项目编号 201410486023

武汉大学国家大学生创新创业训练计划项目结题报告

病例队列设计及其应用

院（系）名 称：数学与统计学院

专 业 名 称 ：统计学

学 生 姓 名 ：吕孝丹 高蓝 金凌

罗正宇 汪源 周慧娟

指 导 教 师 ：丁洁丽 副教授

二〇一五年九月

FINAL REPORT OF PLANNING PROJECT OF INNOVATION AND ENTREPRENEURSHIP TRAINING OF NATIONAL UNDERGRADUATE OF WUHAN UNIVERSITY

Case-Cohort Design and Its Application

College : School of Mathematics and Statistics

Subject : Statistics

Name : LYU Xiaodan, GAO Lan, JIN Ling, LUO Zhengyu, WANG Yuan, ZHOU Huijuan

Director : DING Jieli Associate Professor

Sep 2015

郑 重 声 明

本项目组呈交的中期报告，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我们所知，除文中已经注明引用的内容外，本报告的研究成果不包含他人享有著作权的内容。对本报告所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本报告的知识产权归属于培养单位。

项目组签名：_____ 日期：_____

导师签名：_____ 日期：_____

摘　要

病例队列设计是 Prentice 在 1986 年提出的一种有偏抽样机制，在生存分析尤其是流行病学中应用广泛。该方法能够减少需要观测协变量的研究对象的个数，从而在协变量的观测十分昂贵的情况下比传统的简单随机抽样更节约资源。且在相同样本量下，病例队列分析比传统方法得到的估计量的均方误差更小从而估计更有效。由于研究条件的限制，很多研究收集到的生存数据经常发生删失。Cox 比例风险模型是生存分析中处理右删失数据的常用方法之一，而在数据删失率较高的情况下，病例队列设计下的 Cox 比例风险模型仍十分有效。

本文研究了在 Cox 比例风险模型的基础上病例队列样本和简单随机样本的参数估计，推导了标准差的一种简单估计即经验估计，并将其与自助抽样法得到的标准差估计进行比较。统计模拟的结果说明了病例队列设计的参数估计是无偏的且标准差的经验估计在删失率较高、样本量较大时接近无偏。此外，模拟结果多方面验证了数据删失率较高的情况下病例队列设计比简单随机抽样效率更高。最后我们将病例队列分析方法应用于美国国家肾母细胞瘤研究数据，模型拟合结果表明肿瘤细胞的组织学类型以及肿瘤阶段是影响患者复发时间的显著变量。

关键词：病例队列研究；Cox 比例风险模型；生存分析；临床设计；统计模拟

ABSTRACT

Case cohort design is a biased sampling method put forward by Prentice in 1986，and it’s widely used in survival analysis and epidemiology. Such a design requires less observations of covariates so that it’s especially resource-saving when measuring the covariates is costly. On the other hand, with the same sample size, case cohort study has a smaller mean square error than the traditional simple sampling design. Therefore its estimators are more efficient. It happens a lot that the survival data collected in the study is censored. And in survival analysis, Cox proportional hazard model is one of the most frequently used models dealing with right-censored data. Even if the censoring rate is extremely high, the estimators of Cox model under case-cohort design are still efficient.

We have studied the parameter estimation of case-cohort sample and simple random sample base on Cox model. And we derived the empirical estimator of standard deviation, which simplified the computation compared with the Bootstrap. Our result of statistical simulation shows that for case-cohort study the estimators are unbiased and the empirical estimator is approximately unbiased under high censoring rate and large sample. Moreover, when the censoring rate is high, case-cohort design proves to be more efficient than simple random sampling in many ways. At last, we applied case-cohort design to the data from the clinical trials of the National Wilms Tumor Study and found that histology and tumor stage were significant predictors of the relapse time.

Key words: case cohort study; Cox proportional hazard model; survival analysis; clinical design; statistical simulation

目　录

第 1 章　绪论

1.1　研究背景

Prentice [1] 于1986年在流行病学的队列研究和疾病预防实验领域提出了病例队列设计研究。流行病学研究着重探求疾病的产生于各个重要的协变量之间的关系，即对协变量参数的估计十分感兴趣。在此类研究中，常常从被研究对象中抽取简单随机样本，并利用传统方法如极大似然估计、最小二乘估计等得到参数估计。而很多协变量的观测是非常困难或昂贵的，为了节约成本、提高效率，研究人员往往采取非简单随机抽样的方式减少需要观测协变量的研究对象的个数，其中病例队列设计是一种应用较广的有偏抽样机制。

病例队列设计的设计原理是：首先确定某个人群作为所研究的队列即全队列并收集所有研究对象的原始数据如血液样本，然后在该队列中用随机抽样的方法抽取一个样本即子队列作为对照组（图1.1的 SRS），再收集全队列中所有欲研究疾病的病例作为病例组（图1.1中的 Case），子队列和所有的病例构成了病例队列设计的样本，通过观测样本协变量的值以探索影响病患生存时间的因素。由于病例队列设计样本中的病例组不是随机的，病例队列设计属于有偏的抽样机制，此时传统的分析方法得到的参数估计往往不再相合。Prentice (1986) [1] 针对病例队列设计样本数据的分析提出了伪极大似然估计，并论述了其渐进性质。

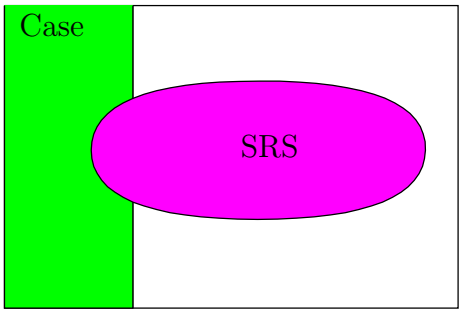

图 1.1　病例队列设计（Case-Cohort Design）

1.2　研究意义

在许多学科领域，如医学、生物学、保险精算学、可靠性工程学、公共卫生学、经济学以及人口统计学等领域，都存在对某给定事件发生的时间进行估计和预测的问题。例如疾病的发生时间，治疗后疾病复发的时间，机械及电子器件

或系统失效时间，经济危机的爆发时间，发行债券的违约时间等。研究事件发生时间的规律问题就是生存分析问题，这类问题的解决通常采用统计学的理论和方法。由于研究时间、经费等条件的限制，很多研究收集到的生存数据发生了删失，即在研究结束时仍未观察到感兴趣事件的发生。病例队列设计也可以应用于流行病学外的其他生存分析问题，尤其对于删失率较高的生存数据。

若不知被观测个体存活时间的精确值，只知其大于或等于 L，则称该个体的观测数据在 L 处右删失。对于右删失情形，我们使用下列记号 :对于某个特定的研究个体，假定其真正的存活时间为 $\tilde{T}$，删失时间为 C_r. 假定所有的 $\tilde{T}$ 相互独立，并且具有相同的分布形式，概率密度函数为 $f(x)$. 当且仅当某个体的生存时间 $\tilde{T}$ 小于等于 C_r 时，我们才能够得知该个体的确切存活时间；如果 $\tilde{T}$ 大于 C_r，那么该个体事件发生时间在 C_r 处删失。我们通常用随机变量组 (T, δ) 表示观测数据，其中 $\delta = I$(观测到个体死亡), $T = \min\left\{\tilde{T}, C_r\right\}$，即当观测数据没有发生删失时，$T = \tilde{T}$，$\delta = 1$；当观测数据发生删失时，$T = C_r, \delta = 0$.

Cox 比例风险模型是生存分析中研究删失数据的经典模型，基于Cox比例风险模型的病例队列设计方法在发病率较低即数据删失率较高的情况下仍然具有较高的估计精度。因此病例队列研究作为一种节约资源、提高效率的统计方法，具有极大的研究意义。

1.3 研究方法

在 Cox 比例风险模型的基础上，我们将研究参数的偏极大似然估计、伪极大似然估计及其大样本性质或渐进分布。有了统计量的渐近分布，当样本量 n 足够大时，就可以用它的近似分布进行涉及分布的推断。通过统计模拟，我们将得到 Cox 比例风险模型下病例队列抽样与简单随机抽样的比较，具体的指标包括相同样本量和删失率下模拟结果得到的估计误差、标准差、均方误差以及置信区间的覆盖率等。

最后，我们将病例队列设计方法应用于美国国家肾母细胞瘤研究数据，找出影响肾母细胞瘤患者存活时间的显著协变量并得到其参数估计、相对风险比及 95% 置信区间。为了进行比较，我们同时模拟从同一总体中抽取相同样本量的简单随机样本，并用 Cox 比例风险模型分析协变量数据得到显著协变量参数的估计。

1.4 研究现状

采用与Prentice（1986）不同的风险集，Self及Prentice（1988）[2]给出了病例队列设计的参数估计及对应的方差估计并证明了相关渐进分布的结果，但其提出的方差估计在计算上十分的复杂，此后一些学者提出了一些简单的估计。Binder（1992）[3]给出了基于 Cox 比例风险模型和抽样方法的一般估计结果。Lin及Ying（1993）[4]讨论了 Cox 比例风险模型协变量不完全测量下的情况并将病例队列设计归为其中的一个特殊情况。Barlow（1994）[5]给出了基于渐进刀切法（Jackknife）的稳健协方差估计。Therneau及Li（1998）[6]介绍了如何用软件中的比例风险回归程序如S-plus的 *coxph* 过程以及SAS的 *phreg* 过程得到方差估计。

此次研究中，我们采用不同于以上方法的另一种方差的简单估计即经验估计，经验估计同时适用于经典的 Cox 比例风险模型及病例队列设计下的 Cox 模型，且两者的差异也仅在于风险集的选择。在经典的 Cox 比例风险模型中，风险集包括所有处于风险中的个体。而在病例队列下的 Cox 模型中，风险集是由所有子队列中处于风险中的个体以及子队列外失效的个体组成。我们采用统计模拟的方法检验经验估计在两个模型下的有效性，同时采用 Efron （1979）[7] 提出的自助抽样法（Bootstrap）计算标准差的估计并进行两种估计方法的比较。

第 2 章　Cox 比例风险模型

1972年 David Cox [8]提出了 Cox 比例风险模型, 简称为 Cox 模型。 Cox 模型是一个半参数模型, 假设个体风险率成比例且具有共同的未知的基线风险率。 Cox 模型可以得到一致有效的协变量效应的估计, 从而成为生存分析中最常用的模型。

2.1　基本概念

生存分析中我们常讨论的函数主要有以下几种：

(1) 生存函数 $S(t)$: 反映个体存活时间超过时间 t 的概率。记个体真正的生存时间为随机变量 $\tilde{T}$, 则有

$$S(t)=P\big(\tilde{T}>t\big), \tag{2.1}$$

$S(t)$ 的图形叫做生存曲线, 陡峭的生存曲线表示较低的生存率或较短的生存时间, 平缓的生存曲线表示较高的生存率或较长的生存时间。 对于删失数据, 生存函数的标准估计式即 Kaplan-Meier 估计又称为乘积限估计, 由 Kaplan 和 Meier (1958) [9]提出, 表达式为

$$\hat{S}(t)=\begin{cases} 1 & t<t_1 \\ \displaystyle\prod_{t_i\leqslant t}\left[1-\frac{d_i}{Y_i}\right] & t\geqslant t_1 \end{cases} \tag{2.2}$$

Greenwood 给出了一个此乘积限估计的方差的估计, 其形式如(2.3)式：

$$\widehat{\mathrm{Var}}\big[\hat{S}(t)\big]=\big[\hat{S}(t)\big]^2\sum_{t_i\leqslant t}\frac{d_i}{Y_i(Y_i-d_i)}, \tag{2.3}$$

(2) 危险率函数 $\lambda(t)$: 个体在时刻 t 瞬间死亡的概率, 描述观察的个体在某时刻存活的条件下, 在下一瞬间死亡的概率。 危险率函数也叫瞬时死亡率、死亡强度、条件死亡率或分年龄死亡率, 在流行病学中称为特定年龄事故率。 危险率函数的定义为

$$\begin{aligned}\lambda(t) &= \lim_{\Delta t\to 0}\frac{P(\text{年龄是}t\text{的个体在}(t,t+\Delta t)\text{中死亡})}{\Delta t} \\ &= \lim_{\Delta t\to 0}\frac{P\big(t\leqslant\tilde{T}<t+\Delta|\tilde{T}\geqslant t\big)}{\Delta t}.\end{aligned} \tag{2.4}$$

(3) 累积危险函数 $\Lambda(t)$: 个体直到时刻 t 的累积危险率, 其表达式如(2.5)式,

$$\Lambda(t)=\int_0^t \lambda(t)\mathrm{d}t. \tag{2.5}$$

累积风险函数与生存函数的关系为：

$$\Lambda(t)=\int_0^t \lambda(u)\mathrm{d}u=\int_0^t \frac{f(u)}{S(u)}\mathrm{d}u=\int_0^t -\frac{[S(u)]'}{S(u)}\mathrm{d}u=-\log[S(t)], \tag{2.6}$$

其中 $f(t)$ 表示个体生存时间所服从的分布密度函数。因此, 对连续的生存时间变量有

$$S(t)=\exp\{-\Lambda(t)\}. \tag{2.7}$$

Cox 模型的基本假设为

$$\lambda(t|Z)=\lambda_0(t)r(\beta'Z(t)), \tag{2.8}$$

其中 $Z(t)$ 为 p 维向量, 表示个体 t 时刻的协变量, $\beta=(\beta_1,\cdots,\beta_p)'$ 是协变量参数, $r(x)$ 为已知的函数且 $r(0)=1$, $\lambda_0(t)$ 为拥有标准协变量即 $Z(t)=0$ 的个体的未知的基准风险率函数。常取 $r(x)=e^x$ 以保证风险率函数 $\lambda(t|Z)$ 大于 0, 代入(2.8)式可得 Cox 模型的常见形式。为简便起见, 下面我们考虑协变量 Z 与时间独立的情形下 Cox 模型, 即

$$\lambda(t|Z)=\lambda_0(t)\exp(\beta'Z). \tag{2.9}$$

假若有两个个体对应的协变量分别为 Z_1, Z_2, 则在此模型下, 这两个个体的风险率之比

$$\frac{\lambda(t|Z_1)}{\lambda(t|Z_2)}=\exp\{\beta'(Z_1-Z_2)\} \tag{2.10}$$

是一个常值, 即风险率是成比例的, 因此 Cox 模型常被称作比例风险模型。(2.10)式的值被称为具有风险因素 Z_1 的个体对风险因素为 Z_2 的个体的相对风险率。特别地, 如果 Z 表示有治疗效应, 即 $Z_1=1$ 表示个体接受了治疗, $Z_2=0$ 表示个体服用安慰剂, 则 $\lambda(t|Z_1)\,/\,\lambda(t|Z_2)=e^{\beta}$ 表示接受治疗的个体相对于服用安慰剂个体所具有的相对风险。

比较(2.9)式与指数分布及威布尔分布的分布密度函数, 我们可以发现 $\lambda_0(t)=1$ 时, $\tilde{T}\sim\mathrm{Exp}(e^{\beta'Z})$;$\lambda_0(t)=2t$ 时, $\tilde{T}\sim\mathrm{Weibull}\left(2,e^{\frac{1}{2}\beta'Z}\right)$。

2.2 偏极大似然估计

将 n 个没有结点的生存数据样本按观测时间从小到大排序得 $(T_i,\,\delta_i,\,Z_i)$, $i=1,\cdots,n$, $T_1<T_2<\cdots<T_n$. 用 $R(t)=\{i:T_i\geqslant t\}$ 表示时刻 t 的风险集, 即 $R(t)$ 中的元

素为在时刻 t 仍可观测到的存活的个体, 记 $R_i = R(T_i)$ 表示在第 i 个观察时间 T_i 处的风险集。 用 $Z_i(t)$ 表示个体 i 在时刻 t 的协变量, Z_i 表示与时间独立的个体 i 的协变量。

为了得到 β 的估计, Cox (1972) 考虑在风险集 R_i 中的个体 i 在时刻 T_i 处死亡的条件概率。事实上, 若个体 i 的生存时间数据未删失, 则

$$\begin{aligned} & P(\text{个体 } i \text{ 在时刻 } T_i \text{ 处死亡} | R_i \text{ 中某个体在 } T_i \text{ 处死亡}) \\ = & \frac{\lambda(T_i|Z_i)}{\sum_{j\in R_i}\lambda(T_i|Z_j)} \\ = & \frac{\lambda_0(T_i)\exp(\beta' Z_i)}{\sum_{j\in R_i}\lambda_0(T_i)\exp(\beta' Z_j)} \\ = & \frac{\exp(\beta' Z_i)}{\sum_{j\in R_i}\exp(\beta' Z_j)}. \end{aligned} \tag{2.11}$$

若个体 i 的生存时间数据删失, 则对条件似然函数无贡献。 因此将所有这些死亡时间的条件概率相乘便得到偏似然函数 $L(\beta)$ 及对数偏似然函数 $l(\beta)$, 其表达式如下:

$$L(\beta) = \prod_{i=1}^{n}\left\{\frac{\exp(\beta' Z_i)}{\sum_{j\in R_i}\exp(\beta' Z_j)}\right\}^{\delta_i}, \tag{2.12}$$

$$l(\beta) = \sum_{i=1}^{n}\delta_i\left\{\beta' Z_i - \log\left[\sum_{j\in R_i}\exp(\beta' Z_j)\right]\right\}, \tag{2.13}$$

协变量与时间相关时将上述两式中的 Z_i, Z_j 替换为 $Z_i(t), Z_j(t)$ 即可。 Cox 证明了在一定条件下 $\hat{\beta}$ 是 β 的相合估计且是渐进正态的, 即

定理 2.1 ($\hat{\beta}$ 的渐近正态性). $\sqrt{n}\big(\hat{\beta} - \beta_0\big) \xrightarrow{d} N(0, \Sigma^{-1})$.

2.3 标准差的经验估计

对于协变量时间独立情形下的 Cox 比例风险模型, 协变量参数 β 的偏似然函数及其对数形式如(2.12)及(2.13)式, 对应的梯度函数和 Hessian 矩阵函数为

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n}\delta_i\left(Z_i - \frac{S_{1i,\beta}}{S_{0i,\beta}}\right), \tag{2.14}$$

$$H(\beta) = -\frac{\partial^2 l(\beta)}{\partial\beta\,\partial\beta'} = \sum_{i=1}^{n}\delta_i\left(\frac{S_{2i,\beta}}{S_{0i,\beta}} - \left(\frac{S_{1i,\beta}}{S_{0i,\beta}}\right)^{\otimes 2}\right). \tag{2.15}$$

其中

$$S_{0i,\beta} = \sum_{l\in R_i} e^{\beta' Z_l}, \tag{2.16}$$

$$S_{1i,\beta} = \sum_{l\in R_i} Z_l\, e^{\beta' Z_l}, \tag{2.17}$$

$$S_{2i,\beta} = \sum_{l\in R_i} Z_l^{\otimes 2}\, e^{\beta' Z_l}. \tag{2.18}$$

由泰勒展开式, 我们可以近似得到

$$\frac{\partial l\big(\hat{\beta}\big)}{\partial \beta} - \frac{\partial l(\beta_0)}{\partial \beta} \approx \frac{\partial^2 l(\beta_0)}{\partial \beta\, \partial \beta'}\big(\hat{\beta} - \beta_0\big), \tag{2.19}$$

其中 β_0 表示参数真值, 对极大似然估计 $\hat{\beta}$ 有 $\frac{\partial l(\hat{\beta})}{\partial \beta} = 0$, 从而

$$\sqrt{n}\big(\hat{\beta} - \beta_0\big) \approx \left(\frac{1}{n}\left(-\frac{\partial^2 l(\beta_0)}{\partial \beta\, \partial \beta'}\right)\right)^{-1}\left(\frac{1}{\sqrt{n}}\frac{\partial l(\beta_0)}{\partial \beta}\right). \tag{2.20}$$

由中心极限定理及强大数定律可得

$$\frac{1}{\sqrt{n}}\frac{\partial l(\beta_0)}{\partial \beta} = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\delta_i\left(Z_i - \frac{S_{1i,\beta_0}}{S_{0i,\beta_0}}\right) - 0\right] \xrightarrow{d} N(0, \Lambda(\beta_0)), \tag{2.21}$$

$$\frac{1}{n}\left(-\frac{\partial^2 l(\beta_0)}{\partial \beta\, \partial \beta'}\right) = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left(\frac{S_{2i,\beta_0}}{S_{0i,\beta_0}} - \left(\frac{S_{1i,\beta_0}}{S_{0i,\beta_0}}\right)^{\otimes 2}\right) \xrightarrow{p} I(\beta_0) \tag{2.22}$$

其中 $\Lambda(\beta_0)$ 及 $I(\beta_0)$ 是与 β_0 有关的常数, 则根据 Slutsky 定理应有

$$\sqrt{n}\big(\hat{\beta} - \beta_0\big) \xrightarrow{d} N(0, \Sigma(\beta_0)), \tag{2.23}$$

其中 $\Sigma(\beta_0) = I^{-1}(\beta_0)\, \Lambda(\beta_0)\, I^{-1}(\beta_0)$.

记

$$S_{i,\hat{\beta}} = \delta_i\left(Z_i - \frac{S_{1i,\hat{\beta}}}{S_{0i,\hat{\beta}}}\right), T_{i,\hat{\beta}} = \delta_i\left(\frac{S_{2i,\hat{\beta}}}{S_{0i,\hat{\beta}}} - \left(\frac{S_{1i,\hat{\beta}}}{S_{0i,\hat{\beta}}}\right)^{\otimes 2}\right),$$

则 $\Lambda(\beta_0)$ 的无偏估计为

$$\hat{\Lambda}(\beta_0) = \frac{1}{n-1}\sum_{i=1}^{n}\left(S_{i,\hat{\beta}} - \frac{1}{n}\sum_{i=1}^{n} S_{i,\hat{\beta}}\right)^2, \tag{2.24}$$

$I(\beta_0)$ 的无偏估计为

$$\hat{I}(\beta_0) = \frac{1}{n}\sum_{i=1}^{n} T_{i,\hat{\beta}} = \frac{1}{n} H\big(\hat{\beta}\big). \tag{2.25}$$

从而有

$$\hat{\Sigma}(\beta_0) = \big(\hat{I}(\beta_0)\big)^{-1}\, \hat{\Lambda}(\beta_0)\, \big(\hat{I}(\beta_0)\big)^{-1}. \tag{2.26}$$

因此由(2.23), 偏极大似然估计 $\hat{\beta}$ 标准差的经验估计 $\mathrm{SE}\big(\hat{\beta}\big)$ 为

$$\mathrm{SE}\big(\hat{\beta}\big) = \sqrt{\frac{\hat{\Sigma}(\beta_0)}{n}}. \tag{2.27}$$

由于上面的推导未涉及 $t \to \infty$ 的变化, 故也适用于协变量与时间相依的情形。

第 3 章　病例队列研究

3.1　基本概念

考虑大小为 n 的全队列 C 以及大小为 $\tilde{n}$ 的子队列 $\tilde{C}$, 则病例队列抽样得到的样本由子队列及全队列中所有的病例组成。 设有大小为 m 的病例队列设计样本, 将其按观测时间从小到大排列后得到 $(T_i, \delta_i, Z_i, \xi_i)$, $i = 1, \cdots, m$, 其中示性变量 $\xi_i = I\big(i \in \tilde{C}\big)$, δ 为删失示性变量, T 为观测时间, Z 为个体协变量数据。

考虑协变量时间独立情形下的 Cox 比例风险模型:

$$\lambda(t) = \lambda_0(t) \exp(\beta' Z), \tag{3.1}$$

其中 $\lambda_0(t), \beta, Z$ 如(2.9)式所定义。

记 $N_i(t) = I$(是否在时刻 t 已观测到个体 i 失效) , 则依据定义删失变量 $\delta_i = I(N_i(T_i) \neq N_i(T_i^-))$。 根据 Self 及 Prentice [2], 时刻 t 处的风险集定义为 $\tilde{R}(t) = D(t) \cup R_{\mathrm{SRS}}(t)$, 其中 $D(t) = \{i | N_i(t) \neq N_i(t^-)\}$ 表示全队列中时刻 t 失效或死亡的个体集合, $R_{\mathrm{SRS}}(t) = \big\{i \in \tilde{C} | T_i \geqslant t\big\}$ 表示时刻 t 子队列存在风险的个体集合。

由图1.1知, 病例队列设计样本可表示为 $S = S_1 \cup S_2 \cup S_3$, 其中 $S_1 = \mathrm{Case} - \tilde{C}$, $S_2 = \tilde{C} - \mathrm{Case}, S_3 = \mathrm{Case} \cap \tilde{C}$ 互不相交。 记 $\tilde{R}_i = \tilde{R}(t_i)$, 对于 S 中的个体 i 有如下结论成立:

$i \in S_1$ 时, $\tilde{R}_i = R_{\mathrm{SRS}}(T_i) \cup \{i\}, \delta_i = 1, \xi_i = 0$, 即有 $\delta_i (1 - \xi_i) = 1$;

$i \in S_2$ 时, $\tilde{R}_i = R_{\mathrm{SRS}}(T_i), \delta_i = 0, \xi_i = 1$, 即有 $(1 - \delta_i)\xi_i = 1$;

$i \in S_3$ 时, $\tilde{R}_i = R_{\mathrm{SRS}}(T_i), \delta_i = 1, \xi_i = 1$, 即有 $\xi_i \delta_i = 1$.

3.2　伪极大似然估计

对于病例队列抽样得到的样本数据, 为了得到协变量参数的估计, 借鉴 Cox 比例风险模型的思想, Prentice (1986) [1] 考虑极大化函数[10]

$$\tilde{L}(\beta) = \prod_{i=1}^{m} \left(\frac{\beta' Z_i}{\sum_{l \in \tilde{R}_i} \exp(\beta' Z_l)} \right)^{\delta_i}, \tag{3.2}$$

或(3.2)式的对数形式

$$\tilde{l}(\beta) = \sum_{i=1}^{m} \delta_i \left(\beta' Z_i - \log\left(\sum_{l \in \tilde{R}_i} \exp(\beta' Z_l) \right) \right). \tag{3.3}$$

(3.2)式被称为伪似然函数, 对应的一阶导函数 $\tilde{U}(\beta)$ 及 Heissian 矩阵 $\tilde{H}(\beta)$ 为

$$\tilde{U}(\beta) = \sum_{i=1}^{m}\left(Z_i - \frac{\sum_{l\in\tilde{R}_i} Z_l \exp(\beta' Z_l)}{\sum_{l\in\tilde{R}_i}\exp(\beta' Z_l)}\right), \tag{3.4}$$

$$\tilde{H}(\beta) = \sum_{i=1}^{m}\left[\frac{\sum_{l\in\tilde{R}_i} Z_l^{\otimes 2}\exp(\beta' Z_l)}{\sum_{l\in\tilde{R}_i}\exp(\beta' Z_l)} - \left(\frac{\sum_{l\in\tilde{R}_i} Z_l\exp(\beta' Z_l)}{\sum_{l\in\tilde{R}_i}\exp(\beta' Z_l)}\right)^{\otimes 2}\right]. \tag{3.5}$$

协变量与时间相关的情形下将上述两式中的 Z_i, Z_j 分别替换为 $Z_i(t), Z_j(t)$ 即可。此外, 由3.1节 $\tilde{R}_i$ 与 $R_{\mathrm{SRS}}(t_i)$ 的关系可得

$$\begin{aligned}
\sum_{l\in\tilde{R}_i}\exp(\beta' Z_l) &= \sum_{l\in R_{\mathrm{SRS}}(t_i)} \delta_l\,(1-\xi_l)\exp(\beta' Z_l) + \delta_i\,(1-\xi_i)\exp(\beta' Z_i) \\
&\quad + \sum_{l\in R_{\mathrm{SRS}}(t_i)} (1-\delta_l)\,\xi_l\exp(\beta' Z_l) + \sum_{l\in R_{\mathrm{SRS}}(t_i)} \xi_l\,\delta_l\exp(\beta' Z_l) \\
&= \sum_{l\in R_{\mathrm{SRS}}(t_i)} \left[\xi_l\,\delta_l + (1-\delta_l)\,\xi_l + \delta_l\,(1-\xi_l)\right]\exp(\beta' Z_l) + \delta_i\,(1-\xi_i)\exp(\beta' Z_i) \\
&= \sum_{l\in R_{\mathrm{SRS}}(t_i)} \exp(\beta' Z_l) + \delta_i\,(1-\xi_i)\exp(\beta' Z_i) \\
&= \sum_{l\in R_i} \xi_l\exp(\beta' Z_l) + \delta_i\,(1-\xi_i)\exp(\beta' Z_i).
\end{aligned} \tag{3.6}$$

其中 R_i 如 Cox 比例风险模型的偏似然函数即(2.12)式中所定义, 对任意 $l \in R_{\mathrm{SRS}}(t_i)$, $\xi_l = 1$, 故 $\xi_l\,\delta_l + (1-\delta_l)\,\xi_l + \delta_l\,(1-\xi_l) = 1$.

将(3.6)式代入(3.2)式可得伪似然函数也可写作

$$\begin{aligned}
\tilde{L}(\beta) &= \prod_{i=1}^{m}\left(\frac{\exp(\beta' Z_i)}{\sum_{l\in R_i}\xi_l\exp(\beta' Z_l) + \delta_i\,(1-\xi_i)\exp(\beta' Z_i)}\right)^{\delta_i} \\
&= \prod_{i=1}^{n}\left(\frac{\exp(\beta' Z_i)}{\sum_{l\in R_i}\xi_l\exp(\beta' Z_l) + \delta_i\,(1-\xi_i)\exp(\beta' Z_i)}\right)^{\delta_i},
\end{aligned} \tag{3.7}$$

比较(3.7)式与(2.12)式可以发现病例队列设计下的 Cox 比例风险模型可看做是加权的 Cox 比例风险模型。

设 $\tilde{\beta}$ 为 β_0 的伪极大似然估计, Self 和 Prentice [2] 证明了在一定条件下 $\tilde{\beta}$ 有一系列渐近性质。

定理3.1 ($\tilde{\beta}$ 的一致性). $\tilde{\beta} \xrightarrow{p} \beta_0$.

定理3.2 ($\tilde{\beta}$ 的渐近正态性).

$$\sqrt{n}\big(\tilde{\beta} - \beta_0\big) \xrightarrow{d} N(0, \Sigma^{-1} + \Sigma^{-1}\Delta\Sigma^{-1}).$$

其中 Σ 以及 Δ 的估计式十分的复杂。

3.3 标准差的经验估计

与2.3节类似地记

$$\tilde{S}_{0i,\beta} = \sum_{l\in\tilde{R}_i} e^{\beta' Z_l} = \sum_{l\in R_i} \xi_l\, e^{\beta' Z_l} + \delta_i\,(1-\xi_i) e^{\beta' Z_i}, \tag{3.8}$$

$$\tilde{S}_{1i,\beta} = \sum_{l\in\tilde{R}_i} Z_l\, e^{\beta' Z_l} = \sum_{l\in R_i} \xi_l\, Z_l\, e^{\beta' Z_l} + \delta_i\,(1-\xi_i) Z_i\, e^{\beta' Z_i}, \tag{3.9}$$

$$\tilde{S}_{2i,\beta} = \sum_{l\in\tilde{R}_i} Z_l^{\otimes 2}\, e^{\beta' Z_l} = \sum_{l\in R_i} \xi_l\, Z_l^{\otimes 2}\, e^{\beta' Z_l} + \delta_i\,(1-\xi_i) Z_i^{\otimes 2}\, e^{\beta' Z_i}. \tag{3.10}$$

从而

$$\tilde{U}(\beta) = \sum_{i=1}^{m} \delta_i \left(Z_i - \frac{\tilde{S}_{1i,\beta}}{\tilde{S}_{0i,\beta}} \right), \tag{3.11}$$

$$\tilde{H}(\beta) = \sum_{i=1}^{m} \delta_i \left(\frac{\tilde{S}_{2i,\beta}}{\tilde{S}_{0i,\beta}} - \left(\frac{\tilde{S}_{1i,\beta}}{\tilde{S}_{0i,\beta}} \right)^{\otimes 2} \right). \tag{3.12}$$

伪极大似然估计 $\tilde{\beta}$ 的标准差的经验估计 $\mathrm{SE}\big(\tilde{\beta}\big)$ 为

$$\mathrm{SE}\big(\tilde{\beta}\big) = \sqrt{\frac{\tilde{\Sigma}(\beta_0)}{m}}. \tag{3.13}$$

其中 β_0 表示 β 的真值, m 表示病例队列设计的样本量,

$$\tilde{\Sigma}(\beta_0) = \big(\tilde{I}(\beta_0)\big)^{-1}\, \tilde{\Lambda}(\beta_0)\, \big(\tilde{I}(\beta_0)\big)^{-1}.$$

记

$$\tilde{S}_{i,\tilde{\beta}} = \delta_i \left(Z_i - \frac{\tilde{S}_{1i,\tilde{\beta}}}{\tilde{S}_{0i,\tilde{\beta}}} \right), \tilde{T}_{i,\tilde{\beta}} = \delta_i \left(\frac{\tilde{S}_{2i,\tilde{\beta}}}{\tilde{S}_{0i,\tilde{\beta}}} - \left(\frac{\tilde{S}_{1i,\tilde{\beta}}}{\tilde{S}_{0i,\tilde{\beta}}} \right)^{\otimes 2} \right),$$

类似地有

$$\tilde{\Lambda}(\beta_0) = \frac{1}{m-1} \sum_{i=1}^{m} \left(\tilde{S}_{i,\tilde{\beta}} - \frac{1}{m} \sum_{i=1}^{m} \tilde{S}_{i,\tilde{\beta}} \right)^2, \tag{3.14}$$

$$\tilde{I}(\beta_0) = \frac{1}{m} \sum_{i=1}^{m} \tilde{T}_{i,\tilde{\beta}} = \frac{1}{m}\, \tilde{H}\big(\tilde{\beta}\big). \tag{3.15}$$

(3.13)式的标准差经验估计也适用于协变量与时间相依情形。

第 4 章　统计模拟

4.1　模拟实验设计

模拟符合如下条件的生存数据集: 抽样方法分别为简单随机抽样和病例队列设计抽样, 样本量 n 分别取 200 和 300 , 参数真值分别取 $\boldsymbol{\beta_0} = (\ln 2, -0.5)'$, $(\ln 2, 0)'$ 以及 $(0, -0.5)'$, 删失率分别取 0.5, 0.7 和 0.9 , 生存时间服从参数为 $e^{\boldsymbol{\beta_0' Z}}$ 的指数分布, 协变量 $Z = (Z_1, Z_2)'$, 其中 Z_1 来自参数为 0.5 的伯努利分布总体, Z_2 来自标准正态分布总体。从而对于每一种抽样方法, 我们各模拟了 18 组数据集。

每产生一组生存数据集 $(T_i, \delta_i, \boldsymbol{Z_i}), i = 1, \cdots, n$, 我们可以用 Newton-Raphson 迭代法求解方程 $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{0}$ 得到参数估计 $\hat{\boldsymbol{\beta}}$ 。其中, 对于简单随机样本我们应用经典 Cox 比例风险模型, 阶梯函数 $\boldsymbol{U}(\boldsymbol{\beta})$ 如(2.14)式所示 ;对于病例队列设计样本我们应用加权比例风险模型, $\boldsymbol{U}(\boldsymbol{\beta})$ 如(3.11)式所示。计算 S 次模拟得到的参数估计值 $\hat{\boldsymbol{\beta}}_{\boldsymbol{1}}, \hat{\boldsymbol{\beta}}_2, \cdots, \hat{\boldsymbol{\beta}}_{\boldsymbol{S}}$ 并求出其平均值 $\hat{\boldsymbol{\beta}}$ 和标准差 SD 。分别采用标准差的经验估计以及 Bootstrap 法的标准差估计, 分别计算两个模型第 s 次模拟得到的 $\hat{\boldsymbol{\beta}}_s$ 的标准差估计 $\mathrm{SE}_s\,(s = 1, \cdots, S)$ 。取置信水平 $\alpha = 0.05$ 得到渐近正态性质下的置信区间 $I_s = \left[\hat{\boldsymbol{\beta}} - 1.96\,\mathrm{SE}_s, \hat{\boldsymbol{\beta}} + 1.96\,\mathrm{SE}_s\right]$, 最后计算 $\mathrm{SE}_1, \cdots, \mathrm{SE}_S$ 的均值作为 $\hat{\boldsymbol{\beta}}$ 标准差的估计 SE , 区间覆盖率 $\mathrm{CP} = \sum_{i=1}^{S} I(\boldsymbol{\beta_0} \in I_s)/S$, 其中 $\boldsymbol{\beta_0}$ 表示参数真值。

4.2　模拟结果

对于标准差的经验估计和 Bootstrap 估计, 分别将设定的数据样本重复模拟 1000 次得到对应的估计偏差 Bias 、标准差 SD 、标准差估计 SE 以及区间覆盖率 CP , 模拟结果见表4.1及4.2, 部分模拟的源程序见附录A。

表 4.1　经验估计模拟结果

样本量	$\boldsymbol{\beta_0}$	删失率	病例队列设计 Bias	SD	SE	CP	Cox模型 Bias	SD	SE	CP
200	(ln2, -0.5)	0.5	0.013	0.286	0.184	0.798	0.010	0.204	0.219	0.967
			-0.016	0.159	0.099	0.772	-0.017	0.112	0.116	0.960
		0.7	-0.007	0.275	0.230	0.901	0.013	0.274	0.282	0.961
			-0.015	0.137	0.119	0.921	-0.002	0.138	0.143	0.962
		0.9	0.015	0.439	0.413	0.939	0.041	0.513	0.511	0.971
			-0.018	0.220	0.197	0.922	-0.008	0.238	0.236	0.935
	(ln2, 0)	0.5	0.003	0.282	0.181	0.801	0.011	0.208	0.218	0.963
			-0.006	0.143	0.090	0.791	0.001	0.108	0.104	0.943
		0.7	0.010	0.256	0.230	0.925	0.006	0.266	0.279	0.959
			0.001	0.132	0.112	0.909	0.003	0.135	0.134	0.945
		0.9	0.023	0.471	0.458	0.962	0.045	0.581	0.574	0.966
			-0.002	0.224	0.207	0.937	-0.020	0.269	0.252	0.932
	(0, -0.5)	0.5	-0.006	0.286	0.175	0.787	0.001	0.208	0.208	0.957
			-0.012	0.156	0.099	0.802	-0.008	0.108	0.115	0.967
		0.7	-0.006	0.262	0.221	0.911	-0.005	0.261	0.269	0.963
			-0.013	0.143	0.119	0.901	-0.001	0.134	0.141	0.962
		0.9	-0.019	0.418	0.387	0.948	0.006	0.461	0.480	0.971
			-0.021	0.210	0.195	0.938	-0.016	0.246	0.236	0.936
300	(ln2, -0.5)	0.5	-0.006	0.203	0.153	0.854	0.004	0.170	0.176	0.963
			-0.011	0.108	0.082	0.858	-0.005	0.089	0.093	0.961
		0.7	0.012	0.215	0.197	0.928	0.007	0.219	0.227	0.963
			-0.013	0.112	0.100	0.922	-0.004	0.111	0.115	0.958
		0.9	0.018	0.356	0.349	0.949	0.033	0.417	0.407	0.955
			0.001	0.169	0.168	0.947	-0.005	0.192	0.192	0.940
	(ln2, 0)	0.5	0.011	0.199	0.153	0.877	0.012	0.171	0.176	0.954
			0.001	0.101	0.074	0.860	0.004	0.084	0.084	0.956
		0.7	0.010	0.207	0.196	0.941	0.026	0.223	0.227	0.948
			0.002	0.108	0.094	0.916	-0.003	0.107	0.108	0.944
		0.9	0.001	0.403	0.389	0.955	0.022	0.477	0.453	0.942
			0.005	0.192	0.179	0.917	-0.002	0.202	0.205	0.947
	(0, -0.5)	0.5	0.007	0.202	0.147	0.864	0.001	0.163	0.168	0.955
			-0.008	0.106	0.082	0.867	-0.003	0.088	0.093	0.968
		0.7	0.003	0.208	0.188	0.935	-0.001	0.211	0.216	0.960
			-0.003	0.110	0.099	0.917	-0.004	0.110	0.114	0.959
		0.9	0.009	0.348	0.330	0.942	0.014	0.389	0.382	0.959
			-0.008	0.174	0.166	0.937	-0.003	0.194	0.192	0.950

表 4.2 Bootstrap模拟结果

样本量	β_0	删失率	病例队列设计 Bias	SD	SE	CP	Cox模型 Bias	SD	SE	CP
200	(ln2, -0.5)	0.5	0.015	0.294	0.298	0.955	0.011	0.212	0.214	0.952
			-0.015	0.163	0.164	0.956	-0.003	0.113	0.111	0.952
		0.7	0.001	0.267	0.278	0.957	0.019	0.281	0.285	0.951
			-0.015	0.144	0.148	0.960	-0.009	0.137	0.143	0.960
		0.9	0.047	0.442	0.457	0.973	0.050	0.499	0.545	0.970
			-0.013	0.213	0.219	0.961	-0.013	0.231	0.253	0.962
	(ln2, 0)	0.5	0.014	0.281	0.289	0.955	0.012	0.214	0.216	0.951
			-0.003	0.144	0.152	0.966	0.004	0.105	0.106	0.946
		0.7	0.028	0.262	0.271	0.962	0.012	0.272	0.285	0.966
			-0.001	0.137	0.137	0.960	0.002	0.134	0.138	0.957
		0.9	0.061	0.475	0.507	0.973	0.039	0.585	0.600	0.952
			-0.004	0.218	0.228	0.961	-0.001	0.266	0.271	0.957
	(0, -0.5)	0.5	0.009	0.275	0.290	0.957	0.010	0.205	0.209	0.955
			-0.023	0.154	0.160	0.962	-0.010	0.104	0.113	0.970
		0.7	0.007	0.269	0.270	0.953	-0.006	0.258	0.274	0.964
			-0.004	0.143	0.147	0.947	-0.006	0.144	0.143	0.944
		0.9	0.015	0.389	0.428	0.973	-0.050	0.506	0.521	0.969
			-0.009	0.215	0.218	0.956	-0.017	0.251	0.253	0.948
300	(ln2, -0.5)	0.5	0.007	0.205	0.207	0.942	0.011	0.172	0.173	0.949
			-0.011	0.108	0.110	0.957	-0.003	0.084	0.089	0.963
		0.7	0.007	0.213	0.221	0.967	0.005	0.228	0.228	0.945
			-0.010	0.116	0.115	0.949	-0.002	0.113	0.114	0.963
		0.9	0.008	0.355	0.377	0.965	0.018	0.416	0.436	0.973
			-0.012	0.174	0.180	0.959	-0.011	0.199	0.201	0.955
	(ln2, 0)	0.5	-0.001	0.200	0.201	0.943	0.005	0.177	0.173	0.947
			0.003	0.104	0.102	0.941	0.002	0.086	0.085	0.947
		0.7	0.004	0.213	0.218	0.959	0.029	0.223	0.229	0.961
			0.002	0.106	0.108	0.964	0.001	0.105	0.110	0.966
		0.9	0.014	0.383	0.421	0.972	0.047	0.465	0.490	0.962
			0.001	0.199	0.189	0.944	0.007	0.214	0.214	0.948
	(0, -0.5)	0.5	0.002	0.195	0.201	0.960	0.005	0.166	0.168	0.940
			-0.014	0.104	0.109	0.963	-0.008	0.089	0.090	0.949
		0.7	-0.004	0.214	0.215	0.961	0.005	0.210	0.219	0.966
			-0.007	0.109	0.114	0.957	0.000	0.113	0.115	0.935
		0.9	-0.004	0.336	0.351	0.972	-0.002	0.368	0.403	0.976
			0.000	0.172	0.179	0.959	-0.007	0.189	0.197	0.960

第 5 章　实例分析

5.1　实例资料

本章实例分析的数据源自美国国家肾母细胞瘤研究组针对儿童的第三次以及第四次临床试验。在两次研究中，研究者共观察了 4028 名肾母细胞瘤患者，到研究结束观察到的病例共 571 例，发病率为 0.14，从而数据删失率为 0.86 。 Breslow 和 Chatterjee (1999) [10]在数据分析中选取了一个样本量为 668 的简单随机样本作为子队列。研究人员记录了每个患者被观测到的疾病复发时间 T 和删失指示变量 δ ，$\delta = 1$ 表示观测到了患者疾病复发，$\delta = 0$ 表示没有观测到患者疾病复发即观测数据发生了删失。该肿瘤细胞的组织学类型包括预后良好型（FH, favorable histology）和预后不良型（UH, unfavorable histology），预后良好型是典型的肾母细胞，通常有较好的治疗效果；预后不良型包括未分化型、透明细胞肉瘤型及横纹肌样肉瘤型。肾母细胞瘤分为四个阶段，第一阶段肿瘤限于肾内，能完整切除；第二阶段肿瘤超出肾脏，但能完整切除；第三阶段肿瘤细胞转移到肺腔内、淋巴结内非血源性残留；第四阶段肿瘤细胞转移到肺部或肝脏。研究人员收集了患者的肿瘤细胞组织学类型、肿瘤阶段及患者年龄三个协变量数据，部分复发时间数据如图5.1（Δ表示删失观测），原始数据见附录B。

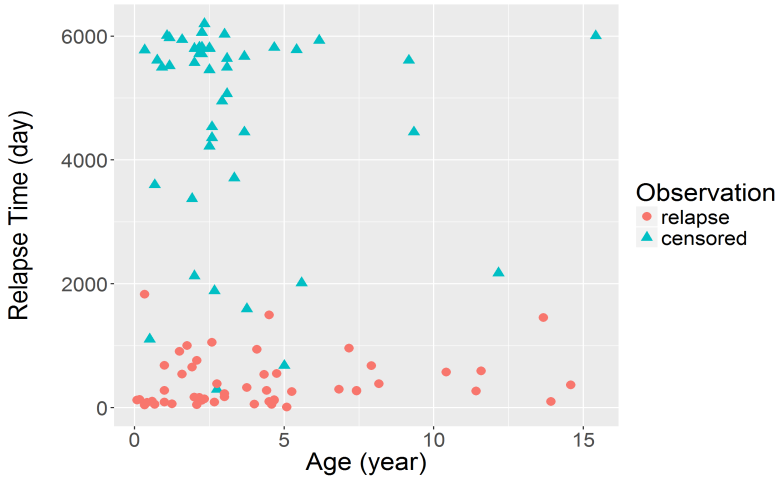

图 5.1　肾母细胞瘤患者的复发时间分布

5.2　实例分析结果

基于 4028 名肾母细胞瘤患者复发时间数据，根据公式(2.2)，计算得到不同肿瘤细胞组织学类型和肿瘤阶段对应的生存函数的 K-M 估计，见图5.2及图5.3。从图5.2可以看出，预后良好组织学（FH）肾母细胞瘤患者的生存概率明显高于

预后不良组织学（UH）肾母细胞瘤患者。从图5.3可以看出，随着肿瘤阶段的深入，肾母细胞患者的生存概率明显降低，只有位于肿瘤第一阶段的患者生存概率高于平均生存概率。

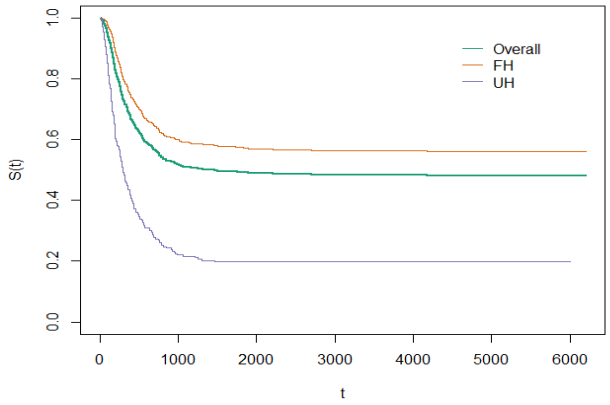

图 5.2　不同组织学类型下生存函数的 K-M 估计

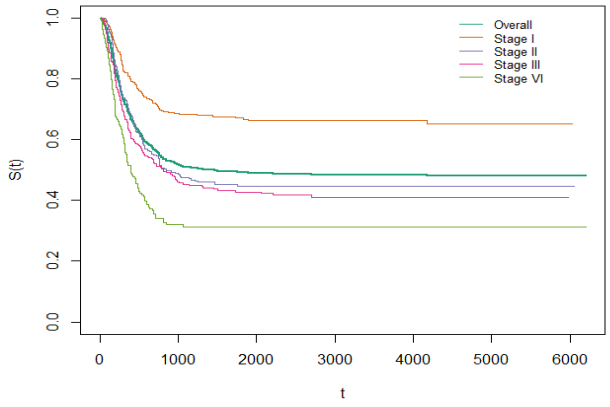

图 5.3　不同肿瘤阶段下生存函数的 K-M 估计

为分析各个协变量对生存概率的影响，首先将收集的协变量数据进行数学表示。令 $Z_1 = 1$ 表示预后良好组织学类型，$Z_1 = 2$ 表示预后不良组织学类型，$Z_2 = i$ $(i = 1, 2, 3, 4)$ 表示患者处于肿瘤第 i 阶段，Z_3 表示患者的年龄。从而，该例的生存数据可表示为 $(T_i, \delta_i, \boldsymbol{Z}_i)$, $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3})'$, $i = 1, \cdots, 4028$，其中 Z_1 与时间独立，Z_2, Z_3 是与时间相关的协变量。下面用比例风险模型分析该组生存数据，即拟合模型为

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t)\, e^{\boldsymbol{\beta}'\boldsymbol{Z}} = \lambda_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3), \tag{5.1}$$

收集病例队列设计样本，即所有的子队列样本及病例，样本量为 1154 （部分子队列中的样本也属于病例），估计结果见表5.1。 为进行比较，同时抽取样本量为1154的简单随机样本并用 Cox 比例风险模型拟合，估计结果见表5.2。

表 5.1　病例队列设计下的估计结果

			经验估计		Bootstrap	
变量	参数	估计值	SE	p value	SE	p value
Histology	β_1	1.429	0.090	$<0.0001^{****}$	0.153	$<0.0001^{****}$
Stage	β_2	0.378	0.042	$<0.0001^{****}$	0.065	$<0.0001^{****}$
Age	β_3	0.041	0.015	$<0.01^{**}$	0.026	0.124

表 5.2　Cox 模型下的估计结果

			经验估计		Bootstrap	
变量	参数	估计值	SE	p value	SE	p value
Histology	β_1	1.855	0.157	$<0.0001^{****}$	0.135	$<0.0001^{****}$
Stage	β_2	0.386	0.069	$<0.0001^{****}$	0.070	$<0.0001^{****}$
Age	β_3	0.025	0.033	0.452	0.032	0.436

若采用标准差的 Bootstrap 估计，则两个模型中协变量参数 β_3 均不显著，用向后回归法剔除影响不显著的变量后两个模型的估计结果见表5.3，表中的 SE 为标准差的 Bootstrap 估计。

表 5.3　剔除 β_3 后的估计结果

		病例队列设计			Cox 模型		
变量	参数	估计值	SE	p value	Estimate	SE	p value
Histology	β_1	1.438	0.150	$<0.0001^{****}$	1.844	0.143	$<0.0001^{****}$
Stage	β_2	0.401	0.062	$<0.0001^{****}$	0.397	0.068	$<0.0001^{****}$

基于以上模型估计结果得到的相对风险比的估计及其95%置信区间为：

表 5.4　相对风险比的估计结果

	病例队列设计		Cox 模型	
	相对风险比	置信区间	相对风险比	置信区间
UH	4.211	[3.136, 5.655]	6.320	[4.624, 8.636]
Stage II	1.493	[1.322, 1.686]	1.487	[1.302, 1.698]
Stage III	2.228	[1.747, 2.842]	2.210	[1.694, 2.883]
Stage VI	3.326	[2.310, 4.790]	3.286	[2.205, 4.895]

从表5.4可以看出， 基于病例队列设计数据的估计结果， 其他条件相同的情况下， 肿瘤细胞组织学为预后不良型的患者复发风险是预后良好型患者的 4.211 倍；第二、三、四阶段的肾母细胞瘤患者的复发风险分别是第一阶段患者的1.493倍、2.228倍和3.326倍。

第 6 章　结论

根据第 4 章统计模拟的主要结果即表4.1及表4.2, 当真实的生存时间服从满足比例风险假定的指数分布时, 病例队列设计及简单随机抽样下的比例风险模型得到的相关统计推断都有较好的性质。 即使是在参数真值 $\boldsymbol{\beta_0}$ 取值不连续以及删失率较高时, 模型得到的参数的极大偏似然估计依然具有较好的优良性。 从图6.1、6.2及6.3可以看出, 估计的优良性体现在以下几个方面：

（1）估计是无偏的, 估计值 $\hat{\boldsymbol{\beta}}$ 与真值 $\boldsymbol{\beta_0}$ 的绝对误差接近 0，绝对误差大部分在 0.05 以内；

（2）估计的标准差 SD 很小, 大部分在 0.5 以内；

（3）区间覆盖率达到置信度, 大部分在 95% 左右。

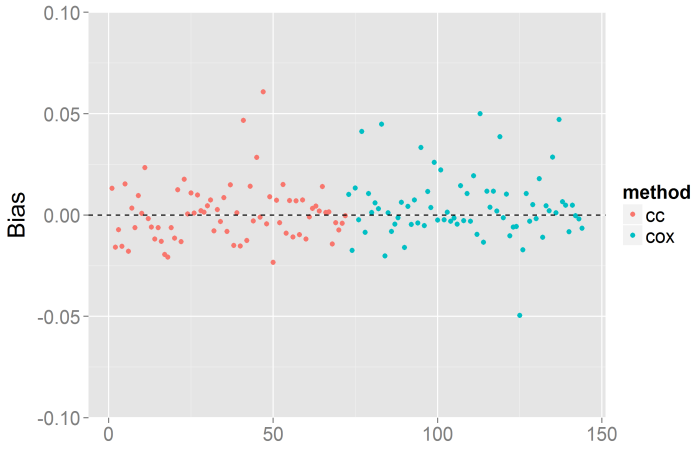

图 6.1　估计值的误差分布

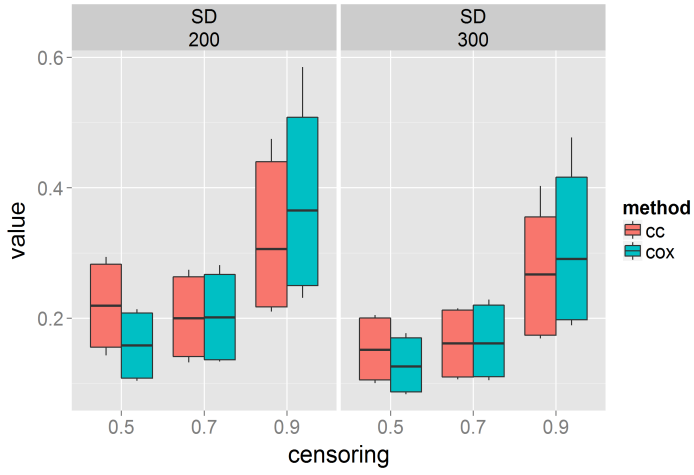

图 6.2　估计值的标准差分布

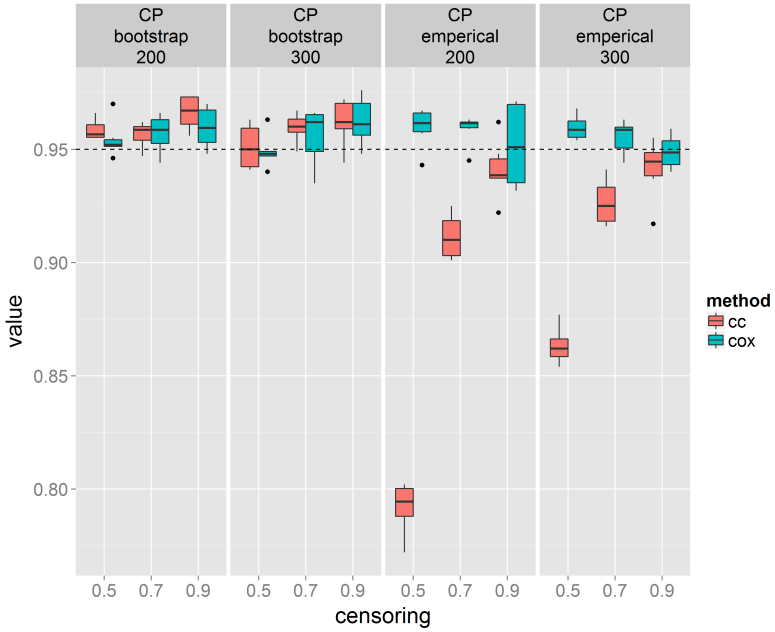

图 6.3　置信区间的覆盖率分布

图6.1表明病例队列设计和简单随机抽样下 Cox 比例风险模型的参数估计都是无偏的。无偏估计的方差等于其均方误差 MSE，依据定义均方误差越小估计越有效。从图6.2可以看出两个模型的标准差或均方误差都有随删失率的增大而增大、随样本量的增大而减小的趋势。相同样本量的条件下，在删失率为 0.5 时 Cox 比例风险模型估计的标准差或均方误差较小、估计更有效，而在删失率为 0.9 时病例队列设计估计的均方误差较小、估计更有效，说明了病例队列设计在删失率较高时估计的有效性。

图6.3表明采用标准差的 Bootstrap 估计时，两个模型的区间覆盖率大部分都在 95% 以上，其中相同样本量及删失率下病例队列设计的区间覆盖率要略高于 Cox 比例风险模型。而采用标准差的经验估计时，病例队列设计的区间覆盖率大部分没达到 95%，而 Cox 比例风险模型的区间覆盖率仍然大部分在 95% 以上。值得注意的是，随着删失率和样本量的增加，病例队列设计的区间覆盖率有明显的上升趋势，而 Cox 比例风险模型的区间覆盖率反而有下降趋势，进一步说明了病例队列设计在删失率较高时估计的有效性。

从图6.4可以看出，标准差的 Bootstrap 估计在两个模型中都与真实值 SD 的误差较小。另一方面，标准差的经验估计在 Cox 比例风险模型中的误差较小、接近 0 ，而在病例队列设计中误差很大，尤其在样本量较小、删失率较小的情况下，但随着删失率和样本量的增加误差有减小的趋势。这说明病例队列设计中

经验估计在小样本下的有偏性可能是经验估计下区间覆盖率没达到 95% 的原因之一。

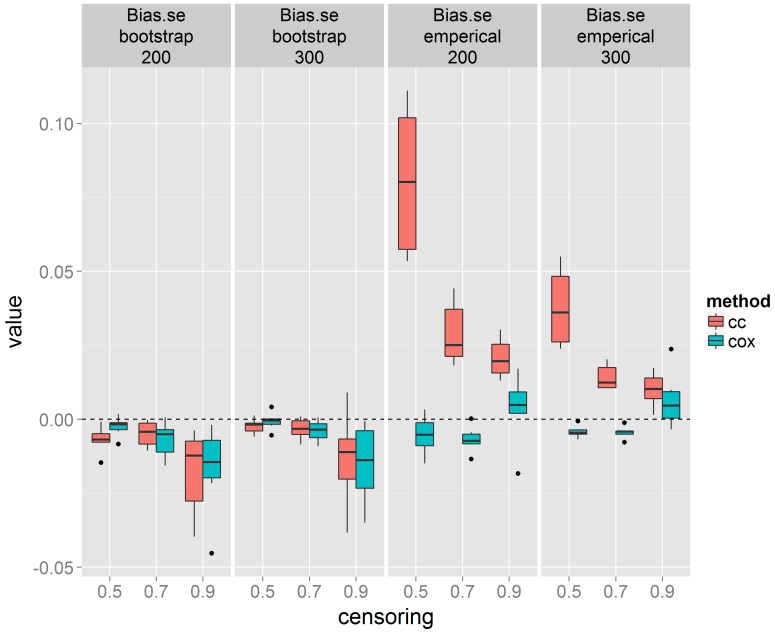

图 6.4　标准差估计的误差分布

有了统计模拟的结果，实例分析得到的参数估计的有效性便有了保证，即参数真值位于置信区间的概率为 95% 。事实上，根据表5.4的结果，处于第一阶段的肾母细胞瘤患者复发风险率最低，存活概率最高，之后依次是第二阶段、第三阶段及第四阶段患者，从而参数估计的结果与不同阶段肾母细胞瘤患者生存函数的 K-M 估计（图5.3）的结果是一致的。

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): 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] BINDER D A. Fitting Cox's Proportional Hazards Models from Survey Data[J]. Biometrika, 1992, 79: 139-47.

[4] BARLOW W E. Robust Variance Estimation for the Case-cohort Design[J]. Biometrics, 1994, 50: 1064-1072.

[5] LIN D Y, YING Z. Cox Regression with Incomplete Covariate Measurements[J]. Journal of the American Statistical Association, 1993, 88: 1341-1349.

[6] THERNEAU T M, LI H. Computing the Cox Model for Case Cohort Designs[J]. Lifetime data analysis, 1999, 5(2): 99-112.

[7] EFRON B. Bootstrap Methods: Another Look at the Jackknife[J]. The annals of Statistics, 1979, 7: 1-26.

[8] COX D R. Regression Models and Life-Tables (with discussion)[J]. Journal of the Royal Statistical Society, 1972, 34: 187-220.

[9] KAPLAN E L, MEIER P. Nonparametric Estimation from Incomplete Observations[J]. Journal of the American Statistical Association, 1958, 53(282): 457-481.

[10] BRESLOW N E, CHATTERJEE N. Design and Analysis of Two-phase Studies with Binary Outcome Applied to Wilms Tumour Prognosis[J]. Applied Statistics, 1999: 457-468.

附录

附录A 部分模拟程序

1. 辅助函数

```
aat <- function (a)   a %*% t (a)
tailsum <- function (a)   sum (a) + a - cumsum (a)
tailsum.cc <- function (a, d, x)   tailsum (a * x) + a * d * (1 - x)
risksum <- function (A)   apply (A, 2, tailsum)
risksum.cc <- function (A, d, x) apply (A * x, 2, tailsum) + A * d * (1 - x)
ccsum <- function (data, coef) {
   #计算病例队列设计中的梯度函数值和 Heissian 矩阵
   data <- as.matrix (data [sort.list (data $ time), ])
   d <- as.vector (data [ , 2])
   x <- as.vector (data [ , 3])
   z <- as.matrix (data [ , 4 : ncol (data)])
   b <- coef
   ezb <- exp (as.vector (z %*% b))
   zezb <- ezb * z
   z2 <- t (apply (z, 1, aat))
   z2ezb <- ezb * z2
   S0 <- tailsum.cc (ezb, d, x)
   S1 <- risksum.cc (zezb, d, x)
   S2 <- risksum.cc (z2ezb, d, x)
   S1S0 <- S1 / S0
   S2S0 <- S2 / S0
   S1S02 <- t (apply (S1S0, 1, aat))
   Gradient <- apply ((z - S1S0) * d, 2, sum)
   Hessian <- matrix (apply ((S2S0 - S1S02) * d, 2, sum), length (b))
   list (S1 = S1, S2 = S2, S1S0 = S1S0, S2S0 = S2S0, S1S02 = S1S02,
         Gradient = Gradient, Hessian = Hessian)
}
```

2. 病例队列设计的SE计算函数

```
ccseest <- function (data, est) {
    #计算标准差的经验估计
    size <- nrow (data)
    d <- data [ , 2]
    z <- data [ , 4 : ncol (data)]
    SS <- ccsum (data, est)
    S1S0 <- SS $ S1S0
    H <- SS $ Hessian
    H.inv <- solve (H / size)
    var.matrix <- var.sample ((z - S1S0) * d)
    var <- diag (H.inv %*% var.matrix %*% H.inv) / size
    se <- sqrt (var)
    se
}
```

附录B 实例原始数据

表 B1　美国国家肾母细胞瘤研究数据

seqno 序号	histol 组织学	stage 肿瘤阶段	study 试验序号	rel 是否复发	edrel 复发时间	age in months 年龄（月）	in.subcohort 是否在子队列中
1	2	1	3	0	6057	25	FALSE
2	1	2	3	0	4121	50	FALSE
3	2	1	3	0	6069	9	FALSE
4	2	4	3	0	6200	28	TRUE
5	2	2	3	0	1244	55	FALSE
6	1	2	3	0	2932	32	FALSE
...	...	...	...	...	...	...	...

完整数据见 R {survival} package data(nwtco) 。