·基础理论与方法·

不完全病例对照研究中对照组部分基因信息缺失基因-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** 探讨不完全病例对照研究中对照组基因信息部分缺失时基因-环境交互作用的估计。**方法** 在 Stata 9.0软件上采用 Monte Carlo 方法模拟不同基因信息缺失比例数据，对缺失数据采用 hot deck 多重填补程序后分析和删除缺失值分析结果进行比较。**结果** 缺失数据＜50％时，hot deck 多重填补后分析和删除缺失值分析对环境主效应、基因主效应以及基因-环境交互作用的估计系数接近完全数据的系数，随缺失比例的增加，两种方法的估计方差均增加，但 hot deck 多重填补估计方差小于删除缺失值分析。**结论** 不完全病例对照研究中，对照组基因信息缺失比例＜50％时，可以用 hot deck 填补方法充分利用已有的信息估计基因-环境的交互作用，提高估计精度。

【关键词】 不完全病例对照研究；基因-环境交互作用；缺失数据

Estimation of gene-environment interaction regarding partial case-control study with missing data on gene information of the controls

*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 discuss the estimation on gene-environment interaction in partial case-control studies when gene information of the controls was partly missing. **Methods** The results of hot deck multiple imputation and listwise deletion analysis were compared when missing data was generated using Monte Carlo method in Stata 9.0. **Results** Coefficients of environment effect, gene effect and gene-environment interaction were respectively estimated by means of hot deck multiple imputation and listwise deletion when approaching to those complete data with missing part less than 50 percent. Both estimated variances of the two methods were increasing with the increased proportion of missing data, but the estimated variance of hot deck multiple imputation was smaller than the one with listwise deletion in each proportion. **Conclusion** Hot deck imputation could be adopted to make full use of existing information to estimate gene-environment interaction in the partial case-control study when missing proportion of gene data of controls was less than 50 percent so as to increase the precision of the estimation.

【**Key words**】 Partial case-control study; Gene-environment interaction; Missing data

复杂疾病(如肿瘤)的发生是由基因和环境因素共同作用的结果[1]。病例对照研究是探讨基因、环境以及基因-环境交互作用的常用方法之一。然而，在实际工作中，病例的基因和环境暴露数据容易获得，对照的环境暴露数据也容易获得，但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 study)[2,3]。对照组基因信息缺失分为两种情况：一是基因信息完全缺失，二是部分基因信息缺失。对第一种情况，笔者已经进行了讨论[3]，采用对数线性模型的方法，可以同时估计环境的主效应和基因-环境交互作用。在部分基因信息缺失的情况下，目前常用的方法是删除含有缺失值的个体，仅采用完全数据进行分析，这种做法损失了信息。

本文针对第二种情况，采用 hot deck 多重填补(multiple imputation, MI)法，在充分利用原始数据的情况下，估计环境的主效应、基因的主效应，以及基因-环境的交互作用，并用 Monte Carlo 模拟试验探讨在不同缺失比例的情况下该估计方法的统计学

基金项目：国家自然科学基金资助项目(30571619)；国家重点基础研究发展计划(973)资助项目(2002CB512910)；江苏省高校自然科学基金重点资助项目(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

性质。

基本原理

1. 多重填补：MI 法是由美国哈佛大学 Rubin[4] 提出，已形成一套比较系统的理论体系。其基本思想是：根据缺失数据的先验分布(prior distribution)，给每个缺失值填补 m($m \geqslant 2$)个填补值(依赖所用的多重填补方法)，构造 m 个"完全"数据集，然后采用相应的完全数据分析法对每一个填补后的新样本进行分析，再综合 m 次分析结果，从而得到未知参数的估计。与一般的缺失数据估计方法相比，多重填补考虑了缺失数据填补的不确定性。

2. 近似贝叶斯自举法：多重填补的方法有多种，如预测均数匹配(predictive mean matching)、趋势得分(propensity score)法、马尔可夫蒙特卡罗(Markov Chain Monte Carlo, MCMC)方法等。本文采用基于近似贝叶斯自举法(approximate Bayesian bootstrap, ABB)的 hot deck 多重填补[5]。

为了方便，定义①全数据集 Y，样本含量为 n，即全部观察值；②完全数据子集 Y_{obs}，样本含量为 n_1，是 Y 中所有具有完整观察数据的个体组成的集合；③不完全数据子集 Y_{mis} 样本含量为 n_0，是 Y 中至少有一个所研究的观察指标数据缺失的个体组成的集合。显然，全数据集由完全数据子集和不完全数据子集组成，即 $Y=(Y_{obs}, Y_{mis})$，$n=n_1+n_0$。

ABB 法估算包括 4 个步骤[6-8]：①在完全数据集 Y_{obs} 中有放回地再抽样(resampling)，样本含量为 n_1，得到数据集 Y^*_{obs}；②在数据集 Y^*_{obs} 中有放回地再抽样，样本含量为 n_0；③用步骤②中得到的 n_0 个观察值替换 Y_{mis} 中的 n_0 个缺失值；④重复上述步骤 m 次，并对 m 次分析结果进行综合。

假设我们感兴趣的是尺度参数的估计 Q，在 logistic 回归中 Q 可以是 OR 的对数值。通过对每一个填补后产生的"完全"数据集进行标准的完全数据分析得到 Q 的估计方差(或标准误)。定义 $Q^{(t)}$ 和 $U^{(t)}$ 分别为第 t($t=1,2,\cdots,m$)个数据集的点估计和估计方差，那么 Q 的点估计、方差等分别为[9]：

点估计(m 个完全数据集估计的均值)

$$\bar{Q}=\frac{1}{m}\sum_{t=1}^{m}Q^{(t)}$$

填补内方差(within-imputation variance)

$$W=\frac{1}{m}\sum_{t=1}^{m}U^{(t)}$$

填补间方差(between-imputation variance)

$$B=\frac{1}{m-1}\sum_{t=1}^{m}(Q^{(t)}-\bar{Q})^2$$

总方差(total variance) $T=W+\left(1+\frac{1}{m}\right)B$

估计系数的 95% CI $\bar{Q}\pm t_{(\upsilon,1-\alpha/2)}\sqrt{T}$

自由度 $\upsilon=(m-1)\left(1+\frac{W}{(1+m^{-1})B}\right)^2$

实例分析

Hwang 等[10] 在 1984－1992 年采用病例对照研究，调查了 Maryland 地区 163 例患儿(其中 69 例单纯性腭裂，114 例唇裂伴或不伴腭裂)和 281 名正常婴儿对照的肿瘤坏死因子(TGF)基因型和母亲吸烟史，分析婴儿 TGF 基因型和母亲吸烟在唇腭裂发生的作用。本文采用了 69 例单纯性腭裂和 281 名正常婴儿对照的 TGF 基因型和其母亲是否吸烟的数据(表 1)。

为说明多重填补的应用，令对照组中个体的基因信息随机产生 30% 的缺失值，数据见表 2。

表1 腭裂与 TGF 和吸烟关系的病例对照研究

吸烟	病例($D=1$)			对照($D=0$)		
	TGF(+) ($G=1$)	TGF(－) ($G=0$)	合计	TGF(+) ($G=1$)	TGF(－) ($G=0$)	合计
+($E=1$)	13	13	26	11	69	80
－($E=0$)	7	36	43	34	167	201
合计	20	49	69	45	236	281

表2 TGF 和吸烟的病例对照研究频数
(基因信息随机产生 30% 的缺失值)

吸烟	病例($D=1$)		对照($D=0$)		
	TGF(+) ($G=1$)	TGF(－) ($G=0$)	TGF(+) ($G=1$)	TGF(－) ($G=0$)	TGF 缺失
+($E=1$)	13	13	10	48	22
－($E=0$)	7	36	24	115	62
合计	20	49	34	163	84

对表 1 的完全数据、表 2 缺失删除后数据、表 2 缺失数据和 hot deck 填补后数据，均采用 logistic 回归分析，结果见表 3。

表3 基因信息随机产生 30% 的缺失值的参数估计

	E	G	$G\times E$
原数据	－0.1347 (0.3538)	－0.0460 (0.4539)	1.8822 (0.6821)
对照组基因信息缺失 30%			
删除缺失数据法	－0.1448 (0.3664)	－0.0707 (0.4701)	1.6393 (0.7040)
多重填补法($m=5$)	－0.1407 (0.3539)	－0.0810 (0.4530)	1.9171 (0.6815)

为了说明某统计方法的可靠性和稳定性，统计学上常采用模拟试验的方法，即在不同条件下模拟产生若干个(通常为 1000 个)数据库，对同一数据库用不同的方法进行分析，然后根据模拟的结果分析不同条件下各种方法所得结果的统计学性质，包括估计系数的均值是否接近真值，估计系数的方差有多大等。为了说明多重填补方法的可靠性和稳定性，本文采用模拟试验的方法，在 Stata 9.0 软件上分别探讨在随机缺失不同比例的情况下，多重填补估计结果的统计性质，并与完全数据的估计结果相比较。

1. 模拟设计思路：分别以表 1 资料对照组的基因信息随机缺失 5%、10%、20%、30%、50% 和 60%(不同比例下种子数不同)，得到新的数据，然后对该数据分别用删除缺失值方法和 hot deck 多重填补法建立 logistic 模型。重复上述过程 1000 次。计算 1000 个回归系数点估计($\bar{Q}_1, \bar{Q}_2, \cdots, \bar{Q}_{1000}$)的均数和总方差的均数($T_1, T_2, \cdots, T_{1000}$)，比较两种方法所得参数的估计值。

2. 模拟运行：结果见表 4。其中，完全数据的估计结果列于表 4 第 1 行。以原完全数据的分析结果为真值，从表 4 可以看出，当对照组基因数据缺失 5%～10% 时，hot deck 多重填补法和删除缺失值法对环境主效应、基因主效应以及基因-环境交互作用的估计系数和估计方差均接近真值；当对照组基因数据缺失 10%～50% 时，hot deck 法对环境主效应估计方差很接近真值，而删除缺失值法的估计方差随着缺失比例的增加而增加，并均高于 hot deck 法；当对照组基因数据缺失＞50% 时，两种方法对基因-环境交互作用的估计与真值偏差均较大。

讨　论

数据缺失是流行病学研究中经常遇到的现象。目前常用的方法是删除含有缺失值的记录。一般认为，在随机缺失的情况下，该法所得到的估计仍然是无偏的(unbiased)。但删除缺失值丢弃了记录中其余的有用信息，造成资源的浪费，当缺失较多时，估计误差较大，影响分析结果的准确性和精确性。如果完全观测与不完全观测之间存在系统差异，则使用完全观测进行的分析就不能代表原来的整体人群，有可能会得到错误的结果。此时多重填补法所得估计仍然是有效的。

本研究结果显示当缺失＜10% 时，删除缺失值法与 hot deck 多重填补估计结果很接近；当缺失比例较大时，hot deck 多重填补精确性优于删除缺失值法；当缺失比例＞50% 时，删除缺失值法和 hot deck 多重填补法均不理想。笔者建议，当研究中缺失＜10% 时，可不必采用缺失值的估计；当缺失比例在10%～50% 之间时，采用多重填补法可以充分利用原始资料的信息；而当缺失比例＞50% 时，即使用 hot deck 多重填补也难以得到满意的效果。

单个填补(one step imputation)没有考虑到缺失数据的不确定性以及填补数据与观察到的数据间可能存在的系统性差异，所以难以提供关于总体参数的准确估计[11]。而多重估计弥补了单个填补的缺陷。与所有的缺失数据估计方法相同，多重填补并没有增加原样本以外的信息，但体现了缺失数据的不确定性，并充分利用了原资料的信息。

表4 1000 次模拟结果中模型回归系数点估计的均数和方差的均数

项目	方法	E	G	$G\times E$
完全数据		−0.1347(0.3538)	−0.0460(0.4539)	1.8822(0.6821)
缺失 5%	删除缺失值法	−0.1358(0.3553)	−0.0430(0.4561)	1.8844(0.6880)
	hot deck 法	−0.1342(0.3538)	−0.0425(0.4540)	1.8827(0.6828)
缺失 10%	删除缺失值法	−0.1339(0.3570)	−0.0453(0.4583)	1.8849(0.6942)
	hot deck 法	−0.1344(0.3538)	−0.0444(0.4541)	1.8864(0.6832)
缺失 20%	删除缺失值法	−0.1331(0.3610)	−0.0451(0.4637)	1.8892(0.7093)
	hot deck 法	−0.1352(0.3538)	−0.0463(0.4541)	1.8997(0.6848)
缺失 30%	删除缺失值法	−0.1346(0.3660)	−0.0388(0.4708)	1.8987(0.7291)
	hot deck 法	−0.1345(0.3538)	−0.0327(0.4547)	1.9093(0.6879)
缺失 50%	删除缺失值法	−0.1322(0.3823)	−0.0441(0.4926)	1.9343(0.7933)
	hot deck 法	−0.1369(0.3538)	−0.0222(0.4556)	1.9685(0.6985)
缺失 60%	删除缺失值法	−0.1252(0.3960)	−0.0276(0.5121)	1.9176(0.8455)
	hot deck 法	−0.1317(0.3540)	−0.0075(0.4565)	1.9685(0.7058)

hot deck 多重填补将 EM 算法和极大似然法的优点与单个填补结合起来，从而产生原始数据的矩阵。hot deck 多重填补和一般 hot deck 填补最大的不同是需要构造 5～10 个“完全数据集”。hot deck 多重估算是采用近似贝叶斯自举法进行再抽样的，属于非参数方法，对数据的分布要求不高，更适用于离散型数据和分布不明的数据[12]。

当某个数据集中含有缺失值的变量比例比较大时，即使采用 hot deck 估算，结果可能也不太理想。另外，hot deck 估算不适用于非随机缺失的数据。

参 考 文 献

[1] Spinka C, Carroll RJ, Chatterjee N. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. Genet Epidemiol, 2005, 29:108-127.

[2] 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.

[3] 柏建岭，荀鹏程，赵杨，等. 不完全病例-对照研究基因-环境交互作用的估计. 中华流行病学杂志，2006，26:72-75.

[4] Rubin DB. Inference and missing data. Biometrika, 1976, 63: 581-592.

[5] Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association, 1986, 81: 366-374.

[6] Allison PD. Missing data (quantitative applications in the social sciences). Thousand Oaks, CA: Sage, 2001.

[7] Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley, 1987.

[8] Rubin DB. The Bayesian bootstrap. Ann Statist, 1981, 9: 130-134.

[9] Royston P. Multiple imputation of missing values. The Stata Journal, 2004, 4: 227-241.

[10] Hwang SJ, Beaty TH, Panny SR, et al. Association study of transforming growth factor alpha (TGF alpha) Taq I polymorphism and oral clefts: indication of gene-environment interaction in a population-based sample of infants with birth defects. Am J Epidemiol, 1995, 141: 629-636.

[11] 曹阳，谢万军，张罗漫. 多重填补的方法及其统计推断原理. 中国医院统计，2003，10:77-81.

[12] 任金马，赵杨，陈峰，等. 配比设计中缺失数据的 hot-deck 估算. 中国卫生统计，2004，21:303-306.

（收稿日期：2007-01-05）

（本文编辑：张林东）

·疾病控制·

天津市主要河道中麦穗鱼华支睾吸虫囊蚴感染情况调查

邸宝华 李昕 徐亚东

华支睾吸虫（肝吸虫）是一种主要分布于亚洲东部地区的人体寄生虫，感染人体的阶段是囊蚴，寄生于第二中间宿主淡水鱼、虾体内，其中以麦穗鱼为主要宿主。为了解天津地区麦穗鱼感染情况，于 2006 年 10－12 月对天津市 8 条主要河道麦穗鱼感染情况进行了调查。麦穗鱼标本采用鱼肉压片法[1]。分别捕获滦河、潮白河、浑河、州河、于桥水库、西青区王兰庄稻地沟河、子牙河、海河的麦穗鱼各 50 条，总计 450 条，通过显微镜检查，其中华支睾吸虫阳性麦穗鱼 29 条，阴性 421 条，总感染率 6.44%。阳性样本的地区分布：天津市 8 条主要河道中，只有流经蓟县的州河与于桥水库有感染的样本，感染率分别为 38%，与 5 年前的调查相比[2]，蓟县州河感染率 100%，二者比较差异有统计学意义（$P<0.01$）；蓟县于桥水库感染率 20%，与 5 年前 92.50% 相比，差异有统计学意义（$P<0.01$），而其他河道中的麦穗鱼感染率均为 0。

调查结果显示，与 5 年前相比天津市 8 条河流麦穗鱼感染率有了明显的下降[2]，而且只有蓟县一个地方有麦穗鱼感染，上次调查蓟县的于桥水库和州河的感染率分别为 92.50% 和 100%；而本次的感染率有明显下降。华支睾吸虫流行的关键因素是人群是否有生吃或半生吃鱼肉的习惯，在烧、烤、烫或蒸全鱼时，可因温度不够、时间不足或鱼肉过厚等原因，未能杀死全部囊蚴[3]。本次调查结果也充分说明了，天津市在这几年的城市建设中对于环境的保护做出的努力，市内的河道水质有了明显的改善以及与天津市民环保意识的加强也有很大的关系。

参 考 文 献

[1] 殷国荣，叶彬. 医学寄生虫学实验指导. 北京：科学出版社，2004：2.

[2] 王毅，曹金钟，祁妙，等. 天津市淡水鱼华支睾吸虫囊蚴感染情况调查. 中国寄生虫病防治杂志，2002，15(6)：插页 2.

[3] 李雍龙. 人体寄生虫学. 6 版. 北京：人民卫生出版社，2004：6.

（收稿日期：2007-04-12）

（本文编辑：尹廉）

作者单位：300203 天津医科大学医学检验学系