生存资料回归模型分析——基于MCMC过程构建生存资料Cox非比例风险回归模型

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流行病与卫生统计学教研室，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：李长平，E-mail：1067181059@qq.com）

【摘要】 本文目的是介绍采用PHREG过程及MCMC过程且基于贝叶斯统计思想分别构建Cox非比例风险回归模型的相关内容及其SAS软件实现。在MCMC过程中，有两种构建模型的方法：一是对观测值进行转置之后，在MODEL语句中使用GENERAL函数；二是不对观测值进行转置，使用MCMC过程中的JOINTMODEL选项。两个过程所得计算结果基本一致。

【关键词】 贝叶斯；生存分析；非比例风险回归模型；马尔科夫蒙特卡洛

中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20200312006

Analysis of regression model of survival data——Cox's non-proportional hazards regression model of survival data based on MCMC procedure

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

(1. *Department of Epidemiology and Health Statistics, 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: Li Changping, E-mail: 1067181059@qq.com*)

【**Abstract**】 This article mainly introduced the related contents of constructing Cox's non-proportional hazards regression model using PHREG procedure and MCMC procedure based on Bayesian theory, and its SAS software implementation. In the MCMC procedure, there were two methods to construct the model, one is to use the GENERAL function in the MODEL statement after transposing the observations, the other is to use the JOINTMODEL option in the MCMC process without transposing the observations. The results were basically consistent.

【**Keywords**】 Bayesian; Survival analysis; Non-proportional hazards regression model; Markov chain Monte Carlo

MCMC过程可应用于基于贝叶斯统计思想的生存资料的Cox比例风险回归模型中，该方法的使用应基于数据满足比例风险假定的前提上[1]。但在实际研究当中，往往会遇到生存资料不满足该假定的情况，此时，可借助SAS软件提供的多种过程步实现相关回归模型的构建。本文将通过MCMC过程对效应随时间变化的时间依赖型变量（即依时协变量）[2]构建基于贝叶斯统计思想的生存资料Cox非比例风险回归模型，并对相关的主要内容加以说明。

1 含依时协变量的扩展Cox模型

经典Cox比例风险回归模型见式（1）：

$$h(t|Z)=h_0(t)\exp(\beta'Z)=h_0(t)\exp\left(\sum\beta_jZ_j\right) \quad (1)$$

β_j是第j个协变量的回归系数，$h_0(t)$是基线风险函数。该模型有两个重要特征：①基线风险函数依赖于t，但是与协变量无关且未知；②风险率依赖于协变量，但与时间t无关。然而，在实际的生存资料中，某些协变量Z_j值会随着时间而变化，即不满足“比例风险假设”，如患者的心率。此时，我们可以在传统的Cox比例风险回归模型中加入该变量与时间的交互项，以描述其对基线风险函数的影响。带有依时协变量的Cox回归模型见式（2）：

$$h[t|Z(t)]=h_0(t)\,exp[\beta'Z(t)] \quad (2)$$

需使用局部似然法来估计β，计算公式见式（3）：

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）

$$L(\beta)=\prod_{i=1}^{n}\left\{\frac{\exp\left[\beta' Z_i(x_i)\right]}{\sum_{j\in R(x_i)} exp\left[\beta' z_j(x_i)\right]}\right\}^{\delta_i} \quad (3)$$

$Z(t)$是时间与协变量的交互项。此时，关于*HR*（风险率）的统计推断与经典的Cox比例风险回归模型类似，唯一的不同是风险率的主要部分$\exp[\beta' Z(t)]$会随着时间而变化，这就是“Cox非比例风险回归模型”。

2 构建Cox非比例风险回归模型

2.1 实例与数据

利用多发性骨髓瘤研究[3]的数据创建数据集Myeloma，所包含变量及解释此处不再赘述，本文所涉及变量见表1。为了简化计算，本例仅将表1中前三个定量自变量纳入考虑，即将它们视为“依时协变量”。

表1 变量赋值表

变量	变量名	赋值
血尿素氮水平	LogBUN	具体数值
血红蛋白水平	HGB	具体数值
血小板水平	Platelet	0=异常，1=正常
生存时间（月）	Time	具体数值
生存状态	Vstatus	0=存活，1=死亡

2.2 采用“PHREG过程”且基于贝叶斯统计思想构建Cox非比例风险回归模型

可以利用PHREG过程中的BAYES语句拟合Cox非比例风险回归模型。

SAS程序如下：

```
proc phreg data=Myeloma;
model Time*VStatus (0) =LogBUN z2 HGB z3 Platelet z4;
z2 = Time*LogBUN;
z3 = Time*HGB;
z4 = Time*Platelet;
bayes seed=1 nmc=10000;
run;
```

【程序说明】MODEL语句等号左边定义生存时间和生存结局变量（括号内为截尾数据标识），右边为各协变量（即自变量）以及时间与各协变量的交互项。BAYES语句指定回归模型使用贝叶斯分析，并设定随机数生成器种子seed=1；为了使初始值对后验推断的影响最小化，需要在Markov Chain达到目标分布后弃掉先前的部分样本，因此，nmc用于设定弃掉先前的部分样本后的迭代次数=10000。

【主要输出结果及解释】

后验汇总和区间

参数	N	均值	标准差	95% HPD 区间	
LogBUN	10000	3.2266	0.8161	1.6450	4.7742
z2	10000	−0.1395	0.0473	−0.2301	−0.0450
HGB	10000	−0.0389	0.0994	−0.2275	0.1605
z3	10000	−0.00406	0.00356	−0.0110	0.00296
Platelet	10000	0.3647	0.7499	−1.0269	1.8955
z4	10000	−0.0411	0.0370	−0.1143	0.0309

这是输出结果的“第1部分”，其中LogBUN、HGB、Platelet分别为样本中的三个协变量，z2、z3、z4分别为各协变量与时间交互项。表中第1列“参数”是拟创建的回归模型中的“自变量”；第2列指随机重复抽样一万次；第3列“均值”是各自变量的回归系数的估计值，而且，其中每个估计值都是一万次随机重复抽样计算所得到的算术平均值；第4列为与“各均值”对应的“标准差”；最后两列为与“各均值”对应的95% HPD（highest posterior density，HPD）区间，即95%最高后验密度置信区间。因此，根据此置信区间是否包含“0”（包含0时表明该变量对结果的影响无统计学意义），可得以下回归方程：

$$h(t)=h_0(t)\,exp(3.2266\times \text{LogBUN}-0.1395\times \text{z2})$$

图1显示LogBUN回归系数的均值在3.0左右波动，随着迭代次数的增加，摆动的幅度基本保持不变，所以有理由认为Markov Chain已经收敛。图2显示变量LogBUN回归系数不存在自相关，后验样本独立。图3显示后验密度核密度图成钟形，左右两边近似对称。因篇幅所限，其他协变量的相关结果从略。

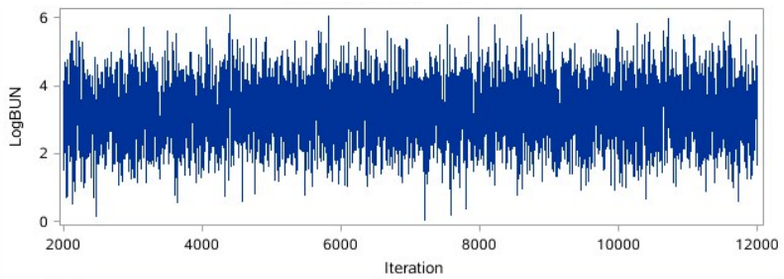

图 1 变量 LogBUN 回归参数的马尔可夫链迭代轨迹图

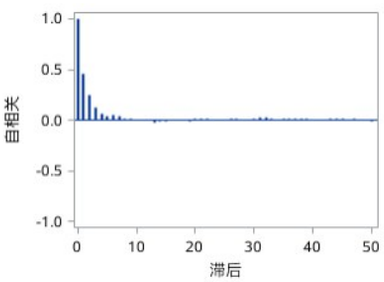

图 2 变量 LogBUN 回归参数的自相关函数图

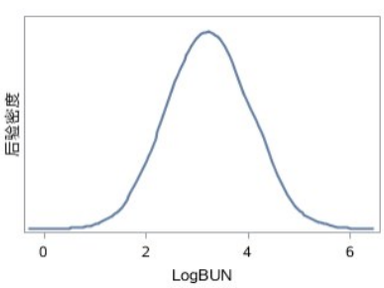

图 3 变量 LogBUN 回归参数的后验密度核密度图

2.3 采用“MCMC 过程”且基于贝叶斯统计思想构建 Cox 非比例风险回归模型

2.3.1 对观测值进行转置处理

【说明】因为 $Z_j(t_i)$ 取决于 t_i，所以每个时期之和 $S=\sum_{l\in R_i} exp\left[\beta' Z_l\left(t_i\right)\right]$ 是当前时间 t_i 和风险集中所有观测结合的结果。风险集 R_i 包含所有生存时间大于等于 t_i 的观测。修改输入的数据集，以使每一行不仅包含当前观察值，而且还包含相应风险集中的所有观察值。这样，当为每个观测值构建对数似然函数时，将得到所有相关数据。为此，首先需创建在不同行具有不同风险集的新数据集[4]，添加变量 stop，此变量是指当前观察值风险集中的观察值数量。其余变量是整个数据集中模型协变量的转置值。由于篇幅所限，SAS 程序从略。

【主要输出结果及解释】

Posterior Summaries and Intervals

Parameter	N	Mean	Standard Deviation	95% HPD Interval	
beta1	50000	3.2397	0.8226	1.6664	4.8752
beta2	50000	−0.1411	0.0471	−0.2294	−0.0458
beta3	50000	−0.0369	0.1017	−0.2272	0.1685
beta4	50000	−0.00409	0.00360	−0.0112	0.00264
beta5	50000	0.3548	0.7359	−1.0394	1.8100
beta6	50000	−0.0417	0.0359	−0.1122	0.0269

此处输出结果（第 3~6 列）与前面输出结果的第 1 部分（第 3~6 列）基本相同，各列的含义相同，此处从略。其中 beta1~beta6 分别对应各协变量及其与时间的交互项，即前面输出结果中第 1 部分的“第 1 列”。因篇幅所限，马尔可夫链迭代轨迹图等从略。

2.3.2 对观测值不进行转置处理

对不独立的数据建模，如果不想对每个观测的数据进行转置处理，还可以使用 JOINTMODEL 选项进行分析。因 SAS 程序过于复杂，此处从略。

3 讨论与小结

3.1 讨论

当生存资料经比例风险假设检验结果判定为不满足 PH 假定，即数据含有依时协变量时，研究者可以选用非比例风险回归模型进行统计分析，进而得到协变量随时间变化的趋势性信息。

PHREG 过程可以用来构建 Cox 非比例风险回归模型，BAYES 语句的应用则可实现回归参数的贝叶斯估计。MCMC 过程则是假设输入的观测值是独

立的，并且联合对数似然函数是各个对数似然函数的总和。因为此过程统计推断的原理是基于贝叶斯统计思想，所以在使用时要为数据指定似然函数，并为参数指定先验分布。如果观测值彼此不独立，则该求和将产生错误的对数似然值。因此，对不满足比例风险假定的数据建模，在MCMC过程中有两种构建模型的方法：一是对观测值进行转置后，在MODEL语句中使用GENERAL函数，此函数可以运行不需要响应变量的蒙特卡洛模拟；二是不对观测值进行转置，则可以使用MCMC过程中的JOINTMODEL选项，其基本思想是将所有必需的数据集变量存储在数组中，并且仅使用数组来构造整个数据集的对数似然函数。此时，MCMC过程将不再在模拟过程中逐步历遍输入数据，所以不再利用数据集变量构造对数似然函数，而是将数据集存储在数组中，并用数组而不是数据集变量来计算对数似然值。

从结果来看，PHREG和MCMC过程中的三种建模方法均可实现基于贝叶斯统计思想构建生存资料Cox非比例风险回归模型，并且所得结果差别不大，只是结果列出形式稍有不同；但从SAS程序上来看，使用MCMC过程所需要编写的SAS程序冗长、复杂，而在PHREG过程中的“bayes语句”（其作用相当于一个很复杂的子程序）大大简化了用户的编程过程。

3.2 小结

对于不满足比例风险假设的生存资料，可以选择SAS软件中的PHREG过程或MCMC过程构建基于贝叶斯统计思想的Cox非比例风险回归模型，前者利用BAYES语句实现贝叶斯分析，后者则根据是否对观测值进行转置处理，提供两种方法实现贝叶斯分析。三种方法建模所得结果基本一致，但基于PHREG过程来实现，SAS程序简单得多。

参考文献

[1] 姚婷婷，刘媛媛，李长平，等. 生存资料回归模型分析——生存资料Cox比例风险回归模型分析[J]. 四川精神卫生，2020，33(1)：27-32.

[2] Zhao J，Zhao Y，Lee AH，et al. A Time-varying covariate approach for survival analysis of paediatricoutcomes[J]. Paediatr Perinat Epidemiol，2017，31(6)：598-602.

[3] Krall JM，Uthoff VA，Harley JB. A step-up procedure for selecting variables associated with survival [J]. Biometrics，1975，31(1)：49-57.

[4] SAS Institute Inc. SAS/STAT 15.1 user's guide[M]. Cary，NC：SAS Institute Inc，2018：6251-6252.

（收稿日期：2020-03-12）

（本文编辑：陈 霞）