·基础理论与方法·

观察性研究中的logistic回归分析思路

冯国双[1,2]

[1]国家儿童医学中心 首都医科大学附属北京儿童医院 大数据和工程研究中心，北京 100045；[2]北京航空航天大学/首都医科大学 北京大数据精准医疗高精尖创新中心，北京 100083

通信作者：冯国双，Email：glxfgsh@163.com

【摘要】 Logistic回归是流行病学研究中常用的方法，然而在实际分析时往往只考虑数据，而不考虑设计，容易导致一些误导性结果和结论。本文主要结合观察性研究的设计目的，探讨logistic回归分析中的具体思路，为logistic回归的实际应用提供借鉴。

【关键词】 Logistic回归；病例对照研究；队列研究

基金项目：北航-首医大数据精准医疗高精尖创新中心计划（BHME-201901）

DOI：10.3760/cma.j.issn.0254-6450.2019.08.025

Logistic regression analysis in observational study

Feng Guoshuang[1,2]

[1]Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Big Data and Engineering Research Center, Beijing 100045, China; [2]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University/Capital Medical University, Beijing 100083, China

Corresponding author: Feng Guoshuang , Email: glxfgsh@163.com

【**Abstracts**】 Logistic regression has been recognized as a commonly used method in epidemiological studies. However, in practice, many people only consider 'data' rather than 'study design' as important issue when working on the analysis, which may easily lead to some misleading results and conclusions. Based on the purpose of observational research during the design of the study, this paper discusses the specific ideas in logistic regression analysis, and provides references for the practical application when logistic regression method is used.

【**Key words**】 Logistic regression; Case-control study; Cohort study

Fund program: Beijing University and Capital Medical University Advanced Innovation for Big Data-Based Precision Medicine Plan

DOI:10.3760/cma.j.issn.0254-6450.2019.08.025

观察性研究在研究设计中占有非常重要的地位，实际应用中比较常见的是病例对照研究和队列研究。尽管其应用广泛，但在数据分析中却存在不少问题。在分析时往往只考虑数据本身，而未能结合研究类型，从而导致结果的偏倚。甚至在已发表的文章中，也存在一些不严谨用语。本文从观察性研究的类型出发，基于不同研究类型的研究目的，以logistic回归分析为例，探讨观察性研究的不同分析思路，希望为医学科研工作者提供一定的参考和借鉴。

1. logistic回归：假定有m个自变量$x_1, x_2, \cdots, x_m$，logistic回归模型的基本形式可表达为：

$$\text{log it}(P)=\ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_mx_m$$

只从数据本身考虑的话，logistic回归模型都是包括一个分类因变量及若干自变量（可以是分类变量，也可以是连续变量），反映了m个自变量对因变量的线性影响。无论对于病例对照研究还是队列研究，这种形式都是不变的。

部分研究在数据分析时，忽略了前期的设计思路，只是简单地把因变量和所有自变量纳入统计软件中相应位置，点击运行直接给出结果。从数据上来看，病例对照研究和队列研究的数据形式完全一样，软件操作过程也并无不同，都是指定因变量和自变量，然后给出参数估计值及统计检验结果。统计软件无法判断研究者采用的是病例对照研究还是队列研究，也并不清楚作者的主要研究目的是什么，只是对指定的变量进行参数估计。而统计分析的思路需要根据研究目的和研究类型而定，对于病例对照研究或队列研究而言，它们的分析思路显然不同。一味依靠统计软件，不仅容易出现一些错误分析思路，也会导致错误的分析结果。

2. 病例对照研究中的logistic回归：从数据分析

的角度来看，病例对照研究大致有两大类目的：一是探索危险因素，二是验证危险因素。

（1）以探索危险因素为目的的分析思路：危险因素的探索常见于临床研究中，通常用于研究初期，此时研究者并不清楚哪些因素可能会影响结局的发生，因此先进行初步探索。根据专业知识和经验收集一些可能的指标，然后从中寻找可能对结局影响较大的因素。例如，探索儿童打鼾的危险因素，研究者并无太多的前期基础，只是为了发现可能与儿童打鼾有关的因素，这种情况下会根据文献报道、专业经验等收集一些可能有关的指标，并从中找出与儿童打鼾有关的部分因素。

危险因素探索的文章中，最常见的表述错误是“校正其他混杂因素”后，发现共k个变量对结局有影响。混杂因素是相对主要研究因素而言，而危险因素探索的研究中，并无明确的主要研究因素，所有变量都是待研究的因素，目的是从这些变量中找出哪些有影响。此类研究中，“校正其他混杂因素”是一种不严谨的表达方式。

对于这种分析思路，需要有一定的分析经验和技巧。实际分析中，需要考虑的几个问题：

①线性问题：由于logistic回归本质上仍属于“线性模型”，因此一定要确认自变量与因变量（logit P）之间是否线性关系，如果不是，需要考虑进行相应的变换，否则可能会产生错误结果。

例1：某研究分析老年人高血压（二分类变量，是或否）的危险因素，研究因素包括gender、age、ox-LDL、Adiponectin、ox-LDL IgG和ox-LDL IgM共6个指标。其中gender为二分类变量，其余变量均为连续变量。如果把6个自变量直接纳入统计软件分析，所得结果见表1。

表1 统计软件直接给出的高血压影响因素分析结果

指标	参数估计值	标准误	t值	P值
sex	−0.513	0.555	−0.93	0.358
age	0.010	0.038	0.25	0.802
ox-LDL	0.001	0.012	0.10	0.922
ox-LDL IgM	0.043	0.033	1.31	0.195
Adiponectin	−0.008	0.026	−0.32	0.749
ox-LDL IgG	−0.745	0.471	−1.58	0.118

可以看出，6个变量均差异无统计学意义。然而对数据重新分析后发现，并不是这些变量对结局均无影响，只是未能发现它们之间的真实关系而已。经仔细观察，发现age和ox-LDL IgM对结局的影响是有统计学意义的，但不是线性影响，而是二次项关系（表2）。

表2 高血压影响因素重新分析后的结果

参数	参数估计值	标准误	$Wald\chi^2$值	P值
age	2.157	0.608	12.58	0.000
age*age	−0.020	0.006	12.57	0.000
ox-LDL IgM	0.463	0.183	6.42	0.011
ox-LDL IgM*ox-LDL IgM	−0.007	0.003	5.84	0.016

②共线性问题：共线性即自变量之间存在高度相关，从而导致结果不可靠[1]。共线性是大多数回归模型都需要考虑的一个问题，一旦发现该问题，需要采取不同措施来解决。常见的解决方案包括删除某一自变量、主成分分析、Lasso回归等。

例2：某研究分析乳腺增生的危险因素，自变量同时包括妊娠次数（三分类变量，用1、2、3表示相应次数）和流产次数（三分类变量，用0、1、2表示相应次数）。在单因素分析中妊娠次数差异有统计学意义（2 *vs.* 1，P=0.026；3 *vs.* 1，P=0.035），然而多因素分析中则差异无统计学意义（P值分别为0.635、0.594）。分析原因发现，主要是由于妊娠次数和流产次数有较强的共线性，二者相关系数高达0.55，从而导致妊娠次数变得无统计学意义。解决方案采用了删除法，删除妊娠次数变量，保留了流产次数变量。

③单因素和多因素的问题：目前危险因素筛选的一种分析思路：先进行单因素分析，将单因素分析中差异有统计学意义（P<0.05）的变量再纳入多因素分析，选出最终有统计学意义的变量作为危险因素。然而这一思路并非十分可靠，有些情况下可能会出现单因素分析无统计学意义而多因素分析有统计学意义的情况，此时就容易漏掉某些重要的因素。

表3 不同血清学指标的胃癌发生情况

$x1$	$x2$	非胃癌	胃癌
阳性	阳性	2	18
	阴性	14	4
阴性	阳性	12	14
	阴性	12	2

例3：某研究分析两个血清学指标（分别用阳性和阴性表示）对胃癌的影响，数据结果见表3。

该数据采用单因素分析的话，可以发现$x1$差异无统计学意义（P=0.114），而在多因素分析中却变得有统计学意义（P=0.018）。如果只将单因素分析中有统计学意义的变量纳入多因素分析的话，就会漏掉$x1$变量。为什么会出现这种情况，主要是因为$x1$和$x2$之间存在负相关，而$x1$、$x2$与结局之间均为正相关。因此，数据分析过程中，不要盲目套用所谓的“分析套路”，而应结合实际情况具体问题具体分析。

总之，在筛选危险因素时，建议不要仅将单因素分析有统计学意义的变量纳入多因素分析，一定要厘清变量之间的关系，否则容易遗漏重要的变量或纳入无意义的变量。

（2）以验证危险因素为目的的分析思路：验证危险因素，说明研究者在研究开始时已经有明确的主要研究因素，主要目的是为了验证该因素是不是真正的影响因素。基于这种目的，研究者在设计时会突出主要因素，但同时也会收集其他可能的混杂因素。例如，探索肺癌与吸烟的关系，吸烟是主要研究因素，因此问卷调查中会详细设置各种与吸烟有关的问题。考虑到其他因素可能也会影响肺癌发生，因此调查时也会加入其他有关因素的调查，但这些因素不是研究者关心的，只是为了校正这些因素，以便真正明确吸烟与肺癌的关系。

因此，对于这种研究目的关键的问题是，如何控制混杂因素，以便真正明确主要研究因素与结局的关系。混杂因素在流行病学中已有详细定义[2]，不再赘述。从数据分析的角度来看，要判断一个因素是否为混杂因素，可以从两个方面来考虑：第一，分析该因素是否对结局有较大影响，通常可采用χ^2检验或单因素logistic回归来实现；第二，分析该因素在主要研究因素中的分布情况，通常采用χ^2检验来实现。

例4：某研究分析性别与幽门螺杆菌（Hp）的关系，现在考虑吸烟是否为影响二者关系的混杂因素。具体数据见表4。

表4　不同性别、吸烟状况的幽门螺杆菌（Hp）感染情况

性别	吸烟状况	Hp阳性	Hp阴性
男	是	146	343
	否	64	258
女	是	18	61
	否	110	530

首先分析吸烟对结局的影响，采用χ^2检验或单因素logistic回归不难发现，吸烟人群与不吸烟人群相比，Hp阳性的风险更高（OR=1.84，95%CI：1.44～2.35）。其次分析吸烟在性别中的分布，χ^2检验结果显示，男性和女性中吸烟的比例差异有统计学意义（χ^2=396.97，P<0.001），男性的吸烟比例远高于女性。

由此看出，以性别作为主要分析变量，在分析性别与Hp感染时，吸烟可能是影响二者关系的混杂因素，必须加以校正。校正前结果显示，性别对Hp的影响有统计学意义，男性有更高的Hp阳性风险（OR=1.62，95%CI：1.26～2.07）；校正后发现，性别对Hp的影响无统计学意义（OR=1.26，95%CI：0.94～1.68）。

因此，对于以验证危险因素为目的的logistic回归分析，分析思路主要是明确哪些因素可能是混杂因素并加以校正，以发现主要研究因素与结局的真实关系。建议尽量避免的两种思路：①把所有变量都进行校正。除非样本量足够大，否则这种方式不可取。因为纳入的自变量越多，所消耗的自由度越大，用于估计主要研究因素的样本量相对越小，结果的精确度也越低。②采用逐步回归筛选变量。作为主要研究变量，一定要保留在模型中，同时要纳入混杂因素。逐步回归筛选适用于探索危险因素，不适用于验证危险因素。

3. 队列研究中的logistic回归：队列研究绝大多数都是为了验证某一危险因素，这是由研究性质决定的。因为队列研究在一开始就需要指定暴露和非暴露，也就相当于确定了主要研究因素。因此，从数据分析角度来讲，队列研究主要是为了排除混杂因素，与前文介绍的思路并无不同。但队列研究在时间顺序上可以证明研究因素发生在前，结局发生在后，因此其验证能力更强。

由于队列研究有明确的时间先后顺序，此时在说明主要研究因素与结局的关联强度时，可采用RR（risk ratio）而非OR（odds ratio）。队列研究中，当结局发生率很低时（<10%），OR是RR的一个很好的替代指标，此时用logistic回归可直接求得OR值，用来说明暴露的危险度。但如果结局发生率不是很低，OR与RR差别较大，此时用OR来说明危险度可能会有一定的偏倚[3]。

例5：某研究分析Hp感染与胃黏膜病变进展的关系，观察数据见表5。

表5　不同幽门螺杆菌（Hp）感染状况的胃黏膜病变进展

Hp感染状况	病变进展	病变未进展
Hp阳性	29	25
Hp阴性	34	12

本研究如果计算OR值则，OR=2.44（95%CI：1.05～5.70），如果计算RR值则，RR=1.77（95%CI：1.01～3.12）。由于病变进展的发生率较高，两个指标差别较大。

队列研究中RR值的计算通常可采用对数二项分布回归（log-binomial regression）。通常需要借助软件实现，如SAS的proc genmod过程[4]。

4. 小结：本文介绍了病例对照研究和队列研究中logistic回归分析的不同思路，以及常见的一些应

用错误。然而本文的思路并不仅限于logistic回归分析，完全可以推广到其他广义线性模型。例如，队列研究的观察结局如果是计数资料，则可考虑Poisson回归或负二项回归，此时仍需考虑混杂因素的校正问题。因此，本文思路对各种常见的回归模型均有一定借鉴意义，至于模型的选择主要取决于研究结局类型及其分布。

在各种常见的回归分析中，一定要分清研究类型及其目的，到底是探索危险因素还是验证危险因素。危险因素的筛选过程较为复杂，需要考虑较多问题，包括变量筛选方式等；验证危险因素相对较为简单，不需要考虑变量筛选，但要明确混杂因素并加以校正。一定要避免“把数据完全交给软件”这种分析方式，软件主要用来解决计算问题，分析思路必须由研究者来确定。统计分析不是简单的参数估计，而应结合研究类型，明确研究思路，才能给出合理的结果。

利益冲突 所有作者均声明不存在利益冲突

参 考 文 献

[1] Mennard S. Applied logistic regression analysis[M]. Newbury Park, California: SAGE Publications, Inc., 2001.

[2] 徐飙. 流行病学原理[M]. 上海：复旦大学出版社，2007.
Xu B. Epidemic theory[M]. Shanghai: Fudan University Press, 2007.

[3] Stokes ME, Davis CS, Koch GG. Categorical data analysis using the SASsystem[M]. 2nd ed. Cary, NC: John Willy & Sons, Inc., 2000.

[4] 冯国双，刘德平. 医学研究中的logistic回归分析及SAS实现[M]. 2版. 北京：北京大学医学出版社，2015.
Feng GS, Liu DP. Logistic regression analysis and SAS application in medical research[M]. 2nd ed. Beijing: Peking University Medical Press, 2015.

（收稿日期：2019-03-28）
（本文编辑：王岚）

读者·作者·编者

本刊常用缩略语

本刊对以下较为熟悉的一些常用医学词汇将允许直接用缩写，即在文章中第一次出现时，可以不标注中文和英文全称。

OR	比值比
RR	相对危险度
CI	可信区间
P_n	第*n*百分位数
AIDS	艾滋病
HIV	艾滋病病毒
MSM	男男性行为者
STD	性传播疾病
DNA	脱氧核糖核酸
RNA	核糖核酸
PCR	聚合酶链式反应
RT-PCR	反转录聚合酶链式反应
*Ct*值	每个反应管内荧光信号达到设定的阈值时所经历的循环数
PAGE	聚丙烯酰氨凝胶电泳
PFGE	脉冲场凝胶电泳
ELISA	酶联免疫吸附试验
*A*值	吸光度值
GMT	几何平均滴度
HBV	乙型肝炎病毒
HCV	丙型肝炎病毒
HEV	戊型肝炎病毒
HBcAg	乙型肝炎核心抗原
HBeAg	乙型肝炎e抗原
HBsAg	乙型肝炎表面抗原
抗-HBs	乙型肝炎表面抗体
抗-HBc	乙型肝炎核心抗体
抗-HBe	乙型肝炎e抗体
ALT	丙氨酸氨基转移酶
AST	天冬氨酸氨基转移酶
HPV	人乳头瘤病毒
DBP	舒张压
SBP	收缩压
BMI	体质指数
MS	代谢综合征
FPG	空腹血糖
HDL-C	高密度脂蛋白胆固醇
LDL-C	低密度脂蛋白胆固醇
TC	总胆固醇
TG	甘油三酯
COPD	慢性阻塞性肺疾病
CDC	疾病预防控制中心
WHO	世界卫生组织