回归分析中自变量的选择问题

作者：张华，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/6278.html

回归分析中自变量的选择问题
。回归分析是我们经常用到的分析方法，在非RCT的研究中，如果没有做回归分析，可能会让杂志编辑觉得非常low。因此在咨询过程中，很多人拿着数据咨询怎么做回归分析。而作回归分析时因变量一般是明确的，存在比较多的一个问题是：哪些变量应该选入方程模型?有什么原则?对于上面这个问题，有人认为单因素分析时P<0.05的变量应该选入。这个观点有其合理之处，但不全面。

在回答这个问题之前，需要明确三个问题。**第一，回归分析的目的**
。一般我们做回归分析的目的是调整混杂因素的干扰，从这个目的出发，所有可能的混杂因素都应该进行调整，以去除混杂因素的干扰；
第二，混杂作用
存在的前提条件是混杂因素与关
心的研究因素和结局都有相关性
，即混杂因素在研究分组间不平衡。例如年龄可能是很多研究中的混杂因素，但在某研究中如果

两组中的病例年龄完全相等，年龄也不会有混杂作用，即如果单因素分析中P值较大，也不会有混杂作用，因此也不需要进行调整；**第三，统计分析中P值的意义**。在统计分析中P是Probability的首字母，即统计推断中原假设成立的概率。一般的统计分析中，原假设都是相等或不相关。当P<0.05，即原假设成立的概率小于0.05，即认为是小概率事件，即发生的可能性较小，可认为均值(或率)不相等或两变量存在相关性。但P只是一个概率，我们统计学中给它找了一个比较公认的界值：0.05.但如果p=0.06就不会存在相关性吗?显然也不一定。即P>0.05也不能确定混杂作用是否存在。

回到开始提到的问题，我个人一般把回归分为探索性回归分析和验证性回归分析，对二者区别对待。在探索性回归分析中，我们不知道哪些因素是混杂因素，因此我们可以先做单因素分析，将单因素分析中“ 有意义 ”的因素作为自变量进行回归分析。这里的有意义不是P<0.05，我认为可以放宽到P<0.1或者P<0.2;如果样本量很大时，也可以再放宽P。对于验证性回归分析，我们已经知道某些因素是混杂因素，某些因素可能是混杂因素。此时已经确定是混杂因素的因素，无论单因素分析P是多大，都应该进入自变量;对于不确定的混杂因素，可以根据单因素分析的P值进行选择，标准同上。