DOI:10.15923/j.cnki.cn22-1382/t.2017.2.02

区间删失数据的 3 种统计模型分析及其 SAS 实现

张倩倩， 王纯杰*， 佟知真， 李纯净

（长春工业大学 基础科学学院，吉林 长春 130012）

摘 要： 借助 SAS9.4 中 PROC ICLIFETEST、PROC ICPHREG 过程步编写宏程序，同步实现了区间删失数据的生存函数估计、广义 Log-Rank 检验和 PH 类型回归模型的统计推断。结合回溯研究中 368 个样本 HIV-1 感染时间的区间删失数据给出实证分析。

关键词： 区间删失；ICLIFETEST；广义对数秩检验；ICPHREG；宏程序

中图分类号： O 212.3 **文献标志码：** A **文章编号：** 1674-1374(2017)02-0111-06

SAS-based study on three statistical models for interval censored data analysis

ZHANG Qianqian， WANG Chunjie*， TONG Zhizhen， LI Chunjing

(School of Basic Sciences, Changchun University of Technology, Changchun 130012, China)

Abstract: Applying the procedures such as PROC ICLIFETEST and PROC ICPHREG in SAS9.4, we realize the statistic deduction for the survival function, Generalized log-rank test and PH regression for the interval censored data by means of macro programming. Interval censored data during the infection time from 368 samples are analyzed in the prospectively study.

Key words: interval censored; ICLIFETEST; generalized Log-rank test; ICPHREG; macro program.

0 引 言

生存分析是对试验或调查得到的人或生物的生存时间数据进行推断，在医学实践中有着广泛应用。一般称给定事件的出现时间为生存时间[1]，分析生存时间数据通常意味着解决 3 个问题：估计生存函数，比较处理组或者生存函数，评估协变量的影响或者依靠生存时间的解释变量。区间删失数据是生存时间中越来越常见的一种数据，在过去几十年里，出现了许多分析区间删失数据的统计方法。Turnbull[2]找到了类似右删失数据下的 Kaplan-Meier 估计的自相合算法来获得生存函数估计；王弄升[3]2012 年利用 SAS 软件中宏程序%EMICM 给出区间删失数据生存函数的估计；Sun[4]等把 Log-Rank 检验推广到区间删失数据中，提出广义对数秩检验；Finkelstein D M[5]给出区间删失数据的 COX 回归模型。但是基于 SAS 软件还没有完整的程序可以同步实现

收稿日期： 2017-01-22

基金项目： 国家自然科学基金(青年基金)资助项目(11301037)；国家自然科学基金资助项目(11571051,11671054)；吉林省教育厅“十三五”规划项目(2016317)

作者简介： 张倩倩(1991—)，女，汉族，山东德州人，长春工业大学硕士研究生，主要从事生存分析方向研究，E-mail:zhangqqxiao@163.com. * 通讯作者：王纯杰(1978—)，女，汉族，辽宁灯塔人，长春工业大学副教授，博士，主要从事生存分析方向研究，E-mail:cjwang2014@126.com.

区间删失数据3种统计分析任务。因此，文中借助SAS9.4中PROC ICLIFETEST[6]、PROC ICPHREG过程步编写宏程序，实现了区间删失数据的生存函数估计、广义Log-Rank检验和PH比例风险类型的回归模型统计推断。

1 模型及方法介绍

1.1 区间删失的数据类型

设T为非负的随机变量，代表研究中个体的生存时间，对于区间删失数据，只能知道T落在某个区间内，即$T\in(L,R]$，在这里$L\leqslant R$。区间删失数据可分为Ⅰ型区间删失与Ⅱ型区间删失。Ⅰ型区间删失数据可以表示$\{C,\delta=I(T\leqslant C)\}$，$C$代表观测时间，$I(\cdot)$是示性函数。Ⅱ型区间删失数据是包括有限区间的区间删失数据(L,R)，假设每个个体观测两次，L,R是两个随机变量，$L\leqslant R$以概率1成立。

1.2 区间删失数据的非参数估计

1.2.1 EMICM算法[1]

SAS软件计算非参数生存函数的方法是EMICM法。该算法是一个把自相合与ICM算法简单结合的混合算法。考虑一个包含n个独立个体带有生存函数为$S(t)$的失效时间的研究。令T_i代表个体$i(i=1,2,\cdots,n)$的生存时间。假设在T_i上的区间删失数据被观测如下：

$$O=\{(L_i,R_i];i=1,2,\cdots,n\}$$

这里$(L_i,R_i]$是观测T_i属于的区间。令$\{s_j\}_{j=0}^{m}$代表$\{0,L_i,R_i;i=1,2,\cdots,n\}$的次序元素。$a_{ij}=I(s_j\in(L_i,R_i])$，$p_j=S(s_{j-1})-S(s_j)$，$j=1,2,\cdots,m$。

则区间删失数据的似然函数为：

$$L_s(p)=\prod_{i=1}^{n}[S(L_i)-S(R_i)]=\prod_{i=1}^{n}\sum_{j=1}^{m}a_{ij}p_j \tag{1}$$

这里$p=(p_1,p_2,\cdots,p_m)'$，$\sum_{j=1}^{m}p_j=1$，$p_j\geqslant 0$ $(j=1,2,\cdots,m)$。

NPMLE算法就是要最大化此似然函数。在自相合算法中对数似然为：

$$l_S(p;T_1,T_2,\cdots,T_n)=\log\Big[\prod_{i=1}^{n}dF(T_i)\Big]=\sum_{j=1}^{m}d_j^{*}\log p_j \tag{2}$$

这里$d_j^{*}=\sum_{i=1}^{n}I(T_i=s_j)$，$j=1,2,\cdots,m$。在ICM算法中，令$\beta_j=F(s_j)$。且$\beta_0=0,\beta_m=1,\beta=(\beta_1,\beta_2,\cdots,\beta_{m-1})'$，因此以上的似然函数可写为：

$$L_s(\beta)=\prod_{i=1}^{n}\sum_{j=1}^{m}a_{ij}(\beta_j-\beta_{j-1}) \tag{3}$$

Wellner，Zhan[7]指出，当NPMLE存在且唯一并且对数似然函数连续可微的时候，EMICM算法收敛到NPMLE。应用EMICM算法需要选择一个收敛准则。这里选择基于Robertson[8]1988年提出的Fenchel优化条件。这种准则下如果满足

$$\left|\sum_{j=1}^{m-1}\hat{\beta}_j^{(l)}\ \frac{\partial}{\partial\beta_j}l_S(\hat{\beta}_j^{(l)})\right|<\varepsilon$$

$$\max\left\{\sum_{u=j}^{m-1}\frac{\partial}{\partial\beta_u}l_S(\hat{\beta}^{(l)});j=1,2,\cdots,m-1\right\}<\varepsilon$$

就停止迭代，并且接受$\hat{\beta}^{(l)}=(\hat{\beta}_1^{(l)},\hat{\beta}_2^{(l)},\cdots,\hat{\beta}_{m-1}^{(l)})'$可以作为$F$的NPMLE。

1.2.2 广义对数秩检验[1]

广义对数秩检验考虑区间删失数据，$\boldsymbol{Z}_i$为受试组的p维矢量，目标是检验原假设$H_0:p+1$维受试组的生存函数都是相同的。在假设H_0下，$\hat{S}_0$是普通生存函数$S_0(t)=P(T_i>t)$的极大似然估计，$s_1<s_2<\cdots<s_m$是$\{L_i,R_i,i=1,2,\cdots,n\}$的跳跃点。

定义$\alpha_{ij}=I(s_j\in(L_i,R_i])$，$s_j\in(L_i,R_i]$，$i=1,2,\cdots,n$，$j=1,2,\cdots,m+1$，$\delta_i=I(R_i\leqslant s_m)$。

定义$\rho_{ij}=I(\delta_i=0,L_i\geqslant s_j)$，如果$H_0$为真，则$S_0(t)$被视为已知，在$s_j$点处所有失效的观测数和在风险数目的估计如下：

$$d_j=\sum_{i=1}^{n}\delta_i\frac{\alpha_{ij}[\hat{S}_0(s_{j-})-\hat{S}_0(s_j)]}{\sum_{u=1}^{m+1}\alpha_{iu}[\hat{S}_0(s_{u-})-\hat{S}_0(s_u)]}$$

$$n_j=\sum_{r=j}^{m+1}\sum_{i=1}^{n}\delta_i\frac{\alpha_{ir}[\hat{S}_0(s_{r-})-\hat{S}_0(s_r)]}{\sum_{u=1}^{m+1}\alpha_{iu}[\hat{S}_0(s_{u-})-\hat{S}_0(s_u)]}+\sum_{i=1}^{n}\rho_{ij}$$

$$j=1,2,\cdots,m$$

类似的对于第 l 组，$l=1,2,\cdots,p+1$，且 $j=1,2,\cdots,m$ 时，在 s_j 处失效数和在风险数的估计为：

$$d_{jl}=\sum_{i}^{l}\delta_i\frac{\alpha_{ij}\left[\hat{S}_0(s_{j-})-\hat{S}_0(s_j)\right]}{\sum_{u=1}^{m+1}\alpha_{iu}\left[\hat{S}_0(s_{u-})-\hat{S}_0(s_u)\right]}$$

$$n_{jl}=\sum_{r=j}^{m+1}\sum_{i}^{l}\delta_i\frac{\alpha_{ij}\left[\hat{S}_0(s_{r-})-\hat{S}_0(s_r)\right]}{\sum_{u=1}^{m+1}\alpha_{iu}\left[\hat{S}_0(s_{u-})-\hat{S}_0(s_u)\right]}+\sum_{i}^{l}\rho_{ij}$$

为了检验 H_0，给出统计量

$$U_r=(U_{r,1},\cdots,U_{r,p+1})'$$

$$U_{r,l}=\sum_{j=1}^{m}\left(d_{jl}-\frac{n_{jl}d_j}{n_j}\right)$$

此检验统计量是服从自由度为 P 的卡方分布。

1.2.3　PH 比例风险模型[1]

令 $S(t;Z)$ 表示带协变量 Z 的生存函数。似然函数正比于 $L=\prod_{i=1}^{n}[S(L_i,Z_i)-S(R_i,Z_i)]$，对所有的 $i=1,2,\cdots,n$。

给定协变量，Z 的比例风险模型形式为 $\lambda(t;Z)=\lambda_0(t)\exp(Z'\beta)$，累积风险函数为 $\Lambda_0(t)=\int_0^t\lambda_0(s)\mathrm{d}s$。在 PH 模型下给定协变量 Z，T 的生存函数为：

$$S(t;Z)=\exp[-\Lambda_0(t)\exp(Z_i'\beta)]=[S_0(t)]^{\exp(Z_i'\beta)}$$

则似然函数的对数形式为：

$$l(\beta,S_0)=\sum_{i=1}^{n}\log\{[S_0(L_i)]^{\exp(Z_i\beta)}-[S_0(R_i)]^{\exp(Z_i'\beta)}\}$$

利用极大似然函数的得分方程估计 β，基准生存函数 S_0 可由连续的阶梯函数来估计。

2　实证分析

2.1　区间删失数据及变量情况

数据来源于美国和欧洲的 16 个研究中心的第 5 中心[9]，此中心采用回溯研究的方法，主要研究带有血友病的病人感染 HIV-1 的风险。血友病人的治疗血液制品来自于成千上万个捐赠者的血浆制成的Ⅷ型凝血因子和Ⅸ型因子，所以这些病人都存在感染 HIV-1 的风险。文中对病人的 HIV-1 感染时间只得到了区间删失数据，且病人依据他们每年获得血液制品的平均剂量被分配到不同的组。来自第 5 中心的 368 个病人的观测数据，不考虑进入研究的病人 HIV-1 抗体状态，不接受或接受低剂量（1～20 000U）的Ⅷ型凝血因子（两组病人的组别记为 NF 和 LDF）见表 1。

两组病人的数量分别为 236 和 13，这些数据以季度为单位，0 代表研究开始时间（1978 年 1 月 1 日）。

表 1　HIV-1 感染数据

组别	序号	L	R	序号	L	R	序号	L	R	序号	L	R
NF	1	55	∞	60	50	∞	119	53	∞	178	0	31
	2	55	∞	61	45	∞	120	54	∞	179	56	∞
	3	56	∞	62	49	∞	121	54	∞	180	56	∞
	4	54	∞	63	52	∞	122	50	∞	181	25	31
	5	53	∞	64	52	∞	123	53	∞	182	35	∞
	6	57	∞	65	57	∞	124	54	∞	183	50	∞
	7	31	33	66	56	∞	125	37	∞	184	56	∞
	8	56	∞	67	55	∞	126	55	∞	185	0	35
	9	56	∞	68	14	49	127	48	∞	186	55	∞
	10	54	∞	69	54	∞	128	51	∞	187	54	∞
	11	56	∞	70	55	∞	129	54	∞	188	25	40
	…	…	…	…	…	…	…	…	…	…	…	…
	59	55	∞	118	24	28	177	54	∞	236	54	∞

续表1

组别	序号	L	R	序号	L	R	序号	L	R	序号	L	R
LDF	1	7	20	34	52	∞	67	51	∞	100	55	∞
	2	9	20	35	0	17	68	33	∞	101	51	∞
	3	0	25	36	0	21	69	17	26	102	0	30
	4	57	∞	37	46	∞	70	14	17	103	50	∞
	5	23	26	38	16	23	71	41	∞	104	45	∞
	6	8	21	39	24	32	72	42	∞	105	8	30
	7	20	26	40	16	24	73	53	∞	106	5	30
	8	25	27	41	53	∞	74	0	26	107	53	∞
	9	24	29	42	12	20	75	49	∞	108	11	41
	10	12	21	43	18	22	76	39	∞	109	52	∞
	11	26	29	44	0	33	77	18	29	110	3	33
	…	…	…	…	…	…	…	…	…	…	…	…
	33	15	19	66	54	∞	99	0	30	132	0	44

2.2 SAS程序

对于上述血友病患者HIV-1感染的区间删失数据，首先要分别建立两组的SAS数据集，然后用ICLIFETEST过程步估计出两个受试组HIV-1感染的时间生存函数，并检验两组生存曲线的区别，最后用ICPHREG过程步建立PH比例风险回归模型。区间删失数据3种生存统计分析可由SAS宏程序%ICLIFETEST-PHREG同步实现。

2.2.1 创建NF组的数据集

```
data NF;
input lTime rTime @@;
Stage=0;
datalines;
55. 55. 56. 54. 53. 57.
56. 56. 54. 56. 54. 55.
…;
run;
```

2.2.2 创建LDF组的数据集

```
data LDF;
input lTime rTime @@;
Stage=1;
datalines;
720 920 025 57.
2326 821 2026 2527
…;
run;
```

2.2.3 估计两组病人的生存概率及绘制生存曲线

```
data zq;
set LDF NF;
run;
proc iclifetest data=zq plot=s(cl) impute(seed=1234);/*ICLIFETEST过程步对应数据，画出带95%置信带的生存曲线图*/
strata stage;/*识别定义分组的变量*/
time (lTime,rTime);/*区间删失的左右观测时间*/
run;
```

2.2.4 两组病人数据的广义Log-Rank检验

```
proc iclifetest data=zq impute(seed=1234);
time (lTime,rTime);
test stage;/*检验的组别*/
run;
```

2.2.5 生存时间与协变量的PH比例风险回归

```
proc icphreg data=zq;
class Stage / desc;
model (lTime,rTime) = Stage / basehaz=pch(intervals=(10));/*协变量Stage与区间删失数据建立PH比例风险模型*/
hazardratio Stage;/*求风险比例*/
run;
```

上述5个程序是分别建立数据集，做区间删失数据的3种统计模型分析的SAS程序，对于临

床试验得到的这类区间删失数据，如果利用 SAS 程序做分析，需要进行上述复杂的 5 个程序的运行，文中给出一个宏程序，同步实现上述模型分析。

2.2.6　区间删失数据同步实现 3 种统计任务的 SAS 宏程序%ICLIFETEST-PHREG

```
%macro iclifetest－phreg(data,var1,var2);
proc iclifetest data=&data plot=s(cl) impute(seed=1234);
strata &var1;
time (lTime,rTime);
run;
proc iclifetest data=&data impute(seed=1234);
time(lTime,rTime);
test &var1;
run;
proc icphreg data=&data;
class &var2/ desc;
model (lTime,rTime) = &var2/ basehaz=pch(intervals=(10));
hazardratio &var1;
run;
%mend;
%iclifetest－phreg(zq,stage,stage)
```

宏程序%ICLIFETEST-PHREG 可以对含二分变量和其他协变量的区间删失数据同步实现生存函数的估计、广义的 Log-Rank 检验及 PH 比例风险类型的回归分析。其语法结构是：data：含二分变量和其他协变量的区间删失数据集；Var1：某一个二分协变量；Var2：做回归分析的所有的协变量集合。

2.3　结果分析

两个不同组的非参数生存估计与生存函数曲线的宏程序结果分别如图 1 和表 2 所示。

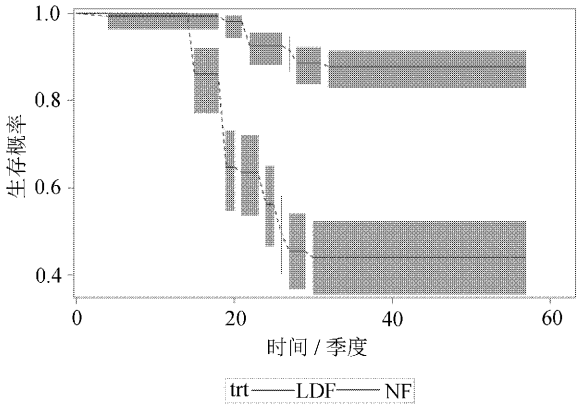

图 1　NF 组与 LDF 组估计的生存函数

表 2　NF 组与 LDF 组生存估计结果

组别	区间		失效概率	生存概率	标准误差
NF	4	18	0.005 5	0.994 5	0.005 4
	19	21	0.019 5	0.980 5	0.010 9
	22	26	0.074 2	0.925 8	0.018 7
	27	27	0.085 6	0.914 4	0.019 9
	28	31	0.115 0	0.885 0	0.021 1
	32	57	0.123 4	0.876 6	0.021 4
LDF	0	14	0.000 0	1.000 0	0.000 0
	15	18	0.139 0	0.861 0	0.036 8
	19	20	0.353 5	0.646 5	0.047 1
	21	23	0.364 8	0.635 2	0.047 2
	24	25	0.437 3	0.562 7	0.047 4
	26	26	0.506 2	0.493 8	0.046 0
	27	29	0.545 6	0.454 4	0.043 9
	30	57	0.560 6	0.439 4	0.043 2

图1中相应的NF组的生存概率要大于LDF组的生存概率，即不接受Ⅷ型凝血因子的血友病实验组的生存概率要高于接受低剂量Ⅷ型凝血因子的实验组。表2中NF组的失效概率要小于LDF组的失效概率。

两组生存概率是否相等的检验结果及回归的显著性检验结果分别见表3和表4。

表3 检验两组生存概率是否相等

权重	卡方值	自由度	$P>$卡方
SUN	91.247 4	1	<0.000 1

表4 回归的显著性检验

最大似然参数估计的分析							
参数	自由度	估计	标准误差	95%置信限		卡方	$Pr>$卡方
Stage	1	1.996	0.219 6	1.565 6	2.426 4	82.62	<0.000 1
Stage	0	0.000					

在显著性水平为0.05的情况下，表3和表4的p值均小于0.000 1，即拒绝原假设，LDF组和NF组的生存概率有明显区别。表4说明该回归模型参数是显著的。分组的风险比估计见表5。

表5 分组的风险比估计

以下对象的危险比：Stage		
说明	点估计	95% Wald 置信限
Stage 1 vs 0	7.360	4.786 11.318

表5结果显示，患有血友病的病人接受低剂量的Ⅷ型凝血因子组感染HIV-1病毒的风险概率要高于不接受组病人感染HIV-1病毒的风险，且是后者的7.360倍。

3 结 语

基于区间删失数据生存函数的估计算法除了EMICM算法还有自相合、ICM、Turnbull算法等。广义对数秩检验也成为检验生存函数是否相等的有用工具。实例说明，NF与LDF治疗组的生存函数有显著差异。回归模型检验出分组协变量对生存时间存在有效的影响。文中给出的SAS宏程序可以针对带有协变量的区间删失数据同步实现生存函数的估计、比较处理组及协变量对生存时间的影响这3种统计推断，对临床工作人员整理、分析实验结果有很大帮助。

参考文献：

[1] Sun J. The statistical analysis of interval-censored failure time data[M]. [S.l.]: Springer Science & Business Media, Inc, 2006.

[2] Turnbull B W. The empirical distribution function with arbitrarily grouped, censored and truncated data[J]. Journal of the Royal Statistical Society, 1976, 38(3): 290-295.

[3] 王弄升，张晋昕，骆福添.含有区间删失数据时的生存函数估计及其SAS实现[J].中国医院统计，2012，19(1)：1-4.

[4] Zhao Q, Sun J. Generalized log-rank test for mixed interval-censored failure time data [J]. Statistics in Medicine, 2004, 23(10): 1621-1629.

[5] Finkelstein D M. A proportional hazards model for interval-censored failure time data [J]. Biometrics, 1987, 42(4): 845-54.

[6] Guo C, Ying S, Gordon J. Analyzing interval-censored data with the ICLIFETEST procedure [J]. SAS Global Forum, 2014, 279: 327-345.

[7] Jon A Wellner, Yihui Zhan. A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data [J]. Journal of the American Statistical Association, 1997, 439(92): 945-959.

[8] Robertson T, Wright F T, Dykstra R L. Order restricted statistical inference [J]. Journal of the American Statistical Association, 1990, 85(410): 111-112.

[9] Kroner B L, Rosenberg P S, Aledort L M, et al. HIV-1 infection incidence among persons with hemophilia in the United States and western Europe, 1978-1990. Multicenter Hemophilia Cohort Study [J]. Journal of Acquired Immune Deficiency Syndromes, 1994, 7(3): 279-86.