中国儿童保健杂志
Chinese Journal of Child Health Care
ISSN 1008-6579,CN 61-1346/R

《中国儿童保健杂志》网络首发论文

题目：Logistic 回归应用的常见问题及其注意事项
作者：李晨，张杨，陈长生
网络首发日期：2020-01-17
引用格式：李晨，张杨，陈长生．Logistic 回归应用的常见问题及其注意事项[J/OL]．中国儿童保健杂志.
http://kns.cnki.net/kcms/detail/61.1346.R.20200116.1007.020.html

Logistic 回归应用的常见问题及其注意事项

李晨，张杨，陈长生

空军军医大学军事预防医学系卫生统计学教研室，陕西 西安 710032

关键词： Logistic 回归；分类资料；优势比

文献标识码：A doi:10.11852/zgetbjzz2019-0012

医学研究尤其是流行病学研究中，常见分析疾病（结局）与多种因素（暴露）之间的定量关系。当结局为分类（二分类、多分类）资料时，为研究多种因素共同作用及其交互作用对结局的定量影响，可以采用 logistic 回归（logistic regression）分析方法。Logistic 回归属于概率型非线性回归的一种多变量分析方法。随着计算机技术的发展，越来越多 Logistic 回归被应用于医学研究，随之也常常出现误用及结果解释不当的问题。本文主要讨论 Logistic 回归应用中有关建模与结果解释方面的常见问题及其注意事项。

1 Logistic 回归模型的建立

在 Logistic 回归模型中，应变量可以是二分类（如发病/未发病、有效/无效、死亡/存活）或多分类（如评分等级 I 级/II 级/III 级、无效/有效/痊愈）。自变量可以有多种形式：连续型变量（如年龄、BMI）、有序分类变量（如教育程度、疾病严重程度）或分类变量（如性别、职业）。

以二分类应变量为例，设 m 个自变量（m 个暴露因素）为 $x_1, x_2 \cdots, xm$ 时发生某研究结局的概率为 P，在 Logistic 模型中，将发生概率 P 与未发生概率 1-P 之比称为优势（odds），其对数记为 $\text{logit}P = \ln\left(\frac{P}{1-P}\right)$，则对于有 m 个自变量的 logistic 回归模型，可表示为

$$\ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m \quad (1)$$

1.1 优势比的意义 对于某暴露因素的两个不同的暴露水平 $X_j=c_1$ 与 $X_j=c_0$（假定其他因素的水平固定不动），公式(1)中回归系数 β_j（$j=1,2,\cdots,m$）表示自变量 x_j 改变一个单位时 $\text{logit}P$ 的改变量，其指数称为优势比（odds ratio，OR），即 $\ln OR_j = \ln\left[\frac{P_1/(1-P_1)}{P_0/(1-P_0)}\right] = \beta_j(c_1 - c_0) = \text{logit}P_1 - \text{logit}P_0$ $OR_j = \exp[\beta_j(c_1-c_0)]$。

注意，OR 的应用及解释存在以下误区：1）认为 $OR_j>1$ 代表 x_j 为危险因素，$OR_j<1$ 代表 x_j 为保护因素。事实上，OR 代表应变量与自变量之间联系的强度，需要根据研究结局进行专业意义的解释。如果研究结局是正性事件，如疾病治愈、生存，则 $OR_j>1$ 代表 x_j 为促进疾病治愈、促进生存的保护因素，而 $OR_j<1$ 代表 x_j 为不利于正性事件的危险因素；当研究结局为负性事件，如疾病发生、进展、死亡等，则 $OR_j>1$ 代表 x_j 为危险因素，$OR_j<1$ 代表 x_j 为保护因素。2）将 OR 与相对危险度（relative risk，RR）含义混淆。RR 是暴露组与未暴露组的研究结局发生率之比，它是一个比值，代表暴露于某个因素的研究结局发生率是未暴露组的多少倍。而 OR 是优势比，可以理解为 x_j 每改变一个单位时，研究结局的发生风险改变量。只有发生率很低的研究结局，即 P 很小时，OR 才近似等于 RR，即

$$\text{OR}=\frac{P_1/(1-P_1)}{P_0/(1-P_0)}\approx\frac{P_1}{P_0}=RR \quad (2)$$

1.2 哑变量的设置 Logistic 回归中最常见的自变量类型为多分类变量，如分娩方式分为顺产、难产、剖腹产，婴儿喂养方式分为母乳喂养、混合喂养、人工喂养。有些研究者将多分类自变量误作为连续型变量引入 Logistic 模型进行分析，这意味着该变量各相邻分类水平间是等距的，显然不符合实际逻辑。例如，有研究对合肥市城区低出生体重儿影响因素的 Logistic 回归分析中，原作者将“父亲职业”这个分类变量分为“体力劳动”、“体力兼脑力劳动”、“脑力劳动三个分类水平，进行 Logistic 回归分析时却将该变量作为有序分类（等级）变量赋值为各分类水平的得分，按连续变量进行处理欠妥。

对这类变量的赋值应采用哑变量的方法，即设立一个参照水平，将有 m 个分类水平的多分类变量转换为 m-1 个哑变量（取值 0 或 1）。如表 1 为 4 分

作者简介：李晨（1986-），女，讲师，医学博士，主要研究方向为医学研究设计与统计分析方法

通讯作者：陈长生 Email：chencs@fmmu.edu.cn

基金资助：国家自然科学基金（81803328，81573251）

类自变量(可转换为3个哑变量)的各水平间的优势比,3个哑变量的偏回归系数分别为b_1、b_2、b_3,则表1中第1行是相对参照水平(第1水平)的优势比,第2行是相对第2水平的优势比,其余类似。

表1 各水平间的优势比

Tab. 1 Odds ratio in different levels

变量分类水平	第1水平(参照)	第2水平	第3水平	第4水平
第1水平(参照)	1	exp(-b_1)	exp(-b_2)	exp(-b_3)
第2水平	exp(b_1)	1	exp(b_1-b_2)	exp(b_1-b_3)
第3水平	exp(b_2)	exp(b_2-b_1)	1	exp(b_2-b_2)
第4水平	exp(b_3)	exp(b_3-b_1)	exp(b_3-b_2)	1

对于各*OR*值是否有统计学意义,应通过相应的假设检验来判断,也可简单地通过其95%*CI*是否包含1来直观判断。

此外,如果研究者想观察分类变量的各暴露水平对研究结局的影响,也可以将有m个分类水平的变量转变为m个哑变量(取值0或1),每个哑变量分别代表有无该水平的暴露(取值1代表有,取值0代表无)。例如,有研究对儿童慢性胃炎、消化性溃疡致病危险因素的Logistic回归分析中,对于膳食模式这个3分类变量(喜爱蔬菜/水果/肉食),可转变为3个哑变量(分别表示"是否喜爱蔬菜"、"是否喜爱水果"、"是否喜爱肉食")。

1.3 Logistic回归模型的变量筛选 Logistic回归模型建立时,如自变量较多,可采用逐步回归法进行变量筛选。不同的筛选方法有时会产生不同的模型。判断某个变量是否显著以及作用大小,与模型中所包含的变量有关。实际工作中衡量某些变量是否选入模型,需要考虑专业背景、研究目的,用以调整的某些重要混杂因素以及模型的可解释性、节约性等。

2 Logistic回归结果的解释

对于Logistic回归模型结果的解释,与参照水平以及哑变量的设置有关。

2.1 参照水平的设置 实际工作中,有些论文作者在Logistic回归模型的结果展示中只标注自变量中文名称和回归系数,未说明各变量的参照水平设置。在回归系数的解释上,也只说明自变量对研究结局是危险或保护因素,未考虑实际专业意义。表2研究儿童注意缺陷多动障碍的非生物学相关因素logistic回归分析结果。

表2 ADHD儿童相关因素的logistic回归分析

Tab. 2 Logistic regression analysis of related factors in children with ADHD

影响因素	β值	*P*值	*OR*值	95%*CI*
不良孕期史	1.025	0.040	2.788	1.049～7.412
儿童视屏年龄<3岁	1.275	0.003	3.577	1.552～8.248
父亲生育年龄(26～35岁)	-1.694	0..15	0.184	0.047～0.723

未说明儿童视屏年龄<3年、父亲生育年龄(26～35岁)分别是与哪个变量水平(即参照水平)作对比,也没有对相应的*OR*值做任何解释,便直接给出结论:儿童视屏年龄<3岁是ADHD儿童的危险因素,父亲生于年龄(26～35岁)是保护因素。

通常,Logistic回归分析以自变量中赋值较小的变量水平为参照水平,但为了更好地解释分析结果,在论文报告中还是需要说明哪个变量水平为参照水平,如表3对儿童超重和肥胖因素的Logistic回归分析研究中,作者列出了每个自变量的参照水平。

表3 儿童超重和肥胖影响因素的Logistic回归分析

Tab. 3 Ordinary Logistic regression analysis of influence factors of overweight and obesity in children

自变量项	项目	*B*值	*S.E.*值	*Wald*值	*P*值	*OR*值
性别	男	0.709	0.142	24.908	0.000	2.032
父亲BMI	肥胖(以正常为参照)	0.681	0.162	17.616	0.000	1.976
母亲BMI	肥胖(以正常为参照)	0.672	0.241	7.814	0.005	1.958
母亲学历	大专及以上(以初中及以下为参照)	0.598	0.276	4.706	0.030	1.819
银屏活动	每周≥7次(以每周≤2次为参照)	0.582	0.261	4.988	0.026	1.79
是否独生	(以非独生为参照)	0.32	0.154	4.331	0.037	1.377
出生体重	2 500～4 000g(以≥4000g为参照)	-0.54	0.188	8.264	0.004	0.583

由表3可知,父亲肥胖的儿童更容易发生超重和肥胖,其风险是父亲体重正常者的1.976倍;而相较于出生体重≥4 000g的儿童,出生体重在2 500-4 000g是儿童超重肥胖的保护因素。

2.2 因素的作用大小 在Logistic回归分析结果的解释中,有些论文作者会直接比较*OR*绝对值的大小来说明不同因素对应变量的作用大小,例如,有研究对儿童哮喘相关因素进行Logistic回归分析,作者根据*OR*值的大小,将发生哮喘的相关因素人为地分为

高危因素、危险因素、低危因素和保护因素；对儿童慢性胃炎、消化性溃疡致病危险因素的 Logistic 回归分析中，作者也按照 *OR* 值的大小，将危险因素进行了排序。通常情况下，各个自变量的度量衡单位不一致，在对各个自变量进行标准化前，Logistic 回归模型的各个自变量所对应的 *OR* 值并不适合直接比较。例如，疾病严重程度每增加 1 级（如Ⅰ级到Ⅱ级）所对应的 *OR* 值与年龄每增加 1 级（如 5～10 岁增至 10～15 岁）所对应的 *OR* 值不适合互相比较。因此，为了比较各因素的相对重要性，可以计算各因素的标准回归系数用于比较。

3 logistic 回归应用的注意事项

3.1 变量的取值形式 对同一资料的分析，变量采用不同的取值形式，参数的含义、量值及符号都可能发生变化。在做影响因素分析时，若自变量是一个定量指标，最好将其按变量值的大小分成几组（如分 4 组），按顺序取值为 1，2，…，k，否则参数的实际意义不够明确。例如对于年龄变量，exp$b$ 表示每增加一岁时的优势比，实际意义不大；如果是白细胞数就显得有些荒谬了。这种情况如果将年龄或白细胞数分成几个不同的水平，更容易解释，在赋值处理上也比较灵活，分析时既可以按赋值得分处理，也可以将其化作 k-1 个哑变量，并在分析中对差别不大的水平做一些必要的合并。

3.2 Logistic 回归的样本含量 Logistic 回归的所有统计推断都是建立在大样本基础上的，因此要求有足够的样本含量。一般来说，样本含量至少是自变量个数的 15～20 倍。关于样本含量的具体确定，已经有一些工具表可供医学科研工作者参考。经验上病例和对照的人数应至少各有 30～50 例，方程中变量的个数愈多需要的例数相应也愈大。对于配对资料，样本的匹配组数应为纳入方程中的自变量个数 p 的 20 倍以上，即 $n \geqslant 20p$ 。

3.3 多分类应变量的 Logistic 回归模型 当应变量为有序多分类变量时，如流行病学中一些慢性病的危险因素研究，观察结果为“无、轻、中、重”；临床试验的疗效评价，结果为“治愈、显效、好转、无效”；临床影像诊断按“－，±，＋，＋＋”不同等级进行分类的资料，均可以采用有序 Logistic 回归模型进行分析，假设应变量有 g 个等级水平，则有序 logistic 回归模型包括 g-1 个方程，且自变量在 g-1 个模型中对累计概率的优势比影响相同，即 g-1 个方程中各自变量的回归系数相同（即平行性假设），不同类别累计概率的差别体现在常数项上。因此在拟合有序 Logistic 回归模型时，需要对 g-1 个方程对应的累计概率曲线的平行性进行检验，即检验各自变量在不同累计概率模型中的回归系数是否相同。如果检验结果 $P>0.10$，说明满足了平行性假设；否则，不满足平行性假设，说明不适合进行有序 Logistic 回归分析，而应该采用无序多分类 Logistic 回归模型进行分析。

当应变量是无序多分类变量时，应采用无序多分类 Logistic 模型，该模型需要选取应变量多类别之一作为参照，拟合剩余各类别相对于此参照类别的 Logistic 模型，因此依然包括 g-1 个方程。与有序 Logistic 回归模型不同的是，各类别相对于参照类别的回归方程自变量的回归系数可以不同。

Logistic 回归模型的建模较为复杂，受到诸多因素的影响，如影响因素有：研究资料的数据质量、缺失值、离群值、样本含量的大小、自变量的暴露水平及其共线性等等。应用中需要注意不同资料类型的使用条件、对结果的专业解释也需慎重。

参考文献（略）