哑变量

1. http://ttdoc.cn/article/486.jhtml

本次我们讨论一下什么是哑变量（Dummy variable）。哑变量不是哑巴了的变量，如果是那么应该叫Mute variable更合适。

那么什么是哑变量呢？NormanR. Draper在他的《Applied Regression Analysis》（ISBN 0-471-17082-8）第14章中提出：在统计和计量经济学中，尤其是在回归分析中，哑变量是指使用0或1去代表某种可能影响结局的事情的发生与否。

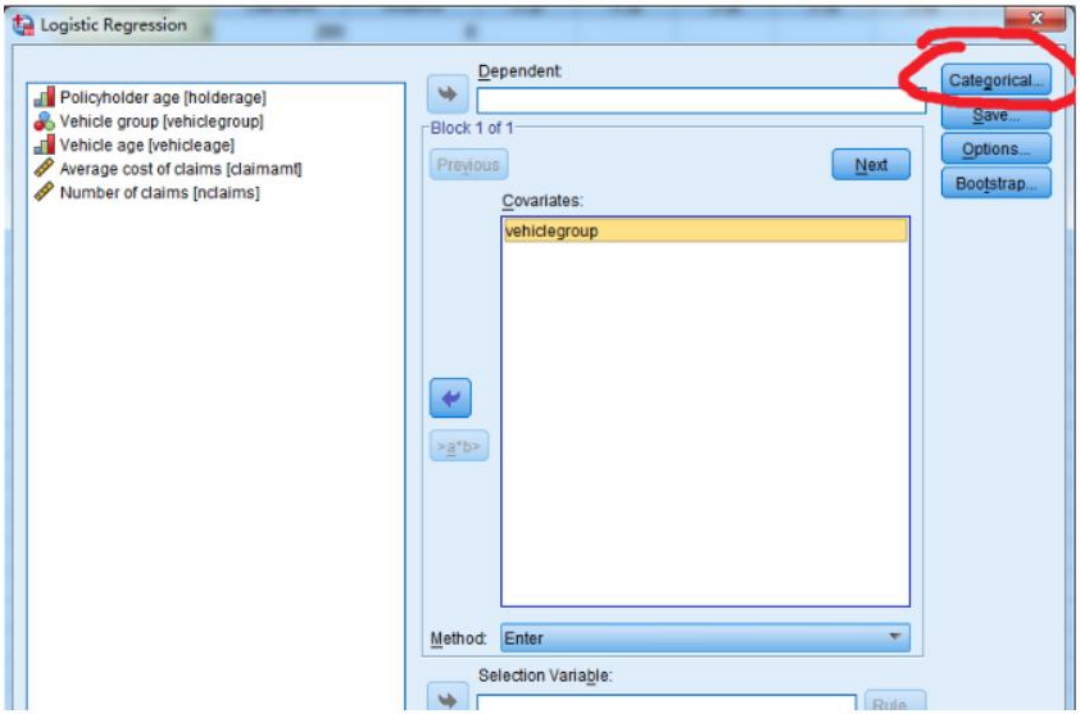
在医学统计分析中，许多变量是可以定量测量的，如身高、血糖、肿块大小等等，但也有一些变量是无法量化的分类变量，如性别、胃癌的病理类型（腺癌、粘液腺癌、印戒细胞癌和特殊类型癌）等。对于性别我们就需要用一个数值来表示它们，常用0和1来表示，如男性为1，女性为0。这样将性别“量化”的方法是为了在统计分析模型中纳入性别的影响，提高模型的精读。广义上说这就是对哑变量（Dummy variable）的应用。

对于dummyvariable的翻译应该叫虚拟变量更为合适，就是用一些数值上虚拟的值（0或1）去代替那些无法直接纳入统计分析的变量。对于性别这种两分类的情况，我们只需要设置0和1即可，那么对于胃癌的病理类型（有4个分类）呢，我们就需要用一系列数值来表示了，它应该设置成如下结构：

	D1	D2	D3
腺癌	0	0	0
粘液腺癌	1	0	0
印戒细胞癌	0	1	0
特殊类型癌	0	0	1

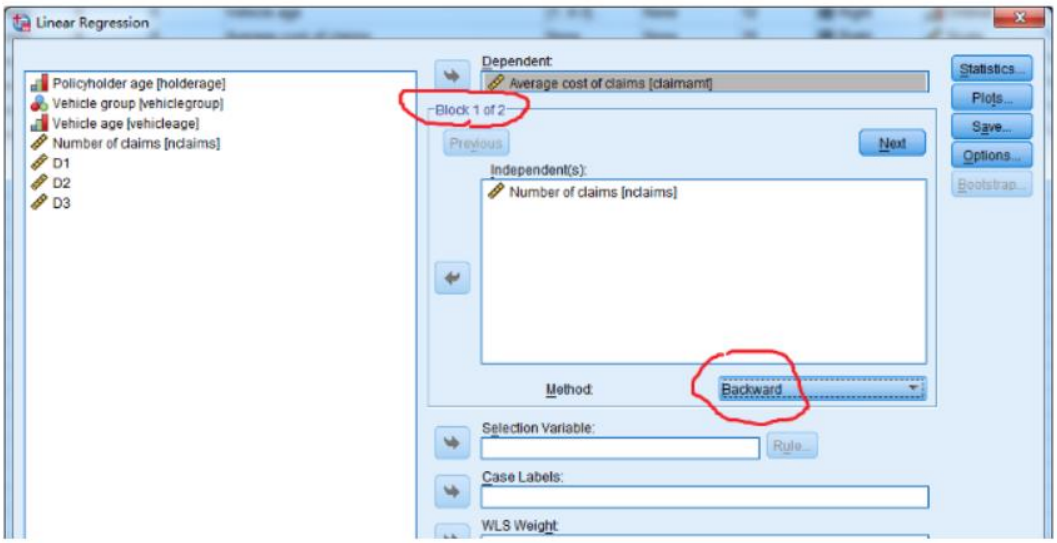
也就是说我们引进了3个变量来表达上述4种病理类型，3个变量分别是D1、D2和D3。为什么不引入4个变量呢？伍德里奇在《计量经济学导论》中说“如果某个定性变量有m种互相排斥的类型，则模型中只能引入m-1个虚拟变量，否则会陷入虚拟变量陷阱，产生完全共线性。”感兴趣的可以了解一下。不感兴趣的，知道某一个变量如果有m个互斥的分类设成m-1个哑变量就ok了。

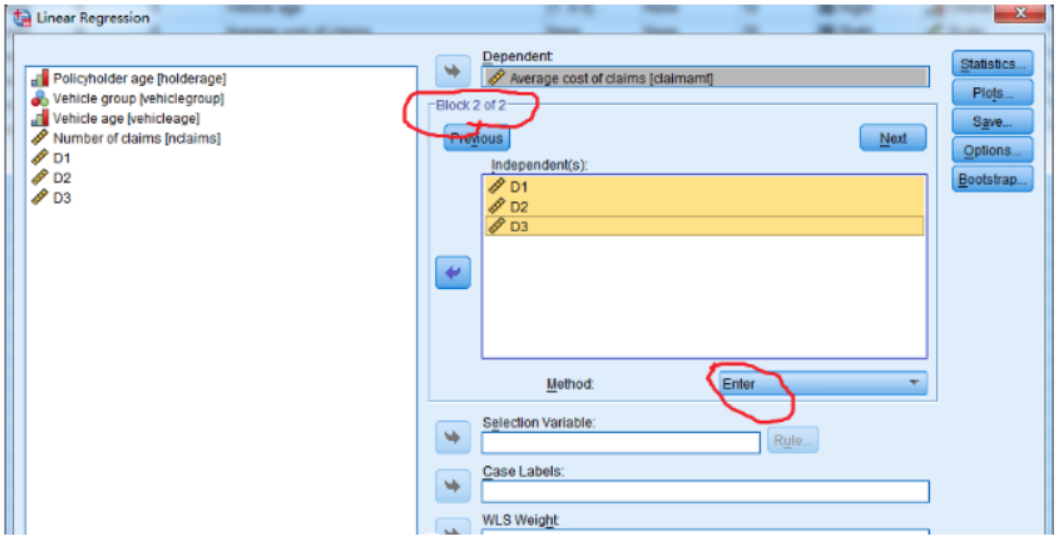
看完上面的介绍，大家应该知道了哑变量是干什么的，设置哑变量就是为了将分类变量数量化，然后纳入分析模型进行分析。哑变量的引入，扩大了回归分析中自变量的纳入范围，也就是说分类变量也可以纳入回归分析啦。熟悉SPSS的人会说，在logistic回归中有define categorical variables过程去定义哑变量，那么在线性回归中怎么办呢？



SPSS线性回归中纳入哑变量的方法就只能靠我们自己去人工设定了，待设定完了之后再将设定好的哑变量纳入方程。以胃癌的病理类型为例，设定方法如下：设定为D1D2D3三个哑变量。

由于一组哑变量表示的是同一个变量，所以它们进入方程时需要“同进同出”，这时候在变量的选择上我们就需要使用SPSS纳入自变量的BLOCK功能，即将哑变量和非哑变量分别纳入不同的BLOCK，且哑变量所在BLOCK的变量筛选方法为enter。其纳入的过程如下：





在设置哑变量时大家需要注意的是：

1、哑变量的设置也算是引入了更多的自变量，需要满足回归样本量和自变量之间的关系，也就是说设置了哑变量的同时，无形地增加了对样本量的要求；

2、如果原分类变量是有序变量的情形，如病情分为轻、中、重三种，可以设置哑变量，也可以按连续变量进行处理。如果样本量足够，设置哑变量；如果样本量不够，按连续变量处理。

2） Dummy variable (statistics)

https://en.wikipedia.org/wiki/Dummy_variable_(statistics)

In statistics and econometrics, particularly in regression analysis, a dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. They can be thought of as numeric stand-ins for qualitative facts in a regression model, sorting data into mutually exclusive categories (such as smoker and non-smoker).