·科研方法专题·

生存资料回归模型分析——Cox比例风险假设的图形检验法

宋德胜[1],李长平[1,2],刘媛媛[1],崔 壮[1*],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:崔 壮,E-mail:cuizhuang@tmu.edu.cn)

【摘要】 本文目的是介绍目前使用图形检验比例风险的常用方法。经典的Cox比例风险回归模型要求生存资料满足比例风险假设,而在临床资料中,这个假设往往并不成立。鉴于此,本文首先阐述了比例风险假设的概念;然后介绍了一些检验比例风险假设是否成立的常用图示方法,主要包括Kaplan-Meier生存曲线图、ln[-ln(St)]生存时间关系图、缩放Schoenfeld残差与时间的关系图、SAS软件PHREG过程中ACCESS语句的PH和RESAMPLE选项产生的模拟路径图;最后,基于SAS软件并通过实例演示上述方法的实现。

【关键词】 生存分析;比例风险假设;SAS软件;Cox回归模型
中图分类号:R195.1 文献标识码:A doi:10.11886/scjsws20200312003

Analysis of regression model of survival data——the test of the Cox's proportional hazards assumption

Song Desheng[1], *Li Changping*[1,2], *Liu Yuanyuan*[1], *Cui Zhuang*[1*], *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: *Cui Zhuang*, *E-mail*: *cuizhuang@tmu.edu.cn*)

【Abstract】 The purpose of this study was to introduce the current common methods of examining the assumption of proportional hazards using graphs. The classic Cox proportional hazards model requested the assumption of constant proportional hazards, which was often not true in clinical data. Given that, this article firstly interpreted the concept of the assumption of proportional hazards. Then introduced some common methods to test whether the proportional hazards were constant, which including graphical methods (Kaplan-Meier survival curve, plot about ln[-ln(St)] against survival time, scaled Schoenfeld residual against time diagram, simulation path diagram generated by the PH and RESAMPLE options in the ACCESS statement of the PHREG procedure). Finally, the implementation of the above methods in SAS software was demonstrated through an example.

【Keywords】 Survival analysis; Proportional hazards assumption; SAS software; Cox regression model

在临床实践中,调查者会在指定的时间段内随访研究对象直到其发生预先指定的观察事件。然而,部分研究对象会在随访期间出于某种原因而退出研究,例如,出现非观察事件导致的死亡或者研究对象主动要求退出等。这些提前退出随访的情况被称之为截尾。这种截尾会导致收集的数据不完整。因此,传统的参数回归模型并不适用于处理生存资料。目前,生存分析中最常用的回归模型是英国统计学家D. R. Cox于1972年提出的Cox半参数回归模型。该模型不要求生存时间满足特定的概率分布,但要求生存资料满足比例风险假定。

1 比例风险概念

经典的Cox比例风险回归模型存在一个假设:不论基线风险如何,在基线以后的任何时间点上,分别在影响因素的"暴露水平"与"非暴露水平"条件下的发生事件的风险比是恒定的,换言之,所考

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)

察的影响因素对于所考察事件的效应不会随时间而改变。这就是比例风险恒定假设[1]，简称为PH假设。例如，年龄为50岁时，男性发生心脏病的风险是女性的2倍，那么60岁时，男性的风险仍然是女性的2倍（说明：此例中的“影响因素”为“性别”，其“暴露水平”为“男性”，“非暴露水平”为“女性”）。但有很多临床生存资料并不满足此假设，此时，这种变量（即影响因素）效应的风险称为非比例风险。这种效应随时间变化的变量称为时间依赖型变量，即时依协变量[2]。比例风险恒定更普遍的情形是：设X_i为第i位受试对象的自变量向量、X_j为第j位受试对象的自变量向量，PH假设为第i位受试对象与第j位受试对象风险之比仅与他们的自变量向量的取值之差呈比例关系，而与自变量向量在什么时间点取值无关。

2 使用图示法检验比例风险假定

判断比例风险假设是否成立的一个简单的方法是图示法，即通过观察每一个定性自变量各水平条件下的Kaplan-Meier生存曲线图是否存在交叉，如果存在交叉，则表示该定性自变量不满足比例风险假设。另外，对于特定的定性自变量的各水平组，绘制$\ln\{-\ln[S(t)]\}$与生存时间或生存时间的对数的关系图，如果线段明显不平行，说明该定性自变量不符合比例风险假设。对于连续型自变量，可使用Schoenfeld残差图、Score残差图进行判断，也可以将连续型自变量定性化，然后采取前述的图示法。

对于已经观测到的事件时间，假如已知第i个对象的第k个协变量及其取值，则Schoenfeld残差见式(1)[3]：

$$\hat{r}_{ik} = X_{ik} - \overline{X}_{w_i k} \quad (1)$$

式(1)中，X_{ik}是第i个对象第k个协变量的值，$\overline{X}_{w_i k}$是给定事件时间的风险集中协变量值的加权均数。若Schoenfeld残差值为正，表示在对应的死亡时间点，X的实际值高于预期值。绘制Schoenfeld残差与生存时间的广义线性回归图，若图形呈现非零斜率，表示该变量不满足比例风险假设。

Martingale残差定义见式(2)[4]：

$$\hat{M}_j = N_j(\infty) - \int_0^{\infty} Y_j(t)\exp[\hat{\beta}z_j(t)]d\widehat{H_0}(t) \quad (2)$$

式(2)中，$N_j(t)$代表t时刻个体j是否经历了某事件的指示变量，$Y_j(t)$是t时刻前个体j是否在观察中的指示变量，$\hat{\beta}$是回归系数向量，$z_j(t)$是t时刻，第j个个体的协变量向量，$\widehat{H_0}(t)$是累积基准风险函数的Breslow估计。因此，Martingale残差可能具有超额事件数，并且这些残差的总和等于0。在满足比例风险假设的前提下，如果利用该残差与时间作图，则可以观察到该残差随时间围绕一条水平线波动。

通过SAS的PHREG过程中ASSESS语句绘制累积得分残差与时间的图形以检验比例风险假设。图形中每条曲线的值开始于0且终止于0，这是一种Brownian过程。在比例风险的假设下，该选项产生若干随机路径。将随机产生的路径与实际数据相对比，若变量的实际路径在随机路径范围内，则表示该协变量服从比例风险假设。反之，则不服从比例风险假设。但是，非比例风险假设的形式则不清楚。

3 SAS软件实现

3.1 创建数据集

Krall等[5]对一项多发性骨髓瘤研究的数据进行了分析。65例患者接受了烷化剂治疗，其中48例在研究期间去世，17例存活。创建的数据集Myeloma包含变量如下：Time（预后生存时间）；Vstatus（患者状态，0表示存活，1表示死亡）；在诊断时被认为与生存时间有关的变量，如LogBUN（血尿素氮水平）；HGB（血红蛋白水平）；Platelet（血小板水平，0表示非正常，1表示正常）；Age（年龄）；LogWBC（白细胞水平的对数）；Frac（骨折，0表示未发生，1表示发生）；LogPBM（骨髓浆细胞对数百分比）；Protein（蛋白质水平）；Scalc（血钙水平）。

数据集创建程序如下：

```
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
……
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;

3.2　比例风险假设检验

3.2.1　Kaplan-Meier 生存曲线以及生存函数负对数的对数与时间的对数关系图

以下 SAS 语句使用图示法中的 Kaplan-Meier 生存曲线以及生存函数负对数的对数与时间的对数关系图判断定性变量是否符合比例风险假设。

```
ods html style=htmlbluecml image_dpi=300;
proc lifetest data=Myeloma
plots=(survival(cb=hw test atrisk(outside maxlen=13)) lls);
time Time * VStatus(0);
strata platelet;
run;
proc lifetest data= Myeloma
plots=(survival(cb=hw test atrisk(outside maxlen=13)) lls);
time Time * VStatus(0);
strata frac;
run;
```

【程序说明】ODS 语句指定后续 PROC LIFETEST 产生的图形使用 HTML BLUECML 的样式显示，图形的 DPI 设置为 300。PROC LIFETEST 语句中，PLOTS 选项的 survival 以及 lls 要求绘制 Kaplan-Meier 生存曲线与生存函数负对数的对数与时间的对数关系图。TIME 语句指定生存时间与截尾指示变量（0 表示截尾），STRATA 语句指定需要考察是否满足 PH 假定的定性变量。

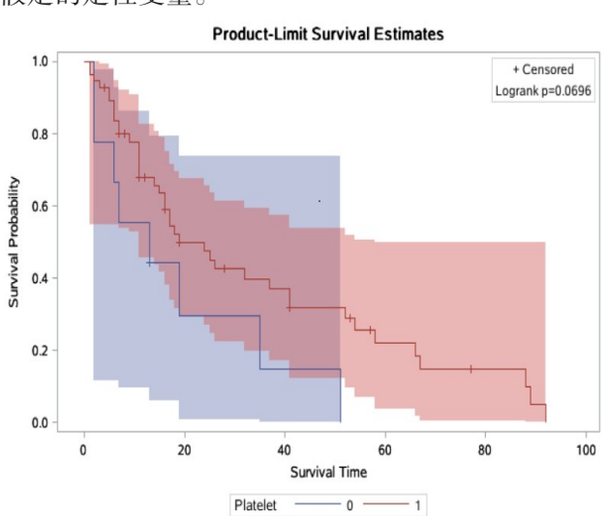

图 1　不同 PLATELET 水平下的生存曲线

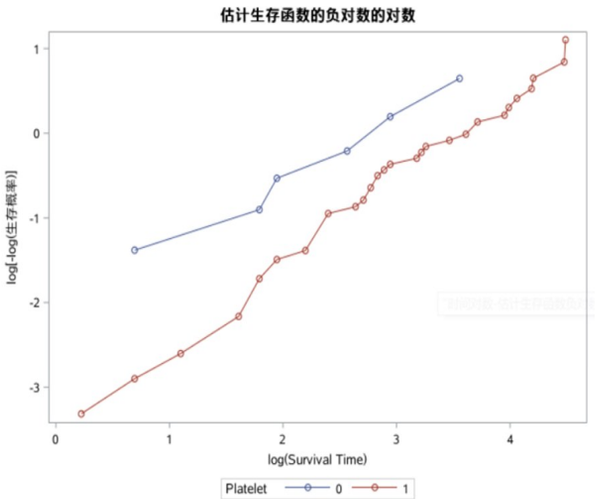

图 2　不同 PLATELET 水平下的 LOG[-LOG(生存函数)]与 LOG(time)的变化趋势

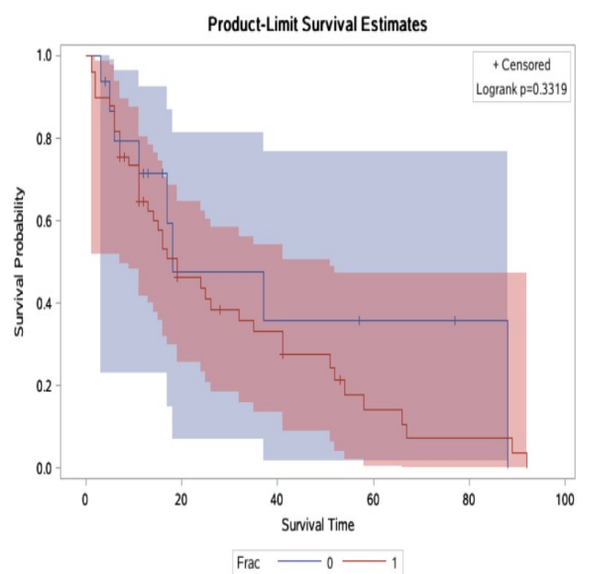

图 3　不同 FRAC 水平下的生存曲线

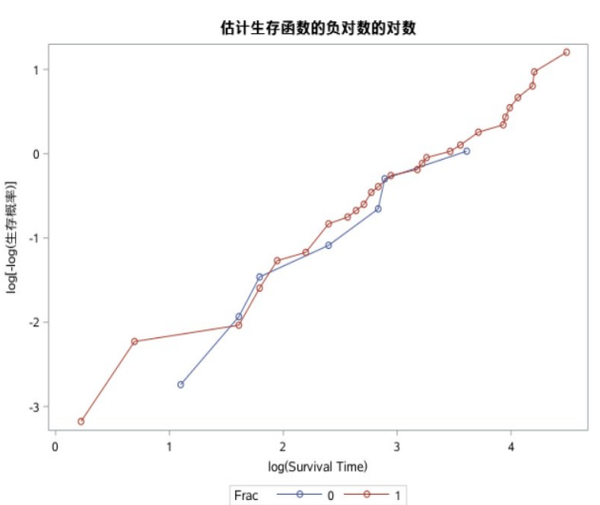

图 4　不同 FRAC 水平下的 LOG[-LOG(生存函数)]与 LOG(time)的变化趋势

图 1 显示的是不同 platelet 水平下的生存曲线；图 2 显示的是不同 platelet 水平下的 LOG[-LOG(生存函数)]与 LOG(time)之间的折线图；图 3 显示的是不同 FRAC 水平下的生存曲线；图 4 显示的是不同 FRAC 水平下的 LOG[-LOG(生存函数)]与 LOG(time)之间的折线图。图 1 中，两条生存曲线无交叉；图 2 中，两条线没有明显交叉的趋势，因此可认为变量 PLATELET 满足比例风险假设；图 3 中，两条生存曲线存在交叉情况；图 4 中，两条曲线亦存在相交，因此可以认为 FRAC 不满足比例风险假设。

3.2.2 Schoenfeld 残差与 log(time)的关系图

SAS 程序如下：

```
proc phreg data=myeloma zph (notest fit=spline transform=LOG);
Class platelet frac;
Model time*vstatus(0) =LogBUN HGB Platelet Age LogWBC Frac LogPBM protein scalc/selection=s;
run;
```

【程序说明】Proc phreg 调用 PHREG 过程进行分析，data 选项指定要分析的数据集是 myeloma，zph 选项(注意：此选项在 SAS 9.3 中无效)要求进行比例风险检验，括号中的 notest 表示不进行相关检验，fit 指定是否呈现光滑曲线拟合结果，本例中使用了 Spline，要求进行惩罚 B 样条曲线拟合。Class 语句指定分类变量为 platelet 和 frac。Model 语句进行模型构建。Time*vstatus(0)中的 time 表示生存时间，当 vstatus=0 时代表截尾。等号右边为一系列自变量。Selection=s 表示进行逐步选择筛选变量。

输出结果见表 1、图 5、图 6。

表 1 基于最大似然法估计回归参数

最大似然估计分析						
参数	自由度	参数估计	标准误差	卡方	Pr>卡方	危险率
LogBUN	1	1.67440	0.61209	7.4833	0.0062	5.336
HGB	1	-0.11899	0.05751	4.2811	0.0385	0.888

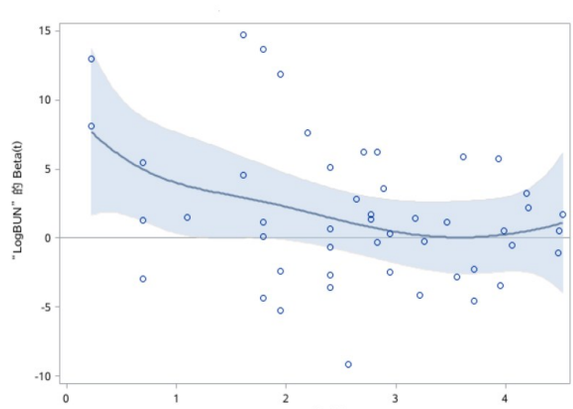

图 5 变量 logBun 的缩放 Schoenfeld 残差与时间对数的关系图

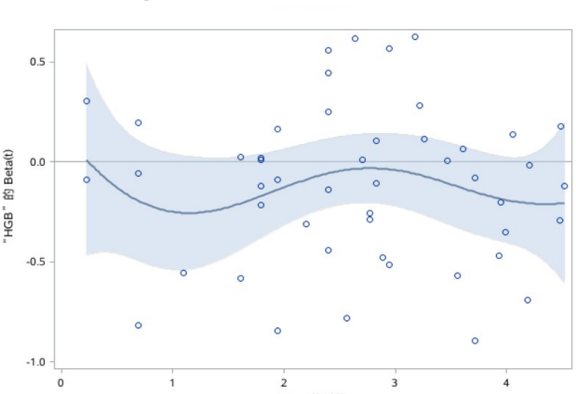

图 6 变量 HGB 的缩放 Schoenfeld 残差与时间对数的关系图

【结果解释】最大似然估计结果显示，逐步回归筛选后，模型中剩余的变量为 logBun 和 HGB。从这两个自变量的缩放 Schoenfeld 残差与时间对数的关系图可以看出，logBun 拟合的曲线斜率明显不为 0，而 HGB 拟合的曲线斜率基本为 0。因此，LogBun 不符合比例风险假设，HGB 符合比例风险假设。

3.2.3 使用 PHREG 过程 ASSESS 语句判断比例风险假设是否成立

SAS 程序如下：

```
proc phreg data=Myeloma;
class platelet frac;
model time*vstatus (0) =LogBUN HGB Platelet Age LogWBC Frac LogPBM Protein Scalc/selelction=s;
assess ph /seed=1234 resample=1000;
run;
```

【程序说明】ASSESS 语句要求进行 Cox 回归模型的充分性检验。通过这个语句，可检验一个或多个协变量的函数形式。PH 选项要求进行比例风险假设。Seed 选项设置了种子数，以保证结果的重现性。对于用户指定的每个协变量[通过 VAR=(变量列表)指定]，ASSESS 语句会绘制已观测的累积 Martingale 残差与解释变量值的关系图，并模拟若干残差图形(通过 NAPTHS=n 指定)。

输出结果见图 7、图 8。

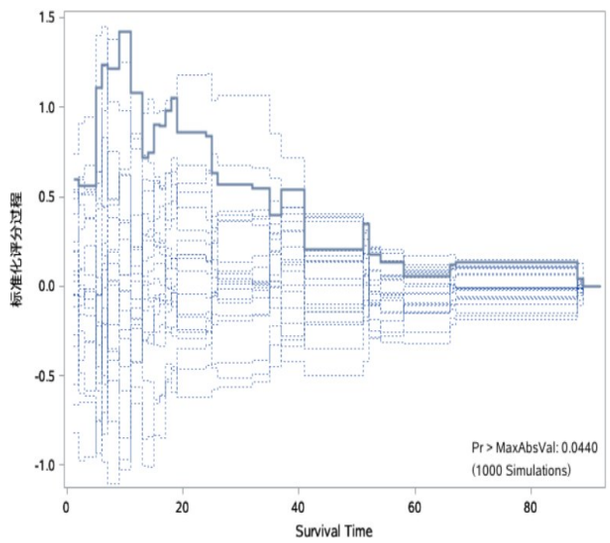

图7　变量LogBun的模拟路径图

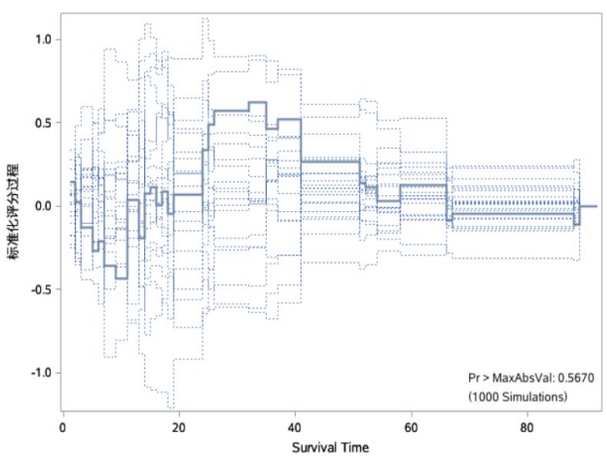

图8　变量HGB的模拟路径图

【结果解释】LogBun变量的实际路径在模拟路径外，而HGB变量的实际路径均在模拟路径范围内；模拟路径图右下角Kolmogorov-type supremum检验结果显示LogBun的$P<0.05$。因此，LogBun不满足比例风险假设；而HGB变量满足比例风险假设。

4　小　结

Cox比例风险回归模型常用于分析生存数据，需满足比例风险假设，但这在实际的生存资料中往往不能满足。因此，本文介绍了在临床试验中简便的比例风险假设的检验方法，即图示法。Kaplan-Meier生存曲线图和LLS生存函数负对数的对数与时间对数的关系图是最常用的用于直观判断分类变量是否满足比例风险假设的工具；对于定量变量，通过观察Schoenfeld残差与时间函数的关系图，可大致判断定量变量是否满足比例风险假设。另外，本文也使用了SAS程序PHREG过程ASSESS语句中的PH选项和RESAMPLE选项检验比例风险假设是否成立。

参考文献

[1] Trinquart L, Jacot J, Conner SC, et al. Comparison of treatment effects measured by the Hazard Ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials[J]. J Clin Oncol, 2016, 34(15): 1813-1819.

[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] 余红梅，何大卫. 检查Cox模型比例风险假定的几种图示法[J]. 中国卫生统计，2000，17(4)：215-218.

[4] 余红梅，何大卫，徐勇勇. 鞅残差在Cox回归模型诊断中的应用[J]. 现代预防医学，2001，28(1)：10-11.

[5] 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：吴俊林）

科研方法专题策划人——胡良平教授简介

胡良平，男，1955年8月出生，教授，博士生导师，曾任军事医学科学院研究生部医学统计学教研室主任和生物医学统计学咨询中心主任、国际一般系统论研究会中国分会概率统计系统专业理事会常务理事、中国生物医学统计学会副会长、北京大学口腔医学院客座教授和《中华医学杂志》等10余种杂志编委；现任世界中医药学会联合会临床科研统计学专业委员会会长、国家食品药品监督管理局评审专家和3种医学杂志编委；主编统计学专著48部、参编统计学专著10部；发表第一作者和通信作者学术论文300余篇、发表合作论文130余篇；获军队科技成果和省部级科技成果多项；参加并完成三项国家标准的撰写工作、参加三项国家科技重大专项课题研究工作。在从事统计学工作的30年中，为几千名研究生、医学科研人员、临床医生和杂志编辑讲授生物医学统计学，在全国各地作统计学学术报告100余场，举办数十期全国统计学培训班，培养20多名统计学专业硕士和博士研究生。近几年来，参加国家级新药和医疗器械项目评审数十项、参加100多项全军重大重点课题的统计学检查工作。归纳并提炼出有利于透过现象看本质的“八性”和“八思维”的统计学思想，独创了逆向统计学教学法和三型理论。擅长于科研课题的研究设计、复杂科研资料的统计分析和SAS与R软件实现、各种层次的统计学教学培训和咨询工作。