

拟合logistic回归方程的步骤和注意事项

作者：张倩 来源：爱科学

本文原地址：<https://www.iikx.com/news/statistics/5964.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

Logistic 回归：实际上属于判别分析，因拥有很差的判别效率而不常用。

1. 应用范围：

适用于流行病学资料的危险因素分析

实验室中药物的剂量-反应关系

临床试验评价

疾病的预后因素分析

2. Logistic 回归的分类：

按因变量的资料类型分：

二分类

多分类

其中二分较为常用

按研究方法分：

条件 Logistic 回归

非条件 Logistic 回归

两者针对的资料类型不一样，后者针对成组研究，前者针对配对或配伍研究。

3. Logistic 回归的应用条件是：

独立性。各观测对象间是相互独立的；

LogitP 与自变量是线性关系;

样本量。经验值是病例对照各 50 例以上或为自变量的 5-10 倍(以 10 倍为宜), 不过随着统计技术和软件的发展, 样本量较小或不能进行似然估计的情况下可采用精确 logistic 回归分析, 此时要求分析变量不能太多, 且变量分类不能太多;

当队列资料进行 logistic 回归分析时, 观察时间应该相同, 否则需考虑观察时间的影响(建议用 Poisson 回归)。

4. 拟合 logistic 回归方程的步骤:

对每一个变量进行量化, 并进行单因素分析;

数据的离散化, 对于连续性变量在分析过程中常常需要进行离散变成等级资料。可采用的方法有依据经验进行离散, 或是按照四分、五分位数法来确定等级, 也可采用聚类方法将计量资料聚为二类或多类, 变为离散变量。

对性质相近的一些自变量进行部分多因素分析, 并探讨各自变量(等级变量, 数值变量)纳入模型时的适宜尺度, 及对自变量进行必要的变量变换;

在单变量分析和相关自变量分析的基础上, 对 P (常取 0.2, 0.15 或 0.3) 的变量, 以及专业上认为重要的变量进行多因素的逐步筛选; 模型程序每拟合一个模型将给出多个指标值, 供用户判断模型优劣和筛选变量。可以采用双向筛选技术: a 进入变量的筛选用 score 统计量或 G 统计量或 LRS(似然比统计量), 用户确定 P 值临界值如: 0.05、0.1 或 0.2, 选择统计量显著且最大的变量进入模型; b 剔除变量的选择用 Z 统计量 (Wald 统计量), 用户确定其 P 值显著性水平, 当变量不显著者, 从模型中予以剔除。这样, 选入和剔除反复循环, 直至无变量选入, 也无变量删除为止, 选入或剔除的显著界值的确定要依具体的问题和变量的多寡而定, 一般地, 当纳入模型的变量偏多, 可提高选入界值或降低剔除标准, 反之, 则降低选入界值、提高删除标准。但筛选标准的不同会影响分析结果, 这在与他人结果比较时应当注意。

在多因素筛选模型的基础上, 考虑有无必要纳入变量的交互作用项; 两变量间的交互作用为一级交互作用, 可推广到二级或多级交互作用, 但在实际应用中, 各变量最好相互独立 (也是模型本身的要求), 不必研究交互作用, 最多是研究少量的一级交互作用。

对专业上认为重要但未选入回归方程的要查明原因。

5. 回归方程拟合优劣的判断(为线性回归方程判断依据, 可用于 logistic 回归分析)

决定系数 (R^2) 和校正决定系数 (Logistic 回归分析简介 - 初学乍练 - 教学科研), 可以用来评价回归方程的优劣。 R^2 随着自变量个数的增加而增加, 所以需要校正; 校正决定系数 (Logistic 回归分析简介 - 初学乍练 - 教学科研) 越大, 方程越优。但亦有研究指出 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, 因此在 logistic 回归中不适合。

Cp 选择法：选择 Cp 最接近 p 或 p+1 的方程(不同学者解释不同)。Cp 无法用 SPSS 直接计算，可能需要手工。1964 年 CL Mallows 提出：

Cp 接近(p+1)的模型为最佳，其中 p 为方程中自变量的个数，m 为自变量总个数。

AIC 准则：1973 年由日本学者赤池提出 AIC 计算准则，AIC 越小拟合的方程越好。

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 、偏差 (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 、偏差 (deviance)、Hosmer-Lemeshow (HL) 指标值均服从 χ^2 分布， χ^2 检验无统计学意义 ($P > 0.05$) 表示模型拟合的较好， χ^2 检验有统计学意义 ($P \leq 0.05$) 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。

6. 拟合方程的注意事项：

进行方程拟合对自变量筛选采用逐步选择法
[前进法(forward)、后退法(backward)、逐步回归法(stepwise)]
时，引入变量的检验水准要小于或等于剔除变量的检验水准；

小样本检验水准 定为 0.10 或 0.15，大样本把 定为 0.05。值越小说明自变量选取的标准越严；

在逐步回归的时可根据需要放宽或限制进入方程的标准，或硬性将最感兴趣的研究变量选入方程；

强影响点记录的选择：从理论上讲，每一个样本点对回归模型的影响应该是同等的，实际并非如此。有些样本点(记录)对回归模型影响很大。对由过失或错误造成的点应删去，没有错误的强影响点可能和自变量与应变量的相关有关，不可轻易删除。

多重共线性的诊断(SPSS 中的指标)：a 容许度：越近似于 0，共线性越强;b 特征根：越近似于 0，共线性越强;c 条件指数：越大，共线性越强；

异常点的检查：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。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。

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 $H(\text{hat matrix diagnosis})$ 、Cook 距离、DFBETA、Score 检验统计量等。这五个指标中，Pearson

残差、Deviance

残差可用来检查特异点，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， 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，其值越大，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，首先应根据专业知识、数据收集的情况，分析其产生原因后酌情处理。如来自测量或记录错误，应剔除或校正，否则处置就必须持慎重态度，考虑是否采用新的模型，而不能只是简单地删除就算完事。因为在许多场合，异常点的出现恰好是我们探测某些事先不清楚的或许更为重要因素的线索。

7. 回归系数符号反常与主要变量选不进方程的原因：

存在多元共线性;

有重要影响的因素未包括在内;

某些变量个体间的差异很大;

样本内突出点上数据误差大;

变量的变化范围较小;

样本数太少。

8. 参数意义

Logistic 回归中的常数项(b_0)表示，在不接触任何潜在危险/保护因素条件下，效应指标发生与不发生事件的概率之比的对数值。

Logistic 回归中的回归系数(b_i)表示，其它所有自变量固定不变，某一因素改变一个单位时，效应指标发生与不发生事件的概率之比的对数变化值，即 OR 或 RR 的对数值。需要指出的是，回归系数的大小并不反映变量对疾病发生的重要性，那么哪种因素对模型贡献最大即与疾病联系最强呢? $(\ln L(t-1) - \ln L(t))$ 三种方法结果基本一致。

存在因素间交互作用时，Logistic 回归系数的解释变得更为复杂，应特别小心。

模型估计出

OR，当发病率较低时，OR RR，因此发病率高的疾病资料不适合使用该模型。另外，Logistic 模型不能利用随访研究中的时间信息，不考虑发病时间上的差异，因而只适于随访期较短的资料，否则随着随访期的延长，回归系数变得不稳定，标准误增加。

9. 统计软件

能够进行 logistic 回归分析的软件非常多，常用的有 SPSS、SAS、Stata、EGRET (Epidemiological Graphics Estimation and Testing Package) 等。

更多 统计方法 请访问 <https://www.iikx.com/news/statistics/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发