SPSS:病例-对照匹配（case-control matching）的实现过程

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/2059.html

匹配
(Matching)又称配比，是指为每一个处理组的研究对象选择在某些特征上一致的对照组对象。匹配的目的是控制某些因素对处理效应的影响，从而评价处理因素对结局的真实作用。经典的匹配方式是首先选定一些匹配变量，然后逐一对变量进行匹配，为每一个病例找出同期的合适的对照，这个过程就像找对象。近年来还有一种匹配方法被广泛应用，即**倾向性评分匹配**(Propensity Score Matching，PSM)。它是通过某种模型求得多个协变量的综合倾向性得分，再按照得分是否接近进行匹配。

如果上面的过程不好理解，那么举个例子。比如替某女孩找对象，该女孩列出以下3个条件：1年龄要与自己相差不大(±2岁)，2民族与自己一致，3学历与自己一致。那么如果是经典的匹配方法，首先是按照这些条件删选男生，然后在符合条件的男生中随机抽取一个男孩介绍给这个女孩。如果是PSM呢，首先给出一个评分方法，比如总分=0.8*年龄+2.3*民族+1.6*学历，然后对该女孩和众多男孩分别求得分，找出一个得分与该女孩最为接近的男孩，介绍他们认识。如果得分接近的有很多怎么办?还是选择一个。

PSM从理论来说比经典的病例-对照匹配方法要更为合理和科学，因为其配对过程中考虑了每个因素不同的影响能力(权重)。但是其实现过程中也存在一些问题，比如：如何选择合适的匹配因素?采用何种方法训练模型?采用何种匹配方法(邻近匹配、卡钳匹配、马氏距离匹配等)?本文不谈PSM，说说**经典的匹配方法，如何通过SPSS实现**。

例：目前有病例895人，对照2532人，要求按照年龄、职业和BMI组别进行匹配，即为每一个病例找一个跟他在年龄、职业和BMI水平上一样的对照，进行病例对照研究。

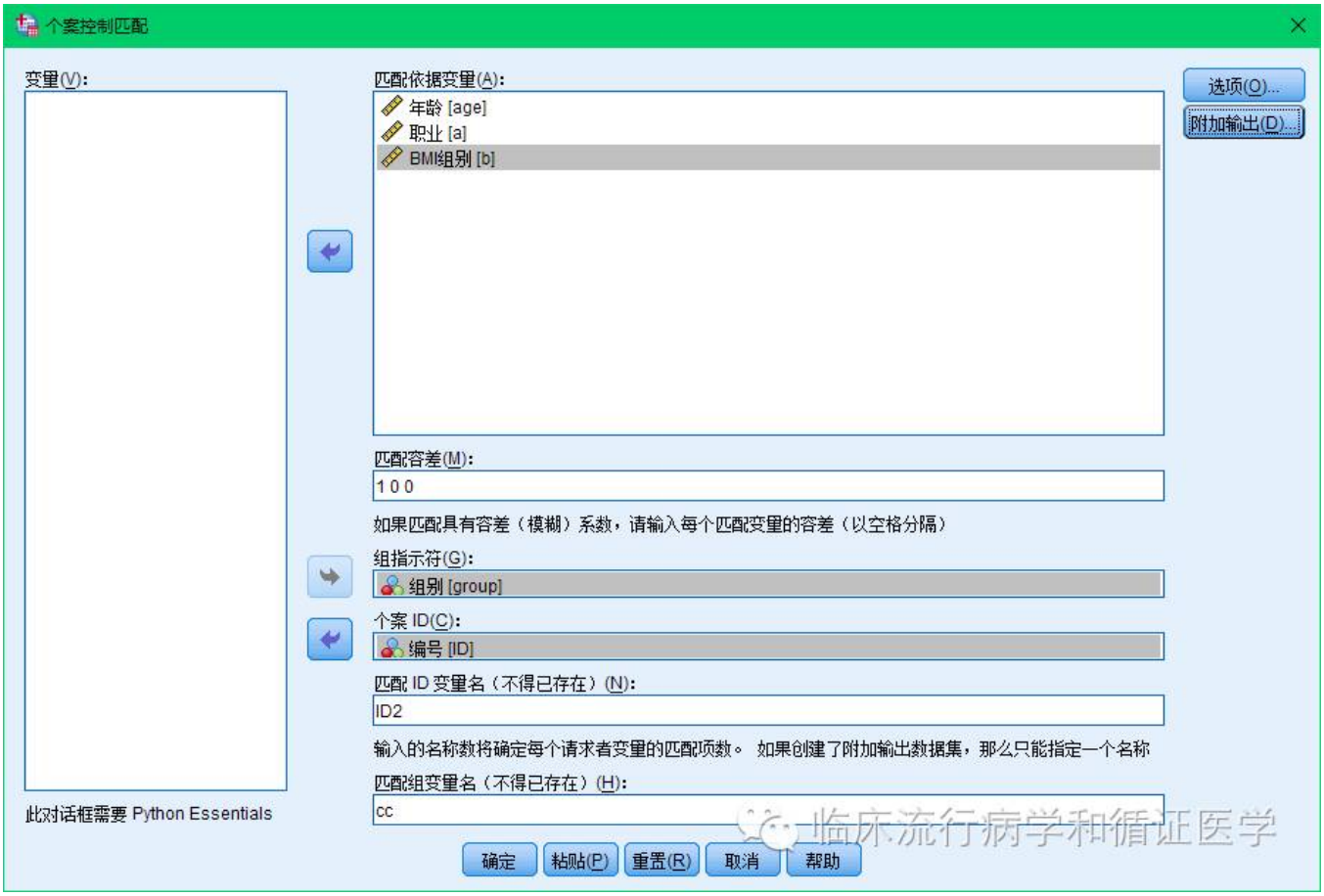
在没有匹配之前，其数据是这样的，在年龄、职业和BMI组别上两组间差异也有统计学意义，如下：

组别

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	对照组	2532	73.9	73.9	73.9
	病例组	895	26.1	26.1	100.0
	Total	3427	100.0	100.0	

在SPSS中点击如下按键，它的英文是Case-control matching，不知道为什么在中文版中翻译成了个案控制匹配，难道不应该是病例-对照匹配吗?

出现如下对话框，并按照要求选入变量：



在上图中匹配容差是指你允许的各个变量的差别，此处年龄选1岁，其他变量容差为0，即不允许有差别。另外匹配ID和匹配组变量名，你需要自己给他随便起个名字，此处起名为ID2和cc。

然后点击选项，你可以选择是否进行放回抽样，以及匹配的方式，另外你可以设置随机种子数，保障匹配过程可以重新，如下图：

再点击“ 附加输出 ”，选中创建新的匹配项数据集，并给数据集起个名，如下图：

这些都做完点确定，会出现以下结果：

Case Control Matching Statistics

Match Type	Count
Exact Matches	621
Fuzzy Matches	83
Unmatched Including Missing Keys	191
Unmatched with Valid Keys	191
Sampling	without replacement
Log file	none
Maximize Matching Performance	no

表示精确匹配了621个，模糊匹配了83个，共匹配上704例，没有匹配上的病例有191个。同时会生成一个名为new的数据集，该数据集就是被选中的对照组数据库，同时在原来的数据集里生成

了刚才咱们命名的ID2变量，该变量有值者就是被匹配上的病例组。

将上述两个数据库合并，并删除那些ID2变量为缺失的个案，就形成了匹配后的数据库。我们再来看一下匹配后的情况，组间在年龄、职业和BMI分组上差异也没有统计学意义了。

组别

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	对照组	704	50.0	50.0	50.0
	病例组	704	50.0	50.0	100.0
	Total	1408	100.0	100.0	

SPSS:病例-对照匹配(case-control matching)的实现过程