·循证医学中的医学统计学问题·

Logistic回归分析的样本量确定

高永祥， 张晋昕

（中山大学公共卫生学院，广州 510080）

［摘要］ Logistic回归是一种广泛使用的统计模型。在实际应用中，有很多研究者往往忽视Logistic回归对样本量的要求，或者凭“纳入的研究对象人数充分”草草带过样本量问题，这些做法使主要影响因素与结局间关系的探索未能结合研究设计阶段对两类错误的设定。本文介绍三种Logistic回归样本量计算方法，并辅以实例说明，帮助研究者合理完成研究的设计与实施。

［关键词］ EPV；样本量；Logistic回归

［中图分类号］ R195.1 ［文献标识码］ A DOI：10.12019/j.issn.1671-5144.2018.02.015

Determination of Sample Size in Logistic Regression Analysis

GAO Yong-xiang, ZHANG Jin-xin

(*School of Public Health*, *Sun Yat-sen University*, *Guangzhou* 510080, *China*)

Abstract: Logistic regression is a widely used statistical model. In practice, many researchers tend to ignore the requirements of the logistic regression on sample size or take over the sample size due to “enough subjects were included in the study populaton”, which fails to explore the relationship between the primary outcome and main impact indicator and ignores two types of errors. This paper introduced three Logistic regression sample size calculation methods, supplemented by examples to help researchers reasonably complete the design and implementation of the study.

Key words: EPV; sample size; Logistic regression

Logistic回归（logistic regression）模型被广泛应用于各学科领域，如医学、社会科学、机器学习等，主要适用于因变量是分类变量的情况，尤其当因变量属于0-1变量。该模型采用的参数估计方法是极大似然估计（maximum likelihood estimate，MLE），这就需要足够的样本量来保证参数估计的准确性，而样本量的估计又是常常困扰研究者的一个问题，以下将汇总二分类Logistic回归分析中几种常用的样本量确定方法。

1 经验方法

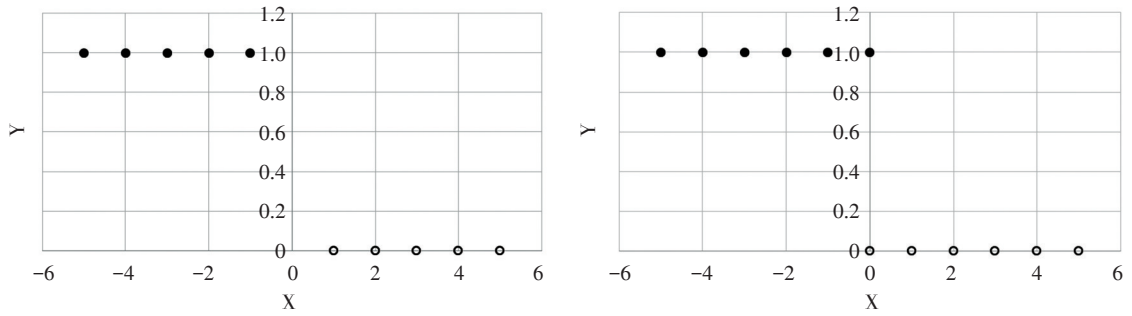
目前广泛使用的方法是EPV（events per variable）的方法，即每个自变量的事件数，其中事件表示因变量中个数较少的那一类[1]。例如调查胃癌发病与3种生活因素（X_1代表不良饮食习惯，X_2代表喜吃卤食和盐渍食物，X_3代表精神状况）的关系[2]，若胃癌患者占的比例为20%，那么当假设EPV=10时，由于有3个协变量，所以所需胃癌患者例数为10×3=30，总共需要的样本量（胃癌患者和健康对照）为30÷20%=150例。当EPV过少时，容易出现分离（separation）现象。此现象出现在自变量若大于某个常数，变量则仅与一个自变量相关联。例如当X为连续型变量时，若$X\leqslant 0$时，有Y恒为1，则出现完全分离（complete separation）现象（见图1a），此时参数估计无法收敛，得不到回归系数的估计值。另一情形是，当$X<0$，Y恒为1，但当$X=0$时Y兼有观察值0和1，这时会出现拟完全分离（quasi-complete separation）现象（见图1b），此时极大似然估计值异常大。统计学模拟研究表明[1]，在Logistic回归中推荐的经验

［基金项目］ 广东省自然科学基金资助项目（2016A030313365）

［作者简介］ 高永祥（1993-），男，山东潍坊人，硕士研究生，从事统计学方法及其医学应用研究。

［通讯作者］ 张晋昕，Tel：020-87332453；E-mail：zhjinx@mail.sysu.edu.cn

准则是 EPV 至少为 10，才能保证结果稳健。另外一个比较常用的经验准则是样本量为协变量个数的 10~15 倍[3]。具体应用时可以综合考虑两种经验准则。



a. 完全分离　　b. 拟完全分离

图 1　Logistic 回归中自变量与结局变量间的分离现象

2　公式法

Whittemore 1981 年提出了罕见事件 Logistic 回归样本量估算公式[4]，随后 Hsieh 对 Whittemore 的公式进行了扩展[5]，在 1998 年提出了一个便于一般应用者实施的简单方法。建议借用样本均值比较和样本频率比较的样本含量计算公式来估算单因素 Logistic 回归所需的样本量，再用方差膨胀因子对其修正便得到多因素 Logistic 回归所需的样本量[6]。

单因素 Logistic 回归中，当 X 为连续型变量并且服从正态分布时，样本量的计算公式为：

$$n=\frac{(Z_{1-\alpha/2}+Z_{\beta})^2}{[p_1(1-p_1)]b^2} \quad (1)$$

式(1)中 p_1 为 X 取均值条件下 $Y=1$ 发生的频率，b 为要度量的效应大小，亦即 X 所对应回归系数的估计值。

当 X 为二分类变量时，样本量的计算公式为：

$$n=\frac{\left\{Z_{1-\alpha/2}[p(1-p)/B]^{1/2}+Z_{1-\beta}[p_0(1-p_0)+p_1(1-p_1)(1-B)/B]^{1/2}\right\}^2}{(p_0-p_1)^2(1-B)} \quad (2)$$

式(2)中 p 为总的阳性结局发生频率，B 为 X=1 的个体在总观察人数中所占的比例(流行病学研究中对应于暴露比例)，p_0 和 p_1 分别为 X=0 和 X=1 时的阳性结局发生频率。

多因素 Logistic 回归样本量计算公式为：

$$n_p=\frac{n_1}{1-R^2_{1,234\cdots p}} \quad (3)$$

式(3)中的 $R^2_{1,234\cdots p}$ 就是以最主要的暴露因素 X_1 为因变量，X_2，……，X_p 为自变量做线性回归得到的决定系数 R^2，n_1 为单因素 Logistic 回归所需的样本量。其实，$1/(1-R^2_{1,234\cdots p})$ 被统计学家定义为一个重要参数——方差膨胀因子(variance inflation factor，VIF)，故多因素 Logistic 回归的样本量即为最主要的暴露因素所对应单因素 Logistic 回归所需的样本量 n_1 乘以该因素对应的方差膨胀因子 VIF。

实例 1　某课题组拟探索非甾体抗炎药相关上消化道出血是否与吸烟之间存在关系，现计算研究所需样本量。假设 α=0.05(双侧)，β=0.10(单侧)。

根据该课题组的回顾性分析，已知 B=0.48，p_0=0.43，p_1=0.58，p=0.50，$Z_{1-\alpha/2}$=1.96，$Z_{1-\beta}$=1.28，代入公式(2)可得 $n\approx$464。

实例 2　假设在实例 1 中除了吸烟因素外，还考虑饮酒、冠心病史、慢性胃炎史等可能影响上消化道出血的因素，在这里我们最关心的暴露因素为是否吸烟，并且已知吸烟与上述因素(自变量)之间的 R^2 为 0.07，则根据公式(3)可得多因素 Logistic 回归所需样本量为 $n\approx$499。

3　软件实现

借助公式(1)~(3)进行手工计算，麻烦且易出错。此处介绍如何利用开源软件 R i386 3.2.1 和商业软件 PASS11 完成 Logistic 回归样本量的估算，R 和 PASS 实际上也是基于前述公式实现的，同样以实例 1 和实例 2 进行说明，程序见表 1 和表 2。R 的计算结果，实例 1 为 464，实例 2 约为 499；PASS 的计算结果，实例 1 为 463，实例 2 为 498。忽略计算精度，二者结果基本类似。

除了 R 软件、PASS 可以用来计算样本量以外，还有 nQuery 等软件可供读者使用，在此不做详细说明，周映雪等[7]具体介绍了用 nQuery Advisor7.0

表1 Logistic回归样本量估算的R程序及其说明

程序语句	说 明
install.packages('powerMediation')	安装R包"powerMediation"
library(powerMediation)	加载R包"powerMediation"
n<-SSizeLogisticBin(p1=0.43,p2=0.58,B=0.48,alpha= 0.05,power=0.9)	通过SSizeLogisticBin函数计算单因素Logistic回归样本量(实例1)
ρsquare<-0.2	指定公式(3)中 $R^2_{1,234\cdots p}$ 的值
n<-n/(1-ρsquare)	通过公式(3)计算最终样本量(实例2)

表2 Logistic回归样本量估算的PASS操作过程及说明

操作选项	说 明
在顶端依次点击Procedures▶Regression▶Logistic Regression	选择Logistic回归
Find(Solve For)—选择N	设置计算样本量
Power—0.9	设置功效值,默认0.9
Alpha—0.05	设置置信水平
P0(baseline Probability that Y=1)—填入0.43	实例1中的$p0$
Use P1 or Odds Ratio—选择P1	也可选择Odds Ratio
P1—填入0.58	实例1中的$p1$
R-Squared of X1 with Other X's—填入0.07*	实例2中的 $R^2_{1,234\cdots p}$,默认0
X1(Independent Variable of Interest)—选择Binary(X=0or1)	设置X为二分类变量
Percent of N with X1=1—填入48	实例1中的B
Alternative Hypothesis—选择Two-Sided	双侧检验

*计算实例1则不用设置R-Square

实现Logistic回归样本量的计算以及提供了相应的SAS代码。不同的软件所采用的公式可能有一定差别,因为统计学家针对Logistic回归样本量估算提出过不同的公式,建议在样本量估算的文档中同时标注公式的出处。

4 结 语

EPV通常被认为是Logistic回归模型中参数估计效果的主要决定因素,在估算样本量时往往被格外重视。但是影响Logistic回归模型中参数估计效果的因素有很多,比如因变量与自变量之间关系的强度、自变量之间的相关性(即共线性)等[8],van Smeden等认为对每个自变量EPV取10作为二分类Logistic回归样本量,低估了合理的样本量水平,建议通过Firth's校正予以改善[9]。Vittinghoff等也认为EPV取10,会致所得样本量偏低[10]。本文建议在采用经验法计算Logistic回归样本量时,应同时兼顾所有自变量不同暴露水平下结局为阳性、阴性者的人数都足够多。

相较于经验法,更提倡使用公式法来估算样本量,并且建议使用影响面较大的权威软件包。本文介绍的两种软件各有利弊,比如R免费,而PASS则可提供更为详尽的输出。

[参 考 文 献]

[1] PEDUZZI P, CONCATO J, KEMPER E, et al. A simulation study of the number of events per variable in logistic regression analysis[J]. J Clin Epidemiol, 1996,49(12):1373-1379.

[2] 方积乾. 医学统计学与电脑实验[M]//上海:上海科学技术出版社, 2012:273-274.

[3] 方积乾. 卫生统计学[M]//北京:人民卫生出版社, 2012:399-400.

[4] WHITTEMORE A S. Sample size for logistic regression with small response probability[J]. J Am Stat Assoc, 1981, 76(373):27-32.

[5] HSIEH F Y. Sample size tables for logistic regression[J]. Stat Med, 1989,8(7):795-802.

[6] HSIEH F Y, BLOCH D A, LARSEN M D. A simple method of sample size calculation for linear and logistic regression[J]. Stat Med, 1998,17(14):1623-1634.

[7] 周映雪, 潘蕾, 陈方尧, 等. 样本量估计及其在nQuery软件上的实现——回归分析(一)[J]. 中国卫生统计, 2013,(5):762-765.

[8] HEINZE G, SCHEMPER M. A solution to the problem of separation in logistic regression[J]. Stat Med, 2002,21(16):2409-2419.

[9] van SMEDEN M, de GROOT J A, MOONS K G, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis[J]. BMC Med Res Methodol, 2016,16(1):163.

[10] VITTINGHOFF E, MCCULLOCH C E. Relaxing the rule of ten events per variable in logistic and Cox regression[J]. Am J Epidemiol, 2007,165(6):710-718.

[收稿日期] 2018-01-16