治理技术专题

定量政治分析方法

Quantitative Analysis II

苏 毓 淞

清华大学社会科学院政治学系

第十讲　倾向值匹配法

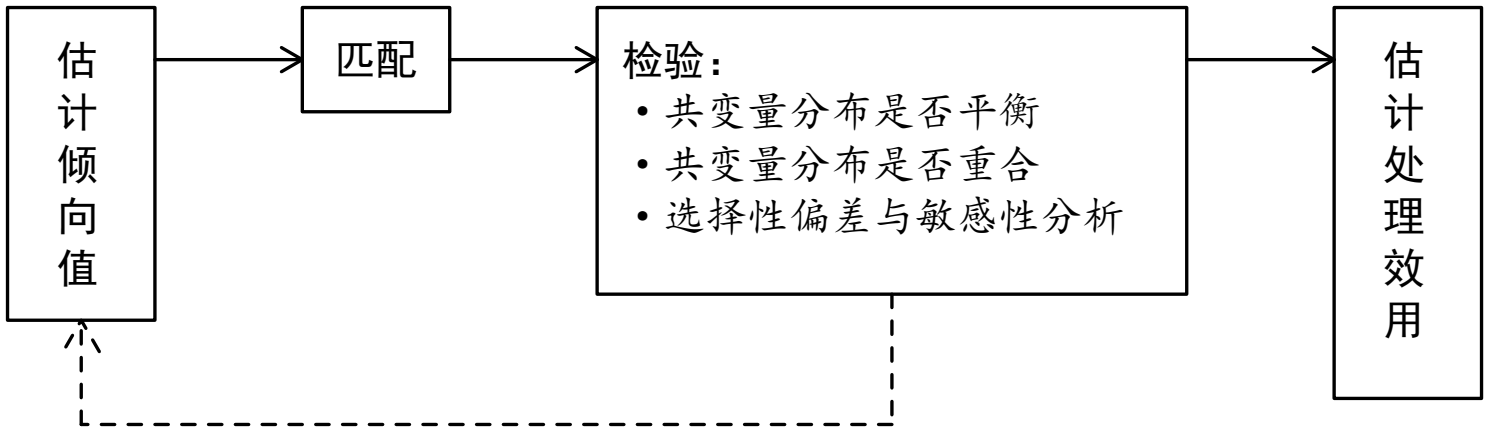
上周内容概述 . . .

- 因果推论方法的基本原理；使用一种可以**平衡 (组间) 数据**的方法，在进行数据分析前，预先将数据整理后，使实验组和对照组的样本符合因果推论的基本要求——两组之间除了处理变量外，其他变量间的差异是平衡的、无差异的。
- “倾向值匹配法”就是众多平衡数据方法中的一种。

倾向值匹配法

- 倾向值匹配法可以分为“倾向值”和“匹配法”两部分来讨论；
  - 匹配法：针对实验组样本和对照组样本进行匹配，找出相似的配对，它们唯一的差别就是一个有受到实验处理，另一个则没有，比较这两个样本结果的差异，即处理效用（因果效用）。
  - 倾向值 (propensity score)：
    - 某一个样本接受到处理的概率。
    - 通常会使用各类回归模型来求解倾向值，而这个回归模型所使用的共变量，也就是我们用来匹配样本的变量，倾向值就是把多维的精确匹配简化成一维的倾向值匹配，匹配的对象和过程得到简化，也更容易找到相似的配对。

倾向值匹配法分析步骤



倾向值匹配法

- 倾向值 $e(\mathbf{X}_i)$ 是某一个样本 (i) 接受到处理的概率，假定这个处理是个二元变量，倾向值是一个接受或不接受处理的指标变量 (indicator variable)，

$$e(\mathbf{X}_i) = \Pr(Z_i = 1 | \mathbf{X}_i)$$

- 当控制接受处理前所有可以观察到的共变量 $\mathbf{X}_i$ 的条件下，单元 i 接受处理 $(Z_i = 1)$ 的概率。

倾向值的性质

- 倾向值能平衡实验组和对照组之间的差异
- 在控制倾向值的情况下，共变量是独立于处理分配的

$$\mathbf{X} \perp\!\!\!\perp Z \mid e(\mathbf{X})$$

- 倾向值所对应各组结果变量的期望值的差值（均值的差），等于其所对应各组间结果变量差值的期望值（差的均值），即平均处理效用 (ATE)。

$$\begin{aligned}
&= \mathsf{E}(Y^1 \mid e(\mathbf{X}), Z = 1) - \mathsf{E}(Y^0 \mid e(\mathbf{X}), Z = 0) \\
&= \underbrace{\mathsf{E}(Y^1 \mid e(\mathbf{X})) - \mathsf{E}(Y^0 \mid e(\mathbf{X}))}_{\text{倾向值所对应各组结果变量的期望值的差值}} \\
&= \underbrace{\mathsf{E}(Y^1 - Y^0 \mid e(\mathbf{X}))}_{\text{倾向值所对应各组结果变量差值的期望值}} \\
&= \mathsf{E}(\tau \mid e(\mathbf{X})) = \widehat{\mathsf{ATE}} \mid e(\mathbf{X})
\end{aligned}$$

估计倾向值

- 数学表达式：

$$e(\mathbf{X}_i) = \Phi_\psi(Z_i|\mathbf{X}_i) = \Phi_\psi(\mathbf{X}_i'\beta)$$

- 使用参数回归模型 (parameterized regression models) 来求解倾向值。
- 非参数模型例如分类与回归树 (Classification and Regression Tree) 分析法、随机森林 (Random Forest) 分析法和贝叶斯累加回归树 (Bayesian Additive Regression Trees, BART) 分析法。

估计倾向值

- 如果 Z_i 是个二元型变量，

$$\Phi_\psi(Z_i|\mathbf{X}_i) = \text{logit}^{-1}\left(\mathbf{X}_i'\boldsymbol{\beta}\right)$$

- 如果 Z_i 是个连续型变量，

$$\Phi_\psi(Z_i|\mathbf{X}_i) \sim \mathcal{N}\left(\mathbf{X}_i'\boldsymbol{\beta}, \hat{\sigma}^2\right)$$

- 如果 Z_i 是个定序型变量，

$$\Phi_\psi(Z_i|\mathbf{X}_i) = \text{logit}^{-1}\left(\mathbf{X}_{i,k}'\boldsymbol{\beta} - c_k\right) - \text{logit}^{-1}\left(\mathbf{X}_{i,k-1}'\boldsymbol{\beta} - c_{k-1}\right)$$

- 如果 Z_i 是类别型变量，

$$\Phi_\psi(Z_i|\mathbf{X}_i) \sim \text{Multinomial}\left(\frac{\exp\left(\mathbf{X}_i'\boldsymbol{\beta}\right)}{\sum \exp\left(\mathbf{X}_i'\boldsymbol{\beta}\right)}, 1\right)$$

估计倾向值

- 非二元处理的倾向值，我们通常使用平衡值（balance score）来进行操作。
- 倾向值也属于平衡值的一种。
- 仔细观察上一页的公式，平衡值即 $\mathbf{X}_i'\beta$（共变量乘上回归系数）。

匹配方法主要的不同之处体现为以下四点

1. 使用相同倾向值匹配或使用相似倾向值匹配?
2. 使用相似倾向值匹配时，如何计算样本间的距离（相似程度)?
3. 选择一对一或选择一对多匹配?
4. 如何调整重复被匹配样本的权重?

匹配后估计处理效用公式

不论使用何种匹配方法，我们都可以使用以下公式来表达估计得的匹配后实验组的处理效用 (treatment effect on the treated, TT)：

$$\hat{\tau}_{TT,\text{matched}} = \frac{1}{n^1} \sum_i \left[(y_i | Z_i = 1) - \sum_j \omega_{i,j} (y_j | Z_j = 0) \right]$$

- n^1 代表实验组的样本量
- i 和 j 分别代表实验组和对照样本序的指标，
- $\omega_{i,j}$ 则调整重复被匹配对照组样本的权重；$\omega_{i,j}$ 在不同的方法中有不同的设定。
- 一般来说，权重只用来调整对照组的样本，因为我们只选择丢弃对照组样本，而不丢弃实验组样本，WHY?

匹配后估计处理效用公式

$$
\begin{aligned}
\text{ATT} &= \mathsf{E}(Y^1 - Y^0 | Z = 1) \\
&= \mathsf{E}(Y^1 | Z = 1) - \underbrace{\mathsf{E}(Y^0 | Z = 1)}_{\text{无法观测到的情况}} \\
&= \mathsf{E}(Y^1 | Z = 1, e(\mathbf{X})) - \underbrace{\mathsf{E}(Y^0 | Z = 0, e(\mathbf{X}))}_{\text{匹配的对照组样本}}
\end{aligned}
$$

精确匹配法 (Exact Matching)

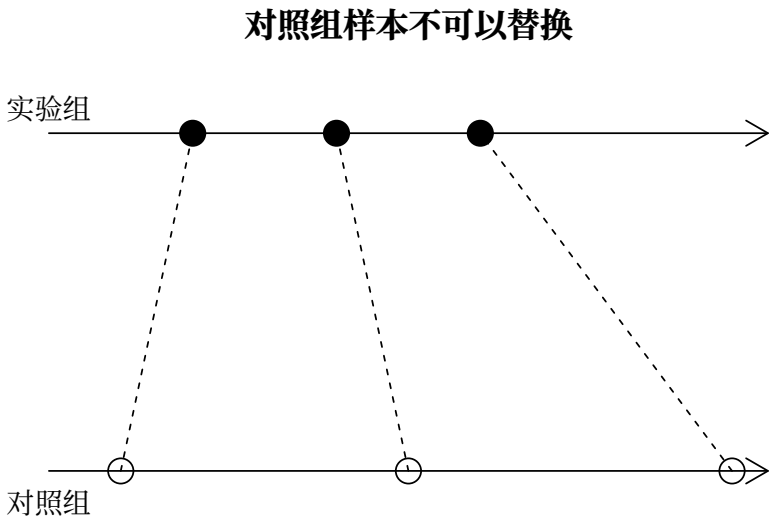
- 精确匹配法 (Exact Matching) 是最基本的匹配法，基本原理就是在实验组和对照组之间，寻找相同的倾向值进行匹配。
- 精确匹配法寻找的是相同的倾向值，所以它无法解决没有相同倾向值的情况；
- 如果对照组的样本找不到匹配，可以选择丢弃该样本，或者将它的权重给定为 0；
- 如果实验组的样本找不到匹配，也可以选择丢弃该样本。
  - 丢弃样本是个昂贵的选择，因为对某样本施予处理，就涉及到额外的成本，当涉及丢弃样本时，除了要考虑样本获得不易，还必须面临由此产生的另一个后果，
  - 即估算出的处理效用低效率的问题。

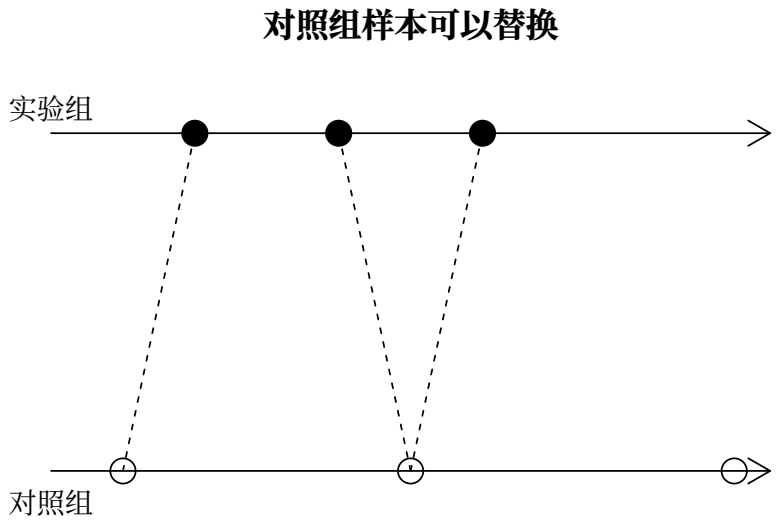
最近邻匹配法 (Nearest-Neighbor Matching)

- 解决了精确匹配法找不到相同倾向值的问题。
- 基本原理是在实验组和对照组间，寻找“相似的”倾向值进行匹配。
  - 将实验组和对照组的样本依照倾向值大小排序，
  - 如果对照组样本可以替换重复使用，则实验组样本匹配倾向值最近似它的对照组样本；
  - 如果对照组样本不能替换重复使用，则须确保每个实验组样本依序匹配对照组的样本，每一个对照组样本只能匹配一个实验组样本。
  - 样本不可重复使用的方法，容易造成匹配度不高的情况。

最近邻匹配法 (Nearest-Neighbor Matching)

样本重复使用与不可重复使用的差别，都有可能制造偏差





最近邻匹配法 (Nearest-Neighbor Matching)

- 为了避免上述匹配不佳的情况发生，**卡尺匹配法** (caliper matching) 限制了实验组和对照组倾向值的最大可容忍差距，超过这个差距的匹配，应选择放弃，
  - 如果发生这种情况，我们陈述因果推论时，就必须如实报告，推论仅仅适用于数据的子集 (subset)。
- **半径匹配法** (radius matching) 即属于卡尺匹配法的一种，

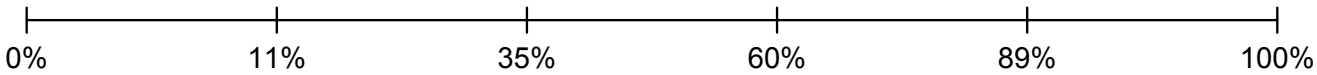
最近邻匹配法 (Nearest-Neighbor Matching)

- **半径匹配法** (radius matching)
  - 任何对照组样本与某一实验组样本的倾向值的绝对差值小于设定的半径大小 r，即 $\mid \hat{e}(\mathbf{X})_j - \hat{e}(\mathbf{X})_i \mid < r$，那么这些对照组样本都会被选择来匹配实验组样本；大于 r 的则选择丢弃，
  - 关于半径大小的设定，目前市面上的软件应用这个方法时，其预设值普遍为 $r = 0.01$，
  - Rosenbaum 建议半径可以取值为 $\frac{1}{4}$ 倾向值的标准差，$r = \frac{1}{4}\sigma_{e(\mathbf{X})}$，其中 $\sigma_{e(\mathbf{X})} = \sqrt{(\sigma_1^2 + \sigma_0^2)/2}$，$\sigma_1^2, \sigma_0^2$ 分别为实验组和对照组样本倾向值的方差。
  - 与卡尺匹配法不同的是，在半径匹配法中，假定实验组样本找不到任何可以匹配的对照组样本（绝对差值大于 r），则选择最邻近的对照组样本进行匹配。

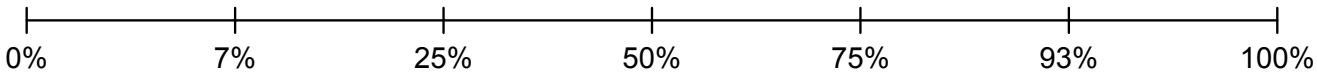
区间匹配法 (Interval Matching)

- 又称子分类匹配法 (subclassification matching)、分层匹配法 (stratification matching)。
- 首先使用分位数法将估计得的倾向值分层或分类据此将数据分成几个区间，然后在各自所属的区间分别进行匹配。

五分位数建议分位点



六分位数建议分位点



区间匹配法 (Interval Matching)

- 每个区间内的样本给予相同的权重 $\omega_{i,j} = 1$，最后将每个区间估计得的处理效用加权平均后，得到总的处理效用，权重是使用各区间内实验组的样本数估算得到的

$$\hat{\tau}_{\text{interval}} = \sum_s \frac{n_s^1 \tau_s}{n_s^1}$$

$$\text{sd}\left(\hat{\tau}_{\text{interval}}\right) = \sqrt{\sum_s \frac{\left(n_s^1 \text{sd}(\tau_s)\right)^2}{\left(n_s^1\right)^2}}$$

- 其中，s 是各区间的指标，n_s^1 是各区间实验组的样本数，τ_s 是各区间估计得的处理效用。
- 在极端情况下，区间匹配法等同于最近邻匹配法。每个区间仅存在一个实验组样本。

核匹配法 (Kernel Matching)

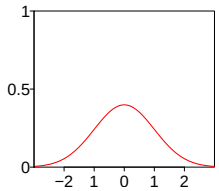
- 使用权重 $\omega_{i,j}$ 调整所有的对照组样本后，将它们匹配每一个实验组样本，权重 $\omega_{i,j}$ 是透过核函数 $\kappa(\cdot)$ 转换计算对照组样本相对于实验组的样本的距离 (倾向值的差值) 所得：

$$\omega_{ij} = \frac{\kappa\left[\frac{\theta_{ij}}{h_n}\right]}{\sum_j \kappa\left[\frac{\theta_{ij}}{h_n}\right]}, \quad \theta_{ij} = \hat{e}(\mathbf{X})_j - \hat{e}(\mathbf{X})_i$$

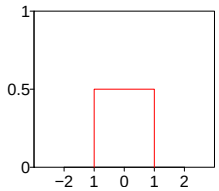
- i, j 分别为实验组和对照组的指标，
- θ_{ij} 是对照组样本倾向值与某一个实验组倾向值的差值 $\hat{e}(\mathbf{X})_j - \hat{e}(\mathbf{X})_i$
- 由于倾向值的差值介于 ± 1 之间，所以中心为 0，分布在 $[-1, 1]$ 之间合适的核函数 $\kappa(\cdot)$ 有 Normal、Uniform、Epanechnikov、Biweight、和 Tricube 几种。
- h 是带宽参数，介于 0 和 1 之间的数。

核匹配法 (Kernel Matching)

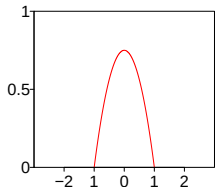
Normal $\kappa(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$



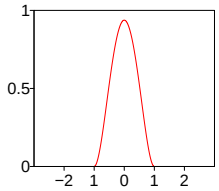
Uniform $\kappa(u) = \frac{1}{2}, |u| < 1$



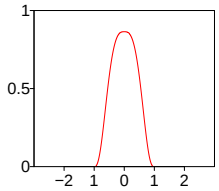
Epanechnikov $\kappa(u) = \frac{3}{4}(1-u^2), |u| < 1$



Biweight $\kappa(u) = \frac{15}{16}(1-u^2)^2, |u| < 1$



Tricube $\kappa(u) = \frac{70}{81}(1-u^3)^3, |u| < 1$



马氏距离 (Mahalanobis distance) 匹配法

- 马氏距离 $\text{md}(\mathbf{X}_i\mathbf{X}_j) = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)'\Sigma^{-1}(\mathbf{X}_i - \mathbf{X}_j)}$
- $\text{md}(\mathbf{X}_i, \mathbf{X}_j)$ 表示任意两个单元 i 和 j 之间的马氏距离，$\mathbf{X}_i, \mathbf{X}_j$ 分别表示单元 i 和 j 的共变量矩阵，Σ 是单元 i 和 j 共变量矩阵的协方差矩阵 (covariance matrix)。
- 使用共变量计算实验组样本 j 与所有对照组样本 j 的距离；之后，从中挑选马氏距离 $\text{md}(\mathbf{X}_i, \mathbf{X}_j)$ 最小的对照组样本进行匹配，
- 然后进行下一个实验组样本匹配对照组样本的工作，直到所有实验组样本都有相匹配的对照组样本为止。
- 最大的问题是随着共变量数量的增加，两个单元间的平均马氏距离也会随之增加，因此增加了找寻匹配的难度，与使用共变量进行精确匹配所面临的难题如出一辙，
- 解决这个问题的办法之一，就是转而使用倾向值匹配法。

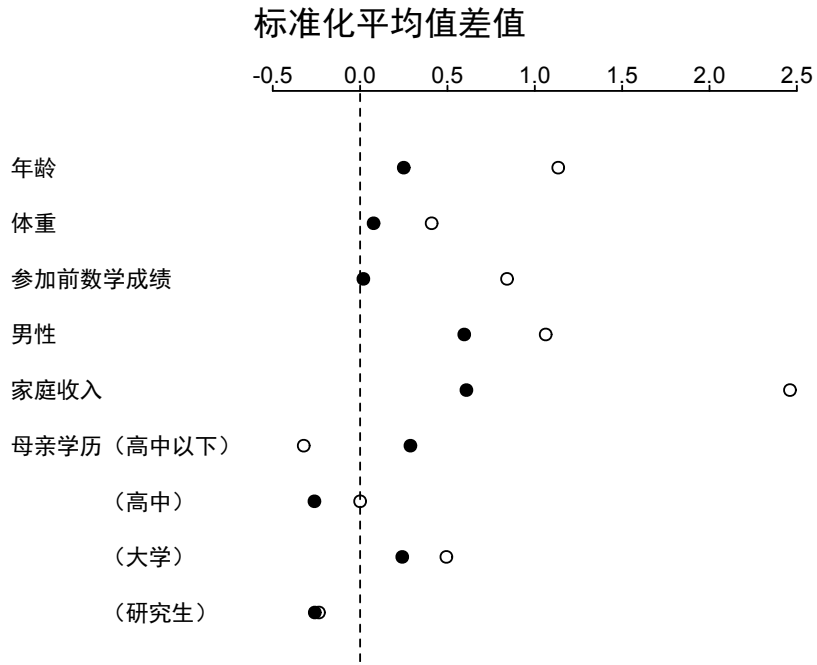
贪婪匹配法与最佳匹配法

- 在对照组样本不重复替换使用的情况下，这些方法也被称为贪婪匹配法 (greedy matching)。
- 贪婪匹配法主要依靠倾向值来估计实验组样本与对照组样本间的相似程度（距离），然后寻找与实验组样本最相近的对照组样本进行匹配。
- 由于它总是试图为实验组样本寻找最相近的对照组样本进行匹配，因此被命名为“贪婪”匹配法。
- 为了改进贪婪匹配法造成可能的偏差，最佳匹配法 (optimal matching) 便应运而生。
- 简单来说，最佳匹配法考量的是最小化所有匹配间距离的总和，因此对于每个实验组样本来说，它们匹配的不可能总是最近邻的样本，有的样本必须匹配第二乃至第三近邻的样本，如此一来，对于所有样本来说，总距离才有可能是最小的。

匹配后检验

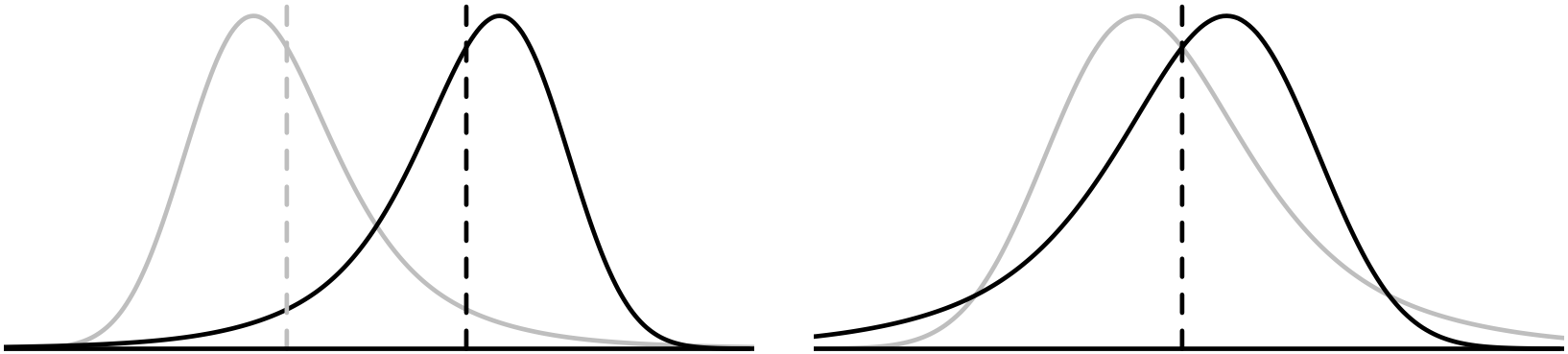
- 共变量分布不平衡分析
- 共变量分布不重合分析
- 选择性偏差与敏感性分析

共变量分布不平衡分析



- 空心点代表匹配前各个共变量的标准化的组间均值的差值，实心点则代表匹配后各个共变量的标准化的组间均值的差值。
- 整体来看，实心点大多分布在 0 点附近，说明匹配后共变量分布较匹配前更为平衡，其标准化的组间均值的差值更趋近于 0。

共变量分布不平衡分析



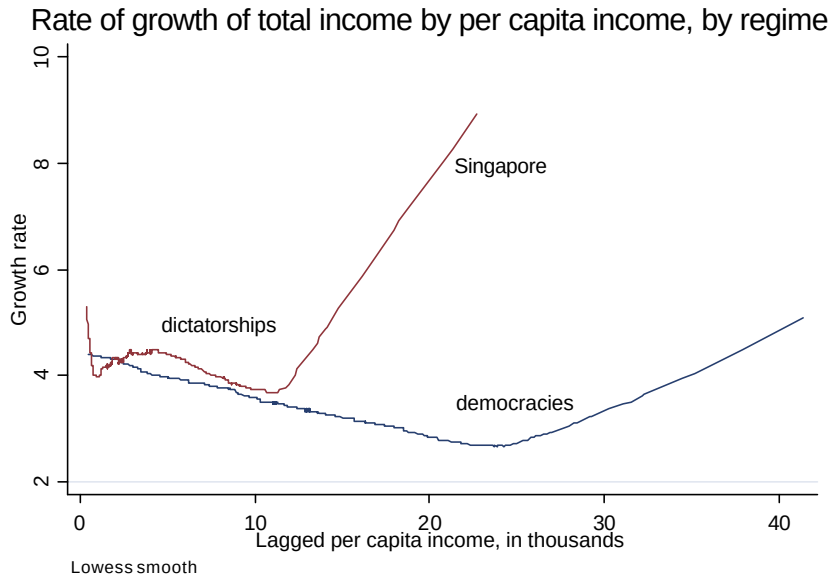
共变量X　　共变量X

- 黑色曲线代表共变量 X 在实验组的分布，灰色曲线则代表共变量 X 在对照组的分布，垂直虚线表示该分布的平均值。
- 左图呈现出共变量 X 在实验组和对照组的分布缺乏平衡，虽然分布覆盖的区域一样，但是分布的平均值和密度函数却不同；
- 右图所显示的是另一种缺乏平衡的情况，虽然分布的均值和覆盖区域相同，但是分布的密度函数却是不同的。

共变量分布不平衡分析

- t 检验法：比较均值和标准差
- KS（Kolmogorov-Smirnov）检验法：比较累计分布函数 (Cumulative Density Function, CDF)

共变量分布不重合分析



- 由于非民主国家中，缺乏人均 GDP 超过 $2,5000$ 美元的国家，所以与民主国家在人均 GDP 这个变量分布上缺乏重合。
- 图中使用 lowess 平滑曲线表示人均 GDP 与 GDP 成长率对应的二维关系，更好的图示了 GDP 成长率响应面在人均 GDP 上的实际分布。

选择性偏差与敏感性分析

- Rosenbaum(2002) 指出当两个受测单元 j 和 k 拥有相同的共变量 **X**，或者相同、相似的倾向值，但他们接触处理的概率 π(倾向值) 却不相同时 ($\pi_j \neq \pi_k$)，就会存在**隐藏性偏差**。
- 对于这样的情况，敏感性分析即是探究不同程度的隐藏性偏差对于处理效用所造成的决定性影响；
- 也就是说，究竟要多大程度的影响才会改变这项研究的因果推论？
- 如果需要很大程度的影响才能改变推论结果，则原来推论结果受到隐藏性偏差的影响是低敏的（这是我们希望看到的结果）。

敏感性分析数学推导

- 单元接受处理发生比：

$$\pi = \text{logit}^{-1}\left(\mathbf{X}'\beta\right) \quad (1)$$

$$\Rightarrow \log\left(\frac{\pi}{1-\pi}\right) = \mathbf{X}'\beta \quad (2)$$

$$\Rightarrow \frac{\pi}{1-\pi} = \exp\left(\mathbf{X}'\beta\right) \quad (3)$$

敏感性分析数学推导

$$\frac{1}{\Gamma} \leq \underbrace{\frac{\left(\frac{\pi_j}{1-\pi_j}\right)}{\left(\frac{\pi_k}{1-\pi_k}\right)}}_{\text{接触处理发生比的比率 (odds ratio)}} \leq \Gamma$$

- 当 $\Gamma = 1$ 时，表示单元 j 和单元 k 接触处理的发生比是相等的 $\frac{\pi_j}{1-\pi_j} = \frac{\pi_k}{1-\pi_k}$，也就是说，它们拥有相同的倾向值 $\pi_j = \pi_k$，而 $\mathbf{X}_j = \mathbf{X}_k$，所以是不存在着隐藏性偏差的；
- 假定 $\Gamma = 2$，而 $\mathbf{X}_j = \mathbf{X}_k$，则表示即便单元 j 和 k 非常相似，但单元 j 接触处理的发生比是单元 k 的两倍，
- 综上所述，Γ 即是测量隐藏性偏差使一项研究推论结果改变的程度；研究者应当考虑在 Γ 不同取值的情况下，因果推论会发生怎样的变化。

敏感性分析操作

1. 假定处理效用是有效的，如果隐藏性偏差造成接触处理发生比的比率不同，则这组配对对应的结果 Y 可能会不同（处理效用加上隐藏性偏差所致），称这个结果的差异为 δ_s；
2. 不同的敏感性分析方法会估算不同的 δ_s 统计数值，大致会用到的数值有差值 $Y_j - Y_k$、差值的正负符号、差值的绝对值 $d_s = |Y_j - Y_k|$ 、差值绝对值在所有 d_s 中的排序（秩），接触处理发生比的比率。
3. 使用 Γ 估计 δ_s 的理论取值。一般来说会令 $\Gamma = 1$，然后逐渐增加 Γ 数值。既然 Γ 是研究者主动赋值，便可以使用它来计算不同 δ_s 的理论取值。
4. 进行无效假设检验 (null hypothesis testing)。重点检验在逐步增加 Γ 后，处理效用 τ 是否有效。
5. 解读 Γ 大小对推论的影响。一般来说，$\Gamma > 2$ 就表示该研究通过了敏感性检验。

敏感性分析实例

Γ	sig^+	sig^-	$\hat{T}^+$	$\hat{T}^-$	CI^+	CI^-
1	< 0.0001	< 0.0001	15	15	9.5	20.5
2	0.0018	< 0.0001	10.25	19.5	4.5	27.5
3	0.0136	< 0.0001	8	23	1	32.5
4	0.0388	< 0.0001	6.5	25	-1	37
4.25	0.0468	< 0.0001	6	25	-1.5	38.5
4.35	0.0502	< 0.0001	6	25.5	-2	38.5
5	0.0740	< 0.0001	5	26.5	-3	42

Γ 是无法观测到的因素对于发生不同接触处理发生比的对数；sig^+ 和 sig^- 分别是 Wilcoxon 符号秩检验显著性水平的上界和下界；$\hat{T}^+$、$\hat{T}^-$、CI^+ 和 CI^- 分别是 Hodges-Lehmann 检验法点估计和 95% 信用区间的上界和下界。

ATE

$$\hat{\tau} = \widehat{\text{ATE}} = \frac{n^1 \times \widehat{\text{ATT}} + n^0 \times \widehat{\text{ATC}}}{n^1 + n^0}$$

处理效用的标准误

1. 使用样本可重复替代的方式，对于样本量为 N 的数据，重新抽出新的样本量为 N 的数据。
2. 进行倾向值匹配法分析，获得平均处理效用。
3. 重复以上步骤 s 次后（次数越多，估计值越稳健，建议次数在 5000 次以上），获得 s 组平均处理效用。
4. 计算这 s 组平均处理效用 τ_s 的标准误差是为平均处理效用的经验标准误差（稳健标准误差），$\text{sd}(\tau_s)$。

非二元处理下的平均处理效用

1. 使用建议的五分位点或六分位点将数据分为六至七个子分类(subclasses)；
2. 使用相应的回归方程分别估计各子分类内处理变量对于结果变量的回归系数，即为各分类内的处理效用；
3. 然后，依照各子分类的样本量，采用加权平均的方式结合各分类所得的处理效用。
4. 或者，分别汇报在各子分类下不同的处理效用。

因果推论？没那么简单！

- 匹配后，共变量的分布可能缺乏平衡、缺乏重合。
- 如果检验过程中，发现缺乏平衡和重合，必须重新回到匹配前的步骤，选择共变量并调整估计倾向值的回归方程和相关的参数设定。
- 研究者必须在匹配后检查这两类问题，并如实使用图表，说明最后分析结果中共变量平衡和重合数据的情况。

因果推论？没那么简单！

- 匹配后，即便共变量的分布平衡和重合状况良好，却可能发生选择性偏差的问题，或未控制应该控制的共变量缺失（遗漏变量偏差）。
- 如果问题无法通过控制先前未控制的共变量解决（因为不存在该共变量的观测值），则研究者必须进行敏感性分析来说明研究结果可以容忍多大程度的偏差，而不改变原先推论的结果。

因果推论？没那么简单！

- 匹配后，就算共变量的分布平衡和重合良好，选择性偏差在可容忍的范围，控制了所有可观测到必要的共变量，还是有可能由于估计倾向值的参数回归式错误而造成偏差。
- 假定处理变量与共变量之间并不是简单的线性关系，而是非线性关系，如果参数回归式没有考虑到这个因素，估计得的倾向值自然不能良好的反应真实的处理分配，据此估计的处理效用自然会有偏差，
- 在这个情况下，研究者必须重新考虑使用非线性的回归方程估计倾向值。

因果推论？没那么简单！

- 匹配后，就算共变量的分布平衡和重合良好，选择性偏差在可容忍的范围，控制了所有可观测到必要的共变量，参数回归方程也适当的反应处理分配与共变量的关系，如果共变量的观测值存在观测值录入错误、缺失值、测量错误等杂讯，研究结果还是不可信。
- 当然，这些问题发生在任何定量分析中，都会影响结果，研究者只能尽量确保共变量观测值的正确性，并减少缺失值对于结果的影响。

因果推论？没那么简单！

- 即便以上的问题都解决了，使用观测性数据进行因果推论的社会科学家们，其研究设计和之后的数据搜集方式仍然是社会科学中因果推论研究的隐患，由于观测性数据及其研究设计本来就不是实验室数据和实验设计，
- 所以研究者在处理分配随机化以及控制其他可能影响处理和结果变量之共变量上，难免存在力所不逮之处，容易被挑出处理分配有选择性偏差，遗漏控制某些重要共变量等等数据上的问题，这种先天不足的数据问题，更是社会科学应用倾向值匹配法进行因果推论的根本性难题。
- 因此，近年来，结合实验研究设计来搜集观测性数据，成为社会科学领域研究的主流。