ± 5.27	27.59 ± 5.07	28.96 ± 4.65	27.67 ± 5.06

表 7: 员工离职研究数据分析结果.

	Naive			IPW			TVW		
	Est.	SE	p-value	Est.	SE	p-value	Est.	SE	p-value
X_1	0.2662	0.0704	0.0002^*	0.2730	0.0676	0.0001^*	0.2979	0.1460	0.0094^*
X_2	−0.1995	0.0749	0.0077^*	−0.2868	0.0718	0.0001^*	−0.3718	0.1216	0.0022^*
X_3	−0.2497	0.0421	$< 0.0001^*$	−0.2210	0.0358	$< 0.0001^*$	−0.2493	0.0634	0.0001^*

总体来说, 三种分析方法得到的估计结果是一致的. 结果表明, 员工的户籍、性别和年龄特征都与其工作年限显著相关. 具体地, 异地员工相比于本地员工离职的风险更高, 分别约为 $e^{0.2662} = 1.3$ 倍 (Naive), $e^{0.2730} = 1.3$ 倍 (IPW) 或 $e^{0.2979} = 1.3$ 倍 (TVW); 男员工相比于女员工离职的风险更高, 分别约为 $e^{0.1995} = 1.2$ 倍 (Naive), $e^{0.2868} = 1.3$ 倍 (IPW) 或 $e^{0.3718} = 1.5$ 倍 (TVW); 年龄越小的员工离职的风险越高. 在三种估计中, 病例队列设计下的 IPW 法和 TVW 法比简单随机抽样下的 Naive 法更有效.

6 总结

病例队列设计作为应用最为广泛的有偏抽样机制之一, 能有效节约研究成本和提高研究效率. 本文研究了病例队列设计在比例风险模型下的应用, 探讨了逆概率加权和与时间相关

加权这两种思想下的估计方程方法, 并综述了其渐近理论. 然后, 我们重点研究上述两种方法在实际中的应用问题, 为上述两种方法编写了一套操作性强的计算方法与可实现这两种方法的应用程序. 模拟研究结果表明在病例队列设计下, 上述两种方法在有限样本下均表现优异, 其估计效率均明显高于传统简单随机抽样设计. 最后, 实际数据分析结果也展示了估计方法在实际中的应用价值.

为了进一步提高估计效率, 未来的工作包括研究基于最优权的推断方法. 相关的判别准则有赤池信息准则 (AIC) (Akaike, 1973[33]); 贝叶斯信息准则 (BIC) (Schwarz, 1978[34]) 或广义交叉验证 (CGV) (Craven & Wanba, 1978[35]). 当感兴趣事件发生率较低时, 病例队列设计的效果更好. 对于删失率较低的生存数据, 相关的研究包括广义病例队列设计 (Cai & Zeng, 2007[10]) 和基于因变量抽样设计 (Ding et al, 2014[36]; Yu et al, 2016[37]) 等等. 这些也将是我们未来的主要工作之一.

参 考 文 献

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

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

[3] Chen Kani, Lo Shaw-Wha. Case-cohort and case-control analysis with Cox's model [J]. Biometrika, 1999, 86: 755–764.

[4] Kang Suhyun, Lu Wenbin, Liu Mengling. Efficient estimation for accelerated failure time model under case-cohort and nested case-control sampling [J]. Biometrics, 2016, 73: 114–123.

[5] Liu Dandan, Cai Tianxi, Lok A, Yingye. Nonparametric maximum likelihood estimators of time-dependent accuracy measures for survival outcome under two-stage sampling designs [J]. Journal of the American Statistical Association, 2018, 113: 522, 882–892.

[6] Lin D Y, Ying Z. Cox regression with incomplete covariate measurements [J]. Journal of the American Statistical Association, 1993, 88: 1341–1349.

[7] Kulich M, Lin D Y. Additive hazards regression for case-cohort studies [J]. Biometrika, 2000, 87: 73–87.

[8] Sun J, Sun L, Flournoy N. Addictive hazards model for competing risks analysis of the case-cohort design [J]. Communication in Statistics – Theory and Method, 2004, 33: 351–366.

[9] Cai Jianwen, Zeng Donglin. Sample size/power calculation for case-cohort studies [J]. Biometric, 2004, 60: 1015–1024.

[10] Cai Jianwen, Zeng Donglin. Power calculation for case-cohort studies with nonrare events [J]. Biometrics, 2007, 63: 1288–1295.

[11] 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: 832–844.

[12] Kong Lan, Cai Jianwen, Sen Pranab K. Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design [J]. Biometrika, 2004, 91: 305–319.

[13] Lu Wenbin, Tsiatis Anastasios A. Semiparametric transformation models for the case-cohort study [J]. Biometrika, 2006, 93: 207–214.

[14] Breslow N E, Wellner J A. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression [J]. Scandinavian Journal of Statistics, 2007, 34: 86–102.

[15] Kang Sangwook, Cai Jianwen, 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: 28–41.

[16] Steingrimsson J A, Strawderman R L. Estimation in the semiparametric accelerated failure time model with missing covariates: improving efficiency through augmentation [J]. Journal of the American Statistical Association, 2017, 112: 519, 1221–1235.

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

[18] Kang Sangwook, Cai Jianwen. Marginal hazard model for case-cohort studies with multiple disease outcomes [J]. Biometrika, 2009, 96: 887–901.

[19] Yan Ying, Zhou Haibo, Cai Jianwen. Improving efficiency of parameter estimation in case-cohort studies with multivariate failure time data [J]. Biometrics, 2017, 73: 1042–1052.

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

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

[22] Horvitz D G, Thompson D J. A generalization of sampling without replacement from a finite universe [J]. Journal of the American Statistical Association, 1951, 47: 663–685.

[23] Kalbfleisch J D, Lawless J F. Likelihood analysis of multi-state models for disease incidence and mortality [J]. Statistics in Medicine, 1988, 7: 147–160.

[24] Borgan O, Langholz B, Samuelsen S, Goldstein L, Pogoda J. Exposure stratified case-cohort designs [J]. Lifetime Data Analysis, 2000, 6: 86–102.

[25] Barlow W. Robust variance estimation for the case-cohort design [J]. Biometrics, 1994, 50: 1064–72.

[26] Hjort N. Bootstrapping Cox's Regression Model [R]. California: Stanford University, Dept. of Statistics, 1985.

[27] Efron B, Tibshirani R. An Introduction to the Bootstrap [J]. Journal of Educational and Behavioral Statistics, 1993, 22: 245–245.

[28] Burr D. A Comparison of Certain Bootstrap Confidence Intervals in the Cox Model [J]. Journal of the American Statistical Association, 1994, 89: 1290–1302.

[29] Breslow N E, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis [J]. Journal of the Royal Statistical Society, Series C, 1999, 48: 457–468.

[30] D'Angio G J, Breslow N, Bechwith J B, et al. Treatment of Wilms' tumor, results of the third national Wilms' tumor study [J]. Cancer, 1989, 64: 349–360.

[31] Green D M, Breslow N E, Bechwith J B, et al. Comparison between single-dose and divided-dose administration of dactinomycin and doxorubicin for patients with Wilms' tumor: a report from the national Wilms' tumor study group [J]. Journal of Clinical Oncology, 1998, 16: 237–245.

[32] 王汉生. 商务数据分析与应用 [M]. 北京: 中国人民大学出版社, 2011.

[33] Akaike H. Information theory and an extension of the maximum likelihood principle [J]. In B. N. Petrov and F. Caski (Eds.), Proceedings of the Second International Symposium on Information Theory, 1973, 267-281.

[34] Schwarz G. Estimating the dimension of a model [J]. The Annals of Statistics, 1978, 6: 461–464.

[35] Craven P, Wahba G. Smoothing noisy data with spline functions [J]. Numerische mathematik, 1978, 31: 377–403.

[36] Ding J, Zhou H, Liu Y, Cai J, Longnecker M P. Estimating effect of environmental contaminants on women's subfecundity for the MoBa study data with an outcome-dependent sampling scheme [J]. Biostatistics, 2014, 15: 636–650.

[37] Yu J, Liu Y, Cai J, Sandler D P, Zhou H. Outcome-dependent sampling design and inference for Cox's proportional hazards Model [J]. Journal of Statistical Planning and Inference, 2016, 178: 24–36.

INFERENCE AND APPLICATION OF CASE-COHORT DESIGN UNDER THE PROPORTIONAL HAZARDS MODEL

ZHANG Jia-qian[1], DENG Li-feng[2], DING Jie-li[1]

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

(*2. College of Mathematics and Systems Science, Shandong University of Science and Technology, Qingdao, Shandong, 266590, China*)

Abstract: A case-cohort design is a cost-effective sampling scheme in large cohort studies. The key idea of such a design is to assemble the measurements of expensive covariates only on a subset of the entire cohort and all the subjects outside the subcohort that experience the event of interest. In this paper, we study the inference methods for case-cohort data under the Cox model. We consider two weighted estimating equation approaches, the inverse-probability and time-varying weighted methods. The asymptotic theories are established. A series of simulation studies are conducted to assess the finite-sample performance of the proposed methods and exhibit the superiority and efficiency of the case-cohort design. Some real data examples are analyzed to illustrate the application of the proposed methods.

Keywords: Case-Cohort design; proportional hazards model; inverse probability weight; time varying weight

2010 MR Subject Classification: 62D05; 62N01; 62N02