第六章 匹配方法

展示人：侯天宇

CIRG2020，2020-11-19

内容框架

- 第五节 倾向得分匹配法操作步骤
- 第六节 倾向得分匹配法实例
- 第七节 匹配方法运用常见问题

第五节 倾向得分匹配法操作步骤

- 估计倾向得分
- 匹配前均衡检验
- 评估共同支撑域条件
- 选择匹配方法
- 匹配后均衡检验
- 计算处置效应

估计倾向得分

- 模型的选择
  - 通常使用Probit模型或Logit模型来估计每个样本接受处置的概率，所得概率即为匹配得分。
  - Probit模型估计方程：

$$\Pr(D_i = 1 \mid \boldsymbol{X}) = \Phi(\boldsymbol{\beta X})$$

其中，$\Phi(\cdot)$是正态分布的累积概率函数。

  - Logit模型估计方程：

$$\Pr(D_i = 1 \mid \boldsymbol{X}) = F(\boldsymbol{\beta X})$$

其中，$F(\boldsymbol{\beta X}) = e^{\boldsymbol{\beta X}}/\left(1 + e^{\boldsymbol{\beta X}}\right)$是logistic分布的累积概率函数。

估计倾向得分

- 变量的选择
- 注：没有标准答案，在实践中遵循以下原则：
  - 倾向得分中应包含能够同时影响处置选择和处置结果的变量
  - 倾向得分中不应包含受处置选择影响的变量，因此特征变量应使用参与处置前的值
  - 倾向得分的估计通常使用Probit和Logit模型，但根本目的并不是要准确估计参与处置的概率，而是通过匹配倾向得分使得处置组和控制组的可观测特征均衡。因此可以在模型中加入一些变量（高阶变量和交叉项）以达到均衡目的。

匹配前均衡检验

- 通常通过检验具有相同倾向得分的处置组和控制组的可观测特征均值是否相同，来检验可观测特征X是否均衡，即：

$$E\left(X_i \mid D_i = 1, ps(X_i)\right) = E\left(X_i \mid D_i = 0, ps(X_i)\right)$$

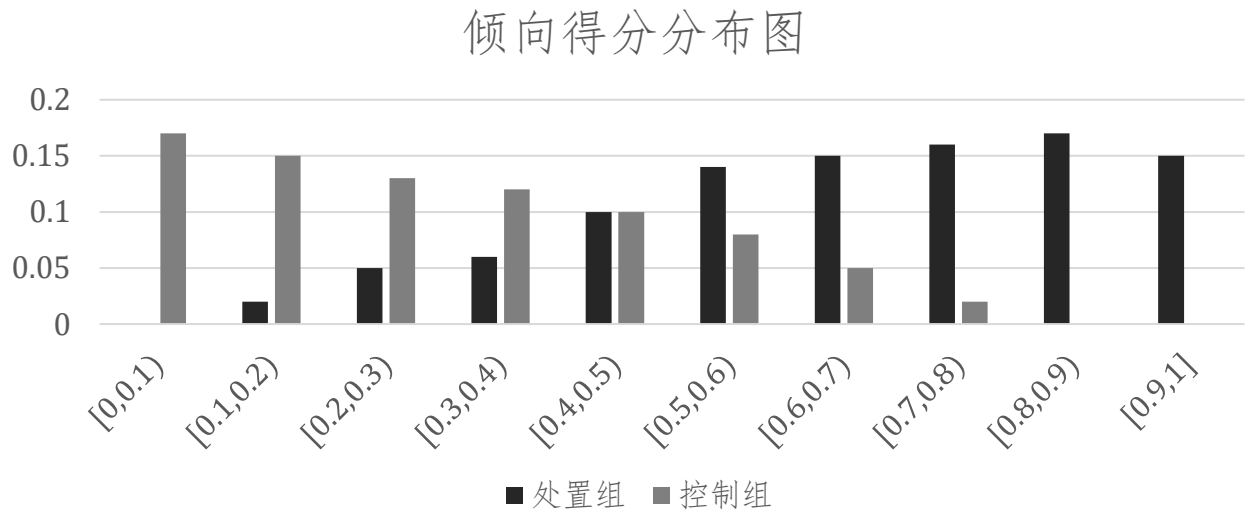
- 分块均衡检验法
  - 根据匹配得分高低将样本分成若干个区间（通常以5等分区间开始），通过t统计值来检验每个区间内处置组和控制组匹配得分是否有差异；如果存在差异，则将区间进一步细分直至没有差异。

匹配前均衡检验

- 分块均衡检验法
  - 检验每个区间内处置组和控制组的可观测特征均值是否相同；如果有显著差异，需要重新调整倾向得分的估计方程（加入高阶变量和交叉项），然后再重新分块。
  - 重复上述步骤，直至每个细分区间内处置组和控制组的可观测特征均值无显著差异。
- 匹配前的分块平衡检验通常与分块匹配法一起使用，如果使用其他匹配方法，则需要在匹配后再次进行均衡检验。

评估共同支撑域条件

- 通过考察处置组和控制组倾向得分的分布，了解共同支撑域条件满足的情况。



  - 在实际运用中，应该考虑只使用有共同支撑域的数据，并检验如果只使用“厚的共同支撑域”样本数据，结果是否稳健。

选择匹配方法

- 分块匹配法
  - 将样本按倾向得分划分为Q个区间，使得每个区间内的处置组和控制组的平均倾向得分和可观测特征都达到均衡。
  - 每个区间内的处置效应为：

$$ATT(q) = \bar{Y}_q^{Treatment} - \bar{Y}_q^{Control}$$

  - 对每个区间的$ATT(q)$进行加权平均，得到平均处置效应为：

$$ATT = \sum_{q=1}^{Q} ATT(q) \frac{N_q^{Treatment}}{N^{Treatment}}$$

  - 缺陷：多变量情形难以保证每个变量在每个块中都均衡，需要接受个别不重要变量的不均衡，因此需要一定的主观判断。

选择匹配方法

- 近邻匹配法
  - 对处置组样本选择控制组中倾向得分最接近的n个样本作为匹配样本。
  - 在实际运用中，控制组样本通常可以重复使用，匹配的平均质量会增加，偏差会减少；代价是方差会变大。
  - 例子：处置组样本倾向得分$(0.6, 0.7)$，控制组样本倾向得分$(0.62, 0.56, 0.3)$，$n = 1$。
    - 可重复使用：$0.6 \to 0.62, 0.7 \to 0.62$
    - 不可重复使用：$0.6 \to 0.62, 0.7 \to 0.56$或$0.7 \to 0.62, 0.6 \to 0.56$
  - 缺陷：即使可以重复使用控制组样本，也存在处置组样本倾向得分和最近控制组样本倾向得分相差较大的可能。

选择匹配方法

- 卡尺匹配法
  - 在近邻匹配的基础上，要求倾向得分差异在一定容忍程度（卡尺）内。
  - 例子：处置组样本倾向得分$(0.6, 0.7)$，控制组样本倾向得分$(0.62, 0.56, 0.3)$，$n = 1$，卡尺为0.05。
    - 可重复使用：$0.6 \rightarrow 0.62, 0.7$无匹配
  - 缺陷：对于如何界定容忍程度没有标准的方法。容忍度过小会增加方差；容忍度过大会增加偏差。

选择匹配方法

- 半径匹配法
  - 允许容忍程度（半径）内的所有样本作为匹配样本。
  - 例子：处置组样本倾向得分$(0.6, 0.7)$，控制组样本倾向得分$(0.62, 0.56, 0.3)$，半径为0.05。
    - 可重复使用：$0.6 \rightarrow (0.62, 0.56), 0.7$无匹配
  - 缺陷：必须确定最大容忍度。容忍度过小会增加方差；容忍度过大会增加偏差。

选择匹配方法

- 核匹配法
  - 对更接近处置组样本的倾向得分的控制组样本赋予更大的权重。如果处置样本i的匹配控制样本有$N_i^{Control}$个，对其中控制样本j赋予的权重为：

$$w_{i,j} = \frac{K\left(\frac{ps(x)_j - ps(x)_i}{h}\right)}{\sum_{j=1}^{N_i^{Control}} K\left(\frac{ps(x)_j - ps(x)_i}{h}\right)}$$

其中，$K(\cdot)$是一个核函数，h是核函数中的平滑参数：带宽。

  - 核函数通常是一个非负、对称且只有单一最大值的密度函数，常用的为Epanechnikov密度函数和Triangle密度函数。
  - 缺陷：带宽h的确定比较困难。

选择匹配方法

- 几种匹配方法小结
  - 在实际应用中，几种方法通常都可以使用以检验结果的稳健性。
  - 除分块匹配法外，其他匹配方法都需要面对不同参数的选择，取决于偏差与方差之间的取舍

选择匹配方法

- 几种匹配方法小结

常见匹配方法对偏差和方差的取舍

匹配方法	特点	偏差	方差
近邻匹配	增加近邻数	+	−
	可重复使用	−	+
卡尺匹配	容忍度增加	+	−
半径匹配	容忍度增加	+	−
核匹配	带宽增加	+	−

匹配后均衡检验

- 标准化偏差
  - 通过比较匹配前后处置组和控制组可观测特征标准化偏差变化来衡量匹配的效果。
  - 匹配前处置组和控制组的特征X_i标准化偏差为：

$$SB_i^{before} = 100 \frac{\bar{X}_{i,treatment}^{before} - \bar{X}_{i,control}^{before}}{\sqrt{0.5\left(\mathrm{Var}\left(X_{i,treatment}^{before}\right) + \mathrm{Var}\left(X_{i,control}^{before}\right)\right)}}$$

  - 匹配后处置组和控制组的特征X_i标准化偏差为：

$$SB_i^{after} = 100 \frac{\bar{X}_{i,treatment}^{after} - \bar{X}_{i,control}^{after}}{\sqrt{0.5\left(\mathrm{Var}\left(X_{i,treatment}^{after}\right) + \mathrm{Var}\left(X_{i,control}^{after}\right)\right)}}$$

匹配后均衡检验

- 标准化偏差
  - 偏差下降度为：

$$BR_i = 1 - \frac{SB_i^{after}}{SB_i^{before}}$$

  - 缺陷：没有标准来衡量偏差降低多少才是效果好的匹配。有学者认为当SB_i^{after}小于20时，匹配结果可以接受。
- t值检验
  - 检验两组的每个特征在匹配后的均值是否有显著偏差。
- F值检验
  - 对所有观测特征在匹配后是否还存在偏差进行共同检验。

计算处置效应

- 除分块匹配法外，其他匹配方法计算ATT的表达式为：

$$ATT = \frac{1}{N^{treatment}} \sum_{i \in I^{Treatment} \cap S_p} \left\{ Y_i - \sum_{j \in I^{Control} \cap S_p} w_{i,j} Y_j \right\}$$

  - 其中，$N^{treatment}$为处置组样本个数，$I^{Treatment}$为处置组集，$I^{Control}$为控制组集，S_p为共同支撑域集，Y_i为处置组里样本i的观测值，Y_j为控制组里样本j的观测值，$w_{i,j}$为匹配的权重。

- 不同匹配方法的区别在于赋予的权重不同。
  - 近邻匹配法$(n = 1)$：最近的一个样本权重为1，其他为0。
  - 半径匹配法：容忍度内包含的n个控制组样本权重均为$1/n$，其他为0。
  - 核匹配法：按核函数值赋予权重。

小结

- 在实际运用中通常使用多种匹配方法和不同参数检验其稳健性，并且尝试不同“共同支撑域”的要求。
- 可见，倾向匹配方法需要主观判断较多，这也是使用该方法的主要弱点。

第六节 倾向得分匹配法实例

- Stata命令总结
- 实例操作

Stata命令总结

- 命令组1：pscore，atts，attnd/attnw，attnr，attk
  - 可进行分块平衡检验，执行步骤详细。
  - 处置结果标准差的估计可使用bootstrap和analytical variance两种方法。其中，bootstrap 考虑到倾向得分也是估计值所造成的误差。
  - 共同支撑域的定义是，样本包含所有处置样本和只在共同倾向得分范围内的控制样本。
- 命令组2：psmatch2，psgraph，pstest
  - 可使用多种匹配方法，提供匹配后特征平衡检验功能。
  - 处置结果标准差的估计可使用bootstrap和analytical variance两种方法。
  - 共同支撑域的定义是，去掉了处置样本中倾向得分低于（高于）控制样本的最小（最大）得分的部分。

Stata命令总结

- 命令组3：teffects psmatch
  - 使用analytical variance进行方差估计，考虑到了倾向得分也是估计值所造成的误差。
- 在实际使用中，可以结合各命令的优点一起使用：
  - 使用pscore进行匹配前的分块平衡检验；
  - 使用psmatch2进行匹配估计和匹配后的平衡检验；
  - 如果方差估计需要考虑到倾向得分也是估计值所造成的误差，可以使用teffects psmatch。

实例操作

- 研究主题：参加就业培训对收入的影响。
- 计算处置组和控制组样本数据特征差别：

```
. tabstat AGE EDUC BLACK HISP MARR NODEGREE RE74 RE75, statistics(mean) by(TREAT)

Summary statistics: mean
  by categories of: TREAT
```

TREAT	AGE	EDUC	BLACK	HISP	MARR	NODEGREE	RE74	RE75
0	34.8506	12.11687	.2506024	.0325301	.8662651	.3052209	19428.75	19063.34
1	25.81622	10.34595	.8432432	.0594595	.1891892	.7081081	2095.574	1532.056
Total	34.22579	11.99439	.2915888	.0343925	.8194393	.3330841	18230	17850.89

实例操作

- 估计倾向得分，匹配前均衡检验，考察共同支撑域

```
. pscore TREAT AGE EDUC BLACK HISP MARR NODEGREE RE74 RE75 U74 U75, logit comsup blockid(block) pscore(myscore)


**************************************************
Algorithm to estimate the propensity score
**************************************************

The treatment is TREAT
```

TREAT	Freq.	Percent	Cum.
0	2,490	93.08	93.08
1	185	6.92	100.00
Total	2,675	100.00	

实例操作

- 估计倾向得分，匹配前均衡检验，考察共同支撑域

Logistic regression

Number of obs = 2675
LR chi2(10) = 929.43
Prob > chi2 = 0.0000
Log likelihood = -207.9341
Pseudo R2 = 0.6909

TREAT	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
AGE	-.105258	.0177476	-5.93	0.000	-.1400427	-.0704734
EDUC	-.0131634	.075999	-0.17	0.862	-.1621188	.1357919
BLACK	2.61126	.3607818	7.24	0.000	1.904141	3.318379
HISP	2.16366	.5928437	3.65	0.000	1.001708	3.325612
MARR	-1.573351	.2833213	-5.55	0.000	-2.128651	-1.018051
NODEGREE	.6403938	.3763713	1.70	0.089	-.0972804	1.378068
RE74	.0000223	.0000311	0.72	0.473	-.0000386	.0000833
RE75	-.0002758	.0000489	-5.64	0.000	-.0003718	-.0001799
U74	3.232991	.4925851	6.56	0.000	2.267542	4.198441
U75	-1.371879	.4599069	-2.98	0.003	-2.27328	-.4704777
_cons	.3714835	1.28189	0.29	0.772	-2.140975	2.883942

实例操作

- 估计倾向得分，匹配前均衡检验，考察共同支撑域

```
Note: the common support option has been selected
The region of common support is [.00028121, .99022711]


Description of the estimated propensity score
in region of common support

                    Estimated propensity score
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0002973       .0002812
 5%     .0003715       .0002822
10%     .0005253       .0002831       Obs                 1,393
25%     .0014721       .0002835       Sum of Wgt.         1,393

50%     .0088792                      Mean           .1327635
                       Largest        Std. Dev.      .2682045
75%     .0707342       .9865269
90%     .5929178       .9892005       Variance       .0719337
95%     .9182312       .9901413       Skewness       2.194775
99%     .9724343       .9902271       Kurtosis       6.438734
```

实例操作

- 估计倾向得分，匹配前均衡检验，考察共同支撑域

```
**********************************************************
Step 1: Identification of the optimal number of blocks
Use option detail if you want more detailed output
**********************************************************


The final number of blocks is 6

This number of blocks ensures that the mean propensity score
is not different for treated and controls in each blocks
```

实例操作

- 估计倾向得分，匹配前均衡检验，考察共同支撑域

```
************************************************************
Step 2: Test of balancing property of the propensity score
Use option detail if you want more detailed output
************************************************************

The balancing property is satisfied

This table shows the inferior bound, the number of treated
and the number of controls for each block
```

Inferior of block of pscore	TREAT 0	1	Total
.0002812	1,068	11	1,079
.1	60	8	68
.2	43	15	58
.4	23	29	52
.6	5	28	33
.8	9	94	103
Total	1,208	185	1,393

```
Note: the common support option has been selected
```

实例操作

- 选择匹配模型，进行匹配后均衡检验，计算处置效应

```
.  attnd RE78 TREAT, pscore(myscore) boot rep(100) comsup


 The program is searching the nearest neighbor of each treated unit.
 This operation may take a while.



ATT estimation with Nearest Neighbor Matching method
(random draw version)
Analytical standard errors

------------------------------------------------------------------------------
n. treat.    n. contr.         ATT     Std. Err.          t
------------------------------------------------------------------------------
      185           60     941.793     1885.669       0.499
------------------------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
nearest neighbour matches
```

实例操作

- 选择匹配模型，进行匹配后均衡检验，计算处置效应

ATT estimation with Nearest Neighbor Matching method
(random draw version)
Bootstrapped standard errors

n. treat.	n. contr.	ATT	Std. Err.	t
185	**60**	**941.793**	**1091.577**	**0.863**

Note: the numbers of treated and controls refer to actual nearest neighbour matches

实例操作

- 选择匹配模型，进行匹配后均衡检验，计算处置效应

```
. bootstrap r(att), reps(100): psmatch2 TREAT AGE EDUC BLACK HISP MARR NODEGREE RE74 RE75 U74 U75, logit outcome(RE78)
>  neighbor(1) common ties
(running psmatch2 on estimation sample)

Bootstrap replications (100)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100

Bootstrap results                               Number of obs      =      2,675
                                                Replications       =        100

      command:  psmatch2 TREAT AGE EDUC BLACK HISP MARR NODEGREE RE74 RE75 U74 U75, logit outcome(RE78) neighbor(1)
                common ties
        _bs_1:  r(att)
```

	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
_bs_1	**783.7636**	**1218.883**	**0.64**	**0.520**	**-1605.204**	**3172.731**

实例操作

- 选择匹配模型，进行匹配后均衡检验，计算处置效应

```
. pstest AGE EDUC BLACK HISP MARR NODEGREE RE74 RE75, both
```

Variable	Unmatched / Matched	Mean Treated	Mean Control	%bias	%reduct |bias|	t-test t	t-test p>|t|	V(T)/ V(C)
AGE	U	25.816	34.851	-100.9		-11.57	0.000	0.47*
	M	25.934	24.44	16.7	83.5	2.01	0.045	1.03
EDUC	U	10.346	12.117	-68.1		-7.69	0.000	0.43*
	M	10.357	11.176	-31.5	53.8	-3.84	0.000	0.98
BLACK	U	.84324	.2506	148.0		18.13	0.000	.
	M	.84066	.74176	24.7	83.3	2.33	0.020	.
HISP	U	.05946	.03253	12.9		1.94	0.053	.
	M	.06044	.13736	-36.7	-185.6	-2.47	0.014	.
MARR	U	.18919	.86627	-184.2		-25.81	0.000	.
	M	.19231	.12088	19.4	89.5	1.88	0.061	.
NODEGREE	U	.70811	.30522	87.9		11.49	0.000	.
	M	.7033	.56593	30.0	65.9	2.74	0.006	.
RE74	U	2095.6	19429	-171.8		-17.50	0.000	0.13*
	M	2130.1	2735.3	-6.0	96.5	-1.09	0.278	0.75
RE75	U	1532.1	19063	-177.4		-17.50	0.000	0.06*
	M	1546.4	2276.6	-7.4	95.8	-2.03	0.043	0.81

第七节 匹配方法运用常见问题

- 无法解决由不可观测特征自选择造成的偏差。
- 倾向得分匹配法和回归都需要估计方程的参数，前者表现为对$ps(\boldsymbol{X})$的估计。
- 可加入不具有明确经济含义的高阶变量和交叉项以满足可观测变量$\boldsymbol{X}$的平衡性要求。
- 由于存在不可观测特征，条件独立假设无法检验，但可以通过其他渠道去“佐证”这一假设。
- 匹配方法得到的结论仅适用于共同支撑域范围里的样本，因此在得出结论时需要谨慎。