生存资料回归模型分析——生存资料参数回归模型分析SAS实现

张甜甜[1],刘红伟[1],刘媛媛[1],李长平[1,2],胡良平[2,3*]

(1. 天津医科大学公共卫生学院,天津 300070;
2. 世界中医药学会联合会临床科研统计学专业委员会,北京 100029;
3. 军事科学院研究生院,北京 100850
*通信作者:胡良平,E-mail:lphu927@163.com)

【摘要】 本文目的是介绍生存资料参数回归模型的SAS实现,包括创建SAS数据集、依据图示法选择模型、拟合参数模型和似然比检验。利用SAS中的LIFEREG过程绘制生存函数关于生存时间的关系图,拟合对应的参数分布回归模型,通过拟合优度检验选择最优的参数回归模型,最后对相关结果进行解释。

【关键词】 生存分析;参数回归;拟合优度检验;似然比检验;Weibull分布
中图分类号:R195.1 文献标识码:A doi:10.11886/scjsws20200106005

Regression analysis of the parametric models for the survival data——the SAS implementation of survival data parametric regression model analysis

Zhang Tiantian[1], *Liu Hongwei*[1], *Liu Yuanyuan*[1], *Li Changping*[1,2], *Hu Liangping*[2,3*]

(1. *School of Public Health, Tianjin Medical University, Tianjin* 300070, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*;
3. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*
**Corresponding author: Hu Liangping, E-mail: lphu927@163.com*)

【**Abstract**】 The purpose of this article was to introduce the SAS implementation of the parametric model for the survival data, including the creation of a SAS data set, a graphical selection method, fitting parametric model, and a likelihood ratio test. The relationship between survival time and survival function was drawn by using the LIFEREG procedure in SAS software, and the corresponding parametric distribution was fitted. The optimal parametric model was selected by the goodness-of-fit test, and the relevant results were explained.

【**Keywords**】 Survival analysis; Parametric regression; Goodness-of-fit tests; Likelihood-ratio tests; Weibull distribution

本期科研方法专题的第三篇文章介绍了生存资料参数回归模型有关的基础知识,包括构建三个常见的生存资料参数回归模型的基本原理、基于图示法判断生存时间服从何种概率分布的方法、基于最大似然估计法求解参数回归模型中的参数和两个参数回归模型拟合优度的比较。本文通过一个实例,利用SAS软件的LIFEREG过程介绍生存资料参数回归模型的SAS实现方法。

1 创建SAS数据集

【例1】本文所用数据是对149例糖尿病患者17年追踪调查数据,包括生存结局、生存时间、随访开始时年龄、体重指数等[1]。由于存在删失数据,此资料为生存资料。Diabetic数据集中生存时间(t,年)、随访开始时年龄(age1,岁)、体重指数(BMI)、诊断出糖尿病时的年龄(age0,岁)、收缩压(SBP,mmHg)和舒张压(DBP,mmHg)是定量自变量;吸烟状况(smk,0表示不吸烟,1表示曾吸烟,2表示吸烟)、心电图读数(ECG,1表示正常,2表示临界,3表示异常)、是否有冠心病(CHD,0表示无,1表示有)和结局(status,0表示截尾,1表示死亡)都是定性变量。利用以下SAS数据步程序,创建名为Diabetic的数据集:

```
data Diabetic;
Input ID status t age1 BMI age0 smk SBP DBP ECG CHD;
MBP = round(SBP*(1/3)+DBP*(2/3),1);
CARDS;
1  0  12.4  44  34.2  41  0  132  96  1  0
2  0  12.4  49  32.6  48  2  130  72  1  0
```

基金项目:国家自然科学基金项目(项目名称:贝叶斯生存分析方法在肝细胞癌肝移植患者预后预测中的应用研究,项目编号:81803333)

```
3  0   9.6  49  22.0  35  2  108  58  1  1
4  0   7.2  47  37.9  45  0  128  76  2  1
5  0  14.1  43  42.2  42  2  142  80  1  0
……
145  1   5.5  75  35.8  66  0  162  78  1  0
146  0  11.0  40  34.0  38  2  132  76  1  0
147  0  73.0  61  19.9  37  0  120  60  2  1
148  1  10.6  62  30.6  49  0  160  86  2  1
149  0  10.5  49  30.8  47  1  146  86  1  0
;
run;
```

【SAS程序说明】因篇幅所限，在CARDS语句后的数据仅列出前5个和后5个观测。因为收缩压和舒张压有一定关联[2]，分析时取加权平均血压（MBP），即令MBP=SBP*(1/3)+DBP*(2/3)，权重系数分别为1/3和2/3。

2 图示法选择模型

2.1 SAS程序

第一部分：

```
proc lifetest method=KM data=diabetic outsurv=surv_data;
time t*status(0);
run;
```

第二部分：

```
data surv_data_new;
set surv_data;
Ls=-log(survival);
lls=log(-log(survival));
lss=log((1-survival)/survival);
lnt =log(t);
run;
```

第三部分：

```
symbol interpol=join value=circle;
axis1 order=0 to 0.4 by 0.1;
axis2 order=0 to 3 by 0.5;
proc gplot data=surv_data_new;
plot ls*t /vaxis=axis1;
plot lls*lnt /haxis=axis2;
plot lss*lnt /haxis=axis2;
run;
quit;
```

【SAS程序说明】第一部分语句产生生存函数的点估计量；第二部分语句根据三个关系式：$\log S(t) = -\lambda t$、$\ln[-\ln S(t)] = \gamma\ln\lambda + \gamma\ln t$ 和 $\log\left[\frac{1-S(t)}{S(t)}\right] = \gamma\ln\lambda + \gamma\ln t$ 分别生成LS、LLS和LSS三个变量；第三部分绘制出三幅关系图：对数生存图、对数-对数生存图和对数失效比生存图，利用这些图形的表现，判断数据是否适合指数分布模型、Weibull分布模型或Log-logistic分布模型（判断方法参见本期科研方法专题第三篇文章，此处从略）。

2.2 主要输出结果

对数生存图、对数-对数生存图和对数失效比生存图分别见图1、图2和图3。

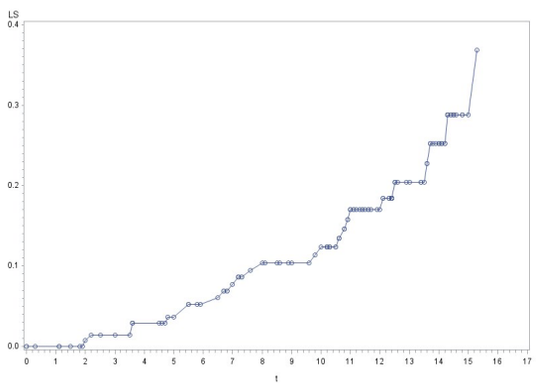

图1 对数生存图

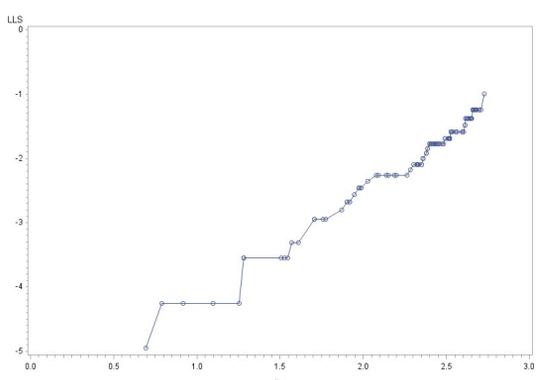

图2 对数-对数生存图

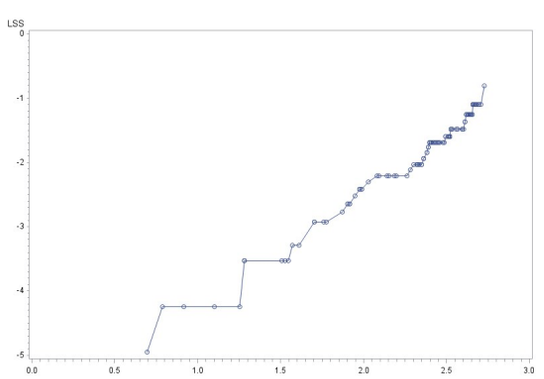

图3 对数失效比生存图

图 1 中的线图有些弯曲，图 2 在整体上呈线性趋势，与图 3 接近。图示法表明，基于现有数据，该生存资料同时适合 Weibull 分布模型和 Log-logistic 分布模型，本文选择 Weibull 分布模型进行拟合（图示法虽然简单直观，但不够精确，若深入探讨该数据是否更加适合 Weibull 分布模型，可参考其他方法[3]）。

3 拟合参数回归模型

3.1 SAS 程序

因为 Weibull 分布回归模型嵌套于广义 Gamma 分布回归模型[4]，为选择最优的模型，本文将分别拟合 Weibull 分布回归模型和广义 Gamma 分布回归模型，根据似然比检验结果来确定最优模型。利用以下 SAS 过程步程序，构建 Weibull 分布回归模型和广义 Gamma 分布回归模型。

```
proc lifereg data=Diabetic;
class CHD smk ECG;
model t*status (0) =age1 BMI age0 MBP CHD smk ECG/DIST=WEIBULL;
run;
proc lifereg data=Diabetic;
class CHD smk ECG;
model t*status (0) =age1 BMI age0 MBP CHD smk ECG/DIST=GAMMA;
run;
```

【SAS 程序说明】采用 LIFEREG 过程分别拟合 Weibull 分布回归模型和广义 Gamma 分布回归模型，class 语句列出离散型自变量，model 语句左边为生存时间和生存结局变量（括号内为截尾值的标记），右边为预测变量（即自变量），包括离散和连续自变量。

3.2 主要输出结果

Weibull 分布回归模型输出结果：

Fit Statistics

-2 Log Likelihood	85. 899
AIC (smaller is better)	107. 899
AICC (smaller is better)	109. 84
BIC (smaller is better)	140. 868

Analysis of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr>ChiSq
Intercept		1	5. 6336	1. 0014	3. 6708	7. 5964	31. 65	<0. 0001
age1		1	-0. 0569	0. 0139	-0. 0841	-0. 0296	16. 73	<0. 0001
BMI		1	-0. 0049	0. 0137	-0. 0317	0. 0219	0. 13	0. 7194
age0		1	0. 0243	0. 0118	0. 0011	0. 0475	4. 23	0. 0397
MBP		1	-0. 0114	0. 0067	-0. 0245	0. 0016	2. 93	0. 0868
CHD	0	1	-0. 5440	0. 3833	-1. 2952	0. 2071	2. 01	0. 1558
CHD	1	0	0	-	-	-	-	-
smk	0	1	0. 1925	0. 2242	-0. 2470	0. 6320	0. 74	0. 3908
smk	1	1	0. 0331	0. 2251	-0. 4082	0. 4744	0. 02	0. 8831
smk	2	0	0	-	-	-	-	-
ECG	1	1	1. 2414	0. 4017	0. 4541	2. 0286	9. 55	0. 0020
ECG	2	1	0. 3495	0. 2312	-0. 1036	0. 8026	2. 29	0. 1306
ECG	3	0	0	-	-	-	-	-
Scale		1	0. 3361	0. 0530	0. 2468	0. 4578	-	-
Weibull Shape		1	2. 9751	0. 4691	2. 1842	4. 0525	-	-

广义 Gamma 分布回归模型输出结果：

Fit Statistics

−2 Log Likelihood	77.65
AIC（smaller is better）	101.65
AICC（smaller is better）	103.961
BIC（smaller is better）	137.617

Analysis of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr>ChiSq
Intercept		1	4.8936	0.6424	3.6346	6.1526	58.03	<0.0001
age1		1	−0.0502	0.0107	−0.0712	−0.0291	21.86	<0.0001
BMI		1	0.0203	0.0109	−0.0011	0.0418	3.44	0.0635
age0		1	0.0079	0.0088	−0.0094	0.0252	0.80	0.3703
MBP		1	−0.0134	0.0053	−0.0237	−0.0031	6.46	0.0110
CHD	0	1	−0.4614	0.2333	−0.9187	−0.0042	3.91	0.0479
CHD	1	0	0	−	−	−	−	−
smk	0	1	0.3575	0.1572	0.0493	0.6657	5.17	0.0230
smk	1	1	0.3561	0.1708	0.0213	0.6909	4.35	0.0371
smk	2	0	0	−	−	−	−	−
ECG	1	1	0.8746	0.2574	0.3701	1.3792	11.54	0.0007
ECG	2	1	0.0234	0.2494	−0.4653	0.5122	0.01	0.9252
ECG	3	0	0	−	−	−	−	−
Scale		1	0.3760	0.0842	0.2425	0.5831	−	−
Shape		1	−2.4742	0.9175	−4.2724	−0.6759	−	−

【SAS 结果说明】Fit Statistics 表是基于 t 为响应变量的最大似然估计得到的统计量，可以用来比较不同协变量的模型。Analysis of Maximum Likelihood Parameter Estimates 表给出了参数的估计，包括预测变量的回归系数和参数分布中参数的估计值。其中“Scale”代表“尺度参数”，“Shape”代表“形状参数”。

4 参数回归模型拟合优度检验

因 Weibull 分布回归模型包含 2 个参数，广义 Gamma 分布回归模型包含 3 个参数，则似然比拟合优度检验[5]的自由度为 1，χ^2 统计量为 8.249（=85.899−77.650）（注意：$\chi^2_{1,0.01}$ = 6.635；显然，8.249>6.635），P<0.01，即两个分布拟合效果之间差异有统计学意义，故采用 −2log$L$ 值最小的分布，即广义 Gamma 分布（−2log$L$ 值为 77.650）。广义 Gamma 分布回归模型输出结果中，只有 age1、MBP、CHD、smk 和 ECG 五个自变量有统计学意义。故仅保留这五个自变量，重新采用广义 Gamma 分布回归模型来进行分析。结果如下：

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr>ChiSq
Intercept		1	5.1270	1.8332	1.5341	8.7199	7.82	0.0052
age1		1	−0.0408	0.0123	−0.0648	−0.0168	11.09	0.0009
MBP		1	−0.0124	0.0075	−0.0270	0.0022	2.76	0.0966
CHD	0	1	−0.4497	0.2648	−0.9686	0.0693	2.88	0.0894
CHD	1	0	0	−	−	−	−	−
smk	0	1	0.4048	0.1725	0.0667	0.7430	5.51	0.0189
smk	1	1	0.2238	0.1949	−0.1582	0.6057	1.32	0.2509

smk	2	0	0	–	–	–	–	–
ECG	1	1	1.0879	0.2881	0.5232	1.6525	14.26	0.0002
ECG	2	1	0.1543	0.4940	−0.8139	1.1224	0.10	0.7548
ECG	3	0	0	–	–	–	–	–
Scale		1	0.4384	0.4092	0.0704	2.7310	–	–
Shape		1	−2.7990	4.7767	−12.1613	6.5632	–	–

广义 Gamma 分布回归模型的分析结果表明，随访开始时年龄的回归系数为负值（−0.0408），说明随访年龄越大，生存时间越短，死亡风险越大；smk 变量 0 水平与 2 水平比较时，对应的 P 值小于 0.05，其回归系数为正值（0.4048），说明不吸烟者生存时间长于吸烟者；ECG 变量 1 水平与 3 水平比较时，对应的 P 值小于 0.001，其回归系数为正值（1.0879），说明心电图正常者生存时间长于心电图异常者。

5 讨论与小结

5.1 讨论

在现有的统计软件中，进行生存资料参数回归模型建模时尚不能筛选自变量，只能依据计算结果，采用手工方法删除没有统计学意义的自变量，这在一定程度上影响了最优回归模型的产生。当自变量中含有两个以上定量自变量时，根据统计分析的经验，尽可能产生出一些派生自变量（例如平方项、交叉乘积项等）参与构建回归模型，可能有助于找到拟合优度更好的回归模型。生存资料参数回归模型的建模策略与 logistic 回归模型建模策略类似，对最后的结果而言，保留在回归模型中的自变量，既要考虑其应具有统计学意义也要考虑其实际意义，即回归系数正负号的含义在专业上是可以得到合理解释的。

5.2 小结

本文通过一个实例，比较详细地介绍了基于 SAS 软件实现参数回归模型的拟合方法；通过图示法大致判断生存时间可能服从何种分布类型；最后，还针对所拟合的两个参数回归模型，进行了拟合优度检验，从而可以确定最优回归模型。需注意的是，当基于图示法确定生存时间符合某种参数分布且其对应的模型属于嵌套模型（特指其参数取某些特定值时，原先复杂的模型就退化成一个相对简单的模型，例如，当威布尔分布模型中的形状参数 $\gamma = 1$ 时，它就退化成指数分布模型了）时，则需要采用似然比检验确定拟合效果最优的模型；否则，需要根据所拟合的两个模型各自的参数数目和其他拟合统计量的数值，综合考虑后再选定相对较优的模型。

参考文献

[1] Lee ET, Wang JW. Statistical methods for survival data analysis [M]. Hoboken: John Wiley & Sons, 2003: 72-76.

[2] 胡良平. 回归建模的基础与要领（Ⅲ）——变量状态与相互间关系[J]. 四川精神卫生, 2018, 31(6): 498-502.

[3] Raqab MZ, Al-Awadhi SA, Kundu D. Discriminating among Weibull, log-normal, and log-logistic distributions[J]. Commun Stat Simul Comput, 2018, 47(5): 1397-1419.

[4] Kleinbaum DG, Klein M. Survival analysis: a self-learning text [M]. 3rd Edition. New York: Springer Science Business Media, 2012: 289-351.

[5] 胡良平. SAS 常用统计分析教程[M]. 2 版. 北京: 电子工业出版社, 2015: 533-537.

（收稿日期：2020-01-06）

（本文编辑：陈 霞）