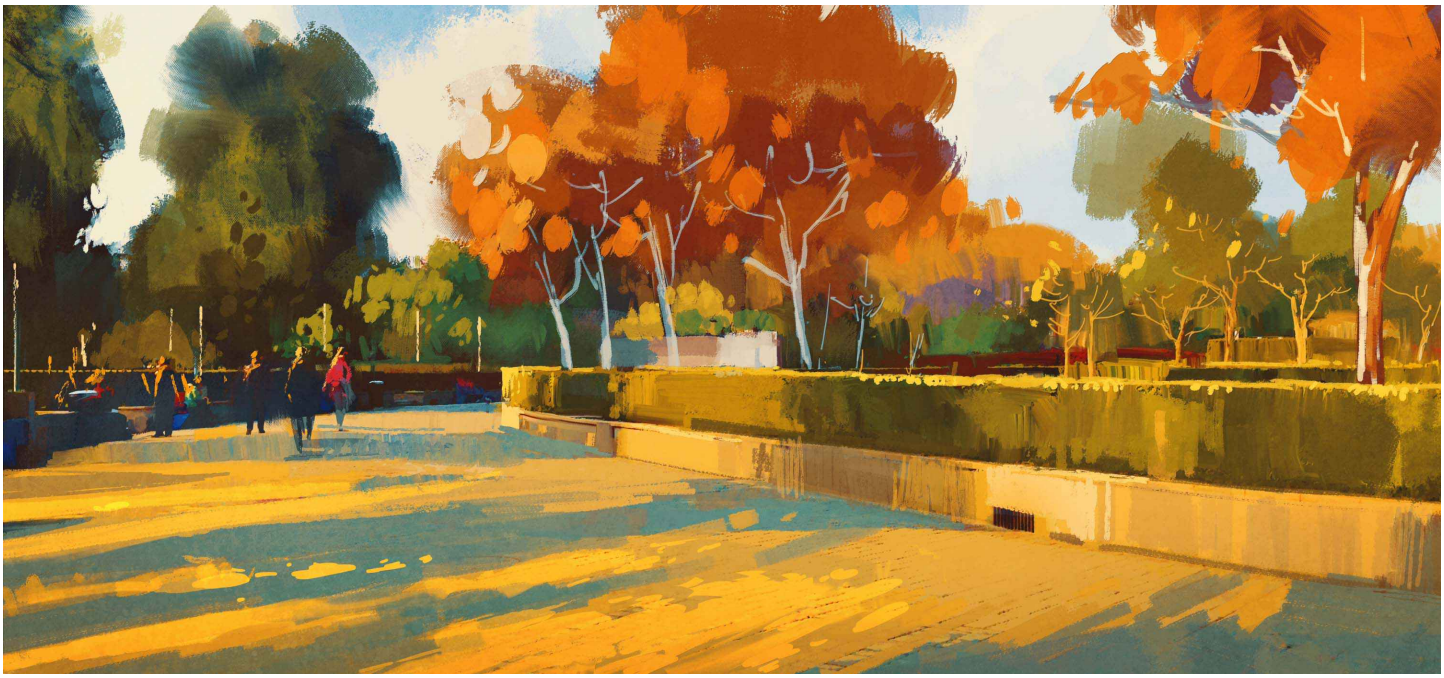
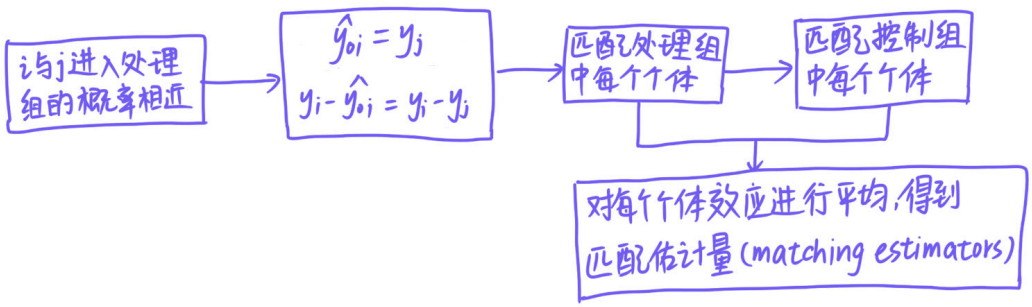


03匹配方法

1. 原理

1.1 匹配（matching）



- 假设个体 `i` 属于处理组，找到属于控制组的某个体 `j` ，是的个体i与个体j的可测变量取值尽可能匹配。
- 优势一：简化分析流程，只需要将被研究对象和他们的匹配对象在响应变量Y上的取值做一个对比即可。
- 优势二：降低对模型的依赖。匹配法的实施过程不存在认为设定好的参数模型，不需要像线性模型那样要求某项随机误差是正态分布，不需要设定方差恒定，不需要假定变量之间的线性关系等。
- **不放回（no replacement）**：每次将匹配成功的个体 `（i，j）` 从样本中去掉，不再参与其余匹配。
- **有放回（with replacement）**：将匹配成功的个体留在样本中，参与其余匹配。

- **允许并列**：将相邻平均值作为估计量。
- **不允许并列**：计算机进行排序。
- 匹配法分为精确匹配和非精确性匹配。

1.2 精确与非精确匹配

- **精确匹配**：匹配的对象要和被匹配的个体的混淆特征的取值要一模一样。如果用 `X` 表示混淆变量， `D` 表示个体间的距离，则精确匹配可表示为：

$$如果\ X_i = X_j，则D_{ij} = 0，否则D_{ij} = \infty$$

- 精确匹配是‘**样本杀手**’，因为精确匹配的是多个混淆变量，因此需要一定的样本量来保证统计。
- **非精确匹配**：要求的匹配关系不一定在混淆因素上取值一模一样，只是近似即可。非精确匹配可以保证样本数量，也是大多数研究使用的方法。

2. 操作过程

2.1 确定距离度量

2.1.1 欧氏距离

- 多维空间中两个点的绝对距离，即每个对应坐标的平方和开根号。例如在二维空间中：

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

```r
# 欧氏距离
iris[1:6,1:4]
dist(iris[1:6,1:4],method = 'euclidean')
```

```r
> iris[1:6,1:4]
# 数据集
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.1         3.5          1.4         0.2
2          4.9         3.0          1.4         0.2
3          4.7         3.2          1.3         0.2
4          4.6         3.1          1.5         0.2
5          5.0         3.6          1.4         0.2
6          5.4         3.9          1.7         0.4

> dist(iris[1:6,1:4],method = 'euclidean')
```

```
# 0.5385165表示第一行数据与第二行数据之间的欧氏距离是0.5385165。
          1         2         3         4         5
2 0.5385165
3 0.5099020 0.3000000
4 0.6480741 0.3316625 0.2449490
5 0.1414214 0.6082763 0.5099020 0.6480741
6 0.6164414 1.0908712 1.0862780 1.1661904 0.6164414
```

2.1.2 马哈拉诺比斯距离

$$D_{ij} = \sqrt{(x_i - x_j)^T \sum^{-1} (x_i - x_j)}$$

- R语言实现

```
# 马氏距离
library(StatMatch)
mahalanobis.dist(iris[1:6,1:4])
```

```
> mahalanobis.dist(iris[1:6,1:4])
# 结果
          1         2         3         4         5         6
1 0.000000 2.237463 2.931554 2.931554 1.216908 2.781191
2 2.237463 0.000000 2.983245 2.983245 3.087399 3.076939
3 2.931554 2.983245 0.000000 3.162278 2.645122 3.146122
4 2.931554 2.983245 3.162278 0.000000 2.645122 3.146122
5 1.216908 3.087399 2.645122 2.645122 0.000000 2.828648
6 2.781191 3.076939 3.146122 3.146122 2.828648 0.000000
```

2.1.3 倾向值

- 个体进入实验组（相对于控制组而言）的概率（`e`）。倾向值是所有混淆变量的一个总结变量。

$$D_{ij} = |e_i - e_j|$$

- 这里倾向值的取值范围是0~1，通过logit变换可以将其转换为正无穷到负无穷。再添加上绝对值来表示距离。

$$logit(e) = \frac{e}{1-e}$$

$$D_{ij} = \left|logit(e_i) - logit(e_j)\right|$$

- 倾向值匹配和马氏距离的综合应用：首先使倾向值在可接受的范围内，进一步使用马哈拉诺比斯匹配或者精确匹配。用公式表示为：

$$如果\left|logit(e_i) - logit(e_j)\right| \leq c,$$

$$则D_{ij} = \sqrt{(x_i - x_j)^T \sum^{-1} (x_i - x_j)}，否则D_{ij} = \infty。$$

2.2 匹配对象个数

2.2.1 匹配个数

- **一对一匹配（one-to-one matching）**：对每个个体寻找一个不同组的最近个体进行匹配。后续分析时的方差会很大。现实研究中使用最多的匹配方式。
- **一对多匹配（one-to-many matching）**：对每个个体寻找多个不同组的最近个体进行匹配。反差相对而言会小，但是接近的样本被匹配进来，匹配的质量受到影响，整体的误差会提高。

2.2.2 样本损失问题

- 匹配之后一定会发生样本损失的问题，但样本损失不会对统计检定力造成大的影响。

1. 统计检定力取决于实验组，控制组哪一个更小。例如我们关心ATT，最理想的情况是实验组比控制组小，保证每一个实验组中的对象都能匹配成功。
2. 通过匹配可以提高样本相似度，从而降低了因为模型不确定性带来的误差，进而提高统计检定力。匹配完成后会把没有匹配的个体删除掉，实际上就是保留有用信息，删除冗余信息。

2.3 匹配对象重复使用

- **贪婪匹配**：某个个体匹配完成后，这个个体就会被拿走，不会参与到后续的匹配中。
- 重复匹配：某个个体匹配完成后还能进行匹配。 存在一个问题，每个匹配对彼此之间不独立。控制组的信息变得非常同质化。

2.4 匹配算法

2.4.1 贪婪匹配

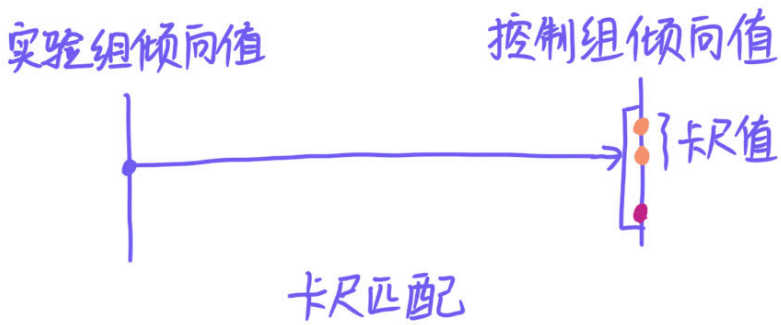
- 某个个体匹配完成后，这个个体就会被拿走，不会参与到后续的匹配中。

2.4.2 最优匹配

- 考虑整体的匹配优度，关注的是整体的平均距离。
- **全匹配（full matching）**：不再关注每两个个体的匹配，而是把所有人看作一个整块，对整体进行切分，尽量保证每一块内部，匹配对象的整体距离最小。最终目的是保证每一个实验组的个体都能匹配上。
- 优点一：不存在顺序效应。
- 优点二：能保证所有实验组的个体都能在控制组中找到匹配的对象。

2.4.3 倾向得分匹配

- 只考虑倾向值大小，是一维水平上的匹配。
- **卡尺（caliper）匹配：**卡尺是一个半径范围，即‘容忍度’，比较常见的卡尺半径值是0.01，0.05。



2.4.4 核函数匹配

- 针对实验组的样本1，我们对控制组按照得分给每个控制组样本权重，越接近0.6赋予越大的权重。

实验组		控制组		
ID	得分	ID	得分	权重
1	0.6	A	0.5	w_1
2	0.5	B	0.63	w_2
3	0.3	C	0.9	w_3

- 与样本1进行匹配的不再是一个单一个体的观测值，而是多个数值的加权平均：

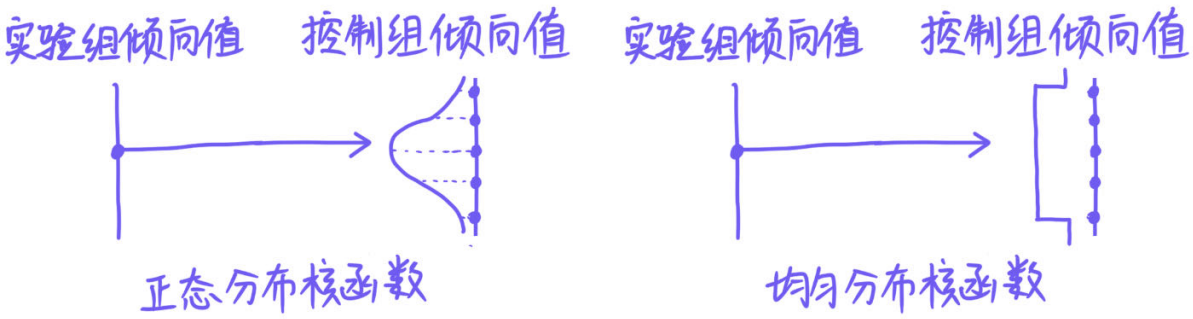
$$0.5w_1 + 0.63w_2 + 0.9w_3$$

- 其中：

$$w_1 + w_2 + w_3 = 1$$

- 权重的分布可以是正态分布，均匀分布，对应的函数即为**核函数（kernel function）**。

- 正态分布核函数与均匀分布核函数，其中均匀分布核函数即卡尺匹配。



- 核函数匹配中每个被研究对象都能用上，能够尽可能保留原始数据

2.4.5 遗传匹配

- 遗传匹配（genetic matching）是倾向值匹配和马哈拉诺比斯匹配的一个综合应用。
- 在遗传匹配中，个体间的距离计算是：

$$D_{ij} = \sqrt{(x_i - x_j)^T (\sum^{-1/2})^T W \sum^{-1/2} (x_i - x_j)}$$

- 与马哈拉诺比斯距离相比，遗传匹配采用的距离度量增加了一个权重矩阵，是为了平衡马哈拉诺比斯距离和倾向值距离对最后的距离取值的作用大小。
- 在计算遗传匹配的距离时，原始数据增加一列，即将估计出的倾向值作为一个新的混淆变量增加到原始数据中去。例如一个研究中有三个混淆变量，差值为 `d` ，倾向值的插值为 `dps` 则遗传匹配的距离为

$$D_{ij} = \sqrt{(d_1, d_2, d_3, d_{ps})(S^{-1/2})^T W S^{-1/2} \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_{ps} \end{bmatrix}}$$

- W是这样一个矩阵：非对角线元素为0，对角线元素是一系列的权重值。

$$\begin{bmatrix} w_1 & 0 & 0 & 0 \\ 0 & w_2 & 0 & 0 \\ 0 & 0 & w_3 & 0 \\ 0 & 0 & 0 & w_4 \end{bmatrix}$$

- 如果 `w4 = 0` , `w1 = w2 = w3 = 1` 则 `Dij` 为传统马哈拉诺比斯距离。
- 如果 `w4 =` 1, `w1 = w2 = w3 =` 0则 `Dij` 为传统倾向值距离。

- 研究者可以通过权重的取值来平衡匹配过程是偏向于多变量的马哈拉诺比斯匹配还是一维的倾向值匹配。

2.4.6 回归调整匹配

- 回归调整匹配将传统的线性模型和匹配方法结合起来，以求尽可能削减因为实验组和控制组之间混淆变量不平衡造成的误差。

3. 通过匹配法识别的三个假设

3.1 条件均值假设

$$\mathrm{E}(Y_1|\mathbf{x},D)=\mathrm{E}(Y_1|\mathbf{x}) \text{ 和 } \mathrm{E}(Y_0|\mathbf{x},D)=\mathrm{E}(Y_0|\mathbf{x}).$$

3.2 重叠假设

$$0<p(\mathbf{x})<1,$$

$$p(\mathbf{x})=\Pr(D=1|\mathbf{x})$$

3.3 平衡假设

$$\{D\perp \mathbf{x}|\text{ Matching }\},$$

4. 优度衡量

4.1 标准均值差

- 衡量匹配的质量：观察匹配完成后，混淆变量在实验组和控制组之间的差异是不是变小了。
- 对于连续性混淆变量而言**标准均值差**：

$$\text{标准均值差}=\frac{\bar{x}_t-\bar{x}_c}{\sqrt{\frac{s_t^2-s_c^2}{2}}}$$

- **T检验统计量**：

$$\text{T检验统计量} = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

- 我们的目的是看实验组和控制组之间的混淆变量是不是在匹配之后变得更加相似，这个相似程度不应该随着样本量的大小而改变。
- 对于离散型混淆变量而言，标准均值差为：

$$\text{标准均值差} = \frac{\hat{p}_t - \hat{p}_c}{\sqrt{\frac{\hat{p}_t(1-\hat{p}_t) + \hat{p}_c(1-\hat{p}_c)}{2}}}$$

4.2 比较倾向值分布的重叠程度

- 通过比较某一部分在实验组和控制组之间的差异的大小来衡量匹配质量
- 例如：在实验组选择倾向值为0.3，比0.3小的在实验组中占比30%，在控制组中比0.3小的占比29%，则说明实验组和控制组的分布很像；再比如在实验组中A和B两点之间占实验组的95%，在控制组中A，B两点之间占控制组的96%，也能说明实验组和控制组分布很像。

4.3 匹配后的分析过程

1. 做组间比较（如T检验）
2. 基于匹配数据做回归模型

4.4 R语言实现

```r
# 匹配分析
library(MatchIt)
library(foreign)
library(cobalt)
data(lalonde)
head(lalonde)
attach(lalonde)

# 精确匹配
out1_exact <- matchit(treat ~ age + educ + race + married + nodegree,
                      data = lalonde, method = 'exact')
# 细分
out2_subc <- matchit(treat ~ re74 + re75 + educ + age,
                     data = lalonde, method = 'subclass',subclass = 5)
```

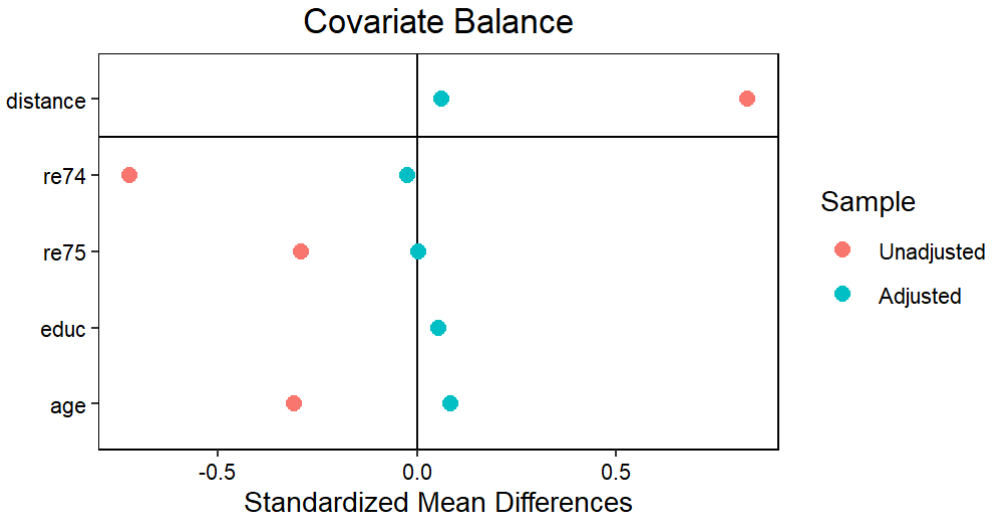
```r
# 最近距离匹配
out3_nearest <- matchit(treat ~ re74 + re75 + educ + age,
                    data = lalonde, method = 'nearest')
# 1：2最优匹配
out4_optimal <- matchit(treat ~ re74 + re75 + educ + age,
                        data = lalonde, method = 'optimal',ratio = 2)
# 全匹配
out5_full <- matchit(treat ~ re74 + re75 + educ + age,
                        data = lalonde, method = 'full')
# 遗传匹配
out6_genetic <- matchit(treat ~ re74 + re75 + educ + age,
                     data = lalonde, method = 'genetic')

# 输出结果
summary(out1_exact)
summary(out2_subc)
summary(out3_nearest)
summary(out4_optimal)
summary(out5_full)
summary(out6_genetic)

# 检验平衡性
love.plot(out6_genetic)
plot(out6_genetic,type = 'jitter')
plot(out6_genetic,type = 'hist')
```

```r
> head(lalonde)
# 数据集
# treat表示是否参加培训项目，re78是响应变量，其余为混淆变量
# re74，re75，分别表示1974，1975年的实际收入，nodegree表示是否是高中文凭
  treat age educ   race married nodegree re74 re75       re78
1     1  37   11  black       1        1    0    0  9930.0460
2     1  22    9 hispan       0        1    0    0  3595.8940
3     1  30   12  black       0        0    0    0 24909.4500
4     1  27   11  black       0        1    0    0  7506.1460
5     1  33    8  black       0        1    0    0   289.7899
6     1  22    9  black       0        1    0    0  4056.4940
```

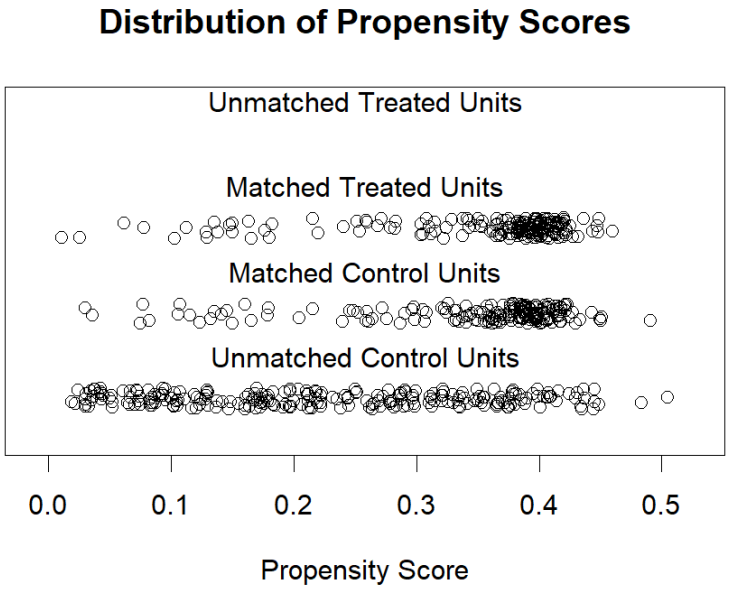
```r
> summary(out1_exact)
# 精确匹配输出结果
Summary of Balance for All Data:
           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF
age              25.8162       28.0303         -0.3094     0.4400    0.0813   0.
```

```
educ             10.3459       10.2354          0.0550     0.4959    0.0347   0.
raceblack         0.8432        0.2028          1.7615          .    0.6404   0.
racehispan        0.0595        0.1422         -0.3498          .    0.0827   0.
racewhite         0.0973        0.6550         -1.8819          .    0.5577   0.
married           0.1892        0.5128         -0.8263          .    0.3236   0.
nodegree          0.7081        0.5967          0.2450          .    0.1114   0.

Summary of Balance for Matched Data:
           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF
age              19.9815       19.9815               0     0.9936         0
educ             10.3333       10.3333               0     0.9936         0
raceblack         0.7778        0.7778               0          .         0
racehispan        0.0370        0.0370               0          .         0
racewhite         0.1852        0.1852               0          .         0
married           0.0370        0.0370               0          .         0
nodegree          0.7593        0.7593               0          .         0

Sample Sizes:
               Control Treated
All             429.       185
Matched (ESS)    40.36      54
Matched          59.        54
Unmatched       370.       131
Discarded         0.         0
```

```
> # 检验混淆变量在匹配前后的平衡性
> love.plot(out6_genetic)
```




```
> plot(out6_genetic,type = 'jitter')
```




```
1 > plot(out6_genetic,type = 'hist')
```