病例对照配对

病例对照匹配目的是配消除匹配因素的混杂，同时也在一定程度上控制了与匹配因素相关的一些潜在因素的混杂。给定一个病例，从数据库里找出满足配对条件的所有可能的对照，然后根据匹配数随机选择对照。如 1：1 匹配，随机选一个作对照；1：2 匹配，随机选 2 个配对。

在具体操作时常遇到可以用来匹配的资源竞争的问题，当可以用来配对的对照组样本量有限时，有些病例不一定能找到配对，这时既要保持随机性，又要尽可能地为更多的病例找到对照，方法不同，结果大为不同。如对照 A 可以与病例 B 配对，也可以与病例 C 配对，但如果与 B 配对了，C 就可能找不到对子了，如果把 A 与 C 配对，B 还有可能找到对子。这时候就应该把 A 配给 C 而不是 B。

本模块采用了智能化的优先顺序进行匹配，尽可能最大限定地为每个病例找到对照，匹配条件包括变量取值范围与倾向性评分匹配。在计算倾向性评分时，对连续性变量可以选择是否要用平滑曲线拟合该变量与应变量的关系。

使用本模块时要给定匹配变量的差异范围，如连续性的年龄变量是匹配变量，匹配时不可能要求病例与对照年龄完全相同，那么相差多大范围可以匹配呢？这就是差异范围。如果按倾向性评分匹配，通常取相差 0.05 的范围，如果供匹配的样本量足够大，可以适当减小差异范围。

本模块可以实现 1:n 匹配，直接修改 n 即可；也可以实现 n:1 匹，将 n 改为 1/n 即可，如要实现 2:1 匹配，输入 n 值为 1/2（注意中间的除号应在英文状态下输入）。

本模块输出完全匹配好的数据，以供下一步分析时直接使用。输出数据中除了用于匹配的变量外，用户可以指定需要输出的变量。如按 SEX，AGE 匹配，输出数据中包含 HEIGHT，WEIGHT 等其它用于下一步分析的变量。

本模块自动对完全匹配的数据进行组间比较，计算两组间各变量的标准差异与 p 值。标准差异的计算公式为：

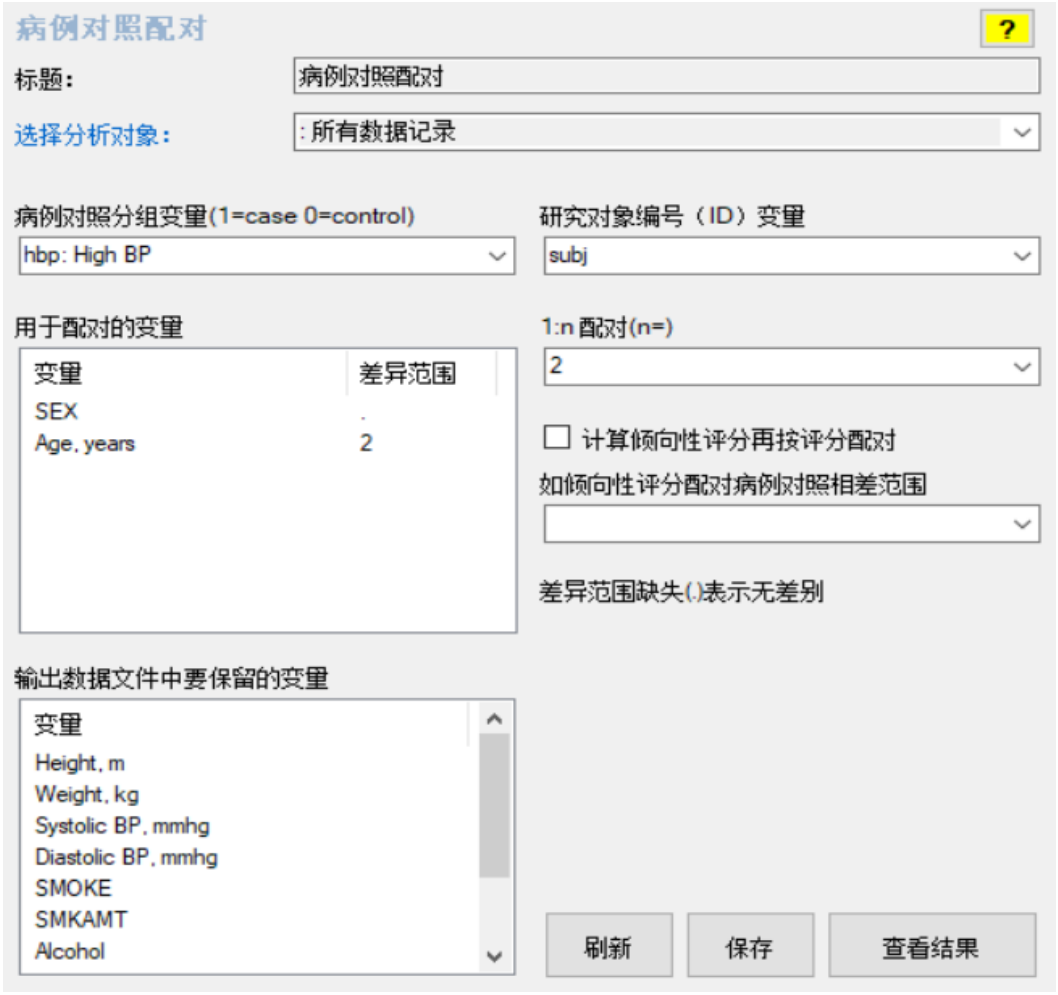
- 连续性变量： abs(Mean1-Mean0)/sqrt((S1+S0)/2)，Mean1 与 Mean0 分别为两组的均数， S1 与 S0 为两组方差。
- 分类型变量： abs(P1-P0)/sqrt((P1*(1-P1)+P0*(1-P0))/2)，P1 与 P0 分别为两组的率。如果变量是 3 分组或以上时，对每个亚组的哑变量分别计算标准差异。如果是两分组变量，两个哑变量的标准差异则完全相同。

两组间各变量的比较的 p 值是用 t 检验与卡方检验得出。

例：选择 demo 的练习项目，DEMO 数据中 HBP 是分组变量，0 为对照组，1 为病例组，NA 代表缺失。每个人有一个唯一编码 SUBJ，根据性别和年龄，对 HBP 变量做 1:2 配对。

调用“数据操作”-》“病例对照配对”模块，选择分组变量、配对变量、配对条件与匹配数、研究对象编号（ID）变量。

右击用于配对的变量修改差异范围，AGE 差异范围为 2，SEX 差异范围缺失（或取 0）表示性别相同，输出文件中包含 HEIGHT,WEIGHT 等变量。



2.运行结果与解释如下：

Found duplicated SUBJ, 2 duplicates were removed before matching

Match (HBP=1) with (HBP=0), 1:2

Using: SEX AGE ,within range: 0 2

发现两个重复的研究对象编号 SUBJ，在匹配之前删除了这两个 SUBJ。

根据 SEX AGE 匹配，差异范围 0 2

Total # of HBP = 1 : 195

Total # of HBP = 0 : 596

100 (HBP = 1) could not found 2 (HBP = 0)

100 个病例没有找到 2 个对照，下面列出了这 100 个 SUBJ

matchset.id	matchset.n	SUBJ.1	SUBJ.0.1	SUBJ.0.2
1	1	12	NA	NA

2	1	32	NA	NA
3	1	82	NA	NA
……				
38	1	807	NA	NA
39	1	813	NA	NA
40	2	17	53	NA
41	2	42	729	NA
……				
98	2	738	708	NA
99	2	791	557	NA
100	2	819	215	NA

上表中 matchset.id 是每个匹配组的编号，matchset.n 是每个匹配组的样本数，如果 matchset.n=1 表示没有匹配到对照，=2 表示匹配到 1 个对照，=3 表示匹配到 2 个对照。从上面列表可以看出 39 个病例没有找到对照，SUBJ 为：12,32,82,…,807,813。

另外 61 个病例分别找到 1 个对照，SUBJ 为：17，42，…,738,791,819。

下面列出这 100 个没有匹配到 2 个对照的匹配组数据：

matchset.id	matchset.n	SUBJ	HBP	SEX	AGE
1	1	12	1	1	68.7
2	1	32	1	2	67.7
…					
39	1	813	1	2	62.2
40	2	17	1	1	60.9
40	2	53	0	1	60.9
…					
39	1	813	1	2	62.2
41	2	42	1	1	60.7
41	2	729	0	1	60.3
…					
100	2	819	1	1	56.2
100	2	215	0	1	54.4

95 (HBP = 1) found 2 (HBP = 0), data saved to .XLS files

Total 285 observations in demo_97_tbl.xls

输出的 xls 文件如上面的输出表，有两种格式。

- 格式一（同表 1 格式）文件名为 XXX_cc.xls

- 格式二（同表 2 格式）文件名为 XXX_dd.xls

Matched data output to: demo_97_tbl.xls

另外输出的完整匹配组数据保存在 demo_97_tbl.xls 中

下面是匹配后两组各变量的标准差异与两组比较的 p 值：

Comparison between matched group

Variables	High BP: no	High BP: yes	Standardized diff.	P value
Age, years	(190) 35.48 ± 9.81	(95) 35.67 ± 9.93	0.0197	0.8751
Height, m	(190) 1.60 ± 0.08	(95) 1.60 ± 0.08	0.0270	0.8296
Weight, kg	(190) 54.78 ± 6.74	(95) 55.47 ± 7.02	0.1006	0.4209
Systolic BP, mmhg	(190) 120.63 ± 10.28	(95) 152.75 ± 14.49	2.5571	<0.0001
Diastolic BP, mmhg	(190) 66.38 ± 7.89	(95) 76.73 ± 8.47	1.2644	<0.0001
SMKAMT	(88) 6.25 ± 3.51	(48) 6.53 ± 4.25	0.0733	0.6746
Body mass index, kg/m2	(190) 21.44 ± 1.96	(95) 21.79 ± 2.33	0.1632	0.1822
SEX			0.0000	0.9999
Male	112 (58.9)	56 (58.9)		
Female	78 (41.1)	39 (41.1)		
SMOKE			0.0951	0.5303
no	102 (53.7)	46 (48.9)		
yes	88 (46.3)	48 (51.1)		
Alcohol			0.0729	0.6743
no	157 (82.6)	75 (79.8)		
yes	33 (17.4)	19 (20.2)		
Occupation			0.1798	0.1931
farmer	83 (43.7)	50 (52.6)		
others	107 (56.3)	45 (47.4)		
Education				0.1232
elementary of lower	75 (39.5)	34 (36.2)	0.0682	
middle	53 (27.9)	37 (39.4)	0.2445	
high or above	62 (32.6)	23 (24.5)	0.1815	

For continuous variables: (N) Mean ± SD, Standardized difference = abs(Mean1-Mean0)/sqrt((S1+S0)/2)
For categorical variables: N (%), Standardized difference = abs(P1-P0)/sqrt((P1*(1-P1)+P0*(1-P0))/2)

计算倾向性评分再配对

选择分组变量、需要配对的变量、研究对象编号（ID）变量。如果匹配条件中用于计算倾向性评分的变量里有连续性变量，可以右击该变量，选择曲线拟合。软件会根据评分进行配对。这里选择匹配条件为分数差值（默认）为 0.05 之内，用户可以手动修改该值。匹配数 1：2。

结果输出如下：

```
Found duplicated SUBJ , 2 duplicates were removed before matching
Calculate propensity score (pp.score) based on formula: HBP ~
SEX+s(AGE)+s(BMI)+OCCU.NEW+EDU.NEW+SMOKE+ALH
Match (HBP=1) with (HBP=0), 1:2
Using: pp.score ,within range: 0.05
Total # of HBP = 1 : 190
Total # of HBP = 0 : 592
96 ( HBP = 1 ) could not found 2 ( HBP = 0 )
```

输出结果表 1 格式同前例，表 2 中没有列出各匹配变量的值，但列出倾向性评分值（pp.score）

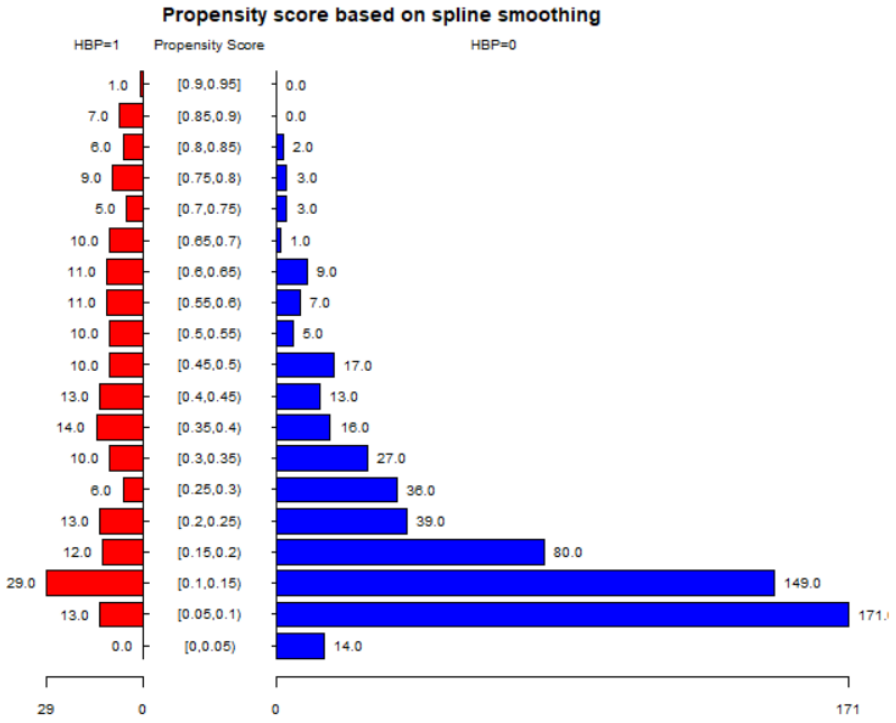
matchset.id	matchset.n	SUBJ	HBP	pp.score
1	1	12	1	0.784699588463251
2	1	32	1	0.794268024664835
……				

94 (HBP = 1) found 2 (HBP = 0), data saved to .XLS files

Matched data output to: demo_98_tbl.xls

Total 282 observations in demo_98_tbl.xls

输出文件格式同前。另外输出倾向性评分分布图如下：



上图标明病例与对照的倾向性评分分布，以判断哪些区间可用于匹配的资源有限。