生存资料回归模型分析——基于MCMC过程构建生存资料Cox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：胡良平，E-mail：lphu927@163.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：一是在MODEL语句中使用LAG函数；二是使用MCMC过程中的JOINTMODEL选项。两个过程所得计算结果基本一致，而PHREG过程的程序相对简洁。

【关键词】 贝叶斯；生存分析；Cox比例风险回归分析；马尔科夫蒙特卡洛
中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20200312005

Analysis of regression model of survival data——construction of Cox's 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: Hu Liangping, E-mail: lphu927@163.com*)

【Abstract】 This article mainly introduced the related contents of constructing Cox's proportional hazards regression model based on Bayesian theory using the PHREG procedure and MCMC procedure, and its SAS software implementation. In the MCMC procedure, there are two ways to build the model, one is to use the LAG function in the MODEL statement, the other is to use the JOINTMODEL option in the MCMC procedure. The results obtained by the two procedures were basically the same, and the programs of the PHREG procedure are relatively simple.

【Keywords】 Bayesian; Survival analysis; Cox's proportional hazards regression analysis; Markov chain Monte Carlo

频率学派和贝叶斯学派是统计学发展历史上两个重要的学派[1]。通常，前者认为随机事件的概率是客观存在并假设固定不变的；而后者则认为此“概率”是随机的而不是固定不变的，并服从于某种分布。也即，经典统计学认为“参数”是固定不变的“常量”，而贝叶斯统计学认为“参数”是随机变量，这是两者的根本分歧所在。1763年由Richard Price整理发表了贝叶斯的成果《An Essay towards solving a Problem in the Doctrine of Chances》提出了贝叶斯公式，并介绍了贝叶斯思想，其核心内容就是对参数的估计并不是单纯取决于客观数据，而是取决于客观数据（包括总体和样本信息）和先验信息的共同作用[2]。随着计算机技术的发展和贝叶斯方法的改进，特别是马尔科夫蒙特卡洛（Markov chain Monte Carlo，MCMC）方法的提出和应用，使得参数后验分布的模拟得以更方便地实现，从而体现出该法在处理小样本数据时的优势[3]。现在，越来越多的新的统计分析方法将经典统计分析方法和贝叶斯思想有机地结合起来，例如，基于贝叶斯理论和生存分析相结合的贝叶斯生存分析在近年来越来越多地被应用于不同的研究领域，尤其是医学科学研究中[4-5]。因此，本文将介绍基于PHREG过程和MCMC过程分别构建贝叶斯统计思想框架下生存资料的Cox比例风险回归模型的相关内容。

1 基于“贝叶斯统计思想”构建Cox比例风险回归模型

姚婷婷等[6]的文章已经介绍了Cox比例风险回归模型，见式（1）：

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）

$$h(t)=h_0(t)exp\left(\beta_1X_1+\beta_2X_2+\cdots+\beta_PX_P\right) \quad (1)$$

式(1)中，X_1、X_2、…、X_p为与生存时间可能有关的自变量(即影响因素)；$h(t)$为具有自变量X_1、X_2、…、X_p的个体在t时刻的风险函数；$h_0(t)$为所有自变量为0时t时刻的基准风险函数；β_1、β_2、…、β_p为各自变量的偏回归系数。偏回归系数的估计需借助偏似然函数，用最大似然估计方法得到。偏似然函数的计算公式见式(2)：

$$L=q_1q_2\cdots q_i\cdots q_k=\prod_{i=1}^{k}q_i=\prod_{i=1}^{k}\frac{exp\left(\beta_1X_{i1}+\beta_2X_{i2}+\cdots+\beta_PX_{iP}\right)}{\sum_{S\in R(t_i)}exp\left(\beta_1X_{s1}+\beta_2X_{s2}+\cdots+\beta_PX_{sP}\right)} \quad (2)$$

式(2)中，q_i为第i个死亡时点的条件死亡概率，其分子部分为第i个个体在$t_i(t_1\leqslant t_2\leqslant\cdots\leqslant t_i\leqslant\cdots\leqslant t_k)$死亡时点的风险函数$h(t_i)$，分母部分为生存时间$T\geqslant t_i$的所有个体(包括死亡和删失)的风险函数之和$\sum_{j=i}^{n}h_j(t)$。

基于贝叶斯统计思想构建生存资料回归模型，即在原模型的基础上利用贝叶斯方法的基本原理对回归参数进行估计的过程。所以，需要先对这些参数指定适当的先验分布，如果先验分布选择不合适，则会对结果产生影响。故文献[7]建议将Cox回归模型系数β_i的先验分布设定为正态分布，本研究也将按此进行先验分布的设置。

2 构建Cox比例风险回归模型

2.1 实例与数据

本文所用数据来自一项骨髓瘤研究，研究者用烷基化剂治疗65例患者，在随访期间内，死亡48例，存活17例。变量赋值情况见表1。

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
年龄(岁)	Age	具体数值
白细胞对数	LogWBC	具体数值
骨折	Frac	0=无，1=有
骨髓中浆细胞的对数百分比	LogPBM	具体数值
蛋白尿	Protein	具体数值
血钙	SCalc	具体数值
生存时间(月)	Time	具体数值
生存状态	Vstatus	0=存活，1=死亡

创建数据集：

```
data Myeloma;
input Time Vstatus LogBUN HGB Platelet Age LogWBC
Frac LogPBM Protein SCalc;
label Time='survival time' VStatus='0=alive 1=dead';
datalines;
1.25 1 2.2175 9.4 1 67 3.6628 1 1.9542 12 10
1.25 1 1.9395 12.0 1 38 3.9868 1 1.9542 20 18
2.00 1 1.5185 9.8 1 81 3.8751 1 2.0000 2 15
……
53.00 0 1.1139 12.0 1 66 3.6128 1 2.0000 1 11
57.00 0 1.2553 12.5 1 66 3.9685 0 1.9542 0 11
77.00 0 1.0792 14.0 1 60 3.6812 0 0.9542 0 12;
run;
```

【说明】完整数据来自文献[8]。因篇幅所限，此处未呈现全部数据。

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

可以利用PHREG过程的BAYES语句拟合Cox比例风险回归模型。

SAS程序如下：

```
proc phreg data=Myeloma;
model Time*VStatus(0)=LogBUN HGB Platelet Age LogWBC
Frac LogPBM Protein Scalc;
Bayes seed=1 nmc=10000 outpost=phout;
run;
```

【程序说明】MODEL语句是PHREG过程的必需语句，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)。使用BAYES语句则要求回归模型的贝叶斯分析是采用Gibbs抽样，同时设定seed为随机数生成器种子，设为1；NMC为退火(退火是指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)后的迭代次

数，设为10000；OUTPOST选项将后验分布样本保存在SAS数据集中以进行以后的处理。

【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	1.7610	0.6593	0.4107	2.9958
HGB	10000	−0.1279	0.0727	−0.2801	0.00599
Platelet	10000	−0.2179	0.5169	−1.1871	0.8341
Age	10000	−0.0130	0.0199	−0.0519	0.0251
LogWBC	10000	0.3150	0.7451	−1.1783	1.7483
Frac	10000	0.3766	0.4152	−0.4273	1.2021
LogPBM	10000	0.3792	0.4909	−0.5939	1.3241
Protein	10000	0.0102	0.0267	−0.0405	0.0637
SCalc	10000	0.1248	0.1062	−0.0846	0.3322

这是输出结果的“第1部分”，第1列“参数”实际上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”实际上是各自变量前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结果的算术平均值；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：

$$h(t) = h_0(t)\,exp\left(1.7610 \times \text{LogBUN}\right)$$

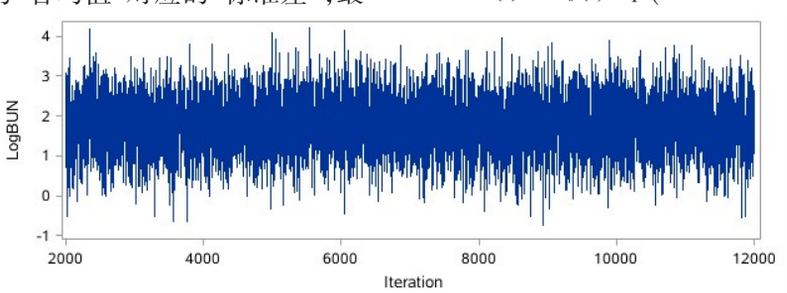

图1　变量LogBUN回归参数的马尔可夫链迭代轨迹图

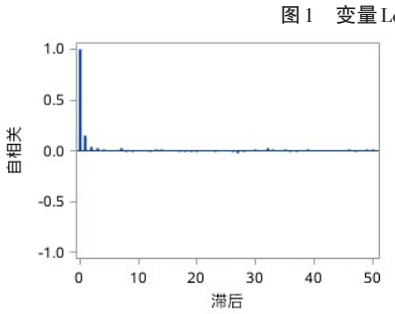

图2　变量LogBUN回归参数的自相关函数图

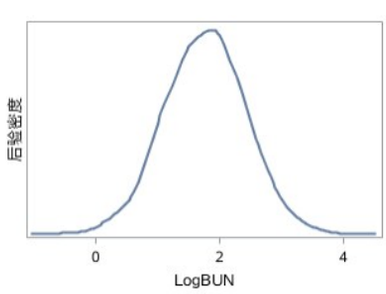

图3　变量LogBUN回归参数的后验密度核密度图

图1显示马尔可夫链已经收敛（因篇幅所限，其他协变量的相关结果此处从略）。

2.3　采用“MCMC过程”且基于贝叶斯统计思想构建Cox比例风险回归模型

2.3.1　利用LAG函数拟合模型

SAS程序如下：

```
proc mcmc data=myeloma Moutpost=outi nmc=
50000 ntu=3000 seed=1;
array beta;
parms beta: 0;
prior beta: ~ normal(0, var=1e6);
bZ = beta1 * LogBUN + beta2 * HGB + beta3 *
Platelet
+ beta4 * Age + beta5 *LogWBC + beta6 *Frac +
beta7 *LogPBM + beta8 * Protein + beta9 *
```

```
SCalc;
    if ind =1 then do; /* first observation */
    S =exp(bZ);
    l =vstatus *bZ;
    v =vstatus;
    end;
    else if (1<ind<&N) then do;
    if (lag1(time) ne time) then do;
    l =vstatus* bZ;
    l = l - v * log(S); /* correct the loglike value */
    v =vstatus; /* reset v count value */
    S = S +exp(bZ);
    end;
    else do; /* still a tie */
    l =vstatus* bZ;
    S = S + exp(bZ);
V = v + vstatus; /* add # of noncensored values */
    end;
    end;
    else do; /* last observation */
    if (lag1(time) ne time) then do;
    l = - v * log(S); /* correct the loglike value */
    S =S +exp(bZ);
    L = l + vstatus* (bZ - log(S));
    end;
    else do;
    S =S +exp(bZ);
    l =vstatus* bZ - (v + vstatus) * log(S);
    end;
    end;
    model general(1);
    run;
```

【程序说明】ARRAY 语句用于将回归系数的名称与协变量、常量相关联。PARMS 语句给出模型中的参数名称，并为其指定初始值。PRIOR 语句指定模型参数的先验分布为正态分布。程序中的 bZ 为似然函数中的回归项 $\beta' Z_j(t_i)$，S 为 $\sum_{l \in R_i} exp\left[\beta' Z_l(t_i)\right]$，即每个观测的风险集项。本例所用似然函数为 Breslow 似然函数。符号“l”为每个观测计算对数似然的公式。IF-ELSE 语句将所有的观测分成三部分，并使用 lag1 函数来检验两个相邻的生存时间 time 是否不同。符号“v”为生存状态（Vstatus）的总和，因为删失数据不进入似然公式的计算，所以需要将其去掉。MODEL 语句用于在给定参数的似然函数的情况下指定数据的条件分布。

【主要输出结果及解释】

Posterior Summaries and Intervals

Parameter	N	Mean	Standard Deviation	95% HPD Interval	
beta1	50000	1.7600	0.6441	0.5117	3.0465
beta2	50000	−0.1308	0.0720	−0.2746	0.00524
beta3	50000	−0.2017	0.5148	−1.2394	0.7984
beta4	50000	−0.0126	0.0194	−0.0512	0.0245
beta5	50000	0.3373	0.7256	−1.1124	1.7291
beta6	50000	0.3992	0.4337	−0.4385	1.2575
beta7	50000	0.3749	0.4861	−0.5423	1.3689
beta8	50000	0.0106	0.0271	−0.0451	0.0616
beta9	50000	0.1272	0.1064	−0.0763	0.3406

这里的输出结果（特指第 3 列到第 6 列）与前面输出结果的第 1 部分（特指第 3 列到第 6 列）是基本相同的，各列的含义相同，此处从略。

其中 beta1~beta9 分别对应各协变量，实际上就是前面输出结果中第 1 部分的“第 1 列”。

2.3.2 利用 JOINTMODEL 选项拟合模型

若利用 PROC 语句中的 JOINTMODEL 选项，则可使对数似然函数适用于整个数据集，而不只是单个的观察值。但在使用此选项前，还要为数据风险集 S 指定包含其中的观察的数量，为此需先创建一个 stop 变量。因篇幅所限，这部分的 SAS 程序和输出结果从略。

3 讨论与小结

3.1 讨论

以 Cox 比例风险回归模型为例，对于满足 PH 假

定的有删失数据的生存资料来说，该模型能够很好地利用偏似然（Partial Likelihood，PL）估计理论识别响应变量的影响因素，但其应用仍需一定的样本量。贝叶斯思想的提出则有效地弥补了小样本的不足。二者相结合的基于贝叶斯统计思想的Cox比例风险回归模型很好地融合了其各自优势。

本研究通过实例，基于贝叶斯统计思想并借助PHREG过程和MCMC过程分别构建了Cox比例风险回归模型，并且分别利用MCMC过程中的LAG函数、JOINTMODEL选项拟合模型。通过对程序及结果的解释比较，不难发现此三种方法均可得到后验样本统计描述指标（包括均值、标准差及95%最大后验密度置信区间）、有效样本大小、马尔可夫链迭代轨迹图等主要结果，而且后验样本的均值等指标在数值上相差不大，MCMC过程中两种方法的结果更是完全一致，只是在结果展示形式上稍有不同。由于PHREG过程本身是实现经典统计学中Cox比例风险模型回归分析的标准过程，这里只需加入BAYES语句用于指定进行贝叶斯估计。因此，通过比较不同过程的具体语句可以发现，PHREG过程相对MCMC过程要简洁，并且可以提供普通的Cox比例风险模型的结果——回归系数的最大似然估计的结果（包括回归系数估计值、标准误差及其95%置信区间），而MCMC过程只是提供了“平均意义”下的类似结果。

3.2 小结

在SAS中，可以借助PHREG过程或MCMC过程构建基于贝叶斯统计思想的Cox比例风险回归模型，两种做法的主要结果相似，但前者程序语句相对简洁。研究者可根据具体情况选择其一进行回归模型的构建。

参考文献

[1] 李晓松. 卫生统计学[M]. 北京：人民卫生出版社，2017：364-365.

[2] SAS Institute Inc. SAS/STAT 15.1 user's guide[M]. Cary，NC：SAS Institute Inc，2018：132.

[3] 林静. 基于MCMC的贝叶斯生存分析理论及其在可靠性评估中的应用[D]. 南京：南京理工大学，2008.

[4] Nasejje JB，Mwambi HG，Achia TN. Understanding the determinants of under-five child mortality in Uganda including the estimation of unobserved household and community effects using both frequentist and Bayesian survival analysis approaches[J]. BMC Public Health，2015，15：1003.

[5] Armero C，Cabras S，Castellanos ME，et al. Bayesian analysis of a disability model for lung cancer survival[J]. Stat Methods Med Res，2016，25(1)：336-351.

[6] 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.

[7] Spiegelhalter D，Thomas A，Best N，et al. Open BUGS user manual（version 3.2.3. Cambridge：MRC Biostatistics Unit[EB/OL]. http://www.openbugs.net/Manuals/Manual.html，2014.

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

（收稿日期：2020-03-12）

（本文编辑：陈 霞）