·问题与探讨·

logistic 回归应用中容易忽视的几个问题

冯国双 陈景武 周春莲

logistic 回归在流行病学研究中应用十分广泛，在病例对照研究和队列研究中，logistic 回归是经常用到的多变量统计分析方法，在随访研究和横断面调查中，logistic 回归的应用也较为普遍[1-5]。与多元线性回归相比，logistic 回归具有许多独特的优点，如对正态性和方差齐性不做要求，对自变量类型不做要求、系数的可解释性等。正是这些优点，使得 logistic 回归成为流行病学研究中广受欢迎的分析工具。1996～2002 年《中华流行病学杂志》发表的文章有 111 篇用到了 logistic 回归。尽管 logistic 回归应用如此广泛，但在具体使用中仍存在不少问题。在 111 篇应用 logistic 回归的文章中，主要存在三个问题：资料的适合、拟合优度检验及回归诊断问题。笔者主要针对这三个问题进行讨论。

1. 资料的适合问题：在应用 logistic 回归方法前，首先应分析该资料用 logistic 模型是否适合，这就是资料的适合问题。在检索到的 111 篇文章中，主要存在两个问题：自变量与 logit p 之间缺乏线性关系判断以及样本含量不足。

logistic 回归模型为 logit $p = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots + \beta_p x_p$。这一模型对自变量类型一般不做规定，但要求自变量与 logit p 之间应符合线性关系。当自变量为分类变量时，可不必考虑线性关系[6]，但当自变量为连续型变量时，则需要检验两者之间的线性关系是否成立。如果不成立，应进行相应的变量变换，如对数变换、指数变换、多项式变换等，使其以恰当的形式进入方程。严格说来，应用 logistic 回归之前必须先检验自变量与 logit p 之间是否具有线性关系，因为如果两者之间的关系是非线性的，参数估计将发生偏差，从而导致结果的不准确以及结论的不可靠。但在 111 篇应用 logistic 回归的文章中却无一篇提及自变量与 logit p 之间的线性关系问题，说明这是一个很容易被忽视的问题，在实际应用中应引起注意。判断自变量与 logit p 之间是否具有线性关系，可用多种方法[7]：比较简单的一种方法是在模型中加入非线性项，如 χ^2、$\ln\chi$ 等，使线性模型变为非线性模型，通过比较非线性模型与线性模型的优劣来判断是否应加入该非线性项，从而判断出自变量与 logit p 是否有非线性关系。还有其他较复杂但更为准确的方法，如将连续变量分为几组，然后用虚拟变量代表这些组别，并以最低的一组作为参照组，然后再用这些虚拟变量代替原先的连续变量，并重新估计模型。具体方法可参考有关文献。

作者单位：261042 潍坊医学院公共卫生与管理学院卫生统计教研室

logistic 回归模型对样本含量有一定的要求，一般经验认为，样本规模至少应是自变量个数的 10 倍以上。当样本含量过少时，估计的方程会显得不稳定，系数或标准误的估计也会出现一些不可思议的数值，从而使方程变得无法解释。在上述 111 篇文章中，约有 10% 出现样本含量过小的问题。如某文章对散发型感染性腹泻进行病例对照研究，采用 1∶1 配对 logistic 回归，样本共 85 对，而研究因素则达 32 个之多，在这种情况下，人们很容易对求出的结果表示怀疑。在有些情况下，如果确实无法获得更多的样本，可以考虑用确切 logistic 回归（exact logistic regression）[8]。确切 logistic 回归主要用于小样本、资料结构不平衡等的数据，可以作为在最大似然法不可靠或者不存在情况下的补充方法。在 SAS 8.0 及以上版本可以通过命令 proc logistic; model……; exact……（省略要进行精确的变量）；来实现确切 logistic 回归的相应输出。

2. 拟合优度检验问题：建立模型并进行假设检验只表明了模型中的回归系数是否具有统计学意义，但并不表明模型拟合的效果如何。要说明这一点，还应对所拟合的模型进行评价，即评价模型的预测值是否与观测值具有较高的一致性，这就是拟合优度检验问题。

拟合优度检验是 logistic 回归分析过程中不可缺少的一部分，拟合的效果好，所做出的结论才更符合事实，若拟合的不好，预测值与实际值差别较大，得出的结论是不可靠的。然而，实际应用中这一点往往被忽略。在检索到的 111 篇文章中，仅有 1 篇提到了模型的拟合优度问题，而且这篇文章中所用的评价指标是决定系数 R^2，这是不恰当的。R^2 是多元线性回归中经常用到的一个指标，表示的是因变量的变动中由模型中自变量所解释的百分比，并不涉及预测值与观测值之间差别的问题。在 logistic 回归中，评价模型拟合优度的指标主要有 Pearson χ^2、偏差（deviance）、Hosmer-Lemeshow（HL）指标、Akaike 信息准则（AIC）、SC 指标等。Pearson χ^2 和 Deviance 主要用于自变量不多且为分类变量的情况，当自变量增多且含有连续型变量时，用 HL 指标则更为恰当。Pearson χ^2、Deviance 和 HL 指标值均服从 χ^2 分布，χ^2 检验显示无统计学意义（$P>0.05$）表示模型拟合的较好，χ^2 检验有统计学意义（$P<0.05$）则表示模型拟合的较差。AIC 和 SC 指标还可用于比较模型的优劣，当拟合多个模型时，可以将不同模型按其 AIC 和 SC 指标值排序，AIC 和 SC 值较小者一般认为拟合得更好。

上述五个指标在 SAS 中均可通过命令实现。如调用命

令 proc logistic; model…… /scale = none aggregate; 可输出 Pearsonχ^2 和 Deviance 值。调用命令 proc logistic; model……/lackfit; 可输出 HL 指标。AIC 和 SC 指标在 SAS 命令 proc logistic; 中自动输出。

3. 回归诊断问题: 即使资料符合 logistic 回归应用的条件, 所求模型的拟合优度也不一定很好, 因为模型中很可能存在相关性较强的几个变量或较为特殊的几个样品, 从而影响模型的拟合效果, 这时就应对模型进行多重共线性诊断以及特殊点的识别等。在检索的 111 篇文章中, 进行回归诊断的不到 5 篇, 而且主要是共线性问题, 说明回归诊断也是一个易被忽视的问题。

logistic 回归与多元线性回归一样也存在多重共线性问题, 其诊断可以用容忍度(tolerance)、方差扩大因子(variance inflation factor, VIF)、条件指数(condition index)、方差比例(proportion of variation)等指标来表示。但是, SAS 的 proc logistic 命令中并不提供输出这些指标的备选项, 需要借助多元线性回归来实现[9]。由于我们关心的只是自变量之间是否存在共线性, 而不是注重自变量与因变量之间的关系, 因而我们可以直接用二分变量 y 来代替 logit$p$ 作为因变量, 进行多元线性回归。通过 SAS 命令 proc reg; model…… / collinoint; 可输出条件指数和方差比例, 用 SAS 命令 proc reg; model…… /tol vif; 可输出容忍度和方差扩大因子。一般认为, 当容忍度 <0.10 或条件指数 >30 可以认为自变量间存在较强的共线性, 在 SAS 输出结果中, 如果某一行的条件指数 >30, 则该行中方差比例 >0.5 的 n 个自变量之间可能存在共线性。解决共线性的方法一般有: ①删除冗余的自变量, 但在实际中往往会因为无法区别有意义的变量与冗余变量而误删, 从而造成模型误设。②增加样本含量, 使标准误减少, 抵消多重共线性的影响。但这种方法只有在多重共线性是由测量误差引起或偶然存在于原始样本而不存在于总体时才适用。③用逐步 logistic 回归, 寻求建立一种最佳回归方程, 但这种方法容易损失一些信息。④用主成分 logistic 回归, 通过主成分变换, 将高度相关的几个变量的信息综合起来参与回归。尽管这种方法仍没有从根本上解决共线性问题, 但不失为一种较好的解决共线性的办法, 也是目前用的较多的一种方法。

特殊样品主要包括特异点(outlier)、高杠杆点(high leverage points)以及强影响点(influential points)。特异点是指残差较其他各点大得多的点; 高杠杆点是指距离其他样品较远的点; 强影响点是指对模型有较大影响的点, 模型中包含该点与不包含该点会使求得的回归系数相差很大[7,10-12]。单独的特异点或高杠杆点不一定会影响回归系数的估计, 但如果既是特异点又是高杠杆点则很可能是一个影响回归方程的"有害"点[12]。对特异点、高杠杆点、强影响点诊断的指标有: Pearson 残差、Deviance 残差、杠杆度统计量 H(hat matrix diagnosis)、Cook 距离、DFBETA 指标等。SAS 中可通过命令 proc logistic; model ……/influence; 输出上述诊断标准。还可通过 SAS 命令 proc logistic; model ……/IPLOT; 绘制残差图对这些特殊点进行直观的描述。这五个指标中, Pearson 残差和 Deviance 残差可用来检查特异点, 在 SAS 结果输出中, 如果某观测值的残差值 >2, 则可认为是一个特异点。杠杆度统计量 H 可用来发现高杠杆点, SAS 结果输出中, H 值大的样品说明距离其他样品较远, 可认为是一个高杠杆点。Cook 距离、DFBETA 指标可用来度量特异点或高杠杆点对回归模型的影响程度。Cook 距离是标准化残差和杠杆度两者的合成指标, Cook 距离越大, 表明所对应的观测值的影响越大。DFBETA 指标值反映了某个样品被删除后 logistic 回归系数的变化, 变化越大(即 DFBETA 指标值越大), 表明该观测值的影响越大。如果模型中检查出有特异点、高杠杆点或强影响点, 对其处理应采取慎重态度[10,11]。首先应检查是否数据搜集或录入中的错误, 如果属于这类错误, 则可以删除。其次应考虑是否忽略了重要的协变量, 是否需要加上交互作用, 样本含量是否足够等其他问题。

参 考 文 献

1 李春波, 何燕玲, 张明园. 抑郁症状对社区亚老龄及老龄人群心身健康结局影响的随访研究. 中华流行病学杂志, 2002, 23: 341-344.

2 陈薇, 杨放, 汪丽萍. 城市居民一般人群艾滋病相关知识、信念、行为现况调查分析. 中华流行病学杂志, 2001, 22: 395-396.

3 郑宏, 于普林, 洪依舒. 我国城乡老年人白内障患病情况调查. 中华流行病学杂志, 2001, 22: 446-448.

4 郭素芳, 王临虹, 尹仁英. 城乡生殖道感染妇女利用卫生的调查. 中华流行病学杂志, 2002, 23: 40-42.

5 徐金华, 王吉耀, 赵耐青, 等. 性病病人健康教育前后性病预防知识改变的流行病学调查. 中华流行病学杂志, 2002, 23: 218-220.

6 陈峰. 医用多元统计分析方法. 北京: 中国统计出版社, 2001. 111-112.

7 王济川, 郭志刚. logistic 回归模型—方法与应用. 北京: 高等教育出版社, 2001. 172-203.

8 刘启军, 曾庆, 周燕荣, 等. 精确 logistic 回归及其 SAS 应用程序. 中华流行病学杂志, 2003, 24: 725-728.

9 赵宇东, 刘嵘, 刘延龄, 等. 多元 logistic 回归的共线性分析. 中国卫生统计, 2001, 17: 259-261.

10 罗登发, 余松林. 条件 logistic 回归模型的残差分析和影响诊断. 中国卫生统计, 1997, 14: 13-15.

11 魏朝晖. logistic 回归诊断. 中国卫生统计, 2001, 18: 112-113.

12 赵清波, 徐勇勇, 夏结来. logistic 回归中高杠杆点的检测. 中国卫生统计, 1997, 14: 17-20.

(收稿日期: 2003-11-19)

(本文编辑: 张林东)