文章编号: 0490-6756(1999) 04-0660-04

Cox模型中共线协变量的分层处理*

林华珍[1],倪宗瓒[2]
(1.四川大学数学学院,成都 610064; 2.华西医科大学,成都 610041)

摘要: 协变量间的共线关系导致不准确的估计及检验,现有的几种方法都存在不同程度的缺陷. 作者给出一种分层处理方法. 从模拟试验的拟合效果看,优于主成分方法、似然比检验及单变量分析方法. 将该方法用于一个实际例子,得到较好的结果.

关键词: Cox模型;共线协变量;分层
中图分类号: O212.1 文献标识码: A

1 引言

在变量筛选及参数估计中,都要求各变量相互独立. 但在很多应用研究中,自变量间不独立,相互间有一定的线性依存关系. 这种线性依存关系常导致: (1)回归系数估计有偏差; (2)标准误差变大; (3)检验结果不准确[1,2]. 更糟的是,分析结果极不稳定,删除少数观察值,甚至改变数据集的顺序都可引起分析结果的改变.

在 Cox模型中,处理共线协变量有以下几种方法: (1)主成分方法[3]: 在信息损失较少的情况下,由原变量综合成彼此独立数目较少的主成分,以主分代替原变量进行模型拟合,它使彼此相关的变量彼此独立. 但主成分方法所剔除的信息不是基于该信息与应变量的关系,因此有可能剔除重要的因素. 并且若原协变量仅有两个且共线,用主成分方法后,原变量的检验结果总是相同[4],这显然不合理. (2)似然比检验: 在共线关系较强时,其检验结果不理想,并且该方法不能得到系数的估计. (3)单变量分析方法: 将相关变量分开分析. 这种方法损失大量信息并且可能有错误结果. 比如, z_1是危险因素, z_2不是危险因素,仅因为 z_1, z_2的共线关系可能使 z_2单变量分析结果有显著性. 上述几种问题我们将以模拟试验说明.

上述几种方法都有不同程度的缺陷. 为此本文提出共线协变量的分层处理方法. 我们首先以两共线协变量为例说明,其主要原理是: 将其中一个变量进行适当分层,分析另一个变量的显著性. 从§ 4的模拟试验的拟合效果看,优于主成分方法、似然比检验及单变量分析. 由于分层是一个关键步骤,在§ 2给出一种简易的分层方法. § 3给出检验方法. 最后是一个实例.

2 连续性变量的分层

Cox模型[5]很好地连接起协变量与生存时间的关系,是处理具有协变量的生存数据的常用模型. 假定 $\lambda(t,z)$表协变量为 z 时 t 时刻的风险函数,我们有

$$\lambda(t,z) = \lambda_0(t)\exp(\beta z), \tag{1}$$

其中 $\lambda_0(t)$为非负基础风险函数, z为 p维自变量, β为 p维未知回归系数.

收稿日期: 1999-03-10
基金项目: 国家自然科学基金(39770662)

为将数据按协变量 x 分层,首先拟合

$$\lambda(t,x)=\lambda_0(t)\exp(\beta x),\tag{2}$$

将数据集按 x 由小到大排列,并将该有序数据集按顺序分为若干子集,记为 $R_1,R_2,\cdots,R_m$,其中 R_1 中 x 的最小. R_m 中 x 的最大,分层后满足: 分别以 $R_1,R_2,\cdots,R_m$ 为数据集进行分析,x 均无显著意义(只要 m 足够大,一般能够满足). 以后按下列步骤将 $R_1,R_2,\cdots,R_m$ 适当合并:

记 $U_{j,i}=R_j\cup R_{j+1}\cup\cdots\cup R_i,\ j=1,2,\cdots,m,\ i=j,j+1,\cdots,m.$

(i) 首先在 $U_{1,m}$ 中按模型(2)作数据分析,若 x 无显著意义,则视 $U_{1,m}$ 为第一层;若 X 有显著意义,则在 $U_{1,m-1},U_{1,m-2}\cdots$,中分析,直到 x 无显著意义,设在 $U_{1,m},U_{1,m-1},\cdots,U_{1,k_1+1}$ 中 x 有显著性,在 U_{1,k_1} 中无显著性,此时视 U_{1,k_1} 为第一层.

(ii) 仿(i),依次在 $U_{k_1+1,m},U_{k_1+1,m-1},U_{k_1+1,m-2},\cdots$ 中分析,至到 x 无显著意义,设在 $U_{k_1+1,m},U_{k_1+1,m-1},\cdots U_{k_1+1,k_1+k_2+1}$ 中 x 有显著性,在 U_{k_1+1,k_1+k_2} 中无显著性,此时视 U_{k_1+1,k_1+k_2} 为第二层.

(iii) 类似可在余下的数据集中找到第三层,第四层,$\cdots$. 设 $R_1,R_2,\cdots,R_m$ 可合并为 r 层,记为 $U_1,U_2,\cdots,U_r$. 在各层内 x 无显著意义,即 x 限于各层内的变化不会导致风险的改变. 对风险而言,各层内的 x 可视为常量. 将 $U_1,U_2,\cdots,U_r$ 中的 x 分别视为常数 $a_1,a_2,\cdots,a_r$.

3　检验方法

设有两个共线协变量 x_1,x_2 满足 Cox模型. 将数据集按 x_1 分层,分为 $U_1,U_2,\cdots,U_r$,得到相应的分层变量 s,即当个体 $j\in U_i$ 时,$s_j=i$. 以 s 为分层变量,x_2 为协变量拟合 Cox模型

$$\lambda(t,x_2)=\lambda_{i0}(t)\exp(\beta x_2),\tag{3}$$

其中 $\lambda_{i0}(t)$ 为第 i 层的基础风险函数. 第 i 层的偏似然函数为

$$L_i=\prod_{k=1}^{n_i}\left(\frac{\exp(\beta x_{2ik})}{\sum_{j\in R(t_{ik})}\exp(\beta x_{2ij})}\right)^{\Delta_{ik}},\tag{4}$$

其中 n_i 为第 i 层的观察数,$i=1,2,\cdots,r$,$(x_{2ik},t_{ik},\Delta_{ik})$ 分别为第 i 层中观察对象 k 的 x_2 观察值、生存时间、截尾指示变量(当观察对象 k 在 t_{ik} 上死亡,$\Delta_{ik}=1$;否则 $\Delta_{ik}=0$),$R(t_{ik})=\{$观察对象 j: $S_j=i$ 且 $t_j\geqslant t_{ik}\}$ 为 t_{ik} 上的风险集,$k=1,2,\cdots,n_i$. 第 i 层的偏似然函数仅涉及到第 i 层的数据. 模型(3)的偏似然函数是各层偏似然函数的积,各层的 x_1 是固定的,因此 x_2 效应估计是固定 x_1 得到的,避免了 x_1,x_2 的共线关系. 类似可求得 x_1 的回归系数的估计及检验.

若有两个以上协变量共线,比如,$x_1,x_2,\cdots,x_p$ 间有很强的共线关系,为得到 x_1 的系数的估计及检验,我们将数据按 $x_2,\cdots,x_p$ 分层,具体的分层方法为: 首先将数据照 § 2的方法按 x_2 分层,假设数据被分为 $R_{21},\cdots,R_{2K}$,依同样的方法,按 x_3 分别将数据 $R_{21},\cdots,R_{2K}$ 进行分层,这样,$R_{21},\cdots R_{21},\cdots,R_{2K}$ 中的每层数据被分为若干层,依此,将各层数据再按 $x_4,\cdots,x_p$ 分层. 假设数据最终被分为 $U_1,U_2,\cdots,U_r$,与前述方法类似,给出分层变量 s,即当个体 $j\in U_i$ 时,$s_j=i$. 以 s 为分层变量,x_1 为协变量拟合 Cox模型,得到其系数的估计及检验.

4　模拟试验

x_1 取自均匀分布 $U(1,5)$,$x_2=3x_1+X$,X 服从标准正态分布 $N(0,1)$,x_1,x_2 的相关系数 $\rho(x_1,x_2)=0.96$,t 时刻的风险函数满足:

$$\lambda(t,x) = \lambda_0(t)\exp(-0.05x_2), \tag{5}$$

$\lambda_0(t)$服从风险率为 1的指数分布.抽 1000个样本,并有 25% 被均匀分布 $U(0,6)$截尾.我们分别用主成分方法、似然比检验、单变量分析、多变量分析及分层方法估计 x_1, x_2的回归系数并作检验.结果见下表:

		系数	标准差	P
主成分方法	x_1	-0.01504498	0.002908936	2.316152e-007
	x_2	-0.04800898	0.009282496	2.316152e-007
似然比检验	x_1			0.99
	x_2			0.1797125
单变量分析	x_1	-0.157	0.0315	5.9e-007
	x_2	-0.0528	0.0102	2.3e-007
多变量分析	x_1	0.0112	0.1273	0.93
	x_2	-0.0563	0.0413	0.17
分层方法	x_1	-0.0711	0.05	0.16
	x_2	-0.0349	0.0159	0.028

在我们的模拟试验中,x_1不是影响因素,x_2是影响因素.用主成分方法后,x_1,x_2的检验结果相同.似然比检验判定 x_1,x_2都不是影响因素,并且不能估计系数.由 x_1,x_2的共线关系单变量分析结果显示 x_1也有显著性.多变量分析结果与似然比检验类似.仅分层方法显示 x_1不是影响因素,x_2是影响因素.

由于分层方法能准确地判定各因素的作用,以此为依据,剔除无关因素,再对有关的危险因素作估计及检验.在我们的模拟试验中,通过分层方法确定 x_1无统计显著性,剔除后,仅用 x_2拟合模型,其系数的估计为: -0.0528, p = 2.3e-007,已相当接近真实值.

5 应用实例

本研究来源于某厂矿,用队列研究方法前瞻观察若干年.共调查 8808人,其中有 8209人符合队列条件,调查结束时有 338人患病.研究的目的是确定高危人群癌症发病与致病因素的关系.通过初步的变量筛选后,确定感兴趣因素如下:

T: 退出队列时已暴露的时间, x_1: 氡子体累积暴露剂量,

x_2: 砷累积暴露剂量, x_3 粉尘累积暴露剂量.

x_1, x_2, x_3是同一环境中存在的有害因素,随着暴露年份的增加,x_1, x_2, x_3也相应增加,它们之间有很强的相关性,其相关系数分别为 $d(x_1,x_2) = 0.745$, $d(x_1,x_3) = 0.615$, $d(x_2,x_3) = 0.849$.另外,其相关性还可以从单因素分析与多因素分析结果的不同中看出.

	单因素分析*		多因素分析**		似然比检验
	系数	P	系数	P	P
x_1	0.00058	3.1e-006	0.000174	0.29	0.2942661
x_2	0.000303	1.7e-008	0.000229	0.0016	0.001745119
x_3	0.000207	0.00032	0.0000784	0.22	0.2206714

* 将相关变量分开分析;** 以 x_1, x_2, x_3为协变量得到的结果.

由于相关性,单因素分析与多因素分析结果都不准确,不能作为判断的依据.似然比检验显示 x_1无危险性,这与有关的流行病学及生物试验的证据相反[6].分层方法的结果见右表.

	分层方法	
	系数	P
x_1	0.000318	0.033
x_2	0.000252	0.00013
x_3	0.0000693	0.3

分层方法显示 x_1是危险因素,与有关的流行病学及生物试验证据吻合.

本文的编程及计算用 *S-plus*[8] 完成.

[参　考　文　献]

[1] Kleimbaum, Kupper and Muller. Applied Regression Analysis and Other Multivariable Methods. Duxbury Press, 1987.

[2]李严洁.多元回归中的多重共线性及其存在的后果 [J].中国卫生统计, 1992 9(1): 24~ 27.

[3] Anderson T W. An Introduction to Multivariate Statistical Analysis, Wiley, 1984

[4]林华珍等.多重共线性变量的回归系数估计及检验 [J].中国公共卫生, 1999, 15(2): 131~ 132

[5] Cox D R. Regression model and life-tables, Journal of the Royal Statistical Society, Series, 1972, 34 187~ 202

[6]卢伟.居室氡气致癌与防治 [M].北京:地质出版社, 1995.

[7] Belsley D A, Kuh E, R. E. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley & Sons, Inc. 1980.

[8] STATSCI INC. S-plus guide to statistical and mathematical analysis, version 3.3, 1995.

TREATMENT FOR COLLINEARITY IN Cox MODEL

LIN Hua-zhen[1], *ZI Zong-zhan*[2]

(1. Mathematical College, Sichuan University, Chengdu 610064; 2. West China University of Medical Science, Chengdu 610041)

Abstract Collinearity create numerical problems, include the inaccurate estimation of regress coefficient and hypothesis test statistics. The methods of principal components, likelihood ratio test and one-way variable analysis are often recommended for treating collinearity problems, but all of them maybe cause inaccurate results. The authors propose a method for treating collinearity survival data. The method changes one of collinearity variables to stratum variable, then analyze another variable. Monte Carlo simulation shows it is an appealing alternative to previously described methods. An example is presented as an illustration.

Key words Cox proportional hazard model; collinearity covariate; stratum variable

(1991 MSC 62J05)