·基础理论与方法·

logistic 回归模型中交互作用的分析及评价

邱宏　余德新　王晓蓉　付振明　谢立亚

【导读】 流行病学病因学研究常运用 logistic 回归模型分析影响因素的作用，并利用纳入乘积项的方法分析因素间交互作用，如有统计学意义表示两因素间存在相乘交互作用，但乘积项若无统计学意义并不表示两因素间相加交互作用或生物学交互作用的有无。文中介绍 Rothman 提出的针对 logistic 或 Cox 回归模型的三个评价相加交互作用的指标及其可信区间的计算，并以 SPSS 15.0 软件应用实例分析得出 logistic 回归模型的参数估计值和协方差矩阵，引入 Andersson 等编制的 Excel 计算表，计算相加交互作用指标及其可信区间，用于评价因素间的相加交互作用，为研究人员分析生物学交互作用提供依据。该方法方便快捷，且 Excel 计算表可在线免费下载。

【关键词】 logistic 回归模型；相加交互作用指标；女性肺癌

Study on the interaction under logistic regression modeling

QIU Hong, Ignatius Tak-sun YU, WANG Xiao-rong, FU Zhen-ming, Shelly Lap Ah TSE. Department of Community and Family Medicine, School of Public Health, Chinese University of Hong Kong, H.K.S.A.R
Corresponding author: Ignatius Tak-sun YU, Email: iyu@cuhk.edu.hk

【Introduction】 When study on epidemiological causation is carried out, logistic regression has been commonly used to estimate the independent effects of risk factors, as well as to examine possible interactions among individual risk factor by adding one or more product terms to the regression model. In logistic or Cox's regression model, the regression coefficient of the product term estimates the interaction on a multiplicative scale while statistical significance indicates the departure from multiplicativity. Rothman argues that when biologic interaction is examined, we need to focus on interaction as departure from additivity rather than departure from multiplicativity. He presents three indices to measure interaction on an additive scale or departure from additivity, using logarithmic models such as logistic or Cox's regression model. In this paper, we use data from a case-control study of female lung cancer in Hong Kong to calculate the regression coefficients and covariance matrix of logistic model in SPSS. We then introduce an Excel spreadsheet set up by Tomas Andersson to calculate the indices of interaction on an additive scale and the corresponding confidence intervals. The results can be used as reference by epidemiologists to assess the biologic interaction between factors. The proposed method is convenient and the Excel spreadsheet is available online for free.

【Key words】 Logistic regression model; Indices of interaction on an additive scale; Female lung cancer

多元统计分析中，交互作用是指某因素的作用随其他因素水平的不同而不同，两因素同时存在时的作用不等于两因素单独作用之和(相加交互作用)或之积(相乘交互作用)。目前多采用在回归方程中纳入因素乘积项的方法进行分析。一般认为，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，而 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[1]。若 logistic 回归模型的乘积项系数不等于零且有统计学意义，表示两因素存在相乘交互作用，但若乘积项无统计学意义，并不表示两因素无相加交互作用，也不表示两因素对某疾病的发生无生物学交互作用。Rothman[2,3]，Hosmer 和 Lemeshow[4] 指出 logistic 或 Cox 回归模型中乘积项分析的不足，从理论上系统探讨了交互作用分析指标的构造和算法。向惠云等[5]曾介绍反映相加交互作用的三个指标和可信区间的计算方法，因其计算过程复杂未得到推广使用。本研究拟以 logistic 回归分析为例，介绍利用 SPSS 软件的分析结果进一步计算交互作用的评价指标，并引入 Andersson 等[6]编制的 Excel 计算表估计可信区间，以期为病因学研究中评价因素间的相加交互作用提供简便快捷的方法，亦为研究人员

作者单位：香港中文大学公共卫生学院社区及家庭医学系
通讯作者：余德新，Email: iyu@cuhk.edu.hk

分析生物学交互作用提供依据。

基本原理

以最简单的两因素两水平为例。假设两暴露因子分别为 A、B，1 表示因素存在，0 表示因素不存在，因变量为疾病的发生与否，其他混杂因素暂不考虑。logistic 回归模型得到的 *OR* 值作为相对危险度（*RR*）的估计值。OR_{00} 表示 A、B 都不存在时发病的 *OR* 值，分析时以此为基准，因此 $OR_{00}=1$；OR_{10} 表示仅 A 存在、B 不存在时发病的 *OR* 值；OR_{01} 表示 A 不存在、仅 B 存在时发病的 *OR* 值；OR_{11} 表示 A、B 共同存在时发病的 *OR* 值。

Rothman 和 Hosmer 用于评价相加交互作用的三个指标，即①相对超危险度比（the relative excess risk due to interaction, *RERI*）$= RR_{11} - RR_{10} - RR_{01} + 1$；②归因比（the attributable proportion due to interaction, *AP*）$=$ RERI$/RR_{11}$；③交互作用指数（the synergy index, *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* 的可信区间应包含 0，*S* 的可信区间应包含 1。

Rothman 用于评价相乘交互作用的指标是：$RR_{11}/(RR_{10}\times RR_{01})$，如果两因素无相乘交互作用，则该指标的可信区间应该包含 1。容易证明，此相乘交互作用指标即 logistic 回归模型中乘积项的 *OR* 值。这也进一步说明 logistic 回归模型中乘积项反映的是相乘交互作用。

1. 交互作用指标的点估计：logistic 回归模型估计 OR_{11}、OR_{10} 和 OR_{01} 可通过以下两种方法得到，代入交互作用指标的计算公式即可得该指标的点估计值。

（1）用两因素 A、B 及乘积项 A×B 构建模型 1。

$$\ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 A+\beta_2 B+\beta_3 A\times B$$

$$\ln(OR_{10})=\ln\left(\frac{odds_{10}}{odds_{00}}\right)=\ln(odds_{10})-\ln(odds_{00})$$
$$=\beta_0+\beta_1-\beta_0=\beta_1\Rightarrow\Rightarrow OR_{10}=e^{\beta_1}$$

$$\ln(OR_{01})=\ln\left(\frac{odds_{01}}{odds_{00}}\right)=\ln(odds_{01})-\ln(odds_{00})$$
$$=\beta_0+\beta_2-\beta_0=\beta_2\Rightarrow\Rightarrow OR_{01}=e^{\beta_2}$$

$$\ln(OR_{11})=\ln\left(\frac{odds_{11}}{odds_{00}}\right)=\ln(odds_{11})-\ln(odds_{00})$$
$$=\beta_0+\beta_1+\beta_2+\beta_3-\beta_0=\beta_1+\beta_2+\beta_3\Rightarrow\Rightarrow OR_{11}$$
$$=e^{\beta_1+\beta_2+\beta_3}$$

评价相乘交互作用的指标是：$OR_{11}/(OR_{10}\times OR_{01})=e^{\beta_1+\beta_2+\beta_3}/(e^{\beta_1}\times e^{\beta_2})=e^{\beta_3}$，说明模型 1 中乘积项的 *OR* 值即反映相乘交互作用。

（2）构造新变量 C 并以三个哑变量的形式纳入，构建模型 2（表 1）。

表1 根据两分类变量 A、B 构造新变量 C 和三个哑变量

A	B	C	Dum10	Dum01	Dum11	*OR* 值
0	0	0	0	0	0	OR_{00}
1	0	1	1	0	0	OR_{10}
0	1	2	0	1	0	OR_{01}
1	1	3	0	0	1	OR_{11}

$$\ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 Dum_{10}+\beta_2 Dum_{01}+\beta_3 Dum_{11}$$

$$OR_{00}=1, OR_{10}=e^{\beta_1}, OR_{01}=e^{\beta_2}, OR_{11}=e^{\beta_3}$$

可见，模型 2 中的 β_1、β_2 分别等同于模型 1 中的 β_1、β_2，而 β_3 等于模型 1 中的 $\beta_1+\beta_2+\beta_3$。

2. 交互作用指标的区间估计：运用 Hosmer 和 Lemeshow[4] 介绍的 Delta 方法估计可信区间，计算所需的因素间方差和协方差项可由 SPSS 的 Multinomial 过程选中“Asymptotic Covariance”得到的协方差矩阵代入计算。本研究引用 Andersson 等[6] 编制的 Excel 计算表，输入模型 1 的 β_1、β_2、$(\beta_1+\beta_2+\beta_3)$ 或模型 2 的 β_1、β_2、β_3 以及因素 A、B 间的方差和协方差，可以方便快捷地得到 *RERI*、*AP* 和 *S* 的估计值及其 95% *CI*，进而评价因素间是否具有相加交互作用。

SPSS 软件的 Multinomial logistic 过程用于无序多分类反应变量的 logistic 回归建模，当因变量为两分类时，Multinomial 过程与 Binary logistic 过程得到的参数估计值结果一致，但前者可以给出因素间的协方差矩阵。

实例分析

以香港女性肺癌的病例对照研究资料为例，分析吸烟和癌症家族史在女性肺癌发生过程中有无交互作用（表 2）。为简化计算，暂不考虑其他因素的作用和混杂因素的影响。

构造乘积项 fhisca × smoke，利用 SPSS 软件的 Binary logistic 过程得到模型 1 的参数估计值（表 3）。或构造新变量 fhisca _ sm，以哑变量形式纳入得到模型 2 的参数估计值（表 4）。

用 SPSS 软件的 Multinomial 过程，因变量选择

以 control 作为参照，对自变量 fhisca _ sm 重新编码，定义 fhisca 和 smoke 都不存在的水平为最高水平(因 SPSS 15.0软件中 Multinomial 默认以自变量的最高水平为参照)，并选中 Asymptotic Covariances 估计的方差、协方差矩阵(表 5)。

表2　香港女性肺癌病例对照研究的癌症家族史和吸烟资料

女性肺癌	fhisca (癌症家族史)	smoke(吸烟)	
		有	无
病例	有	120	540
对照		30	460
病例	无	340	1180
对照		230	2240

注：资料为调查所得，表内数据为实际样本量放大 10 倍

表3　模型 1 的 logistic 回归结果

项目	β	*P* 值	*OR* 值(95% *CI*)
fhisca	0.801	<0.001	2.228(1.932～2.571)
smoke	1.032	<0.001	2.806(2.340～3.365)
fhisca × smoke	0.194	0.405	1.214(0.769～1.917)

表4　模型 2 的 logistic 回归结果

fhisca	smoke	fhisca _ sm	β	*OR* 值(95% *CI*)
0	0	0	-	1.00
1	0	1	0.801	2.228(1.932～2.571)
0	1	2	1.032	2.806(2.340～3.365)
1	1	3	2.027	7.593(5.058～11.399)

表5　两因素间的方差和协方差矩阵

项目	fhisca	smoke	fhisca & smoke
fhisca	0.005 32	0.001 29	0.001 29
smoke	0.001 29	0.008 58	0.001 29
fhisca & smoke	0.001 29	0.001 29	0.042 96

将上述模型 1 的 β_1、β_2、$(\beta_1+\beta_2+\beta_3)$或模型 2 的 β_1、β_2、β_3 以及因素间的方差和协方差输入 Excel 计算表(表 6)，可得到 *RERI*、*AP* 和 *S* 的点估计、95% *CI* 及交互作用示意图(图 1)。

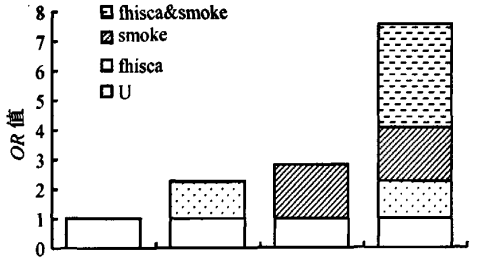
本例模型 1 乘积项 fhisca × smoke 无统计学意义($P=0.405$)，说明两因素无相乘交互作用，癌症家族史和吸烟对香港女性肺癌的发生没有相乘交互作用；Excel 计算表显示 *RERI*、*AP* 的可信区间大于 0，*S* 的可信区间大于 1，图 1 直观显示癌症家族史与吸烟交互作用 *OR* 值的大小，说明癌症家族史和吸烟对香港女性肺癌的发生有相加交互作用(此为协同作用)。*RERI* 和 *S* 意义相同，*AP* 表示全部病例中可归因于两因素交互作用的病例所占的比例，本例 $AP=0.469$，说明全部女性肺癌病例中归因于癌症家族史和吸烟的交互作用所引起的病例占 46.9%。但因本研究分析未考虑其他因素的作用及混杂因素的影响，且分析时为了缩窄可信区间用了实际观察的 10 倍样本量分析，所得结论不一定代表真实情况。

表6　相加交互作用指标 Excel 计算表

	fhisca	smoke	fhisca & smoke
Regr. coefficients	0.801 30	1.031 83	2.027 26
Cov fhisca	0.005 32	0.001 29	0.001 29
Cov smoke	0.001 29	0.008 58	0.001 29
Cov fhisca & smoke	0.001 29	0.001 29	0.042 96

Exposure	*RR*	*Lower*	*Upper*
fhisca	2.228	1.932	2.571
smoke	2.806	2.340	3.365
fhisca & smoke	7.593	5.058	11.399

Measure	*Estimate*	*Lower*	*Upper*
RERI	3.559	0.467	6.650
AP	0.469	0.244	0.693
S	2.173	1.323	3.567



注：U 为不吸烟且无癌症家族史的类别，设为对照组，*OR* = 1

图1　癌症家族史与吸烟交互作用示意图

讨　论

分析交互作用，首先应该清楚统计学交互作用和生物学交互作用的区别[3,7]。统计学交互作用是关于多风险因素的统计模型和参数的定量概念，指在统计模型中纳入乘积项的意义，即随选用模型的不同而不同：线性模型是加法模型，乘积项表示有无相加交互作用，而对于 logistic 或 Cox 等乘法模型，乘积项表示有无相乘交互作用。生物学交互作用是关于多风险因素在发病的生物机制上的定性概念，指两因素皆为病因的前提下是否具有在发病的生物机制上的相互联系，包括协同作用和拮抗作用。生物学交互作用的评价不能等同于统计模型中乘积项的分析。

Rothman[3]认为，生物学交互作用的评价应该

基于相加尺度而非相乘尺度，因此对 logistic、Cox 回归等相乘模型构建了本文介绍的三项指标，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。实际应用中可以用 SPSS 软件计算的模型参数估计值和因素间的协方差矩阵，代入 Andersson 等[6]编制的 Excel 计算表，方便快捷地得出三个相加交互作用指标的点估计和可信区间，为流行病学研究人员分析生物学交互作用提供参考依据。但是相加交互作用是否即表示生物学交互作用，笔者认为还值得商榷，可能是两因素均为病因前提下的相加交互作用可以解释为生物学交互作用。

本研究给出了一个没有相乘交互作用但有相加交互作用的例子，实际分析中还可能碰到有相乘交互作用但不一定有相加交互作用的情况。因为相乘交互作用有统计学意义有两种情况：负相乘（$OR<1$）和正相乘（$OR>1$），有负相乘交互但无相加交互的情况在我们的资料分析中常常可见，也有相关文献为证[8]；而有正相乘交互时，对应的相加交互作用应该也有统计意义，这是从相乘、相加交互的概念上来推论的，还有待证实。

本法适用于两因素两水平时的相加交互作用的评价，当两因素或其中之一为保护因素时，因素变量的编码应以高风险的一类作为暴露，以避免解释上的混乱[3]。例如注射疫苗是某病的保护因素，在分析其与环境因素的交互作用时，将注射疫苗编码为 0，而不注射编码为 1。实例分析多有混杂因素存在，可在拟合 logistic 回归模型时加入混杂因素分析，然后将得到的参数估计值和协方差矩阵代入 Excel 计算表。当因素变量为多分类或连续变量时，三个交互作用指标仍可应用，Knol 等[1]对此有详细阐述。但对于可信区间的估计方法，本研究引入的 Excel 计算表不再适用，Assmann 等[9]提出 Bootstrap 法优于 Hosmer 和 Lemeshow[4]介绍的 Delta 法：Bootstrap 法在原始数据中做重复千次、万次的模拟随机抽样，估计的可信区间更稳定可靠；且当因素为连续变量时，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，对此 Bootstrap 法能做出准确估计而 Delta 法不能。因此用 logistic 回归模型分析两个连续自变量或连续变量与分类变量间的交互作用时，其相加交互作用指标可信区间的估计建议使用 Bootstrap 方法。

参 考 文 献

[1] 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.

[2] Rothman KJ, Greenland S. Modern epidemiology. 2nd eds. Philadelphia: A Wolters Kluwer Company, 1998: 329-342.

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

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

[5] 向惠云，余松林，孙奕，等. 疾病资料多元分析中交互作用指标及可信区间的估计. 中国卫生统计，1999，16：130-133.

[6] Andersson T, Alfredsson L, Kallberg H, et al. Calculating measures of biological interaction. Eur J Epidemiol, 2005, 20: 575-579.

[7] Ahlbom A, Alfredsson L. Interaction: word with two meanings creates confusion. Eur J Epidemiol, 2005, 20: 563-564.

[8] Gustavsson P, Nyberg F, Pershagen G, et al. Low-dose exposure to asbestos and lung cancer: dose-response relations and interaction with smoking in a population-based case-referent study in Stockholm, Sweden. Am J Epidemiol, 2002, 155(11): 1016-1022.

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

（收稿日期：2008-04-28）

（本文编辑：张林东）