运用Cox模型时打结数据的处理方法探讨*

张文丽[1,2]，张　彤[3]，易丹辉[1,2**]，杨宇飞[3**]

（1. 中国人民大学应用统计科学研究中心，北京　100872；2. 中国人民大学统计学院　北京　100872；
3. 中国中医科学院西苑医院　北京　100091）

摘　要：Cox回归模型是目前生存分析中最为广泛使用的方法之一，模型的假设之一是失效时间不存在打结情况，即个体之间有着不同的失效时间。在实际应用当中，生存时间数据存在打结是很常见的。目前有四种常见的处理方法：Exact法，discrete model法，Efron法以及Breslow法。本文研究目的是比较这四种处理方法的优劣。本文采用模拟进行比较，设置了不同的样本量和打结程度，比较四种方法在拟合统计量，计算时间，参数估计精确性等方面的表现，发现Exact法和discrete model拟合统计量结果最好，但计算耗时最久；Efron法以及Breslow法运算较快但是在参数估计方面存在偏差。另外，样本量和打结程度也影响处理的结果，总的来说，当结点数较小时，四种方法之间差别不大。当数据量较大或打结比例较高时，除exact以外的三种近似方法的偏差增加，但运算时间无明显变化，而exact法的运算时间迅速增加。此时如果估计的精确性没有估计时间那么重要，Efron法以及Breslow法是不错的选择，其中，Efron法更为精确而Breslow方法倾向于低估正确的β值。如果时间上没有限制，可以选择Exact法和discrete model，将得到更为精确的结果。

关键词：生存分析　Cox模型　打结数据　部分似然函数

doi：10.11842/wst.2017.09.007　　中图分类号：R33　　文献标识码：A

Cox回归模型，或叫相对风险模型（Relative Risk Model）是目前生存分析中最为广泛使用的方法之一。该模型最早由Cox D.R在1972年提出。该模型无需对基准风险函数做任何的限制，是半参数模型，克服了生存分析传统的参数法和非参数法的局限性，目前广泛运用于不同治疗方法的比较以及各种疾病预后相关因素的研究[1]。

Cox模型在应用中还存在一些问题，其中之一就是数据存在结点时的处理方法。数据存在打结是指有多个个体有相同的失效时间。在实际应用中，由于失效时间往往是以一种离散的方式记录的，得到的数据存在打结是很普遍的。

失效时间不存在打结是Cox模型的一个重要假设，与模型的估计紧密相关，在该假设不满足时仍可以使用Cox模型，但需要对估计的方法进行改进。

目前有四种常见的方式用来处理打结数据，分别是Exact model、discrete model、Efron法以及Breslow法。研究表明，Exact法和discrete model的结果较为精确，但是计算时间较长。Efron法以及Breslow法用时短，但参数估计偏差较大。Breslow方法计算比较简便，是目前大多数软件默认的处理Cox模型打结数据的方法，但是在R软件的survival包中，默认的方法是Efron法。

本文对四种不同的处理打结数据方法的原理进行讨论，并利用模拟讨论在不同样本量水平和打结程度下，在参数估计，估计量效率（efficiency of estimators），拟合统计量，计算时间方面的表现。最后结合具体数

收稿日期：2017-05-18

修回日期：2017-08-23

*　中国人民大学2017年度‘中央高校建设世界一流大学（学科）和特色发展引导专项资金’，负责人：易丹辉；和教育部人文社会科学重点研究基地重大项目(16JJD910002)：基于大数据的精准医学生物统计分析方法及其应用研究，负责人：??。

**　通讯作者：易丹辉，中国人民大学教授，博士生导师，主要研究方向：风险管理与保险、预测与决策。杨宇飞，博士生导师，中国中医科学院西苑医院肿瘤诊治部主任、主任医师，主要研究方向：中西医结合癌症治疗。

据进行展示。

1 方法介绍

Cox模型构建的思路是，所有研究对象的生存情况是多个影响因素共同作用的结果，可用风险函数表示，记为 $\lambda(t;x)$，其中 $x=(x_1,x_2,\cdots)'$ 是基准协变量，在个体进入试验之前或进入试验之时已经测得，T 是绝对连续的失效时间变量[2]。假设研究对象中影响生存的因素均不存在，其生存情况用 $\lambda_0(t)$ 表示，称为基础风险函数，那么，研究对象的实际生存情况，即 $\lambda(t;x)$，是影响因素 x 在基础风险函数的基础上，进一步修改的结果，即风险函数可表达为：

$$\lambda(t;x)=\lim_{h\to 0^+}\frac{P(t\leqslant T<t+h|T\geqslant t)}{h}=\lambda_0(t)r(t;x) \quad (1.1)$$

英国统计学家D. R. Cox于于1972年首次提出把 $r(t;x)$ 构造为指数形式，即将风险函数写作(2.2)式：

$$\lambda(t;x)=\lambda_0(t)\exp(Z(t)'\beta) \quad (1.2)$$

其中 $Z(t)=[Z_1(t),\cdots,Z_p(t)]'$ 是一个可能时间相依的协变量向量，是时间 t 和基准协变量的函数。Cox回归模型可以估计相对风险，例如所建立的回归模型包含组别变量，该变量是一个二分类变量，以1表示治疗组，0表示对照组，则治疗组与对照组风险函数的比值为

$$HR(t|x=1,x=0)=e^{(1-0)\beta}=e^{\beta} \quad (1.3)$$

该比值的含义为，当其他变量相同时，在每个时间点，治疗组的死亡风险都是对照组的 e^{β} 倍。

对于定量数据的协变量，如年龄，HR 的结果是：

$$HR=\frac{\lambda_0(t)\exp(\beta(X_{i1}+m))}{\lambda_0(t)\exp(\beta X_{i1})}=\exp(m\beta) \quad (1.4)$$

即该协变量每增长 m 单位，HR 就要在原来的基础上乘以 $\exp(m\beta)$。

Cox模型的假设之一是失效时间不存在打结，在该假设成立的前提下，参数估计可以通过部分似然(partial likelihood)方法得到[3][4]。

用 z_l 表示第 l 个个体的解释变量，排序的失效时间为 $t_1<\cdots<t_k$。用 $D_i=\{i_1,\cdots,i_{d_i}\}$ 表示在 t_i 时刻失效的个体组成的集合，Q_i 是 $\{i_1,\cdots,i_{d_i}\}$ 的 $d_i!$ 个排列组成的集合，$P=(p_1,\cdots p_{d_i})$ 是 Q_i 中的一个元素，用 d_i 表示在 t_i 时刻失效的个体数，用 R_i 表示在 t_i 时刻的风险集。$R(t_j,P,r)=R(t_j)-\{p_1,\cdots,p_{r-1}\}$。

根据Cox(1972)提出的方法，数据不存在打结时，估计 β 的部分似然的公式为(2.5)式：

$$L(\beta)=\prod_{i=1}^{k}\left\{\frac{\exp[Z_i(t_i)'\beta]}{\sum_{l\in R_i}\exp[Z_l(t_i)'\beta]}\right\} \quad (1.5)$$

利用该方法估计参数时，似然函数的计算依赖于事件发生的顺序，个体的失效时间必须是有序的，如果有两个个体(例如A和B)有着相同的失效时间，在这种情况下，无法确定其中一个(如A)失效时，B是否在该时刻的危险集中。

当数据存在结点时，Kalbfleisch and Prentice(2002)[5]提出的exact法考虑在每个存在打结的时间点上事件发生的所有可能的排序。在 t_j 处将结点分解成各种可能的情形后的平均部分似然为：

$$\frac{1}{d_i!}\exp[s_i(t_i)'\beta]\sum_{P\in Q_i}\prod_{r=1}^{d_i}\left\{\sum_{l\in R(t_i,p,r)}\exp[z_l(t_i)'\beta]\right\}^{-1} \quad (1.6)$$

其中 $s_i(t_i)=\sum_{j=1}^{d_i}Z_{ij}(t_i)$，则对应的平均部分似然成比例于

$$\prod_{i=1}^{k}\left(\exp[s_i(t_i)'\beta]\sum_{P\in Q_i}\prod_{r=1}^{d_i}\left\{\sum_{l\in R(t_i,p,r)}\exp[z_l(t_i)'\beta]\right\}^{-1}\right) \quad (1.7)$$

当每一个失效时间点的结点数目比较大时，上式的计算量会非常大。此时，可以对似然函数进行近似。Breslow(1974)[6]提出的近似似然函数为

$$L(\beta)=\prod_{i=1}^{k}\left\{\frac{\exp\left[\left(\sum_{l\in D_i}z_l\right)'\beta\right]}{\left[\sum_{l\in R_i}\exp(z_l'\beta)\right]^{d_i}}\right\} \quad (1.8)$$

Efron方法(1977)[7]提出的似然函数为

$$L(\beta)=\prod_{i=1}^{k}\left\{\frac{\exp\left[\left(\sum_{l\in D_i}z_l\right)'\beta\right]}{\prod_{j=1}^{d_i}\left[\sum_{l\in R_i}\exp(z_l'\beta)-\frac{j-1}{d_i}\sum_{l\in D_i}\exp(z_l'\beta)\right]}\right\} \quad (1.9)$$

另外，当数据打结的比例较大时，可以考虑将失效时间看作离散变量。Cox(1975)建议利用离散Logistic模型(也叫条件logistic模型)：

$$\frac{d\lambda(t;x)}{1-d\lambda(t:x)}=\exp[Z(t)'\beta]\frac{d\lambda_0(t)}{1-d\lambda_0(t)} \quad (1.10)$$

其中 $d\lambda_0(t)$ 是一个未指定的离散风险函数，在观测失效时间点 $t_1<\cdots<t_k$ 有值，将没有结点情形下的部分似然进行推广，得到以下部分似然函数。

$$\prod_{i=1}^{k}\frac{\exp[s_i(t_i)'\beta]}{\sum_{l\in R_{d_i}(t_i)}\exp[s_l(t_i)'\beta]} \quad (1.11)$$

其中 $R_{d_i}(t_i)$ 是从风险集 $R(t_i)$ 中挑出 d_i 个个体的所有子集组成的集合，$l=(l_1,\cdots,l_{d_i})$ 是 $R_{d_i}(t_i)$ 的一个元素，$s_l(t_i)=\sum_{i\in l}Z_i(t_i)$。

Kalbfleisch 和 Prentice（2002）指出，Efron 法和 Breslow 法对参数的估计存在偏差，且 $\hat{\beta}$ 的方差估计值是不一致的。模拟显示，当 d_i/R_i 的值较大时，Breslow 法对参数的估计会有较大的偏差。结点数很少时，三种方法得到相似的结果，不存在结点时，三种方法会得到完全一样的结果[5]。

2 数据模拟

本文采用模拟比较 Exact 方法和三种近似方法在参数估计，估计量效率（efficiency of estimators），拟合统计量和计算时间方面的差异。对于计算时间，使用每种方法重复 10 次估计，以得到平均计算时间。

生成两组数据，一组是失效时间服从指数分布的生存时间数据，即风险函数为：

$$\lambda(t;x)=\lambda_0(t)=0.02$$

另一组是与该组基础风险函数相同，HR 是 e^{-1} 的生存时间数据，风险函数为

$$\lambda(t;x)=\lambda_0(t)\exp(Z(t)'\beta)=0.02e^{-1}$$

另外，是否删失的设置与生存时间独立，即生成的数据包括的变量为：生存时间，组别（0,1），是否删失（0,1）。

首先在每组 1 000 个个体的样本量水平上进行模拟。利用数据分组制造结点，分别将数据分到 $k=50,200,500$ 个时间区间中去来制造高、中、低三种打结水平。具体方法是用生存时间数据落入的区间的右端点来代替原先的数据以制造结点。

利用 SAS 9.4 中的 PHREG PRocedule 拟合模型。

当样本量为每组 1 000 个个体时，在结点数最多即 $k=50$，平均每个时间点有 20 人的情形下，各种方法的计算时间均小于 0.5 秒，因此不再对计算时间进行记录和比较。

表 1 展示了四种处理方法处理三种打结水平的 $\hat{\beta}$，p 值，标准差结果。可以看出，四种方法的参数估计结果差异不大，Breslow 法与其他三种方法相比，存在着低估参数绝对值的问题，在打结程度高时最为明显（−0.949 02）。所有 p 值结果均显著，选择不同的方法在是否拒绝原假设问题上会得到相同结论。几种方法估计的标准误差异不大，the discrete model 的标准误略大于其他三种方法。

表 1　$n=1\,000$　$\beta=-1$　Cox 模型模拟结果

k−处理方法	$\hat{\beta}$	p	SE
50−Exact	−0.977 48	<0.000 1	0.058 67
50−discrete	−1.006 83	<0.000 1	0.060 49
50−Efron	−0.977 13	<0.000 1	0.058 66
50−Breslow	−0.949 02	<0.000 1	0.058 67
200−Exact	−1.033 18	<0.000 1	0.059 62
200−discrete	−1.041 98	<0.000 1	0.060 15
200−Efron	−1.033 12	<0.000 1	0.059 61
200−Breslow	−1.023 57	<0.000 1	0.059 61
500−Exact	−1.050 84	<0.000 1	0.060 24
500−discrete	−1.055 32	<0.000 1	0.060 50
500−Efron	−1.050 83	<0.000 1	0.060 24
500−Breslow	−1.046 42	<0.000 1	0.060 24

表 2　$n=1\,000$　$\beta=-1$　Cox 模型模拟结果

k−处理方法	SV	−2 LOG L	AIC	SBC
50−Exact	0.060 02	10 667.211	10 667.211	10 667.211
50−discrete	0.060 08	10 667.211	10 667.211	10 667.211
50−Efron	0.060 03	18 497.136	18 497.136	18 497.136
50−Breslow	0.061 82	18 568.013	18 568.013	18 568.013
200−Exact	0.057 71	13 249.199	13 249.199	13 249.199
200−discrete	0.057 73	13 249.199	13 249.199	13 249.199
200−Efron	0.057 7	18 170.855	18 170.855	18 170.855
200−Breslow	0.058 24	18 192.625	18 192.625	18 192.625
500−Exact	0.057 33	14 877.143	14 877.143	14 877.143
500−discrete	0.057 33	14 877.143	14 877.143	14 877.143
500−Efron	0.057 33	18 346.169	18 346.169	18 346.169
500−Breslow	0.057 57	18 356.370	18 356.370	18 356.370

表 2 展示了四种处理方法处理三种打结水平的 SV 和三种拟合统计量结果。其中 SV（standardized measures of variability）的定义为：

$$SV=\frac{\hat{\sigma}_\beta}{|\hat{\beta}|}$$

该统计量可以用来衡量参数估计值的有效性（efficiency of estimators）。

可以看到，Breslow 方法在三种打结程度均有着最高的 SV 值，在参数估计值的有效性方面表现较差，其他三种方法之间不存在明显差异。

在拟合统计量方面，discrete model 和 Exact 法的表

表3 $n=100\,000$ $\beta=-1$ Cox模型模拟结果

k-处理方法	$\hat{\beta}$	p	SE	Real time
100-Exact	-	-	-	-
100-discrete	-1.019 67	<0.000 1	0.006 10	44 min.10 sec.
100-Efron	-0.993 34	<0.000 1	0.005 94	0.60 sec
100-Breslow	-0.968 17	<0.000 1	0.005 94	0.67 sec
500-Exact	-1.001 02	<0.000 1	0.005 95	47 min.27 sec.
500-discrete	-1.009 79	<0.000 1	0.006 00	15 min.39 sec.
500-Efron	-1.000 99	<0.000 1	0.005 95	0.54 sec.
500-Breslow	-0.992 38	<0.000 1	0.005 95	0.59 sec.
1 000-Exact	-0.996 14	<0.000 1	0.005 94	16 min.01 sec.
1 000-discrete	-1.000 50	<0.000 1	0.005 96	7 min.41 sec.
1 000-Efron	-0.996 13	<0.000 1	0.005 94	0.64 sec.
1 000-Breslow	-0.991 81	<0.000 1	0.005 94	0.53 sec.

注：由于计算机内存不足，$k=100$ 时未得到Excat法的结果，用“-”表示

表4 $n=100\,000$ $\beta=-1$ Cox模型模拟结果

k-处理方法	SV	-2 LOG L	AIC	SBC
100-Exact	-	-	-	-
100-discrete	0.005 98	1 135 281.9	1 135 281.9	1 135 281.9
100-Efron	0.005 98	3 099 950.8	3 099 950.8	3 099 950.8
100-Breslow	0.006 14	3 106 182.8	3 106 182.8	3 106 182.8
500-Exact	0.005 94	1 432 886.8	1 432 886.8	1 432 886.8
500-discrete	0.005 94	1 432 886.8	1 432 886.8	1 432 886.8
500-Efron	0.005 94	3 097 018.1	3 097 018.1	3 097 018.1
500-Breslow	0.006 00	3 099 121.2	3 099 121.2	3 099 121.2
1 000-Exact	0.005 96	1 624 697.3	1 624 697.3	1 624 697.3
1 000-discrete	0.005 96	1 624 697.3	1 624 697.3	1 624 697.3
1 000-Efron	0.005 96	3 105 100.5	3 105 100.5	3 105 100.5
1 000-Breslow	0.005 99	3 106 163.6	3 106 163.6	3 106 163.6

注：由于计算机内存不足，$k=100$ 时未得到Excat法的结果，用“-”表示

表5 数据打结情况

时间(月)	事件数	时间(月)	事件数
6	1	17	3
7	1	18	5
8	2	19	5
9	3	20	1
11	4	21	2
12	4	22	1
13	1	23	1
14	6	26	1
15	9	27	1
16	2		

现最好，其次是Efron法，Breslow法的结果最差。

在每组100,000个个体的样本量水平上进行模拟，样本量变大本身也会导致结点数增加，设置了k=100,500,1 000三个打结水平。

表3展示了四种处理方法处理三种打结水平的$\hat{\beta}$，p值，标准差，运算时间。样本量为100,000时，四种方法的结果与真实参数的偏差都不大。所有p值结果均显著，选择不同的方法在是否拒绝原假设问题上会得到相同结论。几种方法估计的标准误差异不大，the discrete model的标准误略大于其他三种方法。在计算时间方面，打结水平变高和样本量变大都会导致Exact方法计算时间迅速增加，表中$k=100$时Exact法结果为- 的原因是SAS日志窗口显示电脑内存不够无法利用该方法进行计算（模拟时所用电脑的处理器型号为：Intel(R) Core(TM) i5-4200U CPU @ 1.60 GHz 1.60 GHz 2.30 GHz，内存为4.00 GB(3.72 GB可用)）。与之形成对比时，即使在样本量为100,000，平均每个时间点有1 000个个体的情况下，Efron法和Breslow法的运算时间仍不超过1秒钟。

表4展示了四种处理方法处理三种打结水平的SV和三种拟合统计量结果。

Breslow方法在三种打结程度均有着最高的SV值，在参数估计值的有效性方面表现较差，其他三种方法之间不存在明显差异。在拟合统计量方面，the discrete model和Exact法的表现最好，Efron法和Breslow法的结果几乎是另外两种方法的3倍。

4 实际应用

某医院采用随机对照临床研究方法，纳入晚期结直肠癌患者60例，经过数据处理有效数据共53例。其中23人的治疗结局为死亡，30人的治疗结局为未死亡，即有56.6%的数据右删失。两组均采用常规治疗(营养支持、化疗、对症、中医)，治疗组在此基础上加用某中药，对照组加用安慰剂胶囊，治疗一段时间后进行随访，观察两组患者的生存期情况。

数据中主要考虑的指标包括：组别(治疗组=1，对照组=0)，性别(女=1，男=0)，年龄，患病阶段(阶段1、阶段2、阶段3、阶段4)，是否死亡，OS生存期(单位：月)。其中OS生存期数据存在结点，具体情况如表5所示。这种情况下，运用Cox模型比较两组的治疗，必须考虑打结数据的处理。表6至表10是四种方法处理该数据的结果。

表6是$\hat{\beta}$结果，由表可以看出，不同方法得到的参数估计值存在差异。Efron法得到的结果与Exact法最

为接近。Breslow法得到的结果与Exact法得到的结果相比，在四个协变量上都有低估参数绝对值的倾向。这一结果与Hertz-Picciotto 和 Rockhill（1977）得到的结论一致。同时可以看到，与Exact法相比，discrete法在四个协变量上都有高估参数绝对值的倾向。

表7是 p 值结果。由表可以看出，Breslow得到的 p 值较大。

表8是参数估计的标准误。由表可以看出，Exact法，Breslow法和Efron法的结果无明显差异，Discrete法得到的标准误与其他三种方法相比略大一些。

表9是SV结果，由表可以看出，Breslow法在四个协变量上的结果均是最差的，另外三种方法不存在明显差异。

表9是拟合统计量结果，由表可以看出，Exact法结果最好，discrete法与Exact法差别不大，Breslow法和Efron法结果较差。

该数据有较多结点，但由于样本量小，四种方法的计算时间都很短，推荐使用Exact法以得到更精确的结果，如果软件（如R）中不包含该方法，则推荐Efron法，该方法估计的结果与Exact法最接近，而其余两种方法偏差较大。

5 结论

样本量较小的情形下（$n \leqslant 1000$），不同程度的打结，四种方法的估计时间都很短，此时Breslow法会低估参数绝对值且偏差较大，Efron法表现好于Breslow法，discrete法会高估参数绝对值。但还是建议使用Exact法来获得最为精确的估计，R软件的survival包没有包括该方法，可以使用SAS软件的PHREG PRocedule，选择“exact”即可。

样本量较大的情况下（$n \geqslant 100{,}000$），Exact法的计算时间迅速增加，打结程度高时个人的电脑可能出现内存不足无法利用SAS运算该方法的情形，此时可以考虑使用discrete法，计算时间不到Exact法的一半。考虑到样本量较大，各种方法的偏差都小，更推荐使用Breslow法和Efron法，在样本量为100,000且打结程度最高时（平均每个时间点有1 000个个体），运算时间仍不超过1秒钟，其中Efron法更加精确。如果时间上没有限制，可以使用Exact法或discrete法，可以得到更好的拟合统计量 结果和更准确的参数估计结果。

表6 实际数据拟合Cox回归模型 $\hat{\beta}$

$\hat{\beta}$	Exact	Discrete	Breslow	Efron
组别	−1.622 99	−1.678 75	−1.545 39	−1.612 22
性别	0.647 65	0.649 54	0.609 18	0.645 79
年龄	−0.044 40	−0.044 62	−0.039 50	−0.043 61
阶段	0.437 59	0.446 85	0.411 73	0.433 58

表7 实际数据拟合Cox回归模型 p

p	Exact	Discrete	Breslow	Efron
组别	0.002 8	0.003 1	0.004 1	0.002 8
性别	0.171 8	0.187 3	0.197 1	0.172 2
年龄	0.032 1	0.045 2	0.049 4	0.032 0
阶段	0.145 3	0.145 7	0.166 6	0.147 8

表8 实际数据拟合Cox回归模型SE

	Exact	Discrete	Breslow	Efron
组别	0.542 54	0.568 45	0.537 80	0.539 79
性别	0.473 99	0.492 56	0.472 31	0.473 08
年龄	0.020 72	0.022 27	0.020 10	0.020 33
阶段	0.300 46	0.307 14	0.297 67	0.299 59

表9 实际数据拟合Cox回归模型SV

	Exact	Discrete	Breslow	Efron
组别	0.334 3	0.338 6	0.348 0	0.334 8
性别	0.731 9	0.758 3	0.775 3	0.732 6
年龄	0.466 7	0.499 1	0.508 9	0.466 2
阶段	0.686 6	0.687 3	0.723 0	0.691 0

表10 实际数据拟合Cox回归模型拟合统计量

statistics	Exact	Discrete	Breslow	Efron
−2 LOG	119.895	120.088	144.704	142.618
AIC	127.895	128.088	152.704	150.618
SBC	132.437	132.630	157.246	155.160

参考文献

1 陈兵, 骆福添. 生存分析中的回归模型. 中国卫生统计, 2006, 23(5): 462-465.

2 金丕焕, 陈峰. 医用统计方法. 复旦大学出版社, 2009: 378-385.

3 Cox D R. Regression Models and Life-Tables. *J Roy Stat Soc*, 1972, 34 (2): 187-220.

4 Cox D.R. *Partial likelihood*. *Biometrika*, 1975, 62(2): 269-276.

5 Kalbfleisch J D, Prentice R L. Marginal likelihoods based on Cox´s regression and life model. *Biometrika*, 1973, 60: 267-279.

6 Breslow N. Covariance analysis of censored survival data, *Biometrics*, 1974, 30: 89-99.
7 Efron B. The efficiency of Cox´s likelihood function for censored data. *J Am Stat Assoc*, 1977, 72, 557-565.
8 Hertz P I, Rockhill B. Validity and Efficiency of Approximation Methods for Tied Survival Times in Cox Regression. *Biometrics*, 1997, 53 (3): 1151-1156.
9 Borucka J. Methods of Handling Tied Events in the Cox Proportional Hazard Model. *Ieee*, 2014, 2(2): 92-106.

Discussion on Methods for Tied Survival Times in Cox Model

Zhang Wenli[1,2], Zhang Tong[3], Yi Danhui[1,2], Yang Yufei[3]

(1. Center for Applied Statistics of Renmin University of China, Beijing 100872, China;
2. School of Statistics, Renmin University of China, Beijing 100872, China;
3. Xiyuan Hospital of China Academy of Chinese Medical Sciences, Beijing 100091, China)

Abstract: Cox regression model is one of the most widely used methods in the survival analysis. One assumption of this model is that there is no tie in the failure times, that is, individual has different failure times. In practical applications, the existence of ties in time data is very common. In this paper, four common methods of dealing with ties in Cox model, including Exact method, discrete model method, Efron method and Breslow method, were compared with simulation. The results showed that Exact method and discrete model were the best, but they took the longest time. Efron method and Breslow method were faster but there was a greater deviation in parameter estimation. Moreover, the sample amount and ties degree also affect the results. In general, when there are a few ties, the difference between four methods was small; and in the case of large datasets or a large number of ties, the bias of three approximation methods increased except Exact method. However, there was no significant change on computational time. While the computational time of the Exact method increased rapidly. Therefore, if the estimation precision is not as important as the estimation time, Efron method and Breslow method will be good choices. Efron method is more preferably as it is more precise. And Breslow method tends to underestimate the true β. If there is no limit in time, Exact method and discrete model can be chosen to achieve more accurate results.

Keywords: Survival analysis, Cox model, tied data, partial likelihood function

（责任编辑：张娜娜，责任译审：王 晶）