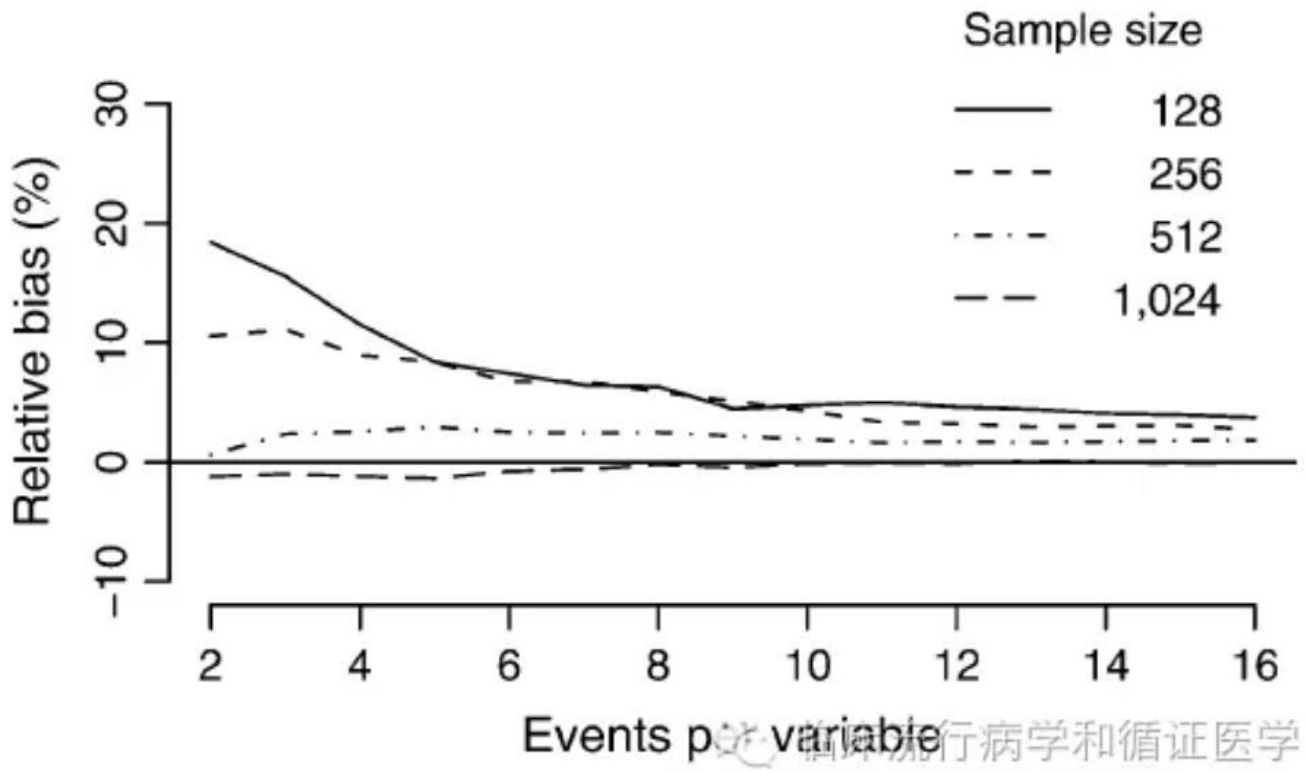
二分类logistic回归中纳入多少自变量合适？

作者：陶立元 赵一鸣 来源：临床流行病学和循证医学

本文原地址：https://www.iikx.com/news/statistics/1980.html

在对临床数据的探索分析工作中，我们经常会
使用**多因素logistic回归**
分析去探索疾病的危险因素，也可以用它来做预测。但是每每在使用logistic回归分析的时候，我们都会纠结应该选哪些作为自变量呢?选多少个合适呢?

选哪些作为自变量，这个问题比较简单。一般情况下，我都是选择那些单因素分析中与因变量有关的自变量进入回归方程。但同时需要提醒，如果某些自变量从作用机制或临床经验上来看，跟因变量之间关系密切，此时也应该将其选入方程，即使单因素分析没有意义。



选谁确定了，剩下的就是选多少个合适了。假设我们的研究对象有m个，需要选择的自变量有n个。如果此时m很大且n很小，那么一般情况都可以选进来;如果此时m相对于n不够大，则不可以

盲目的将n个自变量都丢进方程。

至于m和n之间的关系，有教科书上指出：经验上病例和对照的人数应该至少各有30-50例，方程中自变量的个数越多需要的研究对象例数也越大。

1985年，Harrell等人在其研究论文中指出：从理论上来讲，多因素分析中至少需要的EPV数量为10-20个。EPV(events per variable)，就是每个自变量所需要的事件数。举个例子，也就是研究对象中较少组的数量，除以自变量的个数所得到的。如果m个研究对象中有m1个人有疾病，m2个人无疾病(m1+m2=m)，同时m1小于m2，此时EPV=m1/n(n为自变量个数)。

在1996年，Peter等人针对logistic回归做了计算机模拟试验，探索EPV对logistic回归结果的影响。他们基于一个真实的心血管疾病研究数据，包含673个病人，其中有252人死亡。采用随机抽样的方法，分别设定EPV为2，5，10，15，20和25，计算logistic回归结果并于原始结果比对。其研究结果指出：当EPV大于等于10的时候，回归结果比较稳定，且与原始结果较为一致。当EPV小于10时，其偏回归系数偏倚较大。

到2006年，Eric等人发表题为“ 放松Logistic和Cox回归中10个EPV的规则 ”的论文，文中指出仅仅通过几个计算机模拟试验就确定EPV≥10的规则未免太过保守。作者通过更多数据的计算机模拟，以及对更多影响因素的考虑，指出EPV≥5即可获得可接受的结果，同时EPV过小，可考虑采用bootstrap的方法进行敏感性分析。

针对这个EPV的数量应该多少合适，不同的研究有不同的观点，2009年Karel等人在BMJ发表论文时指出：EPV大于等于10时较为合适，尽管有人提出EPV可以更小。

在此小编建议，在使用logistic回归做危险因素探索的时候考虑EPV≥10，应该是足够的了。注意此处是EPV≥10，而不是m/n≥10。如果拿logistic回归结果来做预测呢?个人觉得可能需要更大的EPV。除此之外，我们还应该考虑其他的预测建模方法(如随机森林等)，同时对预测模型进行严格的评价。