·基础理论与方法·

Logistic回归模型中连续变量交互作用的分析

邱宏　余德新　谢立亚　王晓蓉　付振明

【导读】 Rothman提出生物学交互作用的评价应该基于相加尺度即是否有相加交互作用，而logistic回归模型的乘积项反映的是相乘交互作用。目前国内外文献讨论logistic回归模型中两因素的相加交互作用以两分类变量为主，本文介绍两连续变量或连续变量与分类变量相加交互作用可信区间估计的Bootstrap方法，文中以香港男性肺癌病例对照研究资料为例，辅以免费软件R的实现程序，为研究人员分析交互作用提供参考。

【关键词】 logistic回归模型；连续变量；相加交互作用S指数；方法，Bootstrap

Interaction between continuous variables in logistic regression model *QIU Hong, Ignatius Tak-sun YU, Lap Ah TSE, WANG Xiao-rong, FU Zhen-ming. School of Public Health and Primary Care, Chinese University of Hong Kong, H.K.S.A.R*

Corresponding author: Ignatius Tak-sun YU, Email: iyu@cuhk.edu.hk

【Introduction】 Rothman argued that interaction estimated as departure from additivity better reflected the biological interaction. In a logistic regression model, the product term reflects the interaction as departure from multiplicativity. So far, literature on estimating interaction regarding an additive scale using logistic regression was only focusing on two dichotomous factors. The objective of the present report was to provide a method to examine the interaction as departure from additivity between two continuous variables or between one continuous variable and one categorical variable. We used data from a lung cancer case-control study among males in Hong Kong as an example to illustrate the bootstrap re-sampling method for calculating the corresponding confidence intervals. Free software R (Version 2.8.1) was used to estimate interaction on the additive scale.

【Key words】 Logistic regression model; Continuous variable; Interaction departure from additivity, Synergy index; Bootstrap

流行病学病因学研究在分析因素间的交互作用时，多采用在回归方程中纳入因素乘积项的方法。线性回归模型为相加模型，乘积项反映因素间是否有相加交互作用，logistic回归或Cox回归模型为相乘模型，乘积项反映因素间是否有相乘交互作用，这是统计学意义上的交互作用。目前以Rothman为代表的多数流行病学者认为生物学交互作用的评价应该基于相加而非相乘尺度[1-3]，并对logistic、Cox回归等相乘模型构建了相对超危险度比(the relative excess risk due to interaction, *RERI*)、归因比(the attributable proportion due to interaction, *AP*)和交互作用指数(the synergy index, *S*)3个指标，用于评价因素间是否有区别于相乘交互作用的相加交互作用，进而为生物学交互作用的评价提供依据。我们曾撰文讨论logistic回归模型中两分类变量相加交互作用的分析和评价[4]，但实际资料分析中还常常碰到连续型变量，如年龄、BMI、吸烟量等资料，人为转化成两分类变量导致信息的损失，因此对两连续变量或连续变量和分类变量之间相加交互作用的分析做一补充，以满足实际资料分析的需要。

DOI: 10.3760/cma.j.issn.0254-6450.2010.07.021

作者单位：香港中文大学公共卫生及基层医疗学院

通信作者：余德新，Email: iyu@cuhk.edu.hk

基本原理

记*B*因素不变时*A*因素增加一个单位引起发病的比数比为OR_{10}，*A*因素不变时*B*因素增加一个单位引起发病的比数比为OR_{01}，*A*因素增加一个单位、*B*因素也增加一个单位引起发病的比数比为OR_{11}。Rothman和Hosmer用于评价相加交互作用的指标：

$$RERI = RR_{11} - RR_{10} - RR_{01} + 1$$

$$AP = RERI / RR_{11}$$

$$S = (RR_{11} - 1) / [(RR_{01} - 1) + (RR_{10} - 1)]$$

如果两因素无相加交互作用，则*RERI*和*AP*的可信区间应该包含0，*S*的可信区间应该包含1。

假设用两因素*A*、*B*（可为两分类、多分类或连续变量）及乘积项$A \times B$构建logistic回归模型：$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 A \times B$，则有：$OR_{10} = e^{\beta_1}$，$OR_{01} = e^{\beta_2}$，$OR_{11} = e^{\beta_1+\beta_2+\beta_3}$，用*OR*作为*RR*的估计值代入上述交互作用指标的计算公式即可得到3个指标的点估计。

当因素为两分类变量时，利用Multinomial logistic过程得到因素间的方差和协方差矩阵运用Hosmer和Lemeshow介绍的delta方法估计可信区间[3,4]，但当两因素或其中之一为连续变量时，每改变2个单位或5个单位将导致*RERI*、*AP*和*S*及其可信区间的非线性变化[5]，delta法不再适用，Assmann等[6]提出用Bootstrap法估计可信区间。Bootstrap是以原始样本为基础的再抽样进行统计推断的方法[7]：在原始数据（假设样本含量为n，每个观察单位每次被抽到的概率相等）中进行有放回的模拟随机再抽样，所得新样本称为Bootstrap样本，样本含量仍为n，按既定的公式计算相加交互作用指标的数值。由此重复R次，得到R个Bootstrap样本、R个某相加交互作用指标的数值，再根据该指标频数分布特征采取正态分布原理或百分位数法估计可信区间。Bootstrap再抽样及可信区间的估计可由免费软件R[8]加载Boot软件包实现。

实例分析

以香港男性肺癌的病例对照研究资料为例（表1），分别分析两连续变量吸烟和住宅氡暴露、连续变量吸烟与分类变量肿瘤家族史在男性肺癌发生过程中的交互作用，为简化计算，仅以年龄作为混杂因素纳入模型。

表1 香港男性肺癌病例对照研究资料

变量	病例（n=1208）	社区对照（n=1069）	*P*值
年龄(age)	65.8±9.5	66.2±9.9	0.326
吸烟量(packyr)	44.1±33.8	17.1±29.0	<0.001
住宅氡暴露(radon)	8.9±1.3	8.8±1.3	0.008
肿瘤家族史(fcahis)			
无	784(64.9)	760(71.1)	<0.001
有	239(19.8)	134(12.5)	
不确定/不知道	185(15.3)	175(16.4)	

注：该研究由香港研究局资助（CUHK4460/03M）；住宅氡暴露是根据调查对象的住宅建筑材料、墙面粉刷材料、楼龄、楼层、居住时间、开窗习惯等资料计算出来的累积氡暴露指数[9]，数值越大表示个体累积氡暴露量越大

以下为纳入乘积项的logistic回归模型参数估计及交互作用*S*指数区间估计的*R*程序：

```
interaction<-read.table (file="C:/MyDocuments/interaction/test.dat",header=T,sep="\t")
/*读入以数据文件(.dat)或文本文件(.txt)存储的数据集，命名为"interaction"。
names(interaction) /*查看变量名
"case" "age" "packyr" "radon" "fcahis" "packyr10" /*packyr10为packyr除以10得到，可用于估计吸烟量每增加10个单位的OR值。
logistic1<-glm(case ~ packyr* radon+age, family=binomial, data=interaction) /*建立含乘积项的logistic回归模型，考虑年龄作为混杂因素。
summary(logistic1) /*查看参数估计值，结果见表2。
logistic2 <- glm(case ~ packyr10 * as.factor(fcahis)+age, family=binomial, data=interaction) /*对分类变量fcahis用哑变量纳入logistic回归模型。
summary(logistic2) /*查看参数估计值，结果见表3。以上部分的分析亦可由SPSS或SAS完成。
library(boot) /*安装加载Boot软件包。
s1 <- function(datsam, indices){d<-datsam[indices,]
fitlr <- glm(case ~ packyr * radon+age, family=binomial, data=d)
s <- (exp(fitlr$coef[2]+fitlr$coef[3]+fitlr$coef[5]) - 1) / (exp(fitlr$coef[2]) -1 + exp(fitlr$coef[3]) - 1)} /*…允许Bootstrap从原样本数据再抽样，并自定义从再抽样的样本中计算两连续变量交互作用S指数的功能函数。
outs1<-boot (data=interaction, statistic=s1, R=1000) /*Bootstrap再抽样1000次并计算S指数。
print(outs1) /*输出S指数的点估计值和Bootstrap样本的再抽样误差。
plot(outs1)/*作S指数的频数分布图，呈偏态分布。
boot.ci(outs1, conf=0.95, type="perc") /*用百分位数法估计S指数的95%CI。
s2 <- function(datsam, indices){d<-datsam[indices,]
fitlr <- glm(case ~ packyr10*as.factor(fcahis)+age, family=binomial, data=d)
s <- (exp(fitlr$coef[2]+fitlr$coef[3]+fitlr$coef[6]) -1) / (exp(fitlr$coef[2]) -1+exp(fitlr$coef[3])-1)} /*自定义从再抽样的样本中计算连续变量与分类变量交互作用S指数的功能函数。
outs2<-boot (data=interaction, statistic=s2, R=1000) /*Bootstrap再抽样1000次并计算S指数。
print(outs2) /*输出S指数的点估计值和Bootstrap样本的再抽样误差。
plot(outs2) /*作S指数的频数分布图，呈偏态分布。
boot.ci(outs2, conf= 0.95, type="perc") /*用百分位数法估计S指数的95%CI。
```

表2　吸烟(packyr)和住宅氡暴露(radon)的logistic回归结果

变量	β	P值	OR值(95%CI)
packyr	0.065	<0.001	1.07(1.04～1.10)
radon	0.179	0.001	1.20(1.08～1.33)
Packyr×radon	−0.003	0.025	0.997(0.994～1.000)

表3　吸烟(packyr)和肿瘤家族史(fcahis)的logistic回归结果

变量	β	P值	OR值(95%CI)
packyr10	0.392	<0.001	1.48(1.41～1.55)
fcahis(1)*	0.410	0.028	1.51(1.05～2.17)
packyr10×fcahis(1)	0.075	0.230	1.08(0.95～1.22)

注：fcahis以哑变量纳入模型，fcahis(1)表示有肿瘤家族史的类别

logistic回归结果如表2、3所示，说明吸烟和住宅氡暴露存在负相乘交互作用，联合作用小于单独作用的乘积；而吸烟和肿瘤家族史的联合作用符合相乘模型，没有相乘交互作用。

再按公式计算相加交互作用S指数的大小，并用Bootstrap法估计其可信区间。Bootstrap再抽样后估计的相加交互作用S指数呈偏态分布，用百分位数法估计其95%*CI*：吸烟(packyr)和住宅氡暴露(radon)的S指数为1.033(1.019～1.057)，吸烟(packyr10)和肿瘤家族史(fcahis)的S指数为1.424(1.156～2.006)。S指数>1且可信区间不包含1，说明吸烟和住宅氡暴露、吸烟和肿瘤家族史的联合作用均大于两因素的单独作用之和。

讨　论

Bootstrap估计的误差包括原样本的抽样误差和再抽样误差两部分。原样本的抽样误差不可去除，所以原样本为总体的随机样本，能够代表总体是Bootstrap得以实施的前提条件。只要重复再抽样的次数R足够大，第二部分再抽样误差就会趋于消失，Efron和Tibshirani[10]认为，R取50～200可使再抽样相对误差不超过5%，并建议R取500～1000以得到可靠的百分位数区间估计。本文R取1000，可以认为有效控制了再抽样误差。

本文分析得出的"packyr"和"radon"的*OR*值是指吸烟量和住宅氡暴露指数每增加一个单位引起肺癌发生的*OR*值，实际应用时还可根据需要分析连续变量每改变5个或10个单位时的*OR*值，只需将原变量值除以5或10得到的数值作为新变量代入logistic回归模型即可，进而用相同的Bootstrap法估计其相加交互作用指标的可信区间，本研究分析吸烟和肿瘤家族史的交互作用时吸烟量(packyr10)的处理即采用了该方法。我们编制了S指数区间估计的R程序，读者根据*RERI*和*AP*的计算公式修改功能函数中的部分程序，可以方便得出*RERI*和*AP*的区间估计。

本文在没有考虑年龄以外的其他混杂因素作用的情况下，发现吸烟和住宅氡暴露之间存在正相加、负相乘的交互作用，即吸烟和氡暴露的联合作用大于二者单独作用之和、小于单独作用之积；发现吸烟和肿瘤家族史之间存在正相加交互作用但无相乘交互作用，即吸烟和肿瘤家族史的联合作用符合相乘模型且大于二者单独作用之和。其内在的作用机制有待于进一步探讨。

参 考 文 献

[1] Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia: Lippincott Williams and Wilkins, 2008: 71-83.

[2] Rothman KJ. Epidemiology: An introduction. New York: Oxford University Press, 2002: 168-180.

[3] Hosmer DW, Lemeshow S. Confidence interval estimation of interaction. Epidemiology, 1992(3): 452-456.

[4] Qiu H, Yu IT, Wang XR, et al. Study on the interaction under logistic regression modeling. Chin J Epidemiol, 2008, 29 (9): 934-937. (in Chinese)
邱宏，余德新，王晓蓉，等. Logistic回归模型中交互作用的分析和评价. 中华流行病学杂志，2008，29(9)：934-937.

[5] Knol MJ, van Der Tweel I, Grobbee DE, et al. Estimating interaction on an additive scale between continuous determinants in a logistic regression model. Int J Epidemiol, 2007, 36(5): 1111-1118.

[6] Assmann SF, Hosmer DW, Lemeshow S, et al. Confidence intervals for measures of interaction. Epidemiology, 1996, 7(3): 286-290.

[7] Chen F, Lu SZ, Yang M. Bootstrap estimation and its applications. Chin J Health Stat, 1997, 14: 5-7. (in Chinese)
陈峰，陆守曾，杨珉. Bootstrap估计及其应用. 中国卫生统计，1997，14：5-7.

[8] http://www.r-project.org/

[9] Yu IT, Chiu YL, Au JS, et al. Dose-response relationship between cooking fumes exposures and lung cancer among Chinese nonsmoking women. Cancer Res, 2006, 66(9): 4961-4967.

[10] Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall, 1993.

(收稿日期：2009-12-07)
(本文编辑：张林东)