·基础理论与方法·

样条Cox回归在随访资料分析中的应用

董英 余金明 胡大一

【导读】 介绍借助R软件应用样条Cox回归分析不满足Cox比例风险模型两个基本假定条件的随访资料的方法，可同时估计非线性效应和时协效应。结果表明文中实例涉及的连续型协变量多不符合线性假定，3个变量不符合比例风险假定，应用样条Cox回归控制多个协变量后，踝臂指数每降低0.1，全因死亡的风险比（*HR*）为1.071。随访资料在不满足比例风险Cox回归模型的应用条件时，可选择应用样条Cox回归进行分析。

【关键词】 Cox比例风险模型；限制性立方样条；踝臂指数；全因死亡

Application of spline-based Cox regression on analyzing data from follow-up studies *DONG Ying*[1], *YU Jin-ming*[2], *HU Da-yi*[2]. 1 *Department of Preventive Medicine, Shanghai Traditional Chinese Medicine University, Shanghai* 201203, China; 2 *Clinical Research Institute and Key Laboratory of Public Health Safety, Ministry of Education, School of Public Health, Fudan University*
Corresponding author: YU Jin-ming, Email: jmy@fudan.edu.cn

【**Introduction**】 With R, this study involved the application of the spline- based Cox regression to analyze data related to follow-up studies when the two basic assumptions of Cox proportional hazards regression were not satisfactory. Results showed that most of the continuous covariates contributed nonlinearly to mortality risk while the effects of three covariates were time- dependent. After considering multiple covariates in spline-based Cox regression, when the ankle brachial index (ABI) decreased by 0.1, the hazard ratio (*HR*) for all-cause death was 1.071. The spline-based Cox regression method could be applied to analyze the data related to follow-up studies when the assumptions of Cox proportional hazards regression were violated.

【**Key words**】 Cox proportional hazards regression; Strict cubic spline; Ankle brachial index; All-cause mortality

Cox回归模型用风险函数反映协变量对生存期的影响，能够解决资料中截尾数据的问题，且可同时分析多因素对生存期的影响。但Cox比例风险模型要求满足两个假定，即具有不同回归向量的风险函数之比不随时间而改变，称为比例风险假定（Proportional Hazard, PH），及对数风险或对数累积风险与协变量间的关系应为线性。但在实际研究中，这两个假定往往并不被满足。

针对非线性关系，研究者常对协变量进行变量转换，使得转换后变量与结局事件危险性之间为线性关系，这在实际应用中有时很难做到。有研究者干脆将连续型变量进行分段，即转换成多分类甚至两分类变量拟合Cox模型，但这种处理方式不仅具有主观性，还损失了大量信息，并可能造成偏倚[1]。传统Cox比例风险模型应用中的条件限制及缺陷，已引起研究者的关注，开始探索新的途径进行生存分析。Durrleman和Simon[2]介绍了结合限制性立方样条函数进行Cox回归分析的方法，可放松Cox比例风险模型的两个假定条件，拟合协变量的曲线关系，并自由地估计时协变量的作用[3]，称之样条Cox回归。统计软件的发展，进一步推动了这一方法的应用。目前样条Cox回归模型可方便地在多种统计软件中实现，如SAS[4]、R软件等。本研究在R2.12.2软件中应用样条Cox回归模型分析踝臂指数（ABI）与全因死亡风险的关系，从而控制众多影响因素对全因死亡的影响，评价ABI的独立效应。

基本原理

1. Cox比例风险基本模型及两个基本假定：假定有n个观测，对每个观测i得到观测值(t_i, δ_i, X_i)，其中，t_i为生存时间；δ_i为截尾指示变量，对截尾观测$\delta_i=0$，对非截尾观测$\delta_i=1$；$X_i=(x_{i1}, x_{i2}, \cdots, x_{ip})$为$p$维行向量，表示第$i$观测的第$p$个协变量。Cox比例风险函数的一般形式：

$$h(t;X)=h_0(t)\exp(\beta X) \quad (1)$$

式中，$X=(x_1, x_2, \cdots, x_p)^T$表示$p$维协变量向量，$\beta=$

DOI: 10.3760/cma.j.issn.0254-6450.2012.09.021
作者单位：201203 上海中医药大学预防医学教研室（董英）；复旦大学公共卫生学院临床流行病学研究中心 教育部公共安全重点实验室（余金明、胡大一）
通信作者：余金明，Email: jmy@fudan.edu.cn

$(\beta_1, \beta_2, \cdots, \beta_p)$表示回归系数向量，$h_0(t)$为基准风险函数，则第$i$个个体的风险率：

$$h(t;X_i)=h_0(t)\exp(\beta X_i) \quad (2)$$

在应用Cox模型进行统计推断和预测前，必须首先考察生存资料是否满足该模型的两个基本假定。

（1）比例风险假定：若任意两个体的风险函数之比为风险比（*HR*）或相对危险度（*RR*）：

$$HR=\frac{h_i(t,X_i)}{h_j(t,X_j)}=\frac{h_0(t)\exp(\beta_1X_{i1}+\beta_2X_{i2}+\cdots+\beta_pX_{ip})}{h_0(t)\exp(\beta_1X_{j1}+\beta_2X_{j2}+\cdots+\beta_pX_{jp})}=\exp[\beta_1(X_{i1}-X_{j1})+\cdots+\beta_p(X_{ip}-X_{jp})] \quad (3)$$

此时，*HR*不随时间变化而变化，即在协变量不同状态下，个体的风险比在不同时间点为常数，即传统Cox模型的比例风险假定。

（2）对数线性假定：对模型中的连续型变量，任意个体i的对数风险或对数累积风险与协变量X_i呈线性关系。

$$\log h_i(t)-\log h_0(t)=\beta X_i,\ i=1,2,\cdots,n \quad (4)$$

2. Cox模型中的限制性立方样条函数：回归样条（regression spline）本质上为一个分段多项式，一般要求每个分段点上连续且二阶可导。即设自变量数据的范围在区间$[a,b]$，并根据需要分成k段：$a=t_0<t_1<\cdots<t_{k-1}<t_k=b$，在每个区间$[t_{i-1},t_i)$分别用一个多项式表示，则回归样条$f(x)=S_i(x)$，当$x\in[t_{i-1},t_i)$且$f''(x)$在$[a,b]$区间存在连续性。限制性立方样条（strict cubic spline，RCS）是在回归样条的基础上要求：样条函数在自变量范围两端的两个区间$[t_0,t_1)$和$[t_{k-1},t_k]$内是线性函数。

若在Cox回归中产生一个时间三节点的限制性立方样条函数，首先定义一个新的时协变量$S_i(t)$，t表示时间，即

$$f(t)=\mathrm{RCS}(t,k)=\sum_{i=1}^{k-1}\beta_iS_i(t) \quad (5)$$

产生新的风险函数表达式为

$$h(t;X)=h_0(t)\exp[\beta_1x_1+\beta_2x_1t+\beta_3x_1S_1(t)] \quad (6)$$

变量x_1的单位变化，相应的对数风险比函数（log hazards ratio function，LHRF）为

$$\mathrm{LHRF}=\beta_1+\beta_2t+\beta_3S_1(t) \quad (7)$$

检验$\beta_3=0$是否成立即对应非线性时间依赖的检验，$\beta_2=\beta_3=0$是否成立的检验对应比例风险假定的检验。

实例分析

1. 资料基本情况：研究对象来自上海、北京市8家医院住院患者。纳入标准为年龄>35岁且基线时至少存在2个心血管病危险因素；排除标准为不能测量ABI或其他指标者和ABI>1.4的可疑动脉钙化者。共3732例住院患者进行（2.8±0.8）年的随访，记录随访期间用药、健康事件发生情况及最终生存结局。最终有效随访样本3117例，期间死亡497例。

表1　变量编码或取值范围

指　标	变量名	范围或变量值编码
生存状态	TB	1=死亡，0=存活
年龄（岁）	age	35～96
性别	sex	1=男性，2=女性
是否吸烟	smoke	1=吸烟或曾吸烟，0=从未吸烟
踝臂指数	ABI	0.00～1.39
体重指数（kg/m²）	BMI	9.91～41.67
收缩压（mm Hg）	SBP	80～290
脉压差（mm Hg）	pp	10～190
总胆固醇（mmol/L）	TC	0.70～11.50
甘油三酯（mmol/L）	TG	0.14～16.80
高密度脂蛋白胆固醇（mmol/L）	HDL-C	0.15～6.84
低密度脂蛋白胆固醇（mmol/L）	LDL-C	0.09～7.90
空腹血糖（mmol/L）	FPG	0.00～26.30
尿素氮（μmol/L）	BUN	0.36～560.00
肌酐（μmol/L）	Cre	22～1259
冠心病史	CHD	1=有，0=无
脑血管病史	cerebr	1=有，0=无
糖尿病史	diabetes	1=有，0=无
慢性肾病史	nephr	1=有，0=无
高血压史	hypertension	1=有，0=无
高血脂史	lipid	1=有，0=无
是否服用降压药	TE1	1=是，0=否
是否服用降糖药	TE2	1=是，0=否
是否服用降血脂药	TE3	1=是，0=否

根据文献资料，选择23个潜在的预测变量（表1）。

2. 主要结果：

（1）模型中各变量拟合形式的确定：主要结合图形观察和不同形式拟合模型的拟合优度比较，选择变量拟合模型的最佳形式。在考虑损失自由度情况下，选择χ^2值较大的变量拟合形式。各连续型变量最终在Cox模型中的拟合形式在表2中以黑体字显示。图1为ABI样条图观察，接近线性。

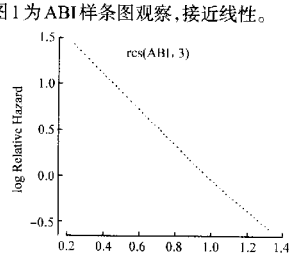

图1　ABI样条图

（2）比例风险假定检验及图形观察：对全因素模型进行比例风险假定的检查，包括图形法观察和假设检验。其中，服用降压药、降血糖药和降血脂药的Schoenfeld残差图见图2（其他变量图略）。可见服

表2 不同变量形式拟合Cox回归模型比较

预测变量	形式	χ^2值	自由度	P值
age	线性	128.3	1	<0.001
	平方	174.1	2	<0.001
	限制性立方样条,3节点	167.8	2	<0.001
	(年龄-60)+:60岁以后线性效应	157.9	1	<0.001
	(年龄-60)²+:60岁以后2次效应	**185.3**	**2**	**<0.001**
	<60,>60	49.9	1	<0.001
ABI	**线性**	**118.1**	**1**	**<0.001**
	限制性立方样条,3节点	121.2	2	<0.001
BMI	线性	80.0	1	<0.001
	限制性立方样条,3节点	**118.4**	**2**	**<0.001**
SBP	线性	1.14	1	0.285
	限制性立方样条,3节点	**12.8**	**2**	**0.002**
pp	线性	3.62	1	0.057
	平方	18.8	2	<0.001
	限制性立方样条,3节点	**22.1**	**2**	**<0.001**
TC	线性	34.6	1	<0.001
	限制性立方样条,3节点	**54.0**	**2**	**<0.001**
TG	线性	26.1	1	<0.001
	限制性立方样条,3节点	**61.8**	**2**	**<0.001**
HDL-C	线性	2.1	1	0.149
	限制性立方样条,3节点	**9.9**	**2**	**0.007**
LDL-C	线性	33.4	1	<0.001
	限制性立方样条,3节点	**41.2**	**2**	**<0.001**
FPG	线性	1.0	1	0.319
	平方	4.0	2	0.133
	限制性立方样条,3节点	**8.1**	**2**	**0.018**
BUN	**线性**	**102.1**	**1**	**<0.001**
	限制性立方样条,3节点	99.0	2	<0.001
Cre	线性	58.7	1	<0.001
	限制性立方样条,3节点	61.4	2	<0.001
	log转换	**79.9**	**1**	**<0.001**

用药物3个变量的Schoenfeld残差随时间的改变,有上升趋势,检验结果显示$P<0.05$,不符合比例风险假定(表3)。而其他变量的Schoenfeld残差随时间改变无变化趋势。因此,进一步在模型中加入服用3种药物的3个变量与时间样条函数交互作用项。

表3 比例风险假定的检验结果

预测变量	rho值	χ^2值	P值
TE1=1	0.195	17.600	<0.001
TE2=1	0.140	8.700	0.003
TE3=1	0.148	9.640	0.002

(3)模型拟合结果:构造时间的限制性立方样条函数为$f(t)$=rcs(time,3),令T=$f(t)$,在模型中加入服用降血压药、降血糖药和降血脂药3个变量与T的交互项。全变量模型估计结果见表4。在控制了众多影响因素后,ABI对死亡风险有影响,计算ABI每降低0.1,其$HR=\exp(0.1\times0.689)=1.071$,即对于心脑血管高危人群,ABI每降低0.1,则全因死亡风险增加0.071倍。结果也表明相关的3个变量依时效应非线性部分均有统计学意义。

讨 论

Cox比例风险模型被广泛应用于多因素生存分析,但是很多研究者并未正确应用该模型,主要反映在对Cox模型两个假定的认识上。只有少数研究者应用时提到了该模型的比例风险假定和对数线性假定,且多数研究没有对假定进行检查,当资料不满足假定时,直接应用即造成估计的偏差。针对这一问题,Goodman和Chandalia[5]指出以往很多大气污染暴露和健康效应关系的半生态研究都是应用Cox比例风险模型进行分析,但构建模型时忽视了Cox模型的两个假定,得到的结论有偏差,美国环境保护署(USEPA)在利用这些相关研究结论时应慎重。我国余红梅和何大卫[6]对Cox回归模型比例风险假定的检查进行了探讨,部分研究者统计分析时能够考虑到Cox回归模型的基本假定而选用其他方法进行数据分析,但该问题仍未引起广泛关注[7,8]。

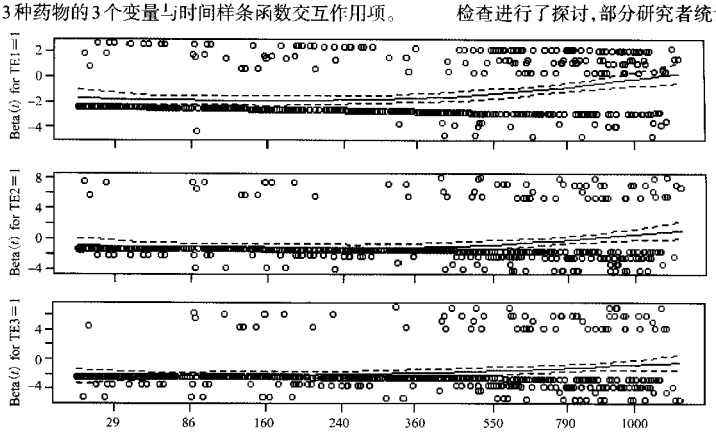

图2 服用3种药物的Schoenfeld残差图

由于限制性立方样条函数拟合资料具有更大的灵活性,样条Cox回归模型可以同时解决非对数线性和时协效应问题,因此得到广泛应用。Kattan[9]对相同资料拟合传统的Cox回归模型、样条

表4　全变量模型估计结果

预测变量	r	s	Z值	P值
sex	0.128	0.120	1.073	0.283
age	0.001	0.000	3.036	0.002
ABI	−0.689	0.211	−3.262	0.001
BMI	−0.067	0.024	−2.790	0.005
BMI′	0.040	0.031	1.322	0.186
SBP	0.001	0.008	0.075	0.941
SBP′	−0.001	0.007	−0.115	0.909
pp	−0.021	0.009	−2.357	0.018
pp′	0.022	0.009	2.551	0.011
TC	−0.170	0.141	−1.209	0.227
TC′	0.018	0.157	0.118	0.906
TG	−0.153	0.165	−0.930	0.352
TG′	0.144	0.211	0.685	0.494
HDL-C	−0.444	0.381	−1.167	0.243
HDL-C′	0.447	0.480	0.930	0.353
LDL-C	0.073	0.183	0.397	0.692
LDL-C′	0.112	0.231	0.484	0.629
glu	−0.025	0.070	−0.363	0.717
glu′	0.164	0.135	1.216	0.224
BUN	0.010	0.017	0.581	0.561
CRE	0.303	0.164	1.844	0.065
cerebr	0.304	0.099	3.060	0.002
diabetes	0.345	0.125	2.771	0.006
smoke	0.233	0.114	2.052	0.040
CHD	0.147	0.101	1.466	0.143
nephr	−0.252	0.143	−1.753	0.080
hypertension	0.544	0.142	3.830	0.000
lipid	0.094	0.125	0.749	0.454
TE1	1.385	0.290	4.774	<0.001
TE2	0.171	0.472	0.364	0.716
TE3	−0.745	0.486	−1.534	0.125
TE1T	−0.001	0.001	−2.499	0.013
TE1T′	−0.005	0.001	−6.392	0.000
TE2T	0.000	0.001	−0.458	<0.001
TE2T′	−0.003	0.001	−2.480	0.013
TE3T	0.002	0.001	2.673	0.008
TE3T′	−0.008	0.001	−6.629	<0.001

注：BMI用限制性立方样条进行估计，估计结果包括两部分：BMI代表线性部分，BMI′代表非线性部分；其他变量同此

Cox回归模型、不同树模型及人工神经网络模型，结果发现样条Cox回归模型在预测能力、分类能力方面均优于这些模型。徐添等[7]尝试应用Buckley-James模型分析不满足比例风险假定的生存资料，但该方法为右删失数据的线性回归模型，要求自变量与应变量有线性关系，这在资料分析中往往不被满足。贺佳等[8]应用BP神经网络预测肝癌患者生存期，多层神经网络往往可以得到很好的预测性能，但不能得到诸如相对风险等指标，对结果的解释较差，适用于以预测为目的的研究。

Goodman和Chandalia[5]评论中也指出了结合样条函数构建Cox模型存在的问题，认为模型过于自由，选择样条函数节点数和节点位置具有主观性，而不同的节点数和位置估计的结果有差异。节点的个数和位置决定着样条曲线的形状，多数学者认为节点数量的选择更为重要，一般认为3～7个节点即可[10]。本文在对资料的分析中，通过图形观察和模型拟合优度的比较综合判断节点的个数，结果表明对多数变量取3节点即可满足分析的需要。当然，并不是所有变量用限制性立方样条函数拟合最为合适，文中也对其他变量形式如2次方项、log转换形式等进行尝试，最终选择各变量的最佳拟合形式。

本文应用样条Cox回归分析ABI对全因死亡的预测价值，控制其他因素后，ABI具有独立的效应，ABI每降低0.1，全因死亡的*HR*为1.071，与Feringa等[11]研究结果(1.08)相近，ABI降低会增加全因死亡发生的风险。模型中控制的多数变量并不满足对数线性假定，部分变量不满足比例风险假定，应用Cox模型前对两个假定条件的检查非常必要。

本文认为在传统Cox模型基础上加入限制性立方样条函数的应用，可以解决非线性问题和时协效应问题，值得应用。但全变量模型中考虑的协变量较多，选择变量的最佳拟合形式工作量很大，当考虑影响因素较少时，这种方法才更加高效。

参 考 文 献

[1] Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med, 2006, 25(1): 127-141.

[2] Durrleman S, Simon R. Flexible regression models with cubic splines. Stat Med, 1989, 8(5): 551-561.

[3] Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. Stat Med, 1994, 13(10): 1045-1062.

[4] Heinzl H, Kaider A. Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. Comput Methods Programs Biomed, 1997, 54(3): 201-208.

[5] Goodman JE, Chandalia JK. Comments on the use of the Cox proportional hazards model to assess the association between air pollutant exposure and health effects in semi-ecological studies. 2010. http://yosemite.epa.gov/sab/sabproduct.nsf/.../Gradient_+Comments_082010.pdf.

[6] Yu HM, He DW. Graphical methods for investigating the proportional hazards assumption in Cox model. Chin J Health Stat, 2000, 17(4): 215-218. (in Chinese)
余红梅，何大卫. 检查Cox模型比例风险假定的几种图示法. 中国卫生统计，2000，17(4)：215-218.

[7] Xu T, Zhao HY, Yan Y, et al. Survival analysis on advanced non-small cell lung cancer with a Buckley-James model. Chin J Epidemiol, 2010, 31(10): 1179-1184. (in Chinese)
徐添，赵洪瑜，严媛，等. 晚期非小细胞肺癌Buckley-James模型生存分析. 中华流行病学杂志，2010，31(10)：1179-1184.

[8] He J, He XM, Liu Q, et al. The prediction of the survival time about hepatoma patients with the BP neural network method. Chin J Health Stat, 2001, 18(1): 17-20. (in Chinese)
贺佳，贺宪民，刘峰，等. BP神经网络预测肝癌患者生存期的研究. 中国卫生统计，2001，18(1)：17-20.

[9] Kattan MW. Comparison of Cox regression with other methods for determining prediction models and nomograms. J Urol, 2003, 170(6): 6-10.

[10] Yu HM, Xu YY, He DW. Assessing proportional hazards assumption in Cox regression models using cubic spline functions. Chin J Health Stat, 2002, 19(1): 20-24. (in Chinese)
余红梅，徐勇勇，何大卫. 利用三次样条函数考察Cox模型比例风险假定. 中国卫生统计，2002，19(1)：20-24.

[11] Feringa HHH, Bax JJJ, Waning VH, et al. The long-term prognostic value of the resting and postexercise Ankle-Brachial index. Arch Intern Med, 2006, 166(5): 529-535.

（收稿日期：2012-05-14）
（本文编辑：张林东）