Rubin因果模型与实验的计量分析

慧航

2024年11月

Rubin因果模型

思想来源：

1. 实验
   1. Neyman (1923) 提出了潜在结果（potential outcomes）
   2. Fisher (1925, 1935)：随机性与推断
2. 经济学
   1. 自选择（Roy, 1951）
   2. Heckman（1974）、Borjas（1987）等应用

Rubin因果模型

Rubin (1974):

- 使用潜在结果定义因果效应
- 在观察数据（observational study）中使用潜在因果的语言。
- **“Rubin因果模型”（Rubin causal model）**或者Neyman-Rubin因果模型

与以往的计量方法区别:

- 以往的计量经济学往往严重依赖经济学模型对于数据生成过程的设定，是一种“经济学方法”;
- 而在使用潜在结果的因果推断语言中，无需对结果的数据生成过程做假定，更多的是一种“统计学方法”。

处理组与控制组

- 因果这一概念是从随机实验中拓展而来。
- 在随机实验中，往往会通过区分控制组（control group）和实验组（experiment group）或处理组（treatment group），比较组间的区别
    - 其中的实验组（处理组）往往会施加某些操纵（manipulation）或者处理（treatment）
    - 而控制组往往不施加额外的处理。
    - 比如，在Jensen和Miller(2011)中，不给优惠券的就是控制组，而得到优惠券的三个组别受到了优惠券这一“处理”，从而是三个处理组。
- 不失一般性，我们首先考虑控制组、处理组的二分类情况，记$w_i = 0$代表i个体属于控制组，而$w_i = 1$代表i个体属于处理组。

潜在结果

- Rubin（1975）强调，没有操纵（manipulation）就没有因果
- 在观察数据中，我们也可以以实验作为基准，假想“如果某个个体受到了某种处理，其结果会是怎样的”。
- 如此，我们可以定义潜在结果$y_i(w)$，即当$w_i = w$时，其结果变量y的取值。
- 在$w_i = 0/1$二分类的情况中，每个个体i都会有$y_i(1)$和$y_i(0)$两个不同的潜在结果。
    - 如果记$w_i = 1$为接收某项职业培训，$w_i = 0$代表没有接收培训，而结果变量y为收入
    - $y_i(1)$即如果个体i接受了培训的收入，而$y_i(0)$则为如果个体i没有接收培训的收入。

反事实

- 然而，现实中，我们无法同时观察到所有的潜在结果：
  - 如果个体i接受了培训，我们可以观察到$y_i(1)$，但是$y_i(0)$是观察不到的；
  - 反过来，如果个体i没有接受培训，$y_i(0)$可被观测但是$y_i(1)$不可被观测。
  - 从而，事实上观察到的变量可以写为

$$
\begin{aligned}
y_i &= w_i y_i(1) + (1 - w_i) y_i(0) \\
&= \begin{cases} y_i(1) & w_i = 1 \\ y_i(0) & w_i = 0 \end{cases}
\end{aligned}
$$

- 相对于事实上观察到的y，没有被观察到的那个结果被称为反事实（counterfacturals），从而如果$w_i = 1$，$y_i(0)$就是反事实；如果$w_i = 0$，$y_i(1)$就是反事实。

因果效应

- 使用潜在结果这一语言，我们就可以定义因果效应（causal effect）或者处理效应（treatment effect），即假设受到处理的潜在结果与假设没有受到处理的潜在结果之差：

$$\tau_i = y_i(1) - y_i(0)$$

- 如前所述，$y_i(1), y_i(0)$中有且仅有一个可以被观测，从而个体i的处理效应τ_i也是不可观测的。
- 从这个角度而言，因果推断的问题本质上是一个“缺失数据”（missing-data）问题，所谓因果推断即如何对反事实进行推断的问题：只要得到了反事实，与事实之间的差距就是因果效应。

RCM v.s. DGP

- 以往我们会通过数据生成过程这一工具进行计量模型建模
  - 比如当我们使用回归模型：

$$y_i = \alpha w_i + x_i'\beta + u_i$$

    时，我们假设了y_i是由w_i和x_i的线性组合决定的
  - 从而$y_i(1) = \alpha + x_i'\beta + u_i$，而$y_i(0) = x_i'\beta + u_i$
  - 缺点：假设太强
    - 比如如果根据这样的数据生成过程，$\tau_i = \alpha$，暗含了所有人的处理效应都是同质的这一假设。

RCM v.s. DGP

- 然而在Rubin因果模型中，我们不对y的数据生成过程做任何假设，而是转而对w_i的决定，即分配机制（assignment mechanism）进行分析。
- 实际上，实验正是通过随机化的分配机制（随机分组）达到因果效应的识别
- 理想状况下，我们的观测数据也应该像实验一样，具有随机化的分配机制，如此就可以对因果效应进行识别。
- 如此，在Rubin因果模型的设定下，我们对于因果效应的定义与分配机制的建模是分开的：
    - 因果效应的定义仅仅与潜在结果有关，与分配机制无关，从而允许处理效应Δ_i的任意的异质性，即不同个体的处理效应可以是不同的。
    - 与此同时，结果变量y的数据生成过程是相对不重要的
        - 甚至很多情况下我们会把$y_i(1)$和$y_i(0)$看做是两个固定的常数而非被可见或者不可见的因素所影响的随机变量
        - 在这种情况下，模型中的随机性完全来自于分配机制w_i的随机性，而不是来自于误差项的随机性

SUTVA假设

- 由于我们严格区分了分配机制和结果变量的数据生成过程，那么隔断w对潜在结果$y(1), y(0)$的影响就非常重要了
  - 如果w对$y(1), y(0)$有影响，我们就无法绕过潜在结果$y(1), y(0)$的数据生成过程。
- 而这一要求即个体处理值稳定假设（**stable unit treatment value assumption，SUTVA，Rubin，1980**）。
- 简单来说，该假设意味着对于$W = (w_1, ..., w_N)'$的分配不会影响$Y^1 = [y_1(1), ..., y_N(1)]'$及$Y^0 = [y_1(0), ..., y_N(0)]'$的取值，即$Y^1, Y^0$对于$W$的变动是不变的。

SUTVA假设

伙伴效应

考虑我们从五个人{☺$_1$, ☺$_2$, ☺$_3$, ☻$_1$, ☻$_2$}中挑选两个进入某个培训项目。这五个人中，☺$_1$, ☺$_2$, ☺$_3$是好朋友而☻$_1$, ☻$_2$是好朋友。由于朋友关系，某个人被选中进入培训项目后，会主动分享培训信息给朋友，从而可能会影响朋友的结果变量。比如下表展示了这样一种情况：如果☺$_3$被选中，他会影响☺$_1$, ☺$_2$的$y(0)$；而☻$_1$和☻$_2$同时被选中，可能由于正向的的伙伴效应，双方的$y(1)$都提高了。由于W的分配影响了Y^1, Y^0，从而SUTVA假设就不满足了。

SUTVA假设

伙伴效应

	W	Y^0	Y^1
☺$_1$	0	1	2
☺$_2$	0	1.2	1.8
☺$_3$	1	0.8	2.1
☻$_1$	1	0.6	1.5
☻$_2$	0	0.8	1.8

(a)

	W	Y^0	Y^1
☺$_1$	0	0.8	2
☺$_2$	0	0.9	1.8
☺$_3$	0	0.8	2.1
☻$_1$	1	0.6	1.9
☻$_2$	1	0.8	2.1

(b)

SUTVA假设

含义：排除了：

1. 一般均衡效应（general equilibrium effects）
2. 伙伴效应（peer effects）
3. ...

如果违背：

1. 重新定义研究个体
2. 直接设定个体之间的交互

分配机制

分配机制：

1. 随机实验：
   1. 处理的分配概率不随着潜在结果而改变
   2. 处理的分配概率为协变量（covariates）的已知函数
2. 无混淆分配（unconfounded assignment）
   - 要求给定协变量，潜在结果与分配独立：

$$w_i \perp\!\!\!\perp (y_i(0), y_i(1)) | x_i$$

   - 与随机实验差别：$P(w_i|x_i)$为未知函数
   - 又称为：
     - selection-on-observable
     - exogeneity
     - conditional independence assumption (CIA)

- 其他
  - selection-on-unobservable
  - 例：Roy Model：$w_i = 1(y_i(1) \geq y_i(0))$

处理效应的定义

不同形式的处理效应：

- 个体处理效应：对于个体i，个体处理效应为$\Delta_i = y_i(1) - y_i(0)$

异质性（heterogeneous effects）：Δ_i随着i的变化而变化。

- 同质处理效应（Homogeneous treatment effects）：$y_i(1) - y_i(0) = \Delta$
  - 例：$y_i = g(x_i) + \alpha w_i + u_i$
- 条件同质处理效应：$\Delta_i = \Delta(X_i)$
  - 例：$y_i = g(x_i, w_i) + u_i \Rightarrow \Delta_i = g(x_i, 1) - g(x_i, 0) = \Delta(x_i)$
  - 例：$y_i = x_i'\beta + w_i x_i'\alpha + u_i$
- 异质处理效应
  - $y_i = g(x_i, w_i, u_i)$

关心的处理效应

感兴趣的处理效应:

1. 平均处理效应（Average Treatment Effects, ATE）： $ATE = \mathbb{E}(\Delta_i)$
2. 处理组平均处理效应（Average Treatment Effects on the Treated, TT）： $ATT = \mathbb{E}(\Delta_i | w_i = 1)$
3. 未处理组平均处理效应（Average Treatment Effects on the Untreated, TUT）： $ATUT = \mathbb{E}(\Delta_i | w_i = 0)$

关心的处理效应

以上定义的是总体处理效应，然而给定样本，我们通常关注给定协变量X_i时的处理效应：

1. $CATE(x_i) = \mathbb{E}(\Delta_i | x_i)$
2. $CATT(x_i) = \mathbb{E}(\Delta_i | x_i, w_i = 1)$
3. $CATUT(x_i) = \frac{1}{N} \sum_{i|w_i=0} \mathbb{E}(\Delta_i | x_i, w_i = 0)$

处理效应的同质性和异质性

几种处理效应关系：

- 同质处理效应：

$$ATE = ATT = ATUT = CATE(x_i) = CATT(x_i) = CATUT(x_i)$$

- 条件同质处理效应：
  - $CATE(x_i) = CATT(x_i) = CATUT(x_i)$
  - 可能$ATE \neq ATT \neq ATUT$，由于x_i的分布随w_i不同
- 异质性处理效应：
  - 如果$y_i(1) - y_i(0) \perp\!\!\!\perp w_i|x_i$：结论同条件同质处理效应
  - 否则$CATE(x_i) \neq CATT(x_i) \neq CATUT(x_i)$

其他处理效应

其他感兴趣的处理效应：

- 分位数处理效应（quantile treament effects）：

$$\tau_q = F_{y(1)}^{-1}(q) - F_{y(0)}^{-1}(q)$$

- 处理效应的分位数：

$$\tilde{\tau}_q = F_{y(1)-y(0)}^{-1}(q)$$

- 最难的参数：$(y_i(1), y_i(0))$的联合分布：

$$P(y_i(1) \leq y_1, y_i(0) \leq y_0)$$

处理效应中的偏误

可能的偏误：由于$y = wy(1) + (1-w)\,y(0) = y(0) + w\,(y(1) - y(0))$，从而：

$$
\begin{array}{rl}
 & \mathbb{E}(y|w=1) - \mathbb{E}(y|w=0) = \\
(ATT) & \mathbb{E}(y_1 - y_0|w=1) + \\
(Selection\ bias) & \mathbb{E}(y_0|w=1) - \mathbb{E}(y_0|w=0)
\end{array}
$$

如果我们关心平均处理效应：

$$
\begin{array}{rl}
 & \mathbb{E}(y|w=1) - \mathbb{E}(y|w=0) = \\
(ATE) & \mathbb{E}(y_1 - y_0) + \\
(Sorting\ Gain) & \mathbb{E}(y_1 - y_0|w=1) - \mathbb{E}(y_1 - y_0) + \\
(Selection\ bias) & \mathbb{E}(y_0|w=1) - \mathbb{E}(y_0|w=0)
\end{array}
$$

偏识别

在最宽松的假设条件下，我们可以放弃点识别（point identification），转而使用偏识别（Partial identification）

- 点识别识别具体的点
- 偏识别仅仅能够给出上下界

偏识别

如果潜在因果$y_i(w_i)$是有界的，比如，$y_i(w_i)=0/1$，由于：

$$\begin{aligned}\mathbb{E}\left(y_i(1)-y_i(0)\right)=&\ \mathbb{E}\left(y_i(1)|w_i=1\right)P\left(w_i=1\right)\\&+\mathbb{E}\left(y_i(1)|w_i=0\right)P\left(w_i=0\right)\\&-\mathbb{E}\left(y_i(0)|w_i=1\right)P\left(w_i=1\right)\\&-\mathbb{E}\left(y_i(0)|w_i=0\right)P\left(w_i=0\right)\end{aligned}$$

因而其下界为：

$$\begin{aligned}\tau_l=&\ \mathbb{E}\left(y_i(1)|w_i=1\right)P\left(w_i=1\right)\\&-P\left(w_i=1\right)-\mathbb{E}\left(y_i(0)|w_i=0\right)P\left(w_i=0\right)\end{aligned}$$

上界为：

$$\begin{aligned}\tau_u=&\ \mathbb{E}\left(y_i(1)|w_i=1\right)P\left(w_i=1\right)\\&+P\left(w_i=0\right)-\mathbb{E}\left(y_i(0)|w_i=0\right)P\left(w_i=0\right)\end{aligned}$$

偏识别

如果我们关心潜在结果的分布，由于:

$$
\begin{aligned}
F_{y_1}(y) &= P(y_i(1) \leq y) \\
&= P(y_i(1) \leq y | w_i = 1) P(w_i = 1) \\
&+ P(y_i(1) \leq y | w_i = 0) P(w_i = 0)
\end{aligned}
$$

因而: $F_{y_1}(y)$的下界为:

$$
P(y_i(1) \leq y | w_i = 1) P(w_i = 1)
$$

而上界为:

$$
P(y_i(1) \leq y | w_i = 1) P(w_i = 1) + P(w_i = 0)
$$

偏识别

其他如:

1. Manski and Pepper(2000)
2. Jun, Lee and Shin (2016)

随机实验与观察研究

根据Cochran（1972），**随机实验**（**randomized experiments**）即分配机制不依赖于个体（包括可观测和不可观测的）特征，而且研究者可以控制处理变量如何分配的设计。

- 在观察研究（observational studies）中，研究者无法控制处理变量的分配。

相对于观察研究，随机实验的优点：

- 通过随机化进行控制，消除选择偏误（selection bias）

内部有效性和外部有效性

- 内部有效性（internal validity）：在研究总体中准确反应因果效应
    - 一个设计良好的实验一般具有内部有效性
- 外部有效性（external validity）：因果效应的泛化（generalization）
    - 在实验和观察研究中，外部有效性都很难保证
    - 外部有效性与处理效应的异质性紧密相关

实验的统计推断：为什么需要单独讨论？

模型中不确定性的来源——两种视角：

- 传统统计学：抽样所带来的误差（抽样误差），然而：
  - 有时我们可以观察到总体
  - 有时总体的界定不是非常清晰
- 另一种视角：由于w_i的随机性导致我们只能观察到$y_i(0), y_i(1)$中的一个
  - 有时可以将手中的样本看做是一个有限总体（内部有效性）

随机试验的设计

- 完全随机化实验（completely randomized experiments）
- 分层（stratified）随机化实验
  - 先将总体分为几个亚组或者层（strata），再在亚组中进行随机分组
  - 比如：在男性和女性两个组别中分别进行随机分组
  - 可能改善有效性
- 整群（clustered）随机化实验
  - 同样将总体分为亚组或者群（cluster），然后随机选择这些亚组整体进入控制组或者处理组
  - 比如：在一个学校中，随机选择班级进入处理组
  - 可以处理溢出效应或者伙伴效应

完全随机化实验的推断

正如我们前面所介绍的，“没有处理效应”有很多种不同的情况，比如：

- 对于所有个体i，处理变量w对y完全没有影响：$y_i(0)=y_i(1)$
- 平均处理效应相同：$\mathbb{E}(y_i(1))=\mathbb{E}(y_i(0))$

注意当我们写下原假设：$H_0: y_i(0)=y_i(1)$时，我们实际上将$y_i(0), y_i(1)$视作固定的，而非随机的，那么$\mathbb{E}(y_i(1))=\mathbb{E}(y_i(0))$的原假设也可以被写成：

$$H_0: \frac{1}{N}\sum_{i=1}^{N}(y_i(1)-y_i(0))=\bar{y}(1)-\bar{y}(0)=0$$

其中N为总体数量。

Fisher's exact p-value

如果我们选取原假设：

$$H_0 : y_i(0) = y_i(1), i = 1, 2, ..., N$$

即完全没有任何处理效应的“清晰原假设”（sharp null hypothesis），可以使用Fisher（1925，1935）的精确p值检验，步骤：

1. 选取一个检验统计量$T(\{w_i, y_i\})$，比如可以选择为两个组别均值的差异：

$$T(\{w_i, y_i\}) = \bar{y}_{w=1} - \bar{y}_{w=0}$$

2. 将w_i进行重新分配，如果有N个样本，其中N_t个处理组和N_c个对照组，那么应该有$\begin{pmatrix} N \\ N_t \end{pmatrix}$种不同的组合：$\{w_i^p\}, p = 1, 2, ..., \begin{pmatrix} N \\ N_t \end{pmatrix}$
3. 针对每种$\{w_i^p\}$的组合，都计算$T(\{w_i^p, y_i\})$，然后计算比例：

$$p = P(|T(\{w_i, y_i\})| \leq |T(\{w_i^p, y_i\})|)$$

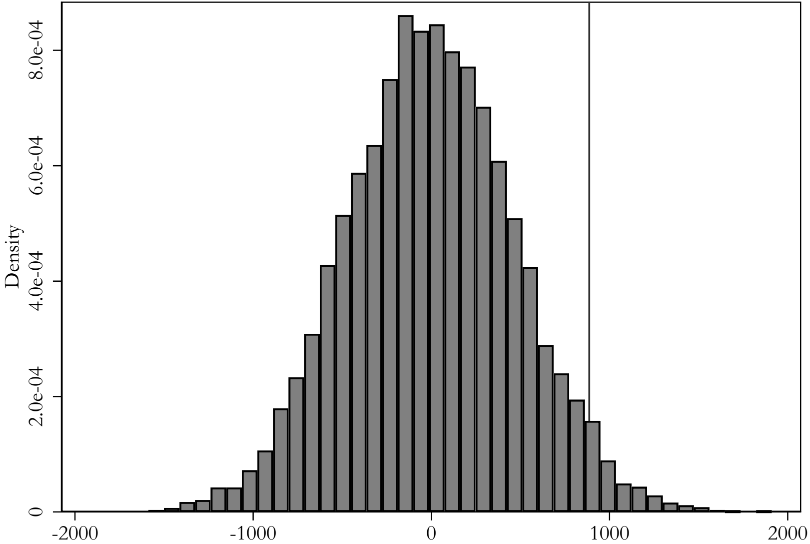
Fisher's exact p-value：原理

- 如果原假设成立，那么w是如何分配的实际上对于两个组之间均值的比较并没有影响：
  - 由于$y_i(0) = y_i(1)$，观察到$y_i(0)$和$y_i(1)$是等价的
- 如果原假设不成立，那么计算得到的p就是原假设成立的条件下，得到比真实数据的检验统计量更极端的概率：
  - 如果原假设成立，有可能会得到更极端的结果，但是这个概率应该很小
  - 如果比$T(\{w_i, y_i\})$更极端的概率很小，意味着在原假设成立的条件下，真实数据已经很极端，从而可以拒绝原假设

Fisher's exact p-value：实施

- 现实中，如果N很大，$\begin{pmatrix} N \\ N_t \end{pmatrix}$会非常非常大
- 解决办法：不需要遍历所有可能性，随机分配M次即可
- fisher_exact_p.do（NSW：National Supported Work）
- 可以拓展到非实验：placebo test

示例



其他的检验统计量

- 除了比较均值之外，还可以比较其他统计量（Imbens和Rubin，2015）
  - 中位数
  - 分位数
  - t统计量
  - 排序
- 其中排序被定义为:

$$R_i = \sum_{j=1}^{N} \mathbb{1}\{y_j < y_i\} + \frac{1}{2}\left(1 + \sum_{j=1}^{N} \mathbb{1}\{y_j = y_i\}\right) - \frac{N-1}{2}$$

gen_rank.ado

平均处理效应的推断

- 如果原假设改成平均处理效应为0：

$$H_0 : \tau = \frac{1}{N}\sum_{i=1}^{N}\left(y_i\left(1\right) - y_i\left(0\right)\right) = \bar{y}\left(1\right) - \bar{y}\left(0\right) = 0$$

注意N为总体数量，$Y_i\left(1\right), Y_i\left(0\right)$被视作常数。

- 自然的点估计量：

$$\hat{\tau} = \bar{Y}_{w=1} - \bar{Y}_{w=0} = \frac{1}{N_t}\sum_{i:w_i=1} y_i - \frac{1}{N_c}\sum_{i:w_i=0} y_i$$

然而：标准误与以往的重复抽样思想下的标准误有所区别。

平均处理效应的标准误

- Imbens和Rubin（2015）计算得到：

$$\mathbb{V}\left(\hat{\tau}\right)=\frac{s_t^2}{N_t}+\frac{s_c^2}{N_c}-\frac{s_{tc}^2}{N}$$

  - 其中：

$$s_{tc}^2=\frac{1}{N-1}\sum_{i=1}^{N}\left[\left(y_i\left(1\right)-y_i\left(0\right)\right)-\left(\bar{y}\left(1\right)-\bar{y}\left(0\right)\right)\right]^2$$

  然而由于反事实观察不到，该项无法计算出

  - 传统的计算方法：

$$\mathbb{V}\left(\hat{\tau}\right)=\frac{s_t^2}{N_t}+\frac{s_c^2}{N_c}$$

  比较两者，传统的t统计量：

$$\frac{\hat{\tau}}{\sqrt{\frac{s_t^2}{N_t}+\frac{s_c^2}{N_c}}}$$

  （绝对值）更大。

实际使用

- 对于同质性处理效应：使用传统t统计量即可
- 如果样本来自于一个无限总体：使用传统t统计量
- 即使两者都不满足：t统计量也更稳健
- Stata：

```
ttest re78, by(treat )
```

或者使用回归：

```
reg re78 treat, r
```

比较协变量的分布

- 在实验中，一个常见的操作是比较处理组和控制组的协变量
  - 检查随机化
  - 观察两个组别之间的细微差别（可能完全是因为随机原因导致的）
- 方法：仍然是进行t检验或者Fisher's exact p-value等，只不过比较协变量而非Y
- cavatiates_balancing.do （fishers_p.ado）

使用回归分析实验

由于:

$$y_i = w_i y_i(1) + (1 - w_i) y_i(0) = y_i(0) + w_i (y_i(1) - y_i(0))$$

考虑：$\tau = \mathbb{E}(y_i(1) - y_i(0))$为平均处理效应，$\alpha = \mathbb{E}(y_i(0))$，那么：

$$\begin{aligned} y_i &= \mathbb{E}(y_i(0)) + w_i \mathbb{E}(y_i(1) - y_i(0)) + \zeta_i + \eta_i w_i \\ &= \alpha + \tau \cdot w_i + \epsilon_i \end{aligned}$$

其中

- $\zeta_i = y_i(0) - \mathbb{E}(y_i(0))$：如果与$w_i$相关：selection bias
- $\eta_i = y_i(1) - y_i(0) - \mathbb{E}(y_i(1) - y_i(0))$：如果与$w_i$相关：sorting gain

如果w_i是完全随机分配的：

$$\mathbb{E}(\epsilon_i | w_i) = \mathbb{E}(\zeta_i | w_i) + \mathbb{E}(\eta_i w_i | w_i) = \mathbb{E}(\zeta_i | w_i) + w_i \mathbb{E}(\eta_i | w_i) = 0$$

实验中的协变量

- 协变量最好的处理方法：在实验设计时，就通过分层随机化等方法进行控制
- 如果实行了完全随机化，在比较时考虑协变量也是有好处的：
  - （可能）增加power
  - （可能）消除bias
  - 可以建模异质性

```
reg re78 treat age education black hispanic married nodegree re75 , r
```

使用交叉项建模异质性

如果有协变量，我们可以先对协变量去平均：

$$\dot{x}_i = x_i - \bar{x}$$

然后使用回归：

$$y_i = \alpha + \tau \cdot w_i + \dot{x}_i' \beta + w_i \dot{x}_i' \delta + \epsilon_i$$

按照如此设定，τ为估计的平均处理效应，而δ则建模了处理效应异质性。（experiment_reg_hetero.do）

匹配的假设

在Unconfoundedness假设条件下，可以方便的得到处理效应的识别。关键假设：

1. Unconfoundedness假设（CIA假设）：

$$w_i \perp\!\!\!\perp (y_i(1), y_i(0)) | x$$

2. 共同支撑假设（Common support assumption, CSA）：

$$0 < P(w_i = 1 | x_i) < 1$$

匹配的原理

CIA意味着均值独立，即：

$$\mathbb{E}\left(y_i\left(1\right)|x_i, w_i\right) = \mathbb{E}\left(y_i\left(1\right)|x_i\right)$$

$$\mathbb{E}\left(y_i\left(0\right)|x_i, w_i\right) = \mathbb{E}\left(y_i\left(0\right)|x_i\right)$$

而CSA需要对于相同的x_i，都有处理组和非处理组：经常需要trimming。

匹配的原理

由于:

$$
\begin{aligned}
\mathbb{E}\left(y_i|w_i=1,x_i\right)-\mathbb{E}\left(y_i|w_i=0,x_i\right)
&=\mathbb{E}\left(y_i\left(1\right)|x_i,w_i=1\right)-\mathbb{E}\left(y_i\left(0\right)|x_i,w_i=0\right)\\
&=\mathbb{E}\left(y_i\left(1\right)|x_i\right)-\mathbb{E}\left(y_i\left(0\right)|x_i\right)\\
&=\mathbb{E}\left(y_i\left(1\right)-y_i\left(0\right)|x_i\right)
\end{aligned}
$$

因而

$$
\mathbb{E}\left[\mathbb{E}\left(y_i|w_i=1,x_i\right)-\mathbb{E}\left(y_i|w_i=0,x_i\right)\right]=\mathbb{E}\left(y_i\left(1\right)-y_i\left(0\right)\right)
$$

同理:

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left(y_i|w_i=1,x_i\right)-\mathbb{E}\left(y_i|w_i=0,x_i\right)|w_i=1\right]=&\\
\mathbb{E}\left(y_i\left(1\right)-y_i\left(0\right)|w_i=1\right)&
\end{aligned}
$$

无混淆分配下的推断：回归方法

方法一：回归

设定

$$\mathbb{E}\left(y_i|w_i=1,x_i\right)=\mathbb{E}\left(y_i\left(1\right)|x_i\right)=\mu_1\left(x_i\right)$$

$$\mathbb{E}\left(y_i|w_i=0,x_i\right)=\mathbb{E}\left(y_i\left(0\right)|x_i\right)=\mu_0\left(x_i\right)$$

平均处理效应：

$$\hat{\tau}_{reg}=\frac{1}{N}\sum_{i=1}^{N}\left[\hat{\mu}_1\left(x_i\right)-\hat{\mu}_0\left(x_i\right)\right]$$

1. 简单线性回归：

$$y_i=\alpha+x_i'\beta+\tau\cdot w_i+\epsilon_i$$

或者：

$$y_i=\alpha+\dot{x}_i'\beta+\tau\cdot w_i+w_i\dot{x}_i'\delta+\epsilon_i$$

2. 非参数回归

无混淆分配下的推断：匹配

或者：

- Nearest Neighbor Matching：
  1. 给定一个正的常数M，比如$M=1$
  2. 令$d(\cdot,\cdot)$为一个距离函数，比如欧几里得距离：

$$d(x_i, x_j) = (x_i - x_j)'(x_i - x_j)$$

或者Mahalanobis距离：

$$d(x_i, x_j) = (x_i - x_j)'\Sigma_x^{-1}(x_i - x_j)$$

- caliper：距离小于一个临界值（caliper）即匹配成功

临近匹配

Nearest-neighbor matching: 对于任意处理组的i，从控制组中找到最近的M个控制组个体，记:

$$J_M(i) = \{l_1(i), ..., l_M(i)\}$$

定义

$$\hat{y}_i(0) = \frac{1}{M} \sum_{m \in J_M(i)} y_m$$

可以使用:

$$\frac{1}{N_1} \sum_{i|w_i=1} [y_i(1) - \hat{y}_i(0)]$$

匹配的细节

实践中，有不同的匹配方案：

1. 选择M，一般而言如果控制组数量远远大于实验组数量，可以使用较多的M
2. 序贯/非序贯
   - 序贯：按顺序一个一个配对
   - 非序贯：所有的实验组和控制组放在一起考虑
3. 贪婪/非贪婪
   - 贪婪：有放回
   - 非贪婪：无放回
4. 先进行分组或者分层（stratification），组内进行匹配
5. 使用propensity score进行排序，进而匹配
6. Stata:
   - psmatch2
   - teffects nnmatch（推荐）

匹配的步骤

倾向得分匹配的步骤：

1. 定义距离
2. 匹配
3. 评估匹配结果：
   1. balancing: t-test、Standardised Bias
   2. unconfoundedness：使用明显无效应的其他的Y
4. 评估政策效应
5. 敏感性分析

评估匹配结果：协变量

评估匹配结果：

1. normalized difference:

$$\hat{\Delta}_{ct} = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{s_c^2 + s_t^2}{2}}}$$

2. t-stat:

$$T_{ct} = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}}}$$

Stata: tebalance

Balancing

Table 16: Summary Statistics for Non-experimental Lalonde Data

Covariate	CPS controls (N_c=15,992) mean	(s.d.)	trainees (N_t=185) mean	(s.d.)	t-stat	nor-dif
Black	0.07	0.26	0.84	0.36	28.6	2.43
Hisp	0.07	0.26	0.06	0.24	-0.7	-0.05
Age	33.23	11.05	25.82	7.16	-13.9	-0.80
Married	0.71	0.45	0.19	0.39	-18.0	-1.23
Nodegree	0.30	0.46	0.71	0.46	12.2	0.90
Education	12.03	2.87	10.35	2.01	-11.2	-0.68
E'74	14.02	9.57	2.10	4.89	-32.5	-1.57
U'74	0.12	0.32	0.71	0.46	17.5	1.49
E'75	13.65	9.27	1.53	3.22	-48.9	-1.75
U'75	0.11	0.31	0.60	0.49	13.6	1.19

	Full Sample nor-dif	Matched Sample nor-dif	ratio of nor-dif
Black	2.43	0.00	0.00
Hispanic	-0.05	0.00	-0.00
Age	-0.80	-0.15	0.19
Married	-1.23	-0.28	0.22
Nodegree	0.90	0.25	0.28
Education	-0.68	-0.18	0.26
E'74	-1.57	-0.03	0.02
U'74	1.49	0.02	0.02
E'75	-1.75	-0.07	0.04

Matching

其他匹配方法:

1. Kernel matching
2. Radius matching
3. Stratification or interval matching
4. Propensity score matching

倾向得分匹配

倾向得分（Propensity score）方法：Rosenbaum and Rubin(1983)证明：

$$w_i \perp\!\!\!\perp \left(y_i\left(1\right), y_i\left(0\right)\right)|x_i \Longleftrightarrow w_i \perp\!\!\!\perp \left(y_i\left(1\right), y_i\left(0\right)\right)|P\left(x_i\right)$$

因而控制倾向得分匹配就足够了。

倾向得分匹配

Rosenbaum and Rubin(1983)提出了三阶段的方法：

1. 估计倾向得分：
$$P\left(w_i|x_i\right)$$
2. 用y_i对w_i和p_i做回归，得到$\hat{\mathbb{E}}\left(y_i|w_i,p_i\right)$
3. 估计ATT：
$$\frac{1}{N_1}\sum_{i|w_i=1}\left[y_i\left(1\right)-\hat{y}_i\left(0\right)\right]$$

注意在使用Propensity Score时，一定要注意Common support假设：trimming！

逆概率加权法

注意到，由于：

$$
\begin{aligned}
\mathbb{E}\left(\frac{w_i y_i}{P(x_i)}\right) &= \mathbb{E}\left(\frac{w_i y_i(1)}{P(x_i)}\right) \\
&= \mathbb{E}\left[\mathbb{E}\left(\frac{w_i y_i(1)}{P(x_i)}|x_i\right)\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}(w_i y_i(1)|x_i)}{P(x_i)}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}(w_i|x_i)\,\mathbb{E}(y_i(1)|x_i)}{P(x_i)}\right] \\
&= \mathbb{E}\left[\mathbb{E}(y_i(1)|x_i)\right] = \mathbb{E}(y_i(1))
\end{aligned}
$$

同理：

$$
\mathbb{E}\left(\frac{(1-w_i)\,y_i}{1-P(x_i)}\right) = \mathbb{E}(y_i(0))
$$

逆概率加权法

因而平均处理效应：

$$\tau_{ATE} = \mathbb{E}\left[\frac{w_i y_i}{P(x_i)} - \frac{(1-w_i)\, y_i}{1-P(x_i)}\right]$$

可以使用：

$$\hat{\tau}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{w_i y_i}{P(x_i)} - \frac{(1-w_i)\, y_i}{1-P(x_i)}\right]$$

进行估计，称为Inverse Propensity Weighting(IPW)。其中$P(x_i)$可以使用Logistic sieve估计量（Hirano, Imbense and Ridder, 2003）。
Stata: teffects ipw

双向稳健的处理效应评估方法

然而IPW方法对倾向得分非常敏感。可以考虑使用Robins等人提出的双向稳健（Double robustness）方法：

1. 结合了回归方法和IPW
2. 只需要$P(x_i)$或者结果方程至少有一个设定正确（双向稳健）

最小化：

$$\min_{\alpha_0,\beta_0}\sum_{i|w_i=0}\frac{\left[y_i-\alpha_0-\beta_0'\left(x_i-\bar{x}_i\right)\right]^2}{1-P\left(x_i;\hat{\gamma}\right)}$$

$$\min_{\alpha_1,\beta_1}\sum_{i|w_i=1}\frac{\left[y_i-\alpha_1-\beta_1'\left(x_i-\bar{x}_i\right)\right]^2}{P\left(x_i;\hat{\gamma}\right)}$$

平均处理效应为：

$$\hat{\tau}_{ATE}=\hat{\alpha}_1-\hat{\alpha}_0$$

Stata: teffects aipw

双向稳健的处理效应评估方法

- 不妨考虑一阶条件（对α_1求导）：

$$-2\sum \frac{w_i\left[y_i-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)\right]}{P\left(x_i;\hat{\gamma}\right)}=0$$

- 考虑其总体形式

$$\begin{aligned}
\mathbb{E}\left(\frac{w_i\left[y_i-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)\right]}{P\left(x_i;\hat{\gamma}\right)}\right) &= \mathbb{E}\left[\mathbb{E}\left(\frac{w_i\left[y_i-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)\right]}{P\left(x_i;\hat{\gamma}\right)}\middle|x_i\right)\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left(w_i\left[y_i-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)\right]|x_i\right)}{P\left(x_i;\hat{\gamma}\right)}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left(w_i\left[y_i(1)-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)\right]|x_i\right)}{P\left(x_i;\hat{\gamma}\right)}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left(w_i|x_i\right)\mathbb{E}\left(y_i(1)-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)|x_i\right)}{P\left(x_i;\hat{\gamma}\right)}\right]
\end{aligned}$$

双向稳健的处理效应评估方法

- 总体一阶条件：

$$\mathbb{E}\left[\frac{\mathbb{E}\left(w_i|x_i\right)\mathbb{E}\left(y_i\left(1\right)-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)|x_i\right)}{P\left(x_i;\hat{\gamma}\right)}\right]=0$$

的两种情况：

  - $P\left(x_i;\hat{\gamma}\right)$正确设定：那么$\alpha_1=\mathbb{E}\left[\mathbb{E}\left(y_i\left(1\right)\right)\right]=\mathbb{E}\left(y_i\left(1\right)\right)$
  - $y_i\left(1\right)$函数形式设定正确：$\mathbb{E}\left(y_i\left(1\right)-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)|x_i\right)=0$成立，$\alpha_1$得到一致估计。

Matching

其他方法：Imai and Ratkovic (2014)：解：

$$\frac{1}{N}\sum_{i=1}^{N} g_\gamma\left(w_i, x_i\right) = \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{w_i}{P\left(x_i;\gamma\right)} - \frac{1-w_i}{1-P\left(x_i;\gamma\right)}\right) f\left(x_i\right)\right] = 0$$

仍然是双向稳健的。此外，Fan et al. (2016)做了推广。