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

姚婷婷[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：李长平，E-mail：1067181059@qq.com）

【摘要】 本文目的是介绍生存资料Cox比例风险回归模型分析的概念、作用及使用SAS软件实现计算的方法。首先介绍相关基本概念，包括“Cox比例风险回归模型简介”“模型假定及其检验”“参数解释”和“参数估计与假设检验”；然后通过一个实例并基于SAS软件演示如何实施生存资料Cox比例风险回归模型分析，内容包括“产生SAS数据集”“绘制生存曲线图”“判断PH假定是否成立”和“算出参数估计值与假设检验结果”。结果表明：当生存资料满足PH假定时，Cox比例风险回归模型可用于生存资料影响因素分析、校正混杂因素后的组间比较以及对每个个体进行预后指数和生存率的预测。

【关键词】 PH假定；生存率曲线；回归分析；Cox比例风险回归模型；生存预测

中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20200106003

Analysis of regression model of survival data——analysis of Cox's proportional hazards regression model of survival data

Yao Tingting[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*: *Li Changping*, *E-mail*: 1067181059@qq.com)

【Abstract】 The purpose of this paper was to introduce the concepts and functions and the calculation methods of the Cox's proportional hazards regression model analysis of survival data by using the SAS software. Firstly, the basic concepts of the regression analysis was introduced, including "introduction to Cox's proportional hazards regression model" "model assumption and tests" "parameter interpretation" and "parameter estimation and hypothesis testing", and then the Cox's proportional hazards regression model analysis was demonstrated through one example by using the SAS software, including "generating SAS data set" "drawing survival curve" "diagnosing whether PH hypothesis to be true" and "calculating parameter estimates and the results of hypothesis testing". The results showed that it could be used for the analysis of influencing factors of survival data, inter-group comparison after correction of confounding factors, and the prediction of prognostic index and survival rate for each individual in the survival data set which met the PH assumption by applying the Cox's proportional hazards regression model.

【Keywords】 PH assumption; Survival curve; Regression analysis; Cox's proportional hazards regression model; Survival prediction

在对生存资料进行分析时，若同时分析众多因素对生存结局和生存时间的影响，需要采用多因素分析方法，而传统的多因素分析方法并不适用，不能同时处理生存结局和生存时间，也不能充分利用删失时间所提供的不完全信息。适用于生存数据的多因素生存分析方法主要有参数回归模型和半参数回归模型两类，参数法需要以特定的分布为基础建立回归模型，应用有其局限性，而半参数法的假定相对较少或较弱，特别是Cox比例风险回归模型（Cox's proportional hazards regression model），不要求生存资料满足特定的分布类型，是目前对生存资料进行多因素分析最常用的方法之一。

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）

1 概　　述

1.1 Cox 比例风险回归模型简介

Cox 比例风险回归模型见式(1)：

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

在式(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)，其中的自变量或影响因素可能是定量的或定性的，在整个观察期内不随时间的变化而变化；$h(t)$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 时刻的风险函数，称为基准风险函数(baseline hazard function)，是未知的；β_1、β_2、…、β_p为各自变量的偏回归系数，是一组未知的参数，需要根据实际的数据来估计。

Cox 模型不直接考察生存函数 $S(t)$ 与自变量的关系，而是利用生存函数 $S(t)$ 与风险函数 $h(t)$ 的关系，将风险函数 $h(t)$ 作为因变量，间接反映自变量与生存函数 $S(t)$ 的关系。该模型右侧可分为两个部分：一部分为 $h_0(t)$，它没有明确的定义，分布无明确的假定，为非参数部分；另一部分是以 p 个自变量的线性组合为指数的指数函数，具有参数模型形式，其中回归系数反映自变量的效应，可通过样本实际观测值来估计。所以 Cox 比例风险回归模型实为半参数模型(semi-parametric model)[1-6]。

1.2 模型假定及其检验

由 Cox 比例风险回归模型可知，任意两个个体风险函数之比，即风险比(hazard ratio，HR)为：

$$HR=\frac{h_i(t)}{h_j(t)}=\frac{h_0(t)exp\left(\beta_1X_{i1}+\beta_2X_{i2}+\cdots+\beta_PX_{iP}\right)}{h_0(t)exp\left(\beta_1X_{j1}+\beta_2X_{j2}+\cdots+\beta_PX_{jP}\right)}$$

$$=\exp\left[\beta_1\left(X_{i1}-X_{j1}\right)+\beta_2\left(X_{i2}-X_{j2}\right)+\cdots+\beta_p\left(X_{ip}-X_{jp}\right)\right]$$

$$i,j=1,2,\cdots,n \quad (2)$$

该比值与 $h_0(t)$ 无关，也与时间 t 无关，即模型中自变量的效应不随时间的改变而改变，具有某种特定预后因素向量的患者的死亡风险与具有另一种特定预后因素向量的患者的死亡风险在所有时间点上都保持一个恒定的比例，这种情形被称为比例风险(proportional hazard)假定，简称 PH 假定。

PH 假定的检验方法包括图示法和检验法。图示法是通过观察散点图中散点的分布或趋势来判断生存时间是否满足或近似满足 PH 假定，主要有 COX-KM 生存曲线、基于累计风险函数的图示法、Schoenfeld 残差图、Score 残差图；检验法是通过构造满足 PH 假设下服从某一已知分布的统计量，基于检验统计量的数值大小和对应的 P 值来判断是否满足或近似满足 PH 假定，主要有时间协变量法、线性相关检验、加权残差 Score 检验、三次样条函数法。因篇幅所限，在此仅介绍 COX-KM 生存曲线和基于累计风险函数的图示法，对其他方法感兴趣的读者可参阅文献[4-6]。

COX-KM 生存曲线法是观察按该变量(指拟考察的自变量)分组的 Kaplan-Meier 生存曲线，若生存曲线明显交叉，则不满足 PH 假定。基于累计风险函数的图示法是以生存时间 t 为横轴，对数生存率 $\ln\left[-ln\hat{S}(t)\right]$ 为纵轴，绘制分类协变量每一组别的生存曲线，若协变量各组别对应的曲线大致平行或等距，则满足 PH 假定。对于连续变量，可将该变量离散后，比较各组的 COX-KM 生存曲线或 $\ln\left[-ln\hat{S}(t)\right]$ 对生存时间 t 的图，也可以将连续变量与对数生存时间的交互作用项放入回归模型中，若交互作用项无统计学意义，则满足 PH 假定。若各协变量(指除时间 t 之外的自变量或影响因素)均满足或近似满足 PH 假定，可直接应用 Cox 比例风险回归模型。

1.3 参数解释

Cox 比例风险回归模型中偏回归系数 β_i 的实际意义是：设 δ_i 代表第 i 个自变量在两个不同个体身上取值差量的绝对值，在其他自变量取值不变的条件下，变量 δ_i 每增加一个单位所引起的风险比的自然对数，即 $\ln HR_i=\beta_i$。

当 $\beta_i>0$ 时，$HR_i>1$，说明 X_i 增加时，风险函数增加，X_i 为危险因素(其真正含义是：此类因素取高水平相对于取低水平风险增大)；当 $\beta_i<0$ 时，$HR_i<1$，说明 X_i 增加时，风险函数下降，X_i 为保护因素(其真正含义是：此类因素取高水平相对于取低水平风险减少)；当 $\beta_i=0$ 时，$HR_i=1$，说明 X_i 增加时，风险函数不变，X_i 为对生存时间无影响的因素。

1.4 参数估计与假设检验

偏回归系数 β_1、β_2、…、β_p 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(3)：

$$L=q_1q_2\cdots q_i\cdots q_k=\prod_{i=1}^{k}q_i=$$

$$\prod_{i=1}^{k}\frac{exp\left(\beta_1X_{i1}+\beta_2X_{i2}+\cdots+\beta_PX_{iP}\right)}{\sum_{S\in R\left(t_i\right)}exp\left(\beta_1X_{s1}+\beta_2X_{s2}+\cdots+\beta_PX_{sP}\right)} \quad (3)$$

式(3)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$t_i\left(t_1 \leqslant t_2 \leqslant \cdots \leqslant t_i \leqslant \cdots \leqslant t_k\right)$死亡时点的风险函数$h\left(t_i\right)$，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$\sum_{j=i}^{n} h_j(t)$，分子分母中的基线风险函数$h_0(t)$抵消了。一般的似然函数包含所有$n$个个体点，而式(3)中只含有k个死亡时点，忽略了删失时点的似然函数，所以称之为偏似然函数(或部分似然函数)。

对偏似然函数取对数，得到对数偏似然函数lnL，求lnL关于$\beta_j\left(j=1,2,\cdots,p\right)$的一阶偏导数为0的解(通常需要采用非线性迭代算法，此处从略)，便可获得β_j的最大似然估计值b_j。

假设检验方法类似于logistic回归分析，有似然比检验、Wald检验和Score检验，详细理论此处从略；使用统计软件计算时，参数估计与假设检验都可方便地实现。

2 实例分析

2.1 问题与数据结构

【例1】30例膀胱肿瘤患者生存资料原始记录见表1。研究者欲分析影响膀胱肿瘤患者生存时间(月)长短的因素，包括年龄、肿瘤分级、肿瘤大小和是否复发，并根据影响因素的取值对不同患者的生存情况进行预测。

表1 30例膀胱肿瘤患者生存资料原始记录

序号	年龄	肿瘤分级	肿瘤大小	复发	随访起点	随访终点	生存时间	结局
1	62	1	0	0	02/10/1996	12/30/2000	59	0
2	64	1	0	0	03/05/1996	08/12/2000	54	1
3	52	2	0	1	04/09/1996	12/03/1999	44	0
4	60	1	0	0	06/06/1996	10/27/2000	53	0
5	59	2	1	0	07/20/1996	06/21/1998	23	1
6	59	1	1	1	08/19/1996	09/10/1999	37	1
7	63	1	1	0	09/16/1996	10/20/2000	50	1
…	…	…	…	…	…	…	…	…
24	61	3	1	0	10/10/1998	06/13/2000	20	1
25	57	3	1	1	01/16/1999	12/20/1999	11	1
26	63	2	0	1	02/17/1999	04/20/2000	14	1
27	72	3	1	1	05/10/1999	05/12/2000	12	1
28	56	3	1	1	09/15/1999	06/17/2000	9	1
29	73	3	1	1	12/19/1999	07/26/2000	7	1
30	54	3	1	1	03/10/2000	09/20/2000	6	1

注：肿瘤分级，I级=1，II级=2，III级=3；肿瘤大小(cm)，≥3.0=1，<3.0=0；是否复发，是=1，否=0；生存结局，死亡=1，截尾=0

表1记录了30例膀胱肿瘤患者的年龄、肿瘤分级、肿瘤大小和是否复发等情况。其中，年龄和生存时间是定量变量，肿瘤分级、肿瘤大小、是否复发和生存结局是定性变量。由于存在截尾数据、生存时间及生存结局，且涉及到多个影响因素，所以，此资料为多因素影响下的一元生存资料。具体地说，生存时间为定量结果变量、生存结局为定性结果变量，它们的信息将被整合在一起参与Cox比例风险模型的建模；而其他变量都属于自变量或影响因素，其中年龄为定量自变量、肿瘤分级为多值有序自变量、肿瘤大小和是否复发为二值自变量。

2.2 创建SAS数据集

创建一个名为“tjfx”的临时数据集，数据步程序如下：

```
data tjfx;
input age grade size relapse t status@@;
cards;
62  1  0  0  59  0
64  1  0  0  54  1
52  2  0  1  44  0
……
56  3  1  1  9   1
73  3  1  1  7   1
54  3  1  1  6   1
;
run;
```

【SAS数据步程序说明】因篇幅所限，此处仅列出部分观测。详细数据见表1。

2.3 利用 SAS/STAT 中 LIFETEST 过程绘制生存曲线

该资料中年龄为定量变量，将年龄转化为二分类变量（<60 岁和≥60 岁），分别按年龄、肿瘤分级、肿瘤大小和是否复发这四个变量的水平分组，采用 Kaplan-Meier 法绘制生存曲线，程序如下：

```
data tjfx;
set tjfx;
if age<60 then agef=1;
else agef=2;
run;
proc lifetest method=PL plots=(s) data=tjfx;
time t*status(0);
strata agef;
run;
proc lifetest method=PL plots=(s) data=tjfx;
time t*status(0);
strata grade;
run;
proc lifetest method=PL plots=(s) data=tjfx;
time t*status(0);
strata size;
run;
proc lifetest method=PL plots=(s) data=tjfx;
time t*status(0);
strata relapse;
run;
```

【SAS 程序说明】生存曲线可用 lifetest 过程绘制，method 用于指定计算生存率的方法，PL 表示生存率的乘积极限估计法，即 Kaplan-Meier 法，plots=(s)用于绘制生存曲线；time 语句为 lifetest 过程的必需语句，设置生存时间变量和生存结局变量，括号内为截尾变量的标示值；strata 语句用于指定分层变量。

【SAS 主要输出结果】

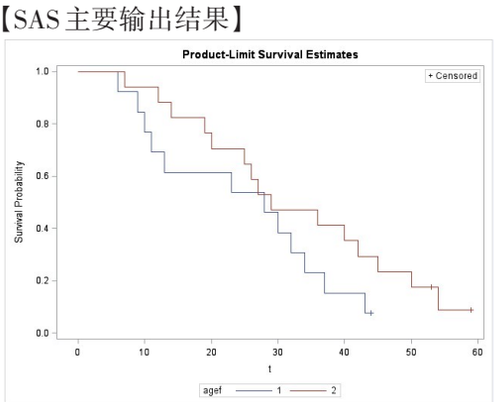

图 1 年龄各水平下的生存曲线

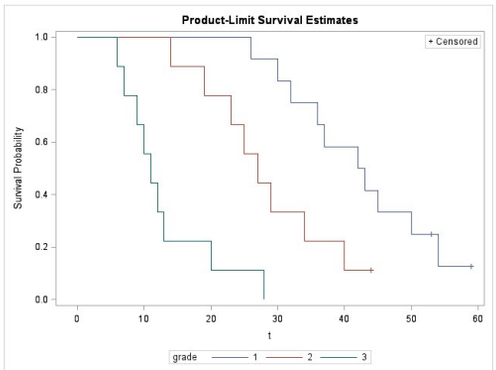

图 2 肿瘤分级各水平下的生存曲线

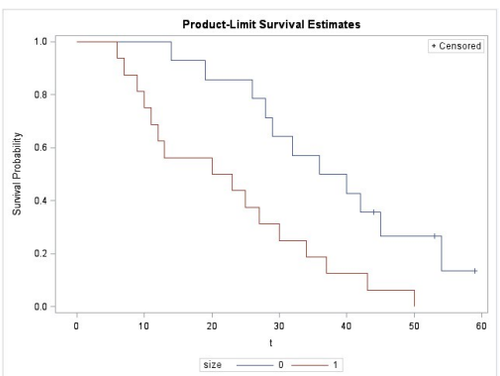

图 3 肿瘤大小各水平下的生存曲线

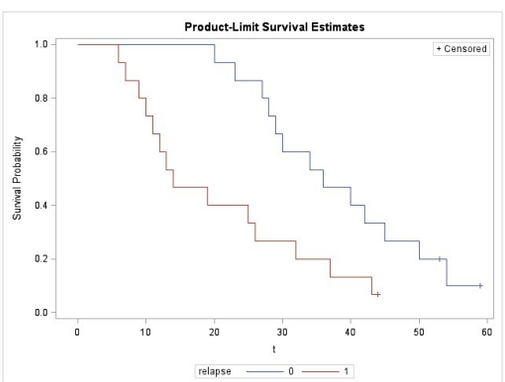

图 4 是否复发的生存曲线

以上四图显示，除两年龄组生存曲线在近 30 月处略有交叉外，其余各图中曲线均基本无交叉，说明四个变量基本满足 PH 假定，可以进行 Cox 比例风险回归模型分析。

2.4 利用 SAS/STAT 中 PHREG 过程拟合 Cox 比例风险回归模型

利用以下 SAS 程序，拟合 Cox 比例风险回归模型，计算每位患者的预后指数 pi 及其所对应生存时间的生存率。

```
proc phreg data=tjfx;
class grade(param=ref ref="1") size(param=ref ref="0") relapse(param=ref ref="0");
```

```
model t*status(0)=age grade size relapse/selection=stepwise sle=0.05 sls=0.05 rl;
output out=report survival=s xbeta=pi/order=data method=PL;
run;
proc print data=report;
run;
```

【SAS程序说明】phreg过程是实现Cox模型回归分析的标准过程，其中class语句可以为分类变量设置哑变量，ref=选项用来指定参考水平，这里需注意的是：肿瘤分级为有序多分类变量，不同的肿瘤分级之间并非是严格的等距关系，因此也将其转化为哑变量进行量化；model语句是必需语句，等号左边为生存时间和生存结局变量（括号内为截尾标识），右边为协变量（即自变量），其中选项selection=forward | backward | stepwise | none | score用来指定变量筛选的方法，分别表示前进法、后退法、逐步法、全回归模型和最优子集法，sle=和sls=分别指定引入和剔除自变量的显著性水平，rl用来指定输出风险比hr的100(1 − α)%置信限；程序中output语句创建一个新的SAS数据集report，含有为每一个观测计算的一些统计量，SAS为每一个统计量定义一个关键字，如生存率和预后指数分别用survival和xbeta表示。选项order=data规定输出的数据集中的观测顺序与输入数据集中的顺序一致。

【SAS主要输出结果及解释】

Model Information

Data Set	WORK.TJFX
Dependent Variable	t
Censoring Variable	status
Censoring Value(s)	0
Ties Handling	BRESLOW
Number of Observations Read	30
Number of Observations Used	30

Class Level Information

Class	Value	Design Variables	
grade	1	0	0
	2	1	0
	3	0	1
size	0	0	
	1	1	
relapse	0	0	
	1	1	

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
30	27	3	10.00

以上是模型的基本信息、分类变量的分类水平信息以及生存结局信息。

Model Fit Statistics

Criterion	Without Covariates	With Covariates
−2 LOG L	143.536	108.447
AIC	143.536	116.447
SBC	143.536	121.630

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	35.0889	4	<0.0001
Score	39.6265	4	<0.0001
Wald	24.4407	4	<0.0001

以上是经逐步回归后，用最终保留下来的协变量建立回归模型计算出的模型拟合统计量及模型检验结果，结果表明模型成立，即因变量与自变量之间的关系可以用所建立的回归方程来表示。

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr>ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
grade	2	1	1.34121	0.56085	5.7187	0.0168	3.824	1.274	11.478
grade	3	1	3.46098	0.76068	20.7014	<0.0001	31.848	7.171	141.441
size	1	1	1.00357	0.47308	4.5001	0.0339	2.728	1.079	6.895
relapse	1	1	1.02450	0.47318	4.6878	0.0304	2.786	1.102	7.042

以上是模型的最大似然估计结果，包括参数估计值、估计值标准误、Waldχ^2值、P值、风险比HR及其95%置信区间。由似然估计结果得出风险函数的表达式为：

$$h(t) = h_0(t)\,exp\left(1.341 \times grade_2 + 3.461 \times grade_3 + 1.004 \times size + 1.025 \times relapse\right)$$

Cox 回归模型计算结果显示：肿瘤分级、肿瘤大小和是否复发为膀胱肿瘤患者生存时间长短的重要影响因素。在肿瘤大小和是否复发保持不变的情形下，II 级肿瘤患者死亡风险是 I 级肿瘤患者死亡风险的 3.824 倍（$e^{1.34121}$），III 级肿瘤患者死亡风险是 I 级肿瘤患者的 31.848 倍（$e^{3.46098}$）；在肿瘤分级和是否复发保持不变的情形下，肿瘤大于或等于 3.0 cm 者死亡风险是肿瘤小于 3.0 cm 者死亡风险的 2.728 倍（$e^{1.00357}$）；在肿瘤分级和肿瘤大小保持不变的情形下，肿瘤复发者死亡风险是未复发者死亡风险的 2.786 倍（$e^{1.02450}$）。

风险函数表达式右边变量的线性组合部分与风险函数成正比，其取值越大，风险越大，反映了一个个体的预后情况，称为预后指数（prognostic index，PI），其值越大，则风险函数 $h(t)$ 的取值就越大，预后越差。此案例的预后指数为：

$$PI = 1.341 \times grade_2 + 3.461 \times grade_3 + 1.004 \times size + 1.025 \times relapse$$

Obs	age	grade	size	relapse	t	status	agef	pi	s
1	62	1	0	0	59	0	2	0.00000	0.24122
2	64	1	0	0	54	1	2	0.00000	0.24122
3	52	2	0	1	44	0	1	2.36571	0.02946
4	60	1	0	0	53	0	2	0.00000	0.48245
5	59	2	1	0	23	1	1	2.34478	0.73482

以上是 report 数据集中的内容，包括患者的基本信息、预后指数及其对应生存时间的生存率（由于篇幅限制，此处仅给出前 5 例患者的信息）。对于 1 号患者，肿瘤分级 I 级，肿瘤小于 3.0 cm，未复发，其预后指数为 0，59 个月的生存率为 24.12%。

3 讨论与小结

3.1 讨论

相对于非参数和参数回归模型而言，半参数回归模型兼有两者的优点，且不要求资料服从特定的概率分布，具有灵活性和稳健性，而且现如今还没有非常精准的方法判定待分析的生存资料中的生存时间服从何种分布，使得 Cox 比例风险回归模型在医学随访研究中得到广泛的应用。虽然 Cox 比例风险回归模型的适用范围广，但它依赖于严格的 PH 假定，若资料不满足 PH 假定，则会较大程度上影响计算的结果，甚至导致错误的结论。因此，在统计分析前，对 PH 假定的检验是重要且必须的。

3.2 小结

本文比较详细地介绍了 Cox 比例风险回归模型、构建模型需要满足的 PH 假定及其判定方法，并通过一个实例，基于 SAS 软件实现了 Cox 比例风险回归模型的构建、校正混杂因素后的组间比较以及对每个个体进行预后指数和生存率的预测。在进行 Cox 回归建模时，需注意只有满足 PH 假定前提下，基于此模型的分析预测才是可靠有效的。其次，还应注意回归所需样本含量的问题，经验估算方法是至少需要相当于协变量个数 10~15 倍的阳性结局事件数，或者根据 Hsieh 和 Lavori 给出的样本量估算公式进行计算，详细信息可参阅文献[7]。

参考文献

[1] 方积乾. 卫生统计学[M]. 7 版. 北京：人民卫生出版社，2012：410-426.

[2] 胡良平. SAS 常用统计分析教程[M]. 2 版. 北京：电子工业出版社，2015：527-532.

[3] 孙振球，徐勇勇. 医学统计学[M]. 4 版. 北京：人民卫生出版社，2014：291-296.

[4] Hess KR. Graphical methods for assessing violations of the proportional hazards assumption in cox regression[J]. Stat Med，1995，14(15)：1707-1723.

[5] 余红梅. Cox 比例风险回归模型诊断及预测有关问题的研究[D]. 西安：第四军医大学，2001.

[6] Ng′andu NH. An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox′s model[J]. Stat Med，1997，16(6)：611-626.

[7] Hsieh FY，Lavori PW. Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates[J]. Control Clin Trials，2000，21(6)：552-560.

（收稿日期：2020-01-06）

（本文编辑：吴俊林）