Propensity Score 知识整理

目录

Propensity Score 知识整理

2018-09 阎相达

符号表示

$Y_{i,1}$，第 i 个样本进入实验组时的输出
$Y_{i,0}$，第 i 个样本进入控制组时的输出
D_i，第 i 个样本进入组的编号
e(X_i)，第 i 个样本的 propensity score 值
n，样本数量
g，分层组的编号
n_g，表示第 g 个分层组的样本数量
w_i，表示控制组中样本的权重
β^T，表示干扰变量的权重

Stable Unit Treatment Value Assumption（SUTVA）
Strongly Ignorable Treatment Assignment Assumption（Strongly Ignorable）
Missing Completely as Random（MCAR）
Missing at Random（MAR）

一、处理效应

将某个实验或某个操作带来的效果称为处理效应（treat effect），在现实中经常需要衡量这个效应的大小是多少。当样本被分为实验组和控制组两组时，即只有一种处理操作，想要获得处理效应可以被定义为(He, Hu, & He, 2016)，

$$\text{Average Treat Effect(ATE)} = \mathrm{E}\left(Y_{i,1}\right) - \mathrm{E}\left(Y_{i,0}\right)$$
$$\text{Average Treat Effect on Treated(ATT)} = \mathrm{E}\left(Y_{i,1}|D_i = 1\right) - \mathrm{E}\left(Y_{i,0}|D_i = 1\right)$$

ATE表达的是对于全部的个体，它们进入实验组与进入控制组后输出的期望的差异。ATT表达的是对于已经进入实验组的个体，它们进入实验组与进入控制组后输出的期望的差异，往往ATT更多的被当作处理效应。

由于进入实验组或进入控制组往往是互斥的，因此$\left(Y_{i,1}\right)$和$\left(Y_{i,0}\right)$均有一部分数据是不可观测的。$\left(Y_{i,0}|D_i = 1\right)$也是无法被观测到的，因此为了计算ATT，要通过一些其他方法得到$\mathrm{E}\left(Y_{i,0}|D_i = 1\right)$。

1.1 随机实验中的处理效应

对于随机实验，含有默认的假设$D_i \perp \left(Y_{i,1}, Y_{i,0}\right)$，$D_i$独立于$(Y_{i,1}, Y_{i,0})$，$(Y_{i,1}, Y_{i,0})$在实验组与控制组中的分布是一致的，即$\mathrm{E}\left(Y_{i,0}|D_i = 1\right) = \mathrm{E}\left(Y_{i,0}|D_i = 0\right)$，此时 ATT 可以表达为，

$$\mathrm{ATT} = \mathrm{E}(Y_{i,1}|D_i = 1) - \mathrm{E}(Y_{i,0}|D_i = 1) = \mathrm{E}(Y_{i,1}|D_i = 1) - \mathrm{E}(Y_{i,0}|D_i = 0)$$

1.2 非随机实验中的处理效应

对于非随机实验，$(Y_{i,1}, Y_{i,0})$在实验组与控制组中的分布是不一致的，因此$\mathrm{E}(Y_{i,0}|D_i = 1) \neq \mathrm{E}(Y_{i,0}|D_i = 0)$。

Strongly ignorable 假设(Rosenbaum & Rubin, 1983)认为$D_i \perp (Y_{i,1}, Y_{i,0})|X_i$，即当$X_i$相等时，$(Y_{i,1}, Y_{i,0})$在实验组与控制组中的分布是一致的，即$\mathrm{E}(Y_{i,0}|D_i = 1, X_i) = \mathrm{E}(Y_{i,0}|D_i = 0, X_i)$，

$$\mathrm{E}(Y_{i,1}|D_i = 1, X_i) - \mathrm{E}(Y_{i,0}|D_i = 1, X_i) = \mathrm{E}(Y_{i,1}|D_i = 1, X_i) - \mathrm{E}(Y_{i,0}|D_i = 0, X_i)$$
$$\mathrm{ATT} = \mathrm{E}[\mathrm{E}(Y_{i,1}|D_i = 1, X_i) - \mathrm{E}(Y_{i,0}|D_i = 0, X_i)]$$

二、Propensity Score 的应用

2.1 Propensity Score 的意义

如果X_i是多维的，那么不太容易找到相同或相近的X_i。因此，propensity score 本质上是一种降维方法，以$\mathrm{e}(X_i)$代表每一个样本点的 propensity score 值，即被分入进实验组的概率值，作为一个标量更加容易找到相同或相近的X_i，propensity score 通过调整实验组和控制组的$\mathrm{e}(X_i)$分布一致，使得干扰变量分布一致。

2.2 Propensity Score 的计算方法

对于两个组的情况（实验组与控制组），以D_i为因变量，X_i作为自变量构建 logistic regression 模型，并用模型的预测值作为$\mathrm{e}(X_i)$。

对于 Propensity Score 的计算,有多种方法被尝试,除了经典的 logistic regression，还包括树结构的机器学习模型、支持向量机、神经网络等等，这些模型不在假设干扰变量具有线性的影响时，有可能取得较好的训练模型。

2.3 基于 Propensity Score 计算 ATT

2.3.1 Propensity Score Matching（PSM）

对于实验组中的每一个样本，在控制组中选择合适的样本进行匹配，在匹配之后可以实现实验组与控制组分布的一致，匹配的具体操作比较多样化，没有哪一种是绝对好的，要根据实际情况进行选择，

（1）匹配的原则是实验组中的样本和控制组中的样本$\mathrm{e}(X_i)$相等或接近，因为$0 < \mathrm{e}(X_i) < 1$，所以两个样本的倾向得分数值差距应该在较小的范围。

（2）对于每一个实验组的样本，可以匹配到控制组中的一个或多个样本，即 1:1 匹配

或 1：n 匹配。

（3）对于每一个控制组的样本，可以选择是否有放回的进行匹配，对于控制组中样本较少时，可以尝试有放回匹配，即一个控制组的样本可以被重复匹配多次。对于没有匹配上的样本，无论是实验组中的还是控制组中的都应该被剔除。

使用 Propensity Score 进行匹配的优势在于，标量相比向量更容易进行比较。缺点在于，如果两个组的 Propensity Score 分布差异太大，导致匹配之后只有很少的样本，会产生其他的选择偏差，以及无法说明匹配上的样本是否具有代表性。

2.3.2 Inverse Propensity Score Weighting（IPSW）

对于实验组的样本输出值，赋与$\frac{1}{\mathrm{e}(X_i)}$，对于控制组的样本输出值，赋与$\frac{1}{1-\mathrm{e}(X_i)}$，同样可以实现$X_i$在实验组与控制组分布的一致，调整后 ATT 的计算方法，

$$\mathrm{ATT} = \sum_{i\in\text{实验组}} \frac{1}{\mathrm{e}(X_i)} * Y_{i,1} - \sum_{j\in\text{控制组}} \frac{1}{1-\mathrm{e}(X_j)} * Y_{j,0}$$

因为$0 < \mathrm{e}(X_i) < 1$，因此样本实际上是被“扩大的”，如果$\mathrm{e}(X_i)$过于接近于 0 或 1 时，会导致分母过小，ATT 值计算有较大误差，从理论上讲，当$\mathrm{e}(X_i)$过于极端时，由于另一个组里面基本不会出现这样的数据，因此这种数据不应该被保留。解决这个问题比较科学的方法是使用 Stabilized Weights，

$$\mathrm{ATT} = \sum_{i\in\text{实验组}} \frac{\Pr(D_i=1)}{\mathrm{e}(X_i)} * Y_{i,1} - \sum_{j\in\text{控制组}} \frac{\Pr(D_i=0)}{1-\mathrm{e}(X_j)} * Y_{j,0}$$

Stabilized Weights 在分子上赋与D_i的先验概率（p<1.0），实现权重的降低，在此基础上，Truncation Weights 对权重范围进行了限制，如只保留分位数 5%-95%内的权重。

2.3.3 Propensity Score Stratification（PSS）

传统的分层分析方法依据X_i进行分层，假设X_i的维度为 n，那么所需要分的层次至少为2^n，因此被分到每个层次中的样本数量会非常少，经常出现只有实验组或只有控制组的样本。使用$\mathrm{e}(X_i)$可以较好的解决这个问题，在更合理的情况下进行更少的分层，保证每层内尽可能同时保留实验组和控制组的数据。

使用 PSS 时，ATT 的计算变为每一层的数据ATT_g的加权，权重为每层中样本的数量比例，计算公式为，

$$\mathrm{ATT}_g = \mathrm{E}\left(Y_{i,1}\middle|\mathrm{i}\in G_g, D_i=1, e(X_i)\right) - \mathrm{E}\left(Y_{i,0}\middle|\mathrm{i}\in G_g, D_i=0, e(X_i)\right)$$

$$\mathrm{ATT} = \sum_g \mathrm{ATT}_g * \frac{n_g}{n}$$

文献证明，将数据按照$e(X_i)$切分为 5-10 层数据是比较合理的。

2.3.4 Covariance Adjustment

除了前面三种计算方法，也可以以D_i作为解释变量构建回归模型来计算 ATT。如果是随机实验，当构建回归模型时，把D_i作为解释变量，由于$D_i \in \{0,1\}$，因此回归模型中D_i的系数可被认为是实验效果。

但是，如果实验是非随机的，直接进行回归会出现较大偏差。基于 Propensity Score 的理论，有两种方法可以被使用，

（1）在方程构建的角度，在方程中同时引入$e(X_i)$作为解释变量，相当于制造了随机实验的环境。

$e(X_i) = f(X_i)$是X_i的函数，可以理解为$e(X_i)$是全部干扰变量的一种表达，其系数可以理解为全部干扰变量对方程的贡献。以最简单的情况为例，只有两个样本，分别属于实验组和控制组，且$e(X_i)$相等，此时D_i的系数代表实验效果，而$e(X_i)$的系数代表干扰变量的贡献。

对于使用$e(X_i)$作为解释变量而不是直接将X_i作为解释变量的意义在于，使用 propensity score 框架，对于X_i可见构建更加复杂的模型，对于$e(X_i)$的估计可以不必具有解释性，而且估计$e(X_i)$时可以使用大量的变量(Hade, 2012)。

（2）在样本重抽样的角度，可以使用 PSM 或 IPSW 方法对样本进行重抽样，在样本层面实现分布的一致，再对模型进行回归。

需要注意的是，使用 Regression Adjustment 是为了估计实验效果的大小，才将D_i引入到方程中，按照线性回归的理解，这种影响是一种“平均”思想的影响，即最开始提到的 ATT。如果不将D_i带入模型中，ATT 影响会被增加到常数项中，但是此时并不能够说明回归方程中其他解释变量系数是没有被影响的。

三、Propensity Score 的多维扩展

上面介绍的 propensity score 只能应用于两个数据组的情况（实验组与控制组），由于提出时间较早，因此发展相对成熟，在医疗、政策领域有非常广泛的应用。

当有多个实验组时，也可以使用 propensity score 的理论框架。对于多维的 propensity score，研究的问题与二维的有所不同。二维条件下，ATT 表示是两个样本组实验效果的差异，但是多维情况下，只能比较多维度之间两两之间的差异，但是这种差异往往不具有传递性或可比性。

3.1 计算多维 propensity score

此时 propensity score 不是单维度的，因此不能使用 logistic regression，比较等价的替换方法是 Multinomial Logistic Regression（也称为 Softmax Regression）。对于每一个样本，$e(X_i) \in R^d$，此时 propensity score 是一个向量。

3.2 匹配多维 propensity score

在对多维 propensity score 进行匹配时，要权衡两个问题，即匹配方法舍弃的一部分信息以及匹配的效率（避免穷举方法）。

3.2.1 欧式距离

3.2.2 Aitchison distance

$\sum_g \mathrm{e}(X_i) = 1$，因此将$\mathrm{e}(X_i)$定义为一种成分数据，此时使用欧式距离来衡量被证明是不太合适的。

$$\mathrm{g}\big(\mathrm{e}(X_i)\big) = \prod_{d=1}^{D} \mathrm{e}(X_i)_d^{\,1/D}$$

$$\mathrm{clr}\big(\mathrm{e}(X_i)\big) = [\ln\left(\frac{\mathrm{e}(X_i)_1}{g\big(\mathrm{e}(X_i)\big)}\right), \ln\left(\frac{\mathrm{e}(X_i)_2}{g\big(\mathrm{e}(X_i)\big)}\right), \dots, \ln(\frac{\mathrm{e}(X_i)_D}{g(\mathrm{e}(X_i))})]$$

$$\mathrm{ait}_{\mathrm{distance}}\left(\mathrm{e}(X_i), \mathrm{e}\big(X_j\big)\right) = |\left|clr\big(\mathrm{e}(X_i)\big), clr\left(\mathrm{e}\big(X_j\big)\right)\right|\,|_2^2$$

Aitchison 保留了相对的关系，在比例一致的情况下，得到的距离更加接近。

3.2.3 Common referent matching（CRM）

CRM 主要应用在 3 维的情况下，以其中一个组为基准，分别让其他两个组与该组进行匹配，以共同被匹配的样本作为最终的匹配数据，使用这种方法，可以保证不同组之间的结果可以具有可比性，缺点是由于两次匹配会导致公共的数据部分很少。

3.2.4 Series of binomial comparisons（SBC）

（1）当比较其中两个实验的效果差异时，只取这两个实验组的数据，同样使用二维 Propensity Score 的方法。但是这种比较没有被证明具有传递性。

（2）当比较其中两个实验的效果差异时，将计算出的多维$\mathrm{e}(X_i)$对应的两维进行归一化，然后进行匹配，这种方法相比第一种考虑了除比较的实验组之外的其他组的部分信息，这种比较也没有传递性质。

3.2.5 Maching learning

（1）K-dimentions Trees,K-D 树是二叉搜索树的多维形态，树的构建即能体现数据之间的距离，通过构建 K-D 树，可以在$\mathrm{O}(\log(\mathrm{N}))$时间复杂度计算结果。文献证明了 K-D 树来进行匹配时，比 CRM 和 SBC 有更好的效果。

（2）K-means。

四、Propensity Score 的替代方法

Propensity Score 本质上是通过平衡X_i在实验组和控制组中的分布使样本满足$D_i \perp \big(Y_{i,1}, Y_{i,0}\big)|X_i$。平衡$X_i$可以有多种方法，除了最早发展的 Propensity Score，后来又有一些非参数方法被提出，

4.1 常数权重

为每一个控制组样本赋与一个常数权重。该类方法的核心思想是为每一个控制组中的样本找到一个权重(Athey, Imbens, & Wager, 2018)，使得控制组样本的干扰变量的分布与实验组分布一致。

最简单的方法是，$\underset{w}{\operatorname{argmin}}(\left|\left|\overline{X_t} - \sum_{i\in \text{Control}} w_i * X_i\right|\right|_2^2 + \left|\left|w\right|\right|_2^2)$。(加入 L2 正则项是因为干扰变量维度少于样本数量时，矩阵不满秩，有无穷多的解。)

4.2 向量权重

为每一个控制组样本赋与一个向量权重。

$$\underset{w}{\operatorname{argmin}}(\left|\left|\beta^T * (\overline{X_t} - \sum_{i\in \text{Control}} w_i * X_i)\right|\right|_2^2)$$

相比前一种方法，增加的β^T相当于为每个干扰变量都赋与了不同的权重(Kuang, Cui, Li, Jiang, & Yang, 2017)。

参考文献

Athey, S., Imbens, G. W., & Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(4), 597-623.

Hade, E. M. (2012). Propensity score adjustment in multiple group observational studies: Comparing matching and alternative methods. The Ohio State University.

He, H., Hu, J., & He, J. (2016). Overview of propensity score methods Statistical Causal Inferences and Their Applications in Public Health Research (pp. 29-48): Springer.

Kuang, K., Cui, P., Li, B., Jiang, M., & Yang, S. (2017). Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing. Paper presented at the Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Lopez, M. J., & Gutman, R. (2017). Estimation of causal effects with multiple treatments: a review and new ideas. Statistical Science, 32(3), págs. 432-454.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.