临床预测模型构建&机器学习(R语言进阶)

第4章 倾向性匹配得分PSM分析与R语言实现

周支瑞

PSM定义

- 通过一定的统计学方法对试验组与对照组进行筛选，使筛选出来的研究对象在某些重要临床特征（潜在的混杂因素）上具有可比性。
- 一般是通过某种统计学模型求得每个观测的多个协变量的综合**倾向性得分**，再按照倾向性得分是否接近进行匹配。
- 最常用的统计模型一般是以分组变量为因变量，其他可能影响结果的混杂因素为协变量构建Logistic回归模型。
- 计算每个观测的倾向得分，按照得分大小进行匹配。

PSM概念解释

- **举个例子：**
  - 比如替某大龄女孩找对象，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？选择多个还是选择一个？这是一个问题！

PSM概念解释

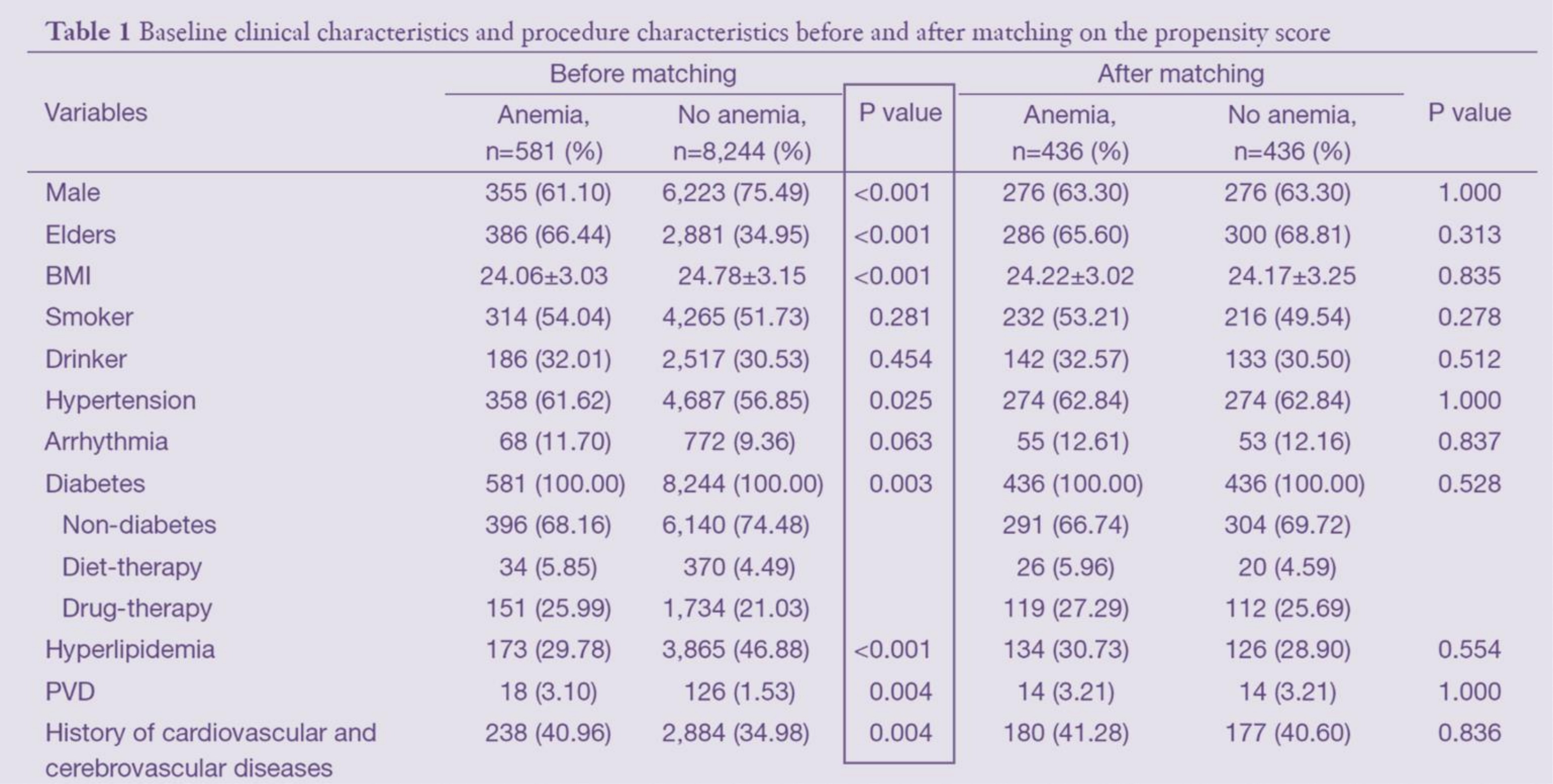


Table 1 Baseline clinical characteristics and procedure characteristics before and after matching on the propensity score

Variables	Before matching		P value	After matching		P value
	Anemia, n=581 (%)	No anemia, n=8,244 (%)		Anemia, n=436 (%)	No anemia, n=436 (%)	
Male	355 (61.10)	6,223 (75.49)	<0.001	276 (63.30)	276 (63.30)	1.000
Elders	386 (66.44)	2,881 (34.95)	<0.001	286 (65.60)	300 (68.81)	0.313
BMI	24.06±3.03	24.78±3.15	<0.001	24.22±3.02	24.17±3.25	0.835
Smoker	314 (54.04)	4,265 (51.73)	0.281	232 (53.21)	216 (49.54)	0.278
Drinker	186 (32.01)	2,517 (30.53)	0.454	142 (32.57)	133 (30.50)	0.512
Hypertension	358 (61.62)	4,687 (56.85)	0.025	274 (62.84)	274 (62.84)	1.000
Arrhythmia	68 (11.70)	772 (9.36)	0.063	55 (12.61)	53 (12.16)	0.837
Diabetes	581 (100.00)	8,244 (100.00)	0.003	436 (100.00)	436 (100.00)	0.528
Non-diabetes	396 (68.16)	6,140 (74.48)		291 (66.74)	304 (69.72)	
Diet-therapy	34 (5.85)	370 (4.49)		26 (5.96)	20 (4.59)	
Drug-therapy	151 (25.99)	1,734 (21.03)		119 (27.29)	112 (25.69)	
Hyperlipidemia	173 (29.78)	3,865 (46.88)	<0.001	134 (30.73)	126 (28.90)	0.554
PVD	18 (3.10)	126 (1.53)	0.004	14 (3.21)	14 (3.21)	1.000
History of cardiovascular and cerebrovascular diseases	238 (40.96)	2,884 (34.98)	0.004	180 (41.28)	177 (40.60)	0.836

PSM实战【案例 1】

- 下图所示数据中共有10个变量，614个观测，试验组185例，对照组429例。treat变量即为分组变量，“1”=试验组(接受职业培训)，“0”=对照组(未接受职业培训)。age, educ, black, hispan, married, nodegree, re74, re75为协变量, re78为结局变量（年总收入）。事实上，倾向性匹配得分分析是要建立一个以分组变量（treat）为因变量，各个协变量（age, educ, black, hispan, married, nodegree, re74, re75）为自变量的回归方程。而结局变量（re78）在PSM过程中几乎不参与建模。

SPSS ? Stata ? R ?

软件名称	优点	不足	推荐级别
SPSS	1. 菜单操作；2. 简便医学；3. 可直接导出匹配好的数据集	1. 仅能实现1:1匹配；2. 导出数据集后需要手动进行均衡性检验；3. 不能导出匹配结果的直观图形	*****
Stata	1. 操作灵活；2. 可实现1:2及以上比例的匹配；3. 自动均衡性检验；4. 导出匹配结果的直观图形	1. 命令行操作，需要一定的Stata基础；2. 匹配完成的数据集导出较麻烦	****
备注：R语言与SAS均可实现PSM, 请根据实际情况选择一种软件即可			

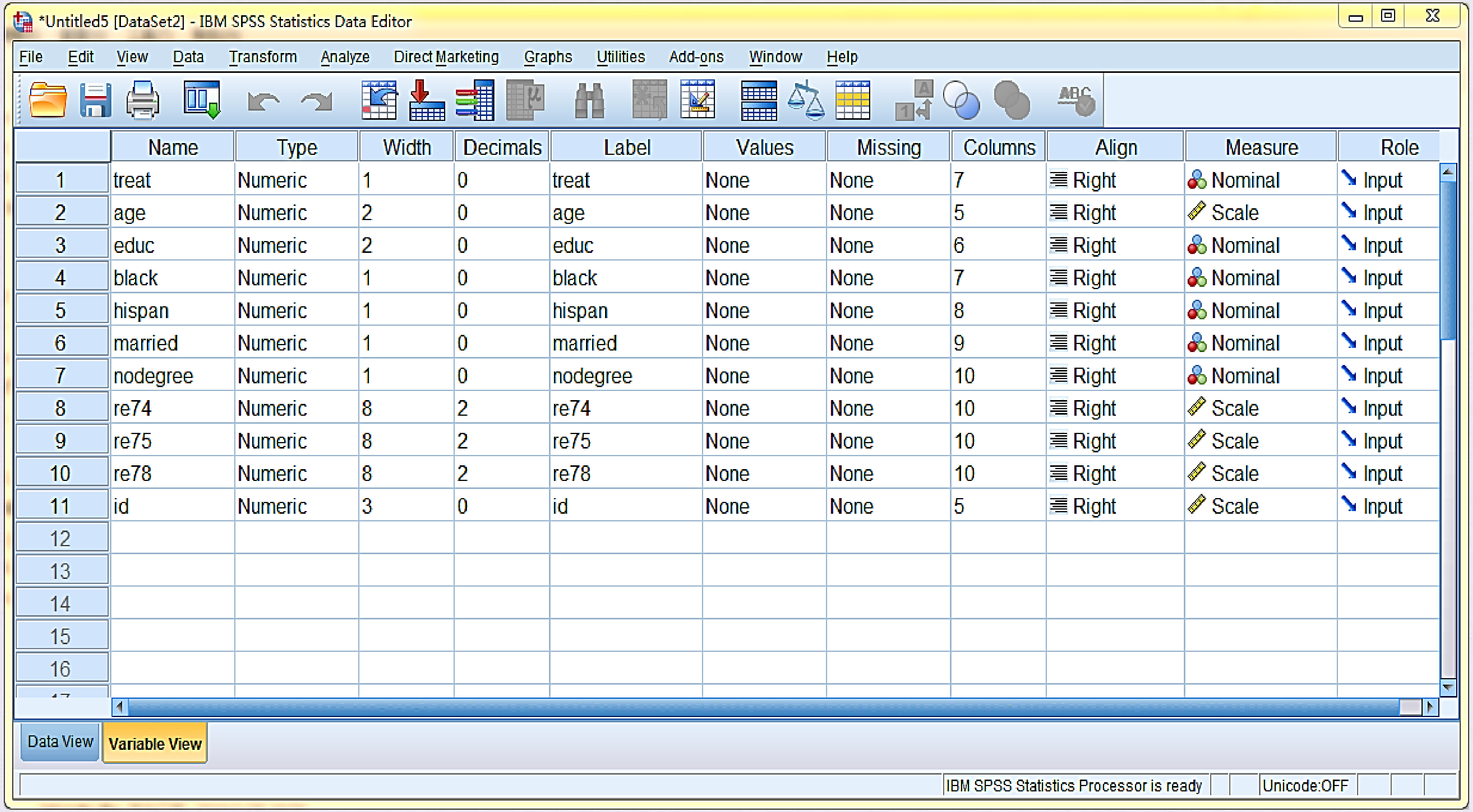
R语言实现PSM代码{MatchIt}

- library(MatchIt)
- data(lalonde)
- head(lalonde)
- f=matchit(treat~re74+re75+educ+black+hispan+age+married+nodegree,data=lalonde,method="nearest")
- #f=matchit(treat~re74+re75+educ+black+hispan+age+married+nodegree,data=lalonde,method="nearest",caliper=0.05)
- summary(f)
- matchdata=match.data(f)
- matchdata
- library(foreign)
- matchdata$id<-1:nrow(matchdata)
- write.dta(matchdata,"d:/matchdata.dta")

R语言实现PSM代码{nonrandom}

```
## plot.pscore {nonrandom}
library(nonrandom)
## STU1
data(stu1)
stu1.ps <- pscore(data = stu1,
               formula = therapie~tgr+age)
plot.pscore(x = stu1.ps,
          main = "PS distribution",
          xlab = "",
          par.1=list(col="red"),
          par.0=list(lwd=2),
          par.dens=list(kernel="gaussian"))
## STU1
data(stu1)
stu1.ps <- pscore(data = stu1,
               formula = therapie~tgr+age)
stu1.match <- ps.match(object = stu1.ps,
                   ratio  = 2,
                   caliper = 0.05,
                   givenTmatchingC = FALSE,
                   matched.by = "pscore",
                   setseed = 38902)
```

PSM实战 -- SPSS操作

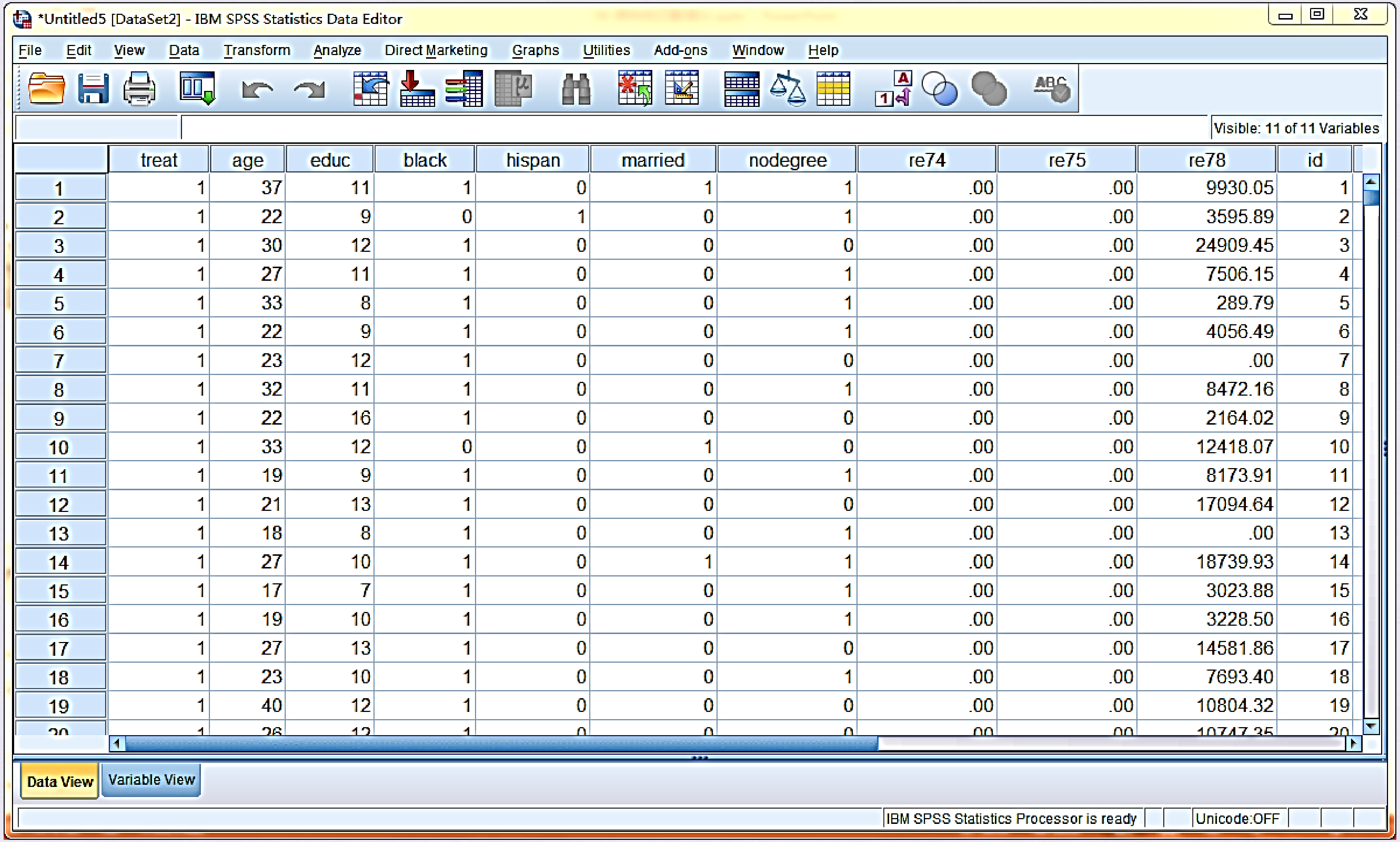


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	treat	Numeric	1	0	treat	None	None	7	Right	Nominal	Input
2	age	Numeric	2	0	age	None	None	5	Right	Scale	Input
3	educ	Numeric	2	0	educ	None	None	6	Right	Nominal	Input
4	black	Numeric	1	0	black	None	None	7	Right	Nominal	Input
5	hispan	Numeric	1	0	hispan	None	None	8	Right	Nominal	Input
6	married	Numeric	1	0	married	None	None	9	Right	Nominal	Input
7	nodegree	Numeric	1	0	nodegree	None	None	10	Right	Nominal	Input
8	re74	Numeric	8	2	re74	None	None	10	Right	Scale	Input
9	re75	Numeric	8	2	re75	None	None	10	Right	Scale	Input
10	re78	Numeric	8	2	re78	None	None	10	Right	Scale	Input
11	id	Numeric	3	0	id	None	None	5	Right	Scale	Input
12											
13											
14											
15											
16											

IBM SPSS Statistics Processor is ready

Unicode:OFF

PSM实战 -- SPSS操作



*Untitled5 [DataSet2] - IBM SPSS Statistics Data Editor

Visible: 11 of 11 Variables

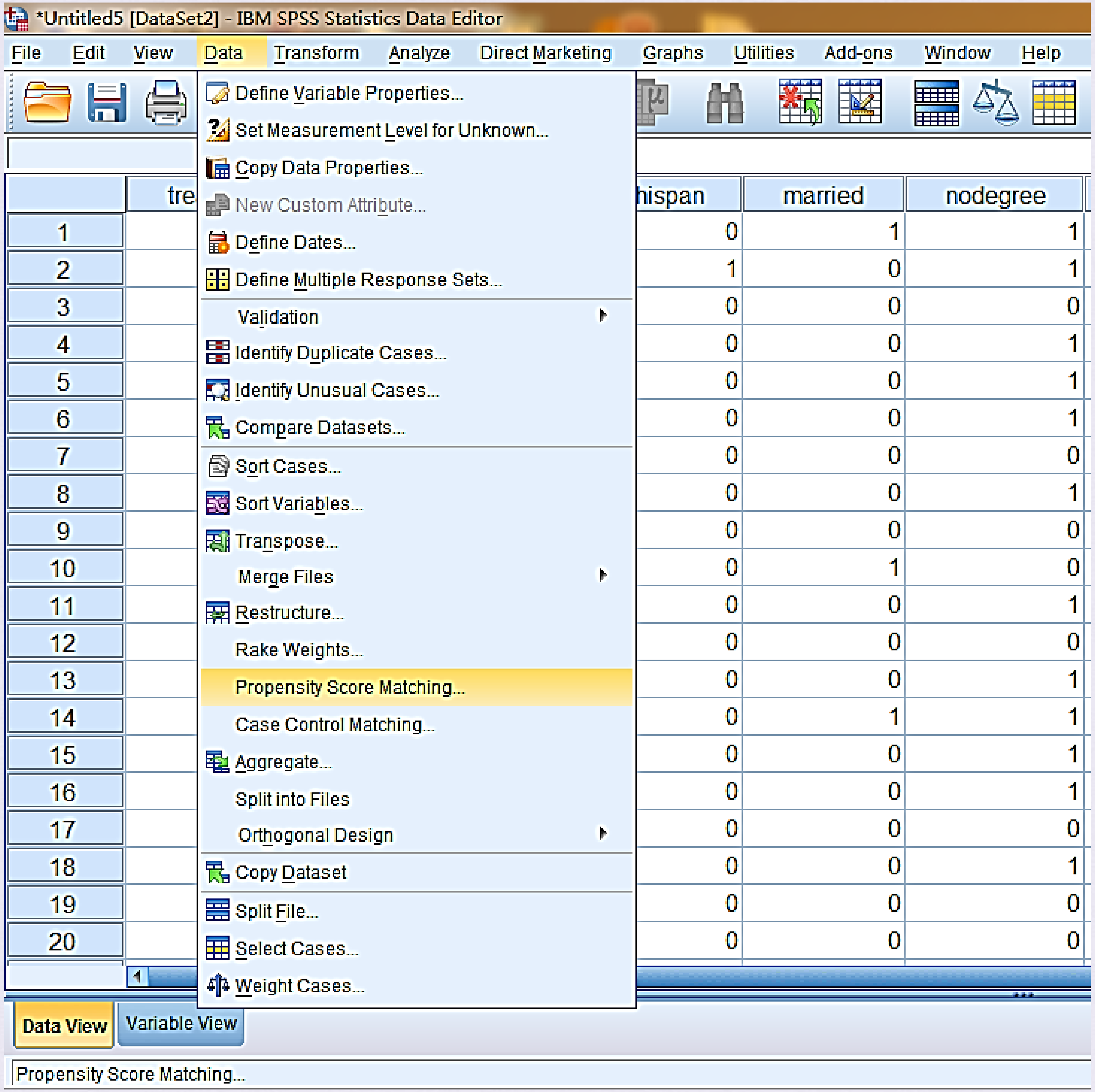
	treat	age	educ	black	hispan	married	nodegree	re74	re75	re78	id
1	1	37	11	1	0	1	1	.00	.00	9930.05	1
2	1	22	9	0	1	0	1	.00	.00	3595.89	2
3	1	30	12	1	0	0	0	.00	.00	24909.45	3
4	1	27	11	1	0	0	1	.00	.00	7506.15	4
5	1	33	8	1	0	0	1	.00	.00	289.79	5
6	1	22	9	1	0	0	1	.00	.00	4056.49	6
7	1	23	12	1	0	0	0	.00	.00	.00	7
8	1	32	11	1	0	0	1	.00	.00	8472.16	8
9	1	22	16	1	0	0	0	.00	.00	2164.02	9
10	1	33	12	0	0	1	0	.00	.00	12418.07	10
11	1	19	9	1	0	0	1	.00	.00	8173.91	11
12	1	21	13	1	0	0	0	.00	.00	17094.64	12
13	1	18	8	1	0	0	1	.00	.00	.00	13
14	1	27	10	1	0	1	1	.00	.00	18739.93	14
15	1	17	7	1	0	0	1	.00	.00	3023.88	15
16	1	19	10	1	0	0	1	.00	.00	3228.50	16
17	1	27	13	1	0	0	0	.00	.00	14581.86	17
18	1	23	10	1	0	0	1	.00	.00	7693.40	18
19	1	40	12	1	0	0	0	.00	.00	10804.32	19
20	1	26	12	1	0	0	0	.00	.00	10747.35	20

IBM SPSS Statistics Processor is ready

Unicode:OFF

PSM实战 -- SPSS操作

1.点击“数据” - “倾向得分匹配”，如下图：

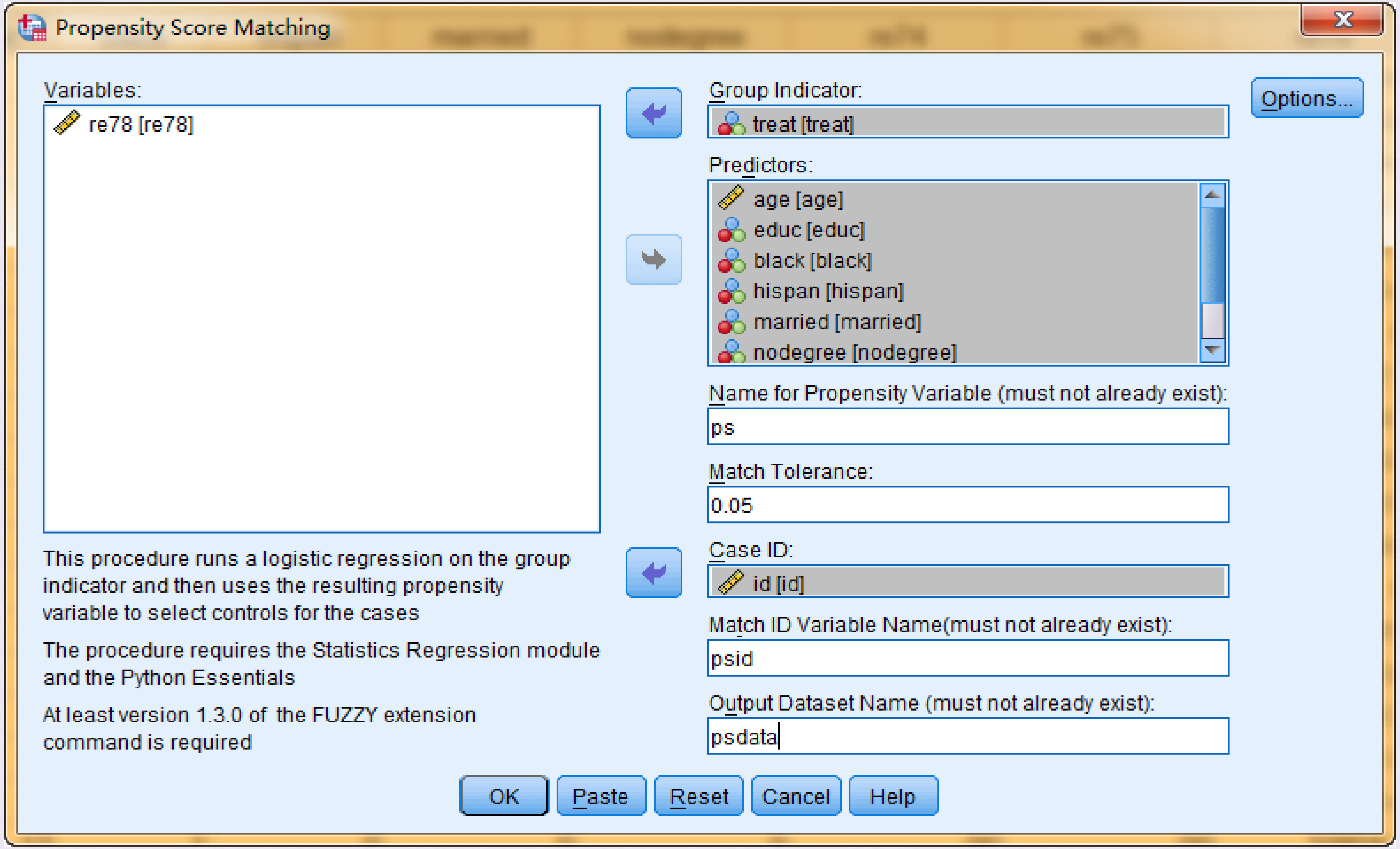


	tre	hispan	married	nodegree
1		0	1	1
2		1	0	1
3		0	0	0
4		0	0	1
5		0	0	1
6		0	0	1
7		0	0	0
8		0	0	1
9		0	0	0
10		0	1	0
11		0	0	1
12		0	0	0
13		0	0	1
14		0	1	1
15		0	0	1
16		0	0	1
17		0	0	0
18		0	0	1
19		0	0	0
20		0	0	0

Propensity Score Matching...

SPSS操作

2. 弹出下图对话框，组指示符选择“treat”，即干预因素，须为二分类变量；预测变量框里选入所有混杂因素，倾向变量名即每个个体的倾向评分得分变量名，可随意填写（字母或字母加数字）、匹配容差可从较小的数值填写，根据情况填写；匹配id变量名，即可输出一个变量，告诉我们每个case的匹配对象的id；数据集名称可自行填写。点击确定可得到匹配结果。



Variables:
re78 [re78]

Group Indicator:
treat [treat]

Predictors:
age [age]
educ [educ]
black [black]
hispan [hispan]
married [married]
nodegree [nodegree]

Name for Propensity Variable (must not already exist):
ps

Match Tolerance:
0.05

Case ID:
id [id]

Match ID Variable Name(must not already exist):
psid

Output Dataset Name (must not already exist):
psdata

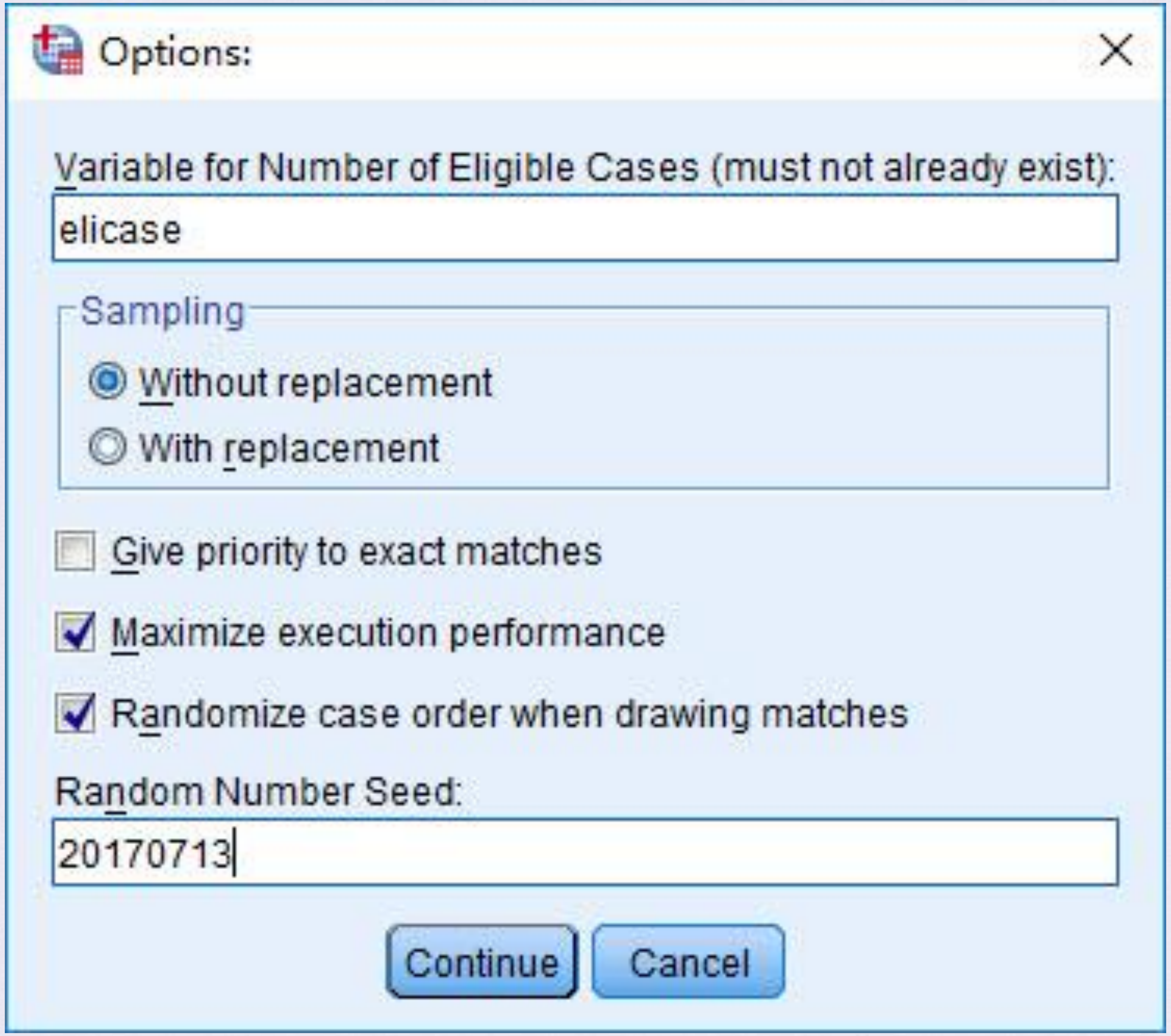
This procedure runs a logistic regression on the group indicator and then uses the resulting propensity variable to select controls for the cases

The procedure requires the Statistics Regression module and the Python Essentials

At least version 1.3.0 of the FUZZY extension command is required

SPSS操作

3.如下图进行“选项设置”。Variable for Number of Eligible Case：输出新的变量（变量名自定义）显示实验组匹配几个满足条件的对照；Sampling: 抽样；Without replacement: 不放回抽; With replacement: 放回抽样; Give priority…: 优先精确匹配; Maxmize execution performance: 最大匹配（系统默认）；Randomize case order when…: 当多个对照符合时随机选择一个对象匹配；Random Number Seed: 设置随机种子数。



Options:

Variable for Number of Eligible Cases (must not already exist): elicase

Sampling

● Without replacement

○ With replacement

□ Give priority to exact matches

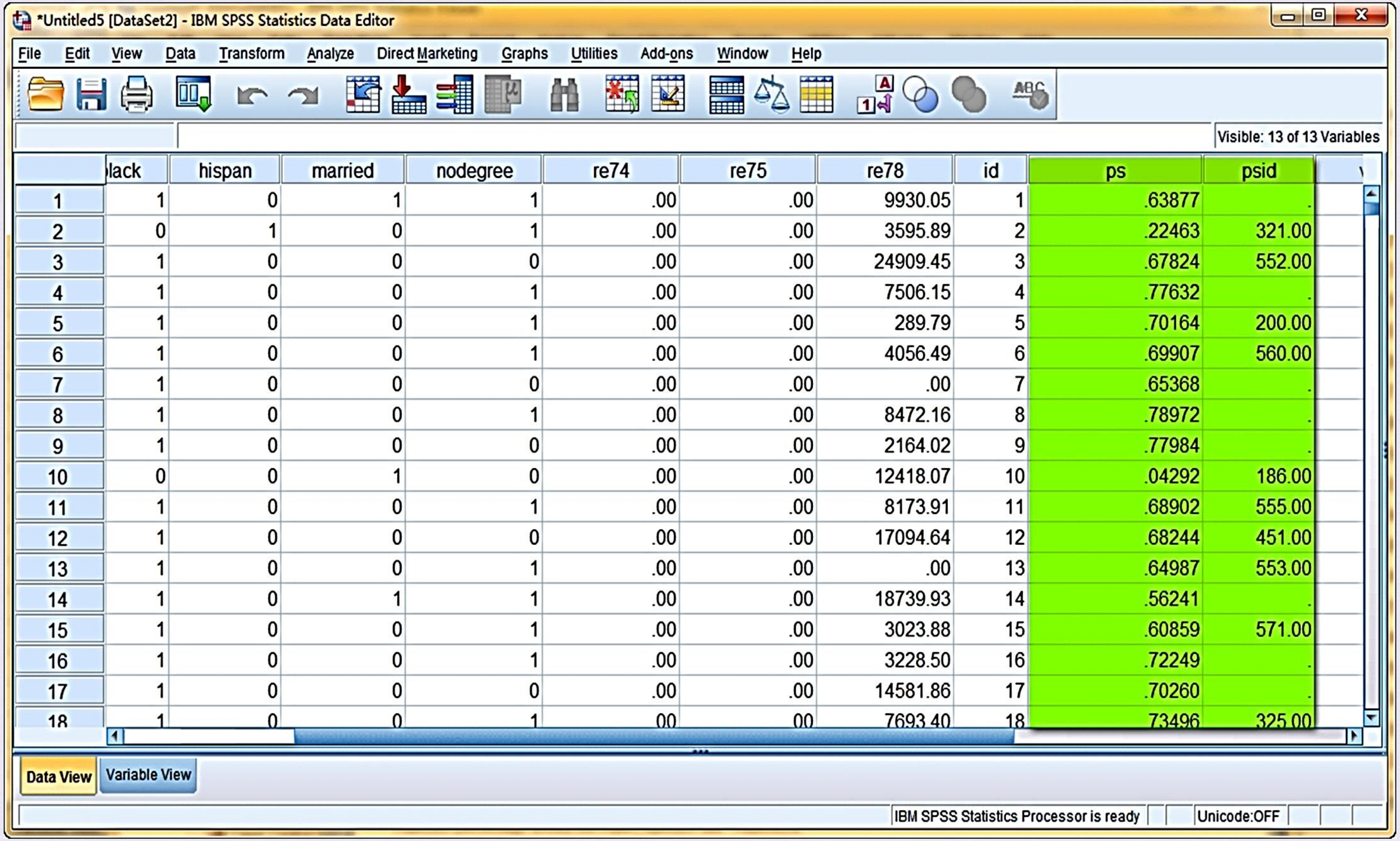
☑ Maximize execution performance

☑ Randomize case order when drawing matches

Random Number Seed: 20170713

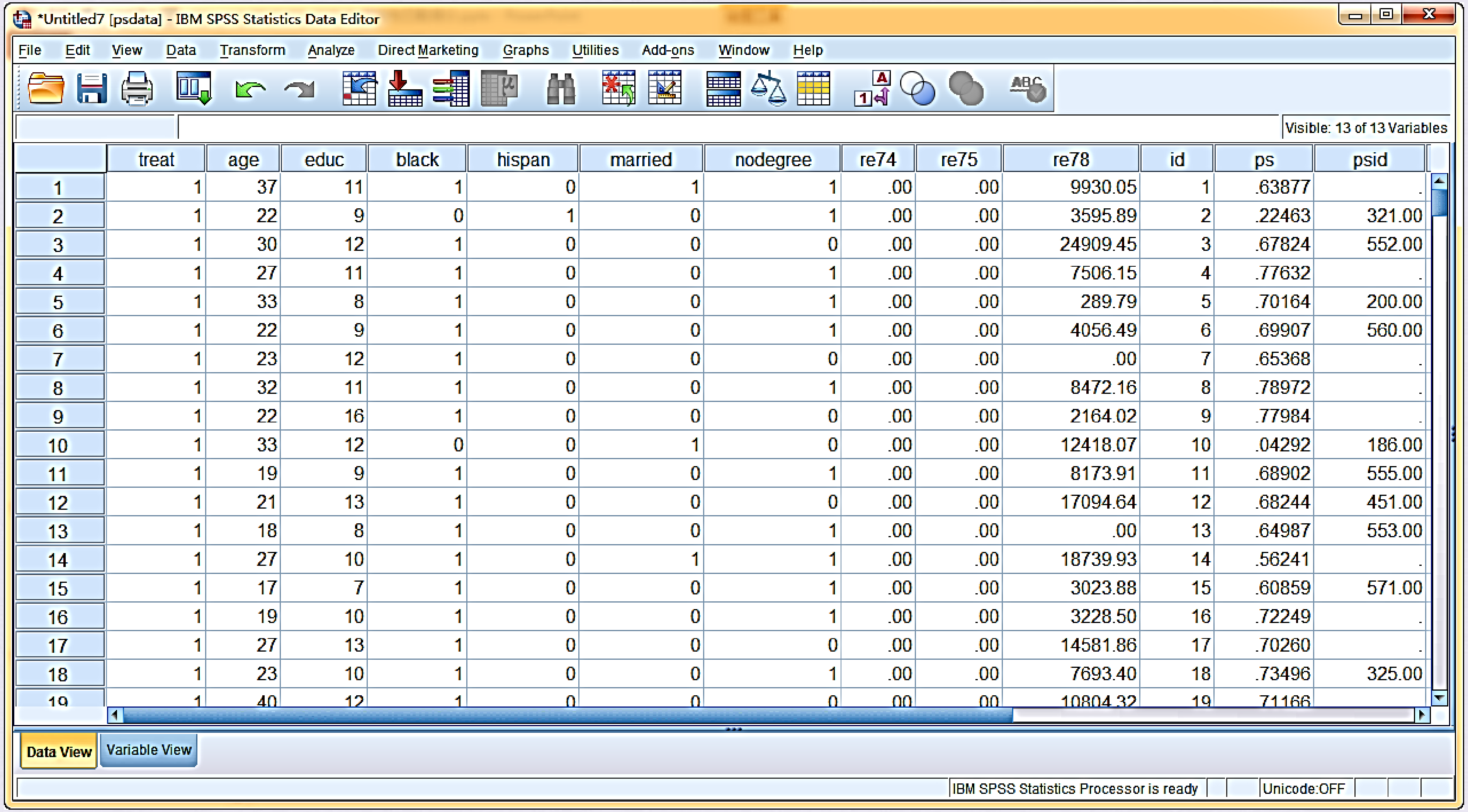
SPSS操作

4. 我们的变量后面多了两个新变量，ps即每个case的倾向得分，psid即匹配对象，第1个未能匹配，第2个匹配对象的id是321，依次类推。

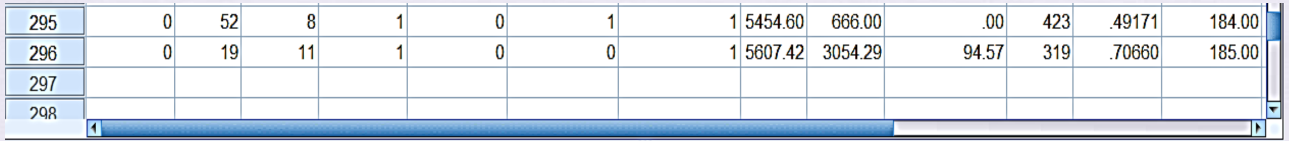

*Untitled5 [DataSet2] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
Visible: 13 of 13 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:OFF

	lack	hispan	married	nodegree	re74	re75	re78	id	ps	psid
1	1	0	1	1	.00	.00	9930.05	1	.63877	.
2	0	1	0	1	.00	.00	3595.89	2	.22463	321.00
3	1	0	0	0	.00	.00	24909.45	3	.67824	552.00
4	1	0	0	1	.00	.00	7506.15	4	.77632	.
5	1	0	0	1	.00	.00	289.79	5	.70164	200.00
6	1	0	0	1	.00	.00	4056.49	6	.69907	560.00
7	1	0	0	0	.00	.00	.00	7	.65368	.
8	1	0	0	1	.00	.00	8472.16	8	.78972	.
9	1	0	0	0	.00	.00	2164.02	9	.77984	.
10	0	0	1	0	.00	.00	12418.07	10	.04292	186.00
11	1	0	0	1	.00	.00	8173.91	11	.68902	555.00
12	1	0	0	0	.00	.00	17094.64	12	.68244	451.00
13	1	0	0	1	.00	.00	.00	13	.64987	553.00
14	1	0	1	1	.00	.00	18739.93	14	.56241	.
15	1	0	0	1	.00	.00	3023.88	15	.60859	571.00
16	1	0	0	1	.00	.00	3228.50	16	.72249	.
17	1	0	0	0	.00	.00	14581.86	17	.70260	.
18	1	0	0	1	.00	.00	7693.40	18	.73496	325.00

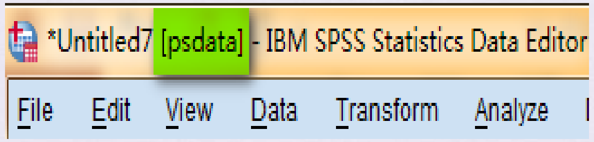
SPSS操作 5. SPSS会将所有匹配的对象重新筛选出来生成一个新的数据，即匹配后的数据。


*Untitled7 [psdata] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
Visible: 13 of 13 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:OFF

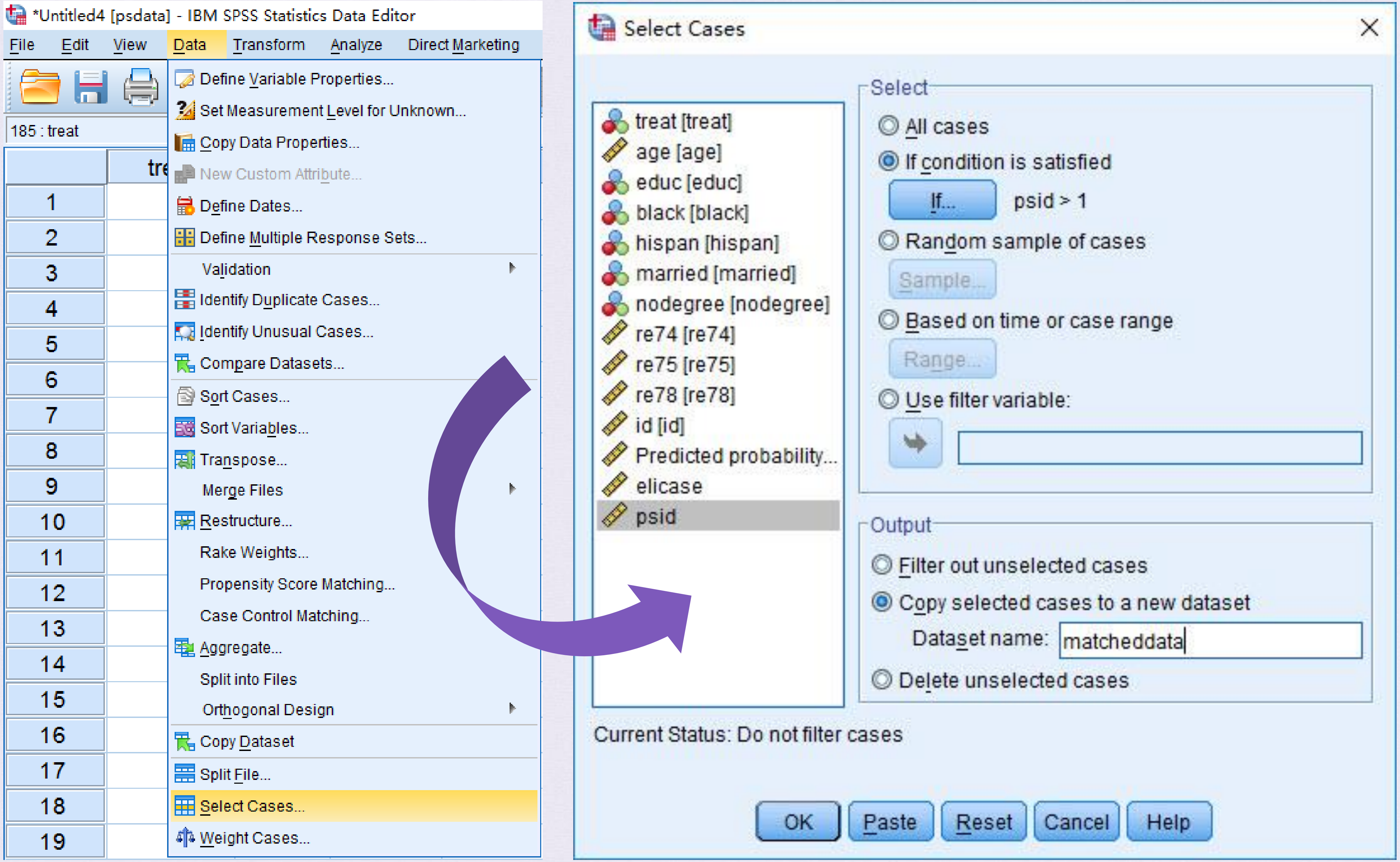
	treat	age	educ	black	hispan	married	nodegree	re74	re75	re78	id	ps	psid
1	1	37	11	1	0	1	1	.00	.00	9930.05	1	.63877	.
2	1	22	9	0	1	0	1	.00	.00	3595.89	2	.22463	321.00
3	1	30	12	1	0	0	0	.00	.00	24909.45	3	.67824	552.00
4	1	27	11	1	0	0	1	.00	.00	7506.15	4	.77632	.
5	1	33	8	1	0	0	1	.00	.00	289.79	5	.70164	200.00
6	1	22	9	1	0	0	1	.00	.00	4056.49	6	.69907	560.00
7	1	23	12	1	0	0	0	.00	.00	.00	7	.65368	.
8	1	32	11	1	0	0	1	.00	.00	8472.16	8	.78972	.
9	1	22	16	1	0	0	0	.00	.00	2164.02	9	.77984	.
10	1	33	12	0	0	1	0	.00	.00	12418.07	10	.04292	186.00
11	1	19	9	1	0	0	1	.00	.00	8173.91	11	.68902	555.00
12	1	21	13	1	0	0	0	.00	.00	17094.64	12	.68244	451.00
13	1	18	8	1	0	0	1	.00	.00	.00	13	.64987	553.00
14	1	27	10	1	0	1	1	.00	.00	18739.93	14	.56241	.
15	1	17	7	1	0	0	1	.00	.00	3023.88	15	.60859	571.00
16	1	19	10	1	0	0	1	.00	.00	3228.50	16	.72249	.
17	1	27	13	1	0	0	0	.00	.00	14581.86	17	.70260	.
18	1	23	10	1	0	0	1	.00	.00	7693.40	18	.73496	325.00
19	1	40	12	1	0	0	0	.00	.00	10804.32	19	.71166	



295	0	52	8	1	0	1	1	5454.60	666.00	.00	423	.49171	184.00
296	0	19	11	1	0	0	1	5607.42	3054.29	94.57	319	.70660	185.00
297													
298													



SPSS操作 6. 如下图所示，筛选最终的成功匹配的数据集



Select

○ All cases

● If condition is satisfied

If... psid > 1

○ Random sample of cases

○ Based on time or case range

○ Use filter variable:

Output

○ Filter out unselected cases

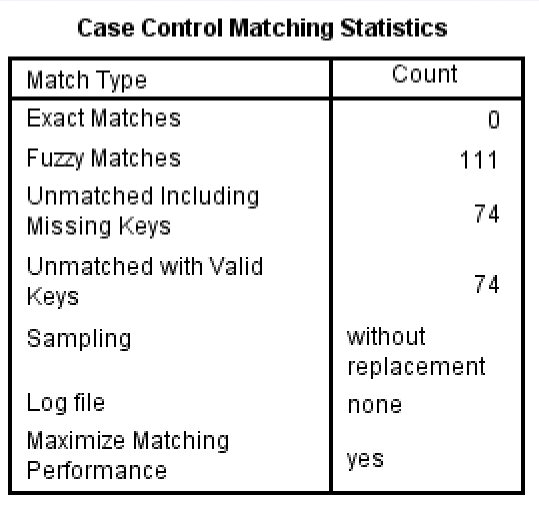
● Copy selected cases to a new dataset

Dataset name: matcheddata

○ Delete unselected cases

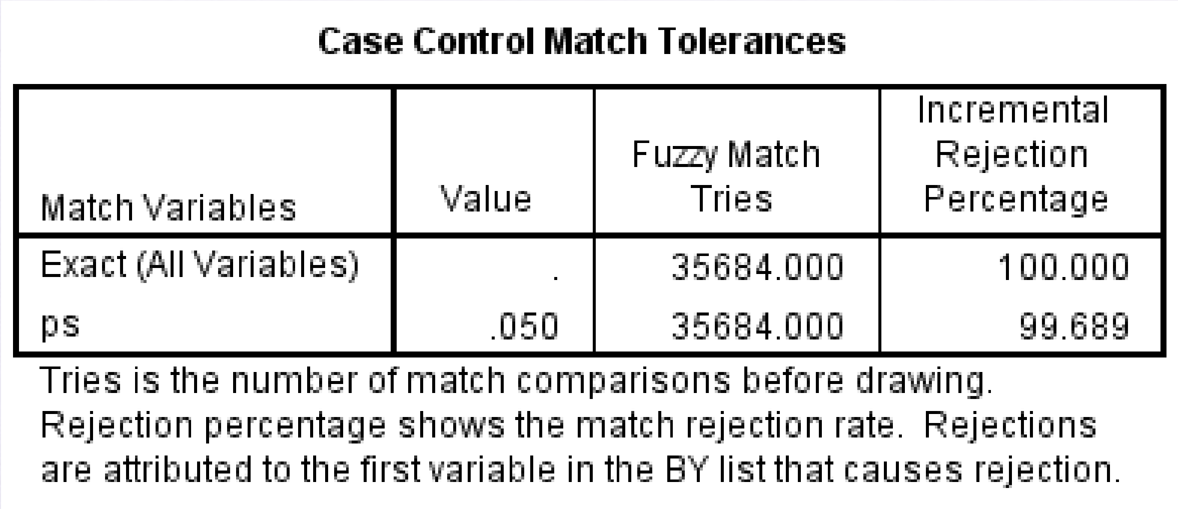
Current Status: Do not filter cases

匹配后的统计结果概览



Case Control Matching Statistics

Match Type	Count
Exact Matches	0
Fuzzy Matches	111
Unmatched Including Missing Keys	74
Unmatched with Valid Keys	74
Sampling	without replacement
Log file	none
Maximize Matching Performance	yes



Case Control Match Tolerances

Match Variables	Value	Fuzzy Match Tries	Incremental Rejection Percentage
Exact (All Variables)	.	35684.000	100.000
ps	.050	35684.000	99.689

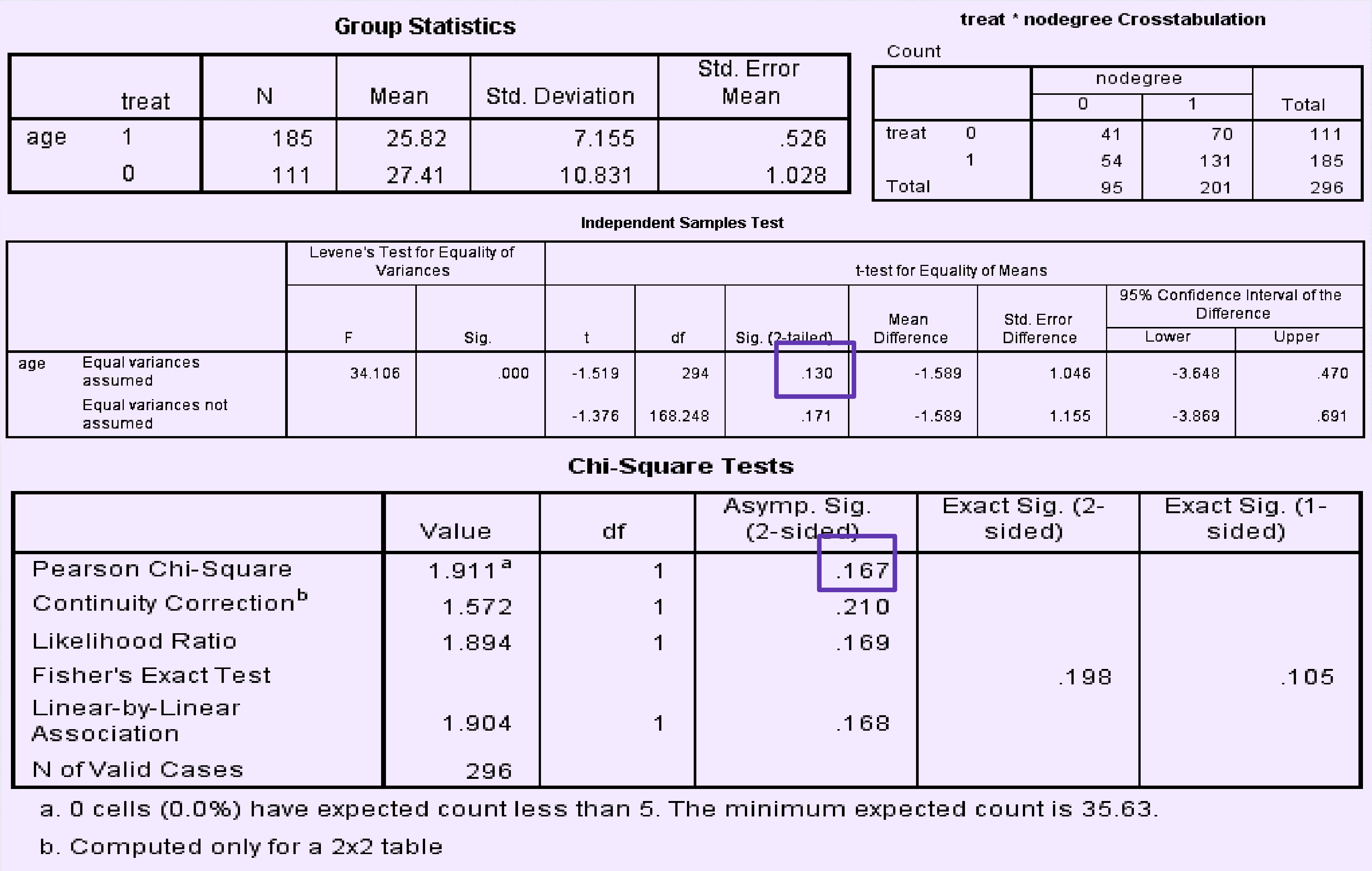
Tries is the number of match comparisons before drawing. Rejection percentage shows the match rejection rate. Rejections are attributed to the first variable in the BY list that causes rejection.

SPSS操作注意事项

在用SPSS做倾向评分匹配时应注意：

- 1. SPSS安装需要同意安装Python Essentials插件，否则无法使用；
- 2. 所有用于分析的变量名和界面填写的变量名必须是英文或英文加数字，不能是中文；
- 3. 匹配容差需要根据实际情况确定，如两组样本量差异较大（两组差10倍以上），可以用较小的容差，如0.001，如较小容差不能匹配，再将容差调大后重试，如0.01或0.05；
- 4. 完成匹配后应对两组进行均衡性检验，成组t检验或者普通卡方检验；
- 5. 如果一个观测有多个匹配对象，程序会从中随机选择，因此每次运行可能得到的匹配结果不同，如果设定固定随机种子数，结果应该一致；
- 6. SPSS程序只能进行1：1匹配，其他比例可通过Stata实现。

完成匹配后均衡性检验(age nodegree)



Group Statistics

	treat	N	Mean	Std. Deviation	Std. Error Mean
age	1	185	25.82	7.155	.526
	0	111	27.41	10.831	1.028

treat * nodegree Crosstabulation

Count

		nodegree		Total
		0	1	
treat	0	41	70	111
	1	54	131	185
Total		95	201	296

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
age	Equal variances assumed	34.106	.000	-1.519	294	.130	-1.589	1.046	-3.648	.470
	Equal variances not assumed			-1.376	168.248	.171	-1.589	1.155	-3.869	.691

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.911[a]	1	.167		
Continuity Correction[b]	1.572	1	.210		
Likelihood Ratio	1.894	1	.169		
Fisher's Exact Test				.198	.105
Linear-by-Linear Association	1.904	1	.168		
N of Valid Cases	296				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 35.63.

b. Computed only for a 2x2 table

PSM实战 -- Stata操作 I

- 1. 安装psmatch2统计包，命令如下：

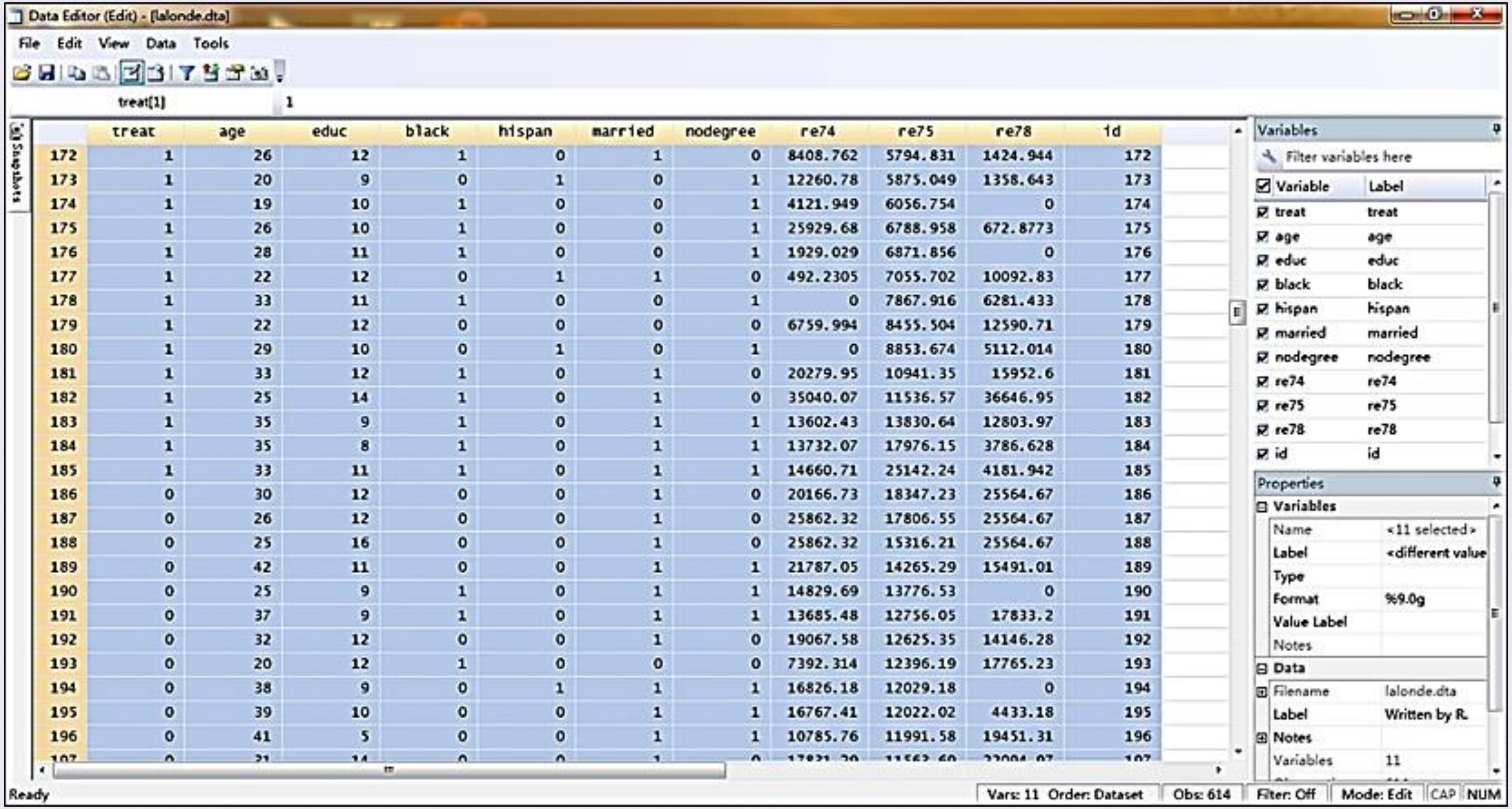
```
.ssc install psmatch2
#需要在联网状态下键入上述命令，然后软件自动搜索对应的程序包进行安装，成功安装后会有以下提示：
checking psmatch2 consistency and verifying not already installed...
installing into .\ado\plus\...
installation complete.（#出现此提示语句表示安装完成）
#为了验证是否成功安装以及查看psmatch2命令的帮助菜单，可在命令窗口键入
.help psmatch2
#如果能顺利弹出帮助文件，表示安装成功，可正常使用。
```

PSM实战 -- Stata操作 II

➢ 2. 数据准备

数据如下图所示，共有10个变量，614个观测，试验组185例，对照组429例。treat变量即为分组变量，“1”=试验组，“0”=对照组。age, educ, black, hispan, married, nodegree, re74, re75为协变量, re78为结局变量（总收入）。事实上，倾向性匹配得分分析是要建立一个以分组变量（treat）为因变量，各个协变量（age, educ, black, hispan, married, nodegree, re74, re75）为自变量的回归方程。而结局变量（re78）在PSM过程中几乎不参与建模。

Stata操作



treat[1] 1

	treat	age	educ	black	hispan	married	nodegree	re74	re75	re78	id
172	1	26	12	1	0	1	0	8408.762	5794.831	1424.944	172
173	1	20	9	0	1	0	1	12260.78	5875.049	1358.643	173
174	1	19	10	1	0	0	1	4121.949	6056.754	0	174
175	1	26	10	1	0	0	1	25929.68	6788.958	672.8773	175
176	1	28	11	1	0	0	1	1929.029	6871.856	0	176
177	1	22	12	0	1	1	0	492.2305	7055.702	10092.83	177
178	1	33	11	1	0	0	1	0	7867.916	6281.433	178
179	1	22	12	0	0	0	0	6759.994	8455.504	12590.71	179
180	1	29	10	0	1	0	1	0	8853.674	5112.014	180
181	1	33	12	1	0	1	0	20279.95	10941.35	15952.6	181
182	1	25	14	1	0	1	0	35040.07	11536.57	36646.95	182
183	1	35	9	1	0	1	1	13602.43	13830.64	12803.97	183
184	1	35	8	1	0	1	1	13732.07	17976.15	3786.628	184
185	1	33	11	1	0	1	1	14660.71	25142.24	4181.942	185
186	0	30	12	0	0	1	0	20166.73	18347.23	25564.67	186
187	0	26	12	0	0	1	0	25862.32	17806.55	25564.67	187
188	0	25	16	0	0	1	0	25862.32	15316.21	25564.67	188
189	0	42	11	0	0	1	1	21787.05	14265.29	15491.01	189
190	0	25	9	1	0	1	1	14829.69	13776.53	0	190
191	0	37	9	1	0	1	1	13685.48	12756.05	17833.2	191
192	0	32	12	0	0	1	0	19067.58	12625.35	14146.28	192
193	0	20	12	1	0	0	0	7392.314	12396.19	17765.23	193
194	0	38	9	0	1	1	1	16826.18	12029.18	0	194
195	0	39	10	0	0	1	1	16767.41	12022.02	4433.18	195
196	0	41	5	0	0	1	1	10785.76	11991.58	19451.31	196

Variables

Filter variables here

☑ Variable	Label
☑ treat	treat
☑ age	age
☑ educ	educ
☑ black	black
☑ hispan	hispan
☑ married	married
☑ nodegree	nodegree
☑ re74	re74
☑ re75	re75
☑ re78	re78
☑ id	id

Properties

⊟ Variables

Name	<11 selected>
Label	<different value
Type	
Format	%9.0g
Value Label	
Notes	

⊟ Data

⊞ Filename	lalonde.dta
Label	Written by R.
⊞ Notes	
Variables	11

Ready | Vars: 11 Order: Dataset | Obs: 614 | Filter: Off | Mode: Edit | CAP | NUM

PSM实战 -- Stata操作 III

- 3. 数据分析及命令解读，命令窗口键入如下命令：

```
.gen tmp = runiform()
.sort tmp
(# 以上两步对所有观测值进行随机排序)
.psmatch2 treat age educ black hispan married nodegree re74 re75, out(re78) logit
neighbor(1) common caliper(.05) ties
.pstest, both
.psgraph
```

PSM实战 -- Stata操作 III

- 3.1 命令解读, 以下是帮助菜单中psmatch2语法格式：

```
psmatch2 depvar [indepvars] [if exp] [in range] [, outcome(varlist) pscore(varname)
neighbor(integer) radius caliper(real) mahalanobis(varlist) ai(integer) population altvariance
kernel llr kerneltype(type) bwidth(real) spline nknots(integer) common trim(real)
noreplacement descending odds index logit ties quietly w(matrix) ate]
```

PSM实战 -- Stata操作 III

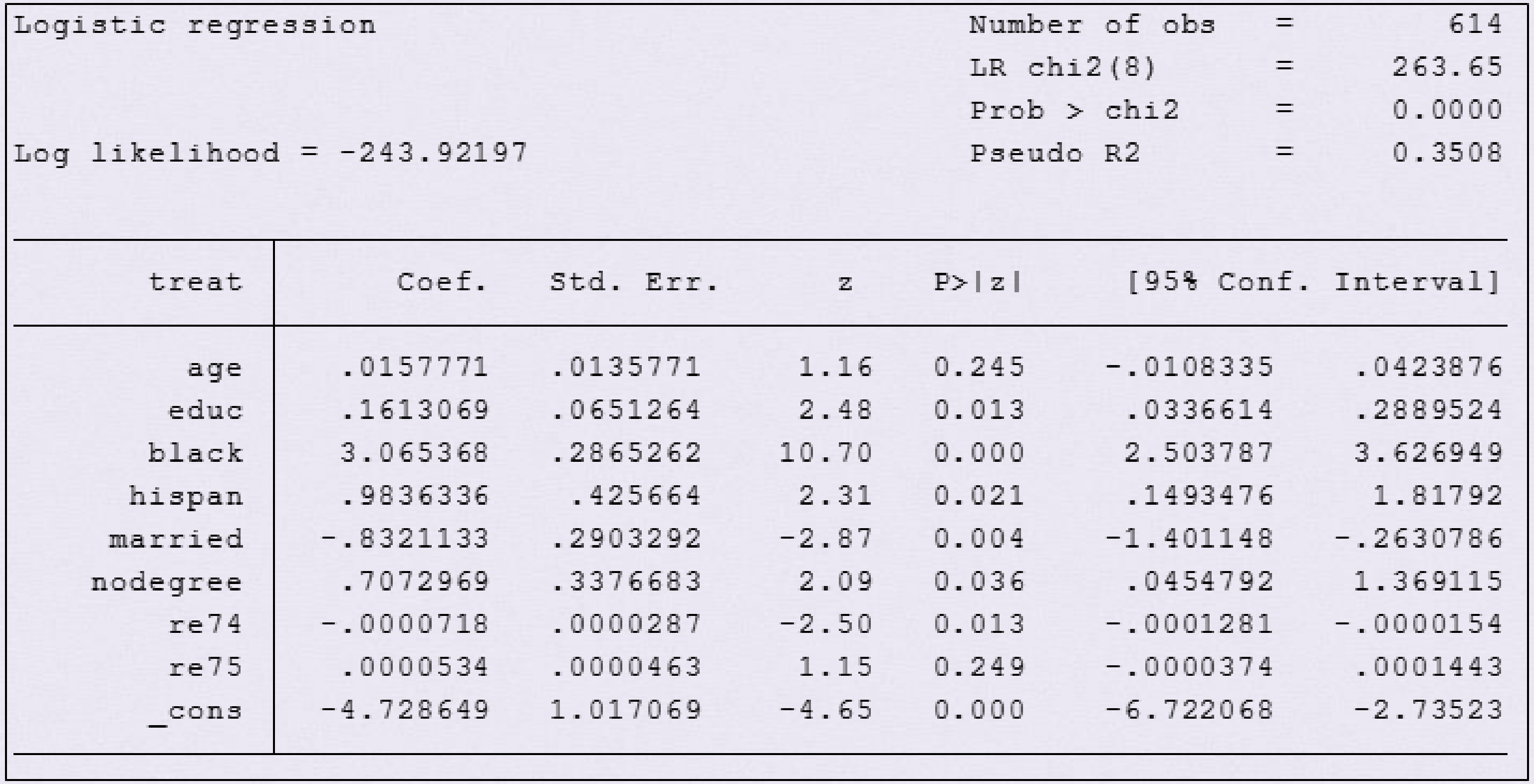
- 3.2. 命令解读:
  - psmatch2 因变量 协变量，[选择项]
  - 本例选择“nearest neighbor matching within caliper”匹配方法。out(re78) 指名结局变量。
  - logit 指定使用logit模型进行拟合，默认的是probit模型。
  - neighbor(1) 指定按照1:1进行匹配，如果要按照1:3进行匹配，则设定为neighbor(3)，本例中因对照组样本量有限，仅适合1:1进行匹配。
  - common 强制排除试验组中倾向值大于对照组最大倾向值或低于对照组最小倾向值的观测。
  - caliper(.05) 试验组与匹配对照所允许的最大距离为0.05。
  - ties 强制当试验组观测有不止一个最优匹配时同时记录。

PSM实战 -- Stata操作 III

- 3.3. 命令解读:
  - pstest, both 做匹配后均衡性检验，理论上说此处只能对连续变量做均衡性检验，对分类变量的均衡性检验应该重新整理数据后运用χ2检验或者秩和检验。但此处对于分类变量也有一定的参考价值。
  - psgraph 对匹配的结果进行图示。

PSM实战 -- Stata操作 IV

- 4. 结果解读
- 4.1 模型拟合结果，此处无太多实际意义。

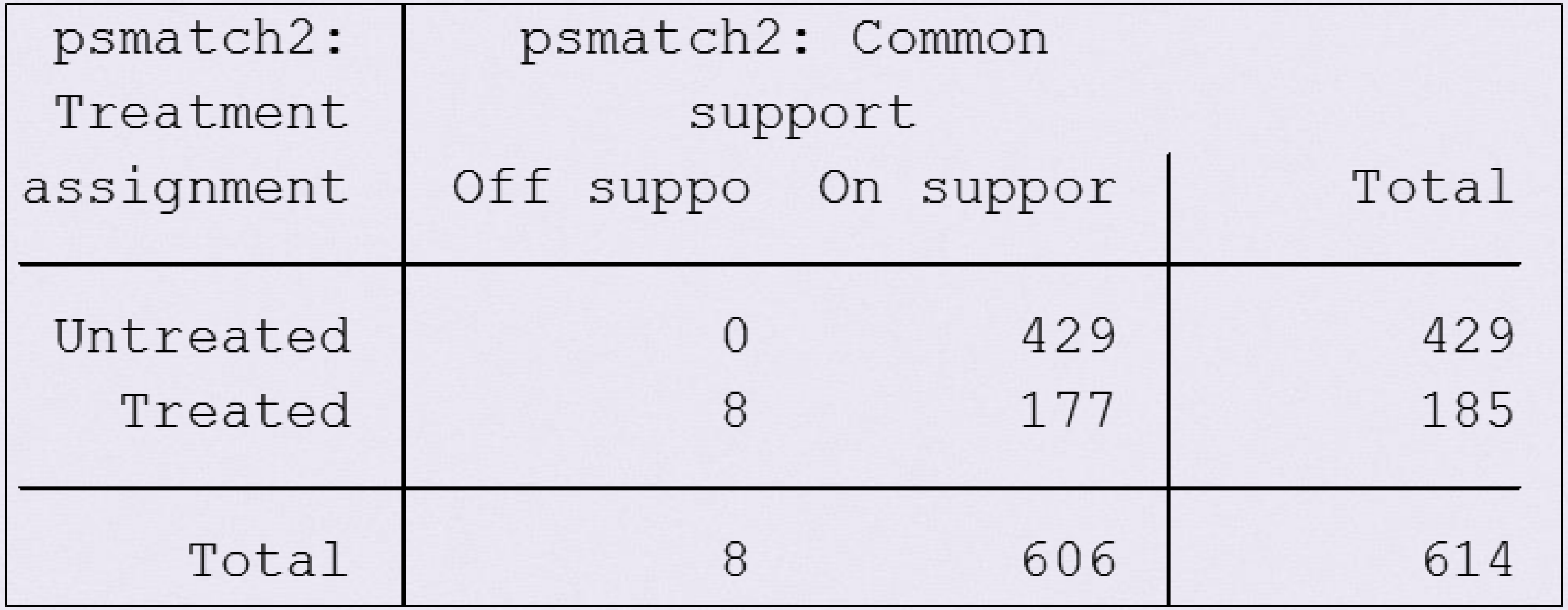


```
Logistic regression                             Number of obs     =        614
                                                LR chi2(8)        =     263.65
                                                Prob > chi2       =     0.0000
Log likelihood = -243.92197                     Pseudo R2         =     0.3508

------------------------------------------------------------------------------
       treat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0157771   .0135771     1.16   0.245    -.0108335    .0423876
        educ |   .1613069   .0651264     2.48   0.013     .0336614    .2889524
       black |   3.065368   .2865262    10.70   0.000     2.503787    3.626949
      hispan |   .9836336    .425664     2.31   0.021     .1493476     1.81792
     married |  -.8321133   .2903292    -2.87   0.004    -1.401148   -.2630786
    nodegree |   .7072969   .3376683     2.09   0.036     .0454792    1.369115
        re74 |  -.0000718   .0000287    -2.50   0.013    -.0001281   -.0000154
        re75 |   .0000534   .0000463     1.15   0.249    -.0000374    .0001443
       _cons |  -4.728649   1.017069    -4.65   0.000    -6.722068    -2.73523
------------------------------------------------------------------------------
```

PSM实战 -- Stata操作 IV

➢ 4.2 试验组可匹配的观测概览，按照命令中设定的匹配规则，试验组有8例患者未能匹配到合适对照。



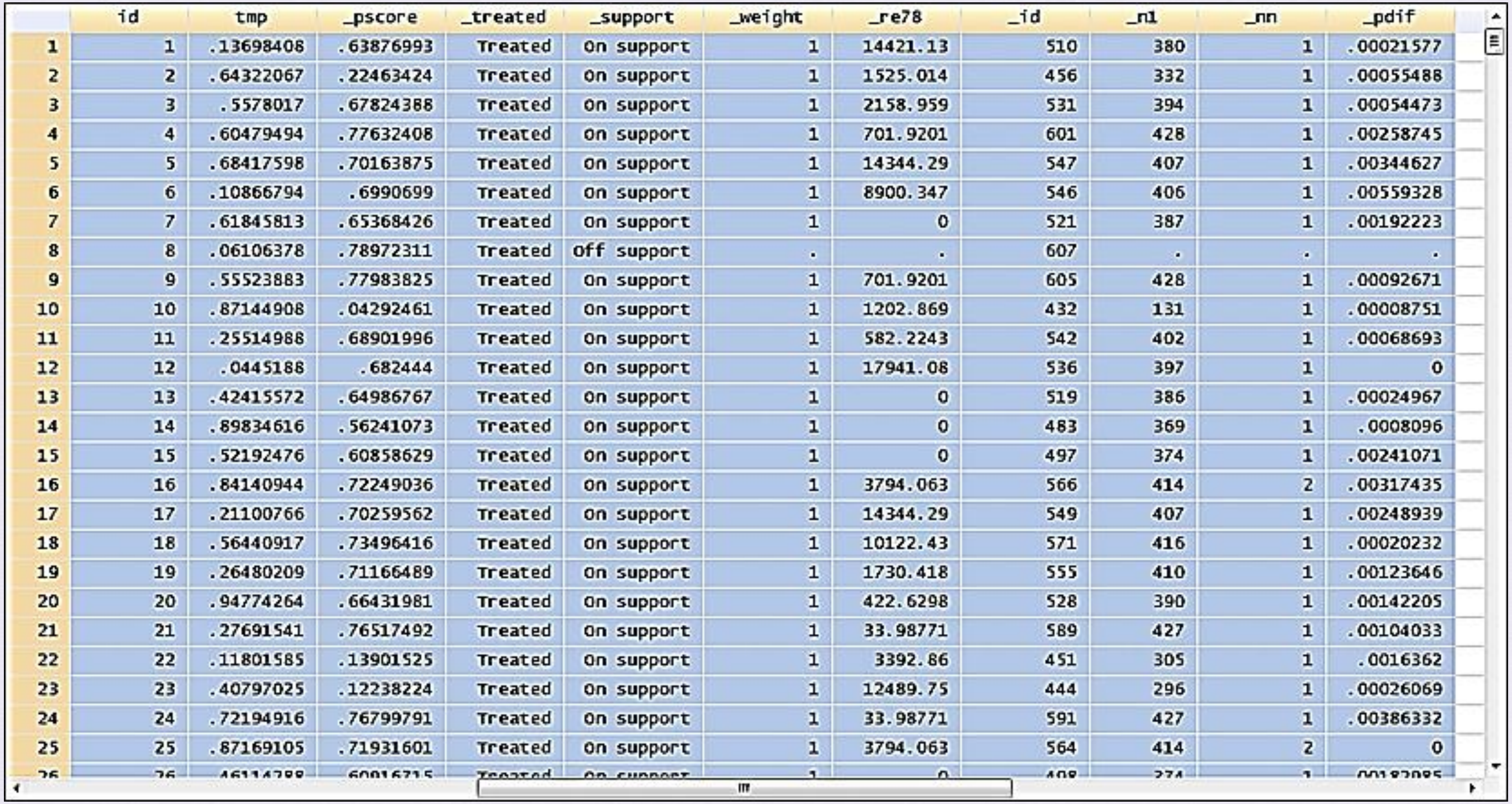
```
 psmatch2: |   psmatch2: Common
 Treatment |        support
assignment | Off suppo  On suppor |     Total
-----------+----------------------+----------
 Untreated |         0        429 |       429
   Treated |         8        177 |       185
-----------+----------------------+----------
     Total |         8        606 |       614
```

PSM实战 -- Stata操作 IV

- 4.3 结果解读的重点应该是对Stata新生成的中间变量的解读。
  - 其中_pscore 是每个观测值对应的倾向值；
  - _id 是自动生成的每一个观测对象唯一的ID（事实上这列变量即是对_pscore 排序）；
  - _treated 表示某个对象是否为试验组；
  - _n1 表示的是他被匹配到的对照对象的_id（如果是1:3匹配，还会生成_n2, _n3）；
  - _pdif 表示一组匹配了的观察对象他们概率值的差。

为了观察方便可以按照id变量进行排序，排序后结果如下图所示：

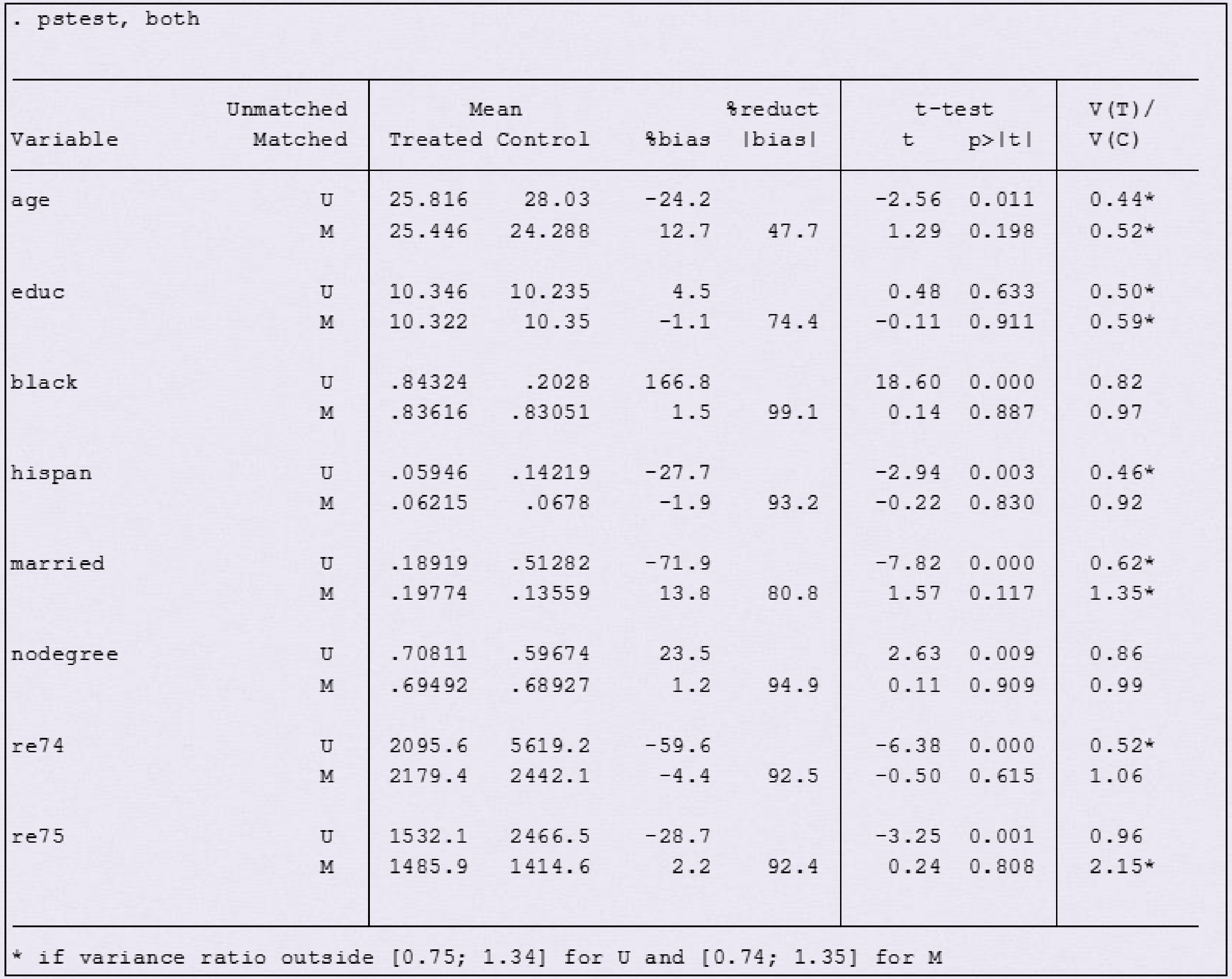
PSM实战 -- Stata操作 IV



	id	tmp	_pscore	_treated	_support	_weight	_re78	_id	_n1	_nn	_pdif
1	1	.13698408	.63876993	Treated	On support	1	14421.13	510	380	1	.00021577
2	2	.64322067	.22463424	Treated	On support	1	1525.014	456	332	1	.00055488
3	3	.5578017	.67824388	Treated	On support	1	2158.959	531	394	1	.00054473
4	4	.60479494	.77632408	Treated	On support	1	701.9201	601	428	1	.00258745
5	5	.68417598	.70163875	Treated	On support	1	14344.29	547	407	1	.00344627
6	6	.10866794	.6990699	Treated	On support	1	8900.347	546	406	1	.00559328
7	7	.61845813	.65368426	Treated	On support	1	0	521	387	1	.00192223
8	8	.06106378	.78972311	Treated	Off support	.	.	607	.	.	.
9	9	.55523883	.77983825	Treated	On support	1	701.9201	605	428	1	.00092671
10	10	.87144908	.04292461	Treated	On support	1	1202.869	432	131	1	.00008751
11	11	.25514988	.68901996	Treated	On support	1	582.2243	542	402	1	.00068693
12	12	.0445188	.682444	Treated	On support	1	17941.08	536	397	1	0
13	13	.42415572	.64986767	Treated	On support	1	0	519	386	1	.00024967
14	14	.89834616	.56241073	Treated	On support	1	0	483	369	1	.0008096
15	15	.52192476	.60858629	Treated	On support	1	0	497	374	1	.00241071
16	16	.84140944	.72249036	Treated	On support	1	3794.063	566	414	2	.00317435
17	17	.21100766	.70259562	Treated	On support	1	14344.29	549	407	1	.00248939
18	18	.56440917	.73496416	Treated	On support	1	10122.43	571	416	1	.00020232
19	19	.26480209	.71166489	Treated	On support	1	1730.418	555	410	1	.00123646
20	20	.94774264	.66431981	Treated	On support	1	422.6298	528	390	1	.00142205
21	21	.27691541	.76517492	Treated	On support	1	33.98771	589	427	1	.00104033
22	22	.11801585	.13901525	Treated	On support	1	3392.86	451	305	1	.0016362
23	23	.40797025	.12238224	Treated	On support	1	12489.75	444	296	1	.00026069
24	24	.72194916	.76799791	Treated	On support	1	33.98771	591	427	1	.00386332
25	25	.87169105	.71931601	Treated	On support	1	3794.063	564	414	2	0
26	26	.46114788	.60016715	Treated	On support	1	0	498	374	1	.00182085

PSM实战 -- Stata操作 IV

➢ 4.4 均衡性检验结果

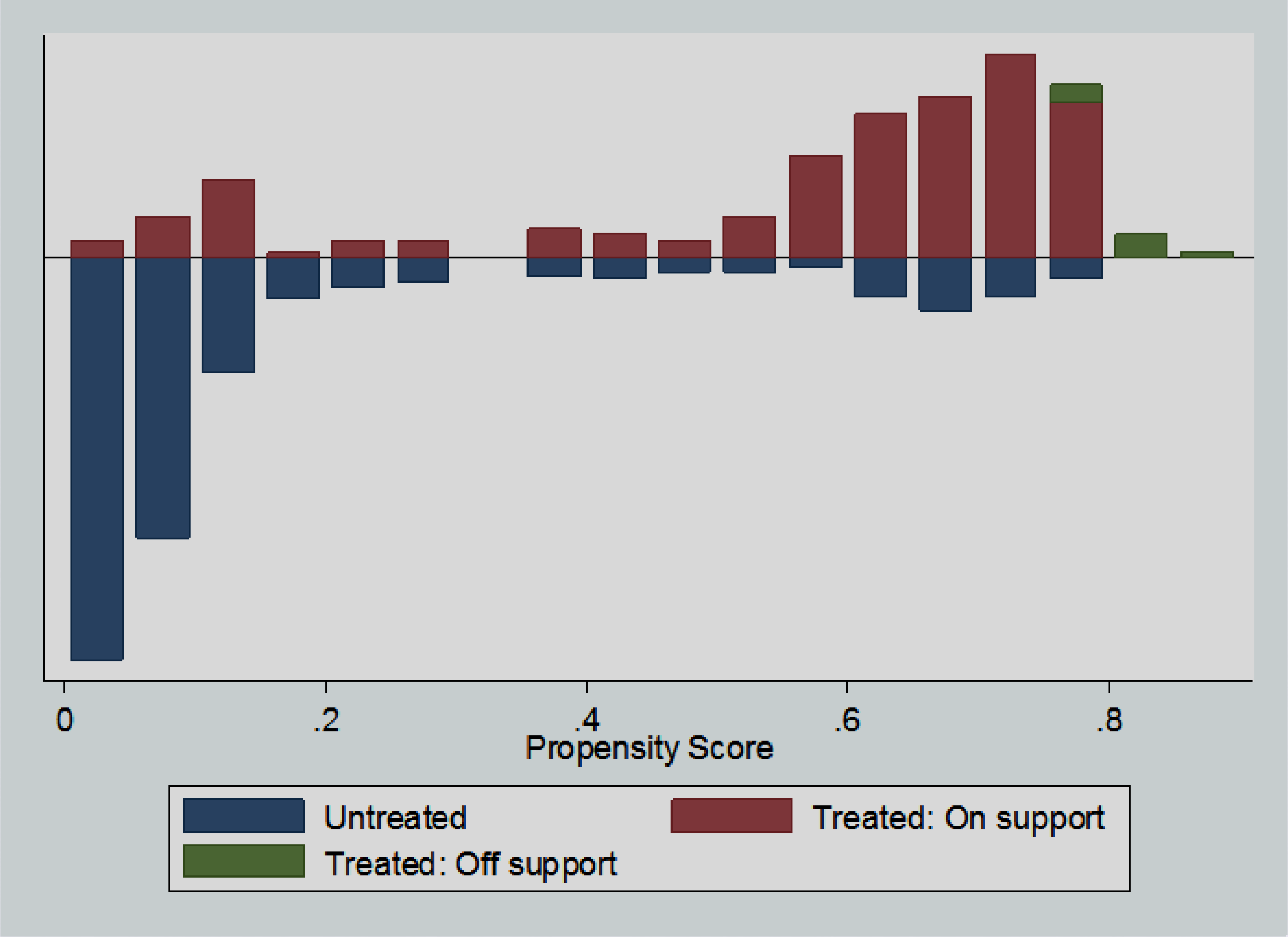


```
. pstest, both

-----------------------------------------------------------------------------------
                Unmatched |       Mean               %reduct |     t-test    |  V(T)/
Variable          Matched | Treated Control    %bias  |bias| |    t    p>|t| |  V(C)
--------------------------+----------------------------------+---------------+----------
age                     U | 25.816    28.03    -24.2         |  -2.56  0.011 |  0.44*
                        M | 25.446   24.288     12.7    47.7 |   1.29  0.198 |  0.52*
                          |                                  |               |
educ                    U | 10.346   10.235      4.5         |   0.48  0.633 |  0.50*
                        M | 10.322    10.35     -1.1    74.4 |  -0.11  0.911 |  0.59*
                          |                                  |               |
black                   U | .84324    .2028    166.8         |  18.60  0.000 |  0.82
                        M | .83616   .83051      1.5    99.1 |   0.14  0.887 |  0.97
                          |                                  |               |
hispan                  U | .05946   .14219    -27.7         |  -2.94  0.003 |  0.46*
                        M | .06215    .0678     -1.9    93.2 |  -0.22  0.830 |  0.92
                          |                                  |               |
married                 U | .18919   .51282    -71.9         |  -7.82  0.000 |  0.62*
                        M | .19774   .13559     13.8    80.8 |   1.57  0.117 |  1.35*
                          |                                  |               |
nodegree                U | .70811   .59674     23.5         |   2.63  0.009 |  0.86
                        M | .69492   .68927      1.2    94.9 |   0.11  0.909 |  0.99
                          |                                  |               |
re74                    U | 2095.6   5619.2    -59.6         |  -6.38  0.000 |  0.52*
                        M | 2179.4   2442.1     -4.4    92.5 |  -0.50  0.615 |  1.06
                          |                                  |               |
re75                    U | 1532.1   2466.5    -28.7         |  -3.25  0.001 |  0.96
                        M | 1485.9   1414.6      2.2    92.4 |   0.24  0.808 |  2.15*
                          |                                  |               |
-----------------------------------------------------------------------------------
* if variance ratio outside [0.75; 1.34] for U and [0.74; 1.35] for M
```

PSM实战 -- Stata操作 IV

➢ 4.5 匹配结果的图示化



PSM实战 -- Stata操作 V

➢ 5. Stata命令汇总

```
.ssc install psmatch2 #安装程序包
.use "F:\lalonde.dta" #调用F盘存储数据
.gen tmp = runiform()
.sort tmp #对所有观测随机排序
.psmatch2 treat age educ black hispan married nodegree re74 re75, out(re78)     logit
neighbor(1) common caliper(.05) ties #PSM分析
.pstest, both #均衡性检验
.psgraph #图示匹配结果
```

有关PSM的几点思考

1. PSM的适用条件：
对照组样本量足够大，对照组与试验组样本量之比5:1以上，确保绝大多数试验组对象可以匹配上合适的对照，最好所有试验组对象均得到良好匹配。
2. PSM与回归的关系
能用PSM的均可以用回归分析，可以用回归的未必可以用PSM，建议同时采用PSM与回归分析处理数据，当两者结果一致的时候说明结果较可信

请在此处输入小标题

感谢观看

THANKS

丁香园特邀讲师 周支瑞