arXiv:2302.03740v1 [stat.ME] 7 Feb 2023

不完整数据的敏感性分析通过未测量混杂

预印本

Heng Chen生物统计学系 吉利德科学公司
Henry.Chen10@gilead.com

Daniel F.Heitjan统计科学系，南卫理公会大学 人口与数据科学系，德克萨斯大学西南医学中心 dheitjan@smu.edu

2023年2月9日

摘要

我们提出一种方法，用于分析频率学派推断对缺失机制可能存在非可忽略性的敏感程度。此类分析通常从选择模型入手，而我们假设缺失是由未测量混杂造成的。我们的模型可以构建敏感性度量，其形式类似于观察性研究中针对未测量混杂的度量。我们定义一个敏感性指数，记为 *MinNI*，即改变感兴趣估计量的均值达到指定幅度所需的最低非可忽略程度。我们将该模型应用于比例的敏感性分析，但这一思路很容易推广到更复杂的情形。

关**键词** 最小非可忽略性指数；非可忽略性；敏感性分析；未测量混杂。

1 引言

Rubin 阐明了缺失机制在从不完整数据中进行频率学派推断时所起的作用。（Rubin，1976）其思路是比较在给定观测到的缺失指示变量时，已观测变量的分布与忽略缺失机制时这些变量的边缘分布。条件完全随机缺失（*MCAR*）足以保证这两个分布相同，因此缺失机制是可忽略的。（Little 和 Rubin，2014）简言之，MCAR 要求在给定假定的完整数据时，观测到的缺失指示变量的条件概率独立于完整数据的取值。Heitjan 将这一分析扩展到粗化数据模型，其中 MCAR 推广为粗化后完全随机。（Heitjan，1994年、1997年）

Rubin 方法从选择模型入手；也就是说，它将感兴趣变量 Y 与缺失指示变量 G 的联合分布参数化为： Y 的边缘分布乘以给定 Y时 G 的条件分布，并将后者称为缺失机制。随后，可以将可忽略性条件识别为对缺失机制参数的限制。若不作出强假设，就无法估计这一联合模型的参数。（Diggle 和 Kenward，1994）另一种较为审慎的做法是进行敏感性分析：在一系列关于非可忽略性参数的假设下，重新估计这些参数，以评估基于可忽略模型所得估计的稳健性。（Copas 和 Li，1997；Verbeke 等人，2001；Troxel 等，2004年；Ma 等，2005年）

在选择模型中，缺失概率被描述为潜在缺失观测值的函数；如果缺失与结局相关（通常指非可忽略性参数不为零），可忽略性便不成立。实际中，或许更适合将这种相关性视为由混杂所致——事实也可能确实如此——即结局和缺失指示变量都与第三个变量相关。如果我们能够识别并测量这个变量，就会有某种形式的条件独立性成立，从而保证可忽略性。如果无法做到，我们便对其形式作出假设，并考察非可忽略性对可测量结局分布的影响。此类模型长期以来一直是观察性研究中敏感性分析的基础；在这类研究中，人们担忧处理指示变量与潜在结局之间存在相关性，进而使标准因果分析产生偏倚。（Cornfield 等，1959年；Rosenbaum 和 Rubin，1983a；Lin 等，1998年；Mitra 和 Heitjan，2007年；Ding 和 VanderWeele，2016；VanderWeele 和 Ding，2017）本文将这种方法应用于不完整数据。

第2节介绍模型并确立一般的可忽略性条件。第3节提出一种响应曲面方法，用于评估感兴趣参数随参数模型中的非可忽略性参数变化的情况。第4节借鉴 Cornfield范式，将产生指定偏倚水平所需的最小非可忽略性程度定义为敏感性指数。第5节将这些方法应用于一项性行为研究的不完整数据。第6节介绍该方法的扩展，进一步的讨论见第7节。

2 模型与方法

2.1 模型与定义

数据由结局变量 $Y=(Y_1,\ldots\ldots,Y_n)$ 及其对应的缺失指示变量向量$G=(G_1,\ldots\ldots,G_n)$组成，其中 $G_i=1$ 表示 Y_i 被观测到，而 $G_i=0$ 表示 Y_i 缺失。假设未测量变量$U=(U_1,\ldots\ldots,U_n)$充当混杂因素，即给定 U 时， Y和 G 条件独立；也就是说，给定 U时， Y和 G的条件分布具有如下性质：对于任意 u、 $f^{Y,\,G|U}(y,g|u)=f^{Y|U}(y|u)f^{G|U}(g|u)$，对所有 y 和 g均成立。因此，联合密度可简化为

$$f^{Y,G,U}(y,g,u)=f^{U}(u)f^{Y|U}(y|u)f^{G|U}(g|u). \tag{1}$$

如果混杂未被测量或未加以控制，就可能在 Y与 G之间诱发相关性。因此，尽管我们试图对 Y的边缘分布进行推断，

$$f^{Y}(y)=\int f^{Y|U}(y|u)f^{U}(u)du, \tag{2}$$

实际上，我们可能只能观测到在给定 $G=g$条件下 Y的条件分布：

$$f^{Y|G}(y|g)=\frac{\int f^{G|U}(g|u)f^{Y|U}(y|u)f^{U}(u)du}{\int f^{G|U}(g|u)f^{Y|U}(y|u)f^{U}(u)dudy}. \tag{3}$$

为此，我们对式 (3) 中的条件分布项建立限制条件，以保证可忽略性；在此语境下，这意味着方程(2)和(3)相同。全文中，我们忽略关于测度为 0 的集合的任何理论考量。

定理1 假设 $G\perp\!\!\!\perp U$或 $Y\perp\!\!\!\perp U$成立。那么，对于任意满足 $f^{G}(g)>0$的 g，忽略式 *(2)* 中缺失机制所得的分布等于式 *(3)* 中的正确分布。

证明。假设 $G\perp\!\!\!\perp U$。那么， $\forall g$，其中 $0<f^{G}(g)<1$,

$$f^{Y|G}(y|g)=\frac{\int f^{Y|U}(y|u)f^{G}(g)f^{U}(u)du}{\int f^{Y|U}(y|u)f^{G}(g)f^{U}(u)dudy}=\frac{\int f^{Y|U}(y|u)f^{U}(u)du}{\int f^{Y|U}(y|u)f^{U}(u)dudy}=f^{Y}(y).$$

类似地，如果 $Y\perp\!\!\!\perp U$，那么 $\forall g$，其中 $0<f^{G}(g)<1$,

$$f^{Y|G}(y|g)=\frac{f^{Y}(y)\int f^{G|U}(g|u)f^{U}(u)du}{\int f^{G|U}(g|u)f^{U}(u)du}=f^{Y}(y).$$

如果 Y或 U是离散的，则可以将定理改述为用求和代替积分。定理中的可忽略性条件比完全随机缺失更强，因为它适用于所有具有正密度的可能缺失模式，而不仅仅是观测到的缺失模式。（Heitjan，1997；Little 和 Rubin，2014）在实践中，相关的条件分布

将是给定观测到的缺失指示变量向量 $\tilde{g}$ 条件下的结局 y 分布，其中 y 的样本空间将被限制为那些可能与 $\tilde{g}$ 一致的取值，称为与 $\tilde{g}$一致。因此，我们提出定理1的一个替代性较弱版本。

考虑以下条件，假设 $\tilde{g}$ 是 g的一个观测值：

1. 缺失是观测可忽略 的，即对于任意可能的 u, $f^{G|Y,\ U}(\tilde{g}|y,\ u)$ 对所有与 $\tilde{g}$一致的 y都取相同的值。
2. $f^{G|U}(\tilde{g}|u)$ 对所有 u都取相同的值。3. 对于任意与 $\tilde{g}$一致的 y， $f^{Y|U}(y|u)$ 对所有 u都取相同的值。

由此得到以下定理：

定理2 在 假设 *1* 以及假设 *2* 或 *3* 中的任一项， $f^{Y}(y) = f^{Y|G}(y|\tilde{g})$ 对于所有与 $\tilde{g}$.

在实践中，也可能存在完全测量的预测变量。在这种情况下，只需进行适当的条件化，这些定理仍然成立，如第6.3节所示。这些定理为混杂模型提供了最简单的一般可忽略性条件。可忽略性通常无法检验，因为 U 通常是假设性的，而 Y_i 只有在 $G_i = 1$时才可获得。因此，我们的思路是在方程（1）的背景下定义不可忽略性参数，然后调整这些参数，以确定它们偏离可忽略模型多远，才会使 $f^{Y}(y)$与 $f^{Y|G}(y|\tilde{g})$之间产生显著差异。

2.2 混杂模型中的敏感性分析

为简化阐述，我们考虑这样一种情形：数据代表 n 独立同分布的案例，其中 U 是一个标量未测量混杂因素。假设 Y具有有限均值和方差，我们首先基于 Y的边际均值与其在被观测条件下的均值进行比较，以开展敏感性分析。

定理1和定理2表明，敏感性参数可以表示 U 与 Y之间以及 U与 G之间的关联。我们将介绍两种方法：第一种方法基于弱参数化模型，在合理范围内改变非可忽略性参数，以传统方式描述偏倚。这将 Rosenbaum 和 Rubin 对观察性研究中混杂问题的敏感性分析扩展至不可忽略缺失数据。（Rosenbaum 和 Rubin，1983a）第二种方法考虑一个最小参数化非参数模型，并将最小非忽略性指数（*MinNI*）定义为造成不可忽略偏倚所需的非可忽略性程度。这与观察性研究中敏感性分析的 Cornfield方法类似。（Cornfield 等，1959）与所有敏感性分析一样，原则上，我们的方法依赖于一位假想专家的判断，其职责是确定 Y中偏倚和非可忽略性参数各自达到不可忽略程度所需的最小值。

3 响应面敏感性分析

3.1 敏感性参数

首先假设 混杂因素 U是二元的。然后，我们部分指定（Y， G， U）的联合分布 a s

$$\Pr[U = 0] = \pi_0,$$

$$\Pr[G = 1|U = u] = h(\gamma_0 + \gamma_1 u),$$

$$\mathbb{E}[Y|U = u] = q(\beta_0 + \beta_1 u),$$

其中， $u \in \{0, 1\}$ 和 $h(\gamma_0 + \gamma_1 u)$ 以及 $q(\beta_0 + \beta_1 u)$ 是连接函数。参数 π_0、 γ_1 和 β_1 描述敏感性程度；由于我们无法观测到 U，它们是不可识别的。

敏感性分析的一种标准方法，是观察关注参数在敏感性参数于合理取值范围内变化时的变化；在本例中，该关注参数为边际均值 $\mathbb{E}[Y]$。在此模型下，用这些敏感性参数表示的 Y的边际均值和条件均值分别为：

$$\mathbb{E}[Y] = q(\beta_0 + \beta_1)(1 - \pi_0) + q(\beta_0)\pi_0 \tag{4}$$

$$\mathbb{E}[Y|G = 1] = \frac{q(\beta_0 + \beta_1)h(\gamma_0 + \gamma_1)(1 - \pi_0) + q(\beta_0)h(\gamma_0)\pi_0}{h(\gamma_0 + \gamma_1)(1 - \pi_0) + h(\gamma_0)\pi_0} \tag{5}$$

3.2 使用指定的不可忽略性参数估计均值

在此模型下，$\Pr[G=1]$ 和 $\mathbb{E}[Y|G=1]$ 可分别根据数据直接估计为 $\hat{p}$ 和 $\hat{\mu}_c$。在固定 π_0、γ_1和 β_1，并观察到 Y 个值的随机样本（其中部分值可能缺失）的情况下，可以很容易地估计γ_0 和 β_0。（Rosenbaum 和 Rubin，1983a）第一个可估计项为

$$\Pr[G=1] = h(\gamma_0+\gamma_1)(1-\pi_0) + h(\gamma_0)\pi_0. \tag{6}$$

因此，我们有两个方程（（5）和（6））以及两个未知数（γ_0 和 β_0）。我们按如下方式计算（4）中的边际均值:

1. 在固定 $\hat{p}$ 和 π_0, γ_1 的条件下，求解方程（6）以得到 $\hat{\gamma}_0$；2. 在固定 π_0, γ_1, β_1 的条件下，求解方程（5）以得到 β_0，其中 $\hat{\gamma}_0$ 来自步骤1，$\hat{\mu}_c$ 则直接根据数据估计；$\hat{}$ 3. 将 β_0 和 π_0, β_1 代入（4），以估计边际均值。

附录8.1给出了两个连接函数均为逻辑函数这一特殊情形的详细内容，该情形适用于二元结局。

定理1和定理2断言，如果 $\gamma_1 = 0$ 或 $\beta_1 = 0$，那么 $\mathbb{E}[Y]$ 和 $\mathbb{E}[Y|G=1]$之间没有差异。因此，如果这些参数取值较小时，$\mathbb{E}[Y]$会出现显著变化，我们就认为结果对这些参数敏感。如果 $\mathbb{E}[Y]$ 关于敏感性参数的响应曲面是平坦的，那么只有当敏感性参数取值较大时，$\mathbb{E}[Y|G=1]$才会发生不可忽略的变化，因此推断对这些参数不敏感。

如果假定的未测量协变量 U 不是二元变量，则对 U 分布的设定会更加复杂，并且可能涉及更多参数。给定 U 的 G 分布以及给定 U 的 Y 分布，均由连接函数进行索引，而对连接函数的设定会引入另一个敏感性来源。因此，对于此应用，半参数模型或非参数模型可能更为令人满意。

4 识别最小可忽略非可忽略性

响应面分析通过描绘非可忽略性对 Y 分布的影响，直接考察偏倚。与之互补的方法是确定能够产生指定变化程度的敏感性参数的最小值——在本例中，即 $\mathbb{E}[Y|G=1]$ 与 $\mathbb{E}[Y]$ 之间预先设定的最大可忽略差异。我们将这些参数值记为 *MinNI*，即最小*Min*imum *N*on*I*gnorability（最小非可忽略性）。此外，我们力求使用参数数量最少的模型开展分析。

4.1 差异尺度中的 MinNI

再次假设存在一个二元混杂因子 U。我们首先注意到

$$\mathbb{E}[Y] - \mathbb{E}[Y|G=1] = (\mathbb{E}[Y|G=0] - \mathbb{E}[Y|G=1])\Pr[G=0]. \tag{7}$$

显然，除非 $0 < \Pr[G=0] < 1$，否则无需进行敏感性分析。将公式（7）中的偏倚按未测量混杂因素 U 展开后，我们发现，$\mathbb{E}[Y|G=0]$ 与 $\mathbb{E}[Y|G=1]$ 之间的差异可以分解为 $\mathbb{E}[Y|U=1]$ 与 $\mathbb{E}[Y|U=0]$ 之间的差异和 $\Pr[U=1|G=1]$与 $\Pr[U=1|G=0]$之间的差异的乘积。详见附录8.2。

将敏感性参数定义为以下两个差异

$$\text{ED}_{YU} = \mathbb{E}[Y|U=1] - \mathbb{E}[Y|U=0],\ \ \text{RD}_{UG} = \Pr[U=1|G=1] - \Pr[U=1|G=0],$$

并注意到

$$|\mathbb{E}[Y] - \mathbb{E}[Y|G=1]| = |\text{ED}_{YU}\text{RD}_{UG}\Pr[G=0]|\,. \tag{8}$$

我们可以 通过为偏倚指定一个可忽略差异的上限，构造一个不敏感区域:

$$|\mathbb{E}[Y] - \mathbb{E}[Y|G=1]| \le k\sigma_{Y|G=1}, \tag{9}$$

其中，$\sigma_{Y|G=1}$ 是给定观测到 Y 时其标准差，k 是根据具体情境定义的正值常数，可能与样本量有关。由（8）和（9）可得非可忽略参数的无差异区域为

$$|\text{ED}_{YU}\text{RD}_{UG}| \le \frac{k\sigma_{Y|G=1}}{\Pr[G=0]}. \tag{10}$$

不等式（10）描述了最大可容忍变化与敏感性参数之间的关系。为定义一个单一的敏感性指数，我们找出满足这一约束且距原点最近的敏感性参数组合。我们将其称为均值的 *MinNI*。对于连续型结果，优化过程为

$$
\begin{aligned}
\text{最小化:} \quad & (\text{ED}_{YU}^2 + \text{RD}_{UG}^2) \\
\text{约束条件:} \quad & |\text{ED}_{YU}\text{RD}_{UG}| \leq \frac{k\sigma_{Y|G=1}}{\Pr[G=0]}; \\
& |\text{ED}_{YU}| \in (0, \infty); \\
& |\text{RD}_{UG}| \in (0, 1).
\end{aligned}
$$

（$|\text{ED}_{YU}|$，$|\text{RD}_{UG}|$）的闭式可行解（即 MinNI）为

$$
\left(\max\left\{ \frac{k\sigma_{Y|G=1}}{\Pr[G=0]}, \sqrt{\frac{k\sigma_{Y|G=1}}{\Pr[G=0]}} \right\}, \min\left\{ 1, \sqrt{\frac{k\sigma_{Y|G=1}}{\Pr[G=0]}} \right\} \right).
$$

对于二元结局，优化过程中 $|\text{ED}_{YU}|$ 的取值范围为 (0，1)，且

$$
\text{MinNI} = \left(\sqrt{\frac{k\sigma_{Y|G=1}}{\Pr[G=0]}}, \sqrt{\frac{k\sigma_{Y|G=1}}{\Pr[G=0]}} \right),
$$

其中 $k\sigma_{Y|G=1} \leq \Pr[G=0]$。如果 MinNI 较大，则忽略缺失数据的抽样推断可能具有较好的稳健性。如果 MinNI 较小，忽略缺失机制可能会导致显著的偏倚。图1展示了对下文第5节所讨论示例的敏感性分析。

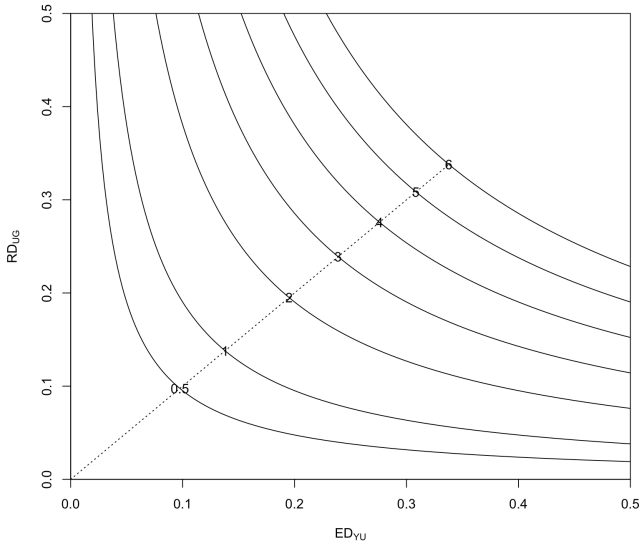


图1：性行为调查数据中 ED_{YU} 和 RD_{UG} 的等偏差图。曲线上的数字表示以标准误单位计的偏倚，相应的 MinNI 值（从左到右）为 (0.10, 0.10)、(0.14, 0.14)、(0.20, 0.20)、(0.24, 0.24)、(0.28, 0.28)、(0.31, 0.31) 和 (0.34,0.34)。

4.2 比率尺度上的 MinNI

对于分类变量　　，在比率尺度上描述偏倚可能更为合适。与公式（7）类似，我们　　e 观察到

$$
\frac{\mathbb{E}[Y]}{\mathbb{E}[Y|G=1]} = \Pr[G=1] + \Pr[G=0]\frac{\mathbb{E}[Y|G=0]}{\mathbb{E}[Y|G=1]}, \tag{11}
$$

其中 $\mathbb{E}[Y|G=1] \neq 0$。将非可忽略性参数定义为这些比值

$$
\text{ER}_{YU} = \frac{\mathbb{E}[Y|U=1]}{\mathbb{E}[Y|U=0]}, \quad \text{RR}_{UG} = \frac{\Pr[U=1|G=1]}{\Pr[U=1|G=0]},
$$

我们得到

$$\frac{\mathbb{E}[Y|G=0]}{\mathbb{E}[Y|G=1]} = \frac{\text{ER}_{YU} - 1 + \frac{1}{\Pr[U=1|G=0]}}{(\text{ER}_{YU} - 1)\text{RR}_{UG} + \frac{1}{\Pr[U=1|G=0]}}. \tag{12}$$

由于我们无法识别 $\Pr[U=1|G=0]$，因此所能做的最好就是求出方程 (12) 中比率的界。该比率的右侧是 $\frac{1}{\Pr[U=1|G=0]} \in (\text{RR}_{UG},\ \infty)$ 的单调函数。原始比率的界不等式表示为

$$\left|\frac{\mathbb{E}[Y]}{\mathbb{E}[Y|G=1]} - 1\right| \leq \left|\frac{(\text{ER}_{YU} - 1)(\text{RR}_{UG} - 1)}{\text{ER}_{YU}\text{RR}_{UG}}\right| \Pr[G=0], \tag{13}$$

其中 $\text{ER}_{YU} \in (-\infty,\ \infty)$ 和 $\text{RR}_{UG} \in (0,\ \infty)$。当 Y 为二元变量时，我们可以将 (9) 两边同时除以 $\mathbb{E}[Y|G=1]$，从而在比率尺度上指定一个无差异区域，得到

$$\left|\frac{\mathbb{E}[Y]}{\mathbb{E}[Y|G=1]} - 1\right| \leq |k\text{CV}_{Y|G=1}|. \tag{14}$$

这里，$\text{CV}^{Y|G=1}$ 是在观测到 Y 的条件下其变异系数。参数 $\Pr[G=0]$、$\sigma_{Y|G=1}$ 和 $\text{CV}^{Y|G=1}$ 均可由数据估计。为保守起见，我们令不等式 (13) 中比率的上界小于 (14) 中指定的可检测差异。不可忽略比率参数的无差异区域于是为

$$\left|\frac{(\text{ER}_{YU} - 1)(\text{RR}_{UG} - 1)}{\text{ER}_{YU}\text{RR}_{UG}}\right| \leq \frac{|k\text{CV}_{Y|G=1}|}{\Pr[G=0]} \tag{15}$$

为获得敏感性指数，我们找出距离 (1，1) 最近的点。假设 ER_{YU} 和 RR_{UG} 均大于 1（不失一般性），优化过程为

$$\begin{aligned}
\text{最小化:} \quad & (\text{ER}_{YU} - 1)^2 + (\text{RR}_{UG} - 1)^2 \\
\text{约束条件:} \quad & \frac{(\text{ER}_{YU} - 1)(\text{RR}_{UG} - 1)}{\text{ER}_{YU}\text{RR}_{UG}} \leq \frac{|k\text{CV}_{Y|G=1}|}{\Pr[G=0]}; \\
& \text{ER}_{YU} \in (1, \infty); \\
& \text{RR}_{UG} \in (1, \infty).
\end{aligned}$$

$(\text{ER}_{YU}, \text{RR}_{UG})$ 的闭式解为

$$\text{MinNI} = \left(\frac{1}{1 - \sqrt{\frac{|k\text{CV}_{Y|G=1}|}{\Pr[G=0]}}}, \frac{1}{1 - \sqrt{\frac{|k\text{CV}_{Y|G=1}|}{\Pr[G=0]}}}\right),$$

其中 $|k\text{CV}_{Y|G=1}| < \Pr[G=0] \leq 1$。其含义与差值尺度相同；参见图2。

5 爱丁堡性行为调查的敏感性分析

5.1 数据

研究人员于1993年对爱丁堡大学随机抽取的6,136名学生进行了调查。最关注的参数是对问题“你是否曾有过性交？”回答“是”的比例；其中2,308名学生（37.6%）拒绝回答。（Raab和Donnelly，1999年；Troxel 等人，2004年；Xie等，2018年）肯定回答的观测比例为0.7320，标准误为0.0072，用以估计 $\mathbb{E}[Y|G=1]$。令人担忧的是，与应答者相比，未应答者的性行为模式可能有所不同，从而在估计关注参数时产生偏倚。下面我们将针对这一比例进行敏感性分析。

5.2 响应曲面敏感性分析

表1显示了偏倚如何随敏感性参数 π_0、β_1和 γ_1变化。图3展示了在 π_0 固定为0.5时，γ_1和 β_1上的等偏差等高线图。我们将 $\pi_0 = 0.5$ 固定为此值，因为它似乎会产生最大的偏倚。

表1中，偏倚的绝对大小相对于估计参数而言较小，在相对尺度上的数值最高约为 3%。然而，对于统计推断而言，这种敏感性相当显著，因为最大偏倚约为名义标准误的 3 倍。图3中的等偏差图表明，γ_1 和 β_1 取中等值时，均值可能发生 2 个标准误的变化。因此，该分析表明，对曾有过性交的学生比例的估计对非可忽略性较为敏感。

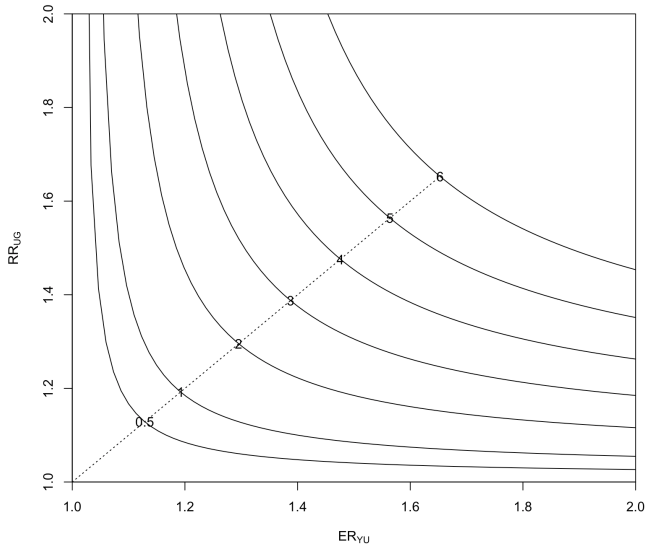


图2：性行为调查数据中 ER_{YU} 与 RR_{UG} 的等偏差图。曲线上的数字表示以标准误单位计的偏倚，对应的 MinNI 值（从左到右）为 (1.13, 1.13)、(1.19, 1.19)、(1.30, 1.30)、(1.39, 1.39)、(1.48, 1.48)、(1.56, 1.56) 和 (1.65, 1.65)。

表1： $\mathbb{E}[Y] - \mathbb{E}[Y|G=1]$ 作为敏感性参数的函数。

		π_0		
$\exp(\beta_1)$	$\exp(\gamma_1)$	0.1	0.5	0.9
2	2	−0.0037	−0.0088	−0.0025
	3	−0.0059	−0.0139	−0.0037
3	2	−0.0061	−0.0138	−0.0036
	3	−0.0097	−0.0218	−0.0053

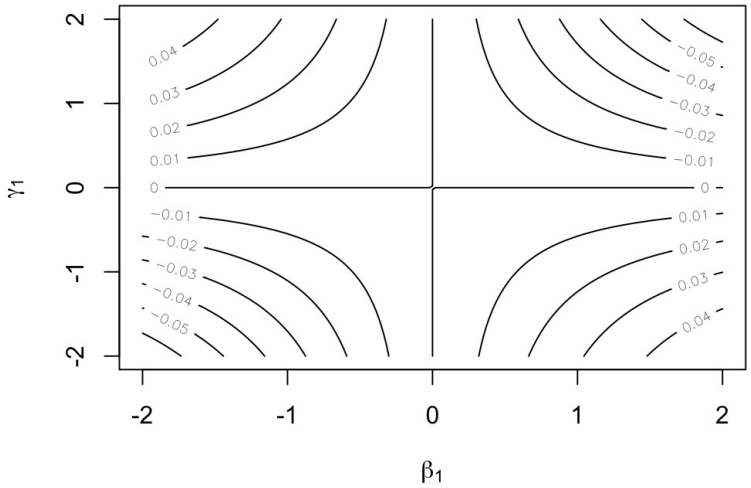


图3：固定 $\pi_0 = 0.5$时， $\mathbb{E}[Y] - \mathbb{E}[Y|G=1]$ 关于 γ_1 和 β_1的等值线。

5.3 MinNI敏感性分析与ISNI分析的比较

这里，我们将可忽略偏倚的最大值设为观测比例（此处为 $k\sigma_{Y|G=1} = 0.0072$）的 1 个标准误，并计算会产生这一程度偏移的敏感性参数最小值。差值尺度下的 MinNI，即 (ED_{YU},RD_{UG})=(0.14,0.14)（见图1），以及比率尺度下的 MinNI，即 (ER_{YU}, RR_{UG})=(1.19,1.19)（来自

图2。在这两种情况下，该指数都很小，表明关于发生过性交的真实比例的抽样推断很敏感。也就是说，即使可忽略模型受到轻微扰动，也可能导致我们的总体比例估计产生显著偏倚，使针对该参数的检验和置信区间不可靠。

我们将此分析与基于似然的ISNI（不可忽略性的局部敏感性指数）敏感性分析进行比较。（Troxel 等，2004年；Xie等，2018年）在ISNI中，关键的敏感性统计量记为 c，它衡量非可忽略性的近似最小标准化幅度，该幅度足以使关注参数的最大似然估计产生1个标准误的变化。通常认为，数值 $c < 1$ 表明存在敏感性。对于爱丁堡数据中回答“是”的比例，我们计算得到 $c = 0.097$，表明敏感性很强，这与我们的频率学分析一致。

5.4 MinNI 对缺失数据比例的依赖性

对非可忽略性的敏感性度量在很大程度上取决于缺失数据比例；事实上，对于存在缺失观测的单变量正态均值，ISNI 度量与缺失比例成正比。（Troxel 等，2004年）为说明这种关系，我们人为改变缺失观测的比例，同时保持肯定回答的观测比例不变。我们将人为设定的缺失比例分别设为 0.1 和 0.2（均低于观测值 0.376），并重复了分析。表2显示了偏倚如何随响应面敏感性参数变化，$\mathbb{E}[Y]$ 的偏倚也不例外。回想一下，响应比例的标准误为 0.0072；显然，当缺失数据比例较小时，除非混杂程度极端，否则敏感性较弱。

表2： $\mathbb{E}[Y] - \mathbb{E}[Y|G = 1]$ 随敏感性参数的变化，且 $\pi_0 = 0.5$。

		缺失比例		
$\exp(\beta_1)$	$\exp(\gamma_1)$	0.1	0.2	0.376
2	2	−0.0023	−0.0046	−0.0088
	3	−0.0035	−0.0071	−0.0139
3	2	−0.0035	−0.0072	−0.0138
	3	−0.0054	−0.0111	−0.0218

表3显示了不同缺失观测比例下，差值尺度和比率尺度敏感性分析的MinNI值。这些数值的含义是，只需较弱程度的混杂，即可在肯定回答的观测比例中引入不可忽略的偏倚。

表3：在不同缺失数据比例下产生一个标准误偏倚的MinNI。

	缺失比例		
尺度	0.1	0.2	0.376
($\lvert ED_{YU}\rvert$,$\lvert RD_{UG}\rvert$)	(0.27,0.27)	(0.19,0.19)	(0.14,0.14)
(ER_{YU},RR_{UG})	(1.46,1.46)	(1.28,1.28)	(1.19,1.19)

6 基本敏感性分析的若干扩展

6.1 分类混杂因素

将我们的分析扩展到具有 $m > 2$ 个水平的未测量混杂因素的情形非常简单。对于差值尺度，将混杂关系表示如下：

$$\mathrm{MD}_{YU} = \max_i \mathbb{E}[Y|U = u_i] - \min_i \mathbb{E}[Y|U = u_i],$$
$$\mathrm{MD}_{UG} = \max_i \left[\Pr\left[U = u_i|G = 1\right] - \Pr\left[U = u_i|G = 0\right]\right].$$

在附录 8.3.1 中，我们推导出的界不等式为

$$|\mathbb{E}[Y] - \mathbb{E}[Y|G = 1]| \leq |(m-1)\mathrm{MD}_{YU}\mathrm{MD}_{UG}\Pr[G = 0]|. \tag{16}$$

为保持保守，我们将差异的上界设为小于观测标准差 $\sigma_{Y|G=1}$ 的 k 个标准差，

$$|\mathrm{MD}_{YU}\mathrm{MD}_{UG}| \leq \frac{k\sigma_{Y|G=1}}{(m-1)\Pr[G = 0]}. \tag{17}$$

敏感性对类别数量的依赖关系与 Ding 和 VanderWeele（2014）的发现相同。对于相对比率尺度，我们将混杂参数表示为

$$\mathrm{ER}_{YU(i)} = \frac{\mathbb{E}[Y|U=u_i]}{\min_i \mathbb{E}[Y|U=u_i]},\ \ \mathrm{RR}_{UG(i)} = \frac{\Pr[U=u_i|G=1]}{\Pr[U=u_i|G=0]},$$

$$\mathrm{MR}_{YU} = \max_i \mathrm{ER}_{YU(i)},\ \ \mathrm{MR}_{UG} = \max_i \mathrm{RR}_{UG(i)}.$$

不失一般性，我们可以假设所有这些参数都大于 1。在附录 8.3.1 中，我们证明界不等式为

$$\left|\frac{\mathbb{E}[Y]}{\mathbb{E}[Y|G=1]} - 1\right| \leq \left|\frac{(\mathrm{MR}_{YU}-1)(\mathrm{MR}_{UG}-1)}{\mathrm{MR}_{YU}\mathrm{MR}_{UG}}\right|. \tag{18}$$

这导出了保守无差异区域

$$\frac{(\mathrm{MR}_{YU}-1)(\mathrm{MR}_{UG}-1)}{\mathrm{MR}_{YU}\mathrm{MR}_{UG}} \leq \frac{k\mathrm{CV}_{Y|G=1}}{\Pr[G=0]}. \tag{19}$$

相应的 MinNI 推导见附录8.3.2。

6.2 方差的敏感性分析

到目前为止，我们只考虑了 Y 均值的偏倚，但偏倚也可能影响方差。用显然的记号，我们将 σ_B^2 定义为随机变量 B 的方差，必要时可带条件。令

$$\mathrm{VD}_{YU} = \sigma^2_{Y|U=0} - \sigma^2_{Y|U=1}, \mathrm{VD}_{UG} = \sigma^2_{U|G=0} - \sigma^2_{U|G=1},$$

我们得到

$$\sigma^2_Y - \sigma^2_{Y|G=1} = \left\{\mathrm{VD}_{YU}\mathrm{RD}_{UG} + \mathrm{ED}^2_{YU}\mathrm{VD}_{UG} + \mathrm{ED}^2_{YU}\mathrm{RD}^2_{UG}\Pr[G=1]\right\}\Pr[G=0]. \tag{20}$$

定理1断言，若 $G \perp\!\!\!\perp U$ 或 $Y \perp\!\!\!\perp U$，则公式（20）中的差值为 0。对于均值比较，若 ED_{YU} 或 RD_{UG} 为 0，则不存在偏倚；但对于方差比较，这一条件并不充分，因为 VD_{YU} 或 VD_{UG} 可能不为 0。通常，感兴趣的主要矩是均值，并且已证明 $\sigma^2_{Y|G=1}$ 的一阶泰勒展开等于 σ^2_Y。（Troxel 等，2004年；Ma 和 Heitjan，2004年）我们可以很容易地推导出类似结果，用于估计给定 X时 Y的条件分布。

6.3 完全测量协变量的分析

许多研究会纳入一组基线变量；如果这些变量未被观测到，就会混杂结局与缺失之间的关联；我们将此类变量记为 X。我们可以很容易地将定理1和定理2推广到给定 X时 Y分布的估计。

定理3 假设 $Y \perp\!\!\!\perp G|(X,U)$，且满足 $G \perp\!\!\!\perp U|X$或 $Y \perp\!\!\!\perp U|X$。那么，对于任意满足 $f^{G,X}(g,x) > 0$的 g 和 x，忽略缺失机制 $f^{Y|X}(y|x)$的分布等于正确分布$f^{Y|G,X}(y|g,x)$。

推广　　定理2，我们定义以下假设，并假定 $\tilde{g}$ 是 G的观测值：

1. 缺失是观测可忽略 的，即对于任意可能的 u和 x, $f^{G|Y,X,U}(\tilde{g}|y,x,u)$ 对所有与 $\tilde{g}$一致的 y 均取相同值。2. 对于任意可能的 x， $f^{G|X,U}(\tilde{g}|x,u)$ 对所有 u均取相同值。3. 对于任意可能的 x 以及任意与 $\tilde{g}$ 一致的 y， $f^{Y|X,U}(y|x,u)$ 对所有 u均取相同值。

定理 4 在假设 *1* 和假设 *2* 或 *3* 之一成立时， $f^{Y|X}(y|x) = f^{Y|X,\ G}(y|x,\tilde{g})$ 对所有与 $\tilde{g}$ 一致的 y 均成立。

我们的分析也很容易推广到这种情形；也就是说，进一步对 X 进行条件化后，便可像上文那样阐明敏感性。假设测量协变量 X 是离散的。对于差值尺度，将这两个混杂关系表示为

$$\mathrm{ED}_{YU(X)} = \mathbb{E}[Y|X,U=1] - \mathbb{E}[Y|X,U=0],$$

$$\text{RD}_{UG(X)} = \Pr[U = 1|X, G = 1] - \Pr[U = 1|X, G = 0].$$

因此，

$$|\mathbb{E}[Y|X] - \mathbb{E}[Y|X, G = 1]| = |\text{ED}_{YU(X)}\text{RD}_{UG(X)}\Pr[G = 0|X]|.$$

上述公式 与公式 (8) 类似。不过，对于比率尺度，下面将展示一种简单的分析 .

$$\frac{\mathbb{E}[Y|X]}{\mathbb{E}[Y|X, G = 1]} = \Pr[G = 1|X] + \frac{\mathbb{E}[Y|X, G = 0]}{\mathbb{E}[Y|X, G = 1]}\Pr[G = 0|X],$$

并将相对比率记为

$$\text{ER}_{YU(X)} = \frac{\mathbb{E}[Y|X, U = 1]}{\mathbb{E}[Y|X, U = 0]},$$

$$\text{RR}_{UG(X)} = \frac{\Pr[U = 1|X, G = 1]}{\Pr[U = 1|X, G = 0]}.$$

因此，

$$\frac{\mathbb{E}[Y|X, G = 0]}{\mathbb{E}[Y|X, G = 1]} = \frac{\text{ER}_{YU(X)} - 1 + \frac{1}{\Pr[U=1|X,G=0]}}{(\text{ER}_{YU(X)} - 1)\text{RR}_{UG(X)} + \frac{1}{\Pr[U=1|X,G=0]}},$$

这与公式 (12) 相呼应。其他所有推导均可直接得出。对 X 进行调整后，边际均值与条件均值之间的总差异可以表示为按 X 加权的差异之和。

7 讨论

我们采用一种将非可忽略性视为未观测混杂后果的模型，设计出一种简单、通用的范式，用于对不完整数据的频率学派推断进行敏感性分析。我们模型中对非可忽略性的解释，与 Cornfield、Rosenbaum 和 Rubin，以及 Ding 和 VanderWeele 针对非可忽略混杂敏感性的分析方法相呼应。（Cornfield 等，1959年；Rosenbaum 和 Rubin，1983a；Ding 和 Vanderweele，2014年）

我们的敏感性分析旨在识别导致 $f^Y(y)$ 与 $f^{Y|G}(y|g)$ 的某种比较出现指定差异的最低非可忽略程度。在此展示的分析中，我们假设结局 Y代表来自某个分布的独立同分布样本，但这并非该方法的必要条件。此外，我们推导了用于比较均值和方差的方程，但该分析应当可以扩展到其他泛函。为保持保守，我们采用比率尺度上的部分识别区域，提出最敏感情形下的 MinNI，以概括非可忽略性的程度。

当存在连续测量协变量 X时，自然可以采用参数化敏感性模型，但这类模型需要指定 X的分布形式。(Lin 等，1998年；Mitra 和 Heitjan，2007年) 如果 X 的维度很大，那么第6.3节所述的 MinNI 非参数分析会因数据稀疏性（即维数灾难）而难以实施。因此，基于选择机制设定的其他非可忽略模型，例如 ISNI分析或倾向评分匹配方法，可能简单得多。(Troxel 等，2004年；Ma 等，2005年；Zhang 和 Heitjan，2006年；Zhao，2019年；Zhao 等，2019年) 在参数化敏感性分析中，一个可能的扩展方向是考察对参数模型设定错误的敏感性，例如，连接 Y与 G 和 U的连接函数设定错误。(Gustafson，2001年)

我们的模型与该领域许多先前研究的不同之处在于，它采用频率学派方法，而非似然/贝叶斯估计。（Heitjan 和 Rubin，1991年；Heitjan，1994年、1997年）原则上，我们可以构建类似的分析，在基于模型的分析中，以混杂机制设定替代选择机制设定作为不完整机制，并评估例如需要多大程度的最小非忽略性，才能使最大似然估计偏离指定幅度。这样，我们的方法便可与ISNI敏感性分析相媲美。

参考文献

Copas, J. B. 和 Li, H.（1997）。非随机样本推断。皇家统计学会会刊：*B*辑（统计方法论），59(1):55–95。

Cornfield, J.、Haenszel, W.、Hammond, E. C.、Lilienfeld, A. M.、Shimkin, M. B. 和 Wynder, E. L.（1959）。吸烟与肺癌：近期证据及若干问题的讨论。《国家癌症研究所杂志》，22(1):173–203。

Dawid, A. P. (1979). 统计理论中的条件独立性。皇家统计学会会刊：*B*辑（统计方法论）, 41(1):1–15.

Diggle, P. and Kenward, M. G. (1994). 纵向数据分析中的信息性脱落。皇家统计学会会刊：*C*辑（应用统计）, 43(1):49–73. Ding, P. and VanderWeele, T. J. (2014). 风险差的广义康菲尔德条件。*Biometrika*, 101(4):971– 977. Ding, P. and VanderWeele, T. J. (2016). 无假设的敏感性分析。*Epidemiology (Cambridge, Mass.)*, 27(3):368. Gustafson, P. (2001). 关于参数模型设定错误的敏感性度量。皇家统计学会会刊：*B*辑（统计方法论）, 63(1):81–94. Heitjan, D. F. (1994). 一般不完整数据模型中的可忽略性。*Biometrika*, 81(4):701–708. Heitjan, D. F. (1997). 可忽略性、充分性与辅助性。皇家统计学会会刊：*B*辑（统计方法论）, 59(2):375–381. Heitjan, D. F. and Rubin, D. B. (1991). 可忽略性与粗化数据。统计年刊, 2244–2253. Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998). 评估观察性研究中回归结果对未测量混杂因素的敏感性。生物统计学, 948–963. Little, R. J. and Rubin, D. B. (1999). 使用半参数无应答模型调整不可忽略脱落：评论。美国统计学会会刊, 94(448):1130–1132. Little, R. J. and Rubin, D. B. (2014). 缺失数据的统计分析。约翰·威立父子出版公司。Ma, G. and Heitjan, D. F. (2004). 频率学派推断对非可忽略性的敏感性。不完整数据视角下的应用贝叶斯建模与因果推断：与 *Donald Rubin* 统计家族同行的重要旅程, 175–186. Ma, G., Troxel, A. B., and Heitjan, D. F. (2005). 纵向建模中不可忽略脱落的局部敏感性指数。医学统计学, 24(14):2129–2150. Manski, C. F. (2003). 概率分布的部分识别。施普林格科学与商业媒体。Mitra, N. and Heitjan, D. F. (2007). 观察性研究中风险比对不可忽略治疗分配的敏感性。医学统计学, 26(6):1398–1414. Raab, G. M. and Donnelly, C. A. (1999). 部分数据缺失时有关性行为的信息。皇家统计学会会刊：*C*辑（应用统计）, 48(2):117–133. Rosenbaum, P. R. and Rubin, D. B. (1983a). 评估观察性研究中对未观测二元协变量的敏感性（结局为二元）。皇家统计学会会刊：*B*辑（统计方法论）, 45(2):212– 218. Rosenbaum, P. R. and Rubin, D. B. (1983b). 倾向评分在因果效应观察性研究中的核心作用。*Biometrika*, 70(1):41–55. Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). 针对重复结局且存在不可忽略无应答的半参数回归。美国统计学会会刊, 93(444):1321–1339. Rubin, D. B. (1976). 推断与缺失数据。*Biometrika*, 63(3):581–592. Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). 使用半参数无应答模型调整不可忽略脱落。美国统计学会会刊, 94(448):1096–1120. Troxel, A. B., Ma, G., and Heitjan, D. F. (2004). 不可忽略性的局部敏感性指数。华人统计学志, 1221–1237. VanderWeele, T. J. and Ding, P. (2017). 观察性研究中的敏感性分析：引入 E值。内科学年鉴。Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., and Kenward, M. G. (2001). 非随机脱落的敏感性分析：一种局部影响方法。生物统计学, 57(1):7–14. Xie, H., Gao, W., Xing, B., Heitjan, D. F., Hedeker, D., and Yuan, C. (2018). 使用*R* 软件包 isni 衡量不可忽略缺失的影响。生物医学中的计算机方法与程序, 164:207–220. Xie, H. and Heitjan, D. F. (2004). 交叉治疗临床试验中因果推断的敏感性分析。临床试验, 1(1):21–30. Zhang, J. and Heitjan, D. F. (2006). 不可忽略粗化的简易局部敏感性分析工具：依赖删失的应用。生物统计学, 62(4):1260–1268.

Zhao, Q. (2019). 配对匹配观察性研究的敏感性值。《美国统计学会会刊》，114(526):713–722。

Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2019). 基于百分位数自助法的逆概率加权估计量敏感性分析。《皇家统计学会会刊：B辑（统计方法论）》。

8 附录

8.1 二元结局下未知参数的估计

假设总体包含 N个单位，其中每个单位 i都对应一个向量 $(Y_i, G_i) \in \{0,1\} \times \{0,1)\}$，表示感兴趣的结局及其相应的缺失指示变量。感兴趣的参数为 $\Pr[Y=1]$。在第3.1节中，我们假设两个连接函数均为 logit函数。假设观测到一个大小为 n 的随机样本，其中有 n_m 个缺失值， n_o 个观测值，且 $n = n_m + n_o$。那么，缺失概率和 $\Pr[Y=0|G=1]$的最大似然估计（MLE）可通过样本中的缺失比例估计，

$$\hat{p}_m = \frac{n_m}{n}$$

and

$$\hat{\mu}_c = 1 - \frac{1}{n_o}\sum_{i:g_i=1}^{n} y_i.$$

按照第3.2节的前两个步骤，并固定 π_0， γ_1， β_1,

$$\Pr[G=0] = \frac{\pi_0}{1+\exp(\gamma_0)} + \frac{1-\pi_0}{1+\exp(\gamma_0+\gamma_1)}. \tag{21}$$

将 $\hat{}p_m$ 代入公式（21），并固定 π_0， γ_1，得到

$$\exp(\hat{\gamma}_0) = \frac{-[(\hat{p}_m-\pi_0)\exp(\gamma_1)+\hat{p}_m+\pi_0-1] + \sqrt{[(\hat{p}_m-\pi_0)\exp(\gamma_1)+\hat{p}_m+\pi_0-1]^2 + 4\exp(\gamma_1)\hat{p}_m(1-\hat{p}_m)}}{2\exp(\gamma_1)\hat{p}_m}.$$

Then

$$\Pr[Y=0|G=1] = \frac{w}{1+\exp(\beta_0)} + \frac{1-w}{1+\exp(\beta_0+\beta_1)}, \tag{22}$$

其中

$$\begin{aligned} w = \Pr[U=0|G=1] &= \frac{\Pr[G=1|U=0]\Pr[U=0]}{\Pr[G=1|U=0]\Pr[U=0] + \Pr[G=1|U=1]\Pr[U=1]} \\ &= \pi_0\left[\pi_0 + \frac{\exp(\gamma_1)[1+\exp(\gamma_0)]}{1+\exp(\gamma_0+\gamma_1)}(1-\pi_0)\right]^{-1}, \end{aligned}$$

并代入 π_0， γ_1 和 $\hat{}\gamma_0$ 来估计 w，记为 $\hat{}w$。将 $\hat{}w$ 和 $\hat{}\mu_c$ 代入公式（22）得到

$$\exp(\hat{\beta}_0) = \frac{-[(\hat{\mu}_c-\hat{w})\exp(\beta_1)+\hat{\mu}_c+\hat{w}-1] + \sqrt{[(\hat{\mu}_c-\hat{w})\exp(\beta_1)+\hat{\mu}_c+\hat{w}-1]^2 + 4\exp(\beta_1)\hat{\mu}_c(1-\hat{\mu}_c)}}{2\exp(\beta_1)\hat{\mu}_c}.$$

8.2 公式8的证明

$$\begin{aligned} \mathbb{E}[Y|G=0] - \mathbb{E}[Y|G=1] &= \mathbb{E}[Y|U=1,G=0]\Pr[U=1|G=0] + \mathbb{E}[Y|U=0,G=0]\Pr[U=0|G=0] \\ &\quad - \{\mathbb{E}[Y|U=1,G=1]\Pr[U=1|G=1] + \mathbb{E}[Y|U=0,G=1]\Pr[U=0|G=1]\} \\ &= \mathbb{E}[Y|U=1]\Pr[U=1|G=0] + \mathbb{E}[Y|U=0](1-\Pr[U=1|G=0]) \\ &\quad - \mathbb{E}[Y|U=1]\Pr[U=1|G=1] - \mathbb{E}[Y|U=0](1-\Pr[U=0|G=1]) \\ &= \mathbb{E}[Y|U=1](\Pr[U=1|G=0] - \Pr[U=1|G=1]) \\ &\quad - \mathbb{E}[Y|U=0](\Pr[U=1|G=0] - \Pr[U=1|G=1]) \\ &= (\mathbb{E}[Y|U=1] - \mathbb{E}[Y|U=0])(\Pr[U=1|G=0] - \Pr[U=1|G=1]) \end{aligned}$$

8.3 分类混杂因素

8.3.1 界不等式 推导（分类混杂因素 混杂因素）

对于差值尺度，MD_{YU} 和 MD_{UG} 已在上一节中用相同的符号定义。不失一般性，定义 U 的 m 水平，使其最小化 $\mathbb{E}[Y|U = u_j]$。为简化推导，定义$|\mathbb{E}[Y|G = 0] - \mathbb{E}[Y|G = 1]|$ 为 D。已推导出 $|\mathbb{E}[Y] - \mathbb{E}[Y|G = 1]| = D \times \text{Pr}[G = 0]$，而根据上述两个定义关系，唯一需要推导的公式是 D。

$$
\begin{aligned}
D &= \left| \sum_{i=1}^{m} \mathbb{E}[Y|U = u_i]\text{Pr}[U = u_i|G = 0] - \sum_{i=1}^{m} \mathbb{E}[Y|U = u_i]\text{Pr}[U = u_i|G = 1] \right| \\
&= \Bigg| \sum_{i=1}^{m-1} \mathbb{E}[Y|U = u_i]\text{Pr}[U = u_i|G = 0] + \mathbb{E}[Y|U = u_m][1 - \sum_{i=1}^{m-1} \text{Pr}[U = u_i|G = 0]] \\
&\quad -\{\sum_{i=1}^{m-1} \mathbb{E}[Y|U = u_i]\text{Pr}[U = u_i|G = 1] + \mathbb{E}[Y|U = u_m][1 - \sum_{i=1}^{m-1} \text{Pr}[U = u_i|G = 1]]\} \Bigg| \\
&= \Bigg| \sum_{i=1}^{m-1} [\mathbb{E}[Y|U = u_i] - \mathbb{E}[Y|U = u_m]]\text{Pr}[U = u_i|G = 0] \\
&\quad - \sum_{i=1}^{m-1} [\mathbb{E}[Y|U = u_i] - \mathbb{E}[Y|U = u_m]]\text{Pr}[U = u_i|G = 1] \Bigg| \\
&\leq \left| \text{MD}_{YU} \sum_{i=1}^{m-1} [\text{Pr}[G = u_i|G = 0] - \text{Pr}[G = u_i|G = 1]] \right| \\
&\leq |(m-1)\text{MD}_{YU}\text{MD}_{UG}|
\end{aligned}
$$

这是不等式（16）。

对于比率尺度，$\text{ER}^{YU(i)}$、$\text{RR}^{UG(i)}$、MR_{YU} 和 MR_{UG} 的定义方式相同，且均假设大于 1。不失一般性，令

$$\mathbb{E}[Y|U = u_1] = \text{min}_i\mathbb{E}[Y|U = u_i], \ \mathbb{E}[Y|U = u_m] = \text{max}_i\mathbb{E}[Y|U = u_i].$$

将 R 记作 $\frac{\mathbb{E}[Y|G=0]}{\mathbb{E}[Y|G=1]}$。那么

$$
\begin{aligned}
R &= \frac{\sum_{i=1}^{m} \mathbb{E}[Y|U = u_i]\text{Pr}[G = u_i|G = 0]}{\sum_{i=1}^{m} \mathbb{E}[Y|U = u_i]\text{Pr}[G = u_i|G = 1]} \\
&= \frac{\text{MR}_{YU}\text{Pr}[U = u_1|G = 0] + \sum_{i=2}^{m-1} \text{ER}_{YU(i)}\text{Pr}[G = u_i|G = 0] + \text{Pr}[U = u_m|G = 0]}{\text{MR}_{YU}\text{Pr}[U = u_1|G = 1] + \sum_{i=2}^{m-1} \text{ER}_{YU(i)}\text{Pr}[G = u_i|G = 1] + \text{Pr}[U = u_m|G = 1]} \\
&= \frac{\sum_{i=1}^{m-1}(\text{ER}_{YU(i)} - 1)\text{Pr}[G = u_i|G = 0] + 1}{\sum_{i=1}^{m-1}(\text{ER}_{YU(i)} - 1)\text{Pr}[G = u_i|G = 1] + 1} \\
&= \frac{(\text{MR}_{YU} - 1)\text{Pr}[U = u_1|G = 0] + \sum_{i=2}^{m-1}(\text{RR}_{YU(i)} - 1)\text{Pr}[G = u_i|G = 0] + 1}{(\text{MR}_{YU} - 1)\text{RR}_{UG(1)}\text{Pr}[U = u_1|G = 0] + \sum_{i=2}^{m-1}(\text{RR}_{YU(i)} - 1)\text{RR}_{UG(i)}\text{Pr}[G = u_i|G = 0] + 1} \\
&\geq \frac{(\text{MR}_{YU} - 1)\text{Pr}[U = u_1|G = 0] + 1}{(\text{MR}_{YU} - 1)\text{RR}_{UG(1)}\text{Pr}[U = u_1|G = 0] + 1} \\
&\geq \frac{\text{MR}_{YU} + \text{RR}_{UG(1)} - 1}{\text{MR}_{YU}\text{RR}_{UG(1)}} \\
&\geq \frac{\text{MR}_{YU} + \text{MR}_{UG} - 1}{\text{MR}_{YU}\text{MR}_{UG}}.
\end{aligned}
$$

当所有相对比率均大于 1 时，$R < 1$:

$$\frac{\text{MR}_{YU} + \text{MR}_{UG} - 1}{\text{MR}_{YU}\text{MR}_{UG}} \leq R \leq 1$$

我们可以将界限重新表述为

$$|R-1| \leq \frac{(\text{MR}_{YU}-1)(\text{MR}_{UG}-1)}{\text{MR}_{YU}\text{MR}_{UG}},$$

由此得到不等式 (18)。当某些相对比率大于 1，而另一些小于 1 时，推导会更加复杂，但可以考虑采用类似技巧，为 $\mathbb{E}[Y]$ 和 $\mathbb{E}[Y|G=1]$ 的比率寻找一个界值。

8.3.2 分类混杂因素的 MinNI

在风险差尺度上的 MinNI 由不等式 (17) 得出的 (MD_{YU},MD_{UG}) 为：对于连续型结果，

$$\left(\max\left\{\frac{k\sigma_{Y|G=1}}{(m-1)\text{Pr}[G=0]}, \sqrt{\frac{k\sigma_{Y|G=1}}{(m-1)\text{Pr}[G=0]}}\right\}, \min\left\{1, \sqrt{\frac{k\sigma_{Y|G=1}}{(m-1)\text{Pr}[G=0]}}\right\}\right),$$

但对于二元结局，

$$\left(\sqrt{\frac{k\sigma_{Y|G=1}}{(m-1)\text{Pr}[G=0]}}, \sqrt{\frac{k\sigma_{Y|G=1}}{(m-1)\text{Pr}[G=0]}}\right),$$

其中 $k\sigma_{Y|G=1} \leq (m-1)\text{Pr}[G=0]$。由不等式 (19) 得到的风险比尺度上的 MinNI，对于 (MR_{YU},MR_{UG})，为

$$\left(\frac{1}{1-\sqrt{\frac{k\text{CV}_{Y|G=1}}{\text{Pr}[G=0]}}}, \frac{1}{1-\sqrt{\frac{k\text{CV}_{Y|G=1}}{\text{Pr}[G=0]}}}\right),$$

其中 $0 \leq k\text{CV}_{Y|G=1} < \text{Pr}[G=0] \leq 1$，这与二元情形相同。