因果推断简介

丁鹏

北京大学 数学科学学院 概率统计系

Email: dingyunyiqiu@163.com

摘要

统计学在“相关”方面的推断取得了很多的成就，但是在因果推断方面取得的成就十分有限。这里从 Yule-Simpson Paradox 讲起，说明用统计学方法做因果推断的困难。然后引入 Rubin Causal Model(RCM)，以及 RCM 在完全随机化试验和观测性研究中如何进行因果推断。这部分将引入因果推断中的一些核心概念，如可忽略性、倾向得分、主分层和工具变量。最后讲因果图(Causal Diagram)，在一个有向无环图中引入了 do 操作，这个图便有了因果的含义。这部分将讲到图上因果作用的识别性准则：前门准则和后门准则。

目录

1 相关与因果的不同: Yule-Simpson Paradox

在高维列联表分析中，有一个很有名的例子，叫做 Yule-Simpson Paradox。此悖论表明，X 和 Y 边缘上正相关，但是给定另外一个变量 Z 后，在 Z 的每一个水平上，X 和 Y 可能负相关。下面表 1 就是一个数值的例子(Pearl，2000)。 由表 1 可以看出，在整个人群中，吃药与康复之间存

表 1: Yule-Simpson Paradox

合并表	康复	未康复	康复率
吃药	20	20	50%
安慰剂	16	24	40%
男性	康复	未康复	康复率
吃药	18	12	60%
安慰剂	7	3	70%
女性	康复	未康复	康复率
吃药	2	8	20%
安慰剂	9	21	30%

在正相关；然而，当用性别对人群分层后发现在男性和女性人群中，吃药与康复都是负相关。这就是 Yule-Simpson Paradox。

上面的例子是人工构造的，在现实中，也存在不少的实例正是 Yule-Simpson Paradox。比如，UC Berkeley 的著名统计学家 Peter Bickel 教授 1975 年在 Science 上发表文章，报告了 Berkley 研究生院男女录取率的差异。他发现，总体上，男性的录取率高于女性，然而按照专业分层后，女性的录取率却高于男性(Bickel 等，1975)。

在流行病学的教科书(如 Rothman 等，2008)中，都会讲到“混杂偏倚”，其实就是 Yule-Simpson Paradox，书中列举了很多流行病学的实际例子。

根据因果图的理论，出现表 1 的原因在于性别、吃药与康复三个变量形成如下的有向无环图(Directed Acyclic Graph, DAG)，见图 1。

由于有 Yule-Simpson Paradox 的存在，观测性研究中很难得到有关因

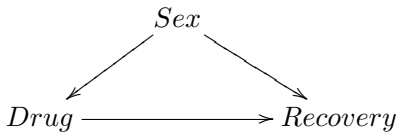

图 1: Yule-Simpson Paradox和DAG

果的结论，除非加上很强的假定，这在后面会谈到。比如，一个很古典的问题：吸烟是否导致肺癌？由于我们不可能对人群是否吸烟做随机化试验，我们得到的数据都是观测性的数据：即吸烟和肺癌之间的相关性(正如表 1) 的合并表。此时，即使我们得到了吸烟与肺癌正相关，也不能断言“吸烟导致肺癌”。这是因为可能存在一些未观测的因素，他既影响个体是否吸烟，同时影响个体是否得癌症。比如，某些基因可能使得人更容易吸烟，同时容易得肺癌；存在这样基因的人不吸烟，也同样得肺癌。此时吸烟和肺癌之间相关，却没有因果作用。

相反的，我们知道放射性物质对人体的健康有很大的伤害，但是铀矿的工人平均寿命却不比常人短；这是流行病学中有名的“健康工人效应”。这样一来，似乎是说铀矿工作对健康没有影响。但是，事实上，铀矿的工人通常都是身强力壮的人，不在铀矿工作寿命会更长。

上面的各种反例已经说明了因果推断的困难。下面介绍 Rubin Causal Model、完全随机化试验中的因果推断、观测性研究中的因果推断、随机化试验存在不依从和死亡删失下的因果推断，以及因果图模型。

2 Rubin Causal Model(RCM) 和随机化试验

因果推断用的最多的模型是 Rubin Causal Model(RCM)和Causal Diagram(Pearl, 1995)。Pearl(2000) 中介绍了这两个模型的等价性，但是就应用来看，RCM 更加精确，而 Causal Diagram 更加直观。这部分主要介绍 RCM，但是也会用 Causal Diagram 来直观的描述所研究的问题。

设 Z_i 表示个体 i 接受处理与否，处理取 1，对照取 0 (这部分的处理变量都讨论二值的，多值的可以做相应的推广)；Y_i 表示个体 i 的结果变量。另外记 $(Y_i(1), Y_i(0))$ 表示个体 i 接受处理或者对照的潜在结果(potential outcome)，那么 $Y_i(1) - Y_i(0)$ 表示个体 i 接受治疗的个体因果作用。不幸

的是，每个个体要么接受处理，要么接受对照，$(Y_i(1), Y_i(0))$ 中必然缺失一半，个体的因果作用是不可识别的。注意，对于个体 i，潜在结果是确定的数；这里的随机性体现在 i 上，i 可以看成通常概率论中样本空间 Ω 中的样本点 ω。但是，在 Z 做随机化的前提下，我们可以识别总体的平均因果作用(Average Causal Effect):

$$ACE(Z \to Y) = E(Y_i(1) - Y_i(0)).$$

这是因为

$$\begin{aligned} ACE(Z \to Y) &= E(Y_i(1)) - E(Y_i(0)) \\ &= E(Y_i(1)|Z_i = 1) - E(Y_i(0)|Z_i = 0) \\ &= E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0), \end{aligned}$$

最后一个等式表明 ACE 可以由观测的数据估计出来。其中第一个等式用到了期望算子的线性性(非线性的算子导出的因果度量很难被识别！)；第二个式子用到了随机化，即 $Z \perp (Y(1), Y(0))$($\perp$表示独立性)。

由此可见，随机化试验对于平均因果作用的识别起着至关重要的作用。

3 观测性研究：可忽略性、倾向得分与回归分析

现实中，很多研究都不是随机化的，比如由于伦理的限制，我们没法对个体吸烟与否进行随机化。观测性研究中，我们通常收集到如下的数据：个体的属性变量 X (如年龄、性别、病史等等)、个体是否接受处理 Z (比如是否吃某种新药、是否吸烟、是否参加某个培训项目或者是否获得教育优惠券等等)，个体的结果变量 Y (如生活质量、是否生病或者考试成绩等等)。由于有 Yule-Simpson Paradox 的存在，我们知道用如下的条件期望之差

$$E(Y \mid Z = 1) - E(Y \mid Z = 0)$$

是不能度量处理的因果作用的。

这里便有一个 ACE 的识别性问题，即通过观测的数据我们能否得到 ACE 的相合估计(严格的定义是同一个观测数据是否对应着唯一的 ACE)。实际中，这需要一个不可验证的假定：可忽略性。

3.1 可忽略性与 ACE 的识别性

随机化实验蕴含着强可忽略性假定，即

$$Z \perp (Y(1), Y(0))$$

这条假定使得我们可以通过观测数据识别 ACE。但是，观测性研究中，个体选择处理与否与其个体属性可能相关(注意 (Y(1),Y(0)) 其实表示的是个体属性)，上面的假定可能被破坏。但是，通常的方法是：收集充分多的个体信息 X，使得如下的强可忽略性假定成立：

$$Z \perp (Y(1), Y(0))|X.$$

可以证明，此时 ACE 可以识别，因为

$$\begin{aligned} ACE &= E(Y(1)) - E(Y(0)) \\ &= E[E(Y(1) \mid X)] - E[E(Y(0) \mid X)] \\ &= E[E(Y(1) \mid X, Z = 1)] - E[E(Y(0) \mid X, Z = 0)] \\ &= E[E(Y \mid X, Z = 1)] - E[E(Y \mid X, Z = 0)]. \end{aligned} \tag{1}$$

上面的推导我们看出，强可忽略性的假定可以减弱到弱可忽略性，即

$$Z \perp Y(1), Z \perp Y(0),$$

同样保证 ACE 可以识别。由 (1) 可以知道估计 ACE 的关键在于两个条件矩 $E[E(Y|X, Z = 1)]$ 和 $E[E(Y|X, Z = 0)]$.

在现有的方法中，有三种模型可以用来估计观测性研究中处理的作用：倾向得分(propensity score)、线性回归和 Heckman Selection Model(或者称 Tobit Model，如Amemiya，1985)。下面分别介绍三种方法。

3.2 倾向得分

如果问题足够的简单，X 为二值变量(如性别)，为了估计 ACE 我们可以直接按照 $X = 1$ 和 $X = 0$ 将人群分成两层，在每层中分别估计平均因果作用 $ACE_{X=1}$ 和 $ACE_{X=0}$，再根据 $X = 1$ 和 $X = 0$ 的比例进行加权即可。但是，事实上，X 的维数可能很高且可能有连续的分量，这时候很难

将人群按照 X 分层，即使分了层，每一层中的人数很少甚至没有人，很难进行估计。

基于这个问题，Rosenbaum and Rubin(1983) 提出了倾向得分的概念，实际上这是一种降维的手段，将高维的 X 降到低维。倾向得分定义成

$$e(X) = P(Z = 1 \mid X)$$

且满足：

(1) $X \perp Z | e(X)$；

(2) 如果有强可忽略性假定，且 $0 < e(X) < 1$，则 $Z \perp (Y(1), Y(0)) | e(X)$ 且 $0 < e(X) < 1$。

上面的第二条性质表明：如果给定 X，处理机制是可忽略的，那么只需要给定一个一维的变量 $e(X)$，处理机制也是可忽略的。Rosenbaum and Rubin(1983) 甚至证明了，倾向得分是最“粗糙”的变量，使得给定这个变量以后，处理机制可忽略。这样一来，我们得到了一种 ACE 的估计方法：

Step 1：先拟合一个 Logistic/Probit 模型，估计每个个体的倾向得分 $\widehat{e}(X)$;

Step 2: 用估计的倾向得分 $\widehat{e}(X)$ 分层，在每一层中估计平均因果作用，再加权平均即可。

上面的方法称为“分层”(stratifying)，Hirano, Imbens and Ridder(2003) 从经验似然的角度指出了另一种“加权方法”(weighting)，并证明这是一种是半参数有效的估计方法，该方法用以下统计量作为ACE的估计：

$$\widehat{ACE} = \frac{1}{N} \sum_{i=1}^{n} \left[\frac{Y_i Z_i}{\widehat{e}(X_i)} - \frac{Y_i (1 - Z_i)}{1 - \widehat{e}(X_i)} \right].$$

事实上，这个估计量和抽样调查理论中的 Horvitz-Thompson 估计(Jun Shao, 2003)是一致的，都是一种逆概加权的估计。

3.3 回归与 Heckman Selection Model

前面指出了，ACE 的估计化成了两个条件矩的估计，因此，有的计量经济学家会倾向于用回归模型来估计条件矩。Wooldrige(2002) 中给出：如果在 X 的每个水平下，平均因果作用都是常数，那么我们可以用如下的回归模型

$$E(Y|Z, X) = \alpha + \beta Z + g(X, \gamma),$$

其中 $\beta = ACE$。当然，也可以在回归模型中加入 Z 和 X 的交互项。

另外一个模型是 Heckman(1979) 提出的 Heckman Selection Model 的推广(见 Greene，2002)。模型为：

$$\begin{aligned} Z^* &= \delta + \theta X + v, \\ Z &= I(Z^* \geq 0), \\ Y &= \alpha + \beta Z + \gamma X + u, \end{aligned}$$

其中 (u, v) 服从联合正态分布。

上面的回归和 Heckman Selection Model 都有较强的模型和参数假定(Heckman Model 甚至假定了分布，这在计量经济中是比较少见的)，而倾向得分的方法更接近于“非参数”的方法，近来也受到了计量经济学的广泛接受。

4 随机化试验“失败”时的因果推断

理想的随机化试验中，处理组和对照组的人完全按照分配行动。但是，现实中，试验中的人会不依从(noncompliance)、提前离开或者未观测到结果变量就已经死亡，这导致了三个问题不依从、非随机的缺失数据和死亡删失。下面分别介绍不依从和死亡删失，而缺失数据的问题在大多数的统计问题中都会存在。

4.1 不依从下的因果推断：主分层与工具变量

这里的记号与 Imbens and Rubin(1997) 保持一致。设总体有 N 个个体，对于个体 i，$Z_i = 1$ 表示个体 i 被随机化分配到处理组，$Z_i = 0$ 表示个体 i 被随机化分配到对照组；$D_i = 1$ 表示个体 i 最终接受处理，$D_i = 0$ 表示个体 i 最终接受对照。当 $Z_i \neq D_i$ 时，就称为“不依从”。用 Y_i 表示个体 i 的结果变量。

对于这样的数据，传统的分析方法有两种。一是 ITT(Intent to treat)，即分析

$$ACE(Z \to Y) = E(Y(1) - Y(0)),$$

即忽略掉不依从的影响，直接估计随机化对结果变量的作用。这种方法的

缺点在于他只能估计随机化的作用，并不是处理的作用。另外的估计方法是 AT(as treated)，直接算实际接受处理与实际接受对照的两群人结果变量均值之差，即

$$E(Y|D=1) - E(Y|D=0).$$

这种方法的缺点在于 D 不是随机化，因此 D 与 Y 之间有混杂(即可能有共同的原因影响个体 i 是否接受处理和其结果变量)，得到的两组均值之差并不是处理的作用。

这里采用 Frangakis and Rubin(2002) 中提出的“主分层(principal stratification)”的框架来分析，实际上 Imbens and Rubin(1997) 用的就是这样的框架。记

$$C_i = \begin{cases} c, & \text{如果 } D_i(0)=0 \text{ 且 } D_i(1)=1 \\ n, & \text{如果 } D_i(0)=0 \text{ 且 } D_i(1)=0 \\ a, & \text{如果 } D_i(0)=1 \text{ 且 } D_i(1)=1 \\ d, & \text{如果 } D_i(0)=1 \text{ 且 } D_i(1)=0, \end{cases}$$

由于 C 是潜在的结果，因此他不受任何处理以及处理后变量的影响，他的地位与协变量(处理前变量)相同。这种用处理后变量的潜在结果进行分层的方法就叫主分层。传统的方法往往推断

$$P(Y \mid Z=1, D=1) \; V.S. \; P(Y \mid Z=0, D=1)$$

和

$$P(Y \mid Z=1, D=0) \; V.S. \; P(Y \mid Z=0, D=0).$$

这种推断方法看上去似乎是“给定 D”的条件下，Z 对 Y 的因果作用。Rubin尖锐的指出了这种推断的错误(Frangakis and Rubin, 2002; Rubin, 2004)。

用 DAG(Directed Acyclic Graph) 来表示这个问题如图 2。

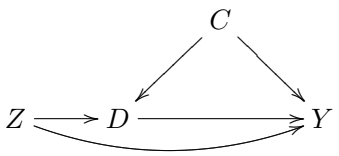

图 2: 随机化不依从

从DAG上很容易看出问题的所在，图 2 中 Z 与 Y 之间本来不存在“混杂(confounding)”，但是一旦条件在 D 上作推断，Z 和 C 变得不独立了，于是 C 成了 Z 和 Y 之间的混杂。流行病学家很早就意识到了“调整”中间变量(其实就是条件的推断)在统计推断中的危险，有个统计学家称这种条件推断带来的混杂为间接混杂(indirect confounding)。

Imbens and Rubin(1997)只考虑如下的因果度量

$$CACE(Z \to Y) = E(Y(Z=1, D(Z=1)) - Y(Z=0, D(Z=0)) \mid C=c),$$

即依从组中，随机化对结果变量的平均因果作用，这个量也就是处理对结果变量的平均因果作用，因为在依从组 $Z = D$。而在其他组中，因果作用是难以定义的，比如

$$E(Y(Z=1, D(Z=1)) - Y(Z=0, D(Z=0)) \mid C=c)(c=n,a,d)$$

只能描述在各个主分层中，随机化对结果变量的平均因果作用，但是并不是实际上处理的作用；因此这样的因果度量是含混不清的。事实上，我们可以证明，除了 $CACE$ 外，其他各层中处理对结果变量的因果作用是不可识别的。计量经济的文献中通常称 CACE 为 LATE(Local Average Treatment Effect)，如 Wooldrige(2002) 等。

在如下的假定：

假定4：单调性假定 $D(1) \geq D(0)$，即 $C = d$ 的人群不存在。

假定5：Exclusion Restriction，即对于人群 $C = n$ 和 $C = a$，有 $D(Z = 1) = D(Z = 0)$ 且 $Y(Z = 1, D(Z = 1)) = Y(Z = 0, D(Z = 0))$，在前面的 DAG 中反映出来就是 Z 到 Y 没有那条直接作用的边。

频率学派证明了 $CACE$ 是可以识别的(Angrist, Imbens and Rubin(AIR), 1996; Imbens and Angrist, 1994)：

$$CACE = \frac{ACE(Z \to Y)}{ACE(Z \to D)} = \frac{E(Y \mid Z=1) - E(Y \mid Z=0)}{E(D \mid Z=1) - E(D \mid Z=0)}.$$

上面的估计称为 Wald 估计。AIR(1996) 证明了这个估计和计量经济中的工具变量(instrumental variable)估计相一致，即在回归模型

$$Y = \alpha + \beta D + \varepsilon$$

中 D 和 ε 之间存在“混杂”(或称存在公共的原因 U，计量经济中称 D 是

内生变量)。此时用 Z 作为工具变量得到两个矩估计方程

$$\begin{aligned} E(Y) &= \alpha + \beta E(D), \\ E(ZY) &= \alpha E(Z) + \beta E(ZD), \end{aligned}$$

可以解出 $\beta = CACE$。

在 Y 存在非随机缺失的时候，$CACE$ 的推断稍微麻烦一些。Frangakis and Rubin（1999)在 Latent Ignorable 的假定下，证明了 $CACE$ 的识别性，并给出了相应的估计方法，此时对应的 DAG 是图 3，其中 R 为 Y 是否缺失的指示变量。

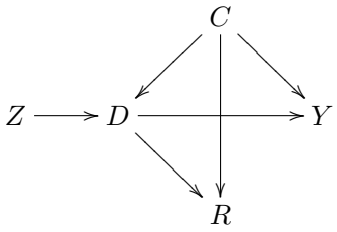

图 3: Latent Ignorable

Chen, Geng and Zhou（2009) 在 Completely Non-ignorable(CN) 的假定下证明了，如果 Y 是离散变量 $CACE$ 也可以是识别(事实上，Y 是连续变量 $CACE$也能识别？)，对应的 DAG 是图 4。

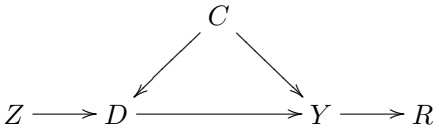

图 4: Completely Non-ignorable

4.2 死亡删失下的因果推断

死亡删失(truncated by death)问题经常出现在临床研究中。在随机化试验中，Z 依然表示随机化试验，中间变量 S 表示病人存活 ($S = 1$) 还是死亡 ($S = 0$)，Y 依然表示结果变量。这里会出现与不依从数据不同的问题，因为对于那些死亡的个体，最终的结果变量是没有意义的—不是

简单的缺失数据问题。这就给因果作用的定义带来了问题。Frangakis and Rubin(2002) 提出的主分层框架正好可以用来定义因果作用。记

$$C_i = \begin{cases} LD, & \text{如果 } S_i(1) = 0 \text{ 且 } S_i(0) = 1 \\ DD, & \text{如果 } S_i(1) = 0 \text{ 且 } S_i(0) = 0 \\ LL, & \text{如果 } S_i(1) = 1 \text{ 且 } S_i(0) = 1 \\ DL, & \text{如果 } S_i(1) = 1 \text{ 且 } S_i(0) = 0, \end{cases}$$

由于主分层中的 DL, DD, LD 层都涉及到了不可定义的结果变量，因此这三层的因果作用是含混不清的；因此我们只考虑 LL 层的平均因果作用，即

$$SACE = E(Y(1) - Y(0) \mid C = LL).$$

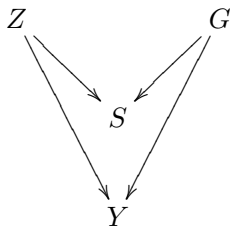

图 5: 死亡删失情形下的DAG

类似的，我们可以画出 DAG 来描述这个问题，如图 5。

根据观测到的 Z 和 S 我们可以把人群分成四组，如下图 6所示。

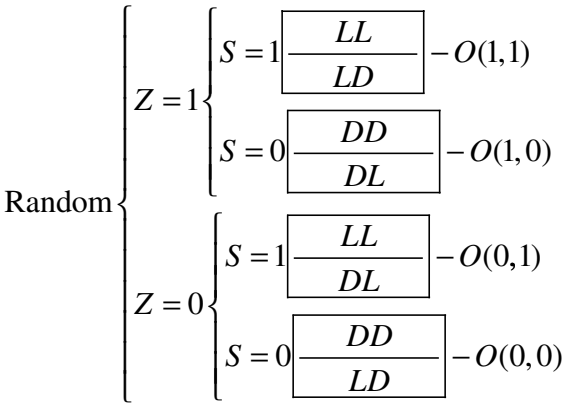

图 6: 死亡删失情形下人群分层

同样的，这里遇到的困难也是混合分布的识别性问题。Zhang, Rubin and Mealli(2009) 的建立如下的参数模型，其模型的识别性强烈的依赖于正态分布的假定。

对主分层 C，假定

$$\pi_{g:i} = P(C_i = g \mid X_i, \theta) = \frac{\exp\{\alpha_g + \beta_g^T X_i\}}{\sum_{g'} \exp\{\alpha_{g'} + \beta_{g'}^T X_i\}}, g \in \{LL, LD, DL, DD\}.$$

对与结果变量建立如下的线性回归模型

$$Y_i(z) \mid G_i = g, X_i, \theta \sim N(\mu_{g,z} + \eta_{g,z}^T X_i, \sigma_{g,z}^2).$$

用 EM 算法算出 MLE 即可。

耿老师的想法是，上面的参数模型本质上没有解决识别性的问题，实际中可以考虑寻找一个工具变量 A 它满足如下的 DAG (图 7)。 此时将 G

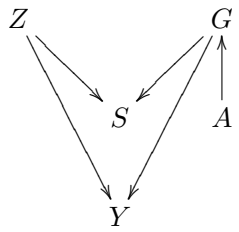

图 7: DAG under Assumption 1 to 4

看成因变量，借鉴 Latent Class Analysis(Goodman, 1976) 的方法证明识别性。先假定 Y 是离散的变量。由于给定 (Z, G)，S, Y, A 相互独立，因此我

们得到：

$$\begin{aligned}
& P(S=1,Y=y,A=a \mid Z=1) \\
= \;& \pi_{LL}P(Y=y \mid Z=1,G=LL)P(A=a \mid G=LL) \\
& \quad + \pi_{LD}P(Y=y \mid Z=1,G=LD)P(A=a \mid G=LD);
\end{aligned}$$

$$\begin{aligned}
& P(S=0,A=a \mid Z=1) \\
= \;& \pi_{DL}P(A=a \mid G=DL) + \pi_{DD}P(A=a \mid G=DD);
\end{aligned}$$

$$\begin{aligned}
& P(S=1,Y=y,A=a \mid Z=0) \\
= \;& \pi_{LL}P(Y=y \mid Z=0,G=LL)P(A=a \mid G=LL) \\
& \quad + \pi_{DL}P(Y=y \mid Z=1,G=DL)P(A=a \mid G=DL);
\end{aligned}$$

$$\begin{aligned}
& P(S=0,A=a \mid Z=0) \\
= \;& \pi_{LD}P(A=a \mid G=LD) + \pi_{DD}P(A=a \mid G=DD).
\end{aligned}$$

这里一共 $2(N_yN_a + N_a - 1)$ 独立方程，$(4-1) + 4(N_y - 1) + 4(N_a - 1)$ 个参数, 其中 N_y 和 N_a 是 Y 和 A 的水平数. 因此这些参数能够是识别的必要条件是

$$\begin{aligned}
2(N_yN_a + N_a - 1) &\geq (4-1) + 4(N_y - 1) + 4(N_a - 1), \\
2N_yN_a + 3 &\geq 4N_y + 2N_a.
\end{aligned}$$

特别的，当 $N_a = 2$ 时，上式变成了

$$3 \geq 4,$$

表明参数不可识别。当 $N_a = 3$ 时，上式变成

$$6N_y + 3 \geq 4N_y + 4,$$

$$N_y \geq 1.5.$$

表明 Y 至少两个水平。因此，当上面方程组对应的 Jacobian 非退化时，参数“局部可识别”，从而ACE也可以识别。

当 Y 连续时定义 $I_y = I(Y > y)$ 即可估计 $P(Y > y)$，从而 ACE 可识别。

注：如果不用主分层的概念，我们猜测计量经济学家会用 Heckman Selection Model(Heckman, 1979) 来处理这个问题，模型如下：

$$\begin{aligned} S^* &= \delta + \tau Z + \theta X + v, \\ S &= I(Z^* \geq 0), \\ Y^* &= \alpha + \beta Z + \gamma X + u, \\ Y &= \begin{cases} Y^*, \text{如果 } S = 1, \\ NA, \text{如果 } S = 0, \end{cases} \end{aligned}$$

其中 (u, v) 服从联合正态分布。这个模型的问题在于，主分层下某些人群的 Y^* 在某些处理下是没有定义的，而不是简单的缺失!

5 贝叶斯观点下的RCM

5.1 截面数据的贝叶斯因果推断

先介绍 Rubin(1978) 中采用的记号。设一次试验有 T 种治疗方案，总体 P 中有 N 的试验个体。我们关心的变量有协变量(又称处理前变量) $X = (X_1, ..., X_c)$；接受哪个处理的指示变量 W，W 可以取 $0, 1, ..., T$，其中 $W = 0$ 表示对照，不接受任何处理；潜在结果变量(又称处理后变量) $Y = (Y^1, ..., Y^T)$，其中每个 $Y^t = (Y_1^t, ..., Y_d^t)$ 表示 Y^t 的 d 个分量。这里通常需要一个**假定**：如果第 i 个个体接受处理 t，那么不管其他个体接受什么处理，观测到的结果变量 Y 唯一。这个假定通常称为 $SUTVA$(Stable Unit Treatment Value Assumption)，又称为“没有交互影响(no interference)”。这个假定可能会不成立，比如在传染性疾病的治疗中，个体i的结果变量会受到其他个体处理的影响。引入了这些符号，就可以定义个体的因果作用，如 $Y_{ki}^1 - Y_{ki}^2$ 表示处理 1 和 2 对于i个体的结果变量 Y_k(Y 的第 k 个分量)的因果作用。

但是，可惜的是，对于任何一个个体 i，上面的个体因果作用中的量都不可能完全被观测；这就是因果推断本质的困难。所以，在 Rubin 的分析框架中，因果推断根本上属于缺失数据统计分析的范畴。因此，分析框架

中还需要引入缺失机制的指示变量 M，其中 $M=1$ 表示被观测，$M=0$ 表示缺失。实际中，如果 $W_i=t$，则 $M_{ki}^j=0, \forall j\neq i, k=1,...,d$。

记所有的变量为 (X,Y,W,M)，其中 (X,Y) 部分观测，(W,M) 完全观测。在一次特定的试验中，观测到的 W 和 M 记为$\tilde{W}$和$\tilde{M}$；而$\tilde{X}=(X_{(0)},\tilde{X}_{(1)}),\tilde{Y}=(Y_{(0)},\tilde{Y}_{(1)})$，其中 $(X_{(0)},Y_{(0)})$ 表示缺失数据，$(\tilde{X}_{(1)},\tilde{Y}_{(1)})$ 表示观测到的数据。所有数据的联合分布是

$$f(X,Y,W,M)=f(X,Y|\pi)k(W|X,Y,\pi)g(M|X,Y,W,\pi),$$

其中 π 是未知的参数，$f(X,Y|\pi)$ 表示给定参数 π 后 (X,Y) 的联合分布，$k(W|X,Y,\pi)$ 表示安排处理的机制(assignment mechanism)，$g(M|X,Y,W,\pi)$ 表示数据缺失或者数据记录的机制(recording mechanism)。

因果推断目的，就是建立观测数据与缺失数据的关系，从而可以通过观测的数据推断缺失数据。为了建立观测数据与完全数据的关系，我们通常需要做如下的假定：

假定1：给定参数π之后，(X,Y) 的各行 iid，即

$$f(X,Y\mid\pi)=\prod_{i=1}^{N}f_*((X_i,Y_i)\mid\pi).$$

假定2：安排处理的机制可忽略(ignorable)，即

$$k(\tilde{W}\mid\tilde{X},\tilde{Y},\pi)=k(\tilde{W}\mid\tilde{X}_{(1)},\tilde{Y}_{(1)}).$$

假定3：数据记录的机制可忽略，即

$$g(\tilde{M}\mid\tilde{X},\tilde{Y},\tilde{W},\pi)=g(\tilde{M}\mid\tilde{X}_{(1)},\tilde{Y}_{(1)},\tilde{W}).$$

注意，Rubin(1978) 中的记号与他后来关于缺失数据的分析框架有所不同，在那里除了将脚标 (1) 和 (0) 换成 obs 和 mis 以外，还引入了有关两个机制的参数，而且可忽略性的假定还与各组参数的先验分布有关，见 Little 和 Rubin(2002)。

贝叶斯因果推断的本质任务是“填补(或预测)”$Y_{(0)}$，即

$$\begin{aligned}&Pre(Y_{(0)}\mid\tilde{X}_{(1)},\tilde{Y}_{(1)},\tilde{W},\tilde{M})\\=\;&\frac{\int\int p(\pi)f(\tilde{X},\tilde{Y}\mid\pi)k(\tilde{W}\mid\tilde{X},\tilde{Y},\pi)g(\tilde{M}\mid\tilde{X},\tilde{Y},\tilde{W},\pi)d\pi dX_{(0)}}{\int\int\int p(\pi)f(\tilde{X},\tilde{Y}\mid\pi)k(\tilde{W}\mid\tilde{X},\tilde{Y},\pi)g(\tilde{M}\mid\tilde{X},\tilde{Y},\tilde{W},\pi)d\pi dX_{(0)}dY_{(0)}}\end{aligned}$$

用上**假定2**和**假定3**，上式变成

$$= \frac{\int\int p(\pi)f(\tilde{X},\tilde{Y}\mid\pi)k(\tilde{W}|\tilde{X}_{(1)},\tilde{Y}_{(1)})g(\tilde{M}\mid\tilde{X}_{(1)},\tilde{Y}_{(1)},\tilde{W})}{\int\int\int p(\pi)f(\tilde{X},\tilde{Y}\mid\pi)k(\tilde{W}\mid\tilde{X}_{(1)},\tilde{Y}_{(1)})g(\tilde{M}\mid\tilde{X}_{(1)},\tilde{Y}_{(1)},\tilde{W})d\pi dX_{(0)}dY_{(0)}}$$
$$= \frac{\int\int p(\pi)f(\tilde{X},\tilde{Y}\mid\pi)d\pi dX_{(0)}}{\int\int\int p(\pi)f(\tilde{X},\tilde{Y}\mid\pi)d\pi dX_{(0)}dY_{(0)}}$$

根据这个式子，可以看出在一定得假设下，因果推断只决定于观测变量以及上面给定 π 关于 (X,Y) 的模型，完全忽略掉了安排处理机制和数据记录机制(这就是“可忽略性”名字的由来)。

进一步的推导，上式可以写成

$$\frac{\int h(\tilde{Y}\mid\tilde{X}_{(0)},\pi)q(\pi\mid\tilde{X}_{(1)})d\pi}{\int\int h(\tilde{Y}\mid\tilde{X}_{(0)},\pi)q(\pi\mid\tilde{X}_{(1)})d\pi dY_{(0)}},$$

其中

$$h(Y|X_{(1)},\pi)=\frac{\int f(X,Y\mid\pi)dX_{(0)}}{\int\int f(X,Y\mid\pi)dX_{(0)}dY}$$

表示给定 X 的观测数据及参数 π 下，Y 的分布。

$$q(\pi\mid X_{(1)})=\frac{p(\pi)\int\int f(X,Y\mid\pi)dX_{(0)}dY}{\int\int\int p(\pi)f(X,Y\mid\pi)dX_{(0)}dYd\pi}$$

表示给定 X 的观测数据下，参数 π 的分布。

当安排处理机制是完全随机化的条件下，机制是可忽略的；但是在一般的观测性研究中(比如吸烟与否)，安排处理机制很难满足可忽略性。可见，随机化试验在因果推断中往往起到至关重要的作用的。在上一部分，我们在频率学派的观点下得到了平均因果作用的估计。但是在这样的贝叶斯框架下，因果作用的识别需要更多的假定！模型 $h(Y\mid X_{(1)},\pi)$ 是关于 Y 所有潜在结果**联合分布**的模型；而在频率学派下，我们只需要算两个边缘分布对应的均值 $E(Y_i(1))$ 和 $E(Y_i(0))$。在实际的统计推断中，更多的假定意味着模型的稳定性会很差(某些假定的破坏甚至会带来灾难性的后果)。当然，这并不是贝叶斯方法用于因果推断的特有问题。事实上，贝叶斯方法往往需要对整个模型的联合分布进行参数化，而频率学派为了追求估计的稳健性可以采用矩估计或者广义矩估计等等方法。这是贝叶斯方法不足的地方。当然在 Rubin(1978) 中，给出了一条结论：上面的随机化试验中，在可忽略性的假定下，当样本量趋于无穷的时候，贝叶斯方法得到的结果渐近与频率学派等价；这意味着，当设定的 $(Y(1),Y(0))$ 的边缘分布时，联合分布对于最后的结果影响将渐近消失。

5.2 不依从和死亡删失下贝叶斯因果推断

离开了前面提到的假定，$CACE$ 是不可识别的。但是 Imbens and Rubin(1997) 坚持认为，即使这样的模型不可识别，但是数据中还是含有参数的信息，因此参数的后验分布还是得到了更新。因此，Imbens and Rubin(1997) 就在这样的思想指导下，不加**假定 4** 和**假定 5** 给出了 $CACE$ 的贝叶斯估计方法，并且声称这是贝叶斯方法至于频率学派的优点。

如图 8 所示的的是整个问题人群分布，按照 Z 和 D 可以将人群分成四组，而事实上每一组又是潜在的两个主分层的混合。因此，本质上，这里的问题就是混合分布的识别问题。混合分布“本质上”时不可以识别的，因此由此也可以看出 Imbens and Rubin(1997) 采用贝叶斯分析框架潜在的问题。我们知道，对于混合的高斯分布，除了几个成分的顺序以外，模型是可以识别的；但是对于离散型变量的混合，其识别性根本无法保证。可以断言，Imbens and Rubin(1997) 对于结果变量为离散型随机变量无能为力。

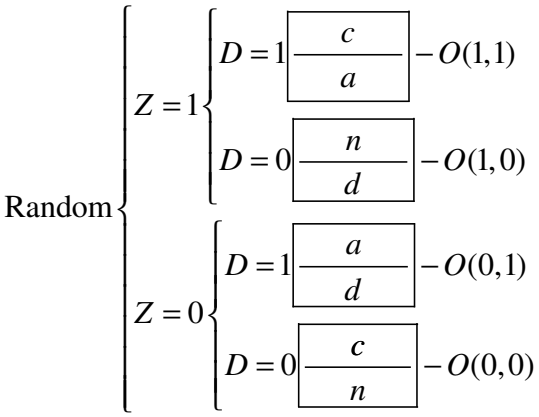

图 8: 不依从情形下人群分层

设 g 表示每个潜在主分层对应的Y的密度函数(一共八个 $g_{CZ}, C = c, n, a, d; Z = 1, 0.$)，$\pi$ 表示整个模型的参数(包含各个主分层的比例 $w_C(C = c, n, a, d)$ 和八个密度函数含的参数)。进过简单的推导或者从

图 8直观的可以得到参数 π 的后验分布如下：

$$\begin{aligned} & P(\pi \mid Z_{obs}, D_{obs}, Y_{obs}) \\ \propto \quad & p(\pi) \\ & \cdot \prod_{i \in O(1,1)} \left(w_c g_{c1}^i + w_a g_{a1}^i\right) \\ & \cdot \prod_{i \in O(1,0)} \left(w_n g_{n1}^i + w_d g_{d1}^i\right) \\ & \cdot \prod_{i \in O(0,1)} \left(w_a g_{a0}^i + w_d g_{d0}^i\right) \\ & \cdot \prod_{i \in O(0,0)} \left(w_c g_{c0}^i + w_n g_{n0}^i\right). \end{aligned}$$

而

$$CACE = \int y g_{c1}(y) dy - \int y g_{c0}(y) dy,$$

因此频率学派用 EM(Expectation-Maximization) 算法，或者贝叶斯学派用 DA(Data Augmentation) 算法都可以得到 $CACE$ 的估计。

对于频率学派通常加的假定，在贝叶斯分析框架下很容易得实现；在这个等价的设某些参数为零或者某些参数相等则对应频率学派的两个假定。

当然，根据前面的参数模型，我们可以写出死亡删失情况下所有参数 θ 的后验分布如下：

$$\begin{aligned} & p(\theta \mid X, Z, S_{obs}, Y_{obs}) \\ \propto \quad & p(\theta) \\ & \cdot \prod_{i \in O(1,1)} \left[\pi_{LL:i} N_i(\mu_{LL,1} + \eta_{LL,1}^T X_i, \sigma_{LL,1}^2) + \pi_{LD:i} N_i(\mu_{LD,1} + \eta_{LL,1}^T X_i, \sigma_{LD,1}^2)\right] \\ & \cdot \prod_{i \in O(1,0)} \left(\pi_{DL:i} + \pi_{DD:i}\right) \\ & \cdot \prod_{i \in O(0,1)} \left[\pi_{LL:i} N_i(\mu_{LL,0} + \eta_{LL,0}^T X_i, \sigma_{LL,0}^2) + \pi_{DL:i} N_i(\mu_{DL,0} + \eta_{DL,1}^T X_i, \sigma_{DL,0}^2)\right] \\ & \cdot \prod_{i \in O(0,0)} \left(\pi_{LD:i} + \pi_{DD:i}\right) \end{aligned}$$

其中先验分布的选取可以和通常的线性回归一样，保证共轭性。

如果需要加一些假定，比如单调性假定(某些层不存在)则直接设定某些参数为零即可，不影响整个贝叶斯分析的框架。

6 因果图(Causal diagram)：do 操作、d 分离、后门准则与前门准则

这部分介绍 Pearl(1995) 发表在 Biometrika 上的工作。

6.1 基本概念

在一个 DAG 中，若所有的节点集合为 $(X_1, ..., X_n)$，则所有变量的联合分布可以有如下的递归分解：

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i \mid pa_i),$$

其中 pa_i 表示 X_i 的父亲集合。要在 DAG 上引入因果的概念，则需要 do 操作，do 的意思可以理解成“干预”(intervention)。在 DAG 中 $do(X_i) = x_i'$ (也可以记做 $\check{x_i'}$)，表示 DAG 指向 X_i 的有向边全部被切断，而 X_i 的取值固定为常数 x_i'，这样得到的新 DAG 的联合分布可以记做 $P(x_1, ..., x_n \mid do(X_i) = x_i')$，注意这不是条件分布！可以证明，干预后的联合分布为

$$P(x_1, ..., x_n \mid do(X_i) = x_i') = \frac{P(x_1, ..., x_n)}{P(x_i \mid pa_i)} I(x_i = x_i').$$

根据 do 操作，便可以定义因果作用，比如二值的变量 Z 对于 Y 的平均因果作用定义为

$$ACE(Z \to Y) = E(Y \mid do(Z) = 1) - E(Y \mid do(Z) = 0),$$

上面 do 操作下的期望，分别对应 do 操作下的分布。这样在 do 操作下定义的因果模型，被著名计量经济学家 Halbert White 称为 Pearl Causal Model(PCM)(White and Chalak, 2009)。Pearl 在其书中写到：“I must take the opportunity to acknowledge four colleagues who saw clarity shining through the do(x) operator before it gained popularity: Steffen Lauritzen, David Freedman, James Robins and Philip David. Phil showed special courage in pringting my paper in Biometrika, the journal founded by causality's worst adversary- Karl Pearson.” (Pearl, 2000) 在书中他还论述了 RCM 和 PCM 的等价性。

6.2 d 分离，后门准则和前门准则

d 分离的定义：设 X，Y，Z 是 DAG 中不相交的节点集合，p 为一条连接 X 中某节点到 Y 中某节点的路径(不管方向)。称 Z 阻断 (block) 路径 p，如果路径 p 上某节点满足如下的条件：

(1) 在路径 p 上，w 点处为 V 结构(或称冲撞点，collider)，且 w 及其后代不在 Z 中；

(2) 在路径 p 上，w 点处不是 V 结构，且 w 在 Z 中。

进一步，称 Z d 分离 X 和 Y，记为 $(X \perp Y|Z)_G$，当且仅当 Z 阻断了 X 到 Y 的所有路径。

下面介绍 Pearl(1995) 的主要工作：后门准则和前门准则。

后门准则：在DAG中，称变量的集合Z相对于变量的有序对 (X_i, X_j) 满足后门准则，如果

(1) Z 中节点不能使 X_i 的后代；

(2) Z 阻断了 (X_i, X_j) 之间所有指向 X_i 的路径(这样的路径可以称为后门路径)。

进一步，称变量的集合Z相对于节点集合的有序对 (X, Y) 满足后门准则，若Z相对于变量的有序对 (X_i, X_j) 满足后门准则，其中 X_i 和 Y_j 是 X 和 Y 中的任意节点。

Pearl(1995) 证明，若存在一个变量集合 Z 相对于 (X, Y) 满足后门准则，那么 X 到 Y 的因果作用是可以识别的，且

$$P(y|do(X) = x) = \sum_z P(y \mid x, z)P(z).$$

从上面可以看出，上面的后门准则和可忽略性假定下 ACE 的识别公式一样：都是用 Z 做调整(adjustment)，先分层再加权求和。这条结论在 Rosenbaum and Rubin(1983) 之后，且流行病学家也都用这样的调整方法控制混杂因素，因此并不新奇。比较新颖的结论是下面的前门准则。

前门准则：在 DAG 中，称节点的集合 Z 相对于有序对 (X, Y) 满足前门准则，如果

(1) Z 切断了所有 X 到 Y 的直接路径；

(2) X 到 Z 没有后门路径；

(3) 所有 Z 到 Y 的后门路径都被 X 阻断。

此时，如果 $P(x,z)>0$，则 X 到 Y 的因果作用可识别，为

$$P(y \mid do(X)=x)=\sum_{z} P(z \mid x) \sum_{x'} P(y \mid x', z) P(x').$$

下面给出一个例子(图 9)。

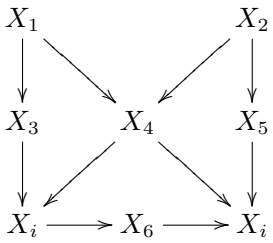

图 9: Back-door criterion and Front-door criterion

上面的 DAG 中，(X_i, X_j) 之间的后门路径被 $\{X_3, X_4\}$ 或者 $\{X_4, X_5\}$ 阻断，而前门路径被 X_6 阻断。上面的两个准则表明，要识别从 X_i 到 X_j 的因果作用，我们不需要观测到左右的变量，只需要观测到切断后门路径或者前门路径的变量即可。

7 未涉及的问题

因果推断方面还有很多其他的方面。比如关于混杂因素的研究(Geng et al, 2002)，替代指标以及代理悖论(Chen, Geng and Jia, 2007; Ju and Geng, 2009)，直接作用和间接作用(Rubin, 2004)，交互作用(Rothman et al, 2008)，以及 DAG 的学习问题(Pearl, 2000)等。

另外，这里只介绍了 RCM 和 PCM，文献中还有一类因因果模型–充分原因模型(Sufficient-Cause Model)–并未涉及，这类模型常常用来讨论交互作用(Rothman et al, 2008)。

参考文献

[] Amemiya,T., 1985, Advanced Econometrics, Cambridge, MA: Harvard University Press.

[] Angrist, J.D., Imbens, G.W. and Rubin,D.B., 1996, Identification of Causal Effects Using Instrumental Variables, *Journal of the American Statistical Association*.

[] Bickel,P.J. et al, 1975, Sex Bias in Graduate Admissions: Data from Berkeley, *Science*.

[] Chen, H., Geng, Z. and Jia, J.,2007, Criteria for surrogate end points.*J. Royal Statist. Soc. B*.

[] Chen Hua ,Geng Zhi and Zhou X.H., 2009, Identifiability and Estimation of Causal Effects in Randomized Trials with Noncompliance and Completely Non-ignorable Missing-Data, *Biometrics*.

[] Frangakis, C.E., and Rubin, D.B., 2002, Principal Stratification in Causal Inference, *Biometrics*.

[] Frangakis, C.E., and Rubin, D.B., 1999, Addressing complications of intention to treat analysis in the combined presence of all or noncompliance and subsequence missing outcomes, *Biometrika*.

[] Geng, Z.， Guo, J. H. and Fung W. K., 2002, Criteria for confounders in epidemiological studies, *J. Royal Statist. Soc. B*.

[] Goodman, L.A., 1974, Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika*.

[] Heckman, J. J. ,1979, Sample Selection Bias as a Specification Error, *Econometrica*.

[] Hirano, K., Imbens, G.W. and Ridder. G., 2003, Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score, *Econometrica*.

[] Imbens, G.W. and Angrist J.D., 1994, Identification and Estimation of Local Average Treatment Effects, *Econometrica*.

[] Imbens, G.W. and Rubin,D.B., 1997, Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance, *The Annals of Statistics*.

[] Ju, C. and Geng, Z. ,2009, Criteria for surrogate endpoints based on causal distributions. To appear in *J. Royal Statist. Soc. B*.

[] Pearl,J., 1996, Causal Diagrams for Empirical Research, *Biometrika*.

[] Pearl,J., 2000, Causality, Cambridge University Press.

[] Rothman et al, 2008, Modern Epidemiology, Lippincott Williams& Wilkins.

[] Rubin, D.B., 1978, Bayesian Inference for Causal Effects: The Role of Randomization, *The Annals of Statistics*.

[] Rubin, D.B., 2004, Direct and Indirect Causal Effects via Potential Outcomes, *Scandinavian Journal of Statistics*.

[] Shao, J., 2003, Mathematical Statistics, 2nd Edition, Springer.

[] White, H. and Drive, G., 2009, Settable Systems: An Extension of Pearl' s Causal Model with Optimization, Equilibrium, and Learning, *Journal of Machine Learning Research*.

[] Wooldrige, 2002, Econometric Analysis of Cross Section and Panel Data, MIT Press.

[] Zhang, J. L., Rubin, D. B., and Mealli, F. ,2008, "Evaluating the Effects of Job Training Programs on Wages through Principal Stratification," in Volume 21 of Advances in Econometrics: Modeling and Evaluating Treatment Effects in Econometrics, ed. by Millimet, D. L., Smith, J. A. and Vytlacil, E., pp. 117-145. Elsevier Science.