学习园地

采用 Logistic 回归分析时需注意的问题

吴振强，王杨，李卫

关键词 Logistic；回归；样本量

Logistic 回归常用于分析二分类因变量（如存活和死亡、患病和未患病等）与多个自变量的关系 [1]。比较常用的情形是分析危险因素与是否发生某疾病相关联。例如，若探讨胃癌的危险因素，可以选择两组人群，一组是胃癌组，一组是非胃癌组，两组人群有不同的临床表现和生活方式等，因变量就为有或无胃癌，即“是”或“否”，为二分类变量，自变量包括年龄、性别、饮食习惯、是否幽门螺杆菌感染等。自变量既可以是连续变量,也可以为分类变量。通过 Logistic 回归分析，就可以大致了解胃癌的危险因素。

Logistic 回归与多元线性回归有很多相同之处，但最大的区别就在于他们的因变量不同。多元线性回归的因变量为连续变量;Logistic 回归的因变量为二分类变量或多分类变量，但二分类变量更常用，也更加容易解释 [1]。

尽管 Logistic 回归在医学研究领域中应用广泛，但在应用中存在很多问题。本文将结合笔者自身的经验，对使用 Logistic 回归常见的问题进行讨论。

1 Logistic 回归的用法

一般而言，Logistic 回归有两大用途，首先是寻找危险因素，如上文的例子，找出与胃癌相关的危险因素；其次是用于预测，我们可以根据建立的 Logistic 回归模型，预测在不同的自变量情况下，发生某病或某种情况的概率(包括风险评分的建立)。

2 用 Logistic 回归估计危险度

所谓相对危险度（risk ratio，RR）是用来描述某一因素不同状态发生疾病（或其它结局）危险程度的比值。Logistic 回归给出的 OR（odds ratio）值与相对危险度类似，常用来表示相对于某一人群，另一人群发生终点事件的风险超出或减少的程度。如不同性别的胃癌发生危险不同，通过 Logistic 回归可以求出危险度的具体数值，例如 1.7，这样就表示，男性发生胃癌的风险是女性的 1.7 倍。这里要注意估计的方向问题，以女性作为参照，男性患胃癌的 OR 是 1.7。如果以男性作为参照，算出的 OR 将会是 0.588 (1/1.7), 表示女性发生胃癌的风险是男性的 0.588 倍，或者说，是男性的 58.8%。撇开了参照组，相对危险度就没有意义了。

Logistic 回归在医学研究中广泛使用的原因之一，就是模型直接给出具有临床实际意义的 OR 值，很大程度上方便了结果的解读与推广。

3 样本量问题

通常回归模型都需要建立在大样本的基础上。在进行 Logistic 回归前，应该考虑当前的样本量是否充足？根据模拟研究，在使用 Logistic 回归时，事件（死亡或患病）个数至少应该是自变量个数的 10 倍以上(这一条也适于 Logostic 其他的应用情况)[2]。例如，观察胃癌的危险因素，比如有性别、年龄和饮食习惯等 9 个研究因素，那就至少需要 90 例胃癌。另一个比较常见的样本量原则是，观测的数量应该至少是自变量数的 20~30 倍，同样如果有 9 个自变量，那么总体样本最好能够达到 180 例以上。建议在进行 Logistic 回归前，结合上述两个原则，从总样本和事件数两个角度共同对模型样本量进行考虑。

4 Logistic 回归中的自变量形式

Logistic 回归的自变量既可以是连续变量，也可为分类变量。总体原则是尽量从实际或专业角度考虑采取何种形式更好。比如年龄,可以取为连续变量，也可以 5 岁、10 岁作为一组，甚至分为老年人和年轻人两组。不同的划分方式决定了结果解读时的差异，比如，在做出胃癌与年龄的关系，如果把年龄

作者单位:100037 北京市，中国医学科学院 北京协和医学院 心血管病研究所 阜外心血管病医院 医学研究统计中心
作者简介：吴振强 硕士 主要从事生物统计学研究 Email wuzhenqiang@mrbc-nccd.com 通讯作者：李卫 Email:liwei@mrbc-nccd.com
中图分类号:R54 文献标识码:C 文章编号：1000-3614（2014）03-0230-02 doi:10.3969/j.issn.1000-3614.2014.03.019

作为连续变量分析，得到危险度为 1.008，其解释为年龄每增加 1 岁，患胃癌的风险就会多出 0.008 倍，这个数据会显得没有太大的临床意义。但如果以 10 岁一组，可能得到的危险度就是 1.6，即年龄每增长 10 岁、患胃癌的风险就增加 60%，这样幅度的相对风险更具有临床实际意义。

如何将连续变量进行划分并没有固定的标准，按照统计学的分位数或具有临床意义的界值划分都是常用的方法。建议在分析时先进行趋势的描述，观察特定的自变量和因变量是何种关系，再结合临床专业角度与统计学考虑，以获得最合理的划分方式。

5 Logistic 回归时单因素分析

在进行 Logistic 回归分析时，是否必须先进行单因素分析，然后才能进行多因素分析？理论上讲，如果样本足够大，且所有的因素之间没有关联，最好把所有的因素都放到方程中，通过全模型法对所有可能的混杂因素同时进行分析，在此基础上进一步通过逐步回归的方法对有显著意义的变量进行筛选，此种情况下可以不做单因素分析。如果样本例数有限，比如，仅有 80 例患者，但是有 20 个因素，这种情况下，最好先进行单因素分析，剔除既无统计学意义，又无临床意义的变量，只分析有意义的变量。

单因素分析时最好将 P 值放宽，比如 0.1 或 0.15 等，避免漏掉一些重要因素 (变量间的相互作用可能导致多因素的结果不同于单因素分析)。当然，也要注意仔细检查各因素间的关联程度，对于高度相关的自变量一般不同时带入模型，例如：收缩压和舒张压。一旦发现因素之间有较强的相关性，建议首先进行筛选，选择最具代表性的变量带入模型。

参考文献

[1] 陈峰 . 医用多元统计分析方法 . 北京 . 中国统计出版社 . 2007. 83-113.

[2] Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996. 49: 1373-1379.

(收稿日期：2014-02-07)

(编辑：常文静)

读者 · 作者 · 编者

本刊已启用稿件管理系统

为顺应当今期刊网络化、数字化的发展趋势，更好地为广大作者、读者提供高质量的服务，《中国循环杂志》社于 2014 年 1 月正式启用稿件管理系统。该系统采用先进的数据库及网络技术，具有强大的数据处理和分析能力。稿件管理系统将协助作者、编辑、审稿专家、编委、总编等相关人员多位一体地进行稿件业务处理，解决编辑部对稿件网络化流程管理的需要，并实现各类查询功能，方便作者及时了解稿件进程、缩短稿件处理周期。

投稿过程中具体注意事项如下：

1 注册及投稿 ① 在浏览器中输入 http://www.chinacirculation.org/，点击“作者在线投稿”第一次使用本系统进行投稿的作者，必须先注册，注册时请务必使用真实邮箱，同时各项信息请填写完整。作者自己设定用户名和密码，该用户名和密码长期有效。本刊的审稿专家投稿，可点击审稿链接进行投稿。② 用户名 (您的真实邮箱) 和密码为您在本刊的登录信息，请牢记！忘记密码时可通过注册邮箱索取密码，密码会发送到您的邮箱。③注册成功后输入“用户名 (您的真实邮箱)”、“密码”，点击“登录”，成功登录。④进入投稿界面，系统会提示首先更新个人注册信息，然后点左上角的“我要投稿”进行投稿。

2 如何查询稿件情况　稿件一经投稿成功，作者可登陆网站关注该稿件的“稿件处理流程”。如有疑问，可打电话向编辑部咨询。稿件投稿成功、退修、退稿等通知会发至投稿人的邮箱中，具体请登录系统查询。

该系统正式启用后，有关投稿及系统操作的相关问题请致电：010-60213898。

《中国循环杂志》编辑部