·基础理论与方法·

运用广义线性模型探讨福建省汉族人群原发性肺癌影响因素

何斐　肖仁栋　俞婷婷　张鑫　刘志强　蔡琳

【导读】 以福建省汉族人群原发性肺癌环境危险因素研究为实例，通过构建广义线性模型，利用λ值、偏差度及Pearson χ^2拟合优度检验等判断方法选择合适模型，进行多因素及因素间交互作用分析，发现自变量之间存在相乘交互作用，进而选用logistic回归模型分析，最终得到吸烟者有7个、非吸烟者有9个单独作用的肺癌危险因素。其中吸烟者中重度吸烟与肺部罹患过炎症性疾病呈正相乘作用，非吸烟者中被动吸烟与少吃新鲜水果呈正相乘作用。运用广义线性模型筛选合适模型并进行交互作用分析，利于全面合理分析流行病学数据。

【关键词】 广义线性模型；肺癌；交互作用；影响因素

Explore the influence factors on primary lung cancer in Fujian province Han population under the use of generalized linear model

He Fei[1], *Xiao Rendong*[2], *Yu Tingting*[1], *Zhang Xin*[1], *Liu Zhiqiang*[1], *Cai Lin*[1]. 1 *Department of Epidemiology and Health Statistics, School of Public Health*, 2 *Department of Thoracic Surgery, the First Clinical Medical College, Fujian Medical University, Fuzhou* 350108, China

Corresponding author: *Cai Lin*, Email: cailin_cn@hotmail.com

This work was supported by grants from the National Natural Science Foundation of China (No. 81172766, No. 81402738).

【Introduction】 The purpose of this study was to use the data on lung cancer in Han Chinese in Fujian province to explore the value of a generalized, linear model and to investigate the impact related to environment factors on lung cancer as well as the independent and interaction effects on the development of lung cancer. SAS 9.2 was used to build a generalized linear model to evaluate the influence factors and interaction of lung cancer on both smokers and non-smokers. Results showed that the relationship of the factors was multiplied. Under the logistic regression analysis, seven risk factors and nine risk factors were noticed in smokers or in non-smokers, respectively. Heavy smokers and lung diseases appeared a positive multiplying effect on smokers while passive smoking and fresh fruits showed positive multiplying effects on non-smokers. The generalized linear models could filter suitable models thus facilitating further research on the interaction between the two. It seemed easy to carry on the comprehensive and rational analysis on related epidemiological data.

【Key words】 General linear model; Lung cancer; Interaction; Influence factors

目前流行病学数据分析的思路最常见的是先描述研究对象的基本特征，后进行单因素分析，再考虑分层分析，最后为多元回归分析[1]。现今对应变量为二分类或多分类而言，logistic回归分析是最常用的方法。但流行病学研究的疾病，特别是肿瘤等慢性病却是多因素多阶段作用的结果，病因极其复杂，个体行为及环境因素等多方共同作用关系尚不明确。所有因素之间的相互作用形式在分析前无法预知，但分析时却经常直接使用相乘模式探讨自变量间的关系，因而可能无法拟合出最优模型。为此本研究以福建省汉族人群原发性肺癌危险因素为例，探讨流行病病因学研究中数据分析的新思路。

基本原理

针对暴露于发病相关因素的疾病发生率构建广义线性模型，即 $\left(\frac{P}{1-P}\right)^{\lambda}=\beta_0+\beta_1x_1+K+\beta_kx_k$，将其进行Box-Cox幂变换得到方程式[2]

$$P=\begin{cases}\dfrac{e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}(\lambda=0)\\ \dfrac{[1+\lambda(\beta_0+\beta_1x_1+\cdots+\beta_kx_k)]^{(\frac{1}{\lambda})}}{1+[1+\lambda(\beta_0+\beta_1x_1+\cdots+\beta_kx_k)]^{(\frac{1}{\lambda})}}(\lambda\neq0)\end{cases}$$

DOI: 10.3760/cma.j.issn.0254-6450.2015.08.027
基金项目：国家自然科学基金（81172766，81402738）
作者单位：350108 福州，福建医科大学公共卫生学院流行病与卫生统计学系（何斐、俞婷婷、张鑫、刘志强、蔡琳），附属第一医院胸外科（肖仁栋）
通信作者：蔡琳，Email：cailin_cn@hotmail.com

通过估计模型中的尺度参数λ大小判断各发病相关因素间的作用形式。λ=0时，判定因素间作用形式为相乘作用模型$RR(Z)=\prod_{i=1}^{k}RR(x_i)$；λ=1时，判定因素间作用形式为相加作用模型$RR(Z)=\sum_{i=1}^{k}RR(x_i)-(k-1)$；λ>1时，判定因素间作用形式为次相加作用模型$RR(Z)<\sum_{i=1}^{k}RR(x_i)-(k-1)$；λ<0时，判定因素间作用形式为超相乘作用模型$RR(Z)>\prod_{i=1}^{k}RR(x_i)$；0<λ<1时，判定因素间作用形式为中间模型，即超相加或次相乘模型。

运用广义线性模型建立2个连接函数$\ln\frac{P}{1-P}$和$\frac{\left(\frac{P}{1-P}\right)^{\lambda}-1}{\lambda}$，并计算模型拟合的偏差度(deviance)及模型的Pearson χ^2值，通过改变λ的大小选择一系列偏差度中的最小值且同时经Pearson χ^2拟合优度检验显示结果有统计学意义($P<0.05$)此时对应λ值者即为选定的λ值。

实例分析

选取2006年1月至2013年1月福建医科大学附属第一医院、协和医院和南京军区福州总医院胸外科及呼吸内科确诊的1 300例新发原发性肺癌患者为病例组，同期选取1 547名到上述医院其他科室探视患者的亲友以及社区人群，在排除肿瘤病史后按相同性别、年龄(±2岁)频数匹配作为对照组。以调查表面访形式，调查一般情况(性别、年龄、婚姻状况、职业、文化程度及BMI)、室内外环境(居住地污染、房屋类型、住房通风、烹调燃料、烹调油烟、烹调用油热度、装修、吸烟、吸烟包年、戒烟年限、被动吸烟)、饮食(进食速度、食生蒜、烹调用油、新鲜水果/蔬菜、维生素、饮酒、饮茶)、疾病史、家族史、个人因素(生活变故及性格)。采用SAS 9.2软件进行统计学分析。两组在性别、年龄、婚姻状况上分布的差异无统计学意义($P>0.05$)，但文化程度、职业及BMI分布的差异有统计学意义($P<0.001$)，见表1。

表1 两组人群一般特征值

特征分类	病例组(n=1 300)	对照组(n=1 547)	χ^2值	P值
性别			0.13	0.71
男性	935(71.92)	1 103(71.30)		
女性	365(28.08)	444(28.70)		
年龄组(岁)			5.18	0.16
<55	420(32.31)	561(36.26)		
55 ~	450(34.62)	507(32.77)		
65 ~	347(26.69)	380(24.56)		
75 ~	83(6.38)	99(6.40)		
婚姻状况			2.85	0.09
在婚	1 221(93.92)	1 428(92.31)		
未婚及其他	79(6.08)	119(7.69)		
文化程度			81.15	<0.000 1
文盲	198(15.23)	151(9.76)		
初中及以下	776(59.69)	768(49.64)		
高中及以上	326(25.08)	628(40.59)		
职业			98.09	<0.000 1
工人	349(26.85)	470(30.38)		
农民	369(28.38)	262(16.94)		
企事业职员	424(32.62)	708(45.77)		
家务及厨师	60(4.62)	49(3.17)		
其他(待业或无业)	98(7.54)	58(3.75)		
BMI(kg/m^2)			90.76	<0.000 1
18.5 ~ 24.0	798(61.38)	842(54.43)		
<18.5	148(11.38)	68(4.40)		
≥24.0	354(27.23)	637(41.18)		
病理类型				
腺癌	691(53.15)			
鳞癌	370(28.46)			
小细胞癌	109(8.38)			
腺鳞癌	42(3.23)			
大细胞癌	32(2.46)			
其他	56(4.31)			

注：括号外数据为例数，括号内数据为构成比(%)

依据原则确定相对危险度模型的结构形式并拟合方程。即λ=0时，符合相乘模型者采用logistic回归模型分析；λ=1时，符合相加模型者采用线性模型分析；λ≠0且λ≠1时，符合其余作用模型者采用广义线性模型分析[3-4]。从表2和图1的数据及所建立模型的偏差度曲线发现，λ=0时，偏差度值均较小，且Pearson χ^2值无统计学意义($P>0.05$)。据此判定本研究人群吸烟和非吸烟者原发性肺癌影响因素之间的关系偏向于相乘关系，因而采用logistic回归模型进行分析。将所有变量(吸烟者：22个；非吸烟者：21个)纳入进行多重共线性诊断，结果显示变量之间不存在共线性。

对吸烟者，将28个自变量及自变量间两两交互项共同纳入logistic多因素分析模型(逐步回归后退法)，偏差度为1.06，Pearson $\chi^2=1.02$(均$P>0.05$)，$R^2=0.36$。结果提示居住地周围有污染企业、居住平房、住房通风情况较差、烹调过程有油烟、直系亲属有肺癌家族史、不常食用新鲜水果及具有悲观消极性格均是肺癌的危险因素，且呈现单独作用模式；对非吸烟人群，将27个自变量及自变量间两两交互项共同纳入logistic多因素分析模型(逐步回归后退法)，偏差度为1.05，Pearson $\chi^2=0.99$(均$P>0.05$)，$R^2=0.30$。结果提示居住地周围有污染企业、居住

表2　构建模型筛选λ值

λ值	吸烟		非吸烟	
	偏差度	Pearson χ^2值	偏差度	Pearson χ^2值
−3.00	2.15	3.40	3.03	4.73
−2.00	2.08	3.13	2.71	3.66
−1.60	1.98	2.80	1.89	1.68
−1.40	1.92	2.58	1.84	1.61
−1.20	1.84	2.34	1.74	1.46
−1.00	1.25	2.00	1.46	1.11
−0.80	1.68	1.95	1.84	1.69
−0.60	1.51	1.57	1.47	1.16
−0.40	1.25	1.12	1.15	0.89
−0.20	1.13	1.05	1.12	1.01
0.00	1.13	1.03	1.11	1.04
0.20	1.13	1.02	1.11	1.08
0.40	1.15	1.00	1.12	1.11
0.60	1.20	0.92	1.21	1.22
0.80	1.42	1.17	1.32	1.44
1.00	1.42	1.14	1.40	1.65
1.20	1.44	1.07	1.36	1.43
1.40	1.49	1.12	1.40	1.50
1.60	1.55	1.21	1.50	1.86
2.00	1.63	1.31	1.45	1.61
3.00	1.62	1.28	1.64	2.38

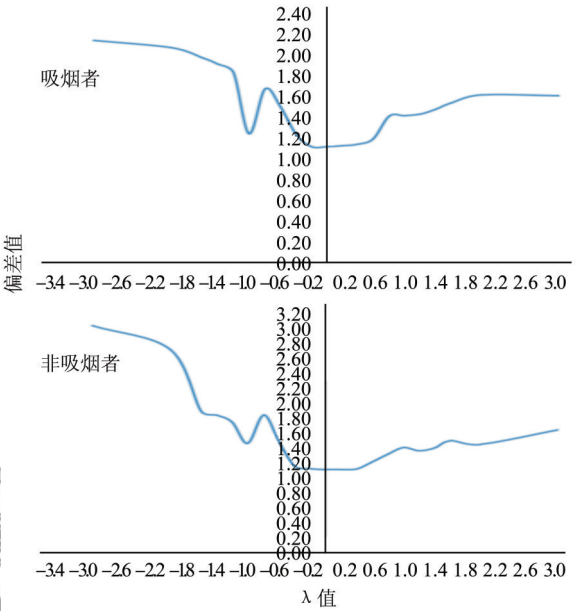

图1　吸烟者和非吸烟者数据拟合因素间相互作用模型的偏差度曲线

平房、住房通风情况较差、进食速度＜10 min/餐、从不服用维生素、烹调过程有油烟、肺部罹患过炎症性疾病、直系亲属有肺癌家族史及不饮茶均是肺癌的危险因素，且呈现单独作用模式（表3）。

对纳入模型的2×2交互作用项有统计学意义且存在交互作用的因素，选用交互作用相对超额危险度（relative excess risk of interaction，*RERI* 或 interaction contrast ratio，*ICR*）、交互作用归因比例（attributable proportion of interaction，*API*）、*S*指数（synergy index，*S*）及其95%*CI*指标估计其联合作用[5-6]，以*ICR*＝0、*API*＝0或*S*＝1即判定为无相加交互作用。结果发现吸烟者中重度吸烟与肺部罹患过炎症性疾病呈正相乘作用，未发现相加交互作用；被动吸烟与新鲜水果对非吸烟人群原发性肺癌呈正相乘作用，未发现相加交互作用（表4）。

讨　论

当前流行病学研究数据分析存在两类意见，一是先进行单因素分析，筛选变量后将有意义变量纳入方程再进行多因素分析，进而考察应变量与自变量及自变量之间的关系；二是把所有因素均放到方程中，以矫正所有可能的混杂因素，直接探讨应变量及自变量之间的关系。应用第一种方法较为普遍，而第二种方法应用时出于拟合模型系数结果的稳定性考虑，一般都提及需要较大样本量（样本量为自变量数量的5～10倍）[7]，但是因为自变量之间可能具有交互作用，分析单一自变量对应变量的贡献时可能会受其他未分析自变量的影响，故先用单因素筛选可能出现偏倚[8]。面对流行病学调查样本量和自变量均增多的情况，涌现出随机森林、树分析、贝叶斯上位效应关联等筛选变量的方法，即通过筛选减少自变量[9]，接近于第二种方法。因此建议在针对当前大样本多变量的流行病学研究中，如计算机软硬件条件允许时，可采用本研究的方法将所有变量纳入方程分析；反之则应先进行变量筛选，再将筛选后的变量全部纳入方程分析。

复杂性状疾病（如肺癌）的发生发展受多因素作

表3　吸烟者和非吸烟者原发性肺癌单独作用影响因素

吸烟者			非吸烟者		
变量	分类	*OR*值(95%*CI*)	变量	分类	*OR*值(95%*CI*)
居住地污染	有	3.25(2.23～4.74)	居住地污染	有	2.61(1.77～3.85)
房屋类型	平房	1.64(1.25～2.14)	房屋类型	平房	1.65(1.25～2.18)
住房通风	较差	2.50(1.72～3.63)	住房通风	较差	1.44(1.03～2.01)
烹调油烟	有	1.47(1.08～2.02)	烹调油烟	有	1.99(1.39～2.84)
肺癌家族史	有	3.91(1.70～9.00)	进食速度	＜10min/餐	1.49(1.06～2.09)
性格	悲观消极	1.94(1.36～2.76)	维生素	不服用	1.52(1.08～2.13)
新鲜水果	不常食用	1.86(1.38～2.52)	饮茶	不饮	2.13(1.61～2.83)
			肺部炎症相关疾病	有	1.92(1.17～3.16)
			肺癌家族史	有	3.26(1.76～6.02)

注：*OR*值经性别、年龄、文化程度、职业及婚姻状况调整

表4　吸烟者和非吸烟者原发性肺癌相加交互作用因素

研究对象	交互项	*RERI*(95%*CI*)	*API*(95%*CI*)	*S*(95%*CI*)
吸烟者	重度吸烟 肺部炎症性疾病	−0.56(−3.69～2.56)	−0.21(−1.53～1.12)	0.75(0.14～4.14)
非吸烟者	被动吸烟 少吃新鲜水果	0.38(−2.92～3.68)	0.13(−0.87～1.13)	1.24(0.21～7.16)

用，在多因素分析控制混杂因素干扰的基础上，需进一步探讨效应修饰因素的影响。Rothman认为效应修饰因素至少存在统计学、生物学和公共卫生学意义上的交互作用形式。在流行病学研究中，常常用相加模型探讨生物学和公共卫生学上的交互和联合作用[10]。本研究在环境危险因素与肺癌关系的多因素分析前，先用比数幂变换的方法探讨危险因素间的联合作用形式，结果显示无论是吸烟还是非吸烟人群，λ=0时，不仅偏差度最小，且模型的拟合优度检验均显示$P>0.05$，应注意当λ=–0.2或0.2时，偏差度也最小，且拟合优度检验结果也显示$P>0.05$，此时建议优先选择λ=0作为判断因素间联合作用形式的依据，毕竟相乘作用下使用的logistic回归模型属于成熟模型，而且便于解释结果。

复杂性状疾病各因素间的作用形式常常不清晰，在未预先判断因素之间作用形式时，直接选择logistic回归模型等相乘结构的统计模型进行分析，极易得到错误结论[11-12]。目前认为当多个因素的共同效应等于每一因素单独效应的乘积时，因素之间联合作用的形式符合相乘模型，此时适宜使用统计学中的相乘模型（如logistic回归模型）；当多个因素的共同效应等于每一因素单独效应相加之和减1时，因素间的联合作用形式符合相加模型，适宜使用统计学中的相加模型（如线性回归模型）；此外如介于相乘和相加模型之间，即次相加至超相乘的中间状态[13]，则推荐通过自定义不同连接函数的方法使用广义线性模型。

广义线性模型是常见的正态线性模型的直接推广[14-15]，可处理多种形式的应变量，而应变量通过连接函数g[E(Y)]与线性预测因子P建立联系，不仅确保线性关系，而且保证预测值落在应变量的各种变化内，可以解决数据过离散的问题。但也有局限性。如果自变量之间的关系可能混合存在相加和相乘时，模型无法进行处理，且模型本身的函数连接形式和参数模型的构建亦有限，尚不能完全涵盖可能出现的情况，反续又发展出广义可加模型。

Rothman等[5]认为，仅通过统计学模型中的相乘项判断交互作用，不能充分挖掘因素间作用的信息，因而提出了相加交互模型的评价指标：*RERI*、*API*和*S*。*RERI*描述归因交互作用对基础危险性的相对大小，*S*与其意义相同，绝对值越大，因素间的交互作用越强，两指标用于评价2个因素同时作用时的联合作用与单独作用之和的差别；*API*则表示因素共同存在时，发生疾病的危险归因于交互作用的比例。这些指标有利于描述因素间的协同或拮抗作用。本研究将变量两两相乘纳入方程后发现，在吸烟者中重度吸烟与肺部炎症性疾病史，非吸烟者中被动吸烟与少吃新鲜水果2个相乘项分别在方程中有统计学意义，进一步探讨该两组因素的相加交互作用，结果*RERI*分别为–0.56和0.38；*API*分别为–0.21和0.13；*S*分别为0.75和1.24（均$P>0.05$），并未发现存在因素间的协同或拮抗作用。有学者提出，分析交互作用时对样本量的要求较高，两因素联合作用的样本量至少是分析每个因素单独作用时样本量的2～4倍[16]，本研究样本量为2 847例，但也未能观察到变量之间的联合作用，考虑与研究中仅分析方程中2×2交互作用项有统计学意义的因素有关，因此建议可将所有变量两两之间组合直接分析相加交互作用，而不经过相乘项进行先期判断，但增加了分析工作量。

综上分析，广义线性模型是筛选自变量与应变量关系的合适模型，并可用于研究自变量之间交互作用，有利于全面合理分析流行病学调查数据。

参 考 文 献

[1] Wang LL, Chen CZ. Way of 'analytical thinking' on data from epidemiological studies [J]. Chin J Epidemiol, 2014, 35(6): 745-748. (in Chinese)
王琳琳，陈常中. 流行病学论文数据分析思路[J]. 中华流行病学杂志，2014，35(6)：745-748.

[2] Guerrero VM, Jhonson RA. Use of the Box-Cox transformantion with binary response models[J]. Biometrika, 1982, 69: 309-314.

[3] Moolgavker SH, Venzon DJ. General relative risk regression models for epidemiology studies[J]. Am J Epidemiol, 1987, 126: 949-961.

[4] Tian J. The Box-Cox transform method of determining λ and SAS program[J]. J Mat Med, 2002, 15(6): 481-484. (in Chinese)
田俊. 比数幂变换的λ确定方法及SAS程序[J]. 数理医药学杂志，2002，15(6)：481-484.

[5] Rothman KJ, Greenland S, Walker AM. Concepts of interaction [J]. Am J Epidemiol, 1980, 112(4): 467-470.

[6] Hosmer DW, Lemeshow S. Confidence interval estimation of interaction [J]. Epidemiology, 1992, 3(5): 452-456.

[7] Kim PH, Chen F. Medical statistical methods[M]. 3rd ed. Shanghai: Fudan University Press, 2009. (in Chinese)
金丕焕，陈峰. 医用统计方法[M]. 3版. 上海：复旦大学出版社，2009.

[8] Hu LP, Li ZJ. Medical statistics and typical errors differentiation [M]. Beijing: Military Medical Science Press, 2003. (in Chinese)
胡良平，李子健. 医学统计学基础与典型错误辨析[M]. 北京：军事医学科学出版社，2003.

[9] Zhang R, Chu M, Zhao Y, et al. A genome-wide gene-environment interaction analysis for tobacco smoke and lung cancer susceptibility[J]. Carcinogenesis, 2014, 35(7): 1528-1535.

[10] Xu B. Epidemiological principles [M]. Shanghai: Fudan University Press, 2007. (in Chinese)
徐飚. 流行病学原理[M]. 上海：复旦大学出版社，2007.

[11] Greenland S. Limitations of the logistic analysis of epidemiologic data [J]. Am J Epidemiol, 1979, 110(6): 693-698.

[12] Liu DS. Comparison among logistic, log binomial and modified Poisson regression model in SAS software [D]. Zhengzhou: Zhengzhou University, 2013. (in Chinese)
刘东升. SAS软件中logistic、log-binomial及稳健Poisson回归模型的比较[D]. 郑州：郑州大学，2013.

[13] Wang XD, Tian J. Generalized linear model analysis of the interaction of factors[J]. Mat Pract Theor, 2010, 40(7): 112-118. (in Chinese)
王晓东，田俊. 因素交互作用分析的广义线性模型[J]. 数学的实践与认识，2010，40(7)：112-118.

[14] Fahrmeir L. Multivariate statistical modeling based on generalized linear models[M]. New York, Springer-Verlag, 1994.

[15] Chen XR. Generalized linear model (1) [J]. Mat Stat Manag, 2002, 21(5): 54-61. (in Chinese)
陈希孺. 广义线性模型(一)[J]. 数理统计与管理，2002，21(5)：54-61.

[16] Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects [J]. Int J Epidemiol, 1984, 13(3): 356-365.

（收稿日期：2015-01-14）
（本文编辑：张林东）