SPSS：二元Logistic回归中自变量的处理和解读——无序多分类变量的处理

作者：writer 来源：临床流行病学和循证医学

本文原地址：https://www.iikx.com/news/statistics/6282.html

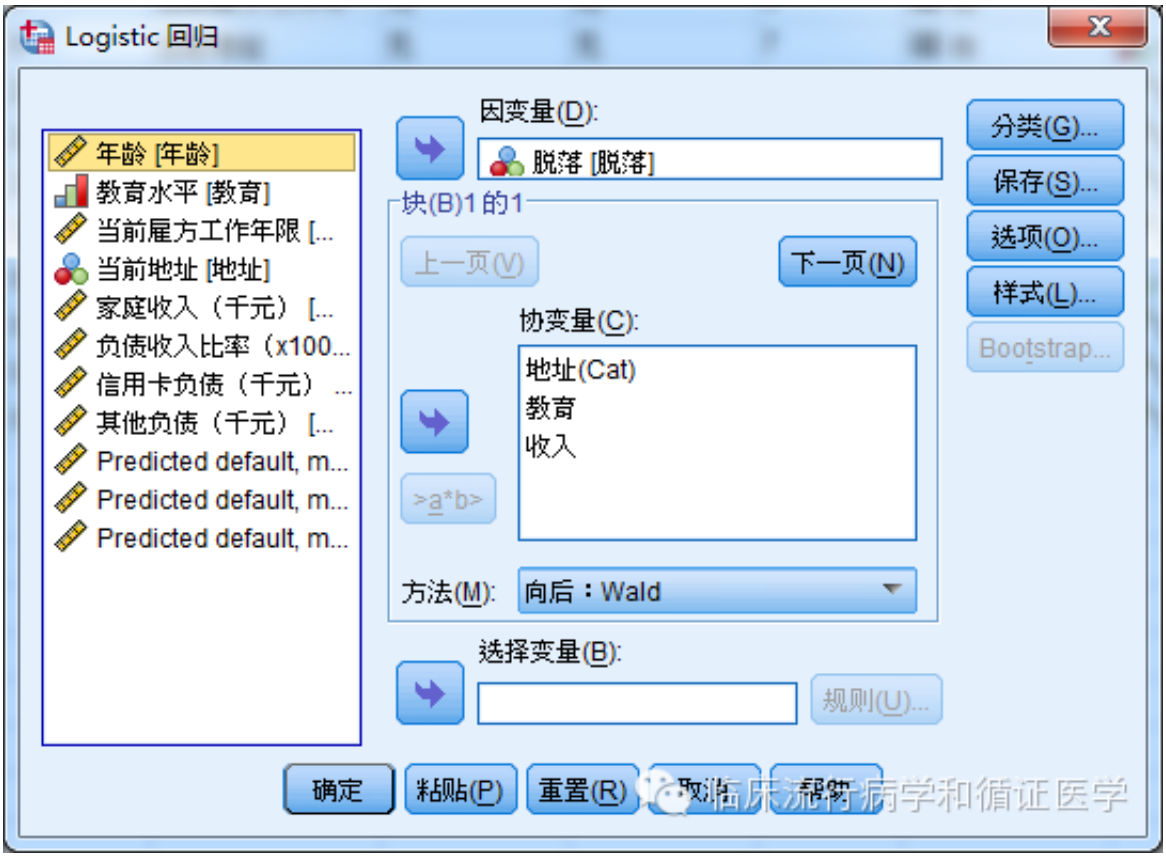
本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

SPSS：二元Logistic回归
中自变量的处理和解读——无序多分类变量的
处理
。在实际进行Logistic回归的过程中，我们会遇到很多现实问题，尤其是在自变量的处理上经常拿不准。比如自变量是否一定是二分类变量?多分类变量作为自变量如何处理?连续变量是否可以当做自变量?多分类变量和连续变量做自变量的时候，结果OR值如何解读等等。

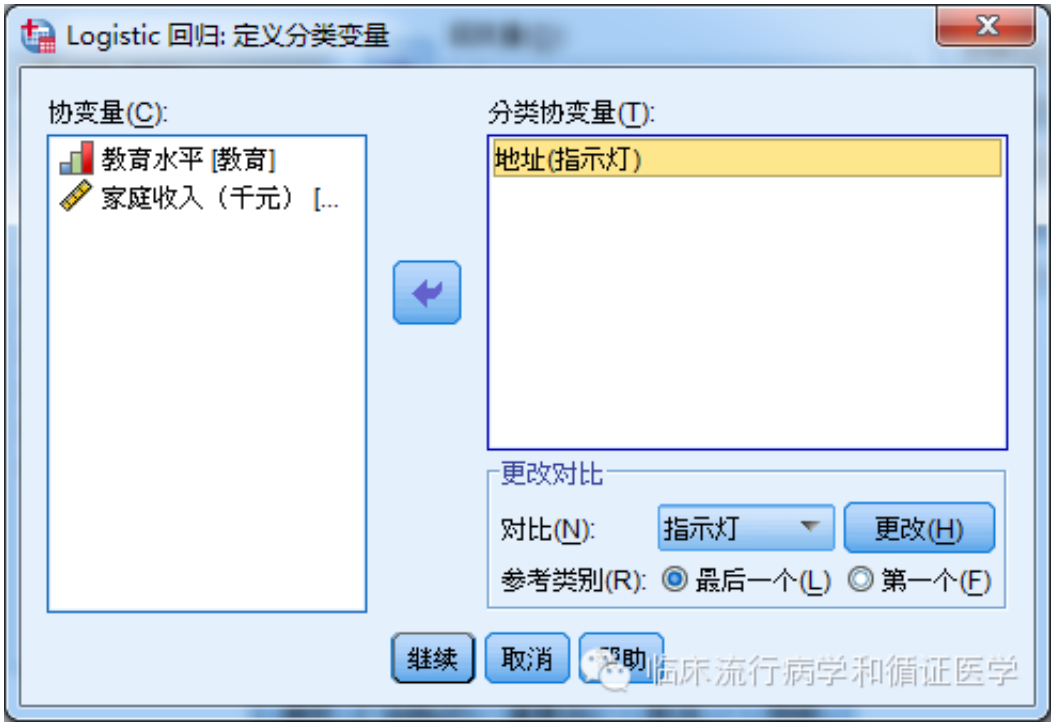
无序多分类变量作为自变量的设置

所谓的**无序多分类变量**
，通常是指变量有多种情况，但各种情况的赋值间不存在大小上的差异。比如用数值代表患者的居住地(1=北京，2=上海，3=广州，4=成都……)。此时数值的大小并没有实际意义，因此我们需要用到前面介绍过的哑变量的概念。在SPSS中，并不用手动生成哑变量，而只是简单的在Logistic回归的“ 分类 ”选项中进行设置就好了。

举个例子：我们希望探讨教育程度、家庭收入、居住地这三个因素(自变量)是否与患者脱落(因变量)之间存在关系。在建立Logistic回归模型分析的时候，我们会这样引入变量：



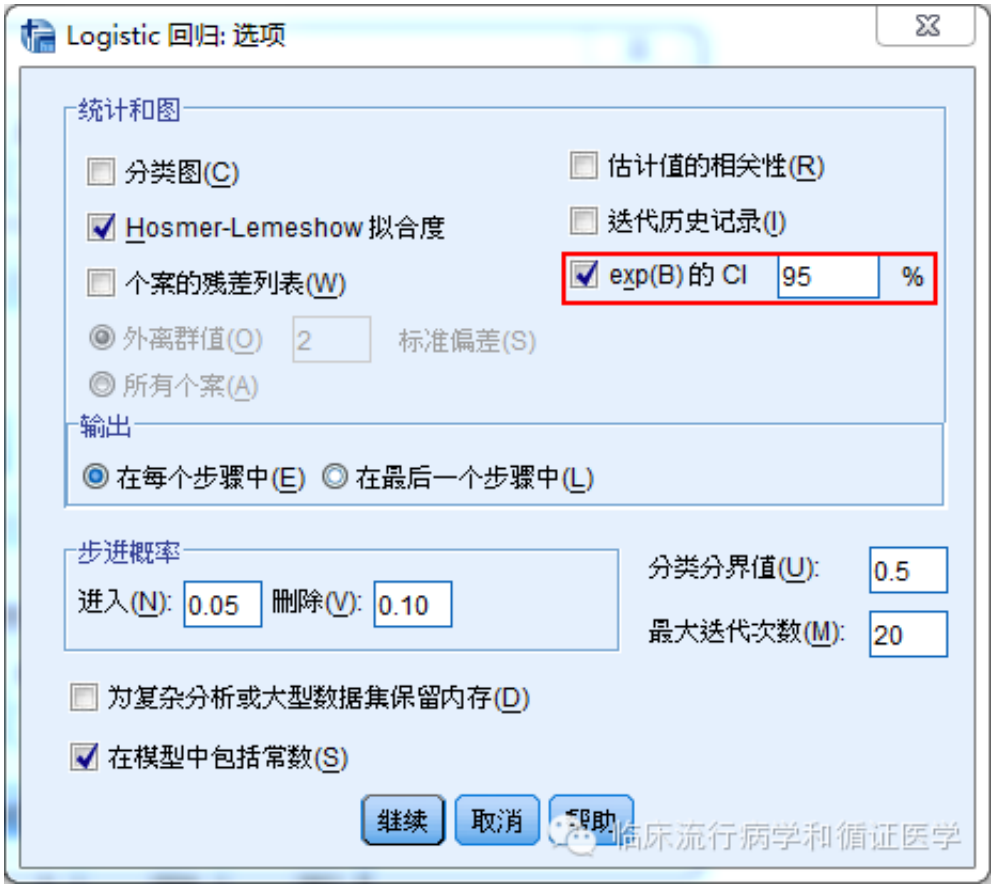
其中“ 地址 ”变量是无序多分类变量。我们要通过上面对话框右上角的“ 分类 ”选项将其转换为哑变量。



如图所示，将地址变量选中，之后通过中间的方向符号丢到右边“ 分类协变量 ”区域。下面“ 更改对比 ”框中的“ 对比(N): ”可以选择“ 指示符 ”或是“ 简单 ”，参考类别选择任意一个皆可。

点击继续，地址变量就已经成功被标记为按哑变量处理了。这一步骤对于无需多分类变量可以说是必须选项，否则SPSS会认为我们的数据的数值之间存在大小的差异，比如2比1大1个单位，4比1大3个单位……导致出现错误的分析结果。

同样别忘了在选项中选择输出OR值的可信区间，如下图：



变量筛选方法我们选择向后的方法(图1)。

结果解读

Hosmer 和 Lemeshow 检验

步长 (T)	卡方	自由度	显著性
1	9.279	8	.319

从Hosmer-Lemeshow检验看，模型的拟合还是不错的，p>0.1。

再看模型中的变量：

方程式中的变量

		B	S.E.	Wald	自由度	显著性	Exp(B)	95% C.I.用于 EXP(B) 下限	95% C.I.用于 EXP(B) 上限
步骤 1[a]	地址			8.759	3	.033			
	地址(1)	-.882	.333	6.992	1	.008	.414	.215	.796
	地址(2)	-.862	.335	6.636	1	.010	.422	.219	.814
	地址(3)	-.961	.336	8.198	1	.004	.382	.198	.738
	教育	.323	.094	11.893	1	.001	1.381	1.149	1.659
	收入	-.007	.003	5.591	1	.018	.993	.987	.999
	常量	-.473	.342	1.911	1	.167	.623		

临床流行病学和循证医学

我们看到，模型中的地址变量已经被当做了4个变量处理了，在结果输出中占了4行。其中第一行是参考类别，也就是我们在“ 分类 ”选项中选择的“ 最后一个 ”或是“ 第一个 ”。在这个例子中，我们选择的是最后一个，也就是“ 地址 ”变量中“ 4 ”对应的地址。具体对应关系如下：

参考类别	原始赋值	对应结果
最后一个	1	地址（1）
	2	地址（2）
	3	地址（3）
	4	地址
第一个	1	地址
	2	地址（1）
	3	地址（2）
	4	地址（3）

结果的解读我们就不多说了，详细解读可参照“ SPSS：Logistic回归(Logistic regression)概述 ”。我们看到“ 地址 ”对应的一系列哑变量中，参照组是没有参数估计和OR值(Exp(B))的。原因很简单，参照嘛，本身就是被别人比的，OR自然也应该是1。其实在Logistic回归中，我们可以吧参照组想象为其他哑变量(地址1~3)的共用“ 0 ”。

也正因此，地址(1)~地址(3)的OR值，其意义也是：当患者来源于“ 地址(1) ”时，其脱落的风险是患者来源于“ 地址 ”(此处为参照项)时的OR倍。

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转发