·基础理论与方法·

不完全病例对照研究基因环境交互作用的估计

柏建岭　荀鹏程　赵杨　于浩　沈洪兵　魏庆义　陈峰

【摘要】 **目的**　介绍不完全病例对照研究中基因与环境交互作用的估计方法。**方法**　分别导出了 logistic 模型、对数线性模型在传统病例对照研究、单纯病例研究、不完全病例对照研究中主效应以及基因与环境交互作用效应的极大似然估计，并通过实例分析其应用价值。**结果**　在传统病例对照研究中，当数据未缺失时，logistic 模型与对数线性模型的结果是等价的。当无对照时，单纯病例研究的 logistic 模型可以估计基因与环境的交互作用。当对照组基因信息缺失但环境信息齐全时，用传统病例对照研究的 logistic 模型无法得到交互作用的估计；用单纯病例研究的 logistic 模型可以估计交互作用，但由于没有充分利用环境的信息，故得不到环境主效应的估计；不完全病例对照研究的对数线性模型，可同时得到交互作用和环境主效应的估计。**结论**　不完全病例对照研究采用对数线性模型既可充分利用对照的环境暴露信息，估计环境的主效应，又可估计基因与环境的交互作用。当基因与环境暴露独立时，其估计值与完全数据是等价的。

【关键词】 基因-环境交互作用；不完全病例对照研究；对数线性模型；logistic 模型

Estimation on gene-environment interaction in the partial case-control study　*BAI Jian-ling*, XUN Peng-cheng, ZHAO Yang, YU Hao, SHEN Hong-bing, WEI Qing-yi, CHEN Feng. *Department of Epidemiology and Biostatistics, School of Public Health, Nanjing Medical University, Nanjing* 210029, China

Corresponding author: *CHEN Feng*, Email: dr.chenfeng@163.com

【**Abstract**】 **Objective**　To introduce the approaches for estimating gene-environment interaction based on partial case-control studies. **Methods**　The effects of logistic model and log-linear model for estimating the main effects and gene-environment interaction effect were estimated by means of maximum likelihood methods in traditional case-control studies, case-only studies and partial case-control studies, respectively. An example was also illustrated. **Results**　In traditional case-control study with complete data, the results of logistic model and log-linear model were equivalent. In case-only study without any information about controls, the logistic model can also efficiently estimate gene-environment interaction. In partial case-control study, environmental information was collected from all of the cases and controls, while genetic information was only collected from cases. For this case-control study with incomplete data, a suitable parameterized log-linear model could simultaneously and efficiently estimate the main effect of environment and gene-environment interaction, whereas the logistic model could not. **Conclusion**　For a partial case-control study, log-linear model could estimate not only the main effect of environment but also gene-environment interaction. If genotype and exposure were independent, estimators from partial case-control were as precisely as those from complete-data case-control studies.

【**Key words**】 Gene-environment interaction; Partial case-control study; Log-linear model; Logistic model

经典的研究设计如队列研究、病例对照研究可以用于分析遗传因素与环境因素的交互作用[1]，单纯病例研究(case-only study)估计遗传和环境暴露的交互作用，在暴露和基因独立的条件下可以提高交互作用的检验效能[2-5]。然而，在实际工作中，病例的基因数据容易获得，对照的基因数据往往难以获得，所得资料不完全，称之为不完全病例对照研究(partial case-control study, incomplete case-control

基金项目：国家重点基础研究发展计划(973)资助(2002CB512910)；江苏省高校自然科学基金重点项目资助(04KJB310081)

作者单位：210029 南京医科大学公共卫生学院流行病与卫生统计学系(柏建岭、荀鹏程、赵杨、于浩、沈洪兵、陈峰)；Department of Epidemiology, University of Texas M. D. Anderson Cancer Center, Houston, TX, USA(魏庆义)

通讯作者：陈峰，Email：dr.chenfeng@163.com

study)[6]。对这种资料，采用传统的 logistic 模型分析只能得到环境的主效应；采用单纯病例研究的 logistic 模型分析只能得到基因与环境的交互作用。这两种分析均没有充分利用资料的信息。本文介绍一种对数线性模型(log-linear model)，可以充分利用对照的环境暴露信息估计环境的主效应，并同时估计基因-环境的交互作用[7]。笔者应用极大似然法，分别导出了 logistic 模型、对数线性模型在传统病例对照研究、单纯病例研究、不完全病例对照研究中可估的主效应和基因与环境交互作用效应及各自的标准估计误差(standard error, SE)的估计，并作了对比研究，通过实例分析介绍其应用价值。

基本原理

不失一般性，在病例对照研究中，考虑两种因素：环境因素(E)和基因因素(G)。患者的基因型分为突变基因型($G=1$)和野生基因型($G=0$)；研究对象环境危险因素分为暴露($E=1$)和未暴露($E=0$)。根据基因和环境信息，可以将观察资料整理成表 1 的格式，称为 $2\times2\times2$ 表，或叉生表(cross table)。

表1 病例对照研究数据整理表

暴露	病例($D=1$)		对照($D=0$)		合计
	基因(+)($G=1$)	基因(-)($G=0$)	基因(+)($G=1$)	基因(-)($G=0$)	
+($E=1$)	X_{111}	X_{101}	X_{011}	X_{001}	X_{0+1}
-($E=0$)	X_{110}	X_{100}	X_{010}	X_{000}	X_{0+0}

表 1 中 X_{dge} 表示 D, G, E 的不同组合时的频数。

[情形 1] 表 1 资料齐全。此时常用 logistic 回归模型[8]

$$\text{logit}P(D|G,E)=\mu_1+\alpha_1E+\beta_1G+\gamma_1EG \tag{1}$$

也可以用对数线性模型

$$\log\mu_{dge}=\mu_0+\alpha_0E+\beta_0G+\gamma_0E+\mu_1D+\alpha_1DE+\beta_1DG+\gamma_1DGE \tag{2}$$

其中，μ_{dge} 表示 D, G, E 的不同组合时的理论频数，$\mu_0, \alpha_0, \beta_0, \gamma_0$ 是对照中基因和环境联合分布的参数。不难验证，logistic 回归模型(1)可以通过对数线性模型(3)得到

$$\text{logit}P=(D|G,E)=\log\mu_{1ge}-\log\mu_{0ge} \tag{3}$$

在模型(1)和模型(2)中，α_1 表示暴露的主效应，β_1 表示基因的主效应，γ_1 表示基因和环境交互作用的效应。用极大似然(maximum likelihood, ML)法估计模型中的参数，此时，两个模型对应的参数估计值及其估计误差是等价的。其 ML 估计均为

$$OR_E=\exp(\alpha_1)=\frac{X_{101}X_{000}}{X_{001}X_{100}}$$

$$SE(\alpha_1)=\sqrt{\frac{1}{X_{101}}+\frac{1}{X_{000}}+\frac{1}{X_{001}}+\frac{1}{X_{100}}}$$

$$OR_G=\exp(\beta_1)=\frac{X_{110}X_{000}}{X_{010}X_{100}}$$

$$SE(\beta_1)=\sqrt{\frac{1}{X_{110}}+\frac{1}{X_{000}}+\frac{1}{X_{010}}+\frac{1}{X_{100}}}$$

$$OR_{GE}=\exp(\gamma_1)=\frac{X_{111}X_{001}X_{010}X_{100}}{X_{011}X_{101}X_{110}X_{000}}$$

$$SE(\gamma_1)=\sqrt{\frac{1}{X_{111}}+\frac{1}{X_{001}}+\frac{1}{X_{010}}+\frac{1}{X_{100}}+\frac{1}{X_{011}}+\frac{1}{X_{101}}+\frac{1}{X_{110}}+\frac{1}{X_{000}}}$$

即在情形 1，两个模型的结果是等价的。其中 OR 值的 SE 等于 OR 值与相应的系数的 SE 之乘积。例如 $SE[OR_{GE}]=OR_{GE}\times SE(\gamma_1)$。余雷同。

[情形 2] 只有病例资料而没有对照。此时就是单纯病例研究[2]。相应的 logistic 模型为

$$\text{logit}P(G=1)=\beta_2+\gamma_2E \tag{4}$$

这种资料也可以用对数线性模型表示

$$\log\mu=\alpha_2E+\beta_2G+\gamma_2GE \tag{5}$$

模型(4)和模型(5)中，γ_2 表示基因和环境交互作用的效应。用 ML 估计模型中参数，两模型的结果是等价的。其 γ_2 的 ML 估计为

$$OR_{GE}=\exp(\gamma_2)=\frac{X_{111}X_{100}}{X_{101}X_{110}},$$

$$SE(\gamma_2)=\sqrt{\frac{1}{X_{111}}+\frac{1}{X_{100}}+\frac{1}{X_{101}}+\frac{1}{X_{110}}}$$

即模型(4)与模型(5)是等价的。特别是当 $\frac{X_{001}X_{010}}{X_{011}X_{000}}$ 等于 1 时(此时，对照中基因型与环境暴露无关)，所得到的交互作用的点估计与模型(1)和模型(2)得到的估计是等价的，但 SE 的值要小一些。

在单纯病例对照研究中，只能估计基因与环境的交互作用，而得不到基因或环境主效应的估计。

[情形 3] 具有不完全病例对照资料。在实际工作中，对照组的基因数据往往难以收集，但其暴露信息可以通过调查表获得。如果病例组个体的基因和环境信息齐全，即表 1 中病例组的 4 个频数($X_{111}, X_{101}, X_{110}, X_{100}$)是已知的，对照组个体关于环境暴露的信息齐全，但缺乏基因信息，即此时表 1 中对照组的 4 个频数 $X_{011}, X_{001}, X_{010}, X_{000}$ 是未知的，但其行边缘合计 X_{0+1}, X_{0+0} 是已知的，则

$X_{0+1} = X_{011} + X_{001}$；$X_{0+0} = X_{010} + X_{000}$

这种设计称为不完全病例对照研究。此时，用一般的 logistic 回归模型只能得到环境的主效应，而病例组基因信息未得到利用。采用对数线性模型可以估计环境的主效应，以及基因-环境的交互作用，但不能分析基因的主效应。相应的模型：

$$\log\mu_{dge} = m_0 + \alpha_0 E + m_1 D + \alpha_3 DE + \beta_3 DG + \gamma_3 DGE \qquad (6)$$

用 ML 估计模型中参数，其中，α_3 表示暴露的主效应，其 ML 估计为

$$OR_E = \exp(\alpha_3) = \frac{X_{101} X_{0+0}}{X_{100} X_{0+1}},$$

$$SE(\alpha_3) = \sqrt{\frac{1}{X_{101}} + \frac{1}{X_{0+0}} + \frac{1}{X_{100}} + \frac{1}{X_{0+1}}}$$

不难证明，当 $\frac{X_{001} X_{010}}{X_{011} X_{000}}$ 等于 1 时（此时对照中基因型与环境暴露无关），暴露的主效应

$$OR_E = \frac{X_{101} X_{0+0}}{X_{100} X_{0+1}} = \frac{X_{101} X_{000}}{X_{100} X_{001}}$$

与模型（1）、（2）所得结果相同，但 SE 的值要小一些。

γ_3 表示基因与环境交互作用的效应，其 ML 估计为

$$OR_{GE} = \exp(\gamma_3) = \frac{X_{111} X_{100}}{X_{101} X_{110}},$$

$$SE(\gamma_3) = \sqrt{\frac{1}{X_{111}} + \frac{1}{X_{100}} + \frac{1}{X_{101}} + \frac{1}{X_{110}}}$$

该估计与情形 2 单纯病例研究之结果等价。

综上所述，对数线性模型可以处理三种情形的资料，而 logistic 回归模型只能适用于第 1 种和第 2 种情形。

实例分析

Vandenbroucke 等[9] 采用病例对照研究，调查 324 人，其中病例 155 例，对照 169 例。分析口服避孕药与 Factor V Leiden 等位基因在静脉血栓发生中的作用（表 2）。

对完全资料，分别采用 logistic 模型和对数线性模型，分别估计环境和基因的主效应，以及交互作用（表 3）。可见，无论是参数的点估计，还是其估计误差，两个模型都是对应相等的。即对完全资料，logistic 模型和对数线性模型的结果是等价的。

表2　Factor V Leiden 等位基因与口服避孕药的病例对照研究频数

口服避孕药	病例数		对照例数		合计
	Factor V(+) ($D=1, G=1$)	Factor V(−) ($D=1, G=0$)	Factor V(+) ($D=0, G=1$)	Factor V(−) ($D=0, G=0$)	
+($E=1$)	25	84	2	63	65
−($E=0$)	10	36	4	100	104
合　计	35	120	6	163	169

表3　拟合模型(1)和模型(2)的 OR 值及 95% 可信区间(CI)

因　素	模型(1)		模型(2)	
	OR 值	95% CI	OR 值	95% CI
口服避孕药	3.704(0.948)	2.242～6.118	3.704(0.948)	2.242～6.118
Factor V	6.944(4.324)	2.049～23.534	6.944(4.324)	2.049～23.534
交互作用	1.350(1.320)	0.199～9.171	1.350(1.320)	0.199～9.171

注：括号内数据为 SE 值

假设 $\begin{array}{|c|c|}\hline X_{011} & X_{001} \\ \hline X_{010} & X_{000} \\ \hline\end{array}$ 缺失但 $X_{011} + X_{001}$ 和 $X_{010} + X_{000}$ 已知，可采用单纯病例研究的 logistic 模型（此时丢掉了对照的信息），也可采用不完全病例对照研究的对数线性模型（此时利用了对照中关于环境暴露的信息）。结果见表 4。可见两种模型估计的基因-环境交互作用的结果是等价的。但对数线性模型同时估计了环境的主效应，而单纯病例的 logistic 模型只能估计交互作用。

表4　拟合模型(4)和模型(6)的 OR 值及 95% CI

因　素	模型(4)		模型(6)	
	OR 值	95% CI	OR 值	95% CI
口服避孕药	–	–	3.733(0.949)	2.268～6.146
Factor V	–	–	–	–
交互作用	1.071(0.454)	0.467～2.459	1.071(0.454)	0.467～2.459

注：同表 3

从上述分析不难发现，对数线性模型既可用于完全资料，此时与 logistic 模型结果等价；又可用于不完全资料，此时估计的交互作用等价于单纯病例研究的 logistic 模型，且能估计环境的主效应。需要注意的是，根据不完全资料的对数线性模型得到的点估计 γ_3，等价于根据单纯病例 logistic 模型得到的交互作用点估计 γ_2，但与完全资料所得结果 γ_1 不等，这是因为，虽然在对照中环境暴露与基因型是独立的，但 $X_{011} X_{000} / X_{010} X_{001}$ 一般不会正好等于 1。只有当 $X_{011} X_{000} / X_{010} X_{001}$ 正好等于 1 时，其结果才是等价的。这也是为什么应用中要强调环境暴露与基因型是独立的缘由。

讨　论

传统的病例对照研究是以患有特定疾病的人群

作为病例，以不患该病但具有可比性的人群作为对照，收集两组既往各种可能危险因素的暴露史，测量并比较两组中各因素的暴露比例，并探讨暴露与疾病间是否存在统计学联系的一种流行病学分析研究方法。若要用传统的病例对照研究设计分析基因与环境的交互作用，在收集环境暴露信息的同时，需收集基因信息数据，此时亦可估计遗传与环境各自的主效应。

单纯病例研究是只根据病例的信息估计环境与基因的交互作用，其应用前提是：环境暴露和基因型是相互独立的[10,11]。Piegorsch 等[2]的研究表明，在这样的前提条件下，根据单纯病例研究所得到的交互作用，比病例对照研究所得到的交互作用具有更高的精度，或者在相同的精度要求下，单纯病例研究所需样本量较小[1]。可见单纯病例研究是估计交互作用的有效方法，但不能评价主效应。单纯病例研究设计的应用前提是遗传与暴露相互独立，这种假设对绝大多数遗传与环境因素的研究是合理的[12]。但有些遗传因素，其存在可因生物学代谢而导致暴露相应的偏高或偏低，如饮酒量与醛脱氢酶基因[13]；此外，尚有遗传因素与环境暴露随某一因素（如种族）的改变而同时变化的情形，均不宜采用单纯病例研究设计。

致力于基因-环境交互作用的流行病学家必须充分考虑伦理问题，因为伦理委员会越来越关心基因数据的潜在滥用，在这种气候下，分析的方法如能提高发现交互作用的能力或减少对照的基因数据的滥用是有价值的。实际工作中，对照组基因信息的收集往往比较困难，这是由于对照往往考虑自身不患所研究疾病，对抽血检验这一“创伤”行为配合度较差所致。这种配合度的降低，会导致对照组整体应答率偏低（包括一些环境信息的缺失）。此时，如仅收集对照暴露信息，则可提高对照组的整体应答率，避免对照选择所引起的偏倚，提高精确性，减少成本。这种设计就是不完全病例对照研究。此时利用对数线性模型，可以对暴露的主效应和基因-环境交互作用的效应进行估计。

不完全病例对照研究的应用前提也是假设环境暴露与基因型相互独立。在该前提下，对数线性模型所得到的交互作用的估计，与单纯病例研究得到的估计是等价的。对不完全病例对照研究，对数线性模型所得到的环境的主效应及交互作用的估计是无偏估计（unbiased estimators）。如果该前提不成立，则所得估计就是有偏的。应用时需谨慎。

我们的模拟试验结果表明，在假设环境和基因相互独立的前提下，环境暴露率、基因频率、环境相对风险、基因相对风险、交互作用相对风险以及样本含量一定的条件下，不完全病例对照研究和单纯病例研究对基因-环境交互作用的估计是相同的，且比传统的病例对照研究估计精度要高，可信区间要窄。不完全病例对照研究还可以估计环境的主效应，其精度也高于传统病例对照研究。

对于对照基因信息缺失的不完全病例对照研究，采用传统病例对照研究设计的 logistic 模型只能分析环境的主效应，采用单纯病例设计的 logistic 模型只能分析基因环境的交互作用，而此时若采用对数线性模型分析不完全病例对照研究则可同时分析环境的主效应和基因环境的交互作用，兼有前两者的优势，且不难推广到多个影响因素（协变量）的情形。

参 考 文 献

1 Yang Q, Khoury M, Flanders WD. Sample size requirement in case-only designs to detect gene-environment interaction. Am J Epidemiol, 1997, 146: 713-720.

2 Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med, 1994, 13: 153-162.

3 Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls. Am J Epidemiol, 1996, 144: 207-213.

4 Yang Q, Khoury MJ. Evolving methods in genetic epidemiology. Ⅲ. gene-environment interaction in epidemiologic research. Epidemiol Rev, 1997, 19: 33-43.

5 Andrieu N, Goldstein AM. Epidemiologic and genetic approaches in the study of gene-environment interaction: an overview of available methods. Epidemiol Rev, 1998, 20: 137-147.

6 Liu X, Fallin MD, Kao WH. Genetic dissection methods: designs used for tests of gene-environment interaction. Curr Opin Genet Dev, 2004, 14: 241-245.

7 Umbach DM, Weinberg CR. Designing and analyzing case-control studies to exploit independence of genotype and exposure. Stat Med, 1997, 16: 1731-1743.

8 陈峰. 医用多元统计分析方法. 北京：中国统计出版社，2001.

9 Vandenbroucke JP, Koster T, Briet E, et al. Increased risk of venous thrombosis in oral-contraceptive users who are carriers of factor V Leiden mutation. Lancet, 1994, 344: 1453-1457.

10 易洪刚，陈峰. 单纯病例研究. 国外医学·流行病学传染病学分册，2004，31：60-62.

11 王培稚，沈洪兵，陈峰，等. 文生分析在基因-环境交互作用研究中的应用与意义. 中华流行病学杂志，2005，26：54-57.

12 Hwang SJ, Beaty TH, Liang KY, et al. Minimum sample size estimation to detect gene-environment interaction in case-control designs. Am J Epidemiol, 1994, 140: 1029-1037.

13 Gatto NM, Campbell UB, Rundle AG, et al. Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. Int J Epidemiol, 2004, 33: 1014-1024.

（收稿日期：2005-06-09）
（本文编辑：张林东）