

目录

译者序与前言	i
第一部分 不加模型的因果推断	
第一章 因果效应的定义	3
1.1 个体的因果效应	3
1.2 因果效应的均值	4
1.3 因果效应的量度	6
1.4 随机变异性	7
1.5 因果性与相关性	8
第一章精讲点和知识点	10
第一章图表	13
第二章 随机试验	15
2.1 随机	15
2.2 条件随机	18
2.3 标准化	20
2.4 逆概率加权	21
第二章精讲点和知识点	23
第二章图表	27
第三章 观察性研究	30
3.1 可识别性	30
3.2 互换性	32
3.3 正数性	34
3.4 一致性：首先，定义反事实结局	35
3.5 一致性：其次，将反事实世界和观测到的数据相结合	37
3.6 靶标试验	39
第三章精讲点和知识点	42
第三章图表	46
第四章 效应修饰	48
4.1 效应修饰的定义	48
4.2 通过分层分析识别效应修饰	50
4.3 为什么要关注效应修饰	52

4.4 通过分层调整变量	54
4.5 通过匹配调整变量	55
4.6 效应修饰与变量调整	56
第四章精讲点和知识点	58
第四章图表	62
第五章 交互作用	66
5.1 交互作用需要联合干预	66
5.2 识别交互作用	67
5.3 反事实回应类型和交互作用	68
5.4 充分成因	70
5.5 充分成因的交互作用	72
5.6 反事实框架还是充分成因框架?	73
第五章精讲点和知识点	73
第五章图表	77
第六章 因果效应的图像表示	80
6.1 因果图	80
6.2 因果图与边缘独立性	81
6.3 因果图与条件独立	83
6.4 因果图中的正数性和一致性	85
6.5 偏移的结构性分类	87
6.6 效应修饰的结构	88
第六章精讲点和知识点	90
第六章图表	95
第七章 混杂	97
7.1 混杂的结构	97
7.2 混杂与互换性	99
7.3 混杂与后门准则	100
7.4 混杂与混杂变量	103
7.5 单一世界干涉图	105
7.6 混杂调整	106
第七章精讲点和知识点	108
第七章图表	114
第八章 选择偏移	116
8.1 选择偏移的结构	116

8.2 选择偏移的例子	118
8.3 选择偏移与混杂	120
8.4 选择偏移与删失	121
8.5 如何调整选择偏移	123
8.6 没有偏移的选择	125
第八章精讲点和知识点	126
第八章图表	128
第九章 测量偏移	133
9.1 测量误差	133
9.2 测量误差的结构	134
9.3 混杂变量的测量误差	135
9.4 治疗意向的效应：治疗错分类的影响	136
9.5 依方案效应	138
第九章精讲点和知识点	140
第九章图表	142
第十章 随机变异性	145
10.1 识别与估计	145
10.2 因果效应的估计	147
10.3 超级人群	148
10.4 条件性“准则”	150
10.5 维度的诅咒	152
第十章精讲点和知识点	152
第十章图表	157

第二部分 模型中的因果推断

第十一章 统计模型	161
11.1 数据不会说话	161
11.2 条件均值的参数估计	163
11.3 条件均值的非参数估计	164
11.4 平滑	165
11.5 偏差方差权衡	167
第十一章精讲点和知识点	168
第十一章图表	170
第十二章 逆概率加权和边缘结构模型	172
12.1 因果性问题	172

12.2 使用模型计算逆概率权重	173
12.3 逆概率稳定权重	175
12.4 边缘结构模型	177
12.5 效应修饰与边缘结构模型	180
12.6 删失和缺失值	181
第十二章精讲点和知识点	183
第十二章图表	186
第十三章 标准化和参数 G-公式	188
13.1 标准化	188
13.2 通过模型估计结局均值	189
13.3 根据混杂变量的分布对结局均值进行标准化	190
13.4 逆概率加权还是标准化？	192
13.5 我们应该如何看待估计值？	193
第十三章精讲点和知识点	195
第十三章图表	199
第十四章 G-估算和结构嵌入模型	200
14.1 再谈因果性问题	200
14.2 再谈互换性	201
14.3 结局均值的结构嵌入模型	202
14.4 保序性	203
14.5 G-估算	205
14.6 两个或多个参数的结构嵌入模型	207
第十四章精讲点和知识点	209
第十四章图表	212
第十五章 结局回归与倾向性评分	213
15.1 结局回归	213
15.2 倾向性评分	214
15.3 倾向性分层和标准化	216
15.4 倾向性评分匹配	218
15.5 倾向性评分模型, 结构模型, 以及预测模型	219
第十五章精讲点和知识点	222
第十五章图表	223
第十六章 工具变量	224
16.1 工具变量的三个条件	224

16.2 工具变量的效应估计	226
16.3 工具变量的第四个条件：同质性	228
16.4 另一种第四个条件：单调性	229
16.5 再谈工具变量的三个条件	232
16.6 工具变量与其他方法比较	234
第十六章精讲点和知识点	235
第十六章图表	240
第十七章 因果推断中的生存分析	243
17.1 危害与风险	243
17.2 从危害到风险	245
17.3 删失	247
17.4 生存分析中的逆概率加权	248
17.5 生存分析中的 G-公式	250
17.6 生存分析中的 G-估算	251
第十七章精讲点和知识点	254
第十七章图表	258
第十八章 因果推断中的变量选择	260
18.1 变量选择的不同目的	260
18.2 造成偏移或放大偏移的变量	261
18.3 因果推断与机器学习	263
18.4 机器学习中的双重稳健估计	265
18.5 变量选择永远是一个难题	266
第十八章精讲点和知识点	267
第十八章图表	268
第三部分 复杂纵向数据的因果推断	
第十九章 时异治疗	271
19.1 时异治疗的因果效应	271
19.2 治疗策略	272
19.3 时序性随机试验	273
19.4 时序互换性	275
19.5 部分治疗策略下的可识别性	276
19.6 时异混杂	279
第十九章精讲点和知识点	280
第十九章图表	284

第二十章 治疗-混杂反馈	286
20.1 治疗-混杂反馈的要素	286
20.2 传统方法的不足	287
20.3 为什么传统方法失效了	289
20.4 我们能改进传统方法吗？	290
20.5 过往治疗	291
第二十章精讲点和知识点	293
第二十章图表	294
第二十一章 时异治疗的 G-方法	296
21.1 时异治疗的 G-公式	296
21.2 时异治疗的逆概率加权	299
21.3 时时异治疗的双重稳健估计	302
21.4 时异治疗的 G-估算	304
21.5 删失与时异变量	308
第二十一章精讲点和知识点	310
第二十一章图表	316
第二十二章 靶标试验	318
22.1 再论靶标试验	318
22.2 随机试验的因果效应	319
22.3 观察性研究的因果效应	321
22.4 初始时间	322
22.5 因果分析的统一框架	323
第二十二章精讲点和知识点	325
中英词汇对照表	329

译者序

译者在开始博士课程后接触到了因果推断 (causal inference) 这门课程, 或者说, 一种和传统统计实践思想迥然不同的新理论。1965 年 Bradford Hill 爵士提出了判断因果关系的 9 条 (一说 10 条) 标准, 此后, 流行病学越来越多的注意力开始转移到因果关系的分析上。相应的, 统计学学者也开始探讨因果分析的基本理论与方法。自 1970 年代以来, 哈佛大学 Donald Rubin 教授提出的反事实框架 (counterfactual framework) 和加州大学洛杉矶分校 Judea Pearl 教授提出的贝叶斯网络 (Bayesian network), 成为了因果推断的两大主要思潮。现今的因果分析方法, 绝大多数都是建立在这两大思潮之上。

现今, 因果推断及其分析方法已经被广泛应用于诸多社会学科, 包括流行病学、经济学、心理学、社会学等。然而国内却鲜有系统介绍因果推断体系的教科书。本书由哈佛大学陈增熙公共卫生学院 Miguel A. Hernan 和 James M. Robins 两位教授合著而成, 既是一本流行病学教材, 也是一本因果推断的入门教材。本书在流行病学的框架下, 从随机试验开始讲起, 系统地介绍了因果推断的基本框架和方法。内容从浅到深、循序渐进, 论证犹如数学教科书一般严谨, 也许广大流行病学子第一眼看去, 会有陌生之感。

虽然本书是一本优秀的教材, 但是依然存在一些不足之处。译者冒昧指出其中几点。第一, 中介效应分析 (mediation analysis) 是因果推断的重要内容之一, 然而本书基本没有讨论这一内容。译者猜想, 主要原因可能是本书作者的同事 Tyler VanderWeele 教授, 一位对中介效应分析论著颇丰的学者, 在本书面世之前已经单独出版了一本中介效应分析的专著, 因而为了避免重复, 本书没有纳入中介效应分析。第二, 本书作者都是哈佛大学的学者, 因而本书的主体内容大多来自于哈佛学派的论述。虽然能感受到作者试图兼顾其他学校学者的成果, 然而其对哈佛学派的侧重依然非常明显。同时, 哈佛学派的某些论述并不一定被所有学者认同接受, 而本书却也大篇幅地介绍讨论。所以, 读者在阅读本书时最好配合阅读其他学者的论文, 学会甄别本书内容。第三, 本书假设读者具有一定的统计基础, 因而未对统计知识做过多介绍。本书给出的统计内容不具备知识梯度上的连续性, 读者在阅读时若遇见过于复杂的统计论证, 可以暂时选择跳过, 不要被吓到失去信心。第四, 本书原文语言有时过于重复琐碎。

流行病学是和实际紧密联系的一门学科。因而许多同学在学习因果推断时总会有这样的疑问: 我们不可能实现或验证因果推断的种种前提假设, 那我们还要这一套理论有何用? 这些顾虑很正常, 但不要因为这些顾虑停止学习的步伐。因果推断的种种方法固然理想化甚至不切实际, 但它提出的问题却是常问常新。因果推断能让我们从一种全新的角度思考问题, 并且因果推断框

Causal Inferences: What if ——前言

作者: Miguel A. Hernan, James M. Robins;

翻译: 罗家俊

架下发展出的分析方法, 能够弥补流行病学传统分析方法的不足, 处理一些过往不能分析的情境。

中国境内尚无系统介绍因果推断的教材, 希望此次翻译能弥补国内部分缺陷。译者的翻译基于原书 2020 年 2 月 21 日及之后的网上版本。翻译过程从 2020 年 3 月 29 日开始, 历经一年多, 在 2021 年 4 月 22 日结束最后一章的翻译。在翻译本书的时候, 尽可能直译, 以保持原书的风格。每一章的图表都从原书中摘出, 置于翻译版每一章的末尾。每一章还有作为正文补充的精讲点和知识点, 这些都一并翻译并置于每章的最后。同时, 在原书的边栏中, 会有小字对正文内容进行补充说明。这些小字非常重要, 在翻译时我将它们置于括号内, 在相应段落后另起一段用斜体表示。译者同时将对应的原书页码置于左侧, 以便读者查询。因为个人精力有限, 没有出版社进行帮忙编修, 排版和语言不能尽善尽美, 望读者见谅。

前言: 让因果推断更加严肃¹

vii

不得不承认, 用“因果推断”作为书名, 是有一些自命不凡的意味在其中。因果推断是一项复杂的科学任务, 需要剖析不同根源的证据, 同时应用不同的方法进行分析。不同的学科有不一样的因果推断分析方法, 因而没有一本书能全面概述所有方法。所有“因果推断”相关书籍都只能有选择地介绍作者认为重要的因果推断分析方法。

这个前言的标题就反映了我们的选择。我们希望这本书能帮助科研工作者——尤其是健康与社科领域的科研工作者——在明确说出自己前提假设的情况下, 做出正确的因果推断。不幸的是, 现今的文献充斥了大量问题不明确、前提假设不清晰的研究。这些随意的因果推断态度, 造成了大量的混乱无序。比如, 你能轻易地找到一个研究, 这个研究中的效应估计很难得到合理的阐释, 因为在这个研究的前提假设(如果有说明)下, 它的数据分析方法不能合理地回答它的因果问题(如果有明确表述)。

在这本书中, 我们将强调精准表述因果问题的重要性, 以及区分因果推断中数据与前提假设的必要性。一旦这些准备工作完成, 因果推断将变得不再随意, 这将有助于减少种种混乱。这本书描述了不同的数据分析方法。在特定的前提假设下, 借助从人群中收集来的数据, 这些方法可用来估算我们感兴趣的因果效应。这本书想传达这样一种信息: 因果推断不能也不应该简化成一系列数据分析方法的总和。

本书按照难度分为三个部分。第一部分主要介绍了没有数学模型的因果推断(也即因果效应的非参数化确定); 第二部分主要介绍了数学模型下的因果推断(也即用参数模型估计因果效应); 第三部分主要介绍了复杂纵向数据中的因果推断(也即时异变量的因果效应估算)。在正文中, 为了更好地说明某些主题, 我们穿插了各种精讲点与知识点讲解²。精讲点旨在让所有读者都能理解, 知识点则是为有一定统计基础的读者而设计。这本书全面地介绍了因果推断的相关概念和方法, 目前, 这些概念与方法散见于不同学科的多种期刊之中。我们期望这本书能够吸引任何对因果推断感兴趣的人, 比如流行病学家、统计学家、心理学家、经济学家、社会学家、政治学家、计算机科学家等等。

viii

重要的是, 这本书不是一本哲学书。我们对形而上学中的种种概念, 诸如因果等概念, 持不可知论态度。相反, 我们更关注人群中因果效应的确定与估算, 也即, 量化在不同干预措施下的结局变化。例如, 我们将会讨论如何比较严重心力衰竭患者接受和不接受心脏移植的死亡风险。我们的目标是帮助决策者做出更好地决策——一种可行的因果推断。

¹ 原标题: Towards less casual causal inferences

² 精讲点: fine point; 知识点: technical point。精讲点和知识点的翻译被译者置于每一章的最后。

我们感谢许许多多让这本书成功面世的师友。Stephen Cole, Sander Greenland, Jay Kaufman, Eleanor Murray, Sonja Swanson, Tyler VanderWeele 和 Jan Vandenbroucke 提供了详细的评论。Goodarz Danaei, Kasuke Kawai, Martin Lajous 和 Kathleen Wirth 帮助创建了NHEFS 数据集。第二部分中的示例代码, SAS 部分由 Roger Logan 完成, Stata 部分由 Eleanor Murray 和 Roger Murray 完成, R 部分由 Joy Shi 和 Sean McGrath 完成, Python 部分由 James Fiedler 完成。Roger Logan 还是我们的 LaTeX 导师。Randall Chaput 绘制了第一章和第二章的插图。Josh McKible 设计了本书封面。Rob Calver, 我们耐心的出版商, 鼓励我们完成这本书, 并支持我们在网上免费公开的决定。

此外, 还有许多同事对本书进行了校对, 帮助我们检查了拼写错误, 以及挑出表意不明的段落。我们要特别感谢 Kafui Adjaye-Gbewonyo, Álvaro Alonso, Katherine Almendinger, Ingelise Andersen, Juan José Beunza, Karen Biala, Joanne Brady, Alex Breskin, Shan Cai, Yu-Han Chiu, Alexis Dinno, James Fiedler, Birgitte Frederiksen, Tadayoshi Fushiki, Leticia Grize, Dominik Hangartner, John Jackson, Luke Keele, Laura Khan, Dae Hyun Kim, Lauren Kunz, Martín Lajous, Angeliki Lambrou, Wen Wei Loh, Haidong Lu, Mohammad Ali Mansournia, Giovanni Marchetti, Lauren McCarl, Shira Mitchell, Louis Mittel, Hannah Oh, Ivironke Olofin, Robert Paige, Jeremy Pertman, Melinda Power, Bruce Psaty, Brian Sauer, Tomohiro Shinozaki, Ian Shrier, Yan Song, Øystein Sørensen, Etsushi Suzuki, Denis Talbot, Mohammad Tavakkoli, Sarah Taubman, Evan Thacker, Kun-Hsing Yu, Vera Zietemann, Jessica Young, Dorith Zimmermann。

Causal Inferences: What if——第一章
作者：Miguel A. Hernan, James M. Robins;
翻译：罗家俊

第一部分

不加模型的因果推断

Causal Inferences: What if——第一章
作者：Miguel A. Hernan, James M. Robins;
翻译：罗家俊

第一章 因果效应的定义

3 开始阅读本书的你想必对因果推断很感兴趣。不过，作为一个会思考的人，你已经掌握了因果推断的基本概念。你肯定知道因果的含义是什么。你也清楚地知道因果和相关的区别是什么。在之前的人生中，你都在不断地使用这些知识。实际上，如果你不理解因果概念，你恐怕活不到能读到本书的年纪——甚至活不到学习阅读的年纪。在你是婴儿的时候，当你看到同伴跳入泳池并得到果酱作为奖励，你也会跟着跳进泳池开始学习游泳。在你是少年的时候，当你看到挑战最难赛道的同伴总是赢得每一次滑雪比赛，你也会试着尝试这条最难的赛道提高自己。在你为人父母的时候，当你看到邻居的小孩在服用药物的第二天就不能外出玩耍，你仍会给你生病的孩子服用药物。

既然你已经理解了因果的含义，以及相关性和因果性的区别，就不要指望从本章中得到什么发人深省的深刻见解。本章旨在介绍各种数学符号与表达式，从而用符号的方式表达我们对因果关系的种种直觉。我们需要确保这些数学符号能表达我们的所思所想。这些数学符号对于精准定义因果概念非常有必要。我们将在这本书中使用这些数学符号。

1.1 个体的因果效应

宙斯（Zeus）是我们的病人，他需要接受心脏移植手术。在 1 月 1 号这一天，他接受了手术。5 天后，他死了。不过我们知道（也许他是神，可以再重复一遍），如果宙斯在 1 月 1 号没有接受心脏移植手术，那他 5 天之后还会活着。有了这些信息，我们就能说是移植手术导致了宙斯的死亡。也就是说，心脏移植手术对宙斯的 5 天生存情况有因果效应。

另一个病人赫拉（Hera）在 1 月 1 号也接受了心脏移植手术。5 天后她还活着。不过我们知道（也许她是神，可以再重复一遍），如果赫拉没有在 1 月 1 号接受手术，5 天后她也会活着。因此我们可以说，心脏移植手术对赫拉的 5 天生存情况没有因果效应。

这两个小例子说明了我们人类如何推理因果效应：我们会（在脑海中）比较做 A 这件事和不做 A 这件事的结局。如果两个结局不一样，我们会说，事件 A 对结局有因果效应，这个效应可以是导致某件事发生，也可以是防止某件事发生。如果两个结局一样，我们会说事件 A 对结局没有因果效应。在流行病学、统计学、经济学以及其他社会科学的研究中，事件 A 可以是一种干预措施，可以是暴露情况，也可以是一种治疗方案。

为了让我们脑海中的因果直觉和数学分析相洽，我们需要一些数学符号。让我们先考虑一个二分的治疗变量 A （1: 治疗，0: 不治疗），和一个二分的结局变量 Y （1: 死亡，0: 存活）。在这本书中，对于某一个变量，如果人群中每一个个体的取值不尽相同，我们就把这个变量称作

随机变量。记 $Y^{a=1}$ (读作 $a=1$ 下的 Y) 为治疗变量取值 $a=1$ 时, 我们观察到的结局变量。同

理, 记 $Y^{a=0}$ (读作 $a=0$ 下的 Y) 为治疗变量取值 $a=0$ 时, 我们观察到的结局变量。 $Y^{a=1}$ 和

- 4 $Y^{a=0}$ 都是随机变量。对于我们例子中的宙斯来说, 因为他接受了手术后死亡, 但是不接受手术却能存活, 所以他的 $Y^{a=1} = 1$ 且 $Y^{a=0} = 0$ 。对于赫拉来说, 因为不管她接不接受手术, 她都存活了下来, 所以她的 $Y^{a=1} = 0$ 且 $Y^{a=0} = 0$ 。

(大写字母表示随机变量, 小写字母表示随机变量的特定取值)

现在我们可以对一个个体的因果效应进行定义: 对一个个体而言, 如果 $Y^{a=1} \neq Y^{a=0}$, 我们就说治疗变量 A 对这个个体的结局 Y 有因果效应。对于宙斯来说, 因为 $Y^{a=1} = 1 \neq Y^{a=0} = 0$, 所以治疗对结局有因果效应; 但对于赫拉来说, 因为 $Y^{a=1} = 0 = Y^{a=0}$, 所以治疗没有因果效应。变量 $Y^{a=1}$ 和 $Y^{a=0}$ 被称作潜在结局 (potential outcomes), 或者反事实结局 (counterfactual outcomes)。一些人更偏爱“潜在结局”这一名称, 因为这一名称强调了现实当中, 我们一般只能够观察到一种结局。另一些人则更偏爱“反事实结局”这一名称, 因为这一名称强调了这些结局可能在现实中从来不会出现 (也就是说, 和现实相反)。

(有时我们用 $Y_i^a = 1$ 表示“个体 i 的结局 $Y^a = 1$ ”。严格来说, 当用 i 来指代某个个体时, Y_i^a 不再是一个随机变量, 这是因为我们需要假设每个个体的反事实结局都是命定的 (详见知识点 1.2)。对于个体而言, 存在因果效应被表述为 $Y_i^{a=1} \neq Y_i^{a=0}$)

对于每一个个体, 他的其中一个反事实结局是实际存在的, 也就是现实中他接受的治疗方案下观测到的结局。比如, 宙斯实际上接受了治疗 ($A=1$), 他的反事实结局 $Y^{a=1} = 1$ 也就等同于实际观察到的结局 $Y = 1$ 。换句话说, 如果一个个体接受的治疗 A 是 a , 那么我们观察到的结局 Y 就等同于他的反事实结局 Y^a 。这个等同性可以简洁地表述为 $Y = Y^A$, 其中 Y^A 表示在治疗 A 为 a 的情况下, 我们观察到的结局 Y^a 。等式 $Y = Y^A$ 被称为一致性。

(一致性: 如果 $A_i = a$, 那么 $Y_i^a = Y_i^A = Y_i$)

个体的因果效应被定义为不同反事实结局之间的对比, 但对每个人来说, 只有一个结局能被观测到, 那就是实际中每个个体接受的治疗所对应的结局。其他反事实结局都不能被观测到。因而, 我们虽然不情愿, 但只能得出这样一个结论: 因为数据的缺失, 个体的因果效应是不可被识别的, 也就是说, 个体的因果效应不能用已观测到的数据来表示。

1.2 因果效应的均值

我们需要三份信息来定义个体的因果效应: 我们感兴趣的结局, 我们要比较的事件 $a=1$ 和 $a=0$, 以及我们要比较的反事实结局 $Y^{a=1}$ 和 $Y^{a=0}$ 。然而, 我们已经知道, 识别一个个体的因果效应基本上是不可能的。不过, 我们可以识别一个总体的因果效应, 也就是一个人群中, 每个个体的因果效应的均值。为了定义因果效应的均值, 我们需要三份信息: 我们感兴趣的结局, 我们要比较的事件 $a=1$ 和 $a=0$, 以及一个良定¹的人群——我们将比较人群中个体的 $Y^{a=1}$ 和 $Y^{a=0}$ 。

我们用宙斯的家庭做例子, 假设这个家庭就是我们要研究的人群。表 1.1 展示了这个家庭中 20 个成员在有治疗 ($a=1$) 和没治疗 ($a=0$) 情况下的反事实结局。表中最后一列展示了如果每个人都接受了治疗 (心脏移植), 我们能观察到的结局 $Y^{a=1}$ 。从表中可知, 他们中的 10 个人会在接受心脏移植后死去。也就是说, 如果所有人都接受治疗 ($a=1$), 那么去世的人的比例会是 5 $\Pr[Y^{a=1} = 1] = 10 / 20 = 0.5$ 。同理, 从表 1.1 的第二列中, 我们能知道如果所有人都没有接受治疗, 会有 10 个人去世。也就是说, 如果所有人都接受治疗 ($a=0$), 那么去世的人的比例会是 $\Pr[Y^{a=0} = 1] = 10 / 20 = 0.5$ 。在这里, 我们计算有治疗时的反事实风险所用的方法 (死亡人数除以总人数), 与计算这个人群中所有个体的反事实结局的均值的方法, 是一样的。

现在我们能对因果效应均值给出一个正式的定义: 在一个人群中, 如果 $\Pr[Y^{a=1} = 1] \neq \Pr[Y^{a=0} = 1]$, 那事件 A 就对结局 Y 存在因果效应。在这个定义下, 我们例子中的治疗 A 就对结局 Y 不存在因果效应, 因为有治疗下的死亡风险 $\Pr[Y^{a=1} = 1]$ 和没治疗下的死亡风险 $\Pr[Y^{a=0} = 1]$ 都是 0.5。也就是说, 不管全体接受或不接受心脏移植, 都会有一半的人去世。当在某一个人群中因果效应均值为零的时候 (就如同我们的例子), 我们会说因果效应均值的零假设为真。在我们的定义里风险等于均值, 同时也因为字母 E 经常用来表示人群中的均值 (同时也是英文中期望 Expectation 的首字母), 所以我们可以将人群中非零因果效应均值定义为 $E[Y^{a=1}] \neq E[Y^{a=0}]$, 这样一来, 我们的定义就既能用于二分结局, 也能用于非二分的结局。

在我们的例子中, 心脏移植治疗 A 的因果效应的均值是由两种不同行为 (即接受治疗 6 ($a=1$) 与不接受治疗 ($a=0$)) 的对比决定的。当一个事件牵涉到不止两种行为时 (也即我们的事件不仅只是一个二分事件), 我们需要指明我们感兴趣的对比。比如, “阿司匹林的因果效应”这一说法没有任何意义, 因为我们没有说明剂量, 也没有指明对比。但是, “连续 5 年每

¹ 英文 well-defined, 一般数学中翻译为良定的, 意思是定义清晰, 不模糊。

日口服 150 mg 阿司匹林”和“不服用阿司匹林”的对比却是有意义的，我们能从中这个对比中得到因果效应。只有指明了对比，因果效应才是良定的，哪怕其行为或者其对应的反事实结局可能根本不存在（比如，连续 5 年每天用皮肤吸收 500 mg 阿司匹林）。

人群中因果效应均值为零，并不意味着个体因果效应为零。从表 1.1 中可以知道，治疗对这个人群中的 12 个个体（包括宙斯）有个体因果效应，因为这 12 个人的反事实结局 $Y^{a=1}$ 和 $Y^{a=0}$ 不等。对其中 6 个人（包括宙斯）而言，治疗是有害的 ($Y^{a=1} - Y^{a=0} = 1$)，但对其他 6 个人而言，治疗是有益的 ($Y^{a=1} - Y^{a=0} = -1$)。这对半的比例，不是巧合，而是反映了因果效应均值的定义：人群中因果效应均值 $E[Y^{a=1}] - E[Y^{a=0}]$ ，等于个体因果效应 $Y^{a=1} - Y^{a=0}$ 的均值 $E[Y^{a=1} - Y^{a=0}]$ 。也就是说均值的差等于差的均值。如果在人群中，每个个体的因果效应都为零，也就是对每个个体都有 $Y^{a=1} = Y^{a=0}$ ，我们就说极端因果零假设为真。

在下一章，我们会讨论到，即使不能从数据中估算个体的因果效应，但有时我们依然可以识别人群中因果效应的均值。以下，我们会把“因果效应的均值”简称为“因果效应”，以及“效应均值的零假设”简称为“因果零假设”。

7 1.3 因果效应的量度

在我们的例子中，我们已经知道心脏移植治疗 A 对结局死亡 Y 没有因果效应。因为 $\Pr[Y^{a=1} = 1] = \Pr[Y^{a=0} = 1] = 0.5$ ，所以因果零假设成立。我们还可以用许多其他的方法来表达零效应，包括：

$$(i) \quad \Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1] = 0$$

$$(ii) \quad \frac{\Pr[Y^{a=1} = 1]}{\Pr[Y^{a=0} = 1]} = 1$$

$$(iii) \quad \frac{\Pr[Y^{a=1} = 1] / \Pr[Y^{a=1} = 0]}{\Pr[Y^{a=0} = 1] / \Pr[Y^{a=0} = 0]} = 1$$

这里的 (i)、(ii)、(iii) 分别对应因果性下的风险差、风险比和比值比。

（人群中的因果性风险差，是个人因果性风险差 $Y^{a=1} - Y^{a=0}$ 的均值，换句话说，这个量度也是个体因果效应均值的量度。与之相比，人群中的因果性风险比，不是个人因果性风险比 $Y^{a=1} / Y^{a=0}$ 的均值，也即这个量度是人群中因果效应的量度，但不是个体因果效应均值的量度）

8 假设在我们的人群中, 有另一个事件 A (比如说吸烟) 对我们的另一个结局 Y (比如说肺癌) 有因果效应, 因果零假设不成立, 也就是 $\Pr[Y^{a=1} = 1] \neq \Pr[Y^{a=0} = 1]$ 。在这种情况下, 因果性风险差不是 0、风险比不是 1 以及比值比也不是 1。这些参数只是从不同的尺度去衡量同一个因果效应, 我们称这些参数为效应量度

每一个效应量度都有自己的含义与作用。比如, 在一大群人中, 如果我们进行治疗, 一百万人中会有 1 人得病, 如果我们不进行治疗, 一百万人中会有 3 人得病。这里的因果性风险比是 3, 但是因果性风险差为 0.000002。这里的因果性风险比 (乘法尺度), 是用来计算相对于治疗, 不治疗的人患病风险升高的倍数。这里的因果性风险差 (加法尺度), 则用来计算能够归因于不治疗的病例的绝对数值。应该选择乘法尺度还是加法尺度, 需要视我们的目标而定。

1.4 随机变异性

迄今, 我们的例子都很简单, 只有 20 个人, 因而显得我们效应量度的计算都不怎么可信。在现实中, 我们需要处理的人群, 都比这 20 个人多上许多。

9 在现实中, 研究人员只能通过抽样从人群中收集信息。即使我们知道抽样群体里面每一个人的相关信息, 我们也不可能知道整个人群中治疗取值为 a 时死亡人数的精确比例。换句话说, 我们不能直接计算没有治疗时会死亡的概率 $\Pr[Y^{a=1} = 1]$, 我们只能估算这一概率。

(随机性的第一个来源: 抽样变异性)

在我们表 1.1 只有 20 个人的人群中, 我们用 $\widehat{\Pr}[Y^{a=0} = 1] = 10 / 20 = 0.50$ 表示如果没有治疗时死亡人数的比例。但在我们的抽样人群中, $\widehat{\Pr}[Y^{a=0} = 1]$ 并不一定等于整个更大的人群中没有治疗时死亡人数的比例。比如, 整个人群中, $\Pr[Y^{a=0} = 1] = 0.57$, 但在我们的抽样人群中, 因为抽样变异性带来的随机性, 我们得到 $\widehat{\Pr}[Y^{a=0} = 1] = 0.50$ 。我们将用抽样人群中的 $\widehat{\Pr}[Y^a = 1]$ 去估算治疗取值为 a 时整个人群中的 $\Pr[Y^a = 1]$ 。抽样人群中 \Pr 上的“帽子”表示抽样人群中的比例 $\widehat{\Pr}[Y^a = 1]$ 是对整个人群中对应比例 $\Pr[Y^a = 1]$ 的一个估计。在抽样变异性带来的误差是随机的, 因而大数定律成立的前提下, 我们将 $\widehat{\Pr}[Y^a = 1]$ 称作 $\Pr[Y^a = 1]$ 一致估计, 因为样本的人数越多, $\widehat{\Pr}[Y^a = 1]$ 就会越接近 $\Pr[Y^a = 1]$ 。

(随着样本量逐渐增大, $\hat{\theta} - \theta$ 逐渐接近接近 0, 这时候我们称 $\hat{\theta}$ 是 θ 的一致估计。)

因为我们不能直接计算整个人群中的概率 $\Pr[Y^a = 1]$, 而只能得到样本中的一致估计 $\widehat{\Pr}[Y^a = 1]$, 所以我们不能完全肯定地对是否有因果效应下结论。因此, 我们必须用统计的方法来衡量我们从样本中得到的证据, 进而判断因果零假设 $\Pr[Y^{a=1} = 1] = \Pr[Y^{a=0} = 1]$ 是否成立 (详见第十章)。

(注意: 这里一致估计中的“一致”, 和我们反事实结局中的“一致”, 含义并不一样)

除了抽样变异性, 随机性还可能来自其他地方, 比如某个人的反事实结局并不是固定的。我们将反事实结局 Y^a 定义为一个个体在治疗取值为 a 时的结局。在我们的例子中, 每个个体的反事实结局都是固定不变的。例如宙斯的是 $Y^{a=1} = 1$ 和 $Y^{a=0} = 0$, 也就是说, 如果宙斯接受治疗, 他 100% 会死, 如果他不接受治疗, 他死亡的可能性是 0。我们将这种情形称为命定的结局。但我们可以设想另外一种情景: 如果宙斯接受治疗, 他有 90% 的可能会死, 如果他不接受治疗, 他有 10% 的可能会死。在这种情景下, 宙斯的反事实结局是随机的, 因为不管治不治疗, 他死亡的概率都不是 1 或 0。相对应的, 表 1.1 中 $Y^{a=1}$ 和 $Y^{a=0}$ 的值就会变成概率。更进一步, 我们会认为每个个体的概率都不一样, 因为每个个体都是不一样的。我们将这种情形称为非命定的结局。这就如同物理中的量子力学, 和经典力学相反, 量子力学认为所有的结局都是不确定的。如果这个观点成立, 那不管我们收集多少关于宙斯的数据, 宙斯的结局总是不确定的。

(随机性的第二个来源: 非命定的反事实结局)

10 因此, 在我们的因果推断中, 随机性可能来源于抽样变异性或者非命定的反事实结局。在第十章之前, 我们会选择忽略随机性。我们会假设反事实结局都是命定的, 以及我们收集了人群中每一个人的数据。

在第十章之前, 我们的计算都会是完全确定的。在第十章, 我们会证明, 不管这个世界是随机的还是命定的, 我们在整个人群中的统计估计值和置信区间都会和因果效应等价。虽然在实际研究样本中, 因果效应均值的置信区间会受到反事实结局是否确定的影响, 但我们更感兴趣的是整个人群, 而不是一两个研究样本。

1.5 因果性与相关性

显然, 实际研究中的样本数据不会和我们表 1.1 的一样。在现实中, 我们不可能同时观测到宙斯治疗时的结局 $Y^{a=1}$, 和不治疗时的结局 $Y^{a=0}$ 。我们只能观测到宙斯治疗时, 或者不治疗时的结局。我们将观测到的结局记为 Y , 表 1.2 记录了我们观测到的每个个体的治疗取值 A 和结局 Y 。

表 1.2 的数据可以用来计算在治疗取值为 a 的人群中，结局为 Y 的人的比例。例如，在表 11.2 中，有 13 个人接受了治疗 ($A=1$)，其中有 7 个人死亡 ($Y=1$)。因此接受治疗的人的死亡风险为 $\Pr[Y=1|A=1]=7/13$ 。在这里，条件概率 $\Pr[Y=1|A=a]$ 被定义为治疗取值为 a 的人当中，结局为 Y 的人的比例。

当 $\Pr[Y=1|A=1]=\Pr[Y=1|A=0]$ 时，治疗 A 与结局 Y 相互独立，也就是 A 和 Y 不相关，或者说 A 不能预测 Y 。独立性用符号 $\perp\!\!\!\perp$ 表示， A 与 Y 相互独立写作 $A \perp\!\!\!\perp Y$ ，或者 $Y \perp\!\!\!\perp A$ 。独立性的其他等价表达如下：

$$(i) \quad \Pr[Y=1|A=1]-\Pr[Y=1|A=0]=0$$

$$(ii) \quad \frac{\Pr[Y=1|A=1]}{\Pr[Y=1|A=0]}=1$$

$$(iii) \quad \frac{\Pr[Y=1|A=1]/\Pr[Y=0|A=1]}{\Pr[Y=1|A=0]/\Pr[Y=0|A=0]}=1$$

(i)、(ii)、(iii) 中等式的左边分别表示相关性下的风险差、风险比和比值比。

$\Pr[Y=1|A=1] \neq \Pr[Y=1|A=0]$ ，则 A 和 Y 相关。在我们的人群中，因为 $\Pr[Y=1|A=1]=7/13 \neq \Pr[Y=1|A=0]=3/7$ ，所以我们的治疗和结局是相关的。这些相关性下的风险差、风险比和比值比（以及其他量度）衡量了相关性（如果存在的话）的大小。这些参数只是从不同的尺度去衡量同一个相关性，我们称这些参数为相关性量度。这些量度也会受到随机变异性的影响。在第十章之前，我们都会忽略这些统计上的问题。

对于一个二分结局，如果我们用 1 和 0 表示有和没有，其风险就等于其在人群中的均值。于是，不管结局 Y 是二分的还是连续的，我们可以把人群中相关性的定义写作 $E[Y|A=1] \neq E[Y|A=0]$ 。对于一个二分变量 A ， Y 和 A 不相关，当且仅当这两个变量在统计上不具关联性。

（对于连续性结局 Y ，我们将均值独立定义为 $E[Y|A=1]=E[Y|A=0]$ 。对于二分结局，独立和均值独立是同一个概念）

在我们 20 人的例子中，我们有如下发现：（1）当我们比较这 20 个人治疗和不治疗的反事实结局时，我们没有发现因果效应；（2）当我们比较能观测到的 13 个人治疗和 7 个人不治疗的结局时，我们发现治疗和结局相关。图 1.1 描绘了因果和相关的区别。这个人群（用方块表示）被分成治疗（白色）和不治疗（灰色）两部分。

12 因果性的定义需要我们把这个方块全部涂成白色或者灰色，然后再进行对比。而相关性只需要我们对比原方块中的白色和灰色两部分。也就是说，因果推断提出的问题是反事实世界究竟

是什么样的, 比如“如果所有人都接受治疗, 死亡风险是多少?”, 或者“如果所有人都不接受治疗, 死亡风险是多少?”。与之相比, 相关性则只是对现实世界提出问题, 比如“在接受治疗的人群中, 死亡风险是多少?”, 或者“在未接受治疗的人群中, 死亡风险是多少?”。

现在我们能用我们定义的数学符号来区分因果性和相关性。 $\Pr[Y = 1 | A = a]$ 是一个条件概率, 表示治疗取值为 a 的人中 (整个人群的一个子集) Y 的风险。 $\Pr[Y^a = 1]$ 则是一个非条件概率, 或者叫边缘概率, 表示的是整个人群中 Y^a 的风险。因此, 相关性是由整个人群中, 实际接受了不同治疗 ($A = 1, A = 0$) 的两个不相交子集的不同风险决定的。而因果性则是由同一个人群在不同治疗下 ($a = 1, a = 0$) 的不同风险决定的。在整本书中, 为了避免混淆, 我们会使用“因果效应”这个稍显冗长的词汇, 因为在大多数时候, “效应”这个词被许多人错误地用来表示相关性而不是因果性。

(区分因果性和相关性非常重要。假设医生更可能给心血管疾病高风险人群开阿司匹林, 服用阿司匹林和不服用阿司匹林对 5 年死亡率的因果性风险比是 0.5, 但是相关性风险比可能是 1.5。如果一个医生知道了相关性风险比, 但不清楚因果性风险比, 他可能就不再给高风险病人开药, 但这是一个错误的做法)

由此可见, 相关性和因果性存在根本区别。在我们表 1.2 的例子中, 治疗和结局存在相关性。但是, 如果我们知道反事实结局的话 (表 1.1), 我们知道这里没有因果性。在我们的例子中, 如果我们知道接受治疗的人都更羸弱, 那我们也就不会对因果性与关联性的不同感到惊讶。在第七章, 我们会进一步讨论“混杂”带来的不同。

因果推断需要如表 1.1 中那样的假想数据, 但我们只能观测到如表 1.2 的数据。下一个问题是: 我们该如何用现实中的数据进行因果推断? 下一章会提供一个可能的答案: 随机试验。

第一章精讲点与知识点

精讲点 1.1: 干扰 (原书第 5 页)

在我们对于反事实结局的定义中, 有一个隐含的前提假设: 一个个体在治疗取值为 a 时的结局, 不受到其他个体治疗取值的影响。比如, 在我们的例子中, 不管赫拉有没有接受治疗, 如果宙斯接受了治疗, 宙斯就会死。我们也可以假设, 赫拉接受了治疗让宙斯感到不爽 (毕竟神话故事中这两人常常相互拆桥), 以至于宙斯在接受手术后没能活下去; 而如果赫拉没有接受治疗的话, 宙斯就会感到开心, 因而宙斯接受手术后能够活下去。在这个场景中, 赫拉的治疗就会对宙斯的结局产生干扰。不同个体间的干扰其实在现实中很常见, 比如在传染病和教育研究中就经常存在干扰。现实生活中, 每个个体的结局, 都会因为社会性互动而受到其他人的干扰。

在干扰存在的时候，个体 i 的反事实结局 Y_i^a 的就不再是良定的，因为这个定义需要依赖其他个体的治疗。“心脏移植对宙斯的结局的因果效应”就不再良定，我们需要进一步说“赫拉没有接受治疗的时候，心脏移植对宙斯的结局的因果效应”，或者“赫拉接受治疗的时候，心脏移植对宙斯的结局的因果效应”。如果还有其他个体对宙斯的结局产生干扰，那我们对因果效应的定义也需要体现其他个体。无干扰这一假设，也被Cox叫作“个体间无交互”（1958），也是Rubin“稳重治疗”假设（Stable unit treatment value assumption, SUTVA）的一部分。详情请参考Struchiner (1995), Sobel (2006), Rosenbaum (2007)以及Hudgens和Halloran (2009)等人所著论文。除非特别说明，在本书中，我们都会假设没有干扰。

精讲点 1.2：不同形式的治疗（原书第 6 页）

在我们对于个体在治疗取值为 a 的反事实结局定义中，有一个隐含的前提假设：治疗 $A = a$ 时，只存在一种治疗形式。比如，我们说宙斯接受心脏移植手术后死亡，这句话就假设了所有的心脏移植手术都由同一个医生用同样的设备遵循同样的步骤进行，也就是心脏移植只有一种形式。但实际上，同一个治疗可能有不同的形式（比如不同的医生），因此，可能出现如果由医生甲进行手术，宙斯会死亡；但如果由医生乙进行手术，宙斯会存活。如果存在不同形式的治疗，个体 i 的反事实结局 Y_i^a 的就不是良定的，因为这个定义需要依赖 a 的不同形式。“心脏移植对宙斯的结局的因果效应”这一表述就不再清晰，我们需要进一步说“由甲进行手术时，心脏移植对宙斯的结局的因果效应”，或者“由乙进行手术时，心脏移植对宙斯的结局的因果效应”。如果还有其他成分（比如手术设备和地点）对宙斯的结局产生干扰，我们就需要把这些成分也包括在我们对因果效应的定义中。

就如无干扰假设（详见精讲点 1.1）一样，不存在多种治疗形式也是 Rubin“稳重治疗”假设（Stable unit treatment value assumption, SUTVA）的一部分。Robin 和 Greenland (2000) 论证了如果不同形式的治疗对结局有相同的因果效应，那反事实结局 Y_i^a 依然是良定的。VanderWeele (2009) 将这一点正式表述为“治疗差异无关紧要”假设。也即，假设就算治疗 $A = a$ 有多种形式，但它们都会导致同样的结局 Y_i^a 。我们将在第三章再次讨论这个问题。在本书中，除非特别说明，我们都会在本书中假设治疗差异无关紧要。

精讲点 1.3：需治数（原书第 8 页）

假设在一亿人中, 如果都接受治疗 ($a=1$) , 有二千万人会在 5 年内死亡; 如果都不接受治疗 ($a=0$) , 会有三千万人在 5 年内死亡。这些信息有以下等价表达:

- 因果性风险差是 $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1] = 0.2 - 0.3 = -0.1$
- 如果我们治疗全部一亿人, 那要比一亿人都不治疗少死亡一千万人。
- 我们可以通过治疗一亿人来拯救一千万人。
- 平均而言, 我们每治疗 10 个病人, 就能拯救 1 个人。

在至少能减少一个 $Y=1$ 病例时, 需要给予治疗 $a=1$ 的平均人数被称作需治数 (NNT)。在我们的例子中, NNT 是 10。在治疗能减少病例时 (即因果性风险差为负时), NNT 等于因果性风险差倒数的绝对值, 即:

$$NNT = \left| \frac{1}{\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]} \right|$$

对于提高病例数的治疗 (即因果性风险差为正), 我们可以相对应地将其定义为“需害数”。NNT 概念由 Laupacis, Sackett 和 Roberts (1988) 引进。如同因果性风险差, NNT 适用于人群之中。关于 NNT 作为一种效应量度的讨论, 可以参考 Grieve (2003) 的论文。

知识点 1.1: 人群中的因果效应 (原书第 7 页)

用 $E[Y^a]$ 表示人群中所有个体在治疗取值为 a 下的反事实结局均值。如果结局是离散的, 则 $E[Y^a] = \sum_y y p_{Y^a}(y)$, 其中 y 是随机变量 Y^a 的所有可能取值, $p_{Y^a}(\cdot)$ 是 Y^a 的概率密度函数, 也即 $p_{Y^a}(y) = \Pr[Y^a = y]$ 。如果结局是二分的, 则 $E[Y^a] = \Pr[Y^a = 1]$ 。如果结局是连续的, 则 $E[Y^a] = \int y f_{Y^a}(y) dy$, 其表示对随机变量 Y^a 的所有可能取值 y 进行积分, $f_{Y^a}(\cdot)$ 是 Y^a 的概率密度函数。对于离散和连续的结局, 一个通用表达式是 $E[Y^a] = \int y dF_{Y^a}(y)$, 其中 $F_{Y^a}(\cdot)$ 是 Y^a 的累积分布函数。如果对于 $a \neq a'$, 有 $E[Y^a] \neq E[Y^{a'}]$, 则我们说人群中因果效应的均值非零。

因果效应的均值被定义为两个反事实结局的均值的对比, 最常用来表示人群中的因果效应。然而, 人群中的因果效应, 也可以定义为反事实结局的其他特征的对比, 比如中位数、方差或者累积分布函数, 只要其能够标志反事实结局的边缘分布。比如, 我们可以用方差来定义人群中的因果效应, 写作 $\text{var}(Y^{a=1}) - \text{var}(Y^{a=0})$ 。对我们表 1.1 中的数据来说, 因为 $Y^{a=1}$ 和 $Y^{a=0}$ 在分布上完全一致 (20 个人中都有 6 个人死), 所以 $\text{var}(Y^{a=1}) - \text{var}(Y^{a=0}) = 0$ 。但是, 和均值不一样的点在

于, 人群中的方差, 并不等于个体因果效应的方差, 即 $\text{var}(Y^{a=1}) - \text{var}(Y^{a=0}) \neq \text{var}(Y^{a=1} - Y^{a=0})$ 。

比如, 对我们表 1.1 中的数据来说, $\text{var}(Y^{a=1} - Y^{a=0}) > 0 = \text{var}(Y^{a=1}) - \text{var}(Y^{a=0})$ 。我们可以通过随机试验的数据得到 $\text{var}(Y^{a=1}) - \text{var}(Y^{a=0})$, 但我们不能得到 $\text{var}(Y^{a=1} - Y^{a=0})$, 因为我们不能同时观测到每一个个体的 $Y^{a=1}$ 和 $Y^{a=0}$, 因而 $Y^{a=1}$ 和 $Y^{a=0}$ 的协方差不能确定。以上这些讨论不仅适用于方差, 也适用于其他非线性的参数值, 比如中位数等。

知识点 1.2: 非命定的反事实 (原书第 10 页)

如果反事实结局是非命定的, 那在治疗取值 a 下的结局均值 $E[Y^a] = \sum_y y p_{Y^a}(y)$, 其中 y 是随机变量 Y^a 的所有可能取值, 概率密度函数 $p_{Y^a}(\cdot) = E[Q_{Y^a}(\cdot)]$, $Q_{Y^a}(y)$ 治疗取值为 a 时 $Y = y$ 的随机概率。在我们正文描述的例子中, 对宙斯而言, $Q_{Y^{a=1}}(1) = 0.9$ 。如果结局是连续的, 则用积分代替表达式中的加权总和。

非命定的反事实结局不会给每个个体的 Y^a 赋一个具体的值, 但会给每个个体的 Y^a 赋一个概率分布函数 $Q_{Y^a}(\cdot)$ 。非命定的反事实结局只是命定的反事实结局的定义的一般化。此时, 人群反事实结局的均值 $E[Y^a] = E\left\{E\left[Y^a | \Theta_{Y^a}(\cdot)\right]\right\} = E\left[\int y d\Theta_{Y^a}(y)\right] = \int y dE\left[\Theta_{Y^a}(y)\right] = \int y dF_{Y^a}(y)$, 其中 $F_{Y^a}(\cdot) = E\left[\Theta_{Y^a}(\cdot)\right]$ 。

如果反事实结局是二分的且非命定的, 人群中的因果性风险比 $\frac{E[Q_{Y^{a=1}}(1)]}{E[Q_{Y^{a=0}}(1)]}$ 就等于个体因果效应 $Q_{Y^{a=1}}(1) / Q_{Y^{a=0}}(1)$ 的加权均值 $E[W\{Q_{Y^{a=1}}(1) / Q_{Y^{a=0}}(1)\}]$, 其中权重 $W = \frac{Q_{Y^{a=0}}(1)}{E[Q_{Y^{a=0}}(1)]}$ 。对于人群中每个个体, 因为结局不是命定的, 所以 $Q_{Y^{a=0}}(1) \neq 0$, $E[Q_{Y^{a=0}}(1)] \neq 0$ 。

第一章图表

Table 1.1

	$Y^{a=0}$	$Y^{a=1}$
Rheia	0	1
Kronos	1	0
Demeter	0	0
Hades	0	0
Hestia	0	0
Poseidon	1	0
Hera	0	0
Zeus	0	1
Artemis	1	1
Apollo	1	0
Leto	0	1
Ares	1	1
Athena	1	1
Hephaestus	0	1
Aphrodite	0	1
Cyclope	0	1
Persephone	1	1
Hermes	1	0
Hebe	1	0
Dionysus	1	0

Table 1.2

	A	Y
Rheia	0	0
Kronos	0	1
Demeter	0	0
Hades	0	0
Hestia	1	0
Poseidon	1	0
Hera	1	0
Zeus	1	1
Artemis	0	1
Apollo	0	1
Leto	0	0
Ares	1	1
Athena	1	1
Hephaestus	1	1
Aphrodite	1	1
Cyclope	1	1
Persephone	1	1
Hermes	1	0
Hebe	1	0
Dionysus	1	0

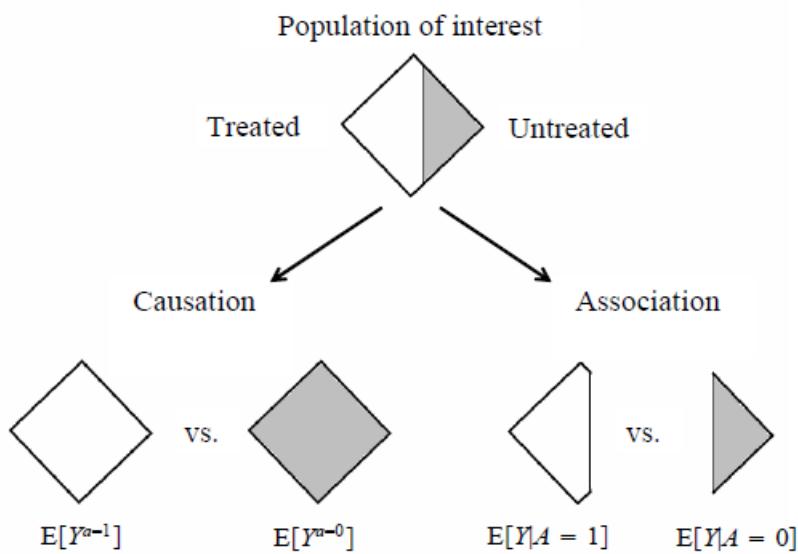


Figure 1.1

第二章 随机试验

13 你在马路中间抬头向天上看的时候，是否也会吸引其他行人向天上看？这个问题包含了所有因果问题的一个重要部分：我们想知道在特定人群中（比如 2019 年马德里的居民），一个行为（你抬头向上看）是否可以影响一个结局（其他人抬头向上看）。假设你需要用科学的方法来研究这个问题，你可能会想：“那我就站在人行道上，看见一个人靠近我就抛硬币。正面朝上我就抬头向上看，背面朝上我就直视前方。我将重复上千次。如果看到我抬头看后也跟着抬头看的行人，比没有看到我抬头看但依然抬头看的人多，那就能说我抬头看对他人是否抬头看有因果效应。我会雇一个助手观察在我抬头看时，行人都在做什么”。当你进行这项试验时，你发现 55% 的行人在看到你抬头看后也抬头看了，但是只有 1% 的人在你直视前方的时候抬头看。

我们把这一研究方式称作随机试验。这是一项试验，因为研究者（你）采取了某种行动（抬头向上看）；这个试验是随机的，因为是否行动是由一个随机机制（抛硬币）决定的。不是所有试验都是随机的。比如，当一个男人靠近时，你选择抬头看，但当一个女人靠近时，你选择直视前方。这样一来，是否行动就由某个固定的规则决定，不再是随机的。当你开展一项非随机试验时，你会发现你的结论不那么可信。如果你是否动作由行人的性别决定，批评者会说也许男性和女性的行为方式不一样（也许女性不喜欢抬头向上看），因此你在这两个性别组中得到的结果不可比。在这一章我们将讨论为什么随机试验能得到令人信服的因果推断。

2.1 随机

在现实世界我们不可能同时观测到宙斯的反事实结局 $Y^{a=1}$ 和 $Y^{a=0}$ ，我们只能知道他在治疗 A 下的结局 Y 。表 2.1 给出了现实世界中我们能观测到的实际情况，对于每一个个体，我们只能观测到一个反事实结局，另一个反事实结局则是缺失的。上一章我们已经讨论过，反事实结局的缺失给我们因果效应的计算带来了麻烦。我们最多只能用表 2.1 的数据来计算相关性量度。

（1923 年，Neyman 已经在随机试验中使用反事实概念来估算因果效应）

如同其他真实世界的研究一样，随机试验也会有缺失值的问题，就像表 2.1 中一样。然而，随机试验能够保证这些缺失值的出现都是随机的。因而，在随机试验中，尽管有缺失值，我们依然能计算不同的效应量度，或者，更严谨地说，我们依然能一致估计不同的效应量度。

14 假设图 1.1 代表近乎无限多的人。对每一个人，我们都扔一次硬币，如果是正面朝上，我们就把他归入到白色一边，如果是背面朝上，我们就把他归入到灰色一边。在这里，我们的硬币也许有些不均匀，因为白色组的人数明显比灰色组的要多，而不是一半一半。下一步，我们给所有白色组的人进行治疗 ($A=1$)，但给所有灰色组的人安慰剂 ($A=0$)。5 天后，这个试验结束

的时候, 我们计算每个组的死亡率, 得到 $\Pr[Y = 1 | A = 1] = 0.3$, $\Pr[Y = 1 | A = 0] = 0.6$ 。所以相关性风险比为 $0.3 / 0.6 = 0.5$, 相关性风险差为 $0.3 - 0.6 = -0.3$ 。我们假设这是一个完美的随机试验: 没有失访, 被试全程配合并遵循治疗指导, 治疗形式只有一种, 并且双盲设计 (详见第九章)。在现实中, 完美的随机试验是不存在的, 但是对于我们学习因果推断却非常有用。接下来我们会考虑更接近现实的随机试验。

假设试验人员误解了研究计划, 因而给灰色组而不是白色组进行了治疗。在试验结束后, 我们发现了这一错误。那两组治疗方案的颠倒会影响我们的结论吗? 一点都不会。我们依然能够在治疗组 (现在是灰色) 中得到 $\Pr[Y = 1 | A = 1] = 0.3$, 在安慰剂组 (现在是白色) 中得到 $\Pr[Y = 1 | A = 0] = 0.6$ 。相关性量度不会有任何改变。因为每个人都是被随机分配到白色组或者灰色组, 因而治疗组的死亡人数比例, 即 $\Pr[Y = 1 | A = 1]$, 不会因白色组或灰色组而有不同。当分组是随机的时候, 哪一组接受了治疗和我们的治疗结果 $\Pr[Y = 1 | A = 1]$ 无关。同样的推理适用于 $\Pr[Y = 1 | A = 0]$ 。用数学术语来说, 我们称这两组是可互换 (exchangeable) 的。

互换性意味着, 治疗组的治疗取值为 a 时的风险 $\Pr[Y^a = 1 | A = 1]$, 应该等于非治疗组的治疗取值为 a 时的风险 $\Pr[Y^a = 1 | A = 0]$, 不管这里的 a 是 0 还是 1。这些 (条件) 风险在不同治疗组中相等带来一个结论: 这些风险也必然等于整个人群在治疗 a 下的 (边缘) 风险, 即 $\Pr[Y^a = 1 | A = 1] = \Pr[Y^a = 1 | A = 0] = \Pr[Y^a = 1]$ 。因为在 $A = 1$ 或 $A = 0$ 时, 治疗为 a 时的反事实风险都是一样的, 所以我们说现实中的治疗 A 并不能影响反事实结局 Y^a 。或者说, 互换性也意味着反事实结局和实际治疗相互独立。于是, 对于所有 a 而言, 都有 $Y^a \perp\!\!\!\perp A$ 。我们高度重视随机性, 这是因为随机性能保证互换性。当治疗组和非治疗组可互换时, 有的人也会称治疗是外源性的, 因此有时外源性会被用作互换性的同义词。

(互换性: 对于任意 a , 有 $Y^a \perp\!\!\!\perp A$)

上一段讨论了, 在互换性成立的情况下, 白色组有治疗时的反事实风险等于整个人群有治疗时的反事实风险。但白色组有治疗时的风险并不是反事实的, 因为白色组就是治疗组。因此, 如果这是一个完美的随机试验, 我们就可以用白色组来计算整个人群在有治疗时的反事实结局

15 $\Pr[Y^{a=1} = 1]$, 我们有 $\Pr[Y^{a=1} = 1] = \Pr[Y = 1 | A = 1] = 0.3$ 。同样的推理也适用于非治疗组: 非治疗组的风险等于整个人群无治疗时的反事实风险, 也即 $\Pr[Y^{a=0} = 1] = \Pr[Y = 1 | A = 0] = 0.6$ 。由此我们可以得到因果性风险比是 0.5, 因果性风险差是 -0.3。在一个完美的随机试验中, 相关性就是因果性。

对于随机试验中的互换性 $Y^a \perp\!\!\!\perp A$, 我们还有另外一种解释。反事实结局 Y^a , 就像基因一样, 在一个人还未被分组的时候, 就已经固定并伴随这个人。 Y^a 表示了这个人如果治疗取值为 a 时的结局, 并不依赖于这个人稍后实际接受的治疗。因为治疗 A 是随机分配的, 它和 Y^a 没有任何关系, 就像和你的基因没有任何关系一样。 Y^a 和我们的基因的区别在于, 你只有在治疗 A 为 a 时, 才能知道 Y^a 的情况。

在我们继续之前, 请确保你能理解 $Y^a \perp\!\!\!\perp A$ 和 $Y \perp\!\!\!\perp A$ 的区别。 $Y^a \perp\!\!\!\perp A$ 表示的是互换性, 意思是反事实结局 Y^a 和我们观测到的实际治疗相互独立。换句话说, 这表示治疗组和非治疗组, 如果都接受同样的治疗方式 ($a = 0$ 或 $a = 1$), 这两组的死亡风险应该是一样的。但是 $Y^a \perp\!\!\!\perp A$ 并不意味着 $Y \perp\!\!\!\perp A$ 。比如, 在一个随机试验中, 互换性 $Y^a \perp\!\!\!\perp A$ 成立并且治疗对结局有因果效应, 这时在我们的观测中, $Y \perp\!\!\!\perp A$ 不成立, 因为治疗方式和观测到的结局相关。

(注意: $Y^a \perp\!\!\!\perp A$ 和 $Y \perp\!\!\!\perp A$ 的含义完全不一样)

(如果存在因果效应, 则有 $Y^{a=1} \neq Y^{a=0}$ 。因为 $Y = Y^A$, 当 $A = a$ 时, 反事实结局 Y^a 就是观测到的结局 Y^A , 而 Y^A 由 A 决定, 所以 Y 和 A 不独立)

在我们的表 2.1 中, 互换性成立吗? 为了回答这个问题, 我们需要验证 $Y^a \perp\!\!\!\perp A$ 对 $a = 1$ 和 $a = 0$ 是否成立。先说 $a = 0$ 。假设我们还知道表 1.1 的数据, 则在 13 个接受治疗的人中, 有
16 $\Pr[Y^{a=0} = 1 | A = 1] = 7/13$, 在剩下 7 个未接受治疗的人中, 有 $\Pr[Y^{a=0} = 1 | A = 0] = 3/7$ 。因为这两个风险不等, 且治疗组的更大, 所以我们说治疗组的预后更差, 并且治疗组和非治疗组不可互换。现在我们在数学上证明了, $Y^a \perp\!\!\!\perp A$ 对 $a = 0$ 不成立(你可以动手算算对 $a = 1$ 是否成立)。因而, 对于这段开头提出的问题, 答案显然是否定的。

但在现实中, 我们只有表 2.1 的数据, 而没有表 1.1 的反事实数据, 所以我们没有足够的数据去计算 $\Pr[Y^{a=0} = 1 | A = 1]$, 也因此, 我们不能回答互换性是否成立这一问题。让我们再进一步思考一下, 假设我们有表 1.1 的反事实数据, 并且已经证明了互换性不成立, 那我们能说, 我们的这项研究不是随机试验吗? 不能。因为两点原因, 我们不能得到不是随机试验这一结论。第一个原因, 你可能已经想到了, 我们只有 20 个人, 人数太少, 以至于我们不可能得到一个确定无疑的结论。在那么小的样本中, 抽样变异性随机误差几乎能解释所有发现。我们将在第十章讨论随机变异性。在此之前, 我们认为我们的样本足够大, 比如表中的每个人代表一亿个人。第二个原因, 就算我们的样本无限大, 且互换性不成立, 它也可能是一项随机试验, 虽然它可能和之前所述的随机试验不一样, 因为研究者可能用不止一枚硬币对被试进行随机分配。

17 2.2 条件随机

表 2.2 展示了从一项心脏移植的随机试验中得到的数据。除了治疗 A (1 表示接受手术, 0 表示没有) 和结局 Y (1 表示死亡, 0 没有) 之外, 表 2.2 还有预后因素 L 的数据 (1 表示重症, 0 轻症), 预后因素数据在分配治疗前测得。我们现在考虑两种互不相容的试验设计, 然后讨论一下表 2.2 的数据是否来自其中一种。

在第一种设计中, 我们随机将人数的 65% 分配到治疗组, 这解释了为什么我们 20 个人中有 13 个人在治疗组。在第二种设计中, 我们先将病人分成重症 ($L=1$) 与轻症 ($L=0$), 然后我们随机从重症中选出 75% 的被试、从轻症中选出 50% 的被试分配到治疗组。这解释了为什么 12 个重症被试中有 9 个在治疗组, 而 8 个轻症被试中有 4 个在治疗组。

这两个试验设计都是随机试验。第一种设计也是上一小节 2.1 中描述的随机试验。在这种试验设计下, 我们仅用一枚硬币来决定被试的分组 (比如背面是治疗组, 正面是非治疗组), 也许这枚硬币不均匀, 有 65% 的可能背面朝上。在第二种设计下, 我们用了不止一枚硬币来决定分组。这里, 我们用了两枚硬币, 重症被试一枚 (75% 概率背面朝上), 轻症被试一枚 (50% 概率背面朝上)。我们将第二种试验设计称作条件随机试验 (conditional randomized experiment), 因为我们的随机分组概率由另一个变量 L 决定。我们将第一种试验设计称作边缘随机试验 (marginal randomized experiment), 因为我们对所有人仅用了一个无条件概率进行随机分组。

如上一小节讨论的, 边缘随机试验能保证治疗组和非治疗组的互换性:

$$\Pr[Y^a = 1 | A = 1] = \Pr[Y^a = 1 | A = 0], \text{ 或对所有 } a, \text{ 有 } Y^a \perp\!\!\!\perp A$$

与之相比, 条件随机试验不能保证治疗组和非治疗组的互换性, 因为每个组有不良预后的人数比例不一样。

在表 2.2 中, 治疗组有 69% 的重症, 非治疗组有 43% 的重症, 因而这个数据不可能是从边缘随机试验中出来的。这种不均衡意味着, 如果治疗组的人没有接受治疗, 他们的死亡风险可能比非治疗组的人更高。也就是说, 治疗 A 与无治疗时的反事实死亡风险有关, 所以互换性 $Y^a \perp\!\!\!\perp A$ 不成立。我们可以说这个试验是随机试验, 但这是一个根据 L 进行随机分组的随机试验。

我们例子中的条件随机试验, 简单来看其实是两个边缘随机试验的组合。我们先在重症被试 ($L=1$) 中开展一次随机试验, 再在轻症被试 ($L=0$) 中开展一次随机试验。让我们先来考虑

第一个在重症中的随机试验。在这个边缘随机试验中, 治疗组和未治疗组在分组时都是重症, 两

18 个组在治疗值为 a 时的反事实死亡风险都应该相等。用数学表达, 就是:

$$\Pr[Y^a = 1 | A = 1, L = 1] = \Pr[Y^a = 1 | A = 0, L = 1], \text{ 或对所有 } a, \text{ 有 } Y^a \perp\!\!\!\perp A | L = 1$$

这里, $Y^a \perp\!\!\!\perp A | L = 1$ 表示在 $L = 1$ 的时候, Y^a 和 A 相互独立。同理, 随机性也能保证在轻症被试中, 治疗组和非治疗组是可互换的, 也即 $Y^a \perp\!\!\!\perp A | L = 0$ 。当 $Y^a \perp\!\!\!\perp A | L = l$ 对所有的取值 l 都成立时, 我们会简写作 $Y^a \perp\!\!\!\perp A | L$ 。因此, 虽然条件随机不能保证无条件的互换性 (或称边缘互换性) $Y^a \perp\!\!\!\perp A$, 但能保证在所有 L 的取值下的有界互换性 $Y^a \perp\!\!\!\perp A | L$ 。总而言之, 随机分配过程要么产生边缘互换性 (第一种试验设计), 要么产生有界互换性¹ (第二种试验设计)。

(有界互换性: 对于所有 a , 有 $Y^a \perp\!\!\!\perp A | L$)

我们已经知道了怎么在边缘互换性下计算各种效应量度。在边缘随机试验中, 因为互换性保证了治疗为 a 时的反事实风险 $\Pr[Y^a = 1]$ 等于现实中治疗为 a 时我们观测到的风险

$$\Pr[Y = 1 | A = a], \text{ 所以因果性风险比就等于相关性风险比, 即 } \frac{\Pr[Y^{a=1} = 1]}{\Pr[Y^{a=0} = 1]} = \frac{\Pr[Y = 1 | A = 1]}{\Pr[Y = 1 | A = 0]}.$$

如果表 2.2 的数据是通过一项边缘随机试验得到的, 那么我们就能直接算出因果性风险比是

$$\frac{7/12}{3/7} = 1.26。现在我们的问题是, 在条件随机试验中, 我们应该如何计算因果性风险比。回想$$

一下, 条件随机试验可以看成人群不同子集中两个或多个边缘随机试验的组合。因而, 我们有两种选择。

(如果 $A = 1$, 那么 $Y^{a=0}$ 就会缺失; 如果 $A = 0$, 那么 $Y^{a=1}$ 就会缺失。在边缘随机试验中, 如果 $\Pr[A = a | L, Y^{a=1}, Y^{a=0}] = \Pr[A = a]$, 则我们称数据完全随机缺失 (*Missing completely at random, MCAR*) ; 如果 $\Pr[A = a | L, Y^{a=1}, Y^{a=0}] = \Pr[A = a | L, Y^a]$, 即条件 $(L, Y^{a=0}, Y^{a=1})$ 下 $A = a$ 的概率由我们观测到的数据决定, 则我们称数据随机缺失 (*Missing at random, MAR*)。其余情况称作非随机缺失 (*not missing at random, NMAR*)。MCAR, MAR 和 NMAR 的概念在 1976 年由 Rubin 引进)

第一种选择, 我们可以在人群的每一个子集中计算因果效应的均值。因为在每一个子集中, 都是边缘随机试验, 因而这一子集中的因果性风险比 $\frac{\Pr[Y^{a=1} = 1 | L = l]}{\Pr[Y^{a=0} = 1 | L = l]}$, 就等于这一子集中的相

¹ 英文原文 conditional exchangeability。Conditional 一词在数学和统计中经常译作“条件”, 比如“条件概率”中的“条件”就是 conditional 一词。这里, 若译作“条件互换性”, 可能在中文中产生歧义。因此译作“有界互换性”。

作者: Miguel A. Hernan, James M. Robins;

翻译: 罗家俊

关性风险比 $\frac{\Pr[Y=1|A=1,L=l]}{\Pr[Y=1|A=0,L=l]}$ 。我们将这种计算每一子集中因果效应的方法称作分层分析。

我们可能会发现每一层的因果风险比不尽相同，在这种情况下，我们会说治疗的效应被 L 修饰，或者存在 L 的效应修饰。

(我们在第四章详细讨论分层分析和效应修饰)

第二种选择，我们可以像之前做的那样，计算在整个人群中因果效应 $\frac{\Pr[Y^{a=1}=1]}{\Pr[Y^{a=0}=1]}$ 的均值。我

们应该计算每一分层的因果效应，还是计算整个人群的因果效应均值，将由我们在实践上和理论上的种种目标决定，我们将在第四章和全书的第三部分讨论这个问题。举一个例子，如果你不能收集 L 的相关数据（也许因为花费高昂），因此你的治疗分配不会依赖于 L ，这种时候也许你对整个人群的因果效应均值，而不是特定层的因果效应感兴趣。第四章前，我们将把我们的精力放在整个人群的因果效应均值上。下面两个小节，我们将讨论怎么在条件随机试验中计算整个人群的因果效应均值。

19 2.3 标准化

我们假想的心脏移植研究是一项条件随机试验：轻症 ($L=0$) 的 8 个被试有 50% 的概率被分配到治疗组，重症 ($L=1$) 的 12 个被试有 75% 的概率被分配到治疗组。首先，然我们先关注 8

个轻症被试。在轻症被试中，治疗组的死亡风险是 $\Pr[Y=1|L=0,A=1]=\frac{1}{4}$ ，非治疗组的死亡

风险是 $\Pr[Y=1|L=0,A=0]=\frac{1}{4}$ 。因为治疗组分配在轻症 ($L=0$) 被试中是随机的，即 $Y^a \perp\!\!\!\perp A|L=0$ ，所以实际治疗组观测到的风险就应该等于所有人都被治疗时的反事实风险，即

$\Pr[Y=1|L=0,A=1]=\Pr[Y^{a=1}=1|L=0]$ ；实际未治疗组观测到的死亡风险等于所有人都未被治疗时的反事实风险 $\Pr[Y=1|L=0,A=0]=\Pr[Y^{a=0}=1|L=0]$ 。同理，在重症被试中，我们有

$\Pr[Y=1|L=1,A=1]=\Pr[Y^{a=1}=1|L=1]=\frac{2}{3}$ ， $\Pr[Y=1|L=1,A=0]=\Pr[Y^{a=0}=1|L=1]=\frac{2}{3}$ 。

我们的目标是计算整个人群中的因果性风险比 $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]$ 。在上一段中我们已经知道有治疗时的死亡风险，在 8 个轻症 ($L=0$) 被试中是 $\frac{1}{4}$ ，在 12 个重症 ($L=1$) 被试中

是 $\frac{2}{3}$ 。因此，整个人群在有治疗时的死亡风险，应该是重症和轻症这两个组的死亡风险根据每组

人数的加权平均。因为有 40% 的被试是轻症 (8 个), 有 60% 的被试是重症 (12 个), 所以加权平

均是 $\Pr[Y^{a=1} = 1] = \frac{1}{4} \times 0.4 + \frac{2}{3} \times 0.6 = 0.5$ 。同理, 无治疗时的死亡风险 $\Pr[Y^{a=1} = 1]$ 也等于 0.5。

因而整个人群中的因果性风险比是 $0.5 / 0.5 = 1$ 。

简言之, 边缘反事实风险 $\Pr[Y^a = 1]$ 是每一层的反事实风险 $\Pr[Y^a = 1 | L = 0]$ 和 $\Pr[Y^a = 1 | L = 1]$ 的加权平均, 其中权重等于每一层 ($L = 0$ 或 $L = 1$) 的人数占比, 即 $\Pr[Y^a = 1] = \Pr[Y^a = 1 | L = 0] \Pr[L = 0] + \Pr[Y^a = 1 | L = 1] \Pr[L = 1]$ 。如果 L 有不止两层, 则有 $\Pr[Y^a = 1] = \sum_l \Pr[Y^a = 1 | L = l] \Pr[L = l]$, 其中 \sum_l 表示对所有可能取值 l 取和。在有界互换性下, 我们可以用观测到的风险 $\Pr[Y = 1 | L = l, A = a]$ 去替代反事实风险 $\Pr[Y^a = 1 | L = l]$ 。因而上述公式可进一步写作 $\Pr[Y^a = 1] = \sum_l \Pr[Y = 1 | L = l, A = a] \Pr[L = l]$ 。等号左边表示的是未观测到的反事实风险, 等号右边表示的是我们用观测到的 L 、 A 和 Y 计算得到的数值。当一个反事实的数值能用已观测到的数据分布来表示时, 就如同我们的这个例子, 我们称这个反事实的数值是可识别的, 反之, 我们称这个反事实的数值是不可识别的。

上述描述的方法, 在流行病学、人口统计学等学科中称为标准化。因果性风险比中的分子 $\sum_l \Pr[Y = 1 | L = l, A = 1] \Pr[L = l]$, 就是用整个人群作为标准进行标准化后得到的风险。有界互换性成立时, 标准化风险能被释作人群中所有人都被治疗时观测到的 (反事实) 风险。

(标准化均值 $\sum_l E[Y | L = l, A = a] \times \Pr[L = l]$)

20 治疗组和非治疗组中的标准化风险, 分别等于有治疗时和无治疗时的反事实风险。因此, 因果性风险比就能用标准化的数值计算, 写作:

$$\frac{\Pr[Y^{a=1} = 1]}{\Pr[Y^{a=0} = 1]} = \frac{\sum_l \Pr[Y = 1 | L = l, A = 1] \Pr[L = l]}{\sum_l \Pr[Y = 1 | L = l, A = 0] \Pr[L = l]}$$

2.4 逆概率加权

上一小节我们通过标准化计算了条件随机试验中的因果性风险比, 这一小节我们将通过逆概率权重来计算因果性风险比。表 2.2 的数据可以用图 2.1 的树状图来表示。图从左至右表示了我们试验的时间顺序。最左边的圆包含了第一次分岔: 8 个轻症 ($L = 0$) 被试和 12 个重症 ($L = 1$) 被试, 括号里的数字表示轻症和重症的概率。接下来, 在 $L = 0$ 这一枝上又分出两枝, 4 个被试在非治疗组 ($A = 0$), 4 个被试在治疗组 ($A = 1$)。非治疗组的条件概率也在括号中

作者: Miguel A. Hernan, James M. Robins;

翻译: 罗家俊

表示, 为 $\Pr[A = 0 | L = 0] = 4 / 8 = 0.5$ 。最右边最上方的圆表示, 在这一枝 ($L = 0, A = 0$) 的 4 个被试中, 3 个存活 ($Y = 0$), 1 个死亡 ($Y = 1$), 即 $\Pr[Y = 0 | L = 0, A = 0] = 3 / 4$,

$\Pr[Y = 1 | L = 0, A = 0] = 1 / 4$ 。图中其他枝干的分析方法同理。圆中的分岔表示与治疗无关的变量。我们接下来用这个树状图来计算我们的因果性风险比。

(图 2.1 是一个用来表示条件随机试验的“完全随机因果性阐释结构树状图”(Fully randomized causally interpreted structured tree graph, FRCISTG)。怎么样, 我们是不是可以拿最烂命名奖?)

21 因果性风险比的分母 $\Pr[Y^{a=0} = 1]$ 是整个人群都无治疗时的反事实死亡风险。现在让我们来计算这个风险。在图 2.1 中, 8 个轻症 ($L = 0$) 被试中有 4 个在非治疗组, 他们中有 1 个死亡。如果这 8 个人都在非治疗组, 会有多少人死亡? 2 个。因为总人数翻倍, 因果性风险比不会改变, 所以死亡人数也会翻倍。同理, 在图 2.1 中, 12 个重症 ($L = 1$) 被试中有 3 个在非治疗组, 他们中有 2 个死亡。如果所有重症被试都在非治疗组, 将会有 8 个死亡。因而, 如果整个人群中的 20 个人都在非治疗组, 会有 $2 + 8 = 10$ 个人死亡。所以因果性风险比的分母

$\Pr[Y^{a=0} = 1] = 10 / 20 = 0.5$ 。图 2.2 的第一个树状图显示了所有人没有被治疗时的情况。当然, 以上计算都需要在 $L = 0$ 和 $L = 1$ 时互换性成立, 也即在每一层中, 治疗组在无治疗时的反事实死亡风险, 和非治疗组观测到的风险一样。

22 因果性风险比的分子 $\Pr[Y^{a=1} = 1]$ 是整个人群都有治疗时的反事实死亡风险。和上一段同理, 我们能算出在整个人群中风险是 $\Pr[Y^{a=1} = 1] = 10 / 20 = 0.5$ 。这也需要在 $L = 0$ 和 $L = 1$ 时互换性成立。将两个结果合起来, 我们有因果性风险比 $\frac{\Pr[Y^{a=1} = 1]}{\Pr[Y^{a=0} = 1]} = \frac{0.5}{0.5} = 1$, 这就是我们的最后结果。

让我们来思考一下为什么这个方法是正确的。图 2.2 的两个树状图是当整个人群中的所有人都接受治疗或未接受治疗时的假想人群。这两个假想人群在有界互换性成立的情况下会是正确的。这两个假想人群可以合在一起创造了一个更大的假想人群, 其中的每个个体都同时接受和未接受治疗。这个假想人群是原人群人数的两倍, 我们称之为虚拟人群 (pseudo-population)。图 2.3 展示了整个虚拟人群。在原人群, 如果有界互换性 $Y^a \perp\!\!\!\perp A | L$ 成立, 则在虚拟人群中, 治疗组和非治疗组是 (无条件) 可互换的, 这是因为虚拟人群中 L 和 A 相互独立。也就是说, 在虚拟人群中, 相关性风险比就等于因果性风险比。

上述方法被称作逆概率加权 (inverse probability weighting, 逆概率也简写作 IP)。我们可以用图 2.1 中 $L = 0$ 的 4 个非治疗组被试来解释为什么称为逆概率。这 4 个被试被用来对应

图 2.3 中虚拟人群里的 8 个人, 也就是说, 这 4 个人每个人的权重是 2, 也就等于 $1/0.5$ 。从图 2.1 中, 我们可知 $L=0$ 时被分配到非治疗组的概率是 0.5。同理, $L=1$ 的 9 个治疗组被试被用来模拟虚拟人群中的 12 个人, 而每个人的权重是 $1.33 = 1/0.75$ 。从图 2.1 中, 我们可知 $L=1$ 时被分配到治疗组的概率是 0.75。因此可以说, 虚拟人群是通过给原人群的每个个体进行加权得到的, 每个个体的权重等于他被分配到实际中其所在组的概率的倒数。图 2.3 中给出了每个组的逆概率权重。

(1952 年, Horvitz 和 Thompson 首次提出在抽样概率不均匀的调查中进行逆概率加权计算)

(逆概率权重: $W^A = 1/f[A|L]$)

在我们的例子中, 逆概率加权得到的结果和标准化得到的结果是一样的——因果性风险比都是 1。这不是巧合, 标准化和逆概率加权在数学上是等价的 (详见知识点 2.3)。实际上, 两种方法都可以视作我们创建了一个虚拟人群, 其中每个个体的治疗取值都是 a 。每个方法用一套不同的概率去创建这个虚拟人群: 逆概率加权用的是给定协变量 L 后接受治疗 A 的条件概率 (如图 2.1 所示), 标准化则是用协变量 L 的概率和给定 A 和 L 后结局 Y 的概率。

因为标准化和逆概率加权两种方法都是模拟了一个不根据变量 (一个或多个) L 决定治疗分组概率的虚拟人群, 所以我们也说这些方法调整了变量 L , 有时我们也说控制了 L 。不过“分析中控制”和“物理上控制”在随机试验中是两个不同的概念。标准化和逆概率加权都可以适用于结局是连续的条件随机试验当中 (详见知识点 2.3)。

那这本书接下去还要讲什么呢? 我们已经有了研究设计 (理想的随机试验), 将其余适当的分析方法 (标准化和逆概率加权) 结合, 从而能计算因果效应的均值。不幸的是, 随机试验在很多方面经常是不合伦理规范的, 不实际的。比如, 伦理委员会也许不会通过我们假想的心脏移植研究。在现实中, 心脏源是紧缺的, 整个社会更倾向将心脏移植给更可能受益的人, 而不是随机地分配。就算不考虑伦理问题, 可行性也是一个问题。在我们的研究中, 双盲设计基本上是不可能的, 同时被分配到治疗组的被试也许因为没有适配的心脏源只能放弃手术。就算这些可行性问题都解决了, 我们还要考虑时效性问题。这项试验需要耗费数年完成, 但也许我们需要在这之前得到结论。因而, 在现实中, 另一个不错的选择, 是进行观察性研究 (observational study)。

第二章精讲点与知识点

精讲点 2.1: 交叉试验 (原书第 16 页)

现在我们想估算闪电魔法 A 对宙斯血压 Y 的个体因果效应。反事实结局 $Y^{a=1}$ 表示宙斯使用闪电魔法后的血压情况, 如果升高则等于 1, 没有升高则等于 0, $Y^{a=0}$ 同理。假设只有在我们要求

作者: Miguel A. Hernan, James M. Robins;

翻译: 罗家俊

的情况下, 宙斯才会使用闪电魔法。昨天早上我们要求宙斯使用了一次闪电魔法 ($a=1$) , 他的血压在之后升高了。今天早上我们要求宙斯不要使用闪电魔法 ($a=0$) , 他的血压没有升高。

在这里, 我们开展了一项交叉试验, 我们依序观察同一个人在不同治疗情况下的结局。有些人可能认为, 因为我们观察到了反事实结局 $Y^{a=1} = 1$ 和 $Y^{a=0} = 0$, 所以我们可以说使用闪电魔法对宙斯的血压有因果效应。然而, 除非下一段中的 3 个强假设为真, 否则这个结论是不正确的。

交叉试验中, 个体在不同的两个时间 $t=0$ 和 $t=1$ 被观察。用 A_{it} 表示个体 i 在时间 t 接受的治疗。用 $Y_{i1}^{a_0, a_1}$ 表示个体 i 在 $t=0$ 时治疗为 a_0 以及 $t=1$ 时治疗为 a_1 的 (命定的) 反事实结局 (结局时间 $t=1$)。 $Y_{i0}^{a_0}$ 表示个体 i 在 $t=0$ 时治疗为 a_0 的 (命定的) 反事实结局 (结局时间 $t=0$)。如果下列 3 个条件成立, 则我们可以识别个体的因果效应 $Y_{it}^{a_t=1} - Y_{it}^{a_t=0}$ 。1) 上一次的治疗不对下一次的治疗造成影响, 即 $Y_{it=1}^{a_0, a_1} = Y_{it=1}^{a_1}$; 2) 治疗对个体的因果效应不受时间的影响, 即

$$Y_{it}^{a_t=1} - Y_{it}^{a_t=0} = \alpha_i; \quad 3) \text{ 无治疗对个体的因果效应不受时间的影响: } Y_{it}^{a_t=0} = \beta_i.$$

如果一个个体在 $t=1$ 的时候有治疗 ($A_{i1}=1$) 但在 $t=0$ 的时候无治疗 ($A_{i0}=0$), 则由一致性, 得到个体因果效应为 $Y_{i1} - Y_{i0} = Y_{i1}^{a_1=1} - Y_{i0}^{a_0=0} = Y_{i1}^{a_1=1} - Y_{i1}^{a_1=0} + Y_{i1}^{a_1=0} - Y_{i0}^{a_1=0} = \alpha_i + \beta_i - \beta_i = \alpha_i$ 。

同理, 如果 $A_{i1}=0$ 且 $A_{i0}=1$, 则 $Y_{i1} - Y_{i0} = \alpha_i$ 就是该个体的因果效应。

条件 1) 也意外着结局 $Y_{it}^{a_t}$ 突然出现, 并在下一个时间它的影响又要完全消失。因此, 交叉试验不能用于心脏移植、死亡等不可逆行为或结局的研究。参见精讲点 3.2。

精讲点 2.2: 风险期 (原书第 21 页)

我们将风险定义为某个时间段内出现特定结局的人数比例。比如, 治疗组的 5 天死亡风险 $\Pr[Y=1 | A=1]$ 就是接受治疗的被试在 5 天的死亡人数比例。本书中, 我们在第一次介绍一个风险的时候, 会说明特定时间段 (比如 5 天), 之后为了方便我们会省略这个时间描述。比如, 我们会说“死亡风险”而不是“5 天内死亡风险”。

不过注明风险期是非常重要的。假设我们进行了一项随机试验来量化在感染了鼠疫细菌的老人中抗生素疗法对死亡率的因果效应。一名研究者分析了数据, 并得到结论说因果性风险比是 0.05, 即平均而言, 抗生素降低了 95% 的死亡率。另一名研究者也分析了数据, 不过却得到结论说因果性风险比是 1, 也即抗生素对死亡率没有影响。这两名研究者都可能是对的。第一名研究者计算的是 1 年内的风险比, 但第二名研究者计算的是 100 年内的风险比。100 年内的风险比当

然会是 1, 因为不管有没有接受治疗, 这群被试都会在 100 年内去世。当我们说一项治疗对死亡率有因果效应的时候, 我们指的推迟了死亡, 而不是阻止了死亡。

知识点 2.1: 完全互换性与均值互换性 (原书第 15 页)

随机性使得对所有 a , 都有 Y^a 共同独立于 A , 这是互换性 $Y^a \perp\!\!\!\perp A$ 的充分条件, 但不是必要条件。更严谨地说, 让 $\mathcal{A} = \{a, a', a'', \dots\}$ 表示人群中的所有治疗取值, $Y^{\mathcal{A}} = \{Y^a, Y^{a'}, Y^{a''}, \dots\}$ 表示所有的反事实结局。根据随机性, 有 $Y^{\mathcal{A}} \perp\!\!\!\perp A$ 。我们将这种共同独立性称作完全互换性。对于一个二分治疗 $\mathcal{A} = \{0, 1\}$, 完全互换性就是 $(Y^{a=1}, Y^{a=0}) \perp\!\!\!\perp A$ 。

如果结局和治疗都是二分的, 互换性 $Y^a \perp\!\!\!\perp A$ 可以写成 $\Pr[Y^a = 1 | A = 1] = \Pr[Y^a = 1 | A = 0]$, 或等价地, $E[Y^a | A = 1] = E[Y^a | A = 0]$ 。我们将这后一个等式称作均值互换性。对连续的结局, 互换性 $Y^a \perp\!\!\!\perp A$ 是均值互换性 $E[Y^a | A = 1] = E[Y^a | A = 0]$ 的充分条件, 但不是必要条件, 因为分布中的其他参数, 比如方差, 可能并不独立于治疗。

要证明 $E[Y^a] = E[Y | A = a]$, 我们不需要完全互换性 $Y^{\mathcal{A}} \perp\!\!\!\perp A$ 或互换性 $Y^a \perp\!\!\!\perp A$ 。只需要均值互换性就足够了。正如正文中所述, 证明分成两步。首先, 根据一致性有 $E[Y | A = a] = E[Y^a | A = a]$, 其次, 根据均值互换性有 $E[Y^a | A = a] = E[Y^a]$ 。对于二分结局来说, 完全互换性和均值互换性是一样的概念。因此, 我们在本章中简称作互换性。

知识点 2.2: 逆概率权重的正式定义 (原书第 23 页)

一个个体的逆概率权重由他的治疗 A 和协变量 L 取值同时决定。比如, 一个治疗组中 $L = l$ 的被试, 他的权重是 $1 / \Pr[A = 1 | L = l]$; 而一个非治疗组中 $L = l'$ 的被试, 他的权重是 $1 / \Pr[A = 0 | L = l']$ 。我们可以用 A 的概率密度函数来表示所有个体的权重。给定 L 时 A 的条件概率密度函数记作 $f_{A|L}[a | l]$, 或者简写作 $f[a | l]$ 。如果 A 和 L 是离散的, $f[a | l]$ 就是条件概率 $\Pr[A = a | L = l]$ 。在条件随机试验中, 对满足 $\Pr[L = l] \neq 0$ 的 l , 有 $f[a | l] > 0$ 。

因为对每个人而言, 其权重的分母都是这个在 A 和 L 特定值下的条件概率密度, 因此这个分母就能用表示随机变量的 A 和 L 来表示 (而不是表示固定取值的 a 和 l), 写作 $f[A | L]$ 。这个表示被用来正式定义逆概率权重 $W^A = 1 / f[A | L]$ 。我们需要一个统一的符号, 而 $\Pr[A = a | L = l]$ 并不适用于所有情况, 比如 A 是连续变量的时候就不适用。

知识点 2.3: 逆概率加权和标准化的等价证明 (原书第 24 页)

假设 A 是离散的, 且取值有限。对满足 $\Pr[L = l] \neq 0$ 的 l , 有 $f[a|l] > 0$ 。条件随机试验能够保证正数性 (positivity) 成立。如果正数性成立, Y 的标准化均值为

$$\sum_l E[Y | A = a, L = l] \Pr[L = l], Y \text{ 的逆概率加权均值为 } E\left[\frac{I(A=a)Y}{f(A|L)}\right], \text{ 其中指示函数 } I(A=a)$$

在 $A = a$ 时取 1, 其他情况下取 0。我们现在来证明正数性成立时标准化均值等于逆概率均值。根

据期望的定义, $E\left[\frac{I(A=a)Y}{f(A|L)}\right] = \sum_l \frac{1}{f[a|l]} \{E[Y | A = a, L = l] f[a|l] \Pr[L = l]\}$
 $= \sum_l \{E[Y | A = a, L = l] \Pr[L = l]\}$, 这个等式中, 在第一步, 我们没必要在所有 A 的可能取值中进行加总, 因为对于不是 a 的 a' , 都有 $I(a' = a) = 0$; 在最后一步, 我们约掉了分子和分母中同时有的 $f[a|l]$ 。这个证明将 A 和 L 视作离散的。对于连续的变量, 我们只需用积分符号替代求和符号即可。

这个证明没有用到反事实的概念。然而, 当我们进一步假设有界互换性成立时, 逆概率加权均值和标准化均值都等于反事实均值 $E[Y^a]$ 。对这个命题, 以下我们提供两种证明。第一种方法, 我们在正文中已经证明了 $E[Y^a]$ 和标准化均值相等, 有:

$$E[Y^a] = \sum_l E[Y^a | L = l] \times \Pr[L = l] = \sum_l E[Y^a | L = l, A = a] \times \Pr[L = l] = \sum_l E[Y | L = l, A = a] \Pr[L = l]$$

这里的第二个等号是根据有界互换性和正数性得到, 第三个等号是根据一致性得到, 第二种方法, 我们将会证明 $E[Y^a]$ 和逆概率加权均值相等。通过一致性得到

$$E\left[\frac{I(A=a)Y}{f(A|L)}\right] = E\left[\frac{I(A=a)}{f(A|L)}Y^a\right]。接下来, 因为正数性保证了 $f[a|l] \neq 0$, 所以有$$

$$\begin{aligned} E\left[\frac{I(A=a)}{f(A|L)}Y^a\right] &= E\left\{E\left[\frac{I(A=a)}{f(a|L)}Y^a | L\right]\right\} = E\left\{E\left[\frac{I(A=a)}{f(a|L)} | L\right]E[Y^a | L]\right\} \text{ (通过有界互换性)} \\ &= E\left\{E[Y^a | L]\right\} \text{ (因为 } E\left[\frac{I(A=a)}{f[a|L]} | L\right] = 1 \text{)} = E[Y^a] \end{aligned}$$

当治疗取值是连续的时候 (虽然在条件随机试验中不太可能) , $E\left[\frac{I(A=a)}{f(A|L)}Y\right]$ 不再等于 $\sum_l E[Y|L=l, A=a]Pr[L=l]$, 因此就算互换性成立, $E\left[\frac{I(A=a)}{f(A|L)}Y\right]$ 对 $E[Y^a]$ 的估计也是有偏的。为了理解这一点, 我们可以将 $f[a|l]$ 视作 $L=l$ 时 A 的一种条件概率密度 (勒贝格测度下), 此时 $E\left[\frac{I(A=a)}{f[a|l]}|L=l\right]$ 等于 0 而不是 1。另一方面, 如果我们继续将 $f[a|l]$ 视作 $Pr[A=a|L=l]$, 则分母 $f[a|L=l]$ 就会是 0, 因此正数性不成立。在 12.4 小节我们将讨论如何把逆概率权重推广到连续治疗中。在知识点 3.1 中, 我们将讨论如果正数性不成立, 以上结论都不成立, 即使 A 是离散的。

第二章图表

Table 2.1

	A	Y	Y^0	Y^1
Rheia	0	0	0	?
Kronos	0	1	1	?
Demeter	0	0	0	?
Hades	0	0	0	?
Hestia	1	0	?	0
Poseidon	1	0	?	0
Hera	1	0	?	0
Zeus	1	1	?	1
Artemis	0	1	1	?
Apollo	0	1	1	?
Leto	0	0	0	?
Ares	1	1	?	1
Athena	1	1	?	1
Hephaestus	1	1	?	1
Aphrodite	1	1	?	1
Cyclope	1	1	?	1
Persephone	1	1	?	1
Hermes	1	0	?	0
Hebe	1	0	?	0
Dionysus	1	0	?	0

Table 2.2

	L	A	Y
Rheia	0	0	0
Kronos	0	0	1
Demeter	0	0	0
Hades	0	0	0
Hestia	0	1	0
Poseidon	0	1	0
Hera	0	1	0
Zeus	0	1	1
Artemis	1	0	1
Apollo	1	0	1
Leto	1	0	0
Ares	1	1	1
Athena	1	1	1
Hephaestus	1	1	1
Aphrodite	1	1	1
Cyclope	1	1	1
Persephone	1	1	1
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0

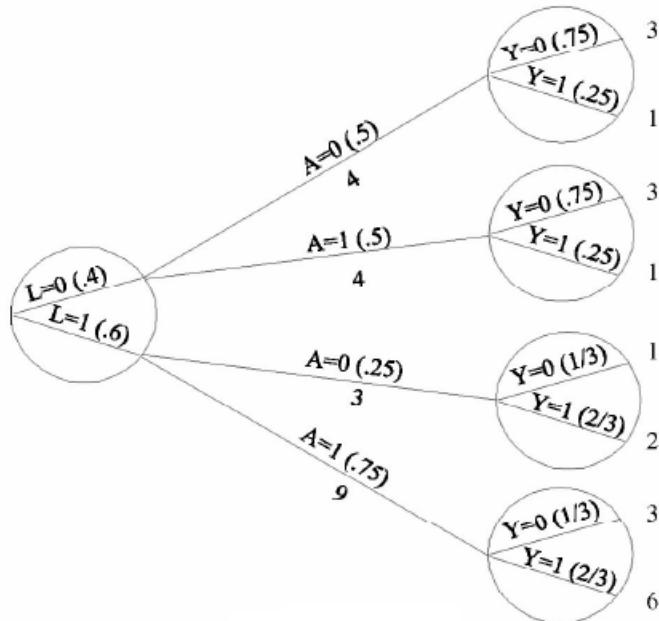


Figure 2.1

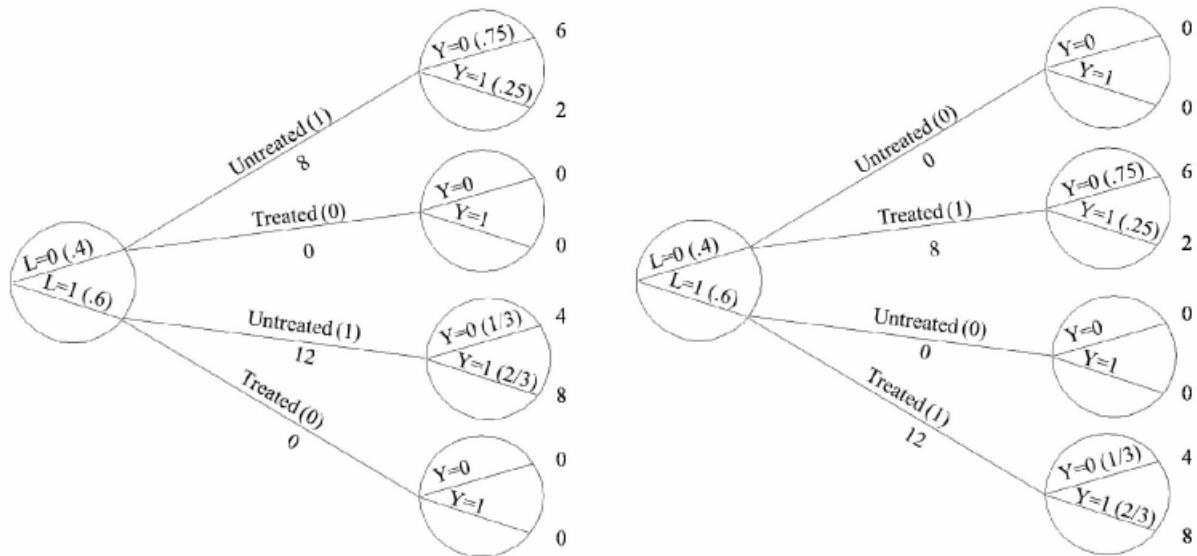


Figure 2.2

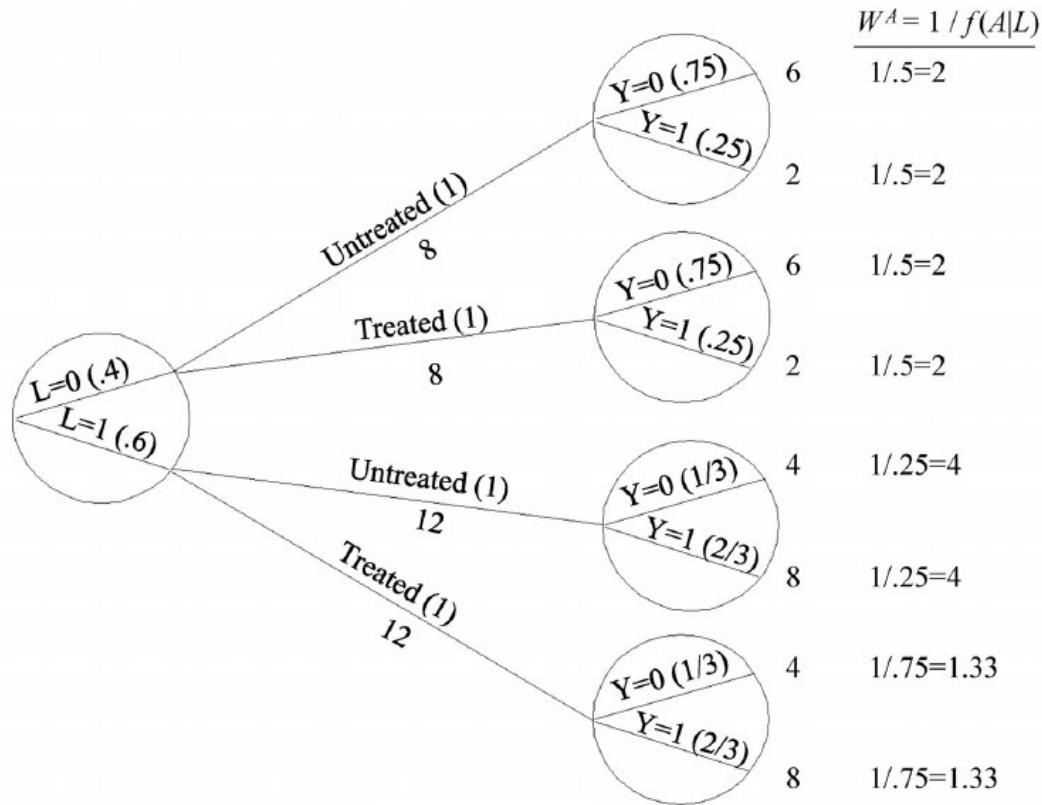


Figure 2.3

第三章 观察性研究

让我们再思考一下上一章开篇提出的问题：“一个人抬头向天上看会影响其他行人也抬头向天上看吗？”上一章我们讨论了随机试验，但你觉得要你抬头看那么多次实在是太累了，并且对你的颈椎也不好。于是你决定采取一种新的研究方法：先找到一个站着但没有抬头向上看的行人，再找一个向第一个人走去但也没有抬头向上看的人，然后观察这两人在接下来 10 秒内的行为。如此这般重复上千次。然后你比较第一个人抬头看之后第二个人也抬头看的人数比例，与第一个人未抬头看但第二个人却也抬头看的人数比例。在这个研究中，研究者观察并记录相关的数据，这样的研究被称作观察性研究。

不过批评者们会说，这两个行人都抬头向上看并不是因为第一个人影响了第二个人，而可能是因为他俩都同时听到了天空中的惊雷声，因此这一研究不足以证明第一个人抬头看会影响第二个人也抬头看。这一批评不适用于随机试验，这也是为什么随机试验对于因果推断理论非常重要的原因。然而，在实践中，用随机试验来估计因果效应经常受到一定的限制。许多科学研究都不是试验。许多人类知识都来源于观察性的研究，比如进化论、板块构造论、全球变暖、甚至还有天文学。想一想我们人类是怎么知道热水能导致烫伤的。这一章将讨论在何种条件下，我们能通过观察性研究得到合理的因果推断。

3.1 可识别性

一个理想的随机试验能够用来识别并量化因果效应均值，这是因为对治疗组的随机分配保证了互换性。比如，在前两章心脏移植和死亡率的研究中，治疗组中的被试如果没有接受移植手术，他们的死亡率应该和实际非治疗组中的死亡率一样。因此，从随机试验中得到的相关性风险比 0.7，也就等于因果性风险比。

(为了方便，在本章中，随机试验都无失访、被试都完全配合研究计划。第八章和第九章将讨论其他情形)

与之相比，观察性研究就没有那么有说服力，主要原因是因为治疗组的分配不是随机的。比如，如果有严重心脏病的病人更可能接受心脏移植手术，那么实际接受治疗的人在未接受治疗的情况下，会比实际未接受治疗的人有更高的死亡风险。因此，观察性研究中得到的相关性风险比，就会是移植的实际有益效果和重症的负面影响相互抵消之后得到的。因而，在观察性研究中，我们观测到的治疗和结局间的相关性，并不一定代表治疗对结局的因果效应。

虽然随机试验对因果推断有根本优势，但我们大多数时候只能用观察性研究来回答因果性的问题。那我们应该怎么做？简单说，我们将观察性研究中的治疗分组视作根据已有的协变量随机

进行的, 然后在此基础上分析数据——虽然我们知道这顶多是一个近似。因此, 观察性研究中因果推断的核心, 是我们能将观察性研究视作条件随机试验。

如果以下条件成立, 则我们可以将观察性研究视为条件随机试验:

1. 我们要比较的治疗方式, 是良定的, 并且能与我们数据中的治疗形式相对应。
2. 治疗取值的条件概率, 虽然不由研究者决定, 但是只由已有的协变量 L 决定。
3. 在控制了 L 的基础上, 任一治疗取值的概率都应该大于 0。

本章, 我们将在观察性研究的背景下讨论这三个条件。第一个条件就是第一章所说的一致性, 第二个条件就是第二章所说的互换性, 第三个条件则是知识点 2.3 中所说的正数性。

我们将会看到这三个条件在大多数情况下很难达成, 这也解释了为什么人们对观测性研究中的因果推断持怀疑态度。然而, 如果观察性研究能和条件随机试验能够类比, 那我们就可以用上一章介绍的方法——逆概率加权和标准化——来计算观察性研究中的因果效应。在此, 我们将这三个条件称作可识别性条件或可识别性假设。比如, 上一章我们用表 2.2 中的数据算出在该条件随机试验中, 因果性风险比是 1。使用从观察性研究中得到的同样数据(现在列于表 3.1 中), 并假设三个可识别性条件成立, 我们依然可以计算出因果性风险比等于 1。

更重要的是, 在一个理想的随机试验中, 试验设计保证了可识别性的三个条件成立。也就是说, 对于条件随机试验, 我们只需要表 3.1 的数据, 而不需要任何条件或假设, 就能计算因果性风险比。与之相比, 要在观察性研究中识别因果效应, 我们需要假设这三个可识别性条件成立, 但显然, 这三个条件不可能时时成立。因而, 观察性研究中的因果推断需要两个重要元素: 数据和可识别性条件。精讲点 3.1 给出了可识别性的详细定义。

(Rubin (1974, 1978) 将 Neyman 关于随机试验的理论扩展到观察性研究。Rosenbaum 和 Rubin (1983) 又将互换性和正数性的结合称为弱可忽略性 (*ignorability*), 将完全互换性和正数性的结合成为强可忽略性)

如果可识别性条件有一条不成立, 那就不能将观察性研究比作条件随机试验。在这种情况下, 也还有其他可能的方法分析观察性研究的数据做出因果推断, 不过这些方法可能需要其他不同的可识别性假设。其中一个方法是用治疗的预测因素——也称作工具变量——在数据分析中代替治疗, 并且假设这个预测因素是根据协变量随机分配的。我们将在第十六章讨论工具变量。

27 将观察性研究视作条件随机变量, 在类比通常被认为是合理的学科(比如流行病学)中, 是常见且通用的。但在其他一些学科(比如经济学)中, 这种类比被认为并不合理, 而更多使用工具变量方法。在第十六章之前, 我们将主要关注能视作条件随机试验的观察性研究, 及其所用的因果推断方法。接下来, 我们将详细介绍可识别性的三个条件。

3.2 互换性

我们已经讨论了许多关于互换性 $Y^a \perp\!\!\!\perp A$ 的内容。在边缘（即无条件）随机试验中，治疗组和非治疗组是可互换的，因为如果实际治疗组没有接受治疗，被试的结局总体而言会和实际非治疗组的一样，反之同理。这是因为随机分组保证了结局的独立预测因素在治疗组和非治疗组之间的分布是相同的。

（不管治疗取值如何，一个变量如果都和结局 Y 相关，则我们称这个变量为结局 Y 的独立预测因素。对于一个二分结局，独立预测因素也经常被称作二分因素）

我们以表 3.1 的数据为例。上一章，我们已经知道这个研究中互换性不成立，这是因为 69%
28 的治疗组被试和 43% 的非治疗组被试是重症 ($L = 1$)。这种独立预测因素分布的不均衡不应该出现在边缘随机试验当中（其实，有时候某些不均衡还是可能会因随机误差出现，但让我们暂时忽略这种误差，让我们仅在假想中进行试验）。

另一方面，如果条件随机试验的治疗组分配是由 L 决定的，那结局的独立预测因素 L 的分布在治疗组和非治疗组就会不均衡，但这种不均衡是由试验设计的本质决定的。表 3.1 的数据可以视作来自一项条件随机试验，其中治疗组和非治疗组不可互换，因为治疗组在试验刚开始的时候就有较差的预后因素。但在每一个 L 的可能取值中，治疗组和非治疗又是可以互换的。因而，我们可以说，在条件随机试验中，有界互换性 $Y^a \perp\!\!\!\perp A|L$ 成立，因为在每一个 L 的可能取值中，其他结局预测因素的分布在治疗组和非治疗组中相等。

回到我们的观察性研究。如果治疗方式不是由研究人员随机分配的，那接受什么样的治疗方式就很可能和其他结局预测因素相关。也就是说，如果条件随机试验一样，在观察性研究中，结局预测因素的分布在治疗组和非治疗组中不等。比如，如果表 3.1 的数据是从一项观察性研究中得到的，那在研究中，医生更可能将稀缺的心脏资源给更需要它的病人，也就是重症 ($L = 1$) 的病人。实际上，如同表 3.1 的数据，如果只有一个结局预测因素的分布在治疗组和非治疗中不同，那我们可以把这样的研究视作：（1）一项观察性研究，其中重症 ($L = 1$) 病人有 75% 概率接受治疗，轻症 ($L = 0$) 病人有 50% 接受治疗；或者（2）一项（非盲）条件随机试验，其中重症 ($L = 1$) 被试有 75% 概率被分配到治疗组，轻症 ($L = 0$) 被试有 50% 被分配到非治疗组。这两种描述在逻辑上是等价的。在任一一种描述中，有界互换性 $Y^a \perp\!\!\!\perp A|L$ 都会成立，我们也就可以用标准化或逆概率加权方法来识别因果效应。

当然，在观察性研究中，一个重要问题是是否只有 L 的分布在接受治疗和未接受治疗的病人中不等。遗憾的是，我们无法回答这个问题。比如，我们观察性研究的一名研究者就强烈认为只

有 L 是不等的, 他的理由如下: “心脏移植手术只会提供给不太可能拒绝这项手术的病人, 也就是说, 有 HLA 基因的心脏只会提供给有亲和基因的病人; 因为 HLA 基因不是死亡风险的预测因素, 因此可以知道在每一种 L 取值下, 是否进行手术在本质上是随机的。”因此, 这位研究中就在有界互换性 $Y^a \perp\!\!\!\perp A|L$ 成立的前提下分析数据。

然而, 在观察性研究中, 我们也只能说在这个“假设”成立的前提下。不管我们的理由多么有说服力, 就像上一段的研究者一样, 在没有随机分配的情况下, 我们都不能保证互换性成立。比如, 可能研究者不知道的是, 医生们更喜欢给不吸烟的病人进行移植手术。如果两个都有相似 HLA 基因且 $L=1$ 的病人, 一个吸烟 ($U=1$), 一个不吸烟 ($U=0$), 那吸烟的人接受手术的概率就会更低。吸烟也是一个重要的结局预测因素。当吸烟的分布在接受治疗的病人 (吸烟的人更少) 和未接受手术的病人 (吸烟的人更多) 中不一样时, 给定 L 时的有界互换性就不再成立。就算我们也收集了吸烟的数据, 还有其他我们未知的结局预测因素可能存在分布不同的情况。

(我们用 U 来表示未知变量。因为未知变量不能用于标准化和逆概率加权方法, 所以只有 L 的数据时, 有界互换性不成立, 我们也就不能识别因果效应)

因此有界互换性 $Y^a \perp\!\!\!\perp A|L$ 在观察性研究中不一定成立。具体而言, 因为我们不知道是否还有除了 L 的未知结局预测因素会影响治疗 A 的分配, 所以有界互换性 $Y^a \perp\!\!\!\perp A|L$ 不一定成立。更可惜的是, 就算有界互换性 $Y^a \perp\!\!\!\perp A|L$ 成立, 我们也不能用我们的数据证明它成立。如果我们没有收集吸烟的数据, 我们怎么能知道吸烟的分布接受了治疗和未接受治疗的病人中不一样? 是否还有其他未知的结局预测因素在这两种中也不一样? 这些问题我们都不能回答。当我们在条件随机试验的假设下分析观察性研究的数据时, 我们只能希望我们的专业知识让我们收集了足够多的数据, 使得这个假设虽不完全正确, 但是接近正确。

(要验证有界互换性, 我们需要证明 $\Pr[Y^a = 1 | A = a, L = l] = \Pr[Y^a = 1 | A \neq a, L = l]$, 但这基本上是不可能的, 因为在观察性研究中我们不可能知道未接受治疗 a ($A \neq a$) 的病人的 Y^a , 因此我们不可能用我们的数据计算这个等式的右侧)

30 研究人员可以用他们的专业知识来提高他们研究中有界互换性假设的可信度。他们可以收集许多相关的变量 L (比如, 既是结局独立预测因素又是治疗决定因素的变量) 的数据, 而不是仅仅如表 3.1 中一样只有一个变量。然后就能假设只要控制了这些变量, 有界互换性就能近似成立。遗憾的是, 不管 L 中有多少个变量, 我们都不能检验这个假设是否成立, 也就使得观察性研究的因果推断不是一件万无一失的事。其因果推断的有效性依赖于研究人员的专业知识。这些专业知识, 表现在通过控制已知的变量从而确保互换性, 将和数据一起共同确认观察性研究中的因果效应。

3.3 正数性

一些研究人员想开展一项试验来计算心脏移植 A 对 5 年死亡率 Y 的平均效应。在这里，不言自明的是，研究人员会将一些被试分配到治疗组 $A=1$ ，而将另一些被试分配到非治疗组 $A=0$ 。没有正经的研究人员会将所有被试都被分配到同一组， $A=1$ 或者 $A=0$ 。如果所有的被试都接受同一种治疗，那我们就不可能计算因果效应。因而，基本可以肯定的是，我们每一个治疗组中，都会有一定数量的被试。换句话说，我们需要确保每一种治疗的概率大于 0。这一性质被称为正数性。

在试验中我们不需要强调太多正数性，因为试验设计肯定能保证正数性成立。在边缘随机试验中， $\Pr[A=1]$ 和 $\Pr[A=0]$ 肯定都为正。在条件随机试验中， $\Pr[A=1|L=l]$ 和 $\Pr[A=0|L=l]$ 在每一种 L 的可能取值下也肯定为正。比如，在我们表 3.1 的条件随机试验数据当中，重症下被分配到治疗组的概率是 $\Pr[A=1|L=1]=0.75$ ，而轻症下则是 $\Pr[A=1|L=0]=0.50$ ，因而，在控制了 L 后，正数性成立。因此，对于所有的 a 有 $\Pr[A=a|L=l]>0$ 时，正数性成立。不过这里正数性的定义并不完整，比如，如果我们的研究人群被限制在 $L=1$ 组中，则没必要考虑 $L=0$ 时的正数性。我们只需要考虑人群中存在的取值 l 下的正数性即可。

(正数性有时也被称作试验-治疗假设)

(正数性定义: 如果 $\Pr[L=l]\neq 0$, $\Pr[A=a|L=l]>0$)

此外，我们仅在对互换性有影响的 L 中考虑正数性。比如，在表 3.1 的条件随机试验中，我们不会问长发被试被分配到治疗组的概率是否大于 0，因为这个试验中的互换性不需要“长发”这个变量（“长发”这个变量不是在控制了 L 和 A 后结局的独立预测因素，更不会影响治疗组的分配）。在我们使用标准化或者逆概率加权计算因果效应时，只需要调整 L 即可，而不需要调整的变量，我们就不必考虑其下的正数性。

31 在观察性研究中，正数性和互换性都不能得到保证。比如，如果医生只治重症 ($L=1$) 病人，也即 $\Pr[A=1|L=0]=0$ ，那正数性就不成立，就如图 3.1 所示。正数性和互换性的区别在于，我们可以用我们的数据来验证正数性（详见第十二章）。比如，在表 3.1 中，如果我们将其视作一项观察性研究，我们就能说 L 下的正数性成立，因为对每一个 L 的可能取值（即 $L=0$ 和 $L=1$ ），都有被试被分配到治疗组和非治疗组（即 $A=0$ 和 $A=1$ ）。我们上一章的讨论只说了

互换性, 但也暗含了正数性假设 (在知识点 2.3 中有说明)。当正数性成立时, 我们之前对标准化和逆概率加权得到的风险的讨论就能成立, 但当正数性不成立时, 这两种方法就不再有意义。为了理解这一点, 可以图 3.1 所表示的情形。如果在 $L=1$ 时没有 $A=0$ 的被试, 那我们就没有信息来模拟实际治疗组未接受治疗时的反事实结局。详见知识点 3.1。

3.4 一致性: 首先, 定义反事实结局

一致性意味着我们观测到的治疗组每个被试的结局, 等于这个被试如果接受了治疗时的反事实结局; 我们观测到的非治疗组每个被试的结局, 等于这个被试如果未接受治疗时的反事实结局。即对于 $A=a$, 有 $Y^a=Y$ 。这个表述似乎是显然的, 一些读者可能会不禁发问: 在什么情况下一致性才会不成立? 毕竟, 如果我服用了阿司匹林 ($A=1$), 然后我死了 ($Y=1$), 不是就说明我的反事实结局 $Y^{a=1}$ 就是 1? 因而许多人认为一致性非常简单。但这种表明上的“简单”具有很强的欺骗性。让我们将一致性分成两个部分: (1) 通过右上角的小 a 定义的反事实结局 Y^a , 和 (2) 反事实结局与我们观测到的结局的联系。这一小节我们将讨论第一个部分。

(Robins 和 Greenland (2000) 认为良定的反事实情境, 或其他数学上等价的概念是因果推断的必要条件)

32 让我们回到心脏移植 A 对 5 年死亡率 Y 的因果效应的随机试验。在招募被试前, 研究人员在研究方案中详细描述了两种治疗方法: 心脏移植 ($A=1$) 和药物治疗 ($A=0$)。比如, 研究人员详细描述了心脏移植组 ($A=1$) 被试的术前准备、麻醉、手术方式、术后护理以及免疫抑制方法。如果这个方案没有详细描述这些过程, 那么每个医生可能会采用不一样的方式进行移植手术或术后的免疫抑制。

(精讲点 1.2 介绍了治疗的不同形式这一概念)

如果同一治疗的不同形式有不一样的因果效应, 那问题就会随之产生。比如, 在我们心脏移植的试验中, 如果手术是由传统方式进行, 那其因果效应势必会和采用新式手术方法有所不同。因此, 当我们说“心脏移植对死亡率的因果效应”时, 我们需要指明治疗 A 的具体形式 a 。如果治疗方式的某一取值 a 不是良定的, 那我们的反事实结局 Y^a 也不是良定的, 于是我们的因果效应

33 $\Pr[Y^{a=1}=1]-\Pr[Y^{a=0}=1]$ 也就不是良定的。理想情况下, 随机试验的研究方案应该清晰定义治疗的每一种可能取值 a , 因而反事实结局 Y^a 是良定的。在观察性研究中, 研究者只能尽可能精确地去定义研究中的 a 。虽然这对医疗干预措施, 比如心脏移植而言相对直接明了, 但是对于现实世界中的其他干预措施而言则不是那么直截了当。

假设一个研究者想研究 40 岁时的肥胖 A 对 50 岁时的死亡率 Y 的因果效应（在这里，肥胖简单定义为 $BMI \geq 30$ ）。首先，他将因果效应定义为研究人群 40 岁时，如果所有人都肥胖下的风险 $\Pr[Y^{a=1} = 1]$ ，和如果所有人都不肥胖下的风险 $\Pr[Y^{a=0} = 1]$ ，两者间的对比。但是，“所有人都肥胖下的风险”具体是什么意思？虽然都是 40 岁时肥胖，但是这背后却隐藏了许许多多的不同类型。有的人在 40 岁之前就已经胖了 20 年，但有的人只胖了 2 年。因此，在这项研究中，我们的肥胖 $A=1$ 有许多种不同的形式，可以由持续时间、肥胖程度等不同的特征决定。因为每一种形式对死亡率都有不同的影响，所以这个研究者需要给出他所想的肥胖的清晰定义。否则，“40 岁时的肥胖对 50 岁时的死亡率的因果效应”就是劣定¹的。

不过，就算这位研究者能在持续时间、程度等各方面清晰定义什么是肥胖 ($A=1$)，这项研究的其他方面也需要精确说明。比如，这位研究者需要说明他需要采取什么干预²措施从而让被试在 40 岁时的肥胖状态一直保持不变。他可以用转基因技术提高被试腰间和大动脉处脂肪，从而让被试一直保持肥胖，或者对被试进行胃切除手术，从而让被试体重不会增长。于是问题随之而来，这些不同的干预措施就算能让肥胖水平保持稳定，但对死亡率都有不同的影响。

（本书的第三部分会重点介绍持续一段时间的干预措施，比如对肥胖的干预。本章我们将忽略干预所持续的时间）

以宙斯为例。假设他在 40 岁时身材正常 ($A=0$)。但他可能在 49 岁时去世 ($Y=1$)，这是因为为了保持体重正常，他常常攀岩（死于攀岩事故），或者做了胃切除手术（死于过量麻醉）。但他可能在 49 岁时没有去世 ($Y=0$)，因为为了保持体重正常，他的饮食非常健康，或者他的基因就决定了他是不容易长胖的体质。于是，宙斯在不肥胖 $a=0$ 时的反事实结局 $Y^{a=0}$ 是什么？我们会说他在一系列导致身材正常 $A=0$ 的情形中死去，但在另一系列导致身材正常 $A=0$ 的情形中他却没有死。在这里， $a=0$ 时的反事实结局 $Y^{a=0}$ 就是劣定的。同理，宙斯肥胖下的反事实结局 $Y^{a=1}$ 也是劣定的。

劣定的反事实结局会让我们的因果性问题变得空洞。如果真的想研究肥胖 $A=1$ 对死亡率的影响，这位研究者还需要花一些工夫，清晰定义一下反事实结局 $Y^{a=0}$ 和 $Y^{a=1}$ 。就像如果我们想研究运动的影响，我们就需要认真定义运动时长、强度、形式等等等。

¹ 英文 ill-defined，本书翻译为劣定的，表示定义模糊不明。

² 英文原文为 treatment。在英文里，试验中对不同组采取的不同做法，都能叫 treatment。原书主要使用 treatment 一词，直译为“治疗”。但在中文中，某些地方译作“治疗”不合语境。因此，译者视情况有时将其翻译为“干预”。

(某些研究很有趣, 它们研究了肥胖对找工作时的歧视的影响。研究者将肥胖的影响定义为雇主看到申请者的简历和照片后叫申请者来面试的比例。这一定义就比我们书中的例子少了许多模糊, 这是因为“雇主对肥胖的印象”这一定义不必考虑什么因素造成了肥胖)

但我们也也要注意到, 有时对治疗方式的过度清晰定义是不必要的。比如, 我们一致同意在操场跑步有益我们的健康, 但是顺时针跑还是逆时针跑, 这就不重要了。当某些特征和反事实结局不相关时, 我们就不必把它们包含进我们的定义中了。也就是说, 我们只需要我们对于治疗方式 a 的定义能消除歧义、足够明确即可。

但如果有人问, “你怎么知道我们的定义是足够良定的呢?”很遗憾, 答案是, 我们不知道。判断一个定义是否良定, 需要研究者用自己的专业知识进行判断。这一判断应在研究人员中取得共识。我们现在认为跑步的方向不会对健康造成影响, 但未来的研究可能会认为我们错了, 比如未来发现, 人体跑步向左倾斜时对健康不利。在历史中的每一个时间点, 撰写研究方案的研究者只能用他当时能有的所有知识来尽可能消除模糊保证精确。然而, 在所有关乎因果的问题中, 模糊总是存在的。一个因果性问题的模糊程度能够用更详尽的叙述减少, 但是不能完全消除。在观察性研究中, 涉及到生物(比如体重或者胆固醇)或者社会(比如社会经济地位)因素的问题时, 模糊程度都会格外地高。

以上讨论阐述了因果推断的一个基本特点: 因果性问题的清晰程度取决于我们的专业知识和我们所拥有的信息。我们现今觉得有意义的因果性问题, 可能在未来发现了能影响结局的更具体因素之后, 就变得模糊而无意义。我们再一次强调, 良定的治疗方式, 其定义取决于大多数研究者的共识, 而这会随时间改变。精讲点 3.3 提供了另一种让因果性问题更精确的做法, 虽然和我们在正文中的讨论在逻辑上等价。

现在, 读者可能已经意识到, 在清晰定义一个治疗方式的过程中, 我们可能会改变最开始的问题。我们最开始只是在讨论一项关于肥胖效果的研究, 但最终我们居然讨论起运动的方式。我们越是想对我们的研究提供一个良定的因果性阐释, 我们离最开始的问题就越远。但这不是一件坏事: 不断地完善我们的问题, 直到它在现有知识下不再模糊不明, 是因果推断的基本要素之一。我们将治疗方式定义得越明确, 我们就越能避免将来研究者互相交流时的歧义和误解, 尤其是当不同的研究得到的数值估计不一样的时候。

我们已经讨论了什么是足够良定的治疗方式, 下一步我们需要解决的问题是对我们研究中的数值估计做出因果性阐释, 这需要一致性的第二个部分。

3.5 一致性: 其次, 将反事实世界和观测到的数据相结合

上一章的研究者在了解了明确治疗方式定义的重要性后, 他决定将自己问题中的干预方式修改成如下所示: 从 18 岁到 40 岁, 对被试进行强制且严格的饮食控制, 从而保证他们的体重不会超过 18 岁的体重 ($a=1$)。具体而言, 每一个被试在 18 岁后每天都要称一次体重。当体重超过 18 岁的基线时, 就会被限制热量摄入, 但不改变热量来源的饮食结构, 直到这个被试恢复到 18 岁时的基线体重。因而, 就算存在一两千克的误差, 基本上没有被试会在 40 的时候超过 18 岁的基线体重。不过这项研究不会对运动进行任何限制。对照组 $a=0$ 则是不进行任何干预。

(这段所述额假想干预方式由 Robins 在 2008 年描绘。这一干预方式仅限于男性, 从而排除了孕期增重的一系列复杂问题)

假设大多数研究者都认为这项新研究的 $a=1$ 和 $a=0$ 是良定的, 因而 $Y^{a=1}$ 和 $Y^{a=0}$ 也不存在模糊不明的情况。此时, 我们就可以将注意力转移到等式 $Y^a = Y$ 和 $A = a$ 当中。

让我们用战神阿瑞斯 (Ares, 宙斯与赫拉的儿子) 来举例。假设阿瑞斯就算没有接受干预, 他 18 岁至 40 岁时的体重都保持得很好, 没有超过 18 岁时的基线体重。这可能是因为他基因好 (来自赫拉), 或者是因为他的运动强度大 (经常上战场)。因而, 我们观测到的干预措施不是 $A=1$, 他的观测结局 Y 也就不是 $a=1$ 时的 $Y^{a=1}$ 。

为了将观测到的结局 Y 和 $Y^{a=1}$ 联系起来, 我们需要保证在分析中, 只有接受了干预 $A=1$ 的被试, 才是接受了干预 $a=1$ 的被试。对未接受干预 $a=0$ 的被试同理。这是因为, 如果我们想用观测到的据去量化因果效应 $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$, 我们就需要我们的被试中, 干预的取值是和 $a=1$ 以及 $a=0$ 是一致, 也就是说, 我们需要 (无条件) 正数性。如果我们不能将反事实世界联系到观测的数据, 只是清晰地定义 a , 并无助于帮我们做出正确的因果推断。

不过, 如果我们的数据不够多, 那完美符合我们干预 a 的定义的被试就可能存在, 这在现实中时常发生。比如, 在我们的肥胖研究中, 我们很可能就只能收集在 40 岁时的体重, 而没有体重史、日常运动量、饮食等信息。

(知识点 3.2 讨论了如果治疗的形式未知的时候, 因果推断的模糊性)

解决这个问题的一个方法是假设这一干预 (或治疗) 的所有不同形式的影响一样, 也就是治疗差异无关紧要假设。某些情况, 这种近似能够接受。比如, 我们想评估高血压和正常血压对中风的不同因果效应, 而研究数据则显示高血压通过药物降低后能和正常血压产生相似的结局。因而, 我们可能会说, 一个精确的“血压”定义对于联系反事实结局和观测结局是不必要的。在另一些情况中, 这个假设则有许多问题。比如, 我们想评估保持体重对死亡风险的影响, 而研究数据则显示某些干预措施可能会提高死亡风险 (比如一致吸烟), 但另一些则会降低 (比如适度运

动)。在实践中,许多观察性研究的分析在做因果推断的时候,尤其涉及同一治疗的多种形式的时候,都暗含了治疗差异无关紧要假设。

(精讲点 1.2 中定义了治疗差异无关紧要假设。对同一治疗 $R = r$ 的任意两种不同形式 $A(r)$ 和 $A'(r)$, 记 $Y_i^{r,a(r)}$ 为个体 i 在 $R = r$ 时 $A(r) = a(r)$ 的反事实结局, 如果有 $Y_i^{r,a(r)} = Y_i^{r,a'(r)}$, 则假设成立)

总而言之,劣定的肥胖会使我们对因果效应估计值的阐释模糊不清(上一小节),但是足够良定的肥胖在我们的现实数据中又不存在的话,也会使我们的因果推断不清不楚(这一小节)。我们需要详细描述在一个人群中存在的实际治疗的各种特征,将其与我们想研究的“治疗”匹配起来。这项任务可能会在试验研究中相对简单直接,但在一些观察性研究中,尤其是涉及生物和社会因素的观察性研究中,这项任务可能会非常困难,甚至不可能。

当然,如果研究人员们一致认为某一治疗的不同形式对结局的影响差不多,那就没有必要在我们的数据中详尽描述所有不同形式的特征。然而专家们也可能出错。我们能做的,是将我们的假设说清楚,然后讨论可能的种种情况,这样就能让其他人来进一步验证我们的结果。下一节,我们将讨论如何实现这种透明化。

(3.4 和 3.5 小节的进一步详细讨论可参阅 Hernan (2016) 和 Robins 与 Weissman (2016) 所著论文)

37 3.6 靶标试验

在整本书中,因果效应一词指两种不同治疗取值下的反事实结局的比较。因此,对每一种因果效应,我们都能用一个假想的随机试验去量化它。这个假想的随机试验,我们称之为靶标试验(target trial)。然而真正开展一项靶标试验是不现实的,我们只能用观察性研究中得到的数据进行因果推断。换句话说,我们将观察性研究中的因果推断过程,视作在模拟一项靶标试验。如果这种模拟是成功的,那我们从观察性研究中得到的数值估计,就和从靶标试验中(如果我们能实际进行的话)得到的没有差别。我们在 3.1 小节中已经讨论过,如果我们能将观察性研究类比作条件随机试验,则我们就可以用上一章介绍的方法,即标准化和逆概率加权,来计算观察性研究中的因果效应。(精讲点 3.4 介绍了如何在观察性研究中计算病例中可归因于治疗的比例。)

(靶标试验,或者其他同义词,是因果推断框架的核心之一。Dorn (1953), Cochran (1972), Rubin (1974), Feinstein (1971) 和 Dawid (1986) 等人都使用过这个概念。Robins (1986) 将这一概念应用于时异性治疗中)

因此, “你想模拟什么样的随机试验”, 是观察性研究中因果推断的核心问题之一。对每一种我们希望在观察性研究中计算的因果效应, 我们需要描述: (1) 靶标试验是什么; (2) 怎么用观察性数据模拟靶标试验。

我们可以通过详细说明靶标试验(假想的)研究计划中的重点要素, 来具体化这项靶标试验。这些因素包括: 入组标准、干预(或治疗)策略、结局定义、随访方式、效果比较以及数据分析。让我们先把关注点放在我们要比较的治疗策略(或者也称干预策略)上。前两小节已经讨论过, 我们需要先明确我们要研究的治疗是什么, 然后再将其联系到我们的实际数据中。

(*Hernan 和 Robins (2016)* 详细讨论了我们需要说明靶标试验的哪些因素, 以及用观察性研究模仿靶标试验的步骤)

一名研究者想研究 40 岁肥胖且不吸烟的人群中, 减肥对死亡率的影响。第一步是尽可能将研究问题定义清楚。比如, 他将研究目标定成每年 BMI 下降 5% 左右对死亡率的影响, 而不考虑究竟是用了什么方式使 BMI 下降。然后, 他再将这个定义转换成一项靶标试验方案中的干预措施, 从而用自己已有的数据来模拟这项靶标试验。

尽可能模拟一项靶标试验, 使得研究者们不能仅仅只做一些非常简单的分析, 比如简单比较 40 岁时肥胖和非肥胖人群的死亡风险。因为这样的简单分析, 其对应的靶标试验仅仅关注被试在 40 岁基线时的瞬间体重。也就是说, 比较反事实世界中, 被试瞬间肥胖后的反事实结局, 以及被试瞬间不肥胖后的反事实结局。这在现实中是不可能的, 因为没有人会瞬间肥胖或者不肥胖。因此我们不能将反事实结局和观测结局联系起来。

(本书作者和他们的研究者曾用观察性数据研究了减肥的影响(见 Daniel 等人 2016 年所著论文)。在这项研究中, 作者们假设减肥的不同方式对结局影响不大, 然后尽可能地清晰定义了干预策略的时间)

38 将观察性研究中的因果推断类比成一项靶标试验这一概念并不被所有人认可。一些研究者认为“ A 对 Y 的因果效应”这个定义, 不管 A 和 Y 是什么(只要 A 在 Y 之前), 就已经是良定了。因此, 在我们肥胖研究的例子中, 这些研究者就认为我们不必去描绘一项不存在的靶标试验。我们认为精确描绘一项靶标试验是我们阐释效应估计数值的必要条件, 但这些研究者则认为没必要去量化一个因果效应。他们认为:

我们不必去知道一项观察性研究中某个因果效应的具体估计值, 我们只需要知道这个因果效应是否存在就足够了。肥胖和死亡率之间的强相关性表明某些对肥胖的干预措施能降低

死亡风险。因此, 我们能得到的有价值信息是, 如果我们对所有肥胖者进行强制干预, 把他们的体重控制在正常范围, 那我们就能预防许多不必要的死亡。

(这一观点的详细例子, 参阅 Pearl (2009), Schwartz (2016) 以及 Clymour 和 Spiegelman (2016) 等人所著论文)

这一观点非常具有吸引力, 但也很危险。接受这一观点会带来一个问题: 劣定的治疗形式会阻碍观察性研究中对互换性和正数性的思考。

39 我们先说互换性。为了正确模拟靶标试验, 研究者需要模拟随机分组的过程, 因为随机分组能保证互换性, 即使是在协变量 L 下的有界互换性。如果我们放弃说明我们研究问题中肥胖干预的各种不同形式, 我们又怎么会去收集影响肥胖的各种变量信息, 从而使得我们的干预组和非干预组拥有有界互换性? 同理, 我们又怎么会去收集影响结局 (死亡率) 的各种风险因素? 这些都会带来疑问: 有界互换性是否成立。

不明确的干预形式进行也会影响正数性。假如我们在计算肥胖对死亡率的影响时调整了包括饮食和运动在内的协变量 L , 可能会出现的是, 在某些变量取值的组合下, 没有一个人是肥胖的, 也就是说, 正数性不成立。我们也可以在协变量 L 的某些特定取值下寻求正数性, 但此时我们的研究人群不再具有代表性。

正数性不成立会导致另一个问题: 对应的靶标试验中的干预措施是不合理的。只是简单比较观察性研究中的肥胖人员和非肥胖人员, 其对应的靶标试验中的干预措施是“让所有人瞬间变得肥胖”, 这在现实中不可能, 也掩盖了肥胖的复杂性, 似乎也让“减肥”变得无所谓了。如同我们之前所说, 一个更合理的, 虽然不是完美的干预措施定义是“每年 BMI 下降 5%”。因而, 将观察性研究中的因果推断与靶标试验相对应, 不仅能帮助我们完善我们的研究问题和研究目标, 还能让我们的因果推断与现实决策更加息息相关。

干预措施不明确导致的问题不能用统计方法解决。本书中介绍的所有因果推断分析方法仅适用于良定的干预 (或治疗) 措施。尽管我们可以用其他不能验证的假设替代互换性 (详见第十六章), 或者放弃正数性从而接受无法验证的模型外推 (详见第十四章), 但是劣定的干预 (或治疗) 定义只能损耗我们因果效应估计的有效性, 而没有其他方法进行弥补。

当一项观察性研究的数据不能用来面膜泥靶标试验的时候, 是不是也就意味着一切都白费了? 不完全是这样。这种情况下, 观测所得的数据, 依然能用来做非因果性的预测。我们知道肥胖的人群有更高的死亡率, 因此肥胖是死亡率的一个预测因素, 换句话说, 肥胖和死亡相关。这将有助于我们辨认什么样的人有较高的死亡率。不过就算我们知道肥胖和死亡相关, 我们却并不

知道肥胖对死亡有没有因果性的影响: 肥胖和死亡的关系, 可能就和携带打火机与肺癌的关系一样。因此, 肥胖和死亡率之间的相关性会进一步催生出种种假设与研究去探寻背后的机制, 但其

40 证据并不足以推荐所有人都去减肥。

(因果推断可以实实在在反事实世界进行的预测。更多关于预测与因果推断区别的讨论, 请参见 Hernan, Hsu 和 Healy 在 2019 年所著论文)

当我们从因果性问题回退到预测性问题时, 我们就回避了许多随机试验不能回答的问题。但另一方面, 因果推断是我们的终极目标, 仅仅预测则不能满足我们的希望。

第三章精讲点和知识点

精讲点 3.1: 因果效应的可识别性 (原书第 27 页)

因果效应的 (非参数化) 识别需要一系列假设。如果这些假设能让观测数据的分布仅和唯一一个效应量度值相容, 则我们认为因果效应可以 (非参数化) 识别。反之, 如果这些假设能让观测数据的分布和多个效应量度值相容, 则我们认为因果效应不可识别。比如, 如果表 3.1 的数据来自一项随机分组概率取决于 L 的条件随机试验 (也即有界互换性 $Y^a \perp\!\!\!\perp A|L$ 天然成立), 则如同上一章所述一样, 这里的因果效应可以识别: 没有其他假设下, 其因果性风险比是 1。然而, 如果表 3.1 的数据来自一项观察性研究, 我们只有在假设了有界互换性 $Y^a \perp\!\!\!\perp A|L$ 成立的情况下, 才能计算出因果性风险比是 1。为了识别观察性研究中的因果效应, 我们需要一些数据之外的假设, 称作可识别性假设。事实上, 如果我们不在数据之外补充假设, 那表 3.1 的数据也可能以下几种情况:

- 如果 L 外的风险因素在已治疗人群中更多, 则因果性风险比小于 1。
- 如果 L 外的风险因素在未治疗人群中更多, 则因果性风险比大于 1。
- 如果 L 外的风险因素在已治疗人群和未治疗人群中一样多, 则因果性风险比等于 1。

本章讨论了因果效应均值非参数化识别的三个可识别性条件。在第十六章, 我们将讨论其他替代方案。

精讲点 3.2: 交叉随机试验 (原书第 29 页)

在精讲点 2.1 中, 我们讨论了什么是交叉试验。在交叉试验中, 我们让每个被试在多个时间段 (比如 $t=0$ 和 $t=1$) 接受不同的治疗, 然后进行观测。我们讨论了交叉试验中个体因果效应可识别的三个条件: 1) 上一次的治疗不对下一次的治疗造成影响, 即 $Y_{it=1}^{a_0, a_1} = Y_{it=1}^{a_1}$; 2) 治疗对个体的因果效应不受时间的影响, 即 $Y_{it}^{a_t=1} - Y_{it}^{a_t=0} = \alpha_i$; 3) 无治疗对个体的因果效应不受时间的影

响: $Y_{it}^{a_i=0} = \beta_i$ 。上一章的交叉试验讨论中没有涉及随机分组。现在让我们讨论交叉随机试验, 也即随机分配每个个体接受到的治疗顺序。

当条件 3) 不能成立的时候, 随机分配就显得非常重要。为了简单起见, 假设每个被试都有 50% 概率被随机分配到 ($A_{i1} = 1, A_{i0} = 0$) 或者 ($A_{i1} = 0, A_{i0} = 1$)。令 $Y_{i1}^{a_i=0} = Y_{i0}^{a_i=0} = r_i$ 。在条件 1) 和 2) 以及一致性成立下, 如果 $A_{i0} = 0$ 且 $A_{i1} = 1$, 则 $Y_{i1} - Y_{i0} = \alpha_i + r_i$; 如果 $A_{i0} = 1$ 且 $A_{i1} = 0$, 则 $Y_{i0} - Y_{i1} = \alpha_i - r_i$ 。因为 r_i 未知, 所以我们不能再识别个体的因果效应。大失, 因为 A_{i0} 和 A_{i1} 是随机分配的, 且都与 r_i 无关, 则 $(Y_{i1} - Y_{i0})A_{i1} + (Y_{i0} - Y_{i1})A_{i0}$ 的均值就是因果效应均值的估计值, 即 $E[\alpha_i]$ 。如果仅条件 1) 成立, 则 $(Y_{i1} - Y_{i0})A_{i1} + (Y_{i0} - Y_{i1})A_{i0}$ 的均值就是在时间 0 和 1 时两个效应均值的均值, 即 $(E[\alpha_{i1}] + E[\alpha_{i0}]) / 2$, 其中 $\alpha_{it} = Y_{it}^{\alpha_i=1} - Y_{it}^{\alpha_i=0}$ 。

总而言之, 如果条件 1) 成立, 即上一次的治疗不对下一次的治疗造成影响, 则交叉试验就能用于估计因果效应的均值。然而, 对绝大多数本书中的例子来说 (甚至现实中的研究来说), 这一假设都不可能成立。

精讲点 3.3: 可能世界 (原书第 35 页)

一些科学哲学家用“可能世界”来定义因果性。现实世界是事物现实样貌的世界。可能世界则是事物可能样貌的世界。让我们想象一个可能世界 a , 其中每个人都接受治疗 a ; 而另一个可能世界 a' , 其中每个人都接受治疗 a' 。如果两个可能世界的结局均值不等, 即 $E[Y^a] \neq E[Y^{a'}]$, 并且这两个可能世界分别是最接近现实世界中每个人都接受 a 或 a' 时的情形, 则这些哲学家认为存在因果效应。

我们引进反事实结局 Y^a 的概念, 用以表示一个个体接受了足够良定的治疗 a 后的结局。这些哲学家们则更喜欢将 Y^a 视作可能世界中的结局, 这个可能世界是最接近现实的、每个人都接受治疗 a 的世界。这两个定义是基本等价的, 唯一区别在于治疗是发生在现实世界, 还是可能世界。一个的难点在于要具体说明治疗 (或干预) 的具体形式, 而另一个则要描绘与现实世界差距最小的可能世界。Stalnaker (1968) 和 Lewis (1973) 讨论了基于可能世界的反事实理论。

精讲点 3.4: 归因比例 (原书第 38 页)

我们已经讨论并了解了因果性风险比 $\Pr[Y^{a=1} = 1] / \Pr[Y^{a=0} = 1]$ 和因果性风险差 $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$, 这两者都是比较 $a=1$ 和 $a=0$ 下的反事实结局。然而, 有些人可能更想比较观测结局和 $a=1$ 或 $a=0$ 下的反事实结局。这一比较能让我们在观察性研究中计算有多少病例的增加 (或预防) 可以归因于我们的治疗 (或干预), 也即如果我们对所有人都进行治疗, 会阻止多少病例的出现。比如, 20 个人参加了一次宴会, 他们中有 10 个人喝了红酒, 另外 10 个人则喝了蜂蜜。第二天, 7 个喝红酒的人和 1 个喝蜂蜜的人生病了。在互换性成立的情况下, 因果性风险比是 $0.7 / 0.1 = 7$, 因果性风险差是 $0.7 - 0.1 = 0.6$ 。后来发现是因为红酒放得太久变质了。现在我们要解决的问题是, 有多少病例能够归因于喝了红酒?

我们总共观察到 8 个人生病, 也就是观测到的总风险是 $\Pr[Y = 1] = 8 / 20 = 0.4$ 。而如果所有人都喝蜂蜜 ($a=0$) 的话, 风险应该是 $\Pr[Y^{a=0} = 1] = 0.1$ 。两者之差为 $0.4 - 0.1 = 0.3$ 。也就是, 多出 30% 的人, 如果喝蜂蜜的话, 就不会生病了。又因 $0.3 / 0.4 = 75\%$, 我们会说有 75% 的病例可以归因于喝了红酒 ($a=1$)。这一超出比例 (excess fraction), 或者也称归因比例, 被定义为:

$$\frac{\Pr[Y = 1] - \Pr[Y^{a=0} = 1]}{\Pr[Y = 1]}$$

精讲点 5.4 讨论了在充分病因模型框架下的超出比例。

一般而言, 超出比例和病因分值 (etiological fraction) 不同。病因分值被定义为在机理上由某一暴露 (或治疗、干预) 引起的病例的百分比, 是另一种形式的归因比例。比如, 如果治疗组有 1 个病例, 但是这个治疗组如果在未治疗的情况下有 7 个病例, 而这 7 个病例不包含之前的那 1 个病例, 那超出比例和病因分值就不一样。此时, 超出比例是病因分值的下限。因为病因分值的定义不依赖超出的病例数, 所以只有在很强的假设下, 才能在随机试验中计算病因分值 (参见 Green 和 Robins 在 1988 年所著论文)。

知识点 3.1: 标准化和逆概率加权中的正数性 (原书第 32 页)

我们已经定义了治疗为 a 时的标准化均值 $\sum_l E[Y | L = l, A = a] \Pr[L = l]$ 。然而, 只有在 $E[Y | L = l, A = a]$ 是良定的时候, 也即对于所有 $\Pr[L = l] \neq 0$ 有 $\Pr[A = a | L = l] > 0$ 的时候, 也就是正数性成立的时候, 我们才能计算这一表达式。

作者: Miguel A. Hernan, James M. Robins;

翻译: 罗家俊

当正数性不成立的时候, 逆概率加权均值 $E\left[\frac{I(A=a)}{f(A|L)}Y\right]$ 就不再等于 $E\left[\frac{I(A=a)}{f(a|L)}Y\right]$ 。具体而言, 因为 $\frac{0}{0}$ 无意义, 因此 $E\left[\frac{I(A=a)}{f(a|L)}Y\right]$ 无意义。另一方面, $E\left[\frac{I(A=a)}{f(A|L)}Y\right]$ 总是良定的, 因为其分母 $f(A|L)$ 不可能等于 0。然而, 在互换性下, 此时反事实均值的估计是有偏的。尤其是

当正数性不成立时, $E\left[\frac{I(A=a)}{f(A|L)}Y\right]$ 就等于

$\Pr[L \in Q(a)] \sum_l E[Y | L = l, A = a, L \in Q(a)] \Pr[L = l | L \in Q(a)]$, 其中集合

$Q(a) = \{l; \Pr[A = a | L = l] > 0\}$ 。因此, 在互换性下,

$$E\left[\frac{I(A=a)}{f(A|L)}Y\right] = E\left[Y^a | L \in Q(a)\right] \Pr[L \in Q(a)]。$$

从 $Q(a)$ 的定义中可知, 当 A 是二分变量且正数性不成立时, $Q(0) \neq Q(1)$ 。这种情况下, 即使互换性成立, $E\left[\frac{I(A=1)}{f(A|L)}Y\right] - E\left[\frac{I(A=0)}{f(A|L)}Y\right]$ 也没有因果推断上的意义, 因为它比较的是两个完全不同的组中的被试。当正数性成立时, 在互换性下, $Q(0) = Q(1)$, 且

$E\left[\frac{I(A=1)}{f(A|L)}Y\right] - E\left[\frac{I(A=0)}{f(A|L)}Y\right]$ 就是因果效应的均值。

知识点 3.2: 似是而非的一致性 (原书第 40 页)

治疗 R 有多种不同的形式。有趣的是, 就算这些不同的形式是劣定的, 一致性在某种程度上依然成立。记 $A_i(r)$ 为个体 i 接受的治疗形式 $R_i = r$ 。对 $R_i \neq r$ 的个体, 我们定义 $A_i(r) = 0$ 。因此 $A_i(r) \in \{0\} \cup A(r)$ 。对所有个体 i , 一致性可以表述为:

$$\text{当 } R_i = r \text{ 且 } A_i(r) = a(r) \text{ 时, } Y_i = Y_i^{r,a(r)}$$

也就是, 对于实际治疗形式 $R = r$ 的个体来说, 其结局等于如果他接受的治疗形式为 $R = r$ 的反事实结局。这一陈述对于实际接受的治疗形式 $R_i = r$ 且 $A_i(r) = a(r)$ 的个体而言是正确的。然而, 使用这一定义是自欺欺人的, 因为正如正文中所讨论的一样, 这一定义无助于我们理解效应估计值, 也阻碍了我们对互换性和正数性的评估。

同理, 我们可以思考一项假想干预: 通过改变肥胖的决定因素让所有被试肥胖, 而我们改变的肥胖决定因素的比例分布, 需要和研究人群中肥胖被试的肥胖决定因素的比例分布一样。这样一来, 每个被试都能随机被分配到一种干预形式, 使得反事实世界中干预形式的分布, 和现实研究人群中肥胖人群肥胖原因的分布完全一样。同理, 我们可以假想出一种对非肥胖的干预措施。

许多观察性研究在比较 $\Pr[Y = 1 | A = 0]$ 和 $\Pr[Y = 1 | A = 1]$ 并作出因果性的阐释时, 都隐含了上述对治疗形式的定义。问题在于, 这样的定义可能并不与我们实际想研究的干预措施相符合。就算我们知道“和非肥胖人群中肥胖决定因素比例分布相同的干预措施能降低 30% 死亡率”并不意味着现实中的干预措施(比如控制饮食)也能降低 30% 死亡率。因为, “肥胖决定因素”可能包括对基因进行干预, 但在现实中我们不能实现, 也就不可能达成 30% 的降低。

第三章图表

Table 3.1

	<i>L</i>	<i>A</i>	<i>Y</i>
Rheia	0	0	0
Kronos	0	0	1
Demeter	0	0	0
Hades	0	0	0
Hestia	0	1	0
Poseidon	0	1	0
Hera	0	1	0
Zeus	0	1	1
Artemis	1	0	1
Apollo	1	0	1
Leto	1	0	0
Ares	1	1	1
Athena	1	1	1
Hephaestus	1	1	1
Aphrodite	1	1	1
Cyclope	1	1	1
Persephone	1	1	1
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0

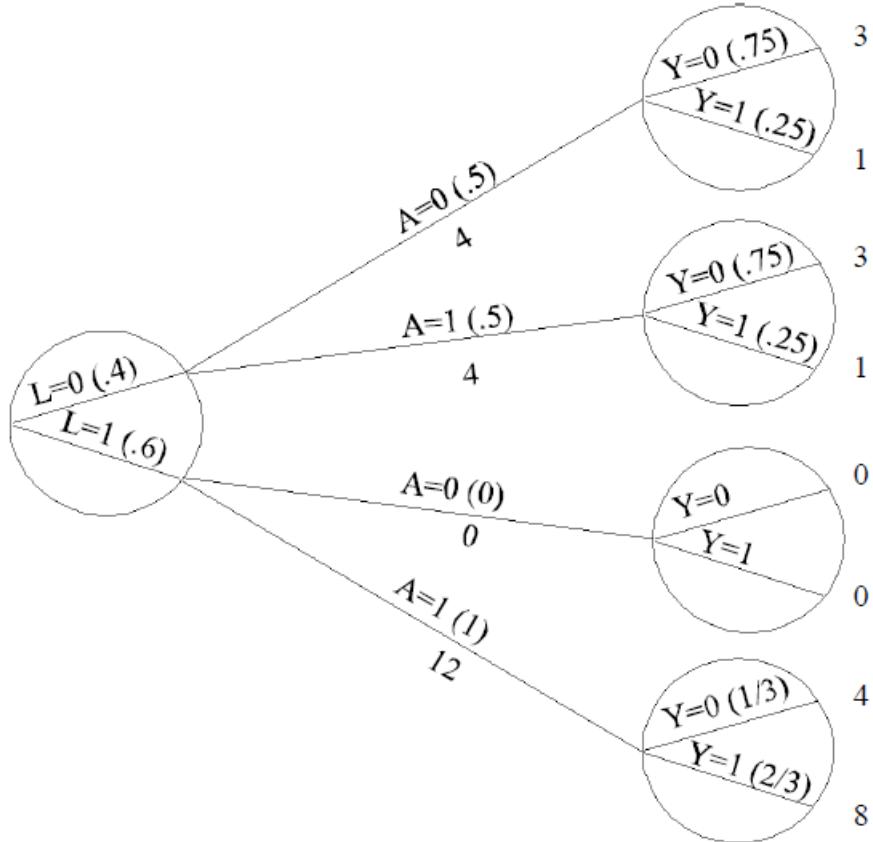


Figure 3.1

第四章 效应修饰

41 我们迄今的讨论都是围绕整个人群中的因果效应。然而, 有一些问题却只关乎一部分人而非所有人。再考虑一下这个因果性的问题: “一个人抬头向天上看会影响其他行人也抬头向天上看吗?”也许你觉得更有意义的做法, 是把行人分成本市居民和外来游客, 然后分别在这两个群体中衡量“抬头看”这个行为的因果效应。

是在整个人群中计算因果效应, 还是在一个子群体中进行计算, 取决于我们的研究目标。在某些情况下, 你并不关心不同群体中的效果差异。比如, 你是一名政策制定者, 正在考虑在全国范围内推行氟化水项目。因为这一项目会影响到全国的所有家庭, 所以你的主要关注点在整个人群中的效果, 而不是某一特定群体中的效果。而当你的干预措施是针对某些特定群体时, 你的关注点可能就是这项干预措施在整个人群不同子群体中的差异。

在这一章我们将会强调, 一项治疗(或干预)并没有一个固定的因果效应均值, 其因果效应均值取决于研究人群中的各种特征。

4.1 效应修饰的定义

我们在本书开始的时候, 计算了宙斯家庭 20 个成员中心脏移植 A 对死亡 Y 的因果效应均值。我们的计算使用了表 1.1 中每个人的反事实结局(通常是无法观测的) $Y^{a=0}$ 和 $Y^{a=1}$ 。在检验了表 1.1 的数据后, 我们得到结论: 因果效应均值为零。这个人群如果所有人都接受心脏移植, 会有一半人去世, $\Pr[Y^{a=1} = 1] = 10 / 20 = 0.5$; 如果没有人接受心脏移植, 也会有一半人去世, $\Pr[Y^{a=0} = 1] = 10 / 20 = 0.5$ 。所以因果性风险差为 $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1] = 0$ 。

现在我们来考虑两个新问题: 在女性中, A 对 Y 的因果效应均值是多少? 在男性中呢? 为了回答这个问题, 我们依然需要表 1.1 的数据。为了方便起见, 我们新建了表 4.1, 其包括表 1.1 的所有信息, 此外还多出一行 V 用来表示性别: $V = 1$ 表示女性, $V = 0$ 表示男性。同时我们重新排列了一下这个表, 这样前 10 行是女性, 后 10 行是男性。

让我们先来计算女性中的因果效应均值。我们要把所有分析都局限在前 10 行, 也就是 $V = 1$ 的个体当中。在这个子群体中, 所有人有治疗时的死亡风险是 $\Pr[Y^{a=1} = 1 | V = 1] = 6 / 10 = 0.6$; 所有人无治疗时的风险是 $\Pr[Y^{a=0} = 1 | V = 1] = 4 / 10 = 0.4$ 。因而, 因果性风险比是 $0.6 / 0.4 = 1.5$, 因果性风险差是 $0.6 - 0.4 = 0.2$ 。也就是, 平均而言, 在女性中, 心脏移植提高了死亡风险。

42

接下来让我们计算男性中的因果效应均值。我们要把所有分析都局限在后 10 行, 也就是 $V = 0$ 的个体当中。在这个子群体中, 所有人有治疗时的死亡风险是

$$\Pr[Y^{a=1} = 1 | V = 0] = 4/10 = 0.4; \text{ 所有人无治疗时的风险是 } \Pr[Y^{a=0} = 1 | V = 0] = 6/10 = 0.6.$$

因而, 因果性风险比是 $0.4/0.6 = 2/3$, 因果性风险差是 $0.4 - 0.6 = -0.2$ 。也就是说, 平均而言, 在男性中, 心脏移植降低了死亡风险。

这个例子很好地说明了, 就算整个人群中的因果效应均值为零, 在这个人群的子群体中, 因果效应均值不一定为零。因而, 在表 4.1 中, 因果效应零假设在整个人群中为真, 但在分开的男性或女性中不为真。这种情况非常巧合, 因为此时男性和女性中的因果效应大小正好一样, 只是方向相反。又因每种性别各占 50% 的比例, 所以在整个人群中, 两个子群体的因果效应相互抵消。虽然这种正好相互抵消的案例很少见, 但是不同群体间的差异性却常见于现实研究, 主要因为每个个体对同一治疗 (或干预) 的反应不尽一致。一个例外是极端因果零假设为真的时候, 此时治疗 (或干预) 对所有个体的因果效应都为零, 也就不存在任何差异, 同时该治疗 (或干预) 在整个人群以及各子群体中的因果效应均值都为零。

现在我们可以给效应修饰下一个定义了。如果 A 对 Y 的因果效应均值在 V 的不同取值下不尽相同, 则我们称 V 是 A 对 Y 因果效应的修饰因子。因为因果效应均值能用不同的效应量度 (比如风险差、风险比) 进行衡量, 因此效应修饰存在与否依赖于所使用的效应量度。比如, 在我们的例子中, 因为性别 V 取值不同时因果性风险比也不同, 因此性别 V 在乘法尺度上是心脏移植 A 对死亡风险 Y 的效应的修饰因子。我们只将不受到治疗 (或干预) A 影响的变量 V 称作修饰因子。

(第六章 6.5 小节详细讨论了对效应修饰因子的结构性分类)

(加法尺度上的效应修饰: $E[Y^{a=1} - Y^{a=0} | V = 1] \neq E[Y^{a=1} - Y^{a=0} | V = 0]$)

(乘法尺度上的效应修饰: $\frac{E[Y^{a=1} | V = 1]}{E[Y^{a=0} | V = 1]} \neq \frac{E[Y^{a=1} | V = 0]}{E[Y^{a=0} | V = 0]}$)

在表 4.1 中, 女性 ($V = 1$) 的因果性风险比大于 1, 男性 ($V = 0$) 的因果性风险比小于 1; 女性 ($V = 1$) 的因果性风险差大于 0, 男性 ($V = 0$) 的因果性风险差小于 0。如果 $V = 1$ 和 $V = 0$ 时因果效应均值的方向相反, 我们称之为质的效应修饰。当质的效应修饰存在时, 加法 (或称加性、加法尺度上) 效应修饰也就意味着乘法 (或称乘性、乘法尺度上) 效应修饰, 反之亦然。如果没有质的效应修饰, 可能某个尺度上存在效应修饰 (比如乘法尺度上), 但在另一个尺度上不存在效应修饰 (比如加法尺度上)。我们用以下一组新的数据来说明这一点: 已知

$$\Pr[Y^{a=0} = 1 | V = 1] = 0.8, \quad \Pr[Y^{a=1} = 1 | V = 1] = 0.9, \quad \Pr[Y^{a=0} = 1 | V = 0] = 0.1,$$

$\Pr[Y^{a=1} = 1 | V = 0] = 0.2$ ，简单计算我们可以知道，在加法尺度上不存在效应修饰

($0.9 - 0.8 = 0.2 - 0.1 = 0.1$)，但在乘法尺度上存在效应修饰 ($0.9 / 0.8 = 1.1 \neq 0.2 / 0.1 = 2$)。

因此，只有在指明了效应量度（风险比或风险差）的情况下，我们才能说是否存在效应修饰。一些研究者更喜欢“效应量度修饰”一词，而不是“效应修饰”，因为他们希望以此强调效应修饰这一概念取决于我们所使用的效应量度。

(我们不会讨论在比值比尺度上的效应修饰，因为比值比这一参数很少且很难用于因果推断)

43 4.2 通过分层识别效应修饰

分层分析是识别效应修饰的最直接方式。为了识别 V 是否修饰 A 对 Y 的效应，我们可以计算 V 的不同取值（每一取值是一个分层）下 A 对 Y 的因果效应。在上一节，我们用表 4.1 的数据计算了性别 V 分层中 A 对 Y 的因果效应。因为这两层中的因果效应不一样（在加法尺度上和乘法尺度上都不一样），所以我们说 V （在加法尺度和乘法尺度上）修饰了 A 对 Y 的效应。

(分层: 在 V 的每一层（每一取值）中计算 A 对 Y 的因果效应。对二分变量 V ，分层因果性风险差是 $\Pr[Y^{a=1} = 1 | V = 1] - \Pr[Y^{a=0} = 1 | V = 1]$ 和 $\Pr[Y^{a=1} = 1 | V = 0] - \Pr[Y^{a=0} = 1 | V = 0]$)

但是表 4.1 的数据并不是研究者在现实中经常遇见的数据。在现实中，我们不可能同时有反事实结局 $Y^{a=0}$ 和 $Y^{a=1}$ 的数据，而只有每个个体观测到的治疗状况 A 和观测到的结局 Y 。因此，我们在没有反事实结局的情况下该怎么使用分层分析去检验是否存在效应修饰？这个问题的答案取决于我们的研究设计。

让我们先考虑一个理想的边缘随机试验。在第二章，我们讨论了在没有随机变异性的情况下，我们依然可以使用观测到的数据计算治疗的因果效应均值，此时因果性风险差

$\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$ 就等于观测到的相关性风险差 $\Pr[Y = 1 | A = 1] - \Pr[Y = 1 | A = 0]$ 。同样的推理可以适用于变量 V 的每一层，这是因为如果治疗分组是随机且无条件的，那互换性在子群体中也依然成立。因此女性中的因果性风险差 $\Pr[Y^{a=1} = 1 | V = 1] - \Pr[Y^{a=0} = 1 | V = 1]$ ，就等

于女性中观测到的相关性风险差 $\Pr[Y = 1 | A = 1, V = 1] - \Pr[Y = 1 | A = 0, V = 1]$ 。对男性同理。因此，在无条件随机分组的理想实验中，研究者只需要直接进行分层分析就能识别 V 的效应修饰作用。分层分析能够用来计算子群体中的因果效应均值，但是不能计算个体的因果效应（参见精讲点 2.1 和 3.2）。

考虑另一个理想的条件随机试验。在 40 个被试中, 重症被试 ($L = 1$) 有 75% 的概率被分配到治疗组, 轻症被试 ($L = 0$) 有 50% 概率被分配到治疗组。这 40 个被试来自两个国家: 其中 20 人来自希腊 ($V = 1$), 另外 20 人来自罗马 ($V = 0$)。表 2.2 (表 3.1 与之相同) 展示了 20 个希腊被试的数据, 表 4.2 展示了 20 个罗马被试的数据。这个研究人群在有治疗时的反事实结局 $\Pr[Y^{a=1} = 1] = 0.55$, 在没有治疗时的反事实结局 $\Pr[Y^{a=0} = 1] = 0.40$ (反事实结局可以由标准化或者逆概率加权得到, 细节留给读者自己计算)。因此, 治疗 A 对死亡 Y 的因果效应均值是: 因果性风险差为 $0.55 - 0.40 = 0.15$, 因果性风险比为 $0.55 / 0.40 = 1.375$ 。因而在整个研究人群中, 治疗会提高死亡风险。

在上一章我们已经讨论过了, 如果这个数据来自于观察性研究, 在有界互换性 $Y^a \perp\!\!\!\perp A|L$ 成立的情况下, 因果效应的计算方法相与之同。

现在我们通过分层分析来检验国籍 V 是否修饰了 A 对 Y 的效应。我们需要分别计算希腊人中 A 对 Y 的效应 $\Pr[Y^{a=1} = 1|V=1] - \Pr[Y^{a=0} = 1|V=1]$, 和罗马人中 A 对 Y 的效应

$\Pr[Y^{a=1} = 1|V=0] - \Pr[Y^{a=0} = 1|V=0]$ 。如果这两个因果性风险差不等, 则 V 在加法尺度上有 44 效应修饰作用。同理, 如果需要检验乘法尺度上的效应修饰作用, 我们需要计算因果性风险比。

计算每一分层 v 中的 $\Pr[Y^{a=1} = 1|V=v]$ 和 $\Pr[Y^{a=0} = 1|V=v]$ 分成两步: 1) 先把人群按照 V 分层, 2) 再根据 L 进行标准化或逆概率加权。我们在第二章中利用标准化计算了希腊人 ($V=1$) 中的因果效应, 并且知道了因果性风险差是 0, 因果性风险比是 1。现在我们要用同样的方法来计算罗马人 ($V=0$) 的因果效应, 我们得到 $\Pr[Y^{a=1} = 1|V=0] = 0.6$ 和 $\Pr[Y^{a=0} = 1|V=0] = 0.3$ 。因此, 罗马人中因果性风险差是 0.3, 因果性风险比是 2。因为这两个效应量度都和希腊人中的不一样, 因此我们说国籍 V 同时在加法尺度和乘法尺度上修饰了治疗 A 对死亡 Y 的效应。这里的修饰作用并没有产生质的变化, 因为这个因果效应在 $V=0$ 和 $V=1$ 时要么有害, 要么为零。

(分层分析的第二步在 V 就是决定有界互换性的 L 时可以忽略, 参见 4.4 节)

现在我们已经知道了, 在我们的研究人群中, 国籍 V 会修饰治疗 A 对死亡 Y 的因果效应。然而, 对修饰作用背后的机理, 我们却没有做任何解释。事实上, 国籍可能不是修饰作用的根本原因, 只是这个根本原因的一个标志物。比如, 希腊的心外科手术技术要强于罗马, 因而研究者就会在数据中发现国籍修饰了治疗的效应。也因此, 一项提高罗马心外科手术技术的干预措施也许

就能消除国籍的修饰作用。当这种情况发生的时候,为了强调两者之间的区别,我们会将国籍称作修饰因子替代物,将心外科手术技术称作因果性修饰因子。

(第六章 6.6 小节用图像的方法表示了修饰因子替代物和因果性修饰因子)

因此,当我们说 V 有效应修饰作用的时候,并不一定百分之百地意味着 V 因果性地改变了治疗(或干预)的效应。为了避免可能的混淆,有一些作者更喜欢用另一个更中性的表达“ V 不同分层中的因果效应异质性”,而不是“ V 的效应修饰作用”。下一章我们将介绍“交互作用”,一个和效应修饰有关,并且能将所涉及的变量和治疗(或干预)效果因果性地联系起来的概念。

4.3 为什么要关注效应修饰

以下几个原因解释了为什么研究者需要检验数据中的效应修饰作用,以及为什么在随机试验中需要收集进行治疗(或干预)前的人群特征变量 V 。

第一点,如果 V 修饰了治疗 A 对结局 Y 的效应,那治疗的因果效应均值就会根据人群中 V 的出现频率而变化。比如,我们表 4.1 的数据显示因果效应均值在女性中有害,但在男性中有益,也就是有质的效应修饰作用。因为两种性别各占人数的 50%,且每个性别中的因果效应大小相等,只是方向相反,所以整个人群中的因果效应均值为零。然而,如果我们研究是在女性占比更高的人群中开展的,我们就会发现整个人群中的因果效应均值是有害的。当非质的效应修饰存在时,因果效应均值的大小可能在不同子群体中不一样,但是方向都会是一样的。非质的效应修饰例子,包括石棉暴露的影响(在吸烟者和不吸烟者中不同)和全民医保的影响(在低收入和高收入家庭中不同)。

也就是说,人群中的因果效应均值取决于人群中每个个体效应均值的分布。因而,治疗 A 对结局 Y 的因果效应均值不可能是固定的,其取决于研究人群的构成特征。

如果在一个人群中计算得到的因果效应能被外推到另一个人群,我们会说这个因果推断在不同人群中具有可移植性(参见精讲点 4.2)。在我们的例子中,心脏移植 A 对死亡 Y 的因果效应在男性和女性、以及希腊人和罗马人中不尽相同。因此,这个人群中的因果效应均值就不能外推到性别和国籍分布不同的另一个人群中。

(一些研究者也将可移植性的缺失称作外在效度的缺失)

修饰因子每层中的条件因果效应相比整个人群中的因果效应更具可移植性,但也不能保证一个人群中的条件因果效应一定等于另一个人群中的条件因果效应。这是因为其他未知的因果性修饰印在在两个人群中可能有不一样的条件分布(其他原因参见精讲点 4.2)。这些未知的修饰因子不是用来达成互换性的变量,而仅仅是结局的风险因素。因此,一个效应在不同人群中的可移

植性, 相较于某一个人群中因果效应的可识别性而言, 是一个更加棘手的难题。因为你不仅要根据影响治疗分组的变量(也许你能从负责分组的研究者那得到这些信息)进行分层从而保证互换性, 还要再根据结局的影响因素进行分层从而检验可能存在效应修饰, 遗憾的是, 们对结局的影响因素知之甚少。

(可移植性不是什么大问题的一个例子: *Smith 和 PeI1 (2003)* 找不到任何变量能修饰高空跳伞对死亡的效应; 因此他们认为在一个特定人群中开展的关于跳伞的随机试验, 其结论可以移植到其他人群中)

因此, 因果效应的可移植性是一个不能被验证的假设, 需要具体问题具体分析。比如, 一些研究者认为给尼日尔每户家庭额外 100 美元带来的健康效应收益, (无论在加法尺度上还是乘法尺度上) 都不能移植到荷兰。但另一些研究者认为在欧洲使用降固醇药物的健康收益, 可以移植到加拿大去。

47 第二点, 检验效应修饰存在与否, 有助于确定能从干预措施中获益最多的群体。在我们表 4.1 的例子中, 治疗 A 对 Y 的因果效应均值为零。然而治疗 A 在男性 ($V = 0$) 中是有益的, 但在女性 ($V = 1$) 中是有害的。如果医生知道性别有质的效应修饰作用, 在没有其他额外信息的情况下, 他只会给男性进行治疗。可能某些情况会稍复杂一些, 就像我们的第二个例子, 其中有乘法效应修饰, 但没有加法效应修饰。此时治疗会在 $V = 0$ 和 $V = 1$ 两个组中都让风险的数值减去 10%, 因而尽管存在乘法尺度上的效应修饰, 治疗在所有病人中都是同样的效果。实际上, 如果 V 的某一分层中存在非零的因果效应, 且每一层的反事实结局 $\Pr[Y^{a=0} = 1 | V = v]$ 都不一样, 那肯定在乘法尺度或加法尺度上存在效应修饰。

加法尺度, 而非乘法尺度, 是我们用来确定哪一群体能从干预措施中获益最多的正确尺度。在加法尺度上不存在效应修饰时, 就算乘法尺度上存在效应修饰, 也无助于事。

(有许多研究者, 包括 Blot, Day, Rothman, Saracci 等人, 将加法尺度上的效应修饰视作更具有公共卫生意义的指标)

在我们的第二个例子中, 乘法效应修饰只有数学上的意义, 它仅仅意味着每一层中无治疗时的风险 $\Pr[Y^{a=0} = 1 | V = v]$ 不一样。因而, 这种情况下, 更有意义的做法, 是报告每一层中的 $\Pr[Y^{a=1} = 1 | V = v]$ 和 $\Pr[Y^{a=0} = 1 | V = v]$, 而不是一个简单差值或比值。

第三点, 也是最后一点, 识别效应修饰有助于我们了解背后的生物学、社会学或其他方面的具体机制。比如, 如果我们知道有没有做过包皮手术对 HIV 感染有修饰作用, 这将有助于我们了解这个病的传染机理。在我们描述两个变量的交互作用之前, 我们要先识别效应修饰。“效应修

饰”和“交互作用”有时被用作同义词。这一章仅关注“效应修饰”，下一章我们将讨论“交互作用”，其虽然和效应修饰有关，但却是一个完全不同的概念。

4.4 通过分层调整变量

本章之前，我们的目标是计算整个人群中的因果效应均值。在没有无条件随机分组的情况下，我们需要对变量 L 进行调整（也称控制），从而保证治疗组和非治疗组的有界互换性。比如，在第二章中，我们得到心脏移植 A 对死亡 Y 的因果效应均值为零，也即因果性风险比 $\Pr[Y^{a=1} = 1] / \Pr[Y^{a=0} = 1] = 1$ 。利用表 2.2 的数据，我们通过标准化或者逆概率加权调整了 L 。

48 在这一章，我们又有了一个新的目标：识别变量 V 的效应修饰作用。为了达成这个目标，我们需要在调整 L 之前先根据 V 进行分层。比如，在这一章，我们先根据国籍 V 将人群分成了希腊人和罗马人，然后再用标准化或者逆概率加权方法调整了 L ，得到心脏移植 A 对死亡 Y 因果效应均值。总而言之，标准化或者逆概率加权是用来调整变量 L ，而分层分析则是用来识别 V 的效应修饰作用。

49 但是分层分析并不总是用来识别 V 的效应修饰作用，它还有其他用处。在实践中，分层分析经常用来替代标准化或逆概率加权对变量 L 进行调整。分层分析被广泛地用于调整变量 L ，以至于许多研究者认为“分层”和“调整”是一对同义词。比如，你让一名流行病学家用表 2.2 中的数据调整 L 后计算 A 对 Y 的效果，他有很大可能会将数据分成两部分，一部分是 $L = 0$ ，另一部分是 $L = 1$ ，然后在这两个数据中分别计算效应量度。比如说，他会分别计算 $l = 0$ 和 $l = 1$ 的风险比，得到 $\Pr[Y = 1 | A = 1, L = l] / \Pr[Y = 0 | A = 1, L = l] = 1$ 。

这种每一分层中的相关性风险比，在 L 的有界互换性下依然具有因果性的阐释意义：它们分别衡量了 $L = 0$ 和 $L = 1$ 两个子群体中的因果效应均值。因此，它们被称作条件效应量度。与之相比，我们在第二章计算得到的边缘（或称无条件）效应量度也等于 1。在这里，这三个风险比——其中两个是条件效应量度，一个是边缘效应量度——都等于 1，这是因为 L 没有效应修饰作用。分层分析必然会得到多个效应量度值（每层一个）。它们表示互不重叠的子群体中的因果效应均值，但没有一个表示整个人群中的因果效应均值。因此，我们并不认为分层分析能像第二章讨论的标准化和逆概率加权一样，可以用于计算整个人群的因果效应均值。因此我们将会把重心放在标准化和逆概率加权上。

（在 L 的有界互换性下，子群体 $L = l$ 中的风险比衡量了 $L = l$ 人群中的因果效应均值，这是因为如果 $Y^a \perp\!\!\!\perp A | L$ ，则有 $\Pr[Y = 1 | A = a, L = l] = \Pr[Y^a = 1 | L = l]$ ）

此外, 与标准化和逆概率加权不同的是, 用分层分析调整变量需要计算变量 L 不同组合下的每一个分层中的效应量度, 这里的 L 指的是所有涉及有界互换性的变量。比如, 如果我们用分层分析计算表 2.2 和表 4.2 数据中心脏移植的因果效应, 需要分别计算 $L=1$ 的罗马人、 $L=1$ 的希腊人、 $L=0$ 的罗马人、和 $L=0$ 的希腊人中的效应量度。此时我们不能忽视 L , 只计算希腊人或罗马人中的效应量度, 因为仅对国籍 V 分层不能保证有界互换性。

也就是说, 用分层分析计算效应量度, 需要我们衡量每一个涉及有界互换性的变量 L 的效应修饰作用, 而不管你是否对这个变量的修饰作用感兴趣。与之相比, 先把数据按照 V 分层, 然后再利用标准化或逆概率加权调整变量 L , 能让研究者将互换性和效应修饰分开处理, 更加简洁。

(Rubins (1986, 1987) 讨论了在某些情况下, 就算互换性、正数性和一致性成立, 时异治疗(或干预)的分层效应量度依然没有因果性阐释意义)

分层分析的另一问题, 是某些效应量度——比如比值比——不具伸缩性 (collapsibility, 参见精讲点 4.3); 以及在涉及时异治疗(或干预)的情况下, 如果时异变量 L 受到上一时间段治疗(或干预)的影响, 不恰当的变量调整会导致偏移 (详见本书第三部分)。

(除了互换性之外, 分层分析也要求正数性成立, 因为如果 $L=l$ 这一层中只有接受治疗或未接受治疗的人, 那我们就不能在这一层中计算因果效应)

有时研究者只会在 L 的某些分层中计算因果效应, 换句话说, 不会计算某些分层中的因果效应, 这被称作限制。对因果推断来说, 分层分析只是简单地将因果效应的计算限制到几个互不重叠的子群体中。在这些子群体中, 互换性成立。当正数性在某些分层中不成立时, 我们会将分析限制在正数性成立的分层中 (参见第三章)。

4.5 通过匹配调整变量

匹配是调整变量的另一种方法。匹配的目标是在人群中构建一个子群体, 变量 L 的分布在这 50 个子群体的治疗组和非治疗组中相同。比如, 在我们表 2.2 的数据中, 对每一个非治疗组的轻症被试 ($A=0, L=0$), 我们随机选取一个治疗组的轻症被试 ($A=1, L=0$) 进行匹配。同理, 对每一个非治疗组的重症被试 ($A=0, L=1$), 我们随机选取一个治疗组的重症被试

($A=1, L=1$) 进行匹配。我们将每一个非治疗组的被试和与之相匹配的治疗组被试称作一个匹配对, 将变量 L 称作匹配变量。假设我们有 7 个匹配对, $L=0$ 的有: Rheia-Hestia, Kronos-Poseidon, Demeter-Hera, Hade-Zeus; $L=1$ 的有: Artemis-Ares, Apollo-Aphrodite, Leto-Hermes。在我们匹配对中, 所有非治疗组的被试都被选中了, 而只有一个治疗组的被试没有被选

中。在由匹配对构成的子群体中, 重症 ($L = 1$) 的比例在治疗组和非治疗组中被有意地构造得一样 (都是 $3/7$)。

(我们对匹配的讨论仅限于队列研究。在病例对照研究中 (第八章有简略讨论), 我们经常对病人和非病人 (即对照) 进行匹配, 而不是对治疗和未治疗进行匹配。即使匹配变量能满足有界互换性, 依照病例进行匹配依然不能在匹配人群中满足治疗和未治疗的互换性。此时, 我们需要通过分层调整匹配变量从而计算每一层中的效应量度)

为了构造匹配人群, 我们将人群中的治疗组替换成原治疗组的一个子群体, 从而使匹配变量 L 的分布在治疗组和非治疗组中一样。在 L 的有界互换性假设下, 匹配能使匹配人群中治疗组和非治疗组的 (无条件) 互换性成立。因此, 我们就能直接在匹配人群中计算因果效应均值。在治疗组中, 死亡风险是 $3/7$, 在非治疗组中, 死亡风险也是 $3/7$, 因此因果性风险比是 1。同时, 匹配也能保证正数性成立, 因为只有治疗组或非治疗组的分层, 都被排除在外了。

通常情况下, 我们会以人数较少的一组 (在我们的例子中是非治疗组) 作为基础人群, 再从另一组 (在我们的例子中是治疗组) 寻找相应的匹配个体。基础人群决定了用以计算因果效应的匹配人群。在上一段, 我们计算了非治疗组中的因果效应, 而此时治疗组的人数比非治疗组的多, 所以我们只能大致计算治疗组的因果效应。同时, 匹配不一定是一对一的 (一个匹配对), 也可以是一对多的 (一个匹配组)。

在实践中, L 可以是含多个变量的向量。此时, 相匹配的治疗组被试和非治疗组被试, 需要同时来自多个变量取值组合而成的同一分层。

我们可以以 L 的任意分布形式进行匹配, 而不一定非要局限于 L 在治疗组或非治疗组中的分布形式。我们可以通过个体匹配 (如上面所述) 或频率匹配达成我们想要的分布形式。在我们的例子中, 如果我们从治疗组中随意选出一群人, 只要最后他们满足有 70% 的人是 $L = 1$, 这样的方法就叫频率匹配。

(当匹配变量的个数增加时, 对一个个体而言, 没有匹配对象的概率就会增加。在这种情况下如何找到最优的匹配对象超出了本书所讨论的范畴)

因为匹配人群只是原人群的一个子群体, 因此其中修饰因子的分布就会和原人群中有所不同。下一小节将会讨论这个问题。

4.6 效应修饰与变量调整

标准化、逆概率加权、分层分析、限制、匹配等方法虽然都能用来计算因果效应均值, 但是 51 得到的是不同类型的因果效应。根据因果效应的不同类型, 这些方法可以分作两类: 标准化和逆

概率加权能用来计算边缘（或称无条件）和条件效应，而分层分析、限制和匹配只能计算人群子群体中的条件效应。这几个方法都要求互换性和正数性成立。如果我们只对 $L = 1$ 时的条件效应感兴趣，只要 $L = 1$ 这一层中的互换性和正数性成立即可。如果需要计算整个人群中的边缘效应，互换性和正数性需要在 L 的所有取值下都成立。

在没有效应修饰的情况下，这几种方法计算得到的效应量度（如风险差或风险比）都应该相同。比如，利用表 2.2 的数据，我们在以前的计算中知道了，在整个人群中（通过标准化和逆概率加权），在重症 ($L=1$) 被试和轻症 ($L=0$) 被试中（通过分层分析），以及在非治疗组中（通过匹配），因果效应都为零。所有方法给出的结果都一样。然而，当有效应修饰存在的时候，这几个方法得到的结果不一定相等。比如表 4.3 的数据，我们需要计算心脏移植 A 对高血压 Z (1 表示有，0 表示没有) 的效应，假设互换性 $Z^a \perp\!\!\!\perp A|L$ 和正数性成立。

52 利用标准化和逆概率加权，我们能得到 $\Pr[Z^{a=1} = 1] / \Pr[Z^{a=0} = 1] = 0.8$ （计算细节留给读者）。利用分层分析可以得到 $L = 0$ 时有 $\Pr[Z^{a=1} = 1 | L = 0] / \Pr[Z^{a=0} = 1 | L = 0] = 2.0$ ， $L = 1$ 时有 $\Pr[Z^{a=1} = 1 | L = 1] / \Pr[Z^{a=0} = 1 | L = 1] = 0.5$ 。利用上一小节描述的匹配对，有 $\Pr[Z^{a=1} = 1 | A = 0] / \Pr[Z = 1 | A = 0] = 1.0$ 。

我们的计算得到了 4 个不同的风险比: 0.8, 2.0, 0.5, 还有 1.0。所有这些都是正确的。在不考虑随机变异性（参见知识点 4.2）的情况下，这些风险比不一样是因为存在质的效应修饰。重症患者中，治疗能使风险翻倍，轻症患者中，治疗能使风险减半。但在整个人群中，治疗却是有益的（ $\Pr[Z^{a=1} = 1] / \Pr[Z^{a=0} = 1] = 0.8$ ），这是因为重症组中非治疗的反事实结局和轻症组中非治疗的反事实结局比值，是轻症与重症比值的两倍还要多，即

$\Pr[Z^{a=0} = 1 | L = 1] / \Pr[Z^{a=0} = 1 | L = 0] > 2 \times \Pr[L = 0] / \Pr[L = 1]$ （参见知识点 4.3）。在非治疗组中的因果效应为零，反映了非治疗组中轻症的比例比整个人群中高。这个例子说明了事先确定人群或者子群体的重要性。

上一章我们说明了只有良定的因果效应才能用于有意义的因果推断。这一章我们说明了只有人群特征清晰的群体才能用于有意义的因果推断。这两个条件在满足一系列前提标准的试验人群中会自动满足。然而，这两个条件在观察性研究中却不会自动满足。在观察性研究中，研究者需要尽可能清晰定义我们关心的因果效应，以及用来计算这个效应的人群。否则，不同方法得到的不同效应估计会产生极大的误解。

(本书第二部分描述了怎么将标准化、逆概率加权和分层分析结合起来，在参数化或半参数化模型中使用。比如，传统的回归模型，就是在协变量组合而成的各个分层中，计算治疗和结局的相关性)

在我们上面的例子中，使用逆概率加权的研究者没必要和使用匹配的研究者争论到底谁的方法更好。他们得到的不同数值仅与他们想研究的问题有关，和他们选取的方法无关。事实上，第53个研究者也能用逆概率加权计算治疗组或非治疗组中的效应量度（参见知识点 4.1）。

最后想强调的是，分层分析能用来计算人群子群体中的因果效应均值，但是不能用来计算个体的因果效应。就如我们之前讨论过的一样，只有在某些极端假设的情况下，我们才能识别个体的因果效应（参见精讲点 2.1 和 3.2）。

第四章精讲点和知识点

精讲点 4.1：治疗组中的因果效应（原书第 44 页）

本章关注的是人群一个子群体中的因果效应均值。其中一个特别的子群体是治疗组 ($A = 1$)。如果 $\Pr[Y^{a=1} = 1 | A = 1] \neq \Pr[Y^{a=0} = 1 | A = 1]$ ，则治疗组中的因果效应均值为零。或者，由一致性，如果 $\Pr[Y = 1 | A = 1] \neq \Pr[Y^{a=0} = 1 | A = 1]$ ，则治疗组中的因果效应均值为零。也就是说，如果治疗组的观测结局，不等于治疗组在未治疗时的反事实结局，则因果效应不为零。此时，治疗组中的因果性风险差是 $\Pr[Y = 1 | A = 1] - \Pr[Y^{a=0} = 1 | A = 1]$ 。治疗组中的因果性风险比，也称作标准化发病率，就等于 $\Pr[Y = 1 | A = 1] / \Pr[Y^{a=0} = 1 | A = 1]$ 。非治疗组中的因果性风险差与因果性风险比的定义与之相似，仅需用 $A = 0$ 替换 $A = 1$ 即可。在计算治疗组（或非治疗组）中的因果效应时，我们比较的就不再是现实观测中的治疗组和非治疗组两个组的结局，而是治疗组（或非治疗组）的观测结局与其对应的反事实结局。图 4.1 展示了此时我们进行比较的两个结局。

如果个体的因果效应在治疗组与非治疗组中分布不同，那么治疗组中的因果效应均值会和整个人群中的因果效应均值不同。也就是说，在计算治疗组中的因果效应时，治疗变量 $A = 1$ 只是一个特征标志，其代表了真正修饰治疗组与非治疗组中因果效应的其他变量。我们可以说一个与试验无关的变量 V 有效应修饰作用——就算变量 V 只是因果性修饰因子的替代物（比如国籍）——但我们不能说治疗 A 有效应修饰作用。

第六章的 6.6 小节讨论了如何用图表示真正的修饰因子及其替代物。本书的大部分内容会主要关注整个人群中的因果效应，因为治疗组与非治疗组中的因果效应，很难直接推广到时异治疗中（参见本书第三部分）。

精讲点 4.2: 可移植性 (原书第 48 页)

我们从一个人群中得到的因果效应，大多时候会被用于对其他人群（称之为目标人群）的决策证据。假设我们在互换性、正数性以及一致性的假设下正确得到了我们研究人群中的因果效应均值，那在目标人群中，这个治疗的效应会是一样的吗？换句话说，我们能把研究人群中的治疗效应移植到目标人群中吗？这个问题的回答取决于两个人群的人群特征。具体而言，一个因果效应从一个人群到另一个人群的可移植性，取决于这两个人群的人群特征是否相似。

- 效应修饰：效应修饰因子会影响治疗的因果效应。如果两个人群的效应修饰因子分布不同，那同一治疗在两个人群中的因果效应均值也会不同。
- 治疗形式：治疗的因果效应也取决于其不同形式在两个人群中的分布。如果分布不同，效应均值也会不同。
- 干扰：在本书正文中，我们都假定没有干扰（参见精讲点 1.1）。然而干扰在现实中确是实际存在的，一个个体的结局可能会影响到其他个体的结局。因此，两个人群中不同的社会交往形式会使治疗的因果效应不同。

可以通过以下方式提高因果推断在不同人群间的可移植性：(1) 将关注点限制在特定分层中的人群；(2) 利用研究人群中每一分层的因果效应，构造符合目标人群特征的因果效应均值。比如，我们书中例子里，涉及了四个分层，即罗马女性、希腊女性、罗马男性、希腊男性。我们可以对每一分层的因果效应进行加权平均，从而构造出满足不同性别比例、国籍比例的目标人群，并得到其中的因果效应均值，而权重就是每一分层在目标人群中所占的比例。然而，我们依然不能保证在构造所得人群中的因果效应均值，等于实际目标人群的真实值，这是因为未知的修饰因子、干扰现象以及治疗形式可能在研究人群与目标人群中分布不同。

精讲点 4.3: 伸缩性与比值比 (原书第 54 页)

乘法效应修饰不存在的情况下，整个人群的因果性风险比等于 V 的每个分层 v 的条件风险比，即 $\Pr[Y^{a=1} = 1] / \Pr[Y^{a=0} = 1] = \Pr[Y^{a=1} = 1 | V = v] / \Pr[Y^{a=0} = 1 | V = v]$ 。更广义上，就算乘法效应修饰存在，整个人群的因果性风险比等于各个分层风险比的加权平均。比如，如果 $V = 1$ 和 $V = 0$ 时，风险分别是 2 和 3，则整个人群中的因果性风险比就会在 2 和 3 之间。人群中的因

果性风险比（或风险差），永远不会超过分层因果比（或风险差）给定的上下限。这一性质被称为伸缩性，也是效应量度的理想性质之一。

但有一些效应量度并不具伸缩性，比值比就是其中之一。

表 4.4 的数据是一项关于海拔 A 和抑郁 Y 的研究，其中有 20 个被试。在 20 个被试在研究开始时都不抑郁。 $A=1$ 表示把这个被试挪到高海拔地区， $A=0$ 则表示没有。 $Y=1$ 表示抑郁， $Y=0$ 则没有。 $V=1$ 表示女性， $V=0$ 表示男性。是否挪动被试是随机的因此互换性 $Y^a \perp\!\!\!\perp A$ 成立。整个人群的因果性风险比是 $\Pr[Y=1|A=1]/\Pr[Y=1|A=0]=2.3$ ，因果性比值比是

$$\frac{\Pr[Y^{a=1}=1]/\Pr[Y^{a=1}=0]}{\Pr[Y^{a=0}=1]/\Pr[Y^{a=0}=0]} = \frac{\Pr[Y=1|A=1]/\Pr[Y=0|A=1]}{\Pr[Y=1|A=0]/\Pr[Y=0|A=0]} = 5.4。 \text{ 风险比和比值比都能用}$$

来衡量同一个因果效应，只是在不同的尺度上衡量。

现在我们来计算每一个性别分层中的风险比和比值比。 $V=0$ 和 $V=1$ 时，条件风险比 $\Pr[Y=1|A=1, V=v]/\Pr[Y=1|A=0, V=v]$ 分别是 2 和 3。 $V=0$ 和 $V=1$ 时，条件比值比 $\frac{\Pr[Y=1|A=1, V=v]/\Pr[Y=0|A=1, V=v]}{\Pr[Y=1|A=0, V=v]/\Pr[Y=0|A=0, V=v]}$ 都是 6。整个人群中的因果性风险比 2.3 位于 2 和 3 之间。然而人群中的比值比是 5.4，小于每一分层中的比值比。当 V 是 Y 的独立风险因子，且 A 与 V 相互独立时，整个人群中的因果性比值比相较每一分层中的非零比值比而言，更接近零值。

整个人群中效应量度能用每一分层的加权表示，这一性质称为伸缩性。风险比和风险差都具有伸缩性，但是比值比——以及更不常用的比值差（odds difference）——不具伸缩性，因而是 Jensen 不等式的一个特例（Samuels, 1981），从而造成一些反直觉的现象，就如上面描述的例子。在极端零假设情况下，比值比具有伸缩性——因为此时边缘效应量度和条件效应量度都是 0。而在结局非常罕见的情况下（比如<10%），比值比近似等于风险比，因而近似具有伸缩性。

比值比不具伸缩性的另一个后果是，我们不能将“互换性缺失”和“条件比值比与边缘比值比不同”等同起来。在我们的例子中，即使治疗组和非治疗组是可互换的，但是比值比却差了 10% 左右 ($1-6/5.4$)。Greenland, Robins 和 Pearl (1999) 所著论文讨论了不具伸缩性和互换性缺失的关系。

知识点 4.1：治疗组中因果效应的计算方法（原书第 46 页）

我们在有界互换性 $Y^a \perp\!\!\!\perp A|L$ 的假设下计算了 $a=0$ 与 $a=1$ 的因果效应均值。计算治疗组中的因果效应均值只需要部分互换性 $Y^{a=0} \perp\!\!\!\perp A|L$ 成立。换句话说, 非治疗组在有治疗时的反事实结局, 和现实治疗组的观测结局无关。同理, 计算非治疗组中的因果效应均值只需要部分互换性 $Y^{a=1} \perp\!\!\!\perp A|L$ 成立。

我们现在来给出如何在部分互换性假设下计算通过标准化和逆概率加权计算反事实均值 $E[Y^a | A = a']$ 。

- 标准化: $E[Y^a | A = a'] = \sum_l E[Y | A = a, L = l] \Pr[L = l | A = a']$ 。Miettinen (1972) 以及 Greenland 和 Rothman (2008) 讨论了如何用标准化计算风险比。

- 逆概率加权: $E[Y^a | A = a'] = \frac{E\left[\frac{I(A=a)Y}{f(A|L)} \Pr[A = a' | L]\right]}{E\left[\frac{I(A=a)}{f(A|L)} \Pr[A = a' | L]\right]}$, 其中权重 $\frac{\Pr[A = a' | L]}{f(A|L)}$ 。

对于二分变量 A , 参见 Sato 和 Matsuyama (2003) 所著论文。更多细节参见 Hernan 和 Robins (2006) 所著论文。

知识点 4.2: 汇总不同分层中的效应量度 (原书第 51 页)

迄今, 我们主要关注的是概念上而非统计上的因果推断, 并假设我们是对整个人群而不是其中的一个样本进行研究。因此我们会说“计算”因果效应, 而不是“估计”因果效应。在现实世界, 我们基本不可能计算整个人群的因果效应, 我们只能通过样本去估计它。因而, 这就需要为我们的估计值计算一个合理的置信区间。

面对不同分层的效应估计, 一个通常的做法是将每一层的效应估计汇总成一个总的效应估计, 从而减小估计值的变异性。这背后的思路很简单: 如果每一层的效应估计都一样的话 (即没有效应修饰), 那把它们合起来我们就能得到一个更精确的效应估计值。有许多不同的方法可以用来汇总每一层的效应估计, 比如 Woolf 方法、Mantel-Haenszel 方法、以及最大似然法。有时候会给每一分层一个权重, 通过加权的方式降低总估计值的变异性。也可以通过在回归模型中放入所有协变量 L 间的乘积项而得到一个总的效应估计值。但是在这个回归模型中, 不能放入治疗 A 和协变量 L 的乘积项, 也即这个模型对于 L 是饱和的 (参见第十一章)。

将不同分层的效应估计汇总是想得到一个更窄的置信区间, 但是一个总的效应估计依然是一个条件效应估计。在我们心脏移植的例子中, 对结局 Z 而言, 汇总的风险比是 0.88 (使用

Mantel-Haenszel 方法)。这一结果只有当分层风险比(2 和 0.5)衡量的是同一因果效应的时候才有意义。比如, 在这两个分层中, 真实风险比是 0.9, 但因为每一层的样本量过小, 我们分别得到了 2 和 0.5。在这种情况下, 利用 Mantel-Haenszel 方法得到一个汇总的风险比就是有意义的, 并且更接近真实值。但是, 如果每一分层中的真实风险比分别是 0.5 和 2, 那汇总这两个风险比就没有太大的意义。在实践中, 很难区分每一层中的风险比到底是有实质差别, 还是因为抽样变异性导致了误差。分层越细, 随机误差也就越大。

知识点 4.3: 边缘风险比和条件风险比 (原书第 52 页)

如果我们已经知道每一分层中的条件风险比 $\Pr[Y^{a=1} = 1 | L = l] / \Pr[Y^{a=0} = 1 | L = l]$, 现在我們想知道, 在什么情况下, 边缘风险比 $\Pr[Y^{a=1} = 1] / \Pr[Y^{a=0} = 1]$ 会小于 1。注意到

$$\Pr[Y^{a=1} = 1] / \Pr[Y^{a=0} = 1] = \sum_l \left\{ \Pr[Y^{a=1} = 1 | L = l] / \Pr[Y^{a=0} = 1 | L = l] \right\} w(l), \text{ 其中}$$

$$w(l) = \left\{ \Pr[Y^{a=0} = 1 | L = l] \Pr[L = l] \right\} / \Pr[Y^{a=0} = 1] \text{ 且 } \sum_l w(l) = 1。 \text{ 经过一些数学变换, 我们能}$$

得到 $\Pr[Y^{a=1} = 1] / \Pr[Y^{a=0} = 1] < 1$ 成立的条件。

比如, 在我们正文的例子中, $L = 1$ 时 $\Pr[Y^{a=1} = 1 | L = l] / \Pr[Y^{a=0} = 1 | L = l] = 0.5$, $L = 0$ 时, $\Pr[Y^{a=1} = 1 | L = l] / \Pr[Y^{a=0} = 1 | L = l] = 2$ 。因此, 当且仅当

$$\frac{\Pr[Y^{a=0} = 1 | L = 1]}{\Pr[Y^{a=0} = 1 | L = 0]} > \frac{2 \Pr[L = 0]}{\Pr[L = 1]} \text{ 时, 边缘风险比小于 1。}$$

第四章图表

Table 4.1

	V	Y^0	Y^1
Rheia	1	0	1
Demeter	1	0	0
Hestia	1	0	0
Hera	1	0	0
Artemis	1	1	1
Leto	1	0	1
Athena	1	1	1
Aphrodite	1	0	1
Persephone	1	1	1
Hebe	1	1	0
Kronos	0	1	0
Hades	0	0	0
Poseidon	0	1	0
Zeus	0	0	1
Apollo	0	1	0
Ares	0	1	1
Hephaestus	0	0	1
Cyclope	0	0	1
Hermes	0	1	0
Dionysus	0	1	0

Table 4.2

Stratum $V = 0$	L	A	Y
Cybele	0	0	0
Saturn	0	0	1
Ceres	0	0	0
Pluto	0	0	0
Vesta	0	1	0
Neptune	0	1	0
Juno	0	1	1
Jupiter	0	1	1
Diana	1	0	0
Phoebus	1	0	1
Latona	1	0	0
Mars	1	1	1
Minerva	1	1	1
Vulcan	1	1	1
Venus	1	1	1
Seneca	1	1	1
Proserpina	1	1	1
Mercury	1	1	0
Juventas	1	1	0
Bacchus	1	1	0

Table 4.3

	<i>L</i>	<i>A</i>	<i>Z</i>
Rheia	0	0	0
Kronos	0	0	1
Demeter	0	0	0
Hades	0	0	0
Hestia	0	1	0
Poseidon	0	1	0
Hera	0	1	1
Zeus	0	1	1
Artemis	1	0	1
Apollo	1	0	1
Leto	1	0	0
Ares	1	1	1
Athena	1	1	1
Hephaestus	1	1	1
Aphrodite	1	1	0
Cyclope	1	1	0
Persephone	1	1	0
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0

Table 4.4

	<i>V</i>	<i>A</i>	<i>Y</i>
Rheia	1	0	0
Demeter	1	0	0
Hestia	1	0	0
Hera	1	0	0
Artemis	1	0	1
Leto	1	1	0
Athena	1	1	1
Aphrodite	1	1	1
Persephone	1	1	0
Hebe	1	1	1
Kronos	0	0	0
Hades	0	0	0
Poseidon	0	0	1
Zeus	0	0	1
Apollo	0	0	0
Ares	0	1	1
Hephaestus	0	1	1
Cyclope	0	1	1
Hermes	0	1	0
Dionysus	0	1	1

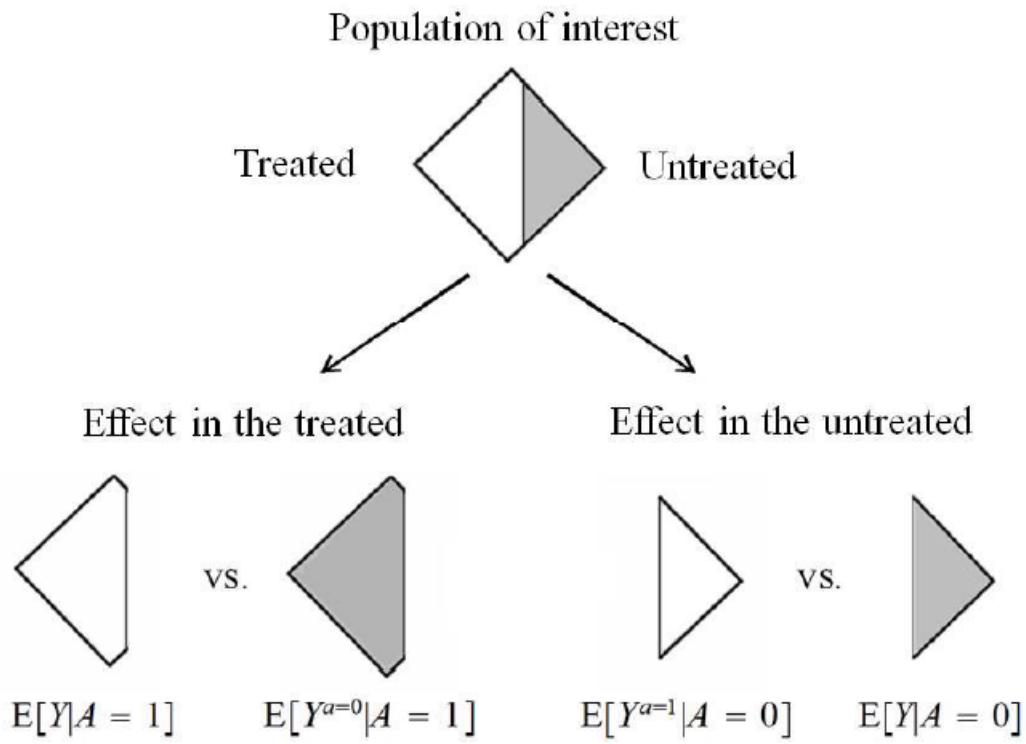


Figure 4.1

第五章 交互作用

55 再思考一下这个因果性的问题: “一个人抬头向天上看会影响其他行人也抬头向天上看吗?” 我们迄今的讨论仅局限于一个单一动作(抬头看)在整个人群或子群体中的因果效应。但是, 有一些问题, 则是关于两个或多个同时动作的因果效应。除了你是否抬头向天上看之外, 我们再增加一项: 你是否赤身裸体地站在大街上。现在我们的问题变成: 如果你穿着衣服站在大街上, 你抬头向天上看会影响其他行人也抬头向天上看吗? 如果你赤身裸体地站在大街上呢? 如果这两个情境中的因果效应有差别, 我们会说这两项动作(抬头看和赤身裸体)相互交互, 从而对结局产生影响。

当多个干预措施存在的时候, 识别是否有交互作用能让我们找出最有效的干预策略。因而, 了解交互作用的基本概念对于因果推断大有帮助。本章, 我们会介绍两个治疗(或干预)间的交互作用, 我们会在反事实结局的框架下, 以及充分成因模型的框架下讨论交互作用。

5.1 交互作用需要联合干预

假设在我们的心脏移植例子中, 在分配心脏移植治疗之前, 每个被试被分为服用维生素组($E = 1$)和不服用维生素组($E = 0$)。于是我们能将被试分成四个组: 有维生素有移植组($E = 1, A = 1$)、有维生素无移植组($E = 1, A = 0$)、无维生素有移植组($E = 0, A = 1$)、无维生素无移植组($E = 0, A = 0$)。于是有四种治疗组合, 每个被试有四个反事实结局:

$Y^{a=1,e=1}$ 、 $Y^{a=1,e=0}$ 、 $Y^{a=0,e=1}$ 、 $Y^{a=0,e=0}$ 。与之前的定义相同, 个体的反事实结局 $Y^{a,e}$ 表示这个个体在 A 和 E 的取值分别为 a 和 e 时我们观测到的结局。我们将这种涉及多种治疗(或干预)的措施称为联合干预。

(当观测到的 E 取值为 e 时, 单一干预 A 所对应的反事实结局 Y^a 就是联合干预的反事实结局 $Y^{a,e}$, 即 $Y^a = Y^{a,E}$ 。对于联合干预, 一致性的定义为 $Y = Y^A = Y^{A,E}$ 。参见知识点6.2)

我们现在可以在反事实结局的框架下对交互作用下一个定义: 对于两种不同的治疗(或干预) A 和 E , 如果 A 对结局 Y 的因果效应在 E 分别为0和1时而不同, 则 A 和 E 之间存在交互作用。比如, 在我们的例子中, 如果心脏移植对死亡率的因果效应在服用维生素组和不服用维生素组中不同, 则心脏移植 A 和维生素 E 之间有交互作用。

用风险差来度量因果效应时, 如果在人群中

56 $\Pr[Y^{a=1,e=1} = 1] - \Pr[Y^{a=0,e=1} = 1] \neq \Pr[Y^{a=1,e=0} = 1] - \Pr[Y^{a=0,e=0} = 1]$, 则我们说 A 和 E 在加法尺度上有交互作用。上述不等式经过简单变化后有

$\Pr[Y^{a=1,e=1} = 1] - \Pr[Y^{a=1,e=0} = 1] \neq \Pr[Y^{a=0,e=1} = 1] - \Pr[Y^{a=0,e=0} = 1]$ 。这两个不等式表明 A 和 E 在交互作用的定义中有相同的地位。

让我们回顾一下效应修饰的定义: 如果 A 对 Y 的因果效应均值在 V 的不同取值下不尽相同, 则 V 是一个效应修饰因子。注意, 当提及效应修饰时, 我们说的是对 A 的因果效应的修饰, 而不是 V 的因果效应。比如, 在上一章, 我们用表 4.1 的数据, 讨论了性别的效应修饰作用, 但我们从未讨论过性别对死亡风险的因果效应。因而, 当我们说 V 修饰了 A 的因果效应时, 我们并不是将 V 和 A 放在同等的地位进行考虑, 而只有 A 才是我们所关注的干预 (或治疗) 变量。换句话说, 效应修饰的定义只涉及反事实结局 Y^a , 与 $Y^{a,v}$ 无关。与之相比, 在交互作用的定义中, A 和 E 具有同等的地位, 就如同上一段的两个等价不等式一样。交互作用的定义涉及的是两个干预 (或治疗) A 和 E 的联合因果效应, 因而涉及的是联合干预下的反事实结局 $Y^{a,e}$ 。

5.2 识别交互作用

在前几章, 我们讨论了在整个人群或其中一个子群体中, 识别治疗 (或干预) A 对结局 Y 的因果效应所需要的三个前提条件, 分别是互换性、正数性和一致性。因为交互作用涉及两个 (或多个) 干预 (或治疗) A 和 E 的联合效应, 因而识别交互作用需要互换性、正数性和一致性对这两个干预 (或治疗) 都成立。

假设研究者是随机、无条件进行维生素 E 分组的, 则正数性和一致性成立, 且服用组 $E = 1$ 和未服用组 $E = 0$ 可互换。也就是说, 如果所有被试都被分配到治疗组 $A = 1$ 和服用组 $E = 1$, 我们观测到的结局, 会和如果现实中所有服用组 $E = 1$ 被试都被分配到治疗组 $A = 1$ 时的结局一样。即边缘风险 $\Pr[Y^{a=1,e=1} = 1]$ 等于条件风险 $\Pr[Y^{a=1} = 1 | E = 1]$ 。因而, 我们可以将加法尺度上 A 和 E 的交互作用写作:

$\Pr[Y^{a=1} = 1 | E = 1] - \Pr[Y^{a=0} = 1 | E = 1] \neq \Pr[Y^{a=1} = 1 | E = 0] - \Pr[Y^{a=0} = 1 | E = 0]$, 这和加法尺度上 E 对 A 的效应修饰定义一样。换句话说, 如果治疗 E 是随机分配的, 则交互作用和效应修饰两个概念重合。因而第四章用来识别 V 对 A 效应修饰的方法, 也可以用来识别 A 和 E 之间的交互作用, 仅需把 V 换成 E 即可。

就算治疗 (或干预) E 不是由研究者分配的, 我们依然需要计算四个不同的边缘风险 $\Pr[Y^{a,e} = 1]$ 来检验 A 和 E 之间的交互作用。此时, 在一些前提假设下, 我们可以用标准化和逆概率加权方法计算这些反事实结局。让我们换一个角度思考这个问题: 不再把 A 和 E 看作取值是

1 或 0 的两个不同变量, 而是将其看作一个有四种不同取值的单一变量 AE (取值分别为 11, 01, 10, 00)。此时, 交互作用的可识别性条件, 就和我们前几章讨论的单一治疗变量的可识别性条件一样。因而, 我们可以在同样的可识别性条件下, 使用同样的方法。唯一的区别就是, 此时我们的治疗 (或干预) 变量取值更多, 对应的反事实结局也就更多。

58 有时研究者会假设有界互换性对 A 成立但是对 E 不成立, 比如在根据 E 进行治疗分组的随机试验中。因为此时不需要任何涉及 E 的假设就能计算 A 在 E 的不同取值下的因果效应, 所以研究者可以检验 E 的效应修饰作用。在我们上一章的讨论中, 我们就没有对国籍 V (注意不是 E) 做任何有关互换性、正数性和一致性的假设。在 4.2 小节, 我们认为 V 是一个修饰因子替代物, 因而不会对结局产生因果性的关系, 也就不会和 A 有交互作用——没有因果效应, 就没有交互作用。但 V 是 A 对 Y 效应的修饰因子, 因为 V 和一个未知的变量相关 (也即 V 是这个未知变量的替代物), 而这个未知变量对 Y 有因果效应并且和 A 有交互作用。因而, 存在这一个变量 V , 其和 A 没有交互作用, 但是修饰了 A 的因果效应。

(A 和 E 之间存在交互作用但是没有效应修饰作用, 这种情况也是存在的, 不过十分罕见, 是因为这种情况需要 A 具备双重效应并且这两种效应能完全相互抵消 (VanderWeele, 2009))

总而言之, 检验 A 和 E 之间的交互作用, 需要互换性、正数性和一致性对于联合治疗 (或干预) AE 成立, 这个联合治疗 (或干预) 有四种不同取值, 分别是 00, 01, 10, 11。此时, 我们就可以用标准化或逆概率加权来估计这两个治疗 (或干预) 的联合效应, 并衡量它们间的交互作用。在本书第三部分我们会讨论到, 如果这两个治疗 (或干预) 出现在不同的时间段, 那某些假设是不必要的。在第一部分的剩下部分 (除去本章), 以及第二部分的大多部分, 我们都将只关注单一治疗 (或干预) A 的因果效应。

在第一章, 我们讨论了命定的和非命定的反事实结局。为了方便, 我们之前所有的讨论只涉及命定的反事实结局。然而, 在我们迄今所有关于因果效应和交互作用的讨论中, 都没有要求反事实结局必须是命定的。但在下一小节, 我们的所有讨论只有在反事实结局是命定的情况下才成立。接下来, 我们假设所有的治疗 (或干预) 和结局都是二分的。

5.3 反事实回应类型和交互作用

每个个体可以根据他们命定的反事实结局分成不同的类型。以表 4.1 (和表 1.1 相同) 数据为例, 其中被试可以被分成四种不同类型: 注定型, 指不管有没有接受治疗, 都会发生我们关注的结局 (表中的 Artemis, Athena, Persephone, Ares); 免疫型, 指不管有没有接受治疗, 都不会发生我们关注的结局 (表中的 Demeter, Hestia, Hera, Hades); 受益型, 指接受了治疗,

就不会发生我们关注的结局, 而不接受就会发生(表中的Hebe, Kronos, Poseidon, Apollo, Hermes, Dyonisus); 受害型, 指接受了治疗, 才会发生我们关注的结局, 而不接受就不会发生(表中的Rheia, Leto, Aphrodite, Zeus, Hephaestus, Cyclope)。我们将每个个体在不同治疗下反事实结局的组合称作回应类型。表 5.1 给出了四种不同的回应类型。

当涉及两种不同的二分治疗 A 和 E 时, 联合治疗共有 4 种不同取值, 因而每个被试有 4 种情形的反事实结局, 也因此总共将有 16 种不同的回应类型。表 5.2 给出了两种治疗下 16 种不同回应类型。本小节将讨论两种治疗情况下, 回应类型和交互作用之间的关系。

(Miettinen (1982) 首次描述了两种治疗下的 16 种不同回应类型)

表 5.2 的第 1 种类型, 在每种情形下反事实结局都是 1, 即

$Y^{a=1,e=1} = Y^{a=1,e=0} = Y^{a=0,e=1} = Y^{a=0,e=0} = 1$ 。换句话说, 治疗 A 或 E 对这个个体的结局都没有影响。

第 16 种类型, 所有的反事实结局都是 0, 即 $Y^{a=1,e=1} = Y^{a=1,e=0} = Y^{a=0,e=1} = Y^{a=0,e=0} = 0$, 同样, 治疗 A 或 E 对这个个体的结局都没有影响。如果一个人群中的所有个体都是第 1 或第 16 类型, 则不管 A 或 E 对结局 Y 都没有因果效应, 极端因果零假设对联合治疗 AE 为真。

在第 4 种类型中, 个体只要接受治疗 E , 就会发生结局 Y , 无论其是否接受治疗 A , 即

$Y^{a=1,e=1} = Y^{a=0,e=1} = 1$ 且 $Y^{a=1,e=0} = Y^{a=0,e=0} = 0$ 。在第 13 种类型中, 个体只要不接受治疗 E , 就会发生结局 Y , 无论其是否接受治疗 A 即 $Y^{a=1,e=1} = Y^{a=0,e=1} = 0$ 且 $Y^{a=1,e=0} = Y^{a=0,e=0} = 1$ 。在第 6 种类型中, 个体只要接受治疗 A , 就会发生结局 Y , 无论其是否接受治疗 E , 即 $Y^{a=1,e=1} = Y^{a=1,e=0} = 1$ 且 $Y^{a=0,e=1} = Y^{a=0,e=0} = 0$ 。在第 11 种类型中, 个体只要不接受治疗 A , 就会发生结局 Y , 无论其是否接受治疗 E , 即 $Y^{a=1,e=1} = Y^{a=1,e=0} = 0$ 且 $Y^{a=0,e=1} = Y^{a=0,e=0} = 1$ 。

在表 5.2 的 16 种类型中, 我们已经发现其中 6 种(第 1、4、6、11、13、16)有一个共同特点: 如果一个个体属于这 6 个类型中的一种, 那对他来说, 治疗 A 对 Y 的因果效应不管 E 的取值如何都是一样的, 同时, 治疗 E 对 Y 的因果效应不管 A 的取值如何也是一样的。如果一个人群是由这 6 种类型的个体构成的, 那有 E 时 A 的因果效应, 和没有 E 时 A 的因果效应就是一样的, 用风险差表达就是 $\Pr[Y^{a=1,e=1} = 1] - \Pr[Y^{a=0,e=1} = 1] \neq \Pr[Y^{a=1,e=0} = 1] - \Pr[Y^{a=0,e=0} = 1]$ 。也就是说, 如果一个人群中的所有个体都是第 1、4、6、11、13 和 16 类型的, 那么 A 和 E 之间在加法尺度上就不存在交互作用。

(Greenland 和 Poole (1988) 注意到 Miettinen 的回应类型分类会因 A 和 E 的编码方式不同而不一样(比如将 0 和 1 调换一下)。因此, 他们将表 5.2 的 16 种类型划分成了 3 类, 从而回应类型不再随编码方式的改变而改变)

如果 A 和 E 之间存在加法交互作用, 那对一个人群中的某些个体而言。我们就不能在不知道 E 的情况下识别 $A = a$ 的反事实结局, 反之亦然。也就是说, 如果存在交互作用, 那人群中的某些个体属于以下三类中的一种:

1. 只会在 4 种不同治疗组合中的一种出现我们关注的结局 (表 5.2 的第 8、12、14 和 15)。
2. 会在 4 种不同治疗组合中的两种出现我们关注的结局, 其中一个组合的两种治疗, 和另一个组合中对应的治疗取值正好相反 (表 5.2 的第 7 和 10)。
3. 会在 4 种不同治疗组合中的三种出现我们关注的结局 (表 5.2 的第 2、3、5 和 9)。

60

另一方面, 如果在加法尺度上 A 和 E 之间不存在交互作用, 则意味着人群中没有一个个体属于上述三种类型, 或者在加法尺度上因为方向相反正好完全抵消。如果人群中第 7 和第 10 类型的个体占比一样, 或者第 8 和第 12 类型的个体占比一样, 就会出现这种抵消。

(对交互作用存在时各效应的抵消情况, 参见 Greenland, Lash 和 Rothman (2008) 所著论文)

上述讨论用反事实回应类型 (参见精讲点 5.1) 解释了“交互作用”的含义。现在我们将借助另一个工具来解释交互作用背后的因果机理。

5.4 充分成因

还是以我们心脏移植的研究为例。只考虑一种治疗 A 时, 我们知道有些人接受了治疗会死, 有些人不接受治疗会死, 有些人不管接不接受治疗都会死, 而还有一些人不管接不接受治疗都不会死。这林林总总的回应类型反映了治疗 A 不是唯一决定结局 Y 的变量。

只考虑在现实中实际接受治疗的人, 他们中有一些人死了, 但有些人没死, 说明仅有治疗不足以充分说明是否会发生结局, 还有一些其他的背景因素也会影响结局。在一个很简化的例子中, 假设只有对麻药过敏 ($U_1 = 1$) 的人在接受心脏移植 ($A = 1$) 后才会死亡, 则我们将治疗 ($A = 1$) 和对麻药过敏 ($U_1 = 1$) 称作结局 Y 的最小充分成因。

61

现在让我们考虑在现实中实际未接受治疗的人, 同样, 他们中有一些人死了, 但有些人没死, 说明仅靠没有治疗不足以充分说明是否会发生结局, 还有一些其他的背景因素也会影响结局。在一个很简化的例子中, 假设只有射血分数小于 20% ($U_2 = 1$) 的人在未接受心脏移植 ($A = 0$) 的情况下才会死亡, 则未接受治疗 ($A = 0$) 和低射血分数 ($U_2 = 1$) 是结局 Y 的另一个最小充分成因。

当然, 也有被试在没有 U_1 和 U_2 的情况下, 不管接不接受心脏移植治疗, 都会去世。这些“注定型”个体也有他们自己的最小充分成因。比如, 所有这些个体在研究开始前都有胰腺癌 ($U_0=1$), 因而不管他们接不接受治疗, 他们都会去世。因而胰腺癌 ($U_0=1$) 就是结局 Y 的另一个最小充分成因。

(我们不能对这些背景因素 U 进行干预, 它们也不会受到治疗 A 的影响)

至此, 我们描述了结局的 3 个充分成因: $A=1$ 且 $U_1=1$, $A=0$ 且 $U_2=1$, $U_0=1$ 。每一个充分成因都有一个或多个成分, 比如第一个中的 $A=1$ 和 $U_1=1$ 。图 5.1 展示了每一种充分成因, 其中一个整圆表示一个充分成因, 圆内的每一部分表示一个成分。充分成因经常用来表示一个结局的充分条件和这个条件的所有成分。

62 用图来表示充分成因可以形象化地展示效应修饰的一个重要推论: 如同第四章所讨论的一样, 治疗 A 的因果效应大小取决于人群中修饰因子的分布状态。让我们来想象两种假想情境。在第一种情境中, 人群中只有 1% 的人有 $U_1=1$ 。在第二种情境中, 人群中有 10% 的人有 $U_1=1$ 。而 U_0 和 U_2 在两种情境中分布相同。如果我们在两种情境中都开展一次心脏移植的试验, 我们会发现第一种情境里心脏移植 A 对死亡 Y 的因果效应均值, 要大于第二种情境, 这是因为第二种情境里有更多的人会在接受治疗后去世。在这个例子中, 3 个充分成因中的一个, $A=1$ 且 $U_1=1$, 在第二种情境里的人数比例是第一种里的 10 倍, 而其余 2 个在两种情境中的出现频率是一样的。

充分成因的概念可以用图像的方法来表示, 而这个图像也可以用来解释交互作用。我们先考虑有两种治疗 A 和 E 时的情况, 于是我们有以下 9 种可能的充分成因:

1. 有 A 时 (不管有没有 E)
2. 无 A 时 (不管有没有 E)
3. 有 E 时 (不管有没有 A)
4. 无 E 时 (不管有没有 A)
5. 同时有 A 和 E 时
6. 有 A 无 E 时
7. 有 E 无 A 时
8. 同时无 A 和 E 时
9. 通过其他途径 (不管有没有 A 或 E)

我们给每一种充分成因补充一个背景因素, 记作 U_0 至 U_8 。图 5.2 展示了两种治疗 A 和 E 下的这 9 种不同充分成因。在下一节我们会进一步用图解释交互作用。

(Greenland 和 Poole (1988) 首次提出了这 9 中充分成因)

- 63 不是每一个涉及一个二分结局和两种不同治疗的情境都有 9 种不同的充分成因。比如在我们的例子中, 可能出现只要服用维生素 $E = 1$, 那不管有没有接受心脏移植手术, 都不会死的情况。此时, 图 5.2 中 $E = 1$ 的 3 个充分成因就不再成立。这 3 个充分成因表示, 在 $U_3 = 1$ 时, 如果被试服用维生素 ($E = 1$) 就会死亡, 也就是说, 如果他们不服用维生素 ($E = 0$), 那他们就不会死亡。

(表示充分成因的图像有时也被称作“因果派”)

5.5 充分成因的交互作用

“ A 和 E 的交互作用”这一说法暗示这两种不同的治疗在机理上共同协作从而造成了结局。然而, 在反事实框架下, 交互作用的定义不涉及任何机理相关的内容(参见精讲点 5.3)。在我们心脏移植的例子中, 如果心脏移植 A 的因果效应在服用维生素 E 和不服用的被试中不一样的话, 那 A 和 E 之间就有交互作用。也就是说, 交互作用是通过反事实结局的对比而定义的, 我们也就能通过随机试验来检验交互作用, 而无需知道背后物理学、化学、生物学、社会学等等各方面的机理。

这一小节我们将介绍交互作用的另一种定义。这一定义不依赖于反事实结局的对比, 而是建立在充分成因之上, 更能用于解释交互作用的机理。

- 64 只有当 A 和 E 同时出现在一个充分成因中, A 和 E 之间才会有交互作用。比如, 一个 $U_5 = 1$ 的个体, 只有同时服用维生素 ($E = 1$) 和接受心脏移植 ($A = 1$) 才死亡, 但只接受两种治疗中的一种则不会。因而, 只要存在一个 $U_5 = 1$ 的个体, A 和 E 之间就存在交互作用。对这样的个体来说, 其反事实结局 $Y^{a=1,e=1} = 1$, 且 $Y^{a=0,e=1} = Y^{a=1,e=0} = 0$ 。

充分成因中的交互作用可以分成协同和拮抗两种。如果在一个充分成因中, $A = 1$ 且 $E = 1$, 则 A 和 E 之间的交互作用是协同的; 而如果 $A = 1$ 且 $E = 0$ (或 $A = 0$ 且 $E = 1$), 则是拮抗的。当然, 也可以认为 A 和 E 之间的协同是 A 和非 E (或非 A 和 E) 之间的拮抗。

(Rothman (1976) 在充分成因的框架下描述了协同和拮抗的概念)

- 与反事实框架下交互作用的定义不同, 充分成因中交互作用的定义需要涉及到 A 和 E 的机理, 也即我们需要相关知识去识别是否存在交互作用。但有时候也会出现这种情况: 就算我们对充分成因和其各成分所知不多, 我们依然发现有交互作用的存在。尤其是当精讲点 5.1 中的不等式成立时, 则存在 A 和 E 之间的协同交互作用。也即我们能实证地检验 A 和 E 的协同交互作用,

但我们不知道 A 和 E 是如何协同产生这个结局的。这一特点和用反事实回应类型定义的交互作用相当（参见精讲点 5.2），同时，这也是因为精讲点 5.1 中的不等式是协同交互作用的充分条件，而不是必要条件。

5.6 反事实框架还是充分成因框架？

充分成因框架和反事实框架能够处理不同的问题。充分成因模型考虑了事件、动作、状态等不同因素，这些不同的因素共同作用从而造成了我们关注的结局。这一模型能够给一个特定的效应作出解释。它能够解决诸如“哪些不同的条件能导致同样的结局？”等问题。而反事实模型则主要关注一个特定的治疗（或干预），并对这个治疗（或干预）的不同因果效应给出解释。反事实框架能够解决诸如“如果我们对一因素进行干预会导致什么样的后果？”等问题。与充分成因模型不同，反事实模型不需要任何涉及作用机理的知识。

（事实上，早在 18 世纪，休谟就提出用反事实框架解决因果性问题）

反事实框架解决的问题是“会是什么？”，而充分成因框架解决的问题是“怎么会发生？”。这本书主要讲述了在什么条件下可以用什么样的方式去估计一项干预的因果效应均值，因而反事实框架是更贴切的选择。充分成因框架有助于我们思考背后的机理。在因果推断中，充分成因模型能帮助我们理解因果效应的大小依赖于背景因素（即修饰因子）的分布，以及效应修饰、交互作用以及协同作用之间的关系，因而它在因果推断中依然有一席之地。

虽然充分成因框架是有益于我们的阐释，但是尚未有方法将其整合到数据分析当中。就其经典形式而言，充分成因框架是命定的，只能用于离散型治疗（或干预）及结局。这一缺陷使其不能应用于连续型变量。近来有研究将充分成因模型扩展到随机过程与有序变量当中，这一扩展也许能拓宽充分成因模型的应用。然而，即使我们能将充分成因模型进一步扩展，在现实中，我们也很少有足够的数据来检验充分成因模型中的细致区别。

（VanderWeele (2020b) 扩展到了 3 层治疗。VanderWeele 和 Robins (2012) 讨论了随机过程中反事实框架和充分成因框架之间的关系）

也许反事实框架是估计因果效应的最佳框架选择。其他可能的框架——诸如因果图、决策论——实质上和反事实框架是等价的。下一章我们将讨论这个问题。

第五章精讲点和知识点

精讲点 5.1：反事实类型与交互作用（原书第 61 页）

根据每个个体的不同反事实类型进行分类有助于我们理解交互作用。比如, 我们只想知道有没有个体在同时接受两种治疗即 $A = 1$ 和 $E = 1$ 时会发生结局, 但在只接受一种治疗时不发生结局, 即 $Y^{a=1,e=1} = 1$ 且 $Y^{a=1,e=0} = Y^{a=0,e=1} = 0$ 。VanderWeele 和 Robins (2007a, 2008) 将充分成因扩展到两种和三种不同治疗的情形, 同时论述了协同交互作用存在的条件。以下不等式是协同交互作用存在的充分条件:

$$\Pr[Y^{a=1,e=1} = 1] - (\Pr[Y^{a=0,e=1} = 1] + \Pr[Y^{a=1,e=0} = 1]) > 0$$

也即, 在一个 A 和 E 随机分配的试验中, 只需要计算上述不等式中的三个反事实结局, 我们就能实证地检验有没有交互作用存在。

上述不等式只是一个充分条件, 而不是必要条件。因而很多时候, 就算有交互作用存在, 上述不等式依然可能不成立。当因果效应是单调的时候 (参见知识点 5.2), 我们可以有一个协同交互作用的稍弱充分条件:

$$\Pr[Y^{a=1,e=1} = 1] - \Pr[Y^{a=0,e=1} = 1] > (\Pr[Y^{a=1,e=0} = 1] + \Pr[Y^{a=0,e=0} = 1])$$

换句话说, 如果 A 和 E 的效应是单调的, 那这个交互作用是超加性的 (参见知识点 5.2), 并且人群中有表 5.2 中第 8 种类型的个体 (单调性排除了第 7 种类型)。这一单调性下协同交互作用的充分条件在 Greenland, Lash 和 Rothman 合著《现代流行病学》 (Modern Epidemiology) 中有讲述。

对于大多数研究问题, 我们更关心是否存在第 8 种类型的个体。这种类型的交互作用也被称为组合上位性 (compositional epistasis)。VanderWeele (2010a) 讨论了组合上位性的实证检验。

精讲点 5.2: 从反事实到充分成因 (原书第 64 页)

反事实回应类型与充分成因也存在对应关系。假设在一个治疗和结局都是二分变量的情境中, 已知的充分成因需要背景因素 U_0 、 U_1 或 U_2 中的至少一个 (图 5.2 的前三个)。下表给出了这个情境中反事实回应类型和充分成因的对应关系:

反事实类型	$Y^{a=0}$	$Y^{a=1}$	所需成分
注定型	1	1	$U_0 = 1$ 或 $\{U_1 = 1 \text{ 且 } U_2 = 1\}$
受益型	1	0	$U_0 = 0$ 且 $U_1 = 0$ 且 $U_2 = 1$
受害型	0	1	$U_0 = 0$ 且 $U_1 = 1$ 且 $U_2 = 0$

$$\begin{array}{cccc} \text{免疫型} & 0 & 0 & U_0 = 0 \text{ 且 } U_1 = 0 \text{ 且 } U_2 = 0 \end{array}$$

每种成分组合都对应一种、且只对应一种反事实回应类型。然而一个反事实回应类型却可以对应多种组合。比如, 注定型就对应包含 $U_0 = 1$ 的所有组合, 或者包含 $\{U_1 = 1 \text{ 且 } U_2 = 1\}$ 的所有组合。

充分成因也可以用来接收互换性 $Y^a \perp\!\!\!\perp A$ 。在一个治疗和结局都是二分变量的情境中, 互换性意味着 $\Pr[Y^{a=1} = 1 | A = 1] = \Pr[Y^{a=1} = 1 | A = 0]$ 且 $\Pr[Y^{a=0} = 1 | A = 1] = \Pr[Y^{a=0} = 1 | A = 0]$ 。

如果一个个体接受治疗会产生我们关注的结局, 那么他是注定型或者受害型, 也即 $U_0 = 1$ 或 $U_1 = 1$; 如果一个个体不接受治疗会产生我们关注的结局, 那么他是注定型或者受益型, 也即 $U_0 = 1$ 或 $U_2 = 1$ 。因而, 互换性成立的一个充分条件是:

$$\begin{aligned} \Pr[U_0 = 1 \text{ or } U_1 = 1 | A = 1] &= \Pr[U_0 = 1 \text{ or } U_1 = 1 | A = 0] \text{ 且} \\ \Pr[U_0 = 1 \text{ or } U_2 = 1 | A = 1] &= \Pr[U_0 = 1 \text{ or } U_2 = 1 | A = 0]。 \end{aligned}$$

更多内容, 请参考 Greenland 和 Brumback (2002), Flanders (2006), 以及 VanderWeele 和 Hernan (2006) 等人所著论文。

精讲点 5.3: 生物学中的交互作用 (原书第 65 页)

在流行病学的讨论中, 充分成因中的交互作用通常指的是生物学上的交互作用 (Rothman, 1980)。选择充分成因这一名称是想表明, 在生物医学实践中, 有的生物机制需要通过两个不同的治疗 (或干预) A 和 E 来实现从而产生我们关注的结局。然而, VanderWeele 和 Robins (2007a) 给出了以下反例。

假设 A 和 E 是同一基因的两个等位基因。两个等位基因上存在有害突变 ($A = 1$ 且 $E = 1$) 的个体不能合成某种重要蛋白质, 从而会在出生一周内死亡。如果两个等位基因上没有突变 ($A = 0$ 且 $E = 0$) 或只有一个等位基因有突变 ($A = 1$ 且 $E = 0$, 或 $A = 0$ 且 $E = 1$), 将会一切正常。因而, 我们会说等位基因 A 和 E 之间有协同作用, 因为 $A = 1$ 且 $E = 1$ 是死亡的一个充分成因。也就是说, 两个等位基因共同造成了结局的发生。然而, 也会有人争论说, 这两个等位基因各自活动, 在生物学上并没有交互作用。

精讲点 5.4: 再论归因比例 (原书第 66 页)

精讲点 3.4 将治疗 A 的超出比例定义为人群中能归因于治疗 A 的病例比例, 同时在精讲点 3.4 给出的例子中, 超出比例是 75%。也就是说, 如果所有人接受的治疗都是 $a=0$, 那就有 75% 的病例不会发生。我们现在再加入第二个治疗 E 。假设 E 的超出比例是 50%, 是不是意味着对 A 和 E 的联合干预, 能阻止 125% 的病例? 当然不是这样。

很明显, 对一项治疗 (或干预) 而言, 超出比例不能超过 100%, 而一项联合治疗 (或干预) 的超出比例也不能超过 100%。我们永远不可能阻止超过 100% 的病例发生。那为什么把两项不同治疗 (或干预) 的单独超出比例相加, 会超过 100% 呢? 充分成因框架可以解答这个问题。

依然以图 5.2 为例。假设宙斯的背景因素 $U_5=1$, 并且同时接受了治疗 $A=1$ 和 $E=1$ 。这种情况下, 撤掉任一治疗, 宙斯都将不再是一个病例。也即, 宙斯是 A 的超出比例 75% 中的一部分, 也是 E 的超出比例 50% 中的一部分。因而, A 和 E 各自的超出比例加起来超过 100% 也就不足为奇了, 因为有些人被单独算了两次。

在充分成因框架下, 如果 A 和 E 都是同一充分成因的组成成分, 那分别讨论 A 和 E 的归因比例就没有太大意义。因而, 讨论某一疾病应该归因于基因还是环境就没有太大意义。比如, 苯丙酮尿症导致的智力低下, 是由于患者吃的某种食物引起的。因而, 食物的归因比例是 100%, 基因的归因比例也是 100%。

知识点 5.1: 加法尺度和乘法尺度上的交互作用 (原书第 57 页)

等式 $\Pr[Y^{a=1,e=1}=1] - \Pr[Y^{a=0,e=1}=1] = \Pr[Y^{a=1,e=0}=1] - \Pr[Y^{a=0,e=0}=1]$ 简单变化可得到:

$$\Pr[Y^{a=1,e=1}=1] - \Pr[Y^{a=0,e=0}=1] = \{\Pr[Y^{a=1,e=0}=1] - \Pr[Y^{a=0,e=0}=1]\} + \{\Pr[Y^{a=0,e=1}=1] - \Pr[Y^{a=0,e=0}=1]\}$$

这一等式是交互作用的另一种定义形式。

上述等式成立, 则我们说 A 和 E 在加法尺度上没有交互作用。同时我们说因果性风险差 $\Pr[Y^{a=1,e=1}=1] - \Pr[Y^{a=0,e=0}=1]$ 是可累加的, 因为它可以写作有 A 无 E 时与有 E 无 A 时效应的相加。与之相反, 如果

$$\Pr[Y^{a=1,e=1}=1] - \Pr[Y^{a=0,e=0}=1] \neq \{\Pr[Y^{a=1,e=0}=1] - \Pr[Y^{a=0,e=0}=1]\} + \{\Pr[Y^{a=0,e=1}=1] - \Pr[Y^{a=0,e=0}=1]\}$$

则在加法尺度上存在交互作用。如果不等号换成大于符号, 我们称为超加性 (superadditive)。如果是小于符号, 我们称为劣加性 (subadditive)。

同理, 我们也可以用风险比来定义乘法尺度上的交互作用。如果有

$$\frac{\Pr[Y^{a=1,e=1} = 1]}{\Pr[Y^{a=0,e=0} = 1]} \neq \frac{\Pr[Y^{a=1,e=0} = 1]}{\Pr[Y^{a=0,e=0} = 1]} \times \frac{\Pr[Y^{a=0,e=1} = 1]}{\Pr[Y^{a=0,e=0} = 1]}, \text{ 则 } A \text{ 和 } E \text{ 在乘法尺度上存在交互作用。如}$$

果是大于号, 则是超乘性 (supermultiplicative), 如果是小于号, 则是劣乘性 (submultiplicative)。

知识点 5.2: 因果效应的单调性 (原书第 60 页)

如果治疗 A 是二分的, 只有在受益型个体中才有 $Y^{a=1} < Y^{a=0}$ 。对于其他三种类型, 有 $Y^{a=1} \geq Y^{a=0}$, 或者说, 个体的反事实结局随 a 单调递增 (或非减)。因而, 当一项治疗并不能改善个体结局时, 也即没有受益型个体时, 所有个体的反事实结局都随 a 单调递增, 我们称为 A 对 Y 的因果效应是单调的。

单调性的概念可以应用于两种不同的治疗 A 和 E 。如果每个个体的反事实结局 $Y^{a,e}$ 都随 a 和 e 单调递增, 则我们说 A 和 E 对 Y 的因果效应是单调的。也即在人群中不存在以下回应类型:

$$(Y^{a=1,e=1} = 0, Y^{a=0,e=1} = 1), \quad (Y^{a=1,e=1} = 0, Y^{a=1,e=0} = 1), \quad (Y^{a=1,e=0} = 0, Y^{a=0,e=0} = 1), \text{ 和} \\ (Y^{a=0,e=1} = 0, Y^{a=0,e=0} = 1).$$

知识点 5.3: 因果效应和充分成因的单调性 (原书第 67 页)

如果治疗 A 和 E 的效应都是单调的, 那么图 5.2 的某些充分成因就不可能存在。假设吸烟 ($A=1$) 不能防止心脏病、不锻炼 ($E=1$) 也不能防止心脏病, 那不可能存在成分中有 $A=0$ 或 $E=0$ 的充分成因。这是因为, 如果一个充分成因包含 $A=0$, 那某些个体 (比如 $U_2=1$) 在 $A=0$ 时会发生结局, 于是等价的, 如果他们接受了 $A=1$, 那结局就被阻止了。 $E=0$ 同理。如果 A 和 E 的因果效应是单调的, 图 5.3 给出了不可能存在的充分成因。

第五章图表

Table 5.1

Type	$Y^{a=0}$	$Y^{a=1}$
Doomed	1	1
Helped	1	0
Hurt	0	1
Immune	0	0

Table 5.2

Type	$Y^{a,e}$ for each a, e value			
	1, 1	0, 1	1, 0	0, 0
1	1	1	1	1
2	1	1	1	0
3	1	1	0	1
4	1	1	0	0
5	1	0	1	1
6	1	0	1	0
7	1	0	0	1
8	1	0	0	0
9	0	1	1	1
10	0	1	1	0
11	0	1	0	1
12	0	1	0	0
13	0	0	1	1
14	0	0	1	0
15	0	0	0	1
16	0	0	0	0

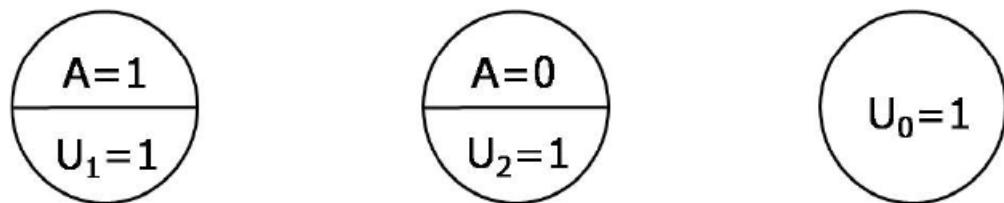


Figure 5.1

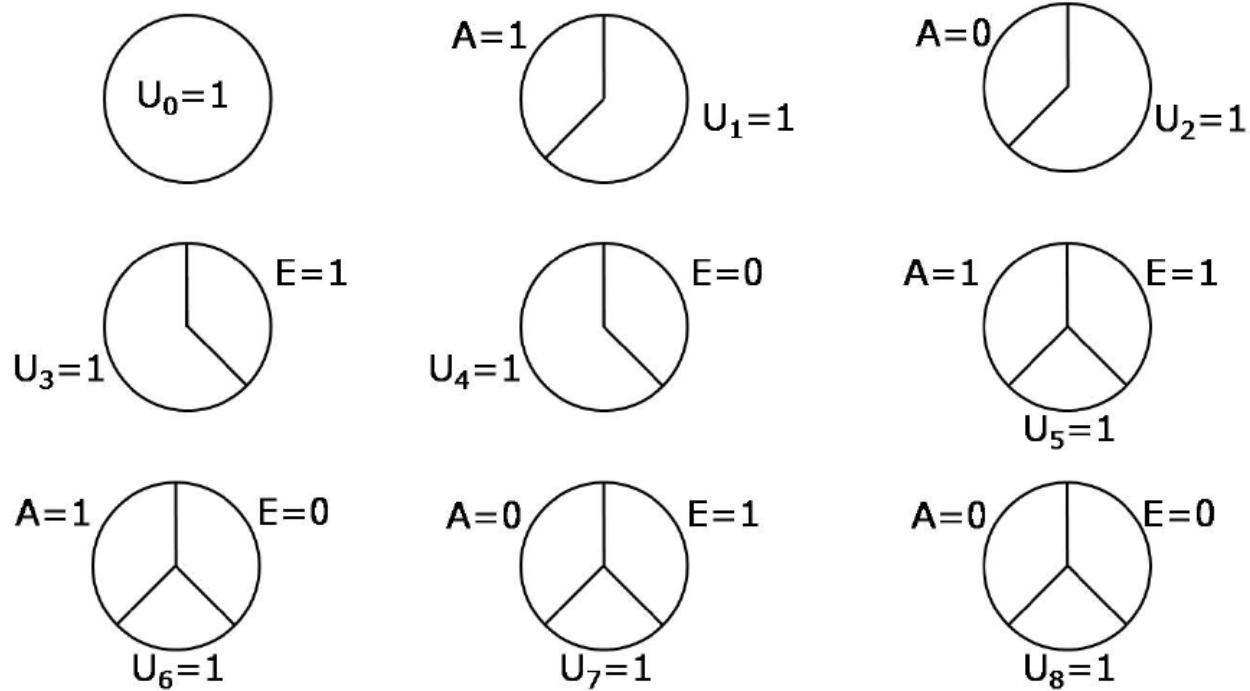


Figure 5.2

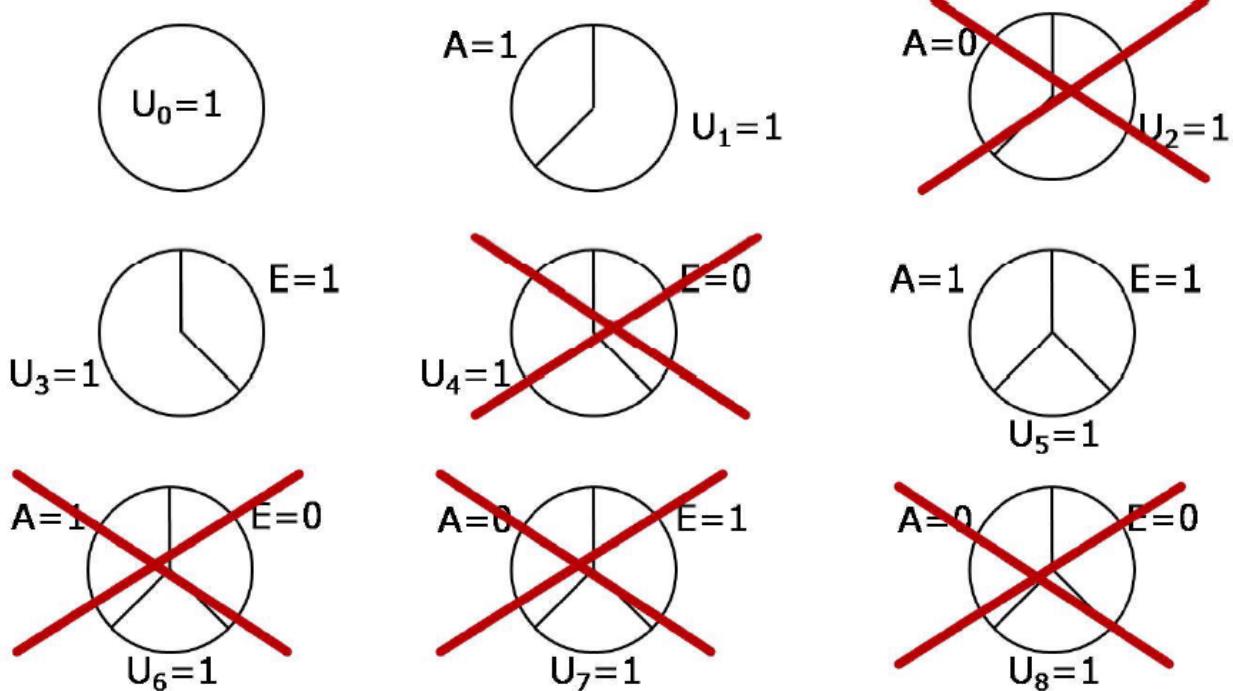


Figure 5.3

第六章 因果效应的图像表示

69 因果推断需要借助研究者的专业知识和不可检验的假设去勾勒治疗、结局以及其他变量之间的关系。前几章主要在简化情境中（比如抬头向天上看、心脏移植）讨论了计算因果效应所需的条件和方法，其目标旨在介绍能应用于复杂现实的种种理念。因为我们讨论的情境过于简单，从而无需详细描绘各变量间的因果关系。现在，我们将要把目光转向更复杂的现实世界，因而，清晰描述我们知道什么、我们假设了什么就变得非常重要。

本章将介绍一种图像工具，其能定性地展示我们所知道的东西，以及我们在因果推断中所做出的前提假设。图像能够清晰直观地展示我们知道了什么和我们假设了什么，从而能避免种种误解，提高研究者之间的交流效率。在因果推断中使用图像，能让我们更好地遵守这一重要准则：在你做出结论之前，先说出你假设了什么。

6.1 因果图

本章将会讲述用以表达因果推断关键概念的因果图。现代因果图的理论起源于计算机科学和人工智能。本章和接下来三章将会利用因果图来描述因果问题中的种种概念。

(Pearl (2009)，以及 Spirtes, Glymour 和 Scheines (2000) 等人的书详细讨论了这一主题)

图 6.1 由 3 个节点和 3 条边组成。每个节点表示一个随机变量 (L , A , Y)，每条边都有一个方向。我们按照惯例，从左至右表示时间顺序，因而， L 在时间上先于 A 和 Y ， A 在时间上先于 Y 。如同前几章， L 表示重症程度， A 表示心脏移植， Y 表示死亡。

从 A 到 Y 的箭头表示我们知道至少在一个个体中， A 对 Y 有直接因果效应（即中间不存在任何其他变量）。同理，如果 A 和 Y 之间不存在任何箭头，那对任何个体， A 对 Y 都没有直接因果效应。在我们的图 6.1 中，有箭头从 L 指向 A ，表示重症程度会影响接受心脏移植的概率。一个标准的因果图不会区分因果效应是有益的还是有害的。同时，在图 6.1 中，变量 Y 有两个诱因，而我们的因果图不会表示这两个诱因是否有交互作用。

类似图 6.1 的因果图被称为有向无环图 (Directed acyclic graphs, 简称为 DAG)，大多时候，会被简称为 DAG 图。“有向”指的是每条边都有一个方向：因为箭头是从 L 指向 A ，所以 L 是 A 的一个诱因，而不是反过来。“无环”指的是没有一个封闭的环形回路：任何变量都不能是自己的诱因，也不能间接通过其他变量成为自己的诱因。

除了因果推断，有向无环图在其他领域还有许多应用。本书重点在于有向无环图在因果推断中的应用。在因果性有向无环图中，如果控制了变量 A 的直接诱因，那除去 A 作为诱因的变量，

A 和图中的其他变量相互独立。这是因果性有向无环图的基本定义之一。这一假设被称为因果性马尔科夫假设 (causal Markov assumption) , 同时暗示了在一个因果性 DAG 图中, 如果任意两个变量有共同诱因, 那这个共同诱因也必须出现在同一图中。知识点 6.1 给出了因果性 DAG 图的正式定义。

在我们心脏移植的随机试验中, 分配到治疗组 A 的概率取决于重症程度 L 。因而 L 是治疗组分配 A 和死亡 Y 的共同诱因, 其必须出现在图中, 就如我们的图 6.1 所表示的一样, 此时这是一项条件随机试验。假设治疗组的分配不再取决于重症程度 L , 也即 L 不再是 A 和 Y 的共同诱因, 也就没必要出现在图中, 图 6.2 反映了这一情况, 此时这是一项边缘随机试验。

图 6.1 也可以表示一项观察性研究。其中, 我们假设心脏移植 A 有一个上游变量, 即重症程度 L , 同时假设死亡 Y 没有其他的诱因。如果 Y 还有其他诱因, 就算没有相关数据, 也必须包含在图中, 因为其可能是 A 和 Y 的共同诱因。在下一章, 我们将论证: 将图 6.1 视作一项观察性研究, 其实就是用图像的方法表示有界互换性 $Y^a \perp\!\!\!\perp A|L$ 。

许多人觉得因果推断的图像表示相比于反事实框架更直观, 也更容易理解。其实, 这两者是紧密联系在一起的。具体而言, 每一幅图都和一个反事实模型相关 (参见知识点 6.2)。反事实
71 模型只是把图像的直观用数学的语言描述了出来。传统的因果图并没有包含反事实变量, 因而两者的联系非常隐晦。最近发展出来的另一种有向无环图——单一世界干涉图 (Single world intervention graph, 简称 SWIG 图) ——通过将反事实变量融入到图中, 统一了反事实框架和因果推断的图像表示。我们将在第七章再介绍 SWIG 图。

(Richardson 和 Robins (2013) 共同发展完善了单一世界干涉图 (SWIG))

因果图可以简单地表示我们现有的知识、我们的假设、以及我们问题的定性结构。此外, 因果图还能表示各变量之间的可能关联, 其能精确地用来区分相关性和因果性, 使得其成为广受研究者喜爱的重要工具。接下来, 我们将介绍使用因果图来推断相关性和因果性。我们重在给出概念性的介绍, 而非严谨的数学推导。

6.2 因果图与边缘独立性

考虑以下两个例子。在第一个例子中, 你知道阿司匹林 A 能预防心脏病 Y , 即 $\Pr[Y^{a=1} = 1] \neq \Pr[Y^{a=0} = 1]$ 。图 6.2 表示了一项随机试验, 其中阿司匹林 A 随机、无条件地进行分配。在第二个例子中, 你知道携带打火机 A 其实对肺癌 Y 没有因果效应 (无论是预防性的还是诱发性的), 即 $\Pr[Y^{a=1} = 1] = \Pr[Y^{a=0} = 1]$, 然而吸烟 L 对携带打火机 A 和肺癌 Y 同时具有因果

效应。图 6.3 表示了第二个例子所描述的情境。在图 6.3 中, 没有箭头从 A 指向 Y , 表明携带打火机对肺癌没有因果效应; 而 L 是 A 和 Y 的一个共同诱因。

我们在画图 6.2 和图 6.3 的时候运用到了我们对这几个变量之间因果关系的理解。同时, 除去因果关系, 这两幅图也表示了各变量之间的相关性。在这两幅图中, A 和 Y 都是相关的, 不管这两者之间有没有箭头。这一点蕴含了因果图论的重要推论之一。

我们先来看图 6.2 表示的随机试验。在直觉上, 我们当然会期待由箭头连接的两个变量 A 和 Y 相关。这也是因果图论的结论之一: 如果 A 对 Y 有因果效应, 那么 A 和 Y 相关。这一点也和现实相一致: 在一个理想的无条件随机试验中, 因果性 $\Pr[Y^{a=1} = 1] \neq \Pr[Y^{a=0} = 1]$ 意味着相关性 $\Pr[Y = 1 | A = 1] \neq \Pr[Y = 1 | A = 0]$, 反之亦然。在因果图中, 因果性-相关性的对应体现在两个变量间的连线上。与因果性不同, 相关性是两个变量间的对称关系。相关性存在的时候, 其与箭头的方向无关。在图 6.2 中, 我们可以说 A 与 Y 相关, 也可以说 Y 与 A 相关, 这两个说法等价。

(DAG 图中, 两个变量 R 与 S 之间的路径, 指的是由一连串箭头组成、将 R 和 S 连接在一起的路线。这条路线中的变量不会出现两次。如果这些箭头的方向都是一样的, 那么这就是一条因果性的路径。如果不都一样, 则不是因果性的路径)

72 我们再来看图 6.3 表示的观察性研究。我们知道携带打火机 A 对肺癌 Y 没有因果效应。那携带打火机 A 和肺癌 Y 相关吗? 也即如果 $\Pr[Y^{a=1} = 1] = \Pr[Y^{a=0} = 1]$, 那

$\Pr[Y = 1 | A = 1] = \Pr[Y = 1 | A = 0]$ 成立吗? 为了回答这个问题, 让我们想象一下研究者可能用什么样的方法去研究这个问题。研究者可能会问一大群人他们有没有携带打火机, 然后再随访记录未来 5 年这些人是否会患肺癌。赫拉是研究人群中的一员。我们知道赫拉经常携带打火机, 因而她很可能经常吸烟, 从而就有更高概率患上肺癌。所以, 我们会直观地推论到 A 和 Y 相关, 因为

73 携带打火机 ($A = 1$) 的人比不带打火机的人 ($A = 0$) 有更高的概率患上肺癌, 也即

$\Pr[Y = 1 | A = 1] \neq \Pr[Y = 1 | A = 0]$ 。换句话说, 即使 A 对 Y 没有因果效应, 但是知道 A 的信息能提高我们预测结局 Y 的能力。我们不能因为 A 和 Y 相关就结论说 A 对 Y 有因果效应。在这里, 因果图论再一次验证了我们的直觉。用图论的术语来说, A 和 Y 相关, 是因为有信息途经共同诱因 L , 从 A 流动到了 Y (或者从 Y 到 A)。

让我们再考虑第三个例子。假设我们知道某个基因型 A 对我们是否抽烟 Y 没有因果效应, 即 $\Pr[Y^{a=1} = 1] = \Pr[Y^{a=0} = 1]$ 。而基因型 A 和抽烟 Y 都对心脏病 L 有因果效应。图 6.4 表示了这一情况。 A 和 Y 之间没有箭头, 表明 A 对 Y 没有因果效应。同时 A 和 Y 都是 L 的诱因。 L 作为 A 和

Y 共有的结果, 被称为对撞变量 (Collider), 其路径可以表述为 $A \rightarrow L \leftarrow Y$, 在这个路径中, 两个箭头在 L 这个节点处相撞, 故采用对撞这一命名。

我们的问题是 A 和 Y 是否相关。为了回答这一问题, 另一名研究者决定研究基因型 A 是否对吸烟 Y 有因果效应。他先确定了一群孩子的基因型, 然后再随访记录这群孩子长大后是否吸烟。阿波罗是研究人群中的一员。我们知道阿波罗没有这个基因型 ($A = 0$)。那他相较于其他人会更可能抽烟 ($Y = 1$) 吗? 我们知道基因型 A 不会影响是否吸烟 Y 。在这里, 就算知道基因型 A 的信息, 也不能提高我们预测是否吸烟 Y 的能力, 因为吸烟在是基因型 ($A = 1$) 和不是基因型 ($A = 0$) 人群中的比例是一样的, 即 $\Pr[Y = 1 | A = 1] = \Pr[Y = 1 | A = 0]$ 。换句话说, 我们能直观地知道 A 和 Y 不相关, 即 A 和 Y 相互独立, 记作 $A \perp\!\!\!\perp Y$ 。 A 和 Y 都能影响心脏病 L 这一信息, 对判断 A 和 Y 之间的相关性没有一点用。因因果图论再一次验证了我们的直觉。用图论的术语来说, 和其他变量不同, 对撞变量会阻断其所在路径中信息的流动。在路径 $A \rightarrow L \leftarrow Y$ 中, 因为对撞变量 L 阻断了这条路径, 所以 A 和 Y 相互独立。

总而言之, 如果一个变量是另一个变量的诱因, 或两个变量有共同诱因, 那么这两个变量相关, 否则这两个变量相互 (边缘) 独立。下一小节我们将讲述在什么情况下, 如果我们控制了第三个变量 L , 变量 A 和 Y 会相互独立。

6.3 因果图与条件独立

让我们再回顾一下图 6.2、6.3 和 6.4 所描绘的情境, 从而更好地讨论因果图中的条件独立。

在图 6.2 中, 我们料到阿司匹林 A 和心脏病 Y 相关, 因为阿司匹林对心脏病有因果效应。假设我们有了其他新知识: 阿司匹林 A 能影响心脏病 Y 是因为阿司匹林 A 能降低血小板凝集 B 。我们可以将这一新知识画成因果图, 如图 6.5 所示, 血小板凝集 B (1 表示高, 0 表示低) 是 A 对 Y 效应的中介变量。

在因果图中引入第三个变量后, 我们就面临一个新问题: 在 B 的每一层取值中 (也即控制了 B 之后), A 和 Y 还相关吗? 或者, 换句话说: 当我们有了 B 的信息后, 我们依然需要 A 的信息来预测 Y 吗? 为了回答这些问题, 假设我们在一大群人中收集 A 、 B 和 Y 的信息, 然后将分析限制在低血小板凝集 ($B = 0$) 的人群中。图 6.5 中 B 外围的方框就代表了限制的意思。如果将分析限制在 $B = 1$ 的人群中, 我们也需要画一个方框将 B 框起来。

低血小板凝集 ($B = 0$) 的人群有较低的心脏病风险。我们把其中一个个体拎出来, 不管他服用了阿司匹林 ($A = 1$) 还是未服用阿司匹林 ($A = 0$), 我们都知道他的心脏病风险较低, 因为他的血小板凝集程度较低。事实上, 因为阿司匹林只通过血小板凝集程度影响心脏病风险, 所

作者: Miguel A. Hernan, James M. Robins;

翻译: 罗家俊

以当我们知道血小板凝集的信息后, 再知道是否服用阿司匹林无助于我们预测这个个体的心脏病风险。因而在 $B = 0$ 的人群中, 治疗 A 和结局 Y 不相关。同样的道理适用于 $B = 1$ 。即使 A 和 Y 边缘相关, 但在给定 B 的条件下 A 和 Y 相互独立。这是因为在 B 的每一层取值中, 治疗组和非治疗组的心脏病发病风险都一样, 即 $\Pr[Y = 1 | A = 1, B = b] = \Pr[Y = 1 | A = 0, B = b]$ 。在图中, 变量 B 周围的方框阻断了路径 $A \rightarrow B \rightarrow Y$ 。

(我们经常说, 相关性的信息蕴含在缺失的箭头当中)

让我们回到图 6.3。在上一节我们说携带打火机 A 和肺癌 Y 相关, 是因为路径 $A \leftarrow L \rightarrow Y$ 是开放的, 相关性的信息能从 Y 流到 A 。现在, 我们想知道, 在控制了 L 之后, A 和 Y 是否依然相关。这一新问题可以用图 6.6 表示, 其中 L 四周的方框表示我们控制了 L 。假设我们只在不吸烟者 ($L = 0$) 中进行研究。此时, 知道一个人是否携带打火机 ($A = 1$) 无益于预测他的肺癌风险 ($Y = 1$), 因为之前我们所有论点的立足点在于携带打火机的人更可能吸烟。当我们把研究人群限制在吸烟者或不吸烟者中时, 这一论点就不再与我们的问题相关。即使 A 和 Y 是边缘相关, 在给定 L 的条件下 A 和 Y 相互独立。这是因为在 L 的每一层取值中, 携带打火机的人和不携带打火机的人的肺癌发病风险都一样, 即 $\Pr[Y = 1 | A = 1, L = l] = \Pr[Y = 1 | A = 0, L = l]$ 。换句话说,

$Y^a \perp\!\!\!\perp A | L$ 。在图中, 变量 L 周围的方框阻断了路径 $A \leftarrow L \rightarrow Y$ 。

(阻断治疗和结局之间通过共同诱因的信息流动, 从图论的角度论证了使用分层分析能使互换性成立)

最后, 让我们再思考一下图 6.4。在上一小节我们说基因型 A 和成为吸烟者 Y 相互独立, 因为路径 $A \rightarrow L \leftarrow Y$ 被对撞变量 L 阻断了。我们接下来会论述, 如果我们控制了 A 和 Y 的共同后果 L , A 和 Y 就会在此条件下相关。一名研究者想研究基因型 A 和吸烟状态 Y 之间的关系, 他将研究人群限制在有心脏病 ($L = 1$) 的人当中。图 6.7 中, L 周围的方框表示变量 L 在分析时被控制。我们知道基因型 ($A = 1$) 和吸烟 ($Y = 1$) 都是心脏病 ($L = 1$) 的诱因。在有心脏病的人群中, 如果不是基因型, 即 $A = 0$, 那另一个诱因出现的概率就会升高, 所以知道一个人有心脏病且不是基因型, 能为他是否吸烟提供一定的信息。也即, 在患有心脏病的人群中, 非基因型的人更可能是吸烟者, 所以 $L = 1$ 时 A 和 Y 逆向相关。但是, 如果仅凭 A 和 Y 在 L 的每一分层中相关就断定 A 对 Y 有因果效应则是错误的。极端一些, 如果 A 和 Y 是 L 仅有的两个诱因, 则患有心脏病的人群中, 其中一个诱因的缺失一定预示另一个诱因的存在。根据因果图论, 如果我们控制路径 $A \rightarrow L \leftarrow Y$ 中的对撞变量 L , 则我们会打开这条原本被阻断的路径。直观上来说, 两个变量 (诱因) 是否相关, 不应该受到未来发生事件 (它俩的共同后果) 的影响。但如果我们根据共同后果进行分层分析的话, 这两个诱因就会变得相关。

(第八章将会更多讨论控制了共同后果带来的相关性)

我们现在来考虑另一个例子。在图 6.7 的基础上, 我们加入了一个新的变量 C , 如图 6.8 所示。 C 表示利尿剂的使用情况, 是诊断出心脏病之后的治疗措施。因为 C 是 A 和 Y 的共同后果, 所以在 C 的每一取值中, A 和 Y 都会相关。根据因果图论, 如果我们控制了对撞变量 L 作为诱因的变量 C , 我们也能开放原本被阻断的路径 $A \rightarrow L \leftarrow Y$ 。

这一小节和上一小节讲述了在什么情况下, 因果图中的两个变量会相关: 如果一个变量是另一个变量的诱因, 如果两个变量有共同诱因, 如果两个变量有共同后果且数据分析被限制在共同后果(或者共同后果的下游变量)的某一取值中。我们也介绍了因果图论从而帮助我们确定因果图中的两个变量是否(有条件地)独立。我们用以描述并支持这些图论的论点来源于我们的现实例子与因果性直觉。不过这些论点也可以在数学上被严谨地证明。更系统地因果图论介绍, 请参见精讲点 6.1。精讲点 6.2 介绍了忠实行(faithfulness)¹。

(因果图论的数学理论被称为“有向分离”或“D-分离”)

两个变量的相关性也可能来自于误差或随机变异性。与之前讲述的其他三个理由不同, 随机误差导致的相关性会随着样本量的增大而减小。

我们会主要讲述结构性的相关, 而不是随机误差导致的相关。第十章之前, 我们都假设随机变异性不会影响我们的讨论。

6.4 因果图中的正数性和一致性

因为因果图定性地包含了变量间因果关系的知识, 所以它可以用来帮助我们分析因果推断中出现的问题。实际上, 第二章讲述的用以量化因果效应的公式——标准化和逆概率加权——也可以从因果图中推导得到, 它们其中的部分推导也被称为介入算法(do-calculus)²。因此, 我们在第一至第五章中选择的反事实理论, 并不意味着某一种理论框架更优越, 而只是为了统一我们所使用的数学表示方法。

(Pearl (2009) 综述了从图论中发展出来的因果推断定量分析方法)

不管我们用了什么样的表示方法(反事实或图像), 互换性、正数性和一致性都是因果推断所需的前提假设。如果这些条件中有一条不成立, 那数据分析得到的估计值就不能阐释为因果效

¹ Faithfulness 在数学上有时也翻译为单射性, 在此译本中翻译为忠实行。

² 该方法由 Judea Pearl 提出。此处的 do 更多是介入的意思, 指的是既然不能在随机试验中直接测量某值, 那就根据受控试验之外的观测数据来估计它。

76 应。在下一小节(同时还有第七章和第八章), 我们将讨论互换性在图中如何表述。现在我们先来讨论正数性和互换性。

正数性在图中可以大致表述为: 从节点 L 到节点 A 的箭头不是命定的。一致性的第一个部分——良定的干预——意味着从干预 A 到结局 Y 的箭头对应一个可能是假想的、但是不含糊的干预措施。在本书所讨论的因果推断中, 除非特别说明, 否则都会认为正数性成立。而一致性则被表述在数学符号标记中, 因为我们只考虑良定的干预措施。正数性和指向干预变量的箭头有关, 而一致性则只和从干预变量指出的箭头有关。

(更多关于因果图中正数性的讨论, 参见 Richardson 和 Robins (2013) 所著论文)

因而, 相较于因果图中的其他节点, 代表干预的节点有更特殊的地位。一些研究者在因果图中引入决策节点, 从而让这个区别更加明显。虽然决策论也能推导出相同的方法, 但我们在本章 77 的因果图中不引入决策论的内容。虽然有时在图中表示变量 A 的不同形式显得冗杂, 但我们在使用 SWIG 图的时候, 依然会在干预节点处标注其不同的形式, 从而表示不同的反事实变量。我们将在接下来几章讨论 SWIG 图。

(包含决策节点以表示不同干预的因果图, 也被称为影响图 (*influence diagram*))

干预节点的特殊地位也在第二章的树状图和第五章的充分成因图中有所体现。在树状图中, 与非干预变量 L 和 Y 对应的分支包含在圆内。而充分成因图区分了干预措施 A 和背景因素 U 。同时, 我们在第三章所讨论了良定的干预形式, 强调了对于干预变量 A 的特别要求, 而这些要求不^见于其他变量。

78 与之相比, 本章的因果图将图中的所有变量一视同仁——这也是用因果图表示非参数化结构方程模型的要求之一(参见知识点 6.2)。然而, 将图中所有变量同等对待有时存在误导性, 尤其是当某些变量是劣定的时候。在图中, 你可以让结局 Y 或协变量 L 表示“肥胖”。然而, 我们在第三章已经讨论过, 你不能让一个因果图中的干预 A 表示“肥胖”。在因果图中, 代表不同干预形式的节点需要是足够良定的。

假设我们现在要研究一项复合型干预 R (即有不同形式) 的因果效应。 $R=1$ 表示“每天至少锻炼 30 分钟”, $R=0$ 表示“每天锻炼少于 30 分钟”。因此, 只要一个个体每天的锻炼时间超过 30 分钟, 不管是 31 分钟, 还是 32、33、34 分钟等等, 都属于 $R=1$, 因而其治疗取值 $A(r=1)$ 有许多种不同形式。同理, 对于 $R=0$ 的个体, 其锻炼时间可以是 29 分钟, 也可以是 28、27、26 分钟等等, 其治疗取值 $A(r=0)$ 有许多不同形式。根据复合型干预的定义, 多种不同取值 $a(r)$ 能被映射到一个单一值 $R=r$ 。

图 6.10 的因果图包含了一个复合型干预 R , 同时也包含了干预的具体形式 A , 以及 A 和 R 的两个共同诱因 L 和 W 。 A 是一个包含了所有 $A(r)$ 变量的向量。与本章中的其他因果图不同, 图 6.10 的 R 和 A 命定地相关: 如果图中没有从 R 到 Y 的直接箭头, 那 A 中的不同形式就是足够清晰明确的。

尽可能地明晰复合型干预 R 和它的不同形式 $A(r)$, 对于清晰定义因果效应、确定相关数据、以及选择需要调整的变量等方面而言非常重要。

6.5 偏移的结构性分类

“偏移”一词经常在因果推断中出现, 并被用于虽然相关、但本质上完全不同的场景当中 (参见第十章)。当我们的数据无限、但依然不足以识别或计算因果效应时, 我们会说存在系统性偏移 (在本章, 我们假设样本量是无限的, 因而我们所说的偏移都是系统性偏移)。一般而言, 如果干预与结局之间的相关性不是来源于干预的因果效应, 那么我们会将这种相关性称为系统性偏移。因为因果图有助于表示相关性的不同来源, 所以我们能用因果图对系统性误差进行分类, 从而深化我们对偏移的讨论。

我们在前几章讨论过, 偏移的一个主要来源是治疗组与非治疗组之间互换性的缺失。对整个人群中的因果效应均值而言, 当 $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1] \neq \Pr[Y = 1 | A = 1] - \Pr[Y = 1 | A = 0]$ 的时候, 存在 (无条件的) 偏移, 此时 (无条件的) 互换性 $Y^a \perp\!\!\!\perp A$ 不成立。(无条件) 偏移的缺失意味着人群中相关性量度 (比如相关性风险比、相关性风险差) 不等于对应的效应量度 (比如因果性风险比、因果性风险差)。

(系统性偏移存在的时候, 没有一个估计值是一致的。一致估计值的概念参见第一章)

就算零假设成立——即治疗对结局不存在因果效应——互换性的缺失也会导致偏移。也即,就算治疗对结局不存在因果效应, 在数据中治疗依然可能和结局相关。我们把这种情况称为零值下的偏移。在表 3.1 所示的观察性研究中就存在零值下的偏移, 因为此时因果性风险比是 1, 但是相关性风险比是 1.26。任何导致零值下偏移的因果结构, 都会导致其余情况下的偏移 (即治疗对结局有非零的因果效应)。然而, 这个命题反过来并不成立。

(比如, 控制某些变量在非零值因果效应下会产生偏移, 但在零值下不会。参见 Greenland (1977) 和 Hernan (2017) 所著论文, 以及本书第十八章)

对 L 每个分层中的因果效应均值而言, 在某一层中有

$$\Pr[Y^{a=1} = 1 | L = l] - \Pr[Y^{a=0} = 1 | L = l] \neq \Pr[Y = 1 | A = 1, L = l] - \Pr[Y = 1 | A = 0, L = l], \text{ 则存在}$$

80 条件偏移, 也即有界互换性 $Y^a \perp\!\!\!\perp A | L$ 不是对所有 a 和 l 成立。

本章之前, 我们已经多次提及互换性的缺失。然而, 我们却未探讨哪种因果结构会导致互换性的缺失。现在, 在因果图的帮助下, 我们知道了在以下两种不同的因果结构中会出现互换性的缺失:

1. 存在共同诱因: 当干预和结局有共同诱因时, 相关性量度通常会和效应量度不同。许多流行病学家用“混杂”一词表示这一情形。
2. 控制共同后果: 来源于这一结构的偏移被大多数流行病学家称作“选择偏移”。

第七章我们将讨论因共同诱因引起的混杂, 第八章我们将讨论控制共同后果引起的选择偏移。这两种偏移, 都可以是因果效应零值下的偏移。

第九章我们将讨论偏移的另一个来源: 测量误差。迄今我们都认为所有变量——包括干预 A 、结局 Y 和协变量 L ——都是零误差测量的。但在实际中, 都会有或多或少的测量误差存在。由测量误差引起的偏移, 被称为测量偏移, 或者信息偏移。某些测量偏移也会导致因果效应零值下的偏移。

(非结构性) 随机变异性也可能导致另一种形式的偏移。参见第十章)

因此, 在接下来三章我们将讨论不同类型的系统性偏移, 即混杂、选择偏移和测量偏移。在观察性研究和随机试验中都可能出现这些偏移。在前几章, 我们将观察性研究抽象为不完美的随机试验, 同时只考察了没有失访、所有被试遵循规范的理想随机试验。在这些讨论中, 随机试验对偏移的敏感性还不是那么明显。虽然我们近似神话的随机试验有益于讲解与教学, 但实际中的随机试验基本不可能是这样。本书第一部分接下来几章就会详细讲述试验与观察的模糊界限。

在这之前, 我们再用因果图讨论一下效应修饰。

6.6 效应修饰的结构

识别偏移的可能来源是因果图的主要用途之一: 我们可以根据已有的专业知识画出各变量间的结构, 然后找出干预和结局间相关性的可能源头。然而, 因果图却难以描述我们在第四章讨论的效应修饰作用。

现在有一项探究心脏移植 A 对死亡 Y 的因果效应的随机试验。为了简化, 我们假设没有偏移, 因而图 6.2 就足够描述这项研究。此时计算 A 对 Y 的效应不是一件困难的事。因为此时相关

81 性就等于因果性, 所以相关性风险差 $\Pr[Y = 1 | A = 1] - \Pr[Y = 1 | A = 0]$ 就等于因果性风险差

$\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$ 。有研究者想进一步研究各医院的医疗技术是否对心脏移植 A 有效应修饰作用。因而研究者将被试进一步分为接受了高质量手术 ($V = 1$) 和普通质量手术 ($V = 0$) 两组, 然后再如第四章一样计算每一组中的因果效应, 最后确定了 V 在加法尺度上有效应修饰作用。图 6.11 包含了修饰因子 V , 并且有箭头从 V 指向 Y , 但 V 和 A 之间没有箭头 (A 是随机分配的, 因此独立于 V)。这里有两点值得注意。

首先, 如果图 6.11 中不包含 V , 它依然是一幅有效的因果图, 因为 V 不是 A 和 Y 的共同诱因。只是我们的因果性问题涉及了 V (在 V 的每一分层中, A 对 Y 的因果效应都一样吗?) , 所以 V 才被包含在图中。其他 $V - Y$ 之间的变量也可能是效应修饰因子。比如, 图 6.12 中 $V - Y$ 之间表示并发症的中介变量 N , 就是一个修饰因子。

其次, 图 6.11 不一定表示 V 有效应修饰作用。图 6.11 表示 A 和 V 都能影响 Y , 但是并不能定性地区分以下几种 A 和 V 之间的交互作用:

1. 治疗 A 对死亡 Y 的因果效应在 $V = 0$ 和 $V = 1$ 时中方向一样。
2. 治疗 A 对死亡 Y 的因果效应在 $V = 0$ 和 $V = 1$ 时方向相反 (即存在质的效应修饰作用)。
3. 治疗 A 只在 V 的某一层中对 Y 有因果效应, 而在另一层中没有。

在这个例子中, 修饰因子 V 对结局有因果作用。然而许多修饰因子却对结局没有因果作用, 它们只是某些未知变量的替代物, 而这些未知变量才真正的有修饰作用。图 6.13 中包含了一个变量 S , 其表示手术的花费 (1 表示花费高, 0 表示花费低), 并受到手术质量 V 的影响, 但对结局 Y 没有因果效应。如果我们根据 S (而非 V) 进行分层分析, 我们会发现 S 也有效应修饰作用, 虽然真正修饰了 A 对 Y 效应的应该是 V 。 S 是修饰因子替代物, 而 V 因果性修饰因子 (参见 4.2 小节)。因为真正的修饰因子和其替代物在实践中经常难以区分, 效应修饰这一概念就经常模糊地包括两者。正如我们 4.2 小节所讨论的一样, 为了避免混淆, 有的人更偏爱用“因果效应的异质性”这一中性表达, 而不是“效应修饰”一词。有的人会将这里的结论阐述为“因为花费越高手术的效果越明显, 所以花费修饰了心脏移植对死亡的效果”, 从而以此作为证据支持提高医疗收费, 而不是医疗质量。

(VanderWeele 和 Robins (2007b) 通过因果图更细致地讨论了效应修饰的分类)

修饰因子替代物仅仅是一个和因果性修饰因子有关联的变量。图 6.13 描绘的场景中, 因果性修饰因子和其替代物的相关性来自于前者是后者的诱因。然而, 这样的相关性也可以是因为两者有共同诱因, 或者控制了共同后果。比如, 图 6.14 包含了表示居住地的变量 U , 和表示护照国

籍的变量 P 。居住地 U 是治疗质量 V 和国籍 P 的共同诱因。因而 P 也是一个修饰因子替代物。

另一个例子（虽然有些蠢）是如图 6.15 所示，其中包含表示治疗花费的变量 S 和表示只喝瓶装矿泉水的变量 W 。只喝瓶装矿泉水 W 能影响 S ，但并不影响死亡 Y 。当我们分析局限于低治疗花费 ($S=0$) 人群时，只喝瓶装矿泉水 W 就和治疗质量 V 相关，从而 W 也是一个修饰因子替代物。总而言之，修饰因子能通过作为直接诱因、有共同诱因、或者控制共同后果这些方式和其替代物相关。

总的来说，因果图并不能表示两个变量之间的交互作用。然而，如果我们在因果图中加入表示充分成因的增强型节点（即有命定的箭头从干预指向充分成因），我们就能在因果图中包含交互作用的信息。有交互作用存在的时候，控制共同后果会影响相关性的大小和方向。我们将在第八章讨论这一增强型因果图。

第六章精讲点和知识点

精讲点 6.1: 有向分离 (原书第 76 页)

我们根据以下准则定义一条路径是开放的还是阻断的：

1. 如果没有任何变量被控制，一条路径是阻断的，当且仅当两个相反方向的箭头在路径中的某个变量处相撞。在图 6.1 中，路径 $L \rightarrow A \rightarrow Y$ 是开放的，但是路径 $A \rightarrow Y \leftarrow L$ 是阻断的，因为有两个箭头在 Y 处相撞。我们将路径 $A \rightarrow Y \leftarrow L$ 中的 Y 称为对撞变量。
2. 如果一条路径中的非对撞变量被控制，那这条路径被阻断。在图 6.5 中，控制了 B 之后， A 和 Y 之间的路径被阻断。我们用方框将变量框起来，表示我们控制了这个变量。
3. 如果一条路径中的对撞变量被控制，那这个对撞变量不再阻断这条路径。在图 6.7 中，控制了 L 之后， A 和 Y 之间的路径变得开放。
4. 如果控制对撞变量的下游变量，那么也会开放原来对撞变量所在的路径。在图 6.8 中， C 是对撞变量 L 的下游变量。控制了 C 之后， A 和 Y 之间的路径变得开放。

以上 4 点可以用如下陈述进行总结：一条路径，当且仅当其中的非对撞变量被控制、或路径中有对撞变量且对撞变量的下游变量没有被控制，这条路径才会被阻断。如果两个变量之间的所有路径都被阻断，我们称这两个变量被有向分离，反之则称为有向连接。如果有两个变量集合，其中一个集合的每一个变量都和另一个集合的每一个变量有向分离，那我们称这两个集合被有向分离。因而，图 6.1 的 A 和 L 没有被有向分离，因为 L 和 A 之间有一条开放的路径 ($L \rightarrow A$)，即使另一条 ($A \rightarrow Y \leftarrow L$) 被对撞变量 Y 阻断。在图 6.4 中， A 和 Y 被有向分离，因为它俩之间的唯一路径被对撞变量 L 阻断。

统计上的独立性和因果图中的有向分离有密切的关系, 然而这一关系依赖于因果性马尔科夫假设(参见知识点 6.1)。在一个因果性 DAG 图中, 控制了母变量后, 任一变量都和它的非下游变量相互独立。Pearl (1988) 证明了以下定理: 因果性马尔科夫假设成立时, 给定三个不相交的变量集合 A 、 B 和 C , 如果在控制了 C 的情况下 A 和 B 有向分离, 那么在给定了 C 的情况下 A 和 B 在统计上相互独立。这一命题的逆命题——即如果在给定了 C 的情况下 A 和 B 在统计上相互独立, 那么在控制了 C 的情况下 A 和 B 有向分离——是另一个重要假设, 其被称为忠实性假设。我们将在精讲点 6.2 中讨论这一假设。在忠实性假设下, 图 6.5 的 A 在给定 B 的情况下和 Y 相互独立; 图 6.7 的 A 在给定 L 的情况下和 Y 并不相互独立; 图 6.8 的 A 在给定 C 的情况下和 Y 并不相互独立。Pearl (1995) 有向分离的准则规范化, 从而能用来从因果图中推导出相关性陈述。另一套等价的图论准则被称为“教化准则”, 由 Lauritzen 等人提出 (1990)。

精讲点 6.2: 忠实性 (原书第 77 页)

在因果性 DAG 图中, 没有从 A 到 Y 的箭头就表示极端因果零假设成立, 即 A 对人群中的每一个个体的 Y 都没有因果效应。而从 A 到 Y 的箭头 ($A \rightarrow Y$, 如图 6.2 所示), 则表示 A 至少对人群中某一个个体的 Y 有因果效应。因而, 在图 6.2 中, A 对 Y 的因果效应, 以及 A 和 Y 之间的相关性, 都不会为零。然而, 以下陈述却不一定为真: 图 6.2 所代表的情境, 存在既没有因果效应、有没有相关性的可能。比如, 表 4.1 的数据里, 心脏移植 A 提升了女性中的死亡率 Y , 但降低了男性中的死亡率。有益和有害的效果正好相互完全抵消, 因而整体的因果效应均值为零。然而图 6.2 也能代表这一情境, 因为治疗 A 对结局的因果效应在某些个体中不为零。

正式而言, 忠实性假设的定义是: 对于三个不相交的变量集合 A 、 B 和 C (其中 C 可以是空集), 如果在给定了 C 的情况下 A 和 B 相互独立, 那么在控制了 C 的情况下 A 和 B 有向分离。在我们的例子里面, 如果因果图中有现实数据中不存在的非零相关性, 那我们会说数据的联合分布对这个 DAG 图不具有忠实性。在我们的例子里面, 不忠实性是效应修饰的结果——效应修饰让方向相反的两个因果效应完全相互抵消。这种完全抵消很罕见, 因而我们在全书中都会假设忠实性成立。因为不忠实性的例子很罕见, 所以在实践中我们认为有向分离(参见精讲点 6.1)不成立等同于相关。

然而, 有的试验设计可能会破坏忠实性。以第四章 4.5 小节中的前瞻性研究为例。我们在根据 L 进行匹配后计算了 A 对 Y 的因果效应。在匹配人群中, L 和 A 不再相关。也就是说, 被试是因为他们具有特定的 L 和 A 的取值才被选入到匹配人群中。图 6.9 可以用来表示这样一种匹配人群的情境, 其中表示是否被选择的变量 S (1 表示被选中, 0 表示没被选中) 是 A 和 L 的共同

后果。 S 周围的方框表示分析被限制在被选中的人群当中 ($S = 1$)。根据有向分离的准则, 当我们控制了 S 后就有两条开放路径: $L \rightarrow A$ 和 $L \rightarrow [S] \leftarrow A$ 。因而部分人会认为在控制了 S 之后 L 和 A 会相关。然而我们的匹配保证了 L 和 A 不相关 (参见第四章)。这是为什么呢? 这是因为, 来源于开放路径 $L \rightarrow [S] \leftarrow A$ 的相关性, 和来源于 $L \rightarrow A$ 的相关性大小相等, 但是方向相反, 因而正好完全抵消。也就是说, 匹配导致了不忠实性。

最后, 图中各变量之间的命定关系, 也可能破坏忠实性。具体而言, 如果两个变量是由一个或多个有命定箭头的路径相连接, 当这些路径被阻断时, 这两个变量相互独立; 当这些路径开放时, 这两个变量依然可能相互独立。在本书中, 除非特别说明, 我们都会假设忠实性成立。当我们进行数据分析的目标是发现各变量间的因果结构时, 我们也会假设忠实性成立 (参见精讲点 6.3)。

精讲点 6.3: 挖掘因果结构 (原书第 79 页)

在本书中, 我们总是用我们的专业知识 (或者假设) 来构建因果图, 进而指导我们的数据分析。但是反过来呢? 通过数据分析进而掌握各变量的因果结构, 这一过程通常被称为挖掘 (参见 Spirtes (2000) 等人所著论文)。

在挖掘因果结构时, 通常会假设我们所观测到的数据来源于一个未知的 DAG 图。这个 DAG 图中除了我们观测到的变量之外, 还有一些我们不知道、未观测的变量 U 。我们在挖掘因果结构时需要假设忠实性, 从而所观测数据中统计上的相互独立就意味着 DAG 图中因果箭头的缺失。然而, 即使我们假设了忠实性, 挖掘因果结构也经常是不可能的任务。比如, 如果我们在数据中观测到了 B 与 C 之间有很强的相关性, 我们依然不能确定 B 和 C 之间的具体关系。究竟是 B 导致了 C ($B \rightarrow C$), 还是 C 导致了 B ($C \rightarrow B$), 抑或 B 和 C 有未知的共同诱因 ($B \leftarrow U \rightarrow C$), 抑或 B 和 C 有共同后果且共同后果已被控制 ($B \rightarrow [U] \leftarrow C$), 甚至以上几种可能的组合? 如果我们知道 B 和 C 的时间顺序, 我们也只能排除 $B \rightarrow C$ 或者 $C \rightarrow B$ 。

不过, 在某些情况下, 挖掘因果结构是可行的。假设我们无限的数据中有三个变量 Z 、 A 和 Y , 我们知道在时间顺序上 Z 在前, A 其次, Y 最后。我们的数据分析发现这三个变量边缘性相关, 唯一成立的条件独立性为 $Z \perp\!\!\!\perp A \mid Y$ 。因而, 如果我们假设忠实性成立, 那可能的因果结构会是 $Z \rightarrow A \rightarrow Y$, 或者 Z 和 A 有一个共同诱因 U 从而替代 (或者同时存在) 从 Z 到 A 的箭头。这是因为, 如果 Z 是 Y 的母变量, 或者 Z 和 Y 有共同诱因, 或者 A 和 Y 有共同诱因, 那么 Z 和 Y 就不能在给定 A 的时候相互独立 (假设忠实性成立)。因而, 要解释 Y 和 A 之间的相关性, 就必须有

一个箭头从 A 指向 Y 。总而言之, 我们推导而得的 DAG 图意味着 Z 不会是 Y 的母变量, A 和 Y 不可能有共同诱因, 以及 A 对 Y 的因果效应可以定义为 $E[Y|A=1] - E[Y|A=0]$ 。

问题在于我们不可能有无限的数据。Robins 等人 (2003) 论证了由于抽样变异性的存在, 有限样本中的独立性检验不可能区分 “ A 是 Y 诱因” 和 “ A 不是 Y 的诱因” 这两个假设区别。因而, 除了忠实性假设之外, 如果我们不再补充其他假设, 我们不可能有足够的信心在数据中检验因果效应的存在与否。更多因果结构的挖掘, 请参阅 Peters 等人 (2017) 所著书籍。

知识点 6.1: 有向无环图 (原书第 70 页)

我们将有向无环图 (DAG) G 的定义为: 节点表示随机变量 $V = (V_1, \dots, V_M)$ 并通过有向箭头连接, 同时不存在有向环的图。我们用 PA_m 表示 V_m 的母变量集合, 即箭头从自身出发指向 V_m 的变量的集合。如果顺着一系列的箭头方向, 我们能从 V_m 到 V_j , 则我们称 V_m 是 V_j 的下游变量 (即 V_j 是 V_m 的上游变量)。比如, 图 6.1 中, $M = 3$, 并且我们可以把变量分别重新定义成 $V_1 = L$, $V_2 = A$ 以及 $V_3 = Y$ 。 $V_3 = Y$ 的母变量集合 PA_3 就是 (L, A) 。在我们的表示规则里面, 如果 $m > j$, 则 V_m 不可能是 V_j 的上游变量。我们定义: 如果 (对每一个 j) 控制了 V_j 的上游变量 V_j 就和它的非下游变量相互独立, 那变量 V 的分布相对于一个 DAG 图 G 有马尔可夫性 (或等价的, 变量的分布参数由 DAG 图 G 确定)。

如果 DAG 图满足了以下条件, 则其可以视作一个因果性 DAG 图: 1) 如果没有箭头从 V_j 指向 V_m (即 V_j 不是 V_m 的母变量), 则 V_j 对 V_m 没有因果效应; 2) 图中任意两个变量的共同诱因, 就算没有测量, 都必须出现在图内; 3) 任意变量都是其下游变量的诱因。

只有当我们通过一个假设将 DAG 图所表示的因果结构和我们现实中的数据相联系起来, 因果性 DAG 图才有实际作用。这个假设也被称为因果性马尔科夫假设: 如果我们控制了变量 V_j 的直接诱因, 那 V_j 就和图中它不作为诱因的其他变量相互独立。也就是说, 控制了 V_j 的母变量, 和 V_j 和它的非下游变量相互独立。后一陈述在数学上等价于 DAG 图中变量 V 的密度 $f(V)$ 满足马尔科夫分解:

$$f(v) = \prod_{j=1}^M f(v_j | pa_j)$$

知识点 6.2: 反事实模型与因果性 DAG 图 (原书第 72 页)

本书中的因果性 DAG 图 G 都表示一个反事实模型。为了给 G 表示的反事实模型进行定义, 我们需要引进下述标记。对于任一随机变量 W , 用 \mathbf{W} 表示 W 所有可能取值 w 的集合。对任一个有序变量 W_1, \dots, W_m 的集合, 记 $\bar{w}_m = (w_1, \dots, w_m)$ 。记 R 为变量集合 V 的一个子集, 同时让 r 表示 R 的一个取值。从而, V_m^r 表示 $R = r$ 时 V_m 的反事实取值。

V 是 DAG 图 G 中节点 (即变量) 的集合。一个由 DAG 图 G 表示的非参数化结构方程模型 (NPSEM) 有两个重要假设。其一是存在不可观测的随机变量 (误差) ε_m ; 其二是存在一个命定的但不可知的函数 $f_m(pa_m, \varepsilon_m)$, 其中 $V_1 = f_1(\varepsilon_1)$, 且每一个反事实取值上一步的反事实取值 $V_m^{\bar{v}_{m-1}} = V_m^{pa_m}$ 由 $f_m(pa_m, \varepsilon_m)$ 给出。也即, 只有 V_m 的母变量对 V_m 才有直接效应。一个 NPSEM 意味着我们可以对图中的任意变量 V_j 进行干预, 因为 V_j 任一取值 v_j 所对应的反事实结局被假设存在。现实中的取值 V_m 和 $R \subset V$ 时的反事实取值 V_m^r 都可以通过递归的方法从 V_1 和 $V_j^{\bar{v}_{j-1}}$ 得到, 其中 $m \geq j > 1$ 。比如 $V_3^{v_1} = V_3^{v_1, v_2^{v_1}}$, 即当 $V_1 = v_1$ 时 V_3 的反事实取值 $V_3^{v_1}$, 就是 v_2 等于 V_2 的反事实取值 $V_2^{v_1}$ 时的反事实取值 $V_3^{v_1, v_2}$, 同理有 $V_3 = V_3^{V_1, V_2^{V_1}}$ 以及 $V_3^{v_1, v_4} = V_3^{v_1}$ (这是因为 V_4 不是 V_3 的直接诱因)。

Robins (1986) 将 NPSEM 称作 “最精微因果性阐释结构树状图” (Finest causally interpreted structural tree graph, FCIST)。Pearl (2000) 讨论了如何用 DAG 图表示 NPSEM。Robins (1986) 还提出了一个更符合现实的 “因果性阐释结构树状图”, 在其中不是所有变量都能被干预。为了解释的方便, 我们假设所有的变量都能被干预, 即使我们的统计模型并不需要这些假设。

一个 FCIST 模型并不保证知识点 6.1 中所说的因果性马尔科夫假设成立, 因而需要其他统计上的独立性假设。比如, Pearl (2000) 就假设在 NPSEM 中所有 ε_m 相互独立。我们将这一假设和模型合称为 NPSEM-IE (IE 表示 independent error)。与此同时, Robins (1986) 假设当 \bar{v}_{j-1} 是 \bar{v}_{m-1} ($j < m$) 的一个子向量时, $V_m^{\bar{v}_{m-1}} = f_m(pa_m, \varepsilon_m)$ 和 $V_j^{\bar{v}_{j-1}} = f_j(pa_j, \varepsilon_j)$ 独立。我们将这一假设和模型合称为 “最精微完全随机因果性阐释结构树状图” (FFRCISTG)。Robins (1986) 的假设保证因果性马尔科夫假设成立。NPSEM-IE 是一个 FFRCISTG, 但反过来不成立, 这是因为 NPSEM-IE 相较于 FFRCISTG 有更多的假设 (参见 Robins 和 Richardson 在 2011 年所著论文)。

一个 DAG 图能表示一个 NPSEM，但我们需要说明它表示的是哪一种 NPSEM。比如，图 6.2 就对应一个 NPSEM-IE (其中完全互换性($Y^{a=1}, Y^{a=0}$) $\perp\!\!\!\perp A$ 成立)，或者一个 FFRCISTG (其中边缘互换性 $Y^a \perp\!\!\!\perp A$ 成立。)。在本书，只要没有特别说明，我们都假设 DAG 图对应的是 FFRCISTG。

第六章图表

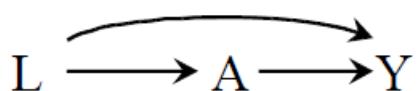


Figure 6.1

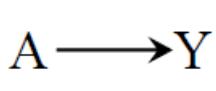


Figure 6.2

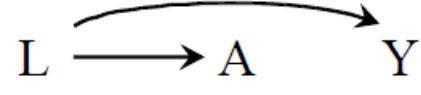


Figure 6.3

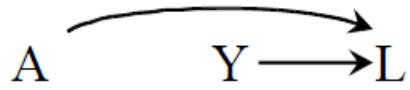


Figure 6.4

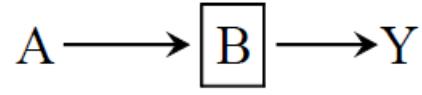


Figure 6.5

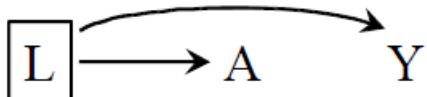


Figure 6.6

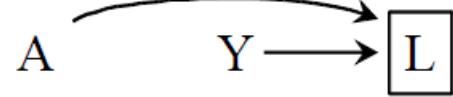


Figure 6.7

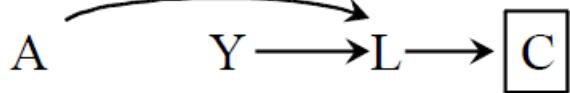


Figure 6.8

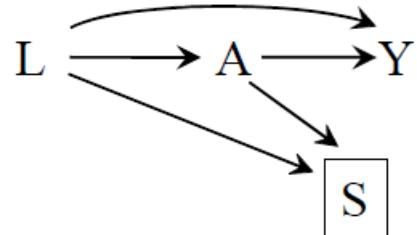


Figure 6.9

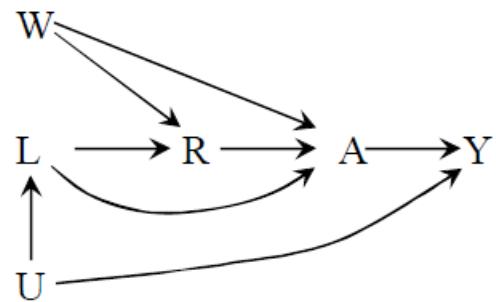


Figure 6.10

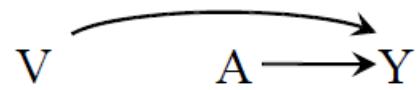


Figure 6.11

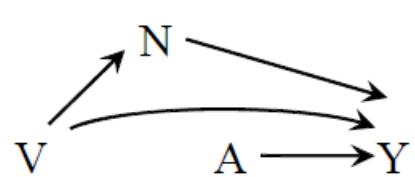


Figure 6.12

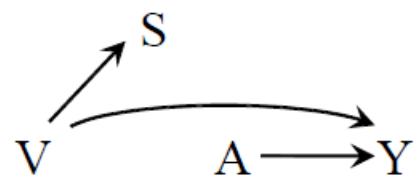


Figure 6.13

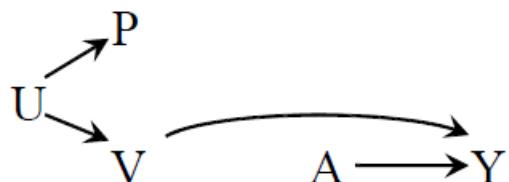


Figure 6.14

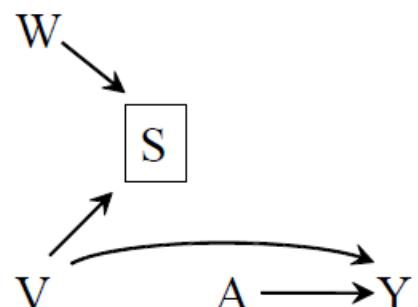


Figure 6.15

第七章 混杂

83 假设某研究者进行了一次观察性研究来探索以下问题：“一个人抬头向天上看会影响其他行人抬头向上看吗？”他发现第一个行人抬头向上看和第二个行人也抬头向上看之间存在相关性。然而，这名研究者发现当行人听到雷声时都会抬头向天上看。因此，到底是第一个行人的行为让第二个行人抬头向天上看，还是雷声让第二个行人向天上看？这名研究者只能说雷声的存在混杂了第一个人抬头向上的因果效应。

在随机试验中，治疗组的分配都是随机的。但是在观察性研究中，治疗与否却受到许多因素的影响。如果这些因素影响了结局出现的风险，那么这些因素的效应就会和治疗的效应混杂在一起。此时我们会说存在混杂，这也是治疗组和非治疗组之间互换性缺失的体现。混杂的存在经常被视为观察性研究的重要缺陷之一。在有混杂的时候，即使我们的研究样本无限大，我们常说的“相关性不等于因果性”也依然是正确的。本章将给出混杂的正式定义，同时将讨论调整混杂的不同方法。

7.1 混杂的结构

混杂的结构，即因为治疗和结局的共同诱因而产生的偏移，可以用因果图进行表示。比如，图 7.1（与图 6.1 相同）就包含了治疗 A 、结局 Y 以及治疗与结局的共同诱因 L 。图 7.1 形象说明了治疗和结局的相关性有两个来源：1) 路径 $A \rightarrow Y$ 所代表的 A 对 Y 的直接因果效应；2) 路径 $A \leftarrow L \rightarrow Y$ 所代表的由共同诱因而产生的偏移。第二条路径也被称为后门路径。

如果图 7.1 不包含共同诱因 L ，那么治疗和结局之间的路径就只剩 $A \rightarrow Y$ ，因而 A 和 Y 之间所有的相关性都来自 A 对 Y 的因果效应。也即，相关性风险比 $\Pr[Y=1|A=1]/\Pr[Y=1|A=0]$ 就等于因果性风险比 $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]$ ，此时相关性就是因果性。而共同诱因 L 的存在是 A 和 Y 之间相关性的另一个来源，我们将其称为 A 对 Y 因果效应的混杂变量。因为混杂以及混杂变量的存在，相关性风险比不再等于因果性风险比，此时相关性不再是因果性。

（在因果性 DAG 图中，后门路径的箭头总是指向治疗（或干预））

观察性研究中有大量混杂的例子：

- 84 • 职业因素：如果身体健康 L 是成为消防队员 A 和低死亡风险 Y 的共同诱因，那 A 对 Y 的因果效应就混杂了 L 的影响。图 7.1 反映了这一情境。这种情况也被称为“健康工人偏移”。

- 临床决策: 如果医生更可能给有心脏病 L 的人开阿司匹林 A , 那阿司匹林 A 对中风风险 Y 的因果效应就混杂了 L 的影响, 这是因为心脏病 L 不仅是开药的指征, 也是中风的风险因素。 L 和 Y 之间的关系, 可以如图 7.1 一样, 也可以如图 7.2 一样, 图 7.2 中心脏病 L 和中风 Y 都是由动脉粥样硬化 U 引起的, 而 U 是一个未测的变量。这种偏移被称为“指征混杂”或者“引流效应”。后一名称一般只用于因为病人个体特征 L 而用药 A 所引起的偏移。

(有些人更喜欢用双向箭头表示未知变量 U 和其他已知变量之间的关系)

- 生活方式: 如果某种生活方式 A (比如锻炼) 和另一个生活方式 L (比如吸烟) 相关, 并且 L 对结局 Y (比如死亡) 有因果效应, 那锻炼 A 对死亡 Y 的因果效应就混杂了 L 的影响。图 7.3 描绘了 L 、 A 和 Y 之间的可能关系, 其中未知变量 U 代表了某种同时影响了锻炼和吸烟的人格特质或者社会因素。然而一个可能出现的问题是, 如果 U 是某种亚健康状态, 并能同时导致锻炼 A 的缺乏和疾病 Y 风险的上升, 那在 L 未知的时候, 这种情况经常被称为逆向因果关系。
- 基因因素: 如果存在一个 DNA 序列 L , 其对某种性状 Y 有因果效应, 同时和另一 DNA 序列 A 相关, 那序列 A 对 Y 的因果效应就混杂了 L 的影响。这一情境依然可以用图 7.3 表示, 并经常被称为连锁不平衡或者人群分层。后一名称经常被用于研究人群中不同种族所引起的偏移。因而 U 能表示种族或者其他与 DNA 序列连锁有关的因素。
- 社会因素: 如果 55 岁的残疾程度 L 影响了未来 65 岁的收入 A 和 75 岁的残疾程度 Y , 那 A 对 Y 的因果效应就混杂了 L 的影响。这一情境可用图 7.1 表示。
- 环境暴露: 如果空气中超细颗粒物 A 的浓度和其他污染物 L 的浓度协同变化, 并且 L 能引起冠心病 Y , 那 A 对冠心病 Y 的因果效应就混杂了 L 的影响。这一情境可以用图 7.3 表示, 其中未知变量 U 可能代表影响不同污染物浓度的天气情况。

(在早期, Yule (1903) 提供了对离散型混杂变量的统计描述, Pearson 等人 (1899) 提供了连续型混杂变量的描述。Yule 将混杂引起的相关性分为“虚构的”, “幻觉的”和“明显的”三种。Pearson 则称之为“伪”相关性。然而, 这些相关性都不是虚假的或者错误的, 它们是真实存在的。由共同诱因引起的相关性是真实的, 虽然它不能阐述为治疗的因果效应。或者, 如 Yule 所言, 这些相关性“不能被赋予解释意义”)

在以上所有例子当中, 偏移的结构都是一样的, 也即其中的偏移都来源于治疗 (或干预) A 和结局 Y 的共同诱因 L 或 U 。我们将源于共同诱因的偏移称为混杂。我们会用其他术语命名源于其他结构的偏移。为了简化, 在本章我们有如下假设: (1) 所有变量都是完美零误差测量的;

85 (2) 数据是从人群中随机抽样而来, 也即 DAG 图中没有表示选择的节点 S ; (3) 不存在随机变异性。我们将在第八章介绍 DAG 图中的选择节点, 在第九章引进表示测量误差的节点, 在第十章讨论随机变异性。

7.2 混杂与互换性

我们用因果图定义了混杂, 用反事实框架定义了互换性。现在我们需要把这两者联系起来。为了简化, 在本章我们假设正数性和一致性成立, 同时我们的 DAG 图中没有变量被控制。

(Greenland 和 Robins (1986, 2009) 详尽地讨论了混杂与互换性的关系)

当互换性 $Y^a \perp\!\!\!\perp A$ 成立时, 就像在边缘随机试验中一样, 我们不需要调整任何变量就能识别治疗的因果效应均值。对一个二分治疗 A , 因果性风险差 $E[Y^{a=1}] - E[Y^{a=0}]$ 就等于相关性风险差 $E[Y|A=1] - E[Y|A=0]$ 。

当互换性 $Y^a \perp\!\!\!\perp A$ 不成立, 但有界互换性 $Y^a \perp\!\!\!\perp A|L$ 成立时, 就如在条件随机试验中一样, 治疗组的分配概率取决于 L , 此时我们依然可以识别治疗的因果效应均值。不过, 就如我们在第二章所讨论的一样, 此时我们需要通过标准化或逆概率加权等方法调整 L 后, 才能识别人群中的因果效应 $E[Y^{a=1}] - E[Y^{a=0}]$ 。同时, 我们在第四章讨论过, 有界互换性成立时, 我们可以通过分层分析识别 L 任一取值 l 分层中的因果效应 $E[Y^{a=1}|L=l] - E[Y^{a=0}|L=l]$ 。

(在有界互换性下,

$$E[Y^{a=1}] - E[Y^{a=0}] = \sum_l E[Y|L=l, A=1] \Pr[L=l] - \sum_l E[Y|L=l, A=0] \Pr[L=l]$$

在实践中, 如果我们认为混杂很可能存在, 那我们就要面对一个严峻的问题: 我们能否找到一系列已测的变量 L , 只要控制了这些变量, 有界互换性就能成立? 回答这一问题并不轻松, 这是因为在现实世界的复杂问题中思考有界互换性 $Y^a \perp\!\!\!\perp A|L$ 不是一件直观的事情。

在本章, 我们将论述, 如果我们知道数据所对应的 DAG 图, 我们就能回答上述问题。假设我们知道数据背后真实的 DAG 图 (不要在意我们是怎么知道的), 那我们应该怎么使用这幅 DAG 图来判断有界互换性取决于哪些变量呢? 我们有两种可行的方法: (1) DAG 图中的后门准则;

(2) 将 DAG 图转化为 SWIG 图。虽然使用 SWIG 图更为直观, 但它也需要更多的前期准备。我们先介绍第一种方法, 然后在第 7.5 小节介绍 SWIG 图方法。

(Pearl (1995, 2000) 首先提出了用后门标准非参数化地识别因果效应)

L 是图中一系列变量的集合, 且这个集合不包含 A 的下游变量。如果控制 L 之后, A 和 Y 之间的所有后门路径都被阻断, 那我们说 L 满足后门准则。在忠实性和知识点 7.1 所给出的假设之下, 有界互换性 $Y^a \perp\!\!\!\perp A|L$ 成立的充要条件是 L 满足后门准则 (证明将在 SWIG 图中给出)。因而, 我们只需遍历所有不含 A 下游变量的变量集合, 就能回答上述问题 (事实上, 现在已有算法大大减少了我们所需要遍历的集合个数)。

86 现在让我们正式介绍一下后门准则 (即互换性) 和混杂的关系。以下两种情境满足后门准则:

1. 治疗和结局没有共同诱因。在图 6.2 中, 治疗和结局没有共同诱因, 因而也就没有后门路径需要被阻断。此时满足后门准则的变量集合为空集, 即不存在混杂。
2. 治疗和结局有共同诱因, 但是一个不含 A 下游变量的变量集合 L 足够阻断所有的后门路径。在图 7.1 中, 满足后门准则的变量集合是 L 。此时存在混杂, 但是不存在需要调整未知变量 (这也不可能实现) 的残余混杂。我们将此称为 “不存在未知混杂”。 L 被称为 “混杂调整的充分集合”。

第一个情境描述了边缘随机试验, 其中的治疗分配是完全随机的, 不取决于任何变量。第二个情境描述了条件随机试验, 其中治疗分配的概率取决于变量 L 。如果 L 是结局 Y 的诱因 (如图 7.1) 或 Y (未知或已知) 诱因的下游变量 (如图 7.2), 那 L 就是一个混杂变量。此时控制 L 就能阻断所有后门路径, 也就保证有界互换性 $Y^a \perp\!\!\!\perp A|L$ 成立。我们也希望第二种情境能用以描述我们的观察性研究。

87 后门准则并不能告诉我们混杂的方向与强度。很可能未阻断的后门路径所引起的相关性很弱 (比如 L 对 A 和 Y 的效果都不强), 那引入的偏移就很小。在实践中, 我们经常要考虑偏移的方向与大小。

7.3 混杂与后门准则

现在我们来讲述应用后门准则去判定 A 对 Y 的因果效应是否可识别, 如果可以, 需要控制哪些变量以保证有界互换性。谨记本章因果性 DAG 图中的节点代表的都是零误差测量的变量。

在图 7.1 中, 治疗 A 和结局 Y 有一个共同诱因 L , 也即 A 和 L 之间有一条通过 L 的开放后门路径。但我们控制住 L 的话, 我们就能阻断这条后门路径。因而, 如果研究者收集了每个个体的 L 信息, 那在给定 L 的情况下, 也就不存在未知或未测的混杂了。

在图 7.2 中, 治疗 A 和结局 Y 有一个未知的共同诱因 U , 也即 A 和 L 之间有一条通过 U 的开放后门路径。理论上, 我们可以通过控制 U 阻断这条后门路径, 从而消除混杂。但与 L 、 A 和 Y

不同的是, U 是一个未知或未测量的变量。不过此时我们也可以通过控制 L 阻断这条后门路径。因而, 那在给定 L 的情况下, 这一情境不存在未知或未测的混杂。

在图 7.3 中, 治疗 A 和结局 Y 有一个未知的共同诱因 U , 我们同样可以通过控制 L 阻断这条后门路径。因而, 那在给定 L 的情况下, 这一情境不存在未知或未测的混杂。

现在来考虑图 7.4。在这幅因果图中, A 和 Y 之间没有共同诱因, 因而也就没有混杂。 A 和 Y 之间通过 L 的后门路径 ($A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$) 是被阻断的, 这是因为这条路径上的 L 是一个对撞变量。因而 A 和 Y 之间的相关性全部来源于 A 对 Y 的因果效应, 此时相关性就是因果性。比如, 假设 A 代表体育锻炼, Y 代表宫颈癌, U_1 代表癌前突变, L 代表癌前突变的诊断测试 (宫颈刮片检查), U_2 代表对自身健康的关心程度, 这一情境就能用图 7.4 表示。此时计算体育锻炼 A 对宫颈癌 Y 的因果效应, 就不需要调整检测 L , 即 A 对 Y 的效应没有混杂。

88 在我们上述四个例子中, 拥有 L 、 A 和 Y 的数据就足以让我们识别 A 对 Y 的因果效应。如果仅有 A 和 Y 的数据不足以识别因果效应, 那我们就把 L 定义为混杂变量 (也即存在结构性混杂)。如果仅有 A 和 Y 的数据就足够识别因果效应, 那 L 就不是混杂变量。这一定义与 L 下的有界互换性等价。

(图 7.1 至 7.4 中混杂的一个非正式定义: 混杂变量是用来调整混杂的任意变量。这一定义不是自说自话, 因为我们之前就给出了混杂的定义。与此类似的定义有: 钢琴家是弹奏钢琴的人。)

因而, 在图 7.1 至图 7.3 中, L 是一个混杂变量, 因为我们需要使用 $\sum_l \Pr[Y=1 | A=a, L=l] \Pr[L=l]$ 来计算 $\Pr[Y^a = 1]$ 。在图 7.2 和图 7.3 中, L 不是 A 和 Y 的共同诱因, 但我们依然说 L 是一个混杂变量, 因为我们需要用 L 去阻断途经未知变量 U 的后门路径。在图 7.4 中, L 不是一个混杂变量, 因为 $\Pr[Y^a = 1] = \Pr[Y=1 | A=a]$ 。

89 有趣的是, 在图 7.4 中, L 下的有界互换性并不成立, 因而 $\Pr[Y^a = 1 | L=l]$ 并不等于 $\Pr[Y=1 | A=a, L=l]$, 即 L 每一分层中的条件因果效应时不可识别的。更进一步, 通过标准化调整 L 得到的 $\sum_l \Pr[Y=1 | A=a, L=l] \Pr[L=l]$ 是对 $\Pr[Y^a = 1]$ 的有偏估计。我们可以使用因果图来理解这一结论: 控制了对撞变量 L 会打开 A 和 Y 之间原本被对撞变量 L 阻断的后门路径 ($A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$), 因而调整 L 会引入偏移。如果调整了 L , 相关性不再是因果性, A 和 Y 之间的相关性就会是 A 对 Y 的因果效应, 以及 A 和 Y 之间后门路径所引起的相关性, 这两者的混合物。这是我们迄今见到的第一个无条件互换性成立, 但是有界互换性不成立的例子: 因

果效应均值是可识别的，但是 L 每一分层中的条件因果效应却是不可识别的。我们将这种每一分层中因果效应的偏移称作选择偏移，这是因为我们选择性地在 L 的某一分层中计算因果效应，而 L 是相互独立却又与 A 和 Y 分别相关的两个变量 U_1 和 U_2 的共同诱因（参见第八章）。

（可以识别无条件的因果效应，但不能识别条件因果效应，这一可能在 Greenland 和 Robins (1986) 所合著的论文中有所论述，并且这一论述未借助因果图。图 7.4 的条件偏移首次由 Greenland (1999) 进行论述，并被称为 M 偏移 (Greenland, 2003)，因为这一结构形似字母 M 。）

（如果 U_1 是 U_2 的诱因，或者 U_2 是 U_1 的诱因，或者一个未知变量 U_3 是两者的共同诱因，那么 A 和 Y 就有共同诱因，此时无条件互换性不再成立， L 下的有界互换性也不成立。）

图 7.5 是图 7.4 的一个变形。区别在于，图 7.5 中，有一个箭头从 L 指向 A 。因为 U_1 是 A 和 Y 的共同诱因，因而这个新箭头的存在创造了一条新的开放后门路径 $A \leftarrow L \leftarrow U_1 \rightarrow Y$ ，从而存在混杂。控制 L 能阻断这条新的后门路径，但是也会开启另一条 L 作为对撞变量的后门路径。

（对撞变量的定义取决于具体路径：在路径 $A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$ 中， L 是对撞变量，但在路径 $A \leftarrow L \leftarrow U_1 \rightarrow Y$ 中， L 不是对撞变量。）

因而，在图 7.5 中，阻断混杂路径会同时开启选择偏移路径，如何消除偏移也就是一件棘手的事情。此时既没有无条件互换性，也没有 L 下的有界互换性。消除图 7.5 中偏移的可能办法有：（1）测量并控制 U_1 和 L 之间，或 U_1 和 Y 之间的变量 L_1 ；（2）测量并控制 U_2 和 L 之间，或 U_2 和 A 之间的变量 L_2 。在第一种情况中，我们有 L_1 下的有界互换性。在第二种情况中，我们有 L 和 L_2 下的有界互换性。比如，图 7.6 就包含了 U_1 和 Y 之间的变量 L_1 ，以及 U_2 和 A 之间的变量 L_2 。如何利用图 7.6 中的变量来识别因果效应，请参见精讲点 7.2。

本小节描绘了互换性缺失的两种结构性原因，这两种原因都来自治疗和结局间开放的后门路径。第一种原因是治疗和结局有共同诱因。第二种原因是控制了治疗和结局的共同后果。我们将第一种称为混杂，将第二种称为选择偏移。另一种结构性地定义混杂的方式是：由 A 和 Y 之间开放的后门路径所引起的偏移被称为混杂。这一定义将图 7.4 中控制 L 后带入的偏移称为混杂而非选择偏移，除此之外这一定义与我们之前所说的定义相同。这一新定义可以等价地表述为：混杂是可以通过对治疗 A 进行随机分组而消除的系统性偏移。图 7.4 中因控制 L 引起的偏移是不会在一项随机试验中出现的，这是因为随机试验中的随机分配会保证 A 和 L 不存在一个共同诱因 U_1 ，因而控制 L 不会开启一条后门路径。

这两个定义的一个有趣区别是: 治疗和结局存在共同诱因 (即结构性定义中的混杂) 基本上是现实研究人群中的一个必然事实, 这与我们所选择的分析方法无关。另一方面, 如果将混杂定义为能通过随机分组消除的偏移, 那混杂是否存在则由我们的分析方法决定。在图 7.4 中, 如果我们不调整 L , 那就没有混杂。但调整了 L , 就会引入混杂。

然而, 选择哪一种混杂的定义方式只是个人的喜好问题, 和实际上的意义无关, 这是因为我们因果效应的可确定性是建立在互换性的基础之上, 而不是混杂的定义之上。下一章我们将更详尽介绍混杂和选择偏移的区别。

7.4 混杂与混杂变量

在上一小节, 我们描述了如何用因果图去确定混杂存在与否, 如果存在, 是否已测的变量集合 L 足以用来调整混杂。这一过程需要事先用我们的专业知识描绘一幅 DAG 图, 图中需要包含治疗 A 和结局 Y 的所有共同诱因, 不管是已测的还是未测或未知的。一旦有了这幅因果图, 我们就能用后门准则去判定需要调整哪些变量。

与此相比, 处理混杂的传统方法大多依赖于观测到的相关性, 而非事先知道的因果知识。传统方法首先将符合某些相关性条件的变量标记为混杂变量, 然后再在分析中调整这些混杂变量。当调整后的估计值和未调整之前的估计值有差别, 则传统方法认为有混杂存在。

(其实, 理论上而言, 我们不需要知道各变量之间的因果性结构, 只需要知道能保证有界互换性的变量集合就行。然而, 知道各变量间的结构性知识, 从而描绘出因果图, 可能是用来知道哪些变量能保证有界互换性的最佳方式)

在传统方法中, 满足以下三个条件的变量被定义为混杂变量: (1) 和治疗相关; (2) 在治疗的每一分层中和结局相关; (3) 不在治疗和结局的因果路径上。然而, 传统定义可能带来不合理的调整。我们将用图 7.1 至图 7.4 来说明为什么。

在图 7.1 中, 变量 L 和治疗 A 相关 (因为其对 A 有因果效应), 并且在 A 的每一分层中和结局 Y 相关 (因为其对 Y 有因果效应), 并且 L 不在治疗 A 对结局 Y 的因果路径上。在图 7.2 中, 变量 L 和治疗 A 相关 (因为其对 A 有因果效应), 并且在 A 的每一分层中和结局 Y 相关 (因为其和 Y 有共同诱因 U), 并且 L 不在治疗 A 对结局 Y 的因果路径上。在图 7.3 中, 变量 L 和治疗 A 相关 (因为其和 A 有共同诱因 U), 并且在 A 的每一分层中和结局 Y 相关 (因为其对 Y 有因果效应), 并且 L 不在治疗 A 对结局 Y 的因果路径上。

因而, 根据传统定义, L 在图 7.1 至图 7.3 都是混杂变量, 都需要被调整。这也是我们根据后门准则得到的结论。在图 7.1 至图 7.3 中, 传统方法和因果图中的后门准则没有什么区别。

我们上一小节讨论了图 7.4, 根据我们的讨论, 图 7.4 的 L 不是一个混杂变量, 也不需要被调整。然而 L 满足传统方法中混杂变量定义的三个条件: L 和治疗相关 (其和 A 有共同诱因 U_2), 在治疗的每一分层中 L 和结局相关 (其和 Y 有共同诱因 U_1), L 不在治疗和结局的因果路径上。因而, 根据传统定义, L 是一个混杂变量, 所以需要被调整, 即使图 7.4 中完全不存在混杂。不过, 我们已经讨论过, 调整 L 会引起选择偏移, 从而导致对因果效应的有偏估计。图 7.7 给出了另一个例子, 其中 L 满足传统定义的三个标准, 但是调整 L 会引入选择偏移。

这些例子说明了相关性标准或者基于统计的标准不足以用来描述混杂。纯粹基于统计考虑的定义只能导致如图 7.4 和图 7.7 的错误: 即使没有结构性混杂的时候依然调整“混杂变量”。为了解决诸如图 7.4 中的问题, 传统方法进行了一些修正, 其中混杂变量第二条标准“在治疗的每一分层中和结局相关”被替换为“是结局的诱因”。这一修正就排除了图 7.4 的 L , 但也导致了新的问题。根据这一修正, 图 7.2 的 L 不再是混杂变量。参见知识点 7.2。

传统方法会误导研究者, 让他们在不应该调整的时候调整某些变量。这一问题的根源在于传统方法没有利用专业知识去考虑混杂的真正来源, 并强制调整被标记为“混杂变量”的变量。如果调整后的效应估计不同于未调整时的效应估计, 则传统方法宣称存在混杂。然而, 调整了非混杂变量而带来的选择偏移、使用不可伸缩的效应量度 (参见精讲点 4.3), 也能改变效应估计。由于这些问题, 根据效应估计的改变去定义混杂这一观点已经被摈弃。

与之相比, 本章介绍的结构性方法, 先确定混杂的来源 (治疗和结局的共同诱因), 然后再依据此去确定调整混杂的充分变量集合。

在结构性方法中, 一个变量是否属于充分集合, 取决于充分集合中的其他变量。比如, 在图 7.2 和图 7.3 中, 变量 U 是未测的, 所以我们需要包含 L 的变量集合去阻断后门路径。因而我们可以说 L 是混杂变量。然而, 如果 U 是已测的, 我们就可以调整 U 来阻断后门路径。此时在给定 U 时, L 就不再是混杂变量。在 DAG 图中, 混杂是一个绝对的概念, 但是混杂变量是一个相对的概念。

(VanderWeele 和 Shpitser (2013) 给出了混杂变量的另一种正式定义)

混杂的结构性方法强调了对观察性数据做出因果推断需要依赖于已有的专业知识。DAG 图形象地表达了这些专业知识, 并且也将研究者的假设也包含在因果性结构这种。当然, 没有人能评判一名研究者给出的 DAG 图是否是正确的, 因而, 也就不能保证研究者所选取的调整变量就能一定消除混杂或不引入选择偏移。不过, 混杂的结构性方法依然有两个重要优点。首先是其排除了研究者假设和行动之间的不一致性。比如, 如果一名研究者假设图 7.4 是一副正确无误的因果

图, 此时没有混杂, 不管其他非结构性定义怎么说, 研究者都没有必要调整变量 L 。其次是研究者的假设能明白无误地体现在因果图中, 因而也就能接受其他研究人员的评判检验。

93 7.5 单一世界干涉图

互换性能用 DAG 图中治疗 A 和结局 Y 之间不存在后门路径、从而没有其余相关性来表示。如果 A 和 Y 之间存在相关性, 互换性也就不再成立。第七至第九章论述了互换性的缺失在 DAG 图中的不同表现形式。

无条件互换性 $Y^a \perp\!\!\!\perp A$ 和后门准则等价, 这一命题看起来似乎有些神奇。因为因果图中并不包含反事实变量, 所以反事实框架中的独立性和后门路径的缺失似乎没有什么联系。因为我们可以从有向分离的方法来检验因果图中的独立性, 自然而然, 我们也就想将反事实变量融入到因果图中, 从而可以直接从图中读出无条件互换性和有界互换性。

这种新类型的图, 被称为单一世界干涉图 (SWIG 图), 其能用图像的方法表示反事实情境。SWIG 图能描绘在假想世界中, 如果所有个体接受的治疗取值都是 a 的时候, 我们能观测到的其他变量以及它们之间的因果关系。也即, SWIG 图, 是描绘“单一干涉”下反事实“世界”的图。与之相比, 标准因果图中的变量都代表的是现实世界中的变量。因而, 将一副标准因果图在某一干涉下进行转化, 就能得到 SWIG 图。以下例子论述了如何进行转化。

(Robins 和 Richardson (2013) 论述了 SWIG 图能克服过往双子因果图 (Balke 和 Pearl, 1994) 所固有的一些缺陷)

94 假设图 7.2 代表了一项观察性研究。如果其中所有个体接受的治疗取值都被固定为 a , 此时的反事实世界就能用图 7.9 的 SWIG 图表示。

在这个 SWIG 图中, 代表治疗的节点被分为左右两部分, 各自代表不同的变量。右侧表示在于干涉之下治疗取值为 a , 并且继承了原 DAG 图中所有从 A 出发的箭头。左侧表示现实世界中如果没有干涉我们能观测到的治疗 A , 也即治疗的自然取值。左侧继承了原 DAG 图中所有指向 A 的箭头。现实世界中其他变量对治疗 A 的因果性影响, 在反事实世界中依然存在。注意到 A 和 a 之间没有箭头, 这是因为 a 是反事实世界中所有个体都会有的取值, 也即在反事实世界中这是一个常数。

即使在所有治疗取值都是 a 的反事实世界中我们不能去测量 A , 我们也依然假设治疗 A 的自然取值是良定的。虽然在某些情境中 A 是可测量的, 比如, 最近有研究显示, 脑电图成像能在个人意识到自己的决策之前 0.5 秒告诉研究者这个人要做出什么样的决策。如果这样的话, A 就能通过脑电图成像进行测量, 这样就还有 0.5 秒进行干涉, 并将治疗取值固定为 a 。

在 SWIG 图中, 结局是 Y^a , 表示 Y 在反事实世界中的取值。因为其余变量在时间上先于 A , 所以它们不会受到干涉的影响, 取值也就和现实世界中我们所观测到的一样, 也即它们不是反事实变量。实际上, 任何 A 的非下游变量都不应被视为反事实变量, 这是因为在忠实性假设下, 治疗对任何非下游变量都没有因果效应。在我们的因果模型中, 因为 Y^a 和 A 之间的所有路径在控制 L 后都能被阻断, 即给定 L 的情况下 Y^a 和 A 有向分离, 所以有界互换性 $Y^a \perp\!\!\!\perp A|L$ 成立。

(在 FFRCIST 模型中, 有向分离也意味着 SWIG 图中统计上的独立性。)

让我们再来看一下图 7.4 和图 7.10 的 SWIG 图。因为在 SWIG 图中 Y^a 和 A 之间的所有路径都被阻断 (就算没有控制 L), 所以边缘互换性 $Y^a \perp\!\!\!\perp A$ 成立。与之相比, 有界互换性 $Y^a \perp\!\!\!\perp A|L$ 并不成立, 这是因为在 SWIG 图中, L 是路径 $A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y^a$ 中的对撞变量, 控制 L 会打开原本被阻断的路径。这也是为什么 $A-Y$ 的边缘相关性也是因果性的, 但是给定 L 后 $A-Y$ 相关性则不是因果性的原因, 因而调整 L 只会导致偏移。这个例子说明了 SWIG 图能将反事实框架和因果图统一在一起。实际上, 在 SWIG 图中, 以下命题显而易见: 在给定 L 的情况下, 当且仅当 L 是 A 的非下游变量且阻断了 A 和 Y 之间的所有后门路径, Y^a 和 A 有向分离。

7.6 混杂调整

95

在没有随机分组的情况下, 因果推断只能依赖于一个不可检验的假设, 即我们所测的变量, 是混杂调整的一个充分集合。换句话说, 我们已有的 A 的非下游变量足以阻断 A 和 Y 之间的所有后门路径。这一假设与 L 下的有界互换性等价, 因而我们就可以用标准化或逆概率加权等方法来计算人群中的因果效应均值。不过, 正如第 4.6 节所讨论的, 标准化和逆概率加权并不是观察性研究中仅有的能用来调整混杂的方法。调整混杂变量 L 的方法可以分为以下两个大类:

- G-方法: 标准化, 逆概率加权, 和 G-估算。这些方法中的 “G” 代表 generalized, 一般化的意思。这些方法主要是利用 L 下的有界互换性去估算整个人群中或某个子群体中 A 对 Y 的因果效应均值。在我们心脏移植的例子中, 我们在 2.4 小节 (标准化) 和 2.5 小节 (逆概率加权) 使用了 G-方法来调整重症 L 带来的混杂。本书第二部分将会讲述基于模型的 G-方法及其延伸: 参数化的 G-公式 (标准化), 边缘结构模型中的逆概率加权, 和嵌入式结构模型中的 G-估算。
- 基于分层的方法: 分层分析 (包括限制) 和匹配。这些方法主要是利用 L 下的有界互换性去估算 L 的各子群体中 A 对 Y 的因果效应。我们在 4.4 小节 (分层分析) 和 4.5

小节(匹配)使用了基于分层的方法来调整重症 L 带来的混杂。本书第二部分将会讲述基于分层的方法及其延伸: 传统的结局回归方法。

(分层分析和匹配的一个常见变种是倾向性评分方法。这一方法用每个个体接受治疗的估计概率 $\text{Pr}[A=1|L]$ 替代变量 L (Rosenbaum 和 Rubin, 1982)。参见第十五章。)

G-方法能模拟一个虚拟人群, 这个人群中所有涉及变量 L 的后门路径都不存在, 从而也就能估计 $A-Y$ 之间的相关性。比如, 逆概率加权模拟的虚拟人群中, 治疗 A 和所有已测的混杂变量 L 相互独立, 也即“删除”了从 L 到 A 的箭头。与之相比, 基于分层的方法并不会删除 L 到 A 的箭头, 而只是在每个子群体中计算条件因果效应, 在图中用一个方框把变量框住进行表示。“删除”箭头的优势将在本书第三部分重点讨论。在有时异治疗、因而也有时异混杂变量的情境中, G-方法是用来调整混杂的最佳方法, 因为此时基于分层的方法可能会造成选择偏移。第二十章将讨论基于分层的方法所带来的偏移。

以上所有方法都需要 L 下的有界互换性。然而, 混杂有时也能用不需要有界互换性的方法进行调整。比如双重差分法(参见知识点 7.3)、工具变量估算法(参见第十六章)、前门准则(参见知识点 7.4)等方法。不过这些方法需要其他假设, 这些假设同有界互换性一样不可被验证。因此, 在实践中, 这些方法效应估计的有效性不能得到绝对的保证。同时, 这些方法也不能处理涉及时异性治疗的情境。因此, 许多研究问题都不会考虑这些方法。对于固定时间的治疗而言, 变量调整方法的选择取决于选择什么样的假设。虽然这些假设都是不可验证的, 但是某些假设在某些情境中更可能成立。

(严格来说, G-估算所需假设是: 给定 L 时未知混杂大小是已知的。有界互换性, 即不存在未知混杂, 是其中的一个特例。因而, G-估算所需要的假设要稍弱一些。参见第十四章。)

在许多观察性研究中, 保证有界互换性成立可能是不实际的, 但是利用我们对因果结构的专业知识能尽可能地达到这一目标(参见第 3.2 小节)。因此, 在观察性研究中, 研究人员尽可能多地测量各种变量 L (这些变量是治疗的非下游变量), 希望以此保证治疗组和非治疗组的有界互换性。研究者的想法是, 即使共同诱因(混杂)可能存在, 已测的变量 L 依然足以阻断所有后门路径(即不存在未知混杂)。然而, 没有人能保证这一做法一定能成功, 这也使得观察性研究中的因果推断是一件冒险的事业。

此外, 专业知识也能用以避免调整某些可能引入偏移的变量。至少, 研究者不应该调整受到治疗或结局影响的变量。当然, 一位深思熟虑的研究者, 在面对同一问题的两个不同因果结构时, 依然可能认为这两个因果结构时同样具有说服力。在这种情况下, 研究者需要进行多次分

析, 并在每次分析中清晰说出自己的假设, 从而保证效应估计的有效性。不幸的是, 没有人能确定我们考虑的因果结构中是否有正确的那一个。在观察性研究中, 不确定性总是在所难免。

(Hernan (2002) 在其所著论文中详细讨论了利用专业知识评估混杂的实践方法)

观察性研究中的混杂给科学研究带来了一定的影响。假设你开展了一项观察性研究来探索心脏移植 A 对死亡 Y 的因果效应, 并假设除了重症 L 之外没有其他未知混杂。一名批评者认为“因为还存在其他混杂, 所以这项观察性研究中的因果推断是靠不住的”。这名批评者的陈述是逻辑性的, 但不是科学性的。因为所有的观察性研究都有混杂, 因而我们的研究也有混杂。如果这名批评者是希望指出你研究中的缺陷, 那么遗憾的是他并没有做到这一点。他的批评不具有任何价值, 因为他只是指出了观察性研究的一个特征, 而这个特征在你开展研究之前所有人都已知晓。

97 一个建设性的批评需要的是一场科学性的对话。比如, 这名研究者引用了其他试验研究或者观察研究, 指出你的研究和其他研究存在矛盾之处, 或者他的批评是诸如“因为吸烟可能是治疗和结局的共同诱因, 但在你的分析中没有被调整, 所以你的结论是不正确的”之类。后者对你“没有其他未知变量”这一假设提出了质疑。证明的重任又再一次回到了你的肩上。此时你需要尝试在分析中调整吸烟。

虽然以上的讨论仅局限于混杂引起的偏移, 但是不存在选择偏移和测量偏移也是有效因果推断的基础之一。然而, 与混杂不同, 随机试验和观察性研究中都可能出现这些偏移。本章介绍了混杂, 下一章将介绍互换性缺失的另一个可能原因: 选择偏移。

第七章精讲点和知识点

精讲点 7.1: 混杂的强度和方向 (原书第 88 页)

你开展了一项研究去探索心脏移植 A 对死亡 Y 的因果效应, 并假设没有未知混杂。一名深思熟虑的批评者会说: “从观察性研究中得到的因果推断可能是不正确的, 因为你没有考虑吸烟带来的混杂”。于是, 一个你要面对的重要问题是: 吸烟带来的混杂是加强了还是减弱了你的效应估计。假设你得到的风险比是 0.6。这位批评者认为吸烟是结局和治疗的一个共同诱因。因为治疗组 ($A=1$) 中有更少的吸烟者 ($L=1$), 所以有人会认为就算 A 对 Y 因果效应的零假设成立, $A=1$ 的人群中也会有更低的死亡率。调整吸烟会使我们的效应估计更接近 1。换句话中, 不调整吸烟会夸大心脏移植 A 的因果效应。

我们可以用因果图中箭头的符号去预测混杂的方向。在图 7.1 中, 假设 L 、 A 和 Y 都是二分变量。如果 L 对 A 的因果效应为正 (即与 $L=0$ 的人群相比, $L=1$ 的人群中 $A=1$ 的概率更高), 那我们就给 L 到 A 的箭头标注一个正号。如果 L 对 A 的因果效应为负 (即与 $L=0$ 的人相

比, $L=1$ 的人群中 $A=1$ 的概率更低), 那我们就给 L 到 A 的箭头标注一个负号。我们也可以同样给 L 到 Y 的箭头标注正负号。如果两个箭头都是正号或者负号, 那混杂就是正向的, 也就意味着不调整 L 会夸大我们的效应估计。如果是一正一负, 那么混杂就是负向的, 也就意味着不调整 L 会降低我们的效应估计。不过, 这一简单方法不适用于复杂的因果图以及非二分变量的情况。更多细节可参见 VanderWeele, Hernan 和 Robins (2008) 所著论文。

尽管我们能确定混杂的方向, 但还有另一个问题, 即混杂的强度。在实践中, 如果偏移的强度不是很大, 对结论不会造成太大影响, 那不管是正向的还是负向的, 一般情况下是可以忽略不计的。只有当混杂变量和治疗之间的相关性很强, 或混杂变量和结局之间的相关性(在治疗的分层中) 很强的时候, 混杂的影响才会较强。对于离散型混杂变量, 偏移的强度取决于混杂变量的分布 (Cornfield 等人, 1959; Walker, 1991)。如果混杂变量未知, 研究者只能猜测偏移的强度是多大。我们可以通过敏感性分析(即在不同的偏移强度假设下进行分析) 进行有根据的猜测, 这一方法能帮助我们量化一个合理偏移的最大可能强度。关于敏感性分析和未知混杂, 请参见 Greenland (1996a), Robins, Rotnitzky 和 Scharfstein (1999), Greenland 和 Lash (2008), VanderWeele 和 Arah (2011) 等人所著的论文。

精讲点 7.2: 条件效应和无条件效应的可识别性 (原书第 90 页)

在任何因果图中, 可识别的因果效应除了治疗和结局, 还取决于其他已测的变量。以图 7.6 为例, 如果我们测量了 L_2 (但没有测 L 和 L_1), 但给定 L_2 的情况下, 无条件互换性和有界互换性都不成立, 因而也就不能识别任何因果效应。如果我们测量了 L_2 和 L , 我们就有给定 L_2 和 L 下的有界互换性, 因而们就能识别:

- L_2 和 L 联合分层中的条件效应: $E[Y | A = a, L = l, L_2 = l_2]$ 。
- 无条件效应: $\sum_{l,l_2} E[Y | A = a, L = l, L_2 = l_2] \Pr[L = l, L_2 = l_2]$ 。
- L 分层中的条件效应: $\sum_{l_2} E[Y | A = a, L = l, L_2 = l_2] \Pr[L_2 = l_2 | L = l]$
- L_2 分层中的条件效应: $\sum_l E[Y | A = a, L = l, L_2 = l_2] \Pr[L = l | L_2 = l_2]$

如果我们仅测量了 L_1 , 那我们就有给定 L_1 下的有界互换性, 也就能识别 L_1 每一分层中的条件效应和无条件效应。如果我们测量了 L_1 和 L , 我们就能识别 L_1 和 L 联合分层中的条件效应。如果我们测量了 L 、 L_1 以及 L_2 , 那我们就能识别这三个变量联合分层中的条件效应。

精讲点 7.3: 混杂变量的替代 (原书第 92 页)

在图 7.8 的因果图中, A 和 Y 有一个未测的共同诱因 U , 所以 A 对 Y 的因果效应中存在混杂。已测的变量 L 是 U 的一个代理变量, 或称替代变量。比如, 未测的社会经济地位变量 U (这个变量经常定义不清) 会混杂体育锻炼 A 对心血管疾病 Y 的因果效应。收入 L 是社会经济地位的一个替代变量。我们是否应该调整变量 L ? 一方面, 我们可以说 L 不是一个混杂变量, 因为它并不在 A 和 Y 的后门路径之中。另一方面, 已测的 L 和 U 相关, 调整 L 能够间接调整变量 U , 从而调整混杂。极端一些, 如果 L 和 U 完美关联, 那么调整 U 和调整 L 就没有任何区别。实际上, 如果 L 是二分的, 且是 U 的无差别错误归类形式 (详见第九章), 在某些弱假设下, 控制 L 能部分阻断后门路径 $A \leftarrow U \rightarrow Y$ (参见 Greenland (1980) 与 Ogburn 和 VanderWeele (2012) 的论文)。因而在大多数情况下, 我们倾向于调整 L 。

我们将调整后能减少混杂偏移, 但是又不在后门路径 (因而也就不能完全消除混杂) 的变量称为混杂变量的替代。调整混杂的一个可能方法是尽可能地多测量混杂变量的替代, 然后再调整这些替代变量。参见第十八章。

精讲点 7.4: 混杂变量不能是治疗的下游变量, 但是可以发生在治疗之后 (原书第 95 页)

让我们考虑图 7.11 的结构。 L 是 A 的下游变量, 并阻断 A 和 Y 之间的后门路径。与图 7.4 和图 7.7 不同, 此时控制 L 并不会引起选择偏移, 因为没有对撞路径被打开。相反, 因为 A 对 Y 的因果效应需要通过中介变量 L , 控制 L 将会完全阻断这条因果路径。这个例子说明了当 L 是 A 的下游变量时, 调整变量 L 会阻断后门路径但是不会消除偏移。

因为有界互换性 $Y^a \perp\!\!\!\perp A|L$ 意味着调整 L 能消除所有的偏移, 因而在上述例子中, 有界互换性不成立, 从而治疗的因果效应均值 $E[Y^{a=1}] - E[Y^{a=0}]$ 不可被识别。这一例子可以用图 7.12 中的 SWIG 图进行说明。图 7.12 描绘了 A 的取值被固定为 a 时的反事实世界, 在其中, 变量 L 被替换为反事实变量 L^a , 也即所有个体在接受治疗 a 时 L 的取值。因为 L^a 能阻断所有从 Y^a 到 A 的路径, 因而我们可以认为 $Y^a \perp\!\!\!\perp A|L^a$ 成立, 但是我们不能确定 $Y^a \perp\!\!\!\perp A|L$ 是否成立, 因为 L 不在图中。(在 FFRCISTG 模型中, 如果某个独立性不能从 SWIG 图中得到, 那我们就不能假设它成立。) 因而, 我们不能保证从 L 、 A 和 Y 的数据中识别 $E[Y^{a=1}] - E[Y^{a=0}]$ 。

这个问题是因为 L 是 A 的下游变量, 而不是因为 L 发生在 A 之后。如果在图 7.11 中不存在从 A 到 Y 的箭头, 那么 L 就不是 A 的下游变量, 也不能阻断所有后门路径。同理, 在图 7.2 的 SWIG 图中, 因为 A 不再是 L 的诱因, 所以我们能用 L 替代 L^a (注意, 此时 Y^a 和 A 在控制 L 的时

候有向分离)。因而即使 L 发生在 A 之后, 调整 L 也能消除所有的偏移。此时重要的是因果图的拓扑结构, 而非各变量的时序关系。Rosenbaum (1984) 和 Robins (1986) 用无图的方式讨论了发生在治疗之后的混杂变量控制。

知识点 7.1: 有界互换性是否意味着后门准则? (原书第 86 页)

L 满足后门准则意味着 L 下的有界互换性成立, 甚至在忠实性假设不成立时, 这一命题依然成立。在正文中, 我们讲述了如果忠实性成立, 这一命题的逆命题成立, 即 L 下的有界互换性成立意味着 L 满足后门准则。这一说法在 FFRCISTG 模型下为真 (参见知识点 6.2)。与之相比, 在 NPSEM-IE 模型中, 即使后门准则不成立, 有界互换性也依然可能成立, 正如一幅仅有节点 A 、 L 、 Y , 以及箭头 $A \rightarrow Y$ 和 $A \rightarrow L$ 的因果 DAG 图。在本书中, 除非特别说明, 否则我们假设 FFRCISTG 模型和忠实性成立。

这一区别在于 NPSEM-IE 模型假设了不同反事实世界之间的独立性, 这一点与 FFRCISTG 模型不同。然而不同反事实世界之间的独立性是一个不可验证的假设, 即使用随机试验也不能验证。这也是为什么 Robins (1986, 1987) 在他的 FFRCISTG 模型中不假设不同反事实世界之间相互独立的一个原因。更多讨论请参见第二十二章。

知识点 7.2: 修补混杂的传统定义 (原书第 93 页)

图 7.4 和图 7.7 中的例子说明了混杂和混杂变量的传统定义会误导研究者, 从而使某些变量调整不仅多余, 而且有害。传统定义的失灵在于其依赖两条不正确的基于统计的标准 (第一、二条) 和一条不正确的因果性标准 (第三条)。为了修补传统定义, 研究者需要做两件事:

1. 将第三条标准替换为: 存在变量 L 和 U , 使得在 L 和 U 的联合分层中有界互换性 $Y^a \perp\!\!\!\perp A | L, U$ 成立。这一新标准较之前的强, 这是因为其意味着 L 不在 A 和 Y 的因果路径上, 且 $E[Y^a | L = l, U = u]$ 可以由 $E[Y | L = l, U = u, A = a]$ 进行明确。
2. 将第一、第二条标准替换为: U 能分解成两个不想交的自己 U_1 和 U_2 (即 $U = U_1 \cup U_2$ 且 $U_1 \cap U_2$ 为空集), 使得 (1) U_1 在 L 的分层中和 A 不相关, (2) U_2 和 Y 在 A 、 L 和 U_1 的联合分层中不相关。 U_1 中的变量可能和 U_2 中的变量相关。 U_1 表示 U 当中和治疗不相关的最大子集。

如果这两个新条件满足, 我们说 U 不是给定 L 下的混杂变量。这些条件有 Robins (1997) 提出, 并被 Greenland, Pearl 和 Robins (1999) 进一步讨论。这两个条件克服了图 7.4 和图 7.7

带来的缺陷, 因为新条件允许研究者将部分变量视作非混杂变量 (Robins, 1997)。将新条件应用到图 7.4 中, 我们会发现不存在混杂。

知识点 7.3: 双重差分和阴性结局对照 (原书第 97 页)

假设我们想计算阿司匹林 A (1 表示服用, 0 表示没有服用) 对血压 Y 的因果效应均值, 但是 A 和 Y 之间有未测的共同诱因 U , 比如心脏病史。如此一来, 我们就不能通过标准化或者逆概率加权计算因果效应。但是在其他假设下, 我们依然有方法对未测混杂进行调整, 比如阴性结局对照 (也被称为“安慰剂检验”)。

假设我们在治疗之前就测量了研究人群中每个个体的结局。我们将治疗之前的结局 C 称为阴性结局对照。如图 7.13 所描绘, U 是 Y 和 C 的共同诱因, 治疗 A 明显不是治疗前结局 C 的诱因。虽然 A 对 C 的因果效应为零, 但是因为混杂变量 U 的存在, $E[C|A=1]-E[C|A=0]$ 不等于零。实际上, $E[C|A=1]-E[C|A=0]$ 衡量的是加法尺度上 A 对 C 的因果效应中混杂的大小。如果 A 对 C 效应中的混杂, 和 A 对 Y 效应中的混杂在加法尺度上强度大小一样, 那我们就能计算治疗组中 A 对 Y 的效应。具体而言, 在混杂相等 (可以表述为

$$E[Y^0|A=1]-E[Y^0|A=0]=E[C|A=1]-E[C|A=0] \text{ 假设下, 因果效应为}$$

$$E[Y^1-Y^0|A=1]=(E[Y|A=1]-E[Y|A=0])-(E[C|A=1]-E[C|A=0])$$

也即, 治疗组中的效应等于 A 和 Y 的相关性减去 A 和 C 相关性。

这一控制混杂的方法被称为双重差分法 (Card, 1990; Meyer 等人, 1995; Angrist 和 Krueger, 1999)。在实践中, 这一方法经常和其他参数化的或非参数化的变量调整方法一起使用 (Abadie, 2005)。然而, 正如 Sofer 等人 (2016) 所述, 双重差分法一般只应用于阴性结局研究中, 它需要测量治疗前和治疗后的结局, 并且需要假设加法尺度上混杂相等。Sofer 等人 (2016) 讨论了其他尺度上的一般化方法, 弱化了混杂相等假设, 并且融入了其他的变量调整。更多有关阴性结局对照和混杂的讨论, 请参见 Lipsitch 等人 (2010) 和 Flanders 等人 (2011) 所著论文。

知识点 7.4: 前门准则 (原书第 98 页)

图 7.14 描绘的情境中, 治疗 A 和二分结局 Y 有一个未测的共同诱因 U , 同时 M 是 A 对 Y 效应的中介变量, 且和 A 或 Y 没有共同诱因。在这种结构中, 因为 U 是一个混杂变量但是我们没有

U 的数据, 所以研究者不能直接使用标准化或逆概率加权等方法来计算反事实风险 $\Pr[Y^{a=1} = 1]$

或 $\Pr[Y^{a=0} = 1]$ 。然而 Pearl (1995) 指出, 我们可以用前门准则来明确 $\Pr[Y^a = 1]$:

$$\sum_m \Pr[M = m | A = a] \sum_{a'} \Pr[Y = 1 | M = m, A = a'] \Pr[A = a']$$

Pearl 将此称为前门调整, 因为这需要 A 和 Y 之间的因果路径通过一个 A 的下游变量 M , 这一点和后门路径不一样。

前门调整的公式证明如下: 注意到 $\Pr[Y^a = 1] = \sum_m \Pr[M^a = m] \Pr[Y^a = 1 | M^a = m]$, 以及在图 7.14 中, 因为 (1) A 对 M 的效应没有混杂 (即 $A \perp\!\!\!\perp M^a$), 且 (2)

$$\Pr[Y^a | M^a = m] = \sum_{a'} \Pr[Y = 1 | M = m, A = a'] \Pr[A = a'], \text{ 所以有}$$

$$\Pr[M^a = m] = \Pr[M = m | A = a]。$$

条件 (2) 的证明如下: 注意到, 因为 (i) $M^a = m$ 时, $Y^a = Y^m$ (在图 7.14 中, A 对 Y 的效应需要通过 M), 以及 (ii) 在 SWIG 图中由有向分离有 $Y^m \perp\!\!\!\perp M^a$ 。最后, 在 SWIG 图中我们仅对 M 进行干涉, 则有界互换性 $Y^m \perp\!\!\!\perp M | A$ 成立, 所以有

$$\Pr[Y^a | M^a = m] = \sum_{a'} \Pr[Y = 1 | M = m, A = a'] \Pr[A = a']。$$

上述证明需要假设反事实结局 Y^m 是良定的。现在我们给出第二种证明, 第二种证明只需假设 Y^a 是良定的。为了进行证明, 我们将图 7.14 中因果性的 DAG 图视作统计上的 DAG 图, 同时利用 SWIG 图中的独立性 $D^a \perp\!\!\!\perp A | N$ (其中 $D = (Y, M)$ 是 A 的下游变量, $N = U$ 是 A 的非下游变量)。

因而:

$$\begin{aligned} & \Pr[Y^a = y] \\ &= \sum_m \sum_u \Pr[Y^a = y, M^a = m, U = u] \\ &= \sum_m \sum_u \Pr[Y^a = y, M^a = m | A = a, U = u] \Pr[U = u] \text{ (by exchangeability)} \\ &= \sum_m \sum_u \Pr[Y = y, M = m | A = a, U = u] \Pr[U = u] \text{ (by consistency)} \\ &= \sum_m \sum_u \Pr[Y = y | M = m, A = a, U = u] \Pr[M = m | A = a, U = u] \Pr[U = u] \\ &= \sum_m \Pr[M = m | A = a] \sum_u \Pr[Y = y | M = m, U = u] \left\{ \sum_{a'} \Pr[U = u | A = a'] \Pr[A = a'] \right\} \\ &= \sum_m \Pr[M = m | A = a] \sum_{a'} \left\{ \sum_u \Pr[Y = y | M = m, A = a', U = u] \Pr[U = u | M = m, A = a'] \right\} \Pr[A = a'] \\ &= \sum_m \Pr[M = m | A = a] \sum_{a'} \Pr[Y = y | M = m, A = a'] \Pr[A = a'] \end{aligned}$$

第七章图表

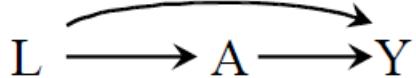


Figure 7.1

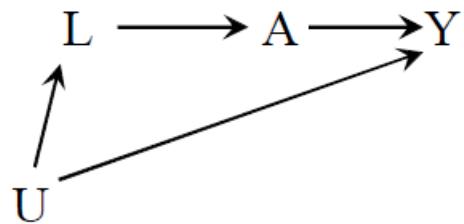


Figure 7.2

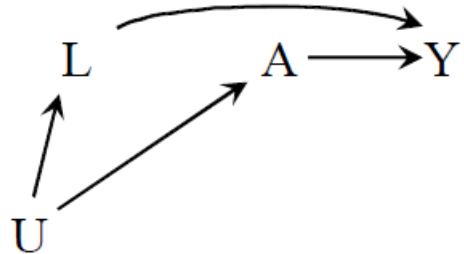


Figure 7.3

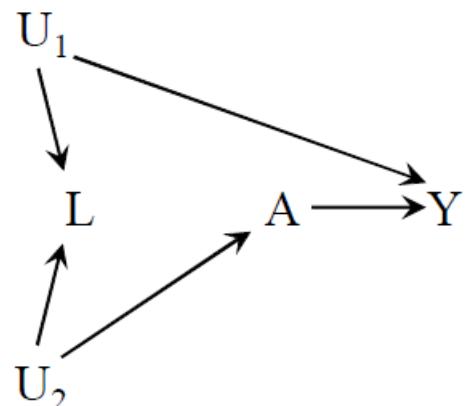


Figure 7.4

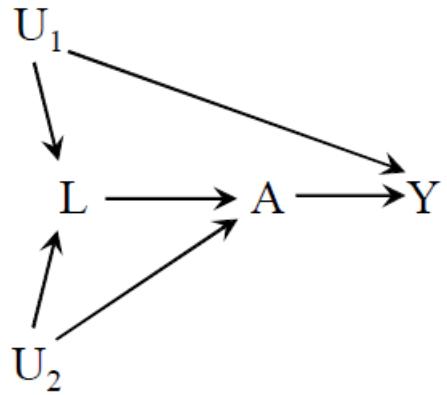


Figure 7.5

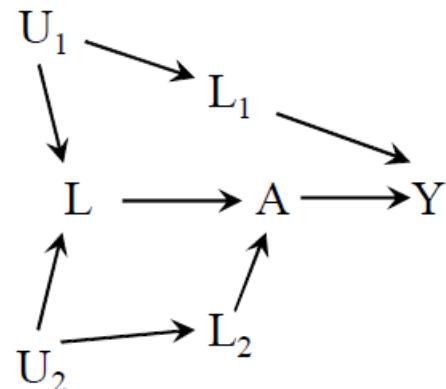


Figure 7.6

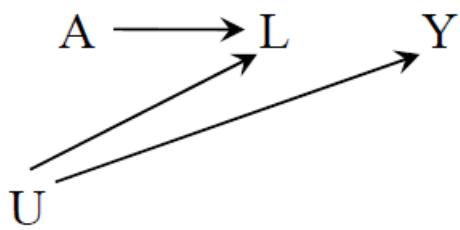


Figure 7.7

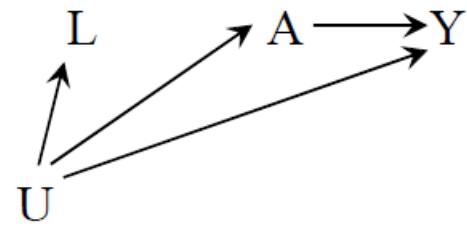


Figure 7.8

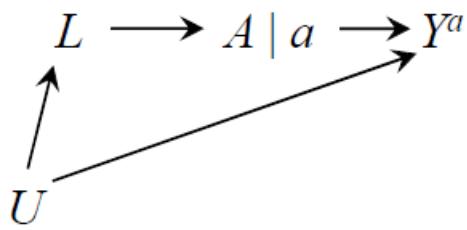


Figure 7.9

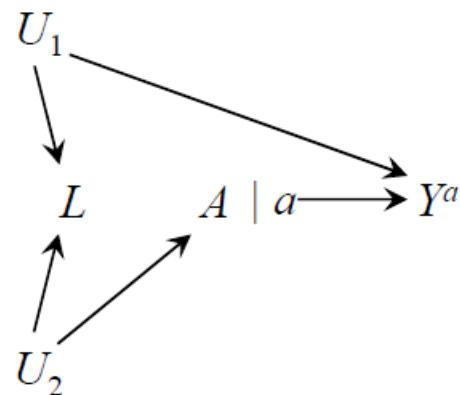


Figure 7.10

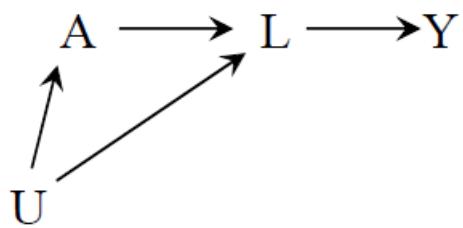


Figure 7.11

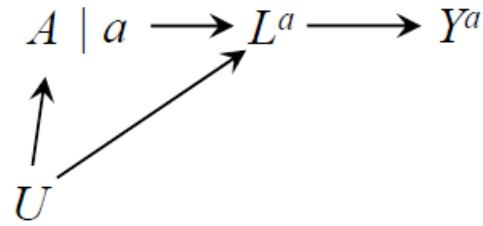


Figure 7.12

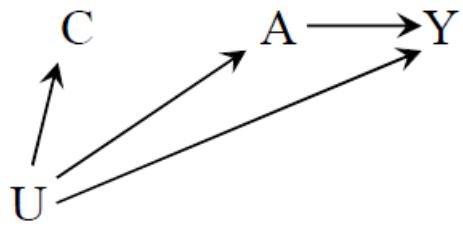


Figure 7.13

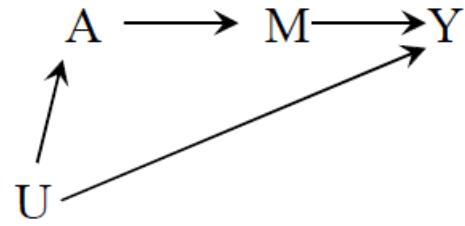


Figure 7.14

第八章 选择偏移

99 假设某研究者进行了一次观察性研究来探索以下问题：“一个人抬头向天上看会影响其他行人抬头向上看吗？”他发现第一个行人抬头向上看和第二个行人也抬头向上看之间存在很强的相关性。这个相关性反映了因果效应吗？这是一项随机试验，因而不太可能存在混杂因素。然而，除了混杂，还可能有另一个问题：数据分析只包括了同意将自己的数据用于研究的行人。比较害羞的行人（因而不太可能抬头向上看）以及感觉受到了戏弄的行人参加这项研究的概率较低。所以，数据分析中的人群有更大概率不会害羞，也因而更可能抬头向上看。因此，不管存不存在因果效应，对研究人群进行选择性地分析，会使得“一个人抬头向天上看”和“另一个行人也抬头向天上看”相关。

因选择性分析而引起的相关性被称作选择偏移。与混杂不同，选择偏移不是由治疗和结局的共同诱因引起的，并且观察性研究和随机试验中都有可能出现选择偏移。与混杂相同的一点在于，选择性偏移也是治疗组和非治疗组互换性缺失的一种体现。本章将讨论选择偏移的定义，并讲述调整选择偏移的方法。

8.1 选择偏移的结构

“选择偏移”一词指代因为对研究人群有选择地分析而引起的种种偏移。此处我们仅关注治疗对结局是零效应时的选择偏移，即零值下的选择偏移（参见第 6.5 小节）。选择偏移可以用诸如图 8.1 的因果图进行表示。图 8.1 中有一个治疗 A ，结局 Y ，及两者的共同后果 C 。假设图 8.1 表示一项探索孕妇叶酸补充剂 A 对胎儿心脏畸形 Y （1 表示有，0 表示没有）因果效应的研究。 C 表示胎儿产前死亡。胎儿心脏畸形能够提高死亡率（箭头从 Y 到 C ），同时叶酸能够降低胎儿心脏畸形和死亡的风险（箭头从 A 到 C 和 Y ）。这一项研究只在活产的胎儿中开展。因此，这项研究只在 $C = 0$ 的人群中进行研究，也即控制了 C ，所以图中 C 被方框框了起来。

(Pearl (1995) 和 Spirtes 等人 (2000) 首次用因果图描绘了因选择而引起的偏移。)

图 8.1 展示了治疗和结局之间相关性的两个可能来源：1) 路径 $A \rightarrow Y$ 所代表的的 A 对 Y 的直接因果效应，2) $A \rightarrow C \leftarrow Y$ 所代表的途经（被控制的）共同后果 C 的后门路径。在数据分析中，控制 C 会导致 A 和 Y 的相关性，被称为控制 C 导致的选择偏移。因为选择偏移的存在，相关性风险比 $\Pr[Y = 1 | A = 1, C = 0] / \Pr[Y = 1 | A = 0, C = 0]$ 就不再等于因果性风险比

100 $\Pr[Y^{a=1} = 1] / \Pr[Y^{a=0} = 1]$ ，此时相关性不是因果性。如果不控制治疗和结局的共同后果（对撞变量） C ，那治疗和结局之间的唯一开放路径就是 $A \rightarrow Y$ ，此时相关性就是因果性。

图 8.2 给出了选择偏移的另一个例子。图 8.2 包含了图 8.1 的所有变量, 外加一个表示父母是否伤心的变量 S (1 表示有, 0 表示没有)。 S 受到出生状态的影响。对胎儿的意外流产或死亡感到悲痛的父母不太愿意参与我们研究, 因而这项研究被局限于不感到悲痛 ($S = 0$) 的父母当中。我们在第六章讨论过, 控制被对撞变量 C 影响的 S 也会打开路径 $A \rightarrow C \leftarrow Y$ 。

图 8.1 和图 8.2 的选择偏移都来源于控制了治疗和结局的共同后果, 在图 8.1 中是 C , 在图 8.2 中是 S 。不管 A 对 Y 有没有因果效应, 这些选择偏移都会产生, 也即存在零值下的选择偏移。混杂 (参见上一章) 和选择偏移都可以是零值下的偏移。图 8.3 至图 8.6 描绘了零值下的选择偏移。

图 8.3 代表了一项 HIV 携带者的随访研究, 旨在探索某种抗逆转录病毒治疗 A 对三年死亡率 Y 的因果效应 (为了简化, 假设没有箭头从 A 到 Y)。未测的变量 U 代表高水平的免疫抑制 (1 表示有, 0 表示没有)。 $U = 1$ 的人死亡风险更高。退出研究或失访的个体在分析中被删失 ($C = 1$)。 $U = 1$ 的个体更可能被删失, 这是因为他们自身疾病的严重程度妨碍他们参与研究。 U 对 C 的影响受到其他因素的中介作用, 包括症状 (比如发烧、腹泻等)、CD4 数量, 以及血液中的病毒载量等, 这些都被包含在 L 之中。当 L 是已测变量时, 其作用我们将在 8.5 小节讨论。在本小节, 我们将 L 视为未测变量。接受治疗的个体更可能经受副作用的影响, 从而使他们退出研究, 这一关系体现在 A 到 C 的箭头中。 C 周围的方框表示我们分析局限于未被删失 ($C = 0$) 的个体中, 这是因为 Y 只在未被删失的个体中有测量。

根据有向分离的准则, 控制对撞变量 C 会打开路径 $A \rightarrow C \leftarrow L \leftarrow U \leftarrow Y$, 因而治疗 A 就会和结局 Y 相关。此时, 因果性风险比等于 1, 相关性风险比不等于 1。图 8.3 可以视作图 8.1 的一个简单变化。 Y 和 C 之间的相关性, 在图 8.1 中来源于 Y 对 C 的直接效应, 而在图 8.3 中则来源于 Y 和 C 的共同诱因 U 。我们可以这样理解这个偏移: 接受了治疗的个体会经受治疗的副作用, 因而也就有更大概率退出研究, 但是如果他们没有退出研究 ($C = 0$), 那影响退出的另一个独立因素 ($U = 1$) 也出现的概率就会降低。因而, 在没有退出研究 ($C = 0$) 的人群中, 就很可能出现 A 和 U 之间的负相关性。因为 U 和结局 Y 正相关, 所以局限于未退出人群中的数据分析会导致 A 和 Y 之间的负相关。

图 8.3 中的选择偏移来源于控制了表示删失的变量 C , 其是 U 和治疗 A 的共同后果, 而 U 是结局 Y 的一个诱因。除此之外, 选择偏移还能来自其他结构。在图 8.4 中, 治疗 A 对症状 L 有直接效应。我们的分析被局限于未删失的个体中, 也就意味着控制了 A 和 U 的共同后果 C , 从而导致治疗和结局之间的相关性。图 8.5 是图 8.3 的一个变种, 而图 8.6 是图 8.4 的一个变种。在这两幅图中, 存在 A 和另一个已测变量的共同后果 W 。 W 表示的未测变量可以是生活方式、人格

特质、教育水平等等, 其能影响治疗(从 W 到 A 的箭头), 和参与研究的态度(图 8.5 中从 W 到 C 的箭头)或报告症状的阈值(图 8.6 中从 W 到 L 的箭头)。

(图 8.5 和图 8.6 都是 M 偏移的例子。)

图 8.1 至图 8.6 都是能引起选择偏移的不同因果结构。在这些例子中, 偏移来源于控制了图中两个变量的共同后果, 也即对撞变量。这两个变量一个是治疗或治疗的诱因, 一个是结局或结局的诱因。我们将这种偏移称为选择偏移。接下来我们将举例讨论这些结构。

(概括而言, 选择偏移能被定义为控制了两个变量的共同后果, 其中一个变量是治疗或同治疗相关, 另一个变量是结局或同结局相关(Hernan, Hernandez-Diaz 和 Robins, 2004)。)

8.2 选择偏移的例子

以下是选择偏移的几个例子:

(其实, 大多数学科并不区分导致互换性缺失的不同结构。不管是什么原因的互换性缺失, 在统计学中被统称为“弱可忽略性”或“可忽略的治疗分组”(Rosenbaum 和 Rubin, 1983), 在其他社会学科中被统称为“可观测数据导致的选择”(Barrow 等人, 1980), 在经济学中被统称为“遗漏变量偏误”或“内生性”(Imbens, 2004)。)

- 有差别的失访: 图 8.3 至图 8.6 描述的情境就是有差别的失访。有时也被称为不对称删失¹引起的偏移。
- 缺失值偏移, 无回应偏移: 图 8.3 至图 8.6 中的变量 C 不仅能表示失访, 也能表示结局出现缺失值(可以是因任意原因)。比如, 某些人可能有缺失值, 这是因为他们不愿意提供自身的信息, 或者他们错过了某些研究随访。不管 Y 出现缺失值的原因为何, 将分析局限于有完整数据($C = 0$)的人群中可能会带来偏移。
- 健康工人偏移: 图 8.3 至图 8.6 也可以用来表示职业暴露中可能出现的偏移。假设我们要研究某化工厂员工中的职业暴露 A 对死亡率 Y 的因果效应。真正的健康状态 U 是一个未测变量。 U 能影响死亡率 Y 和工作状态 C (0 表示正常工作, 1 表示不在工作)。一些体检测量仅局限于正常工作($C = 0$)的人群中。(L 可能是血检结果或其他体检结果。)暴露于化学物质会降低正常工作的概率, 这种关系可能是直接的(比如暴露可能会致残), 就如图 8.3 和图 8.4 所描绘的一样, 也可以通过一个共同

¹ 原文 informative censoring。Informative 是“有信息的、有情报的”意思, 这里指因不同的信息导致的删失。根据常用的“信息不对称”一词, 译作“不对称删失”, 亦满足语境。

102

后果 W (某些高暴露的工作岗位被削减, 从而员工被辞退), 就如图 8.5 和图 8.6 所描绘的一样。

- 自我选择偏移, 志愿者偏移: 图 8.3 至图 8.6 中的 C 也能表示是否同意参与研究 (1 表示同意, 0 表示不同意)。比如, A 是吸烟, Y 是冠心病, U 是心脏病家族史, W 是生活方式 (L 是 U 和 C 之间的中介变量, 比如对心脏病的重视程度)。在这几种结构中, 如果研究仅局限于同意参加研究的人群, 那就可能出现选择偏移。

(Berkson (1955) 首次描述了因自我选择导致的偏移的结构)

- 治疗出现在研究开始前而导致的选择: 图 8.3 至图 8.6 中的 C 可以代表是否被选入研究中 (0 表示没有, 1 表示有), 同时, 治疗 A 出现在研究开始之前。如果治疗影响了参与研究的概率, 那么就很可能存在选择偏移。治疗导致的选择偏移, 可以视作自我选择的一般化。只要一项研究涉及了研究开始前的治疗, 或研究开始前的任意因素, 都可能产生这种情形下的选择偏移。除了选择偏移, 研究开始前的各因素之间也可能有未测或未知混杂, 这些都会影响最后的结果。

(Robins, Hernan 和 Rotnitzky (2007) 首次用因果图描绘了研究开始前的治疗所导致的选择偏移。)

103

除了上述的选择偏移, 以及精讲点 8.1 和知识点 8.1 中的选择偏移, 因果图还被用来描述因共同后果所引起的其他不同偏移。这些例子说明了选择偏移不仅能出现在回溯性研究中 (治疗 A 的数据在结局 Y 出现之后才进行收集), 也出现在前瞻性研究中 (治疗 A 的数据在结局 Y 出现之前进行收集)。此外, 这些例子说明了选择偏移不仅会出现在观察性研究中, 也可能出现在随机试验中。

(比如, 消除确认偏移的某些措施 (Robins, 2001)、估计直接效应的方法 (Cole 和 Hernan, 2002)、以及使用传统方法调整了受到过往治疗所影响的变量 (本书第三部分), 都可能导致选择偏移。)

以图 8.3 和图 8.4 为例, 图中的 A 和其他变量都没有共同诱因, 因此这两幅图都可以用来描述观察性研究或随机试验。在观察性研究或随机试验中, 参与研究的个体都可能失访或在确定结局之前退出研究。此时, 就不能直接计算结局风险 $\Pr[Y = 1 | A = a]$, 因为我们并不知道删失个体 ($C = 1$) 的结局 Y 。因而, 只能计算未删失个体的风险 $\Pr[Y = 1 | A = a, C = 0]$ 。因为删失人群和未删失人群之间可能不存在互换性, 所以将分析限制于未删失人群中可能会导致选择偏移。

因而混杂和选择偏移的一个主要区别是: 随机分组能防止混杂的出现, 但是不能阻挡选择偏移的产生 (因为选择可能出现在随机分组之后)。如果选择出现在治疗分组之前, 那就不会在研

究中引入选择偏移。比如, 只有同意参与随机试验的被试才能加入在随机试验中, 但是又因为这是在随机分组之前进行的选择, 所以这项随机试验不会受到志愿者偏移的影响。因而图 8.3 至图 8.6 不能表示随机试验中的志愿者偏移。图 8.3 和 8.4 被排除是因为治疗不会影响是否同意参加研究 C。图 8.5 和 8.6 被排除是因为随机试验中治疗不可能和其他变量有共同诱因。

8.3 选择偏移与混杂

在本章和上一章, 我们讨论了治疗组和非治疗组不可互换的两个原因: 1) 治疗和结局之间有共同诱因; 2) 控制了治疗 (或其诱因) 和结局 (或其诱因) 的共同后果。我们将前一种称为混杂, 将后一种称为选择偏移。依据结构进行定义虽然不能和传统定义完全吻合, 但能清晰直观地区分混杂和选择偏移。比如, 统计学家和计量经济学家经常将这两者都称为“选择偏移”。他们的理由是这两种偏移都是由于选择造成的: 有选择地进行分析 (结构性定义中的“选择偏移”), 或者有选择地进行治疗 (结构性定义中的“混杂”)。然而, 我们的目标并不是为了让各学科的术语保持一致, 而是想强调造成偏移的不同结构。

(因为同样的原因, 社会学家经常把未测混杂称为“不可观测数据导致的选择”。)

这两种结构都会导致治疗组和非治疗组互换性的缺失, 也即意味着这两种偏移在因果效应为 104 零时都可能出现。比如, 研究者想探索消防员中体育锻炼 A 对心脏病 Y 的因果效应, 图 8.7 描绘了这一情境。为了简化, 我们假设 A 不会影响 Y, 但是研究者不知道这一信息。父母的社会经济地位 L 会影响孩子是否成为消防员 C, 以及 Y (比如通过童年饮食状况)。对运动的喜爱程度 U 是一个未测的变量, 其会影响 A, 但不会影响 Y。L 也不会影响 A。根据我们之前的定义, 此时不存在混杂, 这是因为 A 和 Y 没有共同诱因。因而, 相关性风险比等于因果性风险比。

但是, 这项研究仅局限于消防员 ($C = 0$) 中, 从而研究者计算得到的相关性和因果效应不同, 这是因为在研究中控制了治疗的一个诱因与结局的共同后果, 从而引入了选择偏移。对研究者来说, 是否需要区分混杂与选择偏移还有一定争论, 这是因为不管怎么命名, 研究者都需要调整 L 从而保证治疗组和非治疗组的互换性。这个例子说明了依据结构对偏移进行分类并不总是会改变我们分析的方法。实际上, 许多流行病学家将需要被调整的变量 L 都称为“混杂变量”, 无论互换性的缺失是来源于控制了共同后果, 还是治疗和结局之间存在共同诱因。 105

然而, 依据因果结构对互换性缺失的原因进行分类, 这一方式依然有许多优点。第一, 一个问题的因果结构能帮助研究者决定什么样的分析方法能够减少或者消除偏移。比如, 在涉及时异变量的纵向研究中, 明确因果结构能让我们知道, 分层分析有时会造成选择偏移 (参见本书第三部分)。在此时, G-方法是更佳的选择。第二, 即使知道了因果结构也无益于数据分析方法的选

择(比如我们消防员的例子), 它依然有助于我们的研究设计。比如, 在我们消防员的例子中, 知晓因果结构后, 研究者就会在下一次研究中收集结局 Y 和选择变量 C (即消防员)共同诱因的相关信息(就如7.1小节中第一个混杂的例子一样)。第三, 因控制研究开始前的治疗所导致的选择偏移(比如成为消防员)能解释为什么同一个变量在一项研究中是“混杂变量”, 但在另一项研究中却不是。在我们的例子中, 如果研究人群不局限于消防员, 那也就不用调整父母的社会经济地位 L 。最后, 因果图能促进研究者之间的交流, 降低误解出现的可能性。

(选择不同的专业用语, 一般不会有实际的影响。但是忽视因果结构会导致显然的佯谬。比如, 所谓的辛普森悖论(1951)只是混淆了共同后果与共同诱因的区别。有趣的是, Blyth(1972)也未能理解辛普森悖论的因果结构本质, 从而将其错误解释为一种混杂。因为许多人只读过Blyth的论文而没有读过辛普森的论文, 这一错误影响深远。更多细节参见Hernan, Clayton和Keiding(2011)所著论文。)

我们可以用“健康工人偏移”来解释上一段中的最后一点。我们在前几个小节讨论并知道了这一偏移来源于我们控制了变量 C 。 C 是治疗(或其诱因)和结局(或其诱因)的共同后果。控制 C 导致的偏移可以用图8.3至图8.6来表示。然而, 当我们将一组工人和普通人群相比较时, 其中产生的偏移也能被称为“健康工人偏移”。此时, 这第二种偏移可以用图7.1进行表示, 其中 L 表示健康状态, A 表示是否为工人, Y 表示结局。 L 能影响 A 和 Y , 这是因为健康状态能影响工种与之后的结局。在这种情况下, L 作为共同诱因, 我们会将其视作混杂。因而, 用因果图表示“健康工人偏移”的结构能避免混淆。

上述讨论都忽略了选择偏移的强度与方向。然而, 控制对撞变量所打开的路径可能很弱, 因而不会造成太大偏移。因为选择偏移不是一个有或无的问题, 所以我们需要考虑选择偏移的可能方向与强度(参见精讲点8.2)。

8.4 选择偏移与删失

某研究者开展了一项边缘随机试验, 旨在探索吃芥末 A 对一年内死亡风险 Y 的因果效应均值。**106** 研究中有30名被试被随机分配到吃芥末组($A=1$), 他们的每一餐中都会增添芥末, 直到研究结束, 或结局出现。另外30人则被分到不吃芥末组($A=0$)。一年之后, 每组都有17个人死亡。因而相关性风险比是1。由于是随机分组的, 所以因果性风险比也是1。

然而, 现实是, 研究者不能完全观测到这两组中所有17个人的死亡, 因为有的人会在研究结束之前被删失。如果被试有心脏病($L=1$), 或者被分到吃芥末组($A=1$), 那么其被删失($C=1$)的概率就会高一些。实际上, 吃芥末组中只有9个人没有失访, 与之相比, 不吃芥末

组中有 22 个人没有失访。研究者在吃芥末组中观测到 4 名被试死亡, 在不吃芥末组中观测到 11 名被试死亡。因而, 未删失人群中相关性风险比 $\Pr[Y = 1 | A = 1, C = 0] / \Pr[Y = 1 | A = 0, C = 0]$ 等于 0.89。这一数值和真实的因果性风险比不同, 此时存在因控制共同后果 C 所导致的选择偏移。

图 8.3 描绘了这一芥末试验的情境。 U 代表动脉粥样硬化, 是一个未测变量, 并能同时影响心脏病 L 和死亡 Y 。图 8.3 中 A 和 Y 没有共同诱因, 这和边缘随机试验的设计一致, 因而计算因果效应时就没有必要调整混杂因素。另一方面, C 和 Y 有一个共同诱因 U , 因而存在后门路径 $C \leftarrow L \leftarrow U \rightarrow Y$ 。这条后门路径的存在意味着如果研究者想估计 C 对 Y 的因果效应 (从图中可知是零), 那就必须调整来自于共同诱因 U 的混杂。根据后门准则, 我们可以用已测的变量 L 阻断这条路径。

迄今为止, 我们所说的因果效应, 是“所有人都接受治疗时的风险” $\Pr[Y^{a=1} = 1]$, 和“所有人都不接受治疗时的风险” $\Pr[Y^{a=0} = 1]$ 两者的对比, 并没有涉及 C 。现在, 我们需要考虑删失或选择 (更概括一些) 存在时的因果性对比。如果所有人都未被删失, 那么也就不存在选择偏移。因而, 我们的因果性对比也应该反映删失存在的情境。

用 $Y^{a=1,c=0}$ 表示一个个体在 $A = 1$ 且未被删失 ($C = 0$) 时的反事实结局。同理, $Y^{a=0,c=0}$ 表示 $A = 0$ 且未被删失 ($C = 0$) 时的反事实结局。此时, 我们比较的是“所有人都接受治疗且未被删失时的风险” $\Pr[Y^{a=1,c=0} = 1]$, 和“所有人都不接受治疗时的风险” $\Pr[Y^{a=0,c=0} = 1]$ 。

(我们想计算的是因果性风险比 $E[Y^{a=1,c=0} = 1] / E[Y^{a=0,c=0} = 1]$ 或因果性风险差 $E[Y^{a=1,c=0} = 1] - E[Y^{a=0,c=0} = 1]$ 。)

大多数时候都可以假设删失对结局没有因果效应 (一个例外是失访会导致被试不能接受随后的治疗), 从而可以认为因果效应的定义可以忽略删失的影响, 也即我们可以忽略符号中的上角标 $c = 0$ 。然而, 忽略这个上角标可能掩盖可能存在选择偏移这一事实。实际上, 在考虑有删失的因果效应时, 我们可以将删失 C 视作另一个治疗, 从而我们分析的目标就是计算 A 和 C 联合干预下的因果效应。为了消除治疗 A 效应中的选择偏移, 我们需要调整治疗 C 效应中的混杂。

(在因果图中, 一般没有箭头从删失 C 指向观测到的结局 Y , 此时我们可以用反事实结局 $Y^{c=0}$ 替代 Y , 同时增加从 $Y^{c=0}$ 到 Y 的箭头, 此时就有从 C 到 Y 的箭头。)

因为删失 C 被视为一项治疗, 因而我们需要: (1) 确保可识别性的三个条件, 即互换性、正数性和一致性对 C 和 A 都成立; (2) 使用我们之前讨论过的分析方法去计算删失 C 的因果效应。在可识别性条件下, 我们可以通过分析方法消除选择偏移。下一小节将讨论如何做。

8.5 如何调整选择偏移

虽然在某些情况下能通过试验设计避免选择偏移（参见精讲点 8.1），但是大多数情况下，选择偏移都是不可避免的。就算研究者再尽心尽力，失访、自我选择、数据缺失等情况依然经常发生。此时，我们需要在分析中修正选择偏移。逆概率加权是一个常用的修正方法。该方法赋予每一个被选中者 ($C = 0$) 一个权重 W^C ，从而这个人不仅代表自己，还代表和他相似的人，也即有相同的 L 和 A 取值，但未被选中的人 ($C = 1$)。逆概率权重 W^C 是一个个体被选中的概率 $\Pr[C = 0 | L, A]$ 的倒数。

我们用上一小节的芥末试验来讲述如何使用逆概率加权调整选择偏移。图 8.10 的树状图描述了这次试验的数据。随机分组之前，60 名被试中，40 名有心脏病 ($L = 1$)，其余 20 名没有 ($L = 0$)。分组不取决于 L 的状态，所有人都有 50% 的概率被分配到芥末组 ($A = 1$)。因而在 $A = 1$ 组中有 10 名被试没有心脏病 ($L = 0$)，20 名被试有心脏病 ($L = 1$)。图 8.3 可以表示这一结构。未被删失的概率在树状图的每一分枝中都不一样：($L = 0, A = 1$) 组中是 50%，($L = 1, A = 0$) 组中是 60%。图 8.3 中从 A 和 L 到 C 的箭头表示 A 和 L 对 C 有影响。最后，这幅树状图显示了删失的和未删失的人群中会有多少人会死亡 ($Y = 1$)。当然，在现实世界中，研究者不可能知道删失人群中发生了什么，也正因为如此，研究者只能将分析限制在未删失的人群中，从而开启了后门路径，导致选择偏移。利用图 8.10 中的信息能计算得到整个人群中的风险比是 1。

(本书讲述了如何使用逆概率权重去调整混杂 ($W^A = 1/f(A | L)$) 和选择偏移 ($W^C = 1/\Pr[C = 0 | A, L]$)。当混杂和选择偏移都存在的时候，两个权重的乘积 $W^A W^C$ 能用来同时调整混杂和选择偏移。参见第十二章和本书第三部分。)

108 我们先讲述用逆概率加权调整选择偏移的基本理念。看到图 8.10 的底部。在 20 名被分配到芥末组 ($A = 1$) 但是有心脏病 ($L = 1$) 的被试中，4 人未被删失，16 人失访。也就是说这一组中未被删失的概率是 0.2，即 $\Pr[C = 0 | L = 1, A = 1] = 4 / 20 = 0.2$ 。在逆概率加权分析中，这 16 名被删失的被试权重是 0（因为他们没有出现在分析中），而另外 4 名未被删失的被试权重则是 5，这是他们未被删失的概率的倒数。逆概率加权将 4 名未被删失的被试计算 5 次，用以替代原先的 20 名被试。在其他分枝中重复同样的步骤，就会构建一个虚拟人群，如图 8.11 所示。这个虚拟人群的样本大小和原人群一样大，但没有失访和删失。此时虚拟人群中的相关性风险比是 1，就和整个人群中的因果性风险比相等。

109 如果以下三个可识别性条件成立, 那么虚拟人群中的相关性量度等于原人群的效应量度。

第一, 未被删失人群中的结局均值, 必须等于 A 和 L 分布相同的删失人群中未被观测到的结局均值。如果 A 和 L 足够阻断所有的后门路径, 且选择的概率 $\Pr[C = 0 | L = 1, A = 1]$ 是在 A 和 L 的条件下进行计算的, 那这一条就能满足。遗憾的是, 没有人能保证 L 能满足这一点。因而调整了选择偏移所得结果的因果性阐释, 依然依赖于不可验证的互换性假设。

第二, 逆概率加权要求 L 下未被删失的条件概率大于零。注意到, 这里的正数性仅对未被删失 ($C = 0$) 成立, 而不要求对被删失 ($C = 1$) 成立, 这是因为我们只对未被删失的人群做因果推断, 因而构建一个全部是被删失的虚拟人群也就没有意义。比如, 在 8.10 的树状图中,

110 $\Pr[C = 1 | L = 0, A = 0] = 0$, 但是这个零概率并不会影响我们构建虚拟人群。

第三, 一致性成立, 其中包括良定的治疗 (或干预)。逆概率加权构建的虚拟人群中, 删失 C 被消除, 同时 A 的因果效应和原始人群中的一样。因而, 虚拟人群中的效应量度, 等于如果没有被删失时的效应量度。如果删失是失访或无回应造成的, 那效应量度是良定的。然而, 如果删失是相互矛盾的两个事件造成的, 那么效应量度就不是良定的。比如, 在一项旨在探索某治疗对阿尔茨海默症效应的研究中, 因其他事件导致的死亡 (比如癌症、心脏病等), 就会和阿尔茨海默症构成矛盾事件。如果我们将死亡算作删失, 那会面临一个问题: 我们不知道用什么方法去构建一个虚拟人群, 其中所有因其他事件导致的死亡都能被排除。此外, 也没有什么可行的干预能排除某一死因导致的死亡, 但是不影响其他可能的死因。

(相互矛盾的事件会阻碍我们关心的结局出现。死亡是一个典型的矛盾事件, 因为只有死亡出现, 其他结局就不可能再出现。)

最后, 需要强调的是, 逆概率加权并不是用来调整选择偏移 (如图 8.3 所示) 的唯一方法。研究者也可以通过分层分析 (比如在 L 的每一分层中估计因果效应) 来调整选择偏移。如果控制 L 能够阻断所有的从 C 到 Y 的后门路径, 那么在 L 的每一分层中, 我们都能得到无偏的条件效应估计。此时条件风险比 $\Pr[Y = 1 | A = 1, C = 0, L = l] / \Pr[Y = 1 | A = 0, C = 0, L = l]$ 就是 $L = l$ 中未被删失人群的治疗效应。因果效应是零值时, 如果数据的结构能用图 8.5 来描述, 那么分层分析依然有效。然而, 如果数据是图 8.4 或图 8.6 中所描述的结构, 那么分层分析就不会有效。以图 8.4 为例, 控制 L 能阻断 C 到 Y 的后门路径, 但同时也开启了 A 到 Y 的路径 $A \rightarrow L \leftarrow U \rightarrow Y$ 。此时就算因果效应为零, 分层分析给出的风险比也不会是 1。对图 8.6 同理。与之相比, 逆概率加权能有效地处理图 8.3 至图 8.6 描述的所有情境, 这是因为逆概率加权方法不是在控制了 L 的

情况下估计因果效应, 而是根据每一个个体的治疗和 L 取值赋予他们一个权重, 然后计算无条件的效应量度。

这是第一次出现即使互换性、正数性和一致性成立, 但是分层分析不能有效计算因果效应的情形。我们将会在本书第三部分时异治疗部分讨论更多相似的情形。

8.6 没有偏移的选择

图 8.12 描绘了一项假想研究, 其中涉及一个二分手术变量 A , 某种基因单倍型 E , 以及死亡 Y 。根据有向分离准则, 手术 A 和基因单倍型 E 相互独立, 即接受手术的概率, 在有单倍型和没有单倍型的人群中是一样的。同时, A 和 E 在控制了 Y 时相关, 也即当研究局限于未死亡的人群中时 ($Y = 0$), 接受手术的概率和在基因单倍型的每一分层中不同。

实际上, 控制了 A 和 E 的共同后果 Y , 那至少在 Y 的一个分层中 (比如 $Y = 1$), 相互独立的两个变量 A 和 E 会条件相关, 但 A 和 E 依然可能在其他分层中条件独立 (比如 $Y = 0$ 中)。

假设 A 和 E 能够通过完全不同的机理影响死亡 Y , 并且在这两个机理中, A 和 E 都不能修饰对方对结局的效应。比如, 手术 A 能通过切除肿瘤影响 Y , 而 E 能通过调控脂蛋白胆固醇影响心脏病, 进而影响 Y 。在这种情形中, 我们可以考虑 3 种不同死因导致的死亡: 因为肿瘤导致的死亡 Y_A , 因为心脏病发作导致的死亡 Y_E , 以及其他原因导致的死亡 Y_O 。当 Y_A 、 Y_E 和 Y_O 中某一个等于 1 时, 我们观测到的死亡 Y 等于 1。当 Y_A 、 Y_E 和 Y_O 都等于 0 的时候, Y 等于 0。图 8.13 是图 8.12 的一个延伸, 描绘了此时的情境。我们假设表示具体死亡的变量 (Y_A , Y_E , Y_O) 并没有被测量, 因而已测变量只有图 8.12 中出现的变量 A 、 E 和 Y 。

从 Y_A 、 Y_E 和 Y_O 到 Y 的箭头是命定的。变量 Y 等于 0 的时候, Y_A 、 Y_E 和 Y_O 三个变量也都等于 0。因而, 在图 8.13 中使用有向分离, 我们知道 $Y = 0$ 时 A 和 E 相互独立, 也即 A 和 E 之间因为控制了对撞变量 Y 而打开的路径, 又被 Y_A 、 Y_E 和 Y_O 阻断了。另一方面, $Y = 1$ 时, Y_A 、 Y_E 和 Y_O 的取值有 7 种可能组合 (不再一一列出)。因而 $Y = 1$ 时, A 和 E 相关, 这是因为 Y_A 、 Y_E 和 Y_O 不再阻断因控制对撞变量 Y 而打开的后门路径。

与图 8.13 所示情境不同的是, 图 8.14 至图 8.16 所示情境中, $Y = 0$ 时, A 和 E 不再独立。图 8.14 可以表示 A 和 E 是通过一个共同机制影响死亡, 此时不管 Y 的取值为何, A 和 E 都不会相互独立。同样, 如果 Y_A 和 Y_E 有一个共同诱因 V , 如图 8.15 所示, 那不管 Y 的取值为何, A 和 E 都不会相互独立。最后一种情形, 如果 Y_A 和 Y_O 以及 Y_E 和 Y_O , 分别有共同诱因 W_1 和 W_2 , 如图

作者: Miguel A. Hernan, James M. Robins;

翻译: 罗家俊

8.16 所示, 那不管 Y 的取值为何, A 和 E 都不会相互独立。如果数据的情形可以用图 8.13 进行描述, 那我们会说这个数据遵循乘积生存模型 (参见知识点 8.2)。

图 8.13 中有趣的一点是, 通过增添决定 Y 的 Y_A 、 Y_E 和 Y_O , 我们改进了因果图, 使其能表示 $Y = 0$ 时 A 和 E 相互独立, $Y = 1$ 时 A 和 E 相关。

(我们可以进一步改进因果图, 进而表示第五章所说的充分成因模型 (VanderWeele 和 Robins, 2007c)。)

总而言之, 控制对撞变量会让它的两个诱因相关, 但是这个相关性可能仅局限于这个对撞变量的某一分层。换句话说, 理论上, 如果数据分析仅局限于选择变量的某一分层, 那不一定会引起选择偏移。对撞变量的分层不总是选择偏移的根源。

第八章精讲点和知识点

精讲点 8.1: 病例对照研究中的选择偏移 (原书第 102 页)

图 8.1 可以用来表示病例对照研究中的选择偏移。假设某研究者想探索绝经后雌激素治疗 A 对冠心病 Y 的因果效应。 C 表示研究人群 (用流行病学的术语来说, 叫做基础队列) 中的某个个体是否被病例对照研究选中 (1 表示没有, 0 表示有)。从 Y 到 C 的箭头表示人群中的病例更可能被选中, 这也是病例对照研究的特点之一。在这项病例对照研究中, 研究者更偏爱选择髋骨断裂的女性作为对照 ($Y = 0$)。因为治疗 A 对髋骨断裂有保护作用, 选择髋骨断裂的女性作为对照意味着治疗 A 对选择 C 有因果效应, 在图中表示为 $A \rightarrow C$ 。你也可以在 A 和 C 中间加上一个 F 表示髋骨断裂, 但不改变我们的讨论。

在病例对照研究中, 根据定义, 相关性量度 (一般为治疗对结局的比值比) 仅局限于被选中的人 ($C = 0$) 之中。如果髋骨断裂的人在这项病例对照研究中被过度抽样, 那就会因不当的对照选择而出现偏移。这里的偏移依然是因为控制了共同后果 C 而引起的。我们可以对这个偏移作如下解释: 在被选中的人当中 ($C = 0$), 对照组相比于病例组更可能患有髋骨断裂; 因为雌激素能降低髋骨断裂的风险, 所以对照组更不可能接受雌激素治疗, 因而选中人群中 A 对 Y 的比值比会比整个人群中的比值比大。病例对照研究中的其他形式偏移, 包括发病率-流行率偏移, 都能用图 8.1 及其扩展来表示。更多讨论参见 Hernan, Hernandez-Diaz 和 Robins 的论文 (2004)。

精讲点 8.2: 选择偏移的大小和方向 (原书第 112 页)

我们说选择偏移不是一个有或无的问题。在实践中, 我们需要考虑选择偏移的可能方向和大小。

控制 A 和 E 的共同后果 Y 所引入的相关性方向取决于 A 和 E 对 Y 的相互作用。比如, 假设有一个未知的背景因素 U , 其和 A 或 E 都不相关。在 U 存在的时候, $A=1$ 或 $E=1$ 都会导致死亡, 但如果 U 不存在, 那么 A 或 E 都不能导致死亡。在死亡的人 ($Y=1$) 当中, A 和 E 负相关, 这是因为 $A=0$ 的人更可能是 $E=1$ 。(实际上, 条件比值比 $OR_{AE|Y=1}$ 的趋近于 0, 因为死亡人群中 U 的流行率接近于 1)。当然, 这只是一个特殊的例子用以解释为什么控制了共同后果两个诱因会相关。

再比如, 假设 U 存在的时候, 同时有 $A=1$ 和 $E=1$ 才会导致死亡, U 不存在的时候, A 和 E 与死亡无关。此时, 在死亡人群中, $A=1$ 的人更可能有 $E=1$, 也即 A 和 E 正相关。一副诸如图 8.12 的标准 DAG 图不能区分上述两种情形。而包含了充分成因结构的 DAG 图克服了这一缺陷 (VanderWeele 和 Robins, 2007c)。

除去选择偏移的方向, 另一个问题在于大小。在实践中, 如果偏移太小从而不足以影响我们的强度, 我们一般可以忽略选择偏移, 而不管它的方向为何。一般而言, 一个较大的选择偏移需要对撞变量和其两个诱因有很强的相关性。Greenland (2003) 研究了零值下的选择偏移大小, 并在几个不同情境中称其为对撞变量-分层偏移。

知识点 8.1: 危害比 (Hazard ratio) 中固有的选择偏移 (原书第 104 页)

图 8.8 的 DAG 图能描述一项心脏移植 A 对两个时间点死亡风险 Y_1 和 Y_2 的因果效应。从 A 到 Y_1 的箭头表示心脏移植能降低时间点 1 的死亡风险。 A 和 Y_2 之间没有箭头, 表示 A 对时间点 2 的死亡风险没有影响。也就是说, 只要一个个体在时间点 1 存活下来, 那时间点 2 的死亡风险就不会受到心脏移植的影响。 U 表示一个未测的基因单倍型, 其能降低两个时间点的死亡风险。因为没

有混杂, 所以相关性风险比 $aRR_{AY_1} = \frac{\Pr[Y_1 = 1 | A = 1]}{\Pr[Y_1 = 1 | A = 0]}$ 和 $aRR_{AY_2} = \frac{\Pr[Y_2 = 1 | A = 1]}{\Pr[Y_2 = 1 | A = 0]}$ 就分别表示 A

对两个时间点死亡风险的无偏估计。即使 A 对 Y_2 没有直接效应, aRR_{AY_2} 也会小于 1, 这是因为 A 对整体死亡风险的影响会延续到时间点 2。

让我们来考虑某一具体时间点的危害比。时间点 1 的危害几率就等于在时间点 1 的死亡概率, 因而相关性危害比就等于 aRR_{AY_1} 。然而, 时间点 2 的危害几率等于时间点 1 存活后在时间点

2 的死亡概率。因而时间点 2 的危害比是 $aRR_{AY_2|Y_1} = \frac{\Pr[Y_2 = 1 | A = 1, Y_1 = 0]}{\Pr[Y_2 = 1 | A = 0, Y_1 = 0]}$ 。图 8.8 中 Y_1 四周的方框表示控制了 Y_1 。在时间点 1 的存活人群中, 未接受治疗的个体更可能有 U 代表的基因单倍

型。因而在时间点 2, 接受治疗的个体有更高的死亡风险。换句话说, 控制了 Y_1 , 在时间点 2 治疗 A 和更高的死亡率相关。因而, 时间点 1 的危害比小于 1, 但是时间点 2 的危害比大于 1, 也即发生了反转。我们认为时间点 2 的危害比是一个有偏的估计。这个偏移来源于控制了 A 和 U 的共同后果所引起的选择偏移, 也即打开了 A 和 Y_2 之间的路径 $A \rightarrow Y_1 \leftarrow U \rightarrow Y_2$ 。在生存分析中, 类似于 U , 与治疗无关却是死亡诱因的未测变量, 通常被称为脆弱变量。

与之相比, U 每一层中的条件危害比 $aRR_{AY_2|Y_1=0,U} = 1$, 这是因为控制非对撞变量 U 会阻断路径 $A \rightarrow Y_1 \leftarrow U \rightarrow Y_2$ 。因而, 条件危害比能够正确地表示 A 对 Y_2 没有直接效应。即使 U 和 A 相互独立, 无条件的危害比 $aRR_{AY_2|Y_1=0}$ 和每一分层中的危害比 $aRR_{AY_2|Y_1=0,U}$ 不同, 这一事实表示危害比不具伸缩性 (Greenland, 1996b)。遗憾的是, 虽然 $aRR_{AY_2|Y_1=0,U}$ 是一个无偏估计, 但我们不能计算它, 这是因为 U 是一个未知或未测的变量。在没有 U 数据的情况下。我们不可能知道 A 是否对 Y_2 有直接效应。也就是说, 基于已有的数据, 我们不能知道真正的 DAG 图, 到底是图 8.8 还是图 8.9。以上所有讨论适用于观察性研究和随机试验。

知识点 8.2: 乘积生存模型 (原书第 111 页)

如果给定 A 和 E 时的条件生存概率 $\Pr[Y = 0 | E = e, A = a] = g(e)h(a)$, 我们就说乘积生存模型成立。乘积生存模型等价于以下命题: 生存比 $\frac{\Pr[Y = 0 | E = e, A = a]}{\Pr[Y = 0 | E = e, A = 0]}$ 不取决于 e , 且等于 $h(a)$ 。如果 A 和 E 在乘法尺度上没有交互作用, 且因果结构能用图 8.13 进行表示, 那数据满足乘积生存模型。如果 $\Pr[Y = 0 | E = e, A = a] = g(e)h(a)$, 那 $\Pr[Y = 1 | E = e, A = a] = 1 - g(e)h(a)$ 就不满足乘积死亡模型。因而, 如果 A 和 E 在给定 $Y = 0$ 时相互独立, 那它们在给定 $Y = 1$ 时也相互独立。

第八章图表

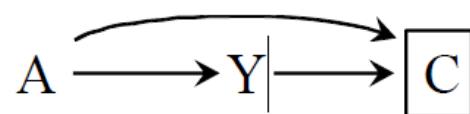


Figure 8.1

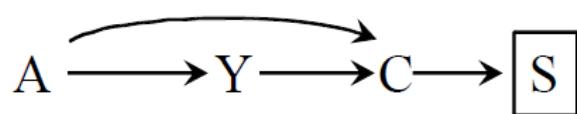


Figure 8.2

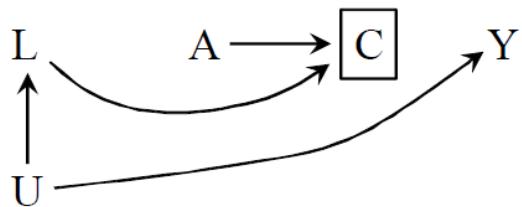


Figure 8.3

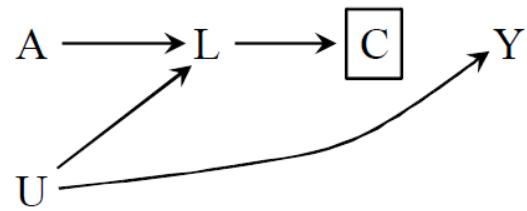


Figure 8.4

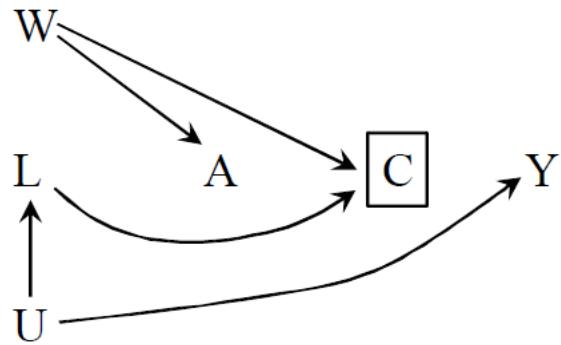


Figure 8.5

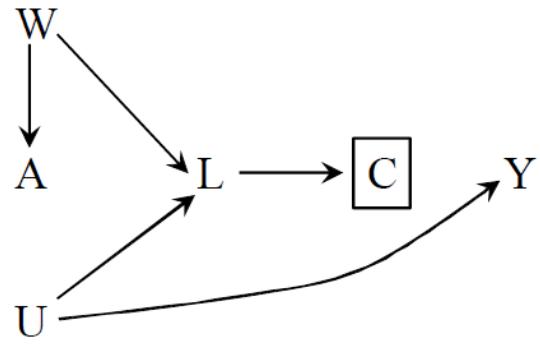


Figure 8.6

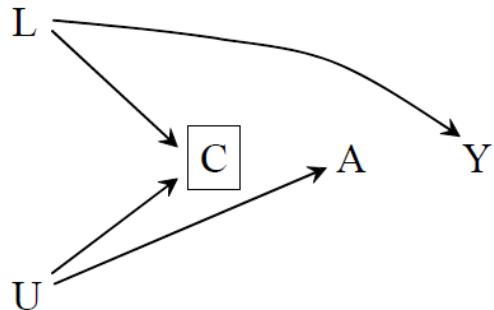


Figure 8.7

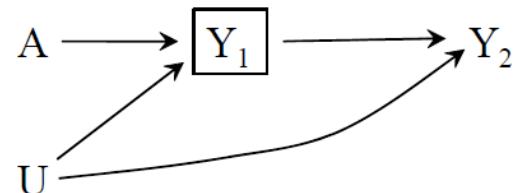


Figure 8.8

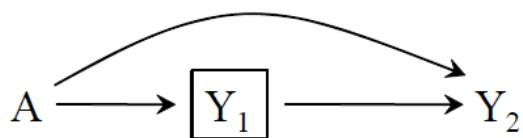


Figure 8.9

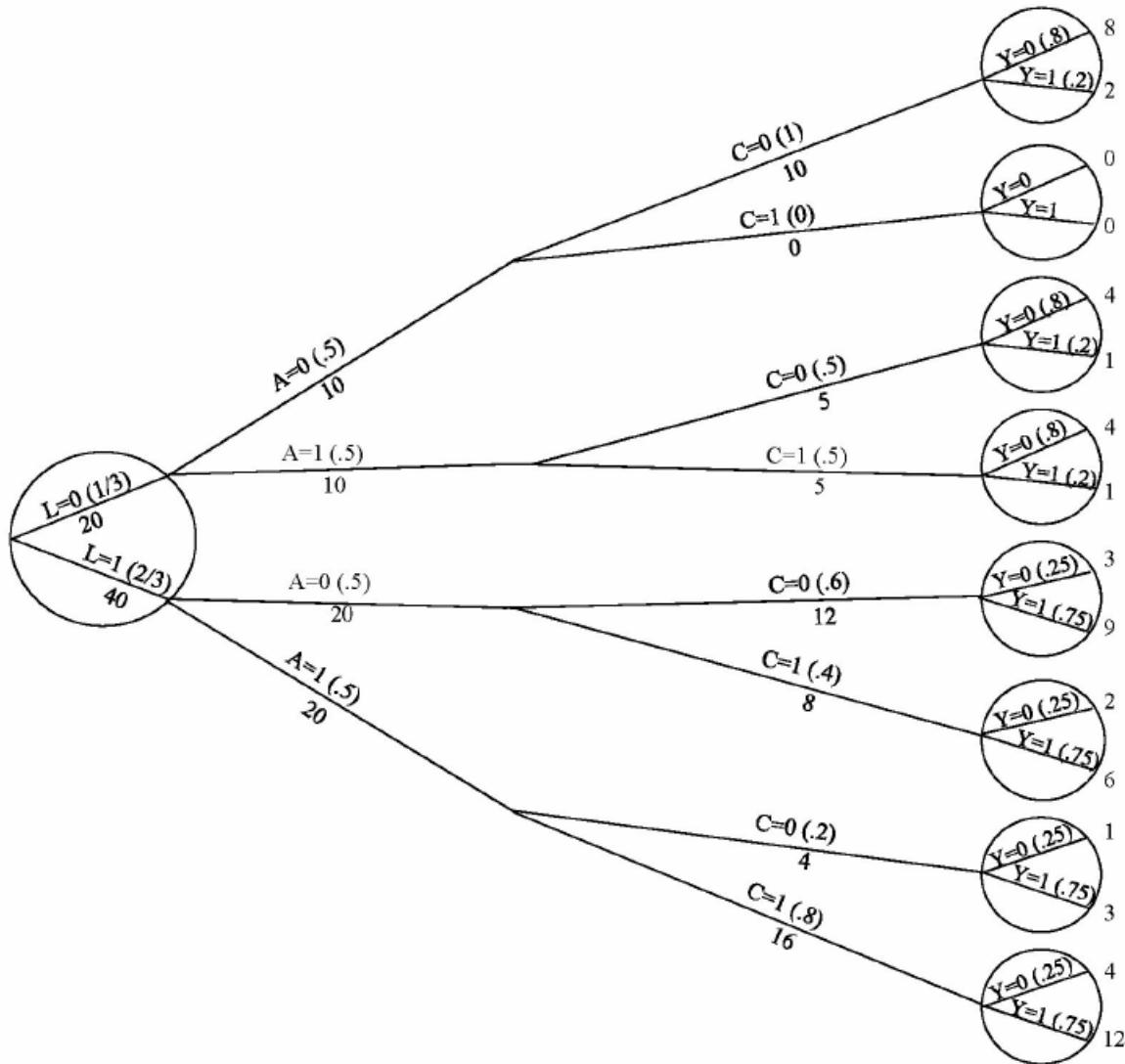


Figure 8.10

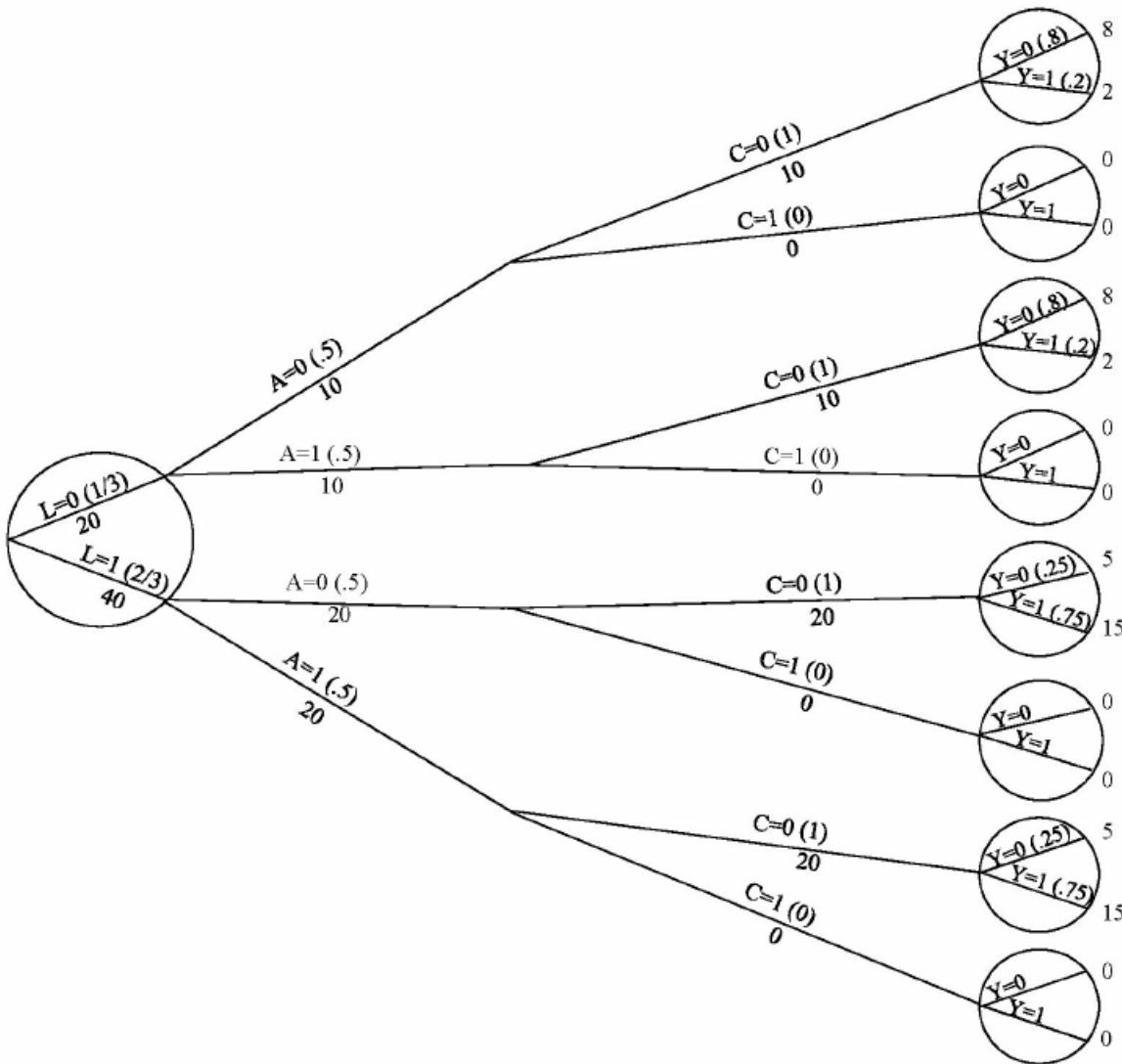


Figure 8.11

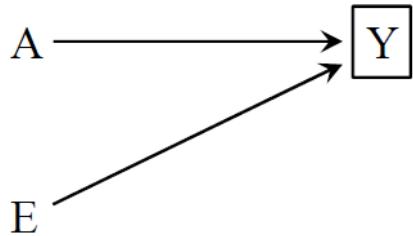


Figure 8.12

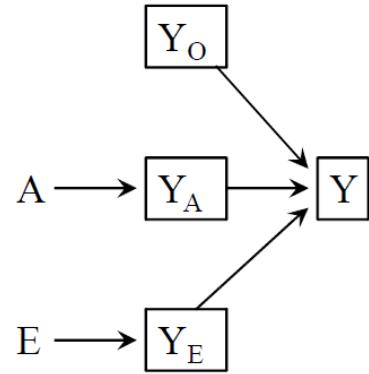


Figure 8.13

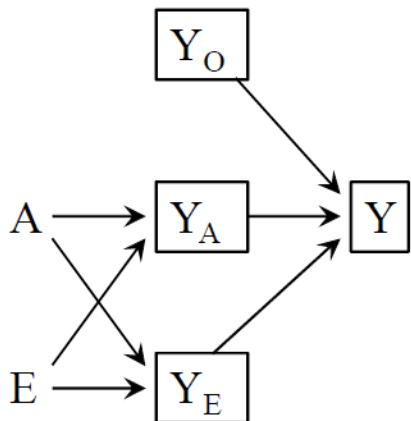


Figure 8.14

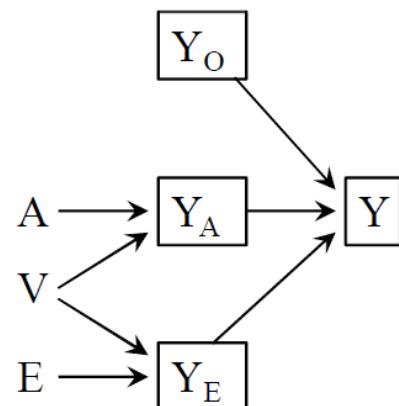


Figure 8.15

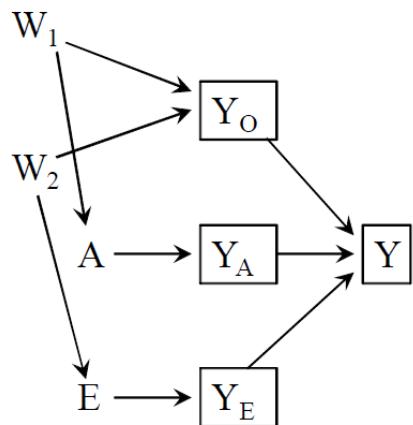


Figure 8.16

第九章 测量偏移

113 假设某研究者进行了一次随机试验来探索以下问题: “一个人抬头向天上看会影响其他行人抬头向上看吗?” 他发现自己抬头向上看和其他行人也抬头向上看之间存在弱相关性。这个弱相关性反映了因果效应吗? 根据随机试验的定义, 研究中不存在混杂。此外, 所有行人的反应都被记录分析, 因而不存在选择偏移。然而, 这项研究还有另一个可能的问题: 负责记录行人行为的试验人员可能会出错。具体而言, 试验人员可能会将“抬头向人看”的人, 记录成“没有抬头”的人。因而, 就算干预(研究者抬头向上看)对结局(其他行人抬头向上看)有很强的因果效应, 对结局的错误划分会稀释干预和(误测的)结局之间的相关性。

一项研究中, 如果数据的测量过程导致了治疗(或干预)和结局之间相关性的增强或减弱, 我们会说存在测量偏移。在任何一种研究设计中, 不管是随机试验还是观察性研究, 测量误差都可能出现。因而, 在我们阐释效应估计的时候, 我们必须将测量偏移考虑在其中。本章将讨论由于测量误差而引起的偏移。

9.1 测量误差

在前几章, 我们假设所有的变量都是完美测量的, 不存在误差。现在我们来考察一项观察性研究, 这项研究旨在估计降胆固醇药物 A 对肝脏疾病 Y 的因果效应。在实际中, 我们认为治疗 A 不可能被完美测量。比如, 我们可以从病历中提取药物的使用信息。但如果医生忘记将自己开的药记录在病历里, 或者病人并没有服药, 那就会存在测量误差。因而, 这项研究中的治疗变量并不是真正的药物使用情况, 而只是药物使用情况的一个测量。我们将这个测量记为 A^* (读作 A 星), 它不完全等于真正的治疗情况 A 。心理学研究有时将 A 称为“意象¹”, 将 A^* 称为“测量”或者“指征”。观察性学科的一个挑战在于研究者需要使用可观测的测量(比如病历中的药物使用情况)对不可观测的意象(比如实际的药物使用)做出逻辑推断。

图 9.1 描绘了变量 A 、 A^* 和 Y 之间的结构。为了方便我们的讨论, 这一情境中不存在混杂和选择偏移。真正的治疗 A 能影响结局 Y 和 A^* 。图中包含一个节点 U_A 表示除了 A 之外还对 A^* 有影响的因素。我们将 U_A 称为 A 的测量误差。 U_A 不是后门路径的一部分, 因此不会构成混杂或者选择偏移, 所以我们在前两章的讨论中没有涉及它。

(对于离散型变量, 测量误差被称为错分类或误分类。)

¹ 原文为 construct, 指被构造出来的概念, 而非实证证据。

除了治疗 A , 结局 Y 也可能存在测量误差。图 9.2 包含了 Y 的测量 Y^* 和 Y 的测量误差 U_Y 。图 9.2 描绘了实践中经常遇到的情形。通常, 我们想计算 A 对 Y 的因果效应均值, 但是 A 和 Y 不能被完美测量, 我们只能依靠它们带有误差的测量 A^* 和 Y^* 来判断 A 对 Y 的因果效应。

图 9.2 所示的情境也没有混杂和选择偏移, 因而相关性就等于因果性。我们能计算任意一个相关性量度并对其作出因果性阐释。比如相关性风险比 $\Pr[Y=1|A=1]/\Pr[Y=1|A=0]$ 就等于因果性风险比 $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]$ 。不过, 在前几章, 我们都暗中假设 A 和 Y 被完美测量。

我们现在需要讨论更现实一点的情境, 也即存在测量误差的情境。此时, 我们并不能保证 A^* 和 Y^* 之间的相关性等于 A 对 Y 的因果效应。一般而言, 相关性风险比 $\Pr[Y^*=1|A^*=1]/\Pr[Y^*=1|A^*=0]$ 不等于因果性风险比 $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]$ 。此时, 我们会说存在测量偏移或者信息偏移。在测量偏移存在的时候, 可识别性的三个条件, 即互换性、正数性和一致性, 就不足以保证我们能计算 A 对 Y 的因果效应。

9.2 测量误差的结构

前两章我们讨论了混杂和选择偏移的结构。概括而言, 混杂来源于治疗和结局之间存在共同诱因, 而选择偏移来源于控制了治疗和结局(或其诱因)的共同后果。测量偏移则来源于存在测量误差, 然而, 测量误差的结构不止一种。这一小节将根据测量误差的两个特性——独立性和差异性——对其进行分类(更正式的定义参见知识点 9.1)。

图 9.2 中包含 A 和 Y 的测量误差 U_A 和 U_Y 。根据有向分离准则, 因为测量误差 U_A 和 U_Y 之间的路径被对撞变量 (A^* 和 Y^*) 阻断, 所以 U_A 和 U_Y 相互独立。比如, 如果药物使用 A 和肝脏病史 Y 都来自时不时出错的电子病历, 那这两者可能相互独立。然而, 在其他情境中, 治疗和结局的测量误差却可能是相互依附的。比如, 通过电话采访回溯性地收集药物使用和肝脏病史的信息, 那 A 和 Y 的测量误差就会受到受访者回忆能力 U_{AY} 的共同影响。

在图 9.2 和图 9.3 的情境中, 治疗的测量误差 U_A 和真实的结局 Y 相互独立。同样, 结局的测量误差 U_Y 和真实的治疗 A 相互独立。此时, 我们说治疗的测量误差对结局来说无差异, 以及结局的测量误差对治疗来说无差异。图 9.4 描绘了独立但是有差异的测量误差, 其中真实的结局影响治疗的测量(有从 Y 到 U_A 的箭头)。以下我们介绍其他有差异的治疗测量误差。

假设我们的结局 Y 是痴呆而非肝脏疾病。药物使用 A 是通过采访得到的。因为痴呆会影响回忆 A 的能力, 所以有箭头从 Y 指向 U_A 。同理, 在一个关于孕期饮酒 A 和出生缺陷 Y 的研究中, 如果饮酒情况是在产后进行采访得到的, 那回忆情况就会受到妊娠结局的影响, 此时也会有从 Y 指向 U_A 的箭头。这两个例子中的测量误差通常被称为回忆偏移。如果 A 是药物使用, 我们在出现肝脏疾病 Y 之后才测量血液中的药物含量 A^* , 并以此代替 A , 那此时出现的偏移和回忆偏移结构相同, 不过这种情况一般会被称为逆向因果偏移。

图 9.5 描绘了独立但是有差异的测量误差, 其中真实的治疗会影响结局的测量 (有从 A 到 U_Y 的箭头)。当医生怀疑药物 A 会引起肝脏疾病 Y , 从而更加频繁监测服用药物 A 的病人时, 就会产生结局的有差异测量误差。图 9.6 和图 9.7 描绘的测量误差相互依附且有差异, 这可以是上述几个例子的综合。

总而言之, 我们讨论了四种不同类型的测量误差: 独立无差型 (图 9.2), 依附无差型 (图 9.3), 独立有差型 (图 9.4 和图 9.5), 以及依附有差型 (图 9.6 和图 9.7)。测量误差的不同结构决定了我们应该用什么样的方法去修正它。比如, 有大量的文献讨论了如何修正独立无差型测量误差。一般而言, 修正测量误差的方法主要依赖于模型假设和验证样本。这些方法超出了本书所讨论的范围。在本书, 我们主要强调变量的测量会引起偏移 (精讲点 9.1 讨论了偏移的方向和大小)。一个与事实相符的因果图需要同时包含混杂、选择和测量引起的偏移。很显然, 消除测量偏移的最佳方法, 是优化我们的测量过程。

116 9.3 混杂变量的测量误差

除去治疗 A 和结局 Y , 混杂变量 L 也可能有测量误差。即使治疗和结局的测量没有误差, 混杂变量的测量误差依然能造成偏移。图 9.8 描述了这一情境, 其中变量 A 表示药物, Y 表示肝脏疾病, L 表示肝炎史。如果一个个体有过往肝炎史, 那他不太可能接受药物 A 治疗, 并且有更高可能发展成其他肝脏疾病 Y 。第七章已经讨论过, 此时存在开放的后门路径 $A \leftarrow L \rightarrow Y$, 不过我们可以通过控制 L 来阻断这个后门路径。我们可以用标准化、逆概率加权等方法调整 L 从而计算 A 对 Y 的因果效应 $\Pr[Y^{a=1} = 1] / \Pr[Y^{a=0} = 1]$ 。

不过以上讨论暗含一个假设, 即 L 的测量完美无误差。如果研究者只能通过问卷调查收集研究人群的过往肝炎史, 那就可能出现测量误差, 这是因为不是所有参与者都会准确无误地提供自己的病史。有些人不想让其他人知道自己得过什么病, 有些人就只是纯粹想不起来了。因而, 研究者能得到的信息是 L 的测量 L^* , 而非真实的 L 。遗憾的是, 后门路径 $A \leftarrow L \rightarrow Y$ 不能通过控

制 L^* 而完全被阻断。此时通过标准化、逆概率加权控制 L^* 得到的效应估计, 会和 A 对 Y 的因果效应 $\Pr[Y^{a=1} = 1] / \Pr[Y^{a=0} = 1]$ 不同, 也即存在测量偏移, 也称信息偏移。

图 9.9 中的后门路径 $A \leftarrow L \leftarrow U \rightarrow Y$ 依然不能完全被 L 的测量 L^* 阻断, 因而 L 的测量误差会导致测量偏移。(注意, 图 9.8 和图 9.9 中都没有包含通常表示测量误差的变量 U_L , 这是因为 U_L 和我们此时的讨论并没有太大关系。)

117 我们也可以将图 9.8 和图 9.9 中的测量偏移视作未测混杂导致的偏移。实际上, 图 9.8 等价于图 7.6。我们可以将 L 视作一个未测或未知变量, 而将 L^* 视作一个混杂替代物(参见精讲点 7.2)。把它叫做未测混杂, 还是测量误差引起的偏移, 只是一个语言上的偏好问题, 和我们的最终目的无关。

混杂变量的测量误差也可能造成明显的效应修饰。比如, 如果所有报告过往肝炎 ($L^*=1$) 的人员和一半报告没有过往肝炎 ($L^*=0$) 的人员实际上都有过往肝炎 ($L=1$)。那么, 真实值和测量值在 $L^*=1$ 时就是完全重合的, 但在 $L^*=0$ 时则不是。假设每个个体治疗 A 对肝脏疾病 Y 都没有因果效应, 也即极端零假设成立。如果研究者将分析限制于 $L^*=1$ 的人群中, 他们就会发现 A 和 Y 不相关, 而当分析被限制于 $L^*=0$ 的人群中, 就会存在 L 带来的混杂, 因而研究者会发现 A 和 Y 相关。如果研究者没有意识到在 $L^*=0$ 时存在测量误差但是在 $L^*=1$ 时不存在, 那么他们就会天真地以为 $L^*=0$ 和 $L^*=1$ 中的相关性都是效应量度, 从而认为存在效应修饰作用。

最后, 如图 9.10 所示, 对撞变量 C 也可能有测量误差。在这种情形中, 控制有误差的测量 C^* 一般也会造成选择偏移。

9.4 治疗意向的效应: 治疗错分类的影响

有一项边缘随机试验旨在探索心脏移植对 5 年死亡率的因果效应。迄今, 我们在本书中用 $A=1$ 表示研究被试被分配到治疗组, 用 $A=0$ 表示被分配到对照组(或称非治疗组)。在一个理想的随机试验中, 如果所有被分配到治疗组的被试都接受了治疗, 所有被分配到非治疗组的都没有接受治疗, 那此时这些数学符号是没有问题的。然而, 在实际中, 这些符号不足以表示如果被试没有遵循分组方案的情形。

在现实随机试验中, 我们需要区分两个治疗变量: 分配的治疗 Z (1 表示被分配到治疗组, 0 表示没有) 和实际接受的治疗 A (1 表示实际接受了治疗, 0 表示没有)。对任一被试, 如果他没有遵循治疗分组的方案, 那 Z 和 A 的取值可能不同。比如, 被分配到心脏移植治疗组

118 ($Z=1$) 的被试可能拒绝接受手术, 从而其实际治疗取值为 $A=0$ 。也可能被试被分配到非治疗

组 ($Z = 0$)，但依然要求接受心脏移植 ($A = 1$)。在这种情形中，如果被试不遵循分组要求，那分配的治疗 Z 就是实际治疗 A 的一种错分类形式。图 9.11 描绘了 Z 、 A 和 Y 的关系 (U 将在下一小节讨论。)

但是，随机试验中的分配治疗 Z 和我们之前所说的 A^* 是有区别的。图 9.1 至图 9.7 中的 A^* 对结局 Y 没有因果效应。 A^* 和 Y 的相关性全部来自于它俩的共同诱因 A 。就算在观察性研究中， A^* 对结局也没有因果效应。而另一方面，从图 9.11 可知，分配治疗 Z 可以通过两个途径对结局 Y 产生因果效应。

首先，分配治疗 Z 对 Y 有因果效应是因为 Z 能直接影响 A 。被分配到治疗组的被试更可能接受治疗，在图中表现为从 Z 到 A 的箭头。如果治疗 A 对结局 Y 有因果效应，那么 Z 就能通过路径 $Z \rightarrow A \rightarrow Y$ 影响 Y 。

其次，分配治疗 Z 能通过不经 A 的路径对 Y 造成影响。比如，被分配到心脏移植组的病人会主动改变他们的生活习惯，从而为手术做好准备，医生也会更加关注这些病人。这些变化在图中表现为从 Z 到 Y 的箭头。

因此，分配治疗 Z 的因果效应不等于实际治疗 A 的因果效应，这是因为 Z 的因果效应不仅仅取决于 $A \rightarrow Y$ (实际治疗的效应) 的强度，还取决于 $Z \rightarrow A$ (遵循治疗分组的程度) 以及 $Z \rightarrow Y$ (被分配到治疗组后的行为变化) 的强度。

大多数时候研究者都想尽可能消除图 9.12 中路径 $Z \rightarrow Y$ 的影响，也即没有直接从 Z 到 Y 的箭头，这也被称为排他性限制 (参见知识点 9.2)。为了达成这一点，研究者故意不让被试知道他们的分组情况。比如，在药物研究中， $Z = 1$ 表示给予被试研究药物， $Z = 0$ 表示给予被试安慰剂，而安慰剂的外形和研究药物一模一样，无法区分。如此一来，被试和他们的医生就不知道他们的“药物”到底是真的研究药物，还是安慰剂。我们将这一类型的研究称为双盲安慰剂对照随机试验。然而，双盲设计并不总是可行的。比如心脏移植的研究中，就不可能有安慰剂组。

(在某些试验中，如果治疗有副作用，那就能知道谁是治疗组，因而双盲设计并不理想。)

再次强调， Z 的因果效应，衡量的不是“实际治疗 A 的因果效应”，而是“分配到治疗组的因果效应”，或者“有治疗意向的因果效应”，这也是为什么随机试验中 Z 的因果效应被称为“治疗意向的效应”。然而，尽管治疗组分配 Z 的效应取决于被试的合作程度，它也是大多数研究者在随机试验中希望计算的效应。那为什么研究者会计算治疗组分配 Z 的效应，而非真正治疗 A 的效应呢？下一小节将讨论这个问题。

9.5 依方案²效应

在随机试验中，依方案效应指的是所有被试都严格遵循试验方案中的治疗组分配时，治疗的因果效应。如果所有的被试都严格遵循治疗组分配，那每个被试的 Z 和 A 就会相同。此时，依方案效应就是治疗分配 Z 或实际治疗 A 的因果效应。我们在第二章的讨论过，理想的随机试验中每个被试都会遵循试验分组的治疗方案，并且治疗组 ($A=1$) 和非治疗组 ($A=0$) 是可互换的，即 $Y^a \perp\!\!\!\perp A$ ，从而相关性就是因果性，此时得到的因果效应也就是依方案效应。

然而在实际中，不是每个被试都遵循试验分组的治疗方案，因而治疗分组 Z 和实际治疗 A 不同。比如，被分到 $Z=0$ 组中的重症被试更可能在研究外去寻求心脏移植 ($A=1$) 治疗。此时 $A=1$ 组就会比 $A=0$ 组有更多重症被试，从而 $A=1$ 和 $A=0$ 就不再可互换， A 和 Y 之间的相关性就不再是因果性。此时得到的相关性量度就不再是依方案效应。

图 9.11 描绘了上一段中的情境，其中预后因素 U 表示重症程度（1 表示重症，0 表示一般）。因为后门路径 $A \leftarrow U \rightarrow Y$ 的存在， A 对 Y 的因果效应中就有混杂，所以计算 A 对 Y 的因果效应时就需要调整混杂。也就是说，计算依方案效应需要我们将随机试验视作一项观察性研究。如果 U 是未测变量，那我们就不能正确估算实际治疗 A 的因果效应。精讲点 9.2 讨论了当影响被试配合程度的预后因素存在且被测量时，能用以计算依方案效应的方法。

120 与之相比，治疗分组 Z 对结局 Y 的因果效应没有混杂。因为 Z 代表的是随机分组，所以互换性 $Y^z \perp\!\!\!\perp Z$ 成立，即使互换性对于实际治疗 A 不成立。图 9.11 中 Z 和 Y 之间没有后门路径，因而， Z 和 Y 的相关性也就意味着 Z 对 Y 的因果效应，不管被试有没有严格遵循分组要求。相关性风险比 $\Pr[Y=1|Z=1]/\Pr[Y=1|Z=0]$ 就等于治疗意向的因果性风险比 $\Pr[Y^{z=1}=1]/\Pr[Y^{z=0}=1]$ 。

(通过 Z 和 Y 之间的相关性去估计治疗意向的效应被称为治疗意向分析。更多讨论参见精讲点 9.4。)

Z 和 Y 之间没有混杂解释了为什么大多数随机试验更偏爱计算治疗意向的因果效应。虽然“接受治疗 A 的意向的因果效应”并不等于“实际治疗 A 的因果效应”，也即依方案效应、我们希望得到的效应，但是治疗意向效应更容易计算。我们将讨论使用治疗意向的理由。更多内容参见精讲点 9.4。

一个常见的理由是治疗意向效应保留了零值。也就是说，如果实际治疗 A 对 Y 的因果效应为零，那么治疗分组 Z 对 Y 的因果效应也是零。零值保留是一个重要性质，因为它能保证在没有因果效应时，我们不会错误地得到不为零的因果效应。更进一步，在极端零值假设和排他性限制

² 原文 per-protocol，意思是“遵循试验方案”。在中文尚无统一翻译。本书主要翻译为“依方案”。

下, 能证明 $\Pr[Y=1|Z=1]/\Pr[Y=1|Z=0]=\Pr[Y^{z=1}=1]/\Pr[Y^{z=0}=1]$ = 1。然而, 如图 9.11 所示, 如果排他性限制不成立, 那么这个等式也就不成立。比如, 试验不是双盲的, 那就可能存在从 Z 到 Y 的箭头, 此时实际治疗 A 的效应是零值, 但是治疗分组 Z 的效应不是零值

(用统计的术语来说, 治疗意向分析提供了一个虽然检功效不强, 但是有效的零假设显著性检验方法。)

121 另一个可能的理由是治疗意向的因果效应相比于依方案效应更接近零。其中逻辑是: 因为不是所有人都会严格遵循试验分组, 所以治疗的效应会减弱。因而治疗意向的风险比

$\Pr[Y=1|Z=1]/\Pr[Y=1|Z=0]$, 就会在实际治疗的因果效应 $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]$ 和 1 之间。因而治疗意向的因果效应就是实际治疗因果效应的一个保守估计。然而, 这一理由有 3 个问题。

第一, 这个理由假设了因果效应的单调性(参见知识点 5.2), 也就是说, 治疗的因果效应对所有个体都是同一个方向。但如果不是这样, 且被试的配合程度不高, 那么依方案效应可能相较于治疗意向效应更接近于零值。比如, 假设被分配到治疗组的被试有 50% 没有配合试验方案, 并且治疗的因果效应在配合和不配合的被试中方向相反, 那此时治疗意向的因果效应就不再是一个保守的效应估计。

第二, 就算治疗的效应是单调的, 即使治疗意向的因果效应在安慰剂对照试验中是一个保守估计, 但在被试被分配到两个不同治疗组的头对头研究³中则不是。假设有一项随机试验, 其中有慢性病的被试被随机分到新型药 ($Z=1$) 和布洛芬 ($Z=0$) 两组中。该试验旨在探索哪一种药对一年随访内疼痛的改善更加有效。但是研究者不知道的是, 这两种药的止痛功效相同, 也就是说, 依方案风险比 $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]$ 是 1。然而, 因为布洛芬副作用更强一些, 所以布洛芬组的被试配合程度更差, 最终导致治疗意向的风险比 $\Pr[Y=1|Z=1]/\Pr[Y=1|Z=0]$ 大于 1, 因而研究者会错误地认为新药更有效。

(一个反对治疗意向分析的相似论点也可以应用于非劣效性试验。非劣效性试验旨在确定一种治疗是否劣于另一种治疗。)

第三, 就算治疗意向的因果效应真是相对保守的, 那么当我们需要衡量药物安全性的时候, 这一特点就会变得十分危险。仅仅因为治疗分组 Z 的因果效应接近于 0, 就结论说实际治疗 A 是

³ 原文 head-to-head trial, 中文一般通俗翻译为“头对头研究”。其中的对照组不再是安慰剂, 而是临床中已使用的其他有效治疗方式。

安全的，这样的结论是危险的。因为被分配到 $Z=1$ 组的被试有可能没有或停止服药，从而使得真实的危害可能会更大。

122

因此，仅仅报告治疗意向的因果效应估计值并不适用与所有随机试验，比如有大量不配合的被试或者评估安全性的试验。遗憾的是，计算依方案效应需要研究人员调整混杂变量。研究人员可以在互换性假设下用我们介绍的方法调整变量，也可以在其他假设下利用工具变量来调整变量（工具变量方法将在第十六章讨论）。

（复杂随机试验中的依方案效应讨论，参见 Hernan 和 Robins (2017) 所著论文。）

在本章，我们对依方案效应的讨论非常简略，尚未引入时异变量。更多时候，随机试验中的治疗方案可能会随时间变化。我们将依方案效应定义为：如果每个被试都严格配合每个时间段分配给他的治疗方案，我们能观察到因果效应。本书第三部分会进一步讨论更常见情境中依方案效应的定义与计算。

总而言之，在随机试验的分析当中，我们需要在可能的未测混杂（依方案效应）和测量偏移（治疗意向效应）两者之间做出权衡选择。仅仅报告治疗意向效应意味着相较于混杂，研究者更愿意接受测量偏移。因而，研究者需要在每一次随机试验中对自己的选择进行说明与辩护。

第九章精讲点和知识点

精讲点 9.1：测量偏移的大小和方向（原书第 116 页）

一般而言，测量误差都会导致偏移，一个例外是 A 和 Y 不相关，同时测量误差是独立且无差的。在图 9.2 中如果没有从 A 到 Y 的箭头，那么 $A-Y$ 和 A^*-Y^* 的相关性都会是零。在其他情况下，测量偏移会使 A^*-Y^* 的相关性偏离 $A-Y$ 的相关性。更糟糕的情况是，如果治疗不是一个二分变量，测量偏移可能使 $A-Y$ 和 A^*-Y^* 的方向相反。当 A^* 可以用 A 的非单调函数进行表示时，那即使是图 9.2 的独立无差测量偏移， $A-Y$ 和 A^*-Y^* 的方向依然可能颠倒。更多细节参见 Dosemeci, Wacholder 和 Lubin (1990)，以及 Weinberg, Unbach 和 Greenland (1994) 所著论文。VanderWeele 和 Hernan (2009) 用因果图进行了讨论。

测量偏移的大小取决于测量误差的大小。也就是说， $U_A \rightarrow A^*$ 和 $U_Y \rightarrow Y^*$ 的强度越大，测量偏移也就越大。因果图并不包含定量信息，因而不能用来描述偏移的大小。

精讲点 9.2：依方案分析（原书第 120 页）

在随机试验中，有两种常用来估计依方案效应的方法，分别称为“实际治疗”和“依方案”分析。

传统的实际治疗分析会比较实际接受治疗 ($A=1$) 和实际未接受治疗 ($A=0$) 中结局 Y 的分布情形, 而不管治疗分组 Z 的情况。很显然, 如果未测的预后因素 U 促使被试实际接受治疗, 那就会存在混杂, 如同图 9.11 和图 9.12 所示。另一方面, 如果 A 和 Y 之间的所有后门路径都可以通过控制已测变量 L 阻断, 如图 9.13 所示, 那么实际治疗分析就能正确估计依方案效应。

传统的依方案分析只包含遵循试验分组方案的被试, 也就是 $A=Z$ 的被试, 然后比较 $Z=1$ 和 $Z=0$ 两组中结局 Y 的分布情形。传统的依方案分析, 只是将治疗意向分析限制于所谓的依方案人群中。一般而言, 这样会导致选择偏移。图 9.14 描绘了这一情境, 其中包含一个表示选择的变量 S 。 $S=1$ 时 $A=Z$ 。如果未能测量和调整变量 L , 那么效应估计就会受到选择偏移的影响。

其实, 实际治疗分析和依方案分析都是用观察性研究的分析方式去分析随机试验, 因而如同观察性分析一样, 需要调整混杂和选择偏移。更多例子及相关讨论, 请参见 Hernan 和 Hernandez-Diaz (2012) 所著论文。

精讲点 9.3: 虚拟治疗意向的分析 (原书第 121 页)

研究者只能在没有失访或删失的情况下直接计算治疗意向的效应。如果出现删失, 某些被试的结局就会缺失, 研究者也就只能将分析限制于没有被删失的人群当中。因而, 研究者只能进行虚拟治疗意向的分析, 计算 $\Pr[Y=1|Z=1,C=0]/\Pr[Y=1|Z=0,C=0]$, 其中 $C=0$ 表示没有被删失。在第八章我们讨论过, 删失会导致选择偏移, 因而虚拟治疗意向的效应估计值就是一个有偏的估计, 并且这个偏移可以是任一方向。在有删失的情况下, 即使是计算治疗意向的效应, 我们也要合理地调整选择偏移。更多讨论, 参见 Little 等人所著论文 (2012)。

精讲点 9.4: 效力与效益⁴ (原书第 122 页)

一些研究者将依方案效应 $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]$ 称作治疗的“效益”, 而将治疗意向的效应 $\Pr[Y^{z=1}=1]/\Pr[Y^{z=0}=1]$ 称作治疗的“效力”。治疗的“效益”更接近于理想随机试验中治疗 A 的因果效应均值。与之相比, 治疗的“效力”对应干预分组 Z 在干预措施未能被完美执行时的因果效应。因而, 经常有人争论说“效力”是一个更实际的效应量度, 因为“效力”还包含了从 Z 到 Y 不经过 A 的其他因果效应, 同时还考虑了不是所有病人都完美遵循医嘱。另一方面,

⁴ 原文“effectiveness versus efficacy”。Effectiveness 和 efficacy 两个单词在日常英文中基本没有什么区别, 这里只是指出在学术用语中的细微区分。此处中文对应翻译为效力和效益, 在日常中文中也基本没什么区别。

治疗的“效益”不能反映现实的实际情况。因而，一项随机试验仅仅报告治疗意向的因果效应是正当合理的，这不仅是因为“效力”更容易计算，还因为“效力”是一个更真实的量度。

然而，这一论点是有问题的。第一，随机试验中被试的配合程度和现实生活中的配合程度不一致。随机试验中的被试更可能配合治疗分组的要求，同时随机试验后一个新药是否有用会影响实际生活中病人的配合程度。第二，这个论点意味着我们进行的不是双盲试验，因为这个论点暗藏的观点是病人和医生清楚地知道他们接受的是什么样的治疗。第三，现实中的病人，更希望知道依方案效应，而非治疗意向的效应。更多讨论，请参见 Hernan 和 Hernandez-Diaz (2012) 所著论文。

知识点 9.1: 独立性和差异性 (原书第 114 页)

记 $f(\cdot)$ 为概率密度函数 (PDF)。如果测量误差 U_A 和 U_Y 的联合 PDF 等于它们边缘 PDF 的乘积，即 $f(U_Y, U_A) = f(U_Y)f(U_A)$ ，那这两个测量误差相互独立。如果 U_A 的 PDF 独立于 Y ，即 $f(U_A | Y) = f(U_A)$ ，那这个测量误差无差，对 U_Y 同理。

知识点 9.2: 排他性限制 (原书第 119 页)

当没有箭头从治疗分组 Z 指向结局 Y 时，也即 Z 对 Y 的所有效应都需要通过实际治疗 A 时，排他性限制成立。令 $Y^{z,a}$ 表示治疗分配为 z 、实际治疗为 a 时的反事实结局。一个更正式的表述为：如果对所有人以及所有 a （包括每个人的实际观测值 A ）都有 $Y^{z=0,a} = Y^{z=1,a}$ ，那排他性限制成立。排他性限制是工具变量方法（参见第十六章）的基本假设之一。

第九章图表

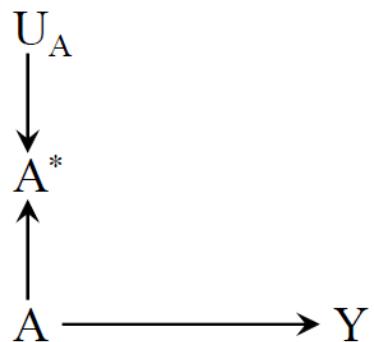


Figure 9.1

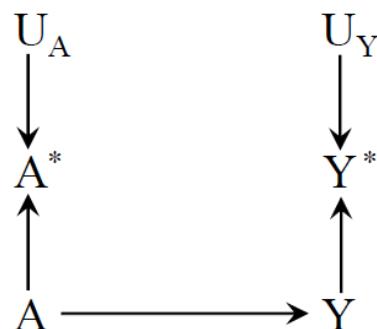


Figure 9.2

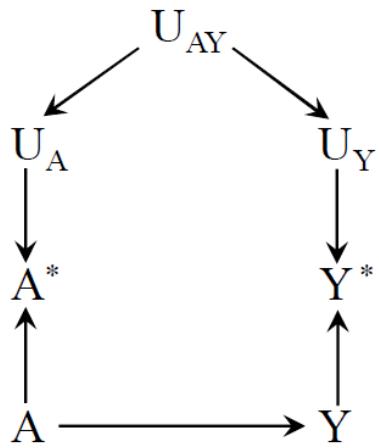


Figure 9.3

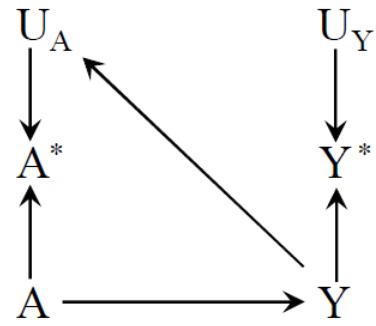


Figure 9.4

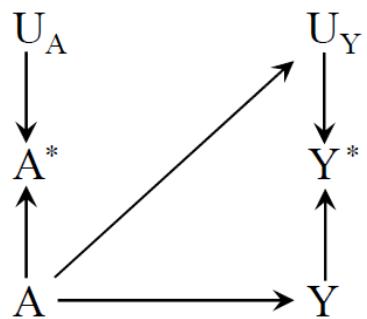


Figure 9.5

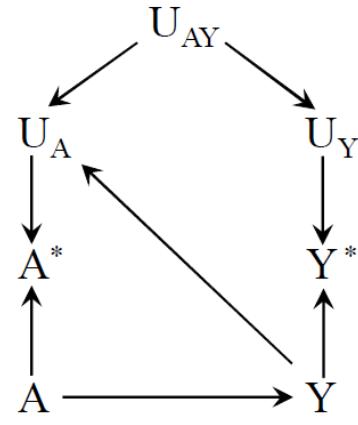


Figure 9.6

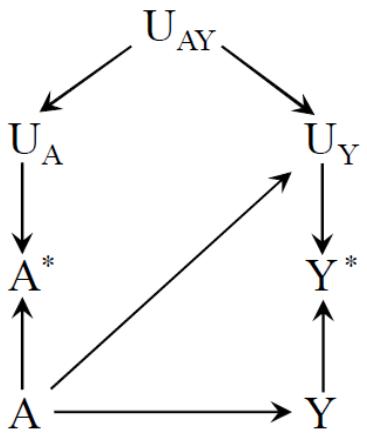


Figure 9.7

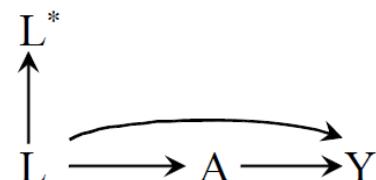


Figure 9.8

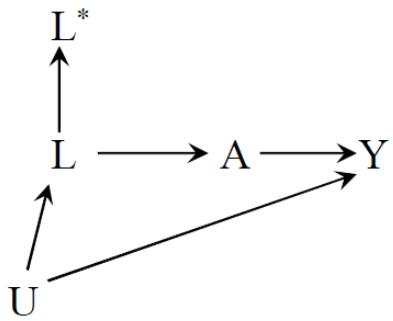


Figure 9.9

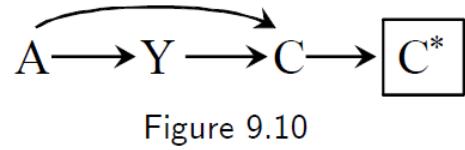


Figure 9.10

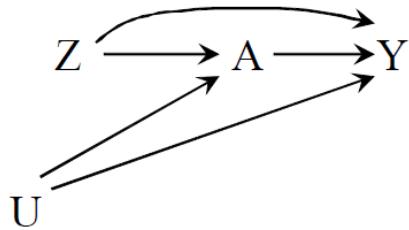


Figure 9.11

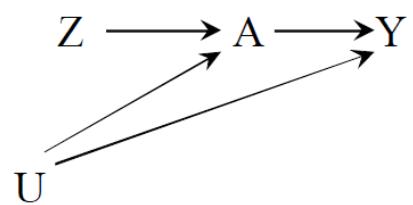


Figure 9.12

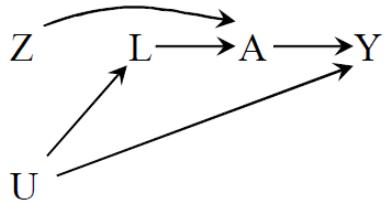


Figure 9.13

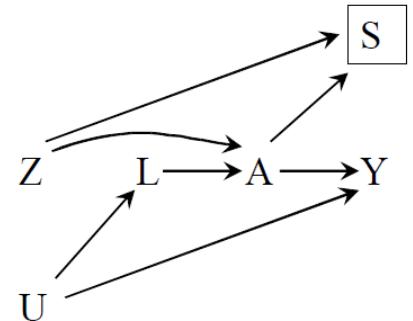


Figure 9.14

第十章 随机变异性

123 假设某研究者进行了一次随机试验来探索以下问题：“一个人抬头向天上看会影响其他行人抬头向上看吗？”他发现自己抬头向上看和其他行人也抬头向上看之间存在相关性。这个相关性反映了因果效应吗？根据随机试验的定义，研究中不存在混杂。此外，所有行人的反应都被记录分析，因而不存在选择偏移，并且所有变量都被完美测量，不存在测量偏移。然而，这项研究还有另一个可能的问题：这项研究只包含了4名行人，每个干预组中都只有2名行人。碰巧有1名“向上看”组的行人是盲人。因而，即使干预真的对结局有很强的因果效应，干预组中有一半的人还是会对这种效应免疫。这项研究的样本量太少，会稀释干预对结局的效应估计。

定性而言，因果推断出现错误的原因有两个：系统性偏移和随机变异性。前三章讨论了三种不同的系统性偏移：混杂，选择偏移，以及测量偏移。本章之前，我们假设所有的讨论都是在大样本中展开的，从而忽略了随机变异性导致的可能偏移。在本章，我们要回到现实。在几乎所有人群研究中，研究人群的样本大小都不能完全排除随机变异性带来的偏移。本章将讨论随机变异性的特征以及我们应该如何处理随机变异性。

10.1 识别与估计

本书前九章的讨论都是在假想的无限大人群中进行的。比如，在第二章心脏移植的例子中，虽然我们只有20个人，但我们将每一个人视作成百上千万与他完全相同的人。这样的假设能让我们忽略随机浮动，并将主要精力放在系统性混杂上。在统计学中，我们之前所关注的问题被称作“识别问题”。

迄今我们只考虑了因果推断中的识别问题，在之前几章我们的目标是去识别（有时我们也说计算）治疗 A 对结局 Y 的因果效应均值。可识别的概念在3.1小节、7.2小节和8.4小节都有所讨论，并且我们介绍了可识别性的三个前提条件，分别是互换性、正数性和一致性。

124 忽略随机变异性虽然有助于我们对系统性偏移的讲解，但却是不切实际的。在现实研究中，研究人群不可能是无限的，随机变异性的影响也就不能忽略。让我们回到最初那个20人的心脏移植研究，其中接受治疗的13个被试，有7个死亡。

假设我们研究中的20个被试是从一个超级人群中抽样得到的，而我们的目标是在这个超级人群中做出因果推断。我们想知道这个超级人群中的 $\Pr[Y=1|A=a]$ ，我们将其称为“待估值

¹”。而我们从样本中计算得到的, 是待估值的一个“估计值”。 $\Pr[Y=1|A=a]$ 对应的估计值被记为 $\widehat{\Pr}[Y=1|A=a]$ 。在我们的样本中 $\widehat{\Pr}[Y=1|A=a]=7/13$ 。我们将这个 $7/13$ 称为点估计, 这一估计值取决于随机抽样得到的 20 个人。

在第一章我们定义过, 如果样本量的增大使得估计值更加接近真实值, 那这个估计就是一致的 (知识点 10.1 给出了正式定义)。因而, 样本中的 $\widehat{\Pr}[Y=1|A=a]$ 是整个超级人群 $\Pr[Y=1|A=a]$ 的一致估计, 也即如果样本量越大, $\Pr[Y=1|A=a]-\widehat{\Pr}[Y=1|A=a]$ 越小。

(更多统计学基础知识请参见其他相关教材。)

即使估计是一致的, 估计值也可能与真实值相差巨大。样本量越小, 这个差距也就可能越大。因而, 样本量越大, 我们对估计值的信心也就越大。在没有系统性偏移的时候, 我们可以使用统计理论去量化我们的信心, 也即计算点估计附近的置信区间。样本量越大, 置信区间也就越窄。我们一般使用 95% Wald 置信区间。其计算过程如下。

首先, 我们需要假设样本是从更大的全人群中随机抽取得到的, 然后需要计算点估计的标准误差。其次, 95% Wald 置信区间的上确界等于点估计加上 1.96 倍标准误差, 下确界等于点估计减去 1.96 倍标准误差, 这是因为在标准正态分布中, 1.96 对应 97.5% 分位数。因而, 对于我们的估计值 $\hat{\theta}$, 置信区间为 $\hat{\theta} \pm 1.96 \times \widehat{se}(\hat{\theta})$ 。比如, 在我们的例子中, $\widehat{\Pr}[Y=1|A=a]=\hat{p}$, 则标准

125 误差是 $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(7/13)(6/13)}{13}} = 0.138$, 我们可以得到置信区间是 $(0.27, 0.81)$ 。95% Wald 置信区间的宽度与中心值因样本而异。

(只有在大样本中, 才能保证 Wald 置信区间中心值 \hat{p} 的有效性。为了方便, 我们假设讨论中的样本都足够大, 因而 Wald 置信区间是有效的。)

我们进行多次抽样, 如果在 95% 的随机样本中, 待估值都被包含在 95% 置信区间内, 那我们说这个置信区间是被校准的; 如果是超过 95% 的随机样本, 那我们说这个置信区间是保守的; 其他情况则置信区间是非保守的。如果置信区间是校准的或者保守的, 也即所有抽样中至少有 95% 的置信区间包含真实值, 那我们说置信区间是有效的。我们会尽可能选择有效且最窄的置信区间。

¹ 原文区分了 estimand, estimate (名词), 和 estimator 的不同含义。Estimand 是我们希望估计的量, 被译为“待估值”。Estimate 是我们的计算结果, 被译为“估计值”。Estimator 是计算估计值的一系列方法的总和, 为保持译文的可读性, 简译为“估计”。不过原文作者有时会混用 estimator 和 estimate。

置信区间的有效性是根据超级人群中的重复抽样进行定义的，但我们的现实研究中只有一次抽样。那为什么我们还要关心其余根本看不见的抽样？一个重要原因是置信区间也意味着：如果我们一生都在不断进行新的研究，在每一次研究中都计算一次 95% 置信区间，当我们结束职业生涯并总结一生的研究时，我们会声称有 95% 的研究包括了真实值，然而不幸的是，我们不能指出是哪 5% 的研究未包括真实值。

重要的是，某一项研究中 95% 置信区间的意思，不是真实值有 95% 的可能落在这个区间内。在我们的例子中，我们不能总结说真实值落在 0.27 至 0.81 之间的概率为 95%。真实值是固定的，从这个角度而言，其落在我们之心区间中的概率是 0 或 1。置信区间只有频率上的阐释。它的置信水平（这里是 95%）指的是在一系列重复研究中置信区间能包含真实值的频率。

（与频率学派 95% 置信区间不同，贝叶斯学派的 95% 可信区间能被阐释为“真实值有 95% 的概率落在区间内”。然而，在贝叶斯学派中，概率不是依据频率进行定义，而是根据信念度进行定义。在本书，我们采用的是频率学派的定义，也即依据频率定义概率。更多关于贝叶斯区间的讨论，参见精讲点 11.2。）

置信区间经常被分为“小样本”和“大样本”置信区间。一个有效的（即保守的或校准的）小样本置信区间对于所有样本大小都有效。校准的小样本置信区间也被称为精确置信区间。而一个有效的大样本置信区间仅对大样本有效。当样本量增加时，校准的大样本 95% 置信区间会更加接近 95%。我们上述提到的 $\Pr[Y = 1 | A = a] = p$ ，其 Wald 置信区间就是一个校准的大样本置信区间。（也存在 p 的有效小样本置信区间，但在现实中基本不会用到。）当样本量很小时，有效的大样本置信区间，诸如我们的 95% Wald 置信区间，可能就不再有效。在本书，除非特别说明，95% 置信区间指的都是大样本置信区间，比如 Wald 区间。参见精讲点 10.1。

（除了 Wald 区间，还有许多有效的大样本置信区间（Casella 和 Berger, 2002）。Wald 区间在小样本时效果不佳，此时研究者可能会更偏爱其他形式的置信区间（Brown 等, 2001）。）

然而，即使在大样本中，也不是所有的一致估计都能用来构建一个 Wald 置信区间。某些人认为如果一个估计能用来构建 Wald 区间，那它就是无偏的，反之，则是有偏的（参见知识点 10.1）。从现在起，我们将“偏移”一词等同于是否能用来构建 Wald 置信区间。同时，我们要牢记，置信区间只是量化了随机误差导致的不确定性，因而在有系统性偏移存在的情况下，我们对置信区间抱有的信心也许过多（参见精讲点 10.2）。

10.2 因果效应的估计

假设我们的心脏移植研究是一个边缘随机试验, 其中的 20 个被试从近乎无效的超级人群中随机抽样得到。假设超级人群中的每个人都被随机分配到 $A=1$ 或 $A=0$, 并且他们所有人都遵循分组方案。如此一来, 治疗组和非治疗组的互换性在这个超级人群中成立, 也即

$$\Pr[Y^a = 1] = \Pr[Y = 1 | A = a], \text{ 此时相关性量度就等于因果性量度。}$$

因为我们的研究样本是超级人群的一个随机样本, 所以 $\widehat{\Pr}[Y = 1 | A = a]$ 是 $\Pr[Y = 1 | A = a]$ 的无偏估计。因为超级人群中互换性成立, 所以 $\widehat{\Pr}[Y = 1 | A = a]$ 也是 $\Pr[Y^a = 1]$ 的无偏估计。因而我们就可以用 $\widehat{\Pr}[Y = 1 | A = 1] = 7/13$ 和 $\widehat{\Pr}[Y = 1 | A = 0] = 3/7$ 去检验零假设

$\Pr[Y^{a=1} = 1] = \Pr[Y^{a=0} = 1]$ 。之后我们可以用标准的统计方法计算出 95% 置信区间, 这适用于从标准化、逆概率加权和分层分析中得到的相关性量度。

我们也可以从另一个角度去思考随机试验中的抽样变异性。假设只有我们研究人群中的被 127 试, 而非超级人群中的每个人, 被分配到 $A=1$ 或 $A=0$ 两个组。因为随机变异性存在, 互换性在我们的研究人群中并不成立。比如每个被试在分组之前都有一个预后因素, 可以分为良好或危险两个种类。可能 $A=1$ 组 13 名被试中的 2 名、 $A=0$ 组 7 名被试中的 3 名, 他们的预后因素分组都是危险, 因此这两个组并不均衡。不过当我们的样本量增加时, 这两个组的不均衡程度就会降低。

在这样的情况下, 因果推断就有两种可能取向。第一种取向, 因为研究者并不知道超级人群的存在, 所以将推断局限于被随机分组的研究样本中。这被称为“基于随机化的推断”, 其中涉及的知识超过了本书讨论的范畴。第二种取向, 研究者知晓研究样本是从一个超级人群中抽取出来的, 因而希望在整个超级人群的范围内做出因果推断。从因果推断的角度来说, 后一种取向在数学上更接近于本小节开头给出的例子, 也即整个超级人群先被随机分组, 然而我们再从中抽样。换句话说, 先随机分组再随机抽样, 其实等价于先随机抽样再随机分组。

在大多数情况下, 我们并不会采用第一种取向。这是因为我们希望将自己的研究结论推广到所有适宜人群当中。比如你进行了一项癌症治疗方法的研究, 在研究结束后, 你更希望将自己的方法推广到所有癌症病人, 而非仅仅局限于自己的研究人群。迄今为止, 我们都假设了研究人群是从一个更大的人群, 也即超级人群中随机抽样出来的。接下来, 我们将讨论超级人群及其相关概念。

10.3 超级人群

在第一章我们已经说过, 随机性有两个来源: 抽样变异性与非命定的反事实。以下将对两者进行讨论。

$\widehat{\Pr}[Y=1|A=1] = \hat{p} = 7/13$ 是超级人群 $\Pr[Y=1|A=a] = p$ 的一个估计值。几乎所有的研究者都还会进一步计算置信区间 $\hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 7/13 \pm \sqrt{\frac{(7/13)(6/13)}{13}}$, 因为大多数研究者认为置信区间能反映随机变异性导致的不确定性。但这些置信区间只有在 \hat{p} 是二项抽样分布的时候才是有效的。因此我们需要知道什么时候 \hat{p} 是二项抽样分布。简单而言, 有如下两种情形。

129

(Robins (1988) 更详尽地讨论了这两种情境。)

- 研究人群是从一个本质上无限的超级人群中抽样得到, 待估值是 $\Pr[Y=1|A=a] = p$ 。站在数学的角度, 从这个无限人群中重复抽样 13 个人, 这 13 个人中出现结局的人数是一个二项随机变量, 其概率为 $\Pr[Y=1|A=a] = p$ 。95% Wald 置信区间就会渐近于 $\Pr[Y=1|A=a] = p$ 。我们迄今考虑的所有模型都是这个。
- 研究人群并不是一个超级人群的随机样本。而是: (1) 每个个体 i 的结局都是非命定 (随机) 的, 其反事实概率为 $p_i^{a=1}$; (2) 个体 i 观测结局等于反事实结局 $Y_i = Y_i^{a=1}$ 的概率为 $p_i^{a=1}$; 以及 (3) $p_i^{a=1}$ 对每一个个体都是相同的。因而, 这 13 个人中产生结局的人数就是一个二项随机变量, 其概率为 p 。95% 置信区间会渐近于 p 。

第一种情形假设了一个超级人群, 第二种则没有。然而第二种是不现实的, 因为产生结局的概率 $p_i^{a=1}$ 不可能对所有人都一样。因为个体间差异, 这 13 个人每一个人都有一个不同的概率。如果 $p_i^{a=1}$ 非常数, 那实际研究人群中的待估值就会是这 13 个人 $p_i^{a=1}$ 的平均值 p 。此时产生结局的人数就不是一个以 p 为概率的二项随机分布, 95% 置信区间也不再渐近于 p 。

因此, 如果研究者报告的是一个二项分布概率 $\Pr[Y=1|A=a] = p$ 的置信区间, 并且承认有个体间差异, 那研究者就已经假设了第一种情形: 研究样本是从一个近乎无限的超级人群中随机抽样得到的, 我们是对这个超级人群做因果推断。在第一种情形中, 不管反事实结局是命定的还是随机的, 这 13 个人中产生结局的人数, 都是一个二项分布变量。

假设超级人群的优势在于, 我们可以回避结局到底是命定的还是随机的这一问题。另一方面, 我们假设的超级人群是虚构的, 甚至可以说不切实际的。在大多数研究中, 研究样本并非从

一个近乎无限的人群中随机抽样得到。那为什么超级人群的概念一直被沿用? 其中一个原因是它简化了我们的统计模型。

第二个原因是基于一般化的考虑。研究者都想将自己从研究人群中得到的结论推广到更大的人群当中。¹³⁰ 因此最简单的方法是假设研究人群是一个更大人群的随机样本, 进而第一种情形中的 95% 置信区间反映的是更大人群中的真实值, 即使这个更大人群是虚构的。换句话说, 第一种情形中的置信区间可以阐述为“如果…会怎样…”。

当然, 如果研究者的目标人群并不是很多, 或者目标人群和研究人群的差别已经不能用抽样变异性来解释了, 那么研究者也就没必要坚持假设第一种情形。不过, 本书中我们依然会假设研究人群是从更大人群中随机抽样得到。我们会讨论在这种情形中随机变异性对因果推断的影响。接下来, 我们先在一个简单的随机试验中进行讨论。

10.4 条件性“准则”

表 10.1 是一项随机试验的数据, 旨在探索某治疗 A (1 表示有, 0 表示没有) 对一年内死亡风险 Y (1 表示有, 0 表示没有) 的因果效应。试验总共有 240 名被试, 每个治疗组都有 120 名。

相关性风险差为 $\Pr[Y=1|A=1] - \Pr[Y=1|A=0] = \frac{24}{120} - \frac{42}{120} = -0.15$ 。假设研究人群来自于一个近似无限的超级人群, 则治疗组和非治疗组可互换, 即 $Y^a \perp\!\!\!\perp A$, 并且相关性风险差等于因果性风险差 $\Pr[Y^{a=1}=1] - \Pr[Y^{a=0}=1]$ 。研究者算得 95% 置信区间为 $(-0.26, -0.04)$, 然后将结果发表, 并总结说治疗能降低死亡风险 15 个百分点。

然而, 这项研究只有 240 个人, 因为一些随机情况, 治疗组和非治疗组并不能互换。随机分组并不能保证两个分组都是可互换的, 只能保证互换性缺失是因为随机差异, 而非系统性误差。实际上, 置信区间的宽度 (这里为 0.22) 反映了随机差异以及我们的知识盲点带来的不确定性。

几个月后, 研究者发现数据中还有第三个变量吸烟 L (1 表示有, 0 表示没有)。根据 L 分层后的数据在表 10.2 中。之前没想到的是, 研究者发现吸烟者中接受治疗的概率 ($80/120$) 是不吸烟者的 ($40/120$) 两倍, 表明治疗组和非治疗组不可互换, 并且需要调整吸烟。当研究者通过分层分析调整吸烟后, 在吸烟者中的相关性风险差是 0, 在不吸烟者中也是 0, 也即治疗在吸烟者和不吸烟者中都没有用。

¹³¹ (知识点 10.5 讲述了调整后的估计方差。此时 95% Wald 置信区间是 $(-0.08, 0.08)$ 。)

这一新发现给研究者带来了困扰。要不然是有一些被试没有被随机分组 (这是研究者的渎职), 要不然是随机分组了依然没有保证互换性, 那真是运气太不好了。那研究者应该撤回他们

的文章吗? 研究者一致认为, 如果这是因为渎职造成的, 那就应该把文章撤回。如果这样的话, 那就存在吸烟导致的混杂, 因而需要调整吸烟。但调查后发现, 这并不是因为有人渎职, 而是随机分组了依然没有保证互换性。那他们还应该撤回文章并修正结果吗?

有一个研究者认为不应该撤回文章, 他的理由如下: “这里的吸烟没有被随机分组。但问题是, 为什么我们会认为调整后的结果好于未调整的结果呢? 也可能其他未测变量 U 不均衡带来的混杂抵消了 L 不均衡的效果, 因而未调整的估计值仍然近似超级人群中的待估值。”另一个研究者认为应该撤回文章, 他的理由如下: “因为 L 给我们的效应估计带来了混杂, 所以我们应该调整 L 。在 L 的每一分层中, 我们都可以视作一个小型的随机试验, 并且能计算置信区间。置信区间的宽度能够反映控制了 L 之后 U 带来的不确定性。”

为了判断哪个研究者是正确的, 我们需要考虑几个重要因素。为了简化, 假设真实的因果性风险差在 L 的每一分层中都相等, 并且我们能进行成百上千次随机试验。我们只选择 L 和 A 如同观测数据中一样正相关的试验。在这些试验中, 我们会发现, 在 L 的每一层中, U 和 A 正相关的次数, 等于 U 和 A 负相关的次数。(即使 U 和 L 在超级人群与研究人群中高度相关, 这一结论依然成立。) 因而, 分层之后, 也即控制 L 之后, 调整后的估计值是无偏的, 但是未调整的估计值是有偏的。但是, 如果我们不控制 L , 也即不施加任何条件, 那综合这多次试验的结果, 我们会发现未调整的和调整后的估计值都是无偏的, 但是调整后的估计值方差更小。也就是说, 调整后的估计, 在有条件时是无偏的, 在无条件时是高效的。因而, 不管是从有条件还是无条件的角度出发, 条件估计的 Wald 置信区间都是一个更好的分析方法, 因而这篇文章应该被撤回。第二名研究者是正确的。

(调整后的估计更高效是因为调整后的估计是在有 L 数据时的最大似然估计。)

我们应该尽可能控制 $L - A$ 之间的相关性, 这一想法在统计学中被称为条件性准则。在统计学中, 学者一般认为 $L - A$ 的相关性从属于我们关注的因果性量度。条件性准则认为任何统计推断都需要调整所有从属性因素(参见知识点 10.2 和 10.3)。上一段的讨论说明了为什么统计学者会采用条件性准则。我们之前对偏移的定义在他们看来是不充分的。他们会认为一个估计是无偏的, 当且仅当在控制了从属性因素后的估计能用来构建一个 Wald 置信区间。知识点 10.5 说明了为什么大多数研究者实际上都在遵循条件性准则。

不过, 频率学派会争论说: “没必要调整 L , 这是因为综合多次试验后, 未调整的估计值也是无偏的。”支持条件性准则的研究者会反击道: “其他假想研究中 $L - A$ 相关性和我的现实研究有什么关系? 在我的研究中, L 是一个混杂变量, 所以我要调整它从而消除混杂。”这个反击

对随机试验和观察性研究都非常有说服力, 尤其是当未测混杂不是很多的时候。然而, 如果存在大量未测混杂, 依然严格遵循条件性准则可能并不是一个明智的选择。

10.5 维度的诅咒

前几节的讨论都建立在渐近理论之上, 也即 L 的分层数远小于样本量大小。在本节, 我们会¹³⁴ 讨论如果 L 的维度非常大, 甚至远大于样本量, 那会发生什么。

假如研究者测量了 100 个治疗之前的二分变量, 那这 100 个二分变量就会构造出 2^{100} 个可能分层, 我们把这个数据成为高维数据。为了简化, 假设 L 没有加法尺度上的效应修饰作用。也即超级人群的风险差 $\Pr[Y=1|A=1,L=l] - \Pr[Y=1|A=0,L=l]$ 在 2^{100} 分层中都是一样的。更进一步, 假设风险差在每一层中都是 0。

研究者会再次争论是否要撤回文章, 并重新报告分层之后的风险差。他们现在同意应该遵循条件性准则, 因为未调整的风险差 -0.15 是一个条件性的有偏估计。然而, 他们也注意到, 当有 2^{100} 个分层时, 调整后估计值的 95% 置信区间要远比未调整估计值的宽。这和 L 只有两层时的结果相反。实际上, 此时调整后估计值的 95% 置信区间太宽了, 以致于基本没有什么意义。

2^{100} 显然远大于样本量 240, 因而 L 的分层中只有一些分层能同时包含治疗组和非治疗组。假设仅有一个分层包含了一名治疗组被试和一名非治疗组被试, 没有其他分层再同时包含治疗组和非治疗组被试。那调整后的估计值 95% 置信区间就会是 $(-1, 1)$, 从而没有丝毫意义, 这是因为在每一分层中, 风险差只能是 -1, 0 或 1, 这取决于每一分层中 Y 的取值。与之相比, 未调整估计值的 95% 置信区间依然是 $(-0.26, -0.04)$, 这是因为它的宽度不会受到协变量数量的影响。这个结果表明了只有在样本量大小与变量个数之比足够大的时候, 调整后的估计才比未调整的估计更加高效。一般而言, 这个比例取 10, 不过具体数值视数据特性而定。¹³⁵

那研究者们应该做什么呢? 如果遵循条件性准则, 他们会在高维数据中遇到麻烦。这被称为维度的诅咒: 控制了 100 个协变量的估计依然是有偏的, 甚至是无意义的。这说明了条件性准则在简单情境能加以运用, 但不应该上升到“准则”的地步, 尤其是在高维模型中。虽然我们只是在随机试验中讨论这个问题, 但对观察性研究同样适用。参见知识点 10.6。

如何解决维度的诅咒是一个很难的问题, 也是研究热点之一。在第十八章我们将总结一下现有研究, 并提出一些实践指导。第十一章至第十七章会介绍因果推断的相关模型。

第十章精讲点和知识点

精讲点 10.1: 可靠的置信区间 (原书第 126 页)

在重复的试验中, 至少有 95% 能覆盖真实值的有效大样本 95% 置信区间, 所需最小样本量取决于真实值大小。如果存在样本大小 n 能保证无论真实值多大 95% 置信区间都能满足上述要求时, 我们称这个有效的大样本 95% 置信区间是一致的, 或者可靠的。因为一致性缺失的时候, 在任意有限样本大小中, 可能存在覆盖真实值的频率小于 95% 的情形, 所以我们希望有一个可靠区间。遗憾的是, 对一个可靠的大样本置信区间, 这个最小值 n 是未知的, 就算用模拟也难以确定。

在本书的剩余部分, 当我们说有效的置信区间时, 我们指的是可靠的大样本置信区间。根据定义, 小样本的置信区间对于任意 n 都是一致或可靠的。

精讲点 10.2: 系统性偏移中的不确定性 (原书第 128 页)

我们常见的置信区间根据估计的标准差计算得到, 因而能反映随机误差导致的不确定性。然而, 系统性误差 (包括混杂、选择偏移和测量偏移) 也可能导致不确定性。样本量越大, 随机误差导致的不确定性在绝对值和相对值上都会越小, 因而这些置信区间, 包括 Wald 区间, 都会低估真实的不确定性。

在样本量增加的时候, 95% 置信区间中的“95%”就会被逐渐夸大, 这是因为置信区间不会考虑系统性偏移导致的不确定性, 而这个不确定性不会随着样本量的增大而减小。于是, 一些研究者认为我们应该使用一个新的名字, 比如“相容性区间”。这一新名字仅能表示在我们模型假设下, 我们计算得到的效应大校和我们的数据相容 (Amrhein 等人, 2019; Greenland, 2019)。相容度概念弱于置信度概念, 因为相容度不要求我们消除所有系统性偏移。

然而不管这个区间叫什么名字, 系统性偏移导致的不确定性经常是科研论文的讨论中心的之一。然而, 许多讨论围绕的都是系统性偏移可能的方向与强度。一些研究者认为需要定量方法来构建同时包含随机误差和系统性偏移的区间。这些方法经常被称为定量偏移分析。

知识点 10.1: 统计推断中的偏移与一致性 (原书第 127 页)

在前几章, 我们讨论了系统性偏移和一致性估计。现在我们来讨论这几者之间的关系。

为了给待估值 θ 的一致估计提供一个正式定义, 假设我们观察了 n 个独立同分布 (i. i. d.) 的随机变量, 其分布 P 在 (我们模型) 分布的集合 \mathcal{M} 之内。如果对于任一 $P \in \mathcal{M}$, 估计值 $\hat{\theta}_n$ 在概率上向 θ 收敛, 即:

$$\text{对任意 } \varepsilon > 0, n \rightarrow 0 \text{ 时, 有 } \Pr_P [\left| \hat{\theta}_n - \theta(P) \right| > \varepsilon] \rightarrow 0, P \in \mathcal{M}$$

那 $\hat{\theta}_n$ 在模型 \mathcal{M} 中就是 $\theta = \theta(P)$ 的一致估计。

如果对任意 $P \in \mathcal{M}$, $E_P[\hat{\theta}_n] = \theta(P)$, 那模型 \mathcal{M} 的估计值 $\hat{\theta}_n$ 就是一个无偏估计。 P 下的偏移被定义为 $E_P[\hat{\theta}_n] - \theta(P)$ 。我们使用 $\hat{\theta}_n$ 而非 $\hat{\theta}$ 是想强调我们的估计取决于样本大小 n 。另一方面, $\theta(P)$ 虽然是未知的, 却是固定的, 取决于 $P \in \mathcal{M}$ 。当 P 代表生成数据的分布时, 我们会省略 P , 比如直接写作 $E[\hat{\theta}_n] = \theta$ 。对于许多参数 θ , 精确无偏的估计是不存在的。

存在系统性偏移的估计不会是一致的, 也不会是无偏的。许多应用研究者 (比如流行病学家) 认为只有当估计能用来构建一个 Wald 置信区间的时候, 它才是无偏的 (Robins 和 Morgenstern, 1987)。在这个定义下, 如果估计值一致渐近于正态分布且无偏 (UANU), 那估计值才会是无偏的, 这是因为, 只有一个满足 UANU 的估计值才能用来构建模型 \mathcal{M} 中 $\theta(P)$ 的 Wald 区间。如果存在数列 $\sigma_n(P)$, 使得 Z 统计量 $(\hat{\theta}_n - \theta(P)) / \sigma_n(P)$ 按照如下定义一致收敛于某标准正态随机变量:

$$\text{对于 } t \in R, \text{ 当 } n \rightarrow \infty, \sup_{P \in \mathcal{M}} \left| \Pr_P \left[n^{1/2} (\hat{\theta}_n - \theta(P)) / \sigma_n(P) < t \right] - \Phi(t) \right| \rightarrow 0$$

则估计值 $\hat{\theta}_n$ 在模型 \mathcal{M} 中是 UANU, 其中 $\Phi(t)$ 标准正态累积分布函数 (Robins 和 Ritov, 1997)。

所有不一致估计和某些一致估计 (参见第十八章) 在这个定义下是有偏的。在正文中, 当我们说一个估计是无偏的, 如果没有另加说明, 我们就认为它是 UANU。

知识点 10.2: 条件性准则的正式定义 (原书第 131 页)

我们观测到的数据似然包含三个要素: 给定 A 和 L 时 Y 的密度, 给定 L 时 A 的密度, 以及 L 的边缘密度。让我们考虑一个剪刀的例子。 L 是一个二分变量, 互换性 $Y^a \perp\!\!\!\perp A | L$ 成立, 分层风险差 $sRD = \Pr[Y=1 | L=l, A=1] - \Pr[Y=1 | L=l, A=0]$ 在 L 的每一层中都是一个常数, 并且我们希望计算的也是分层风险差。这个数据的似然是:

$$\prod_{i=1}^n f(Y_i | L_i, A_i; sRD, p_0) \times f(A_i | L_i; \alpha) \times f(L_i; \rho)$$

其中 $p_0 = (p_{01}, p_{02})$, 而 $p_{0l} = \Pr[Y=1 | L=l, A=0]$ 、 α 和 ρ 都是干扰参数, 且分别与给定 A 的 L 时 Y 的密度、给定 L 时 A 的密度以及 L 的边缘密度相关。

在给定 A 和 L 时, 如果数据的分布取决于我们想要计算的参数, 但 A 和 L 的联合密度不和 $f(Y_i | L_i, A_i; sRD, p_0)$ 共享参数的时候, 我们会说 A 和 L 完全从属于我们想要计算的参数。此时条件性准则陈述如下: 我们应该在控制所有从属性因素的情况下进行我们关注的统计推断。因而, 研

究者需要控制所有从属性因素 $\{A_i, L_i; i=1, \dots, n\}$ 。同理, 如果风险比 (而非风险差) 在 L 的每一分层中也是常数, 那 $\{A_i, L_i; i=1, \dots, n\}$ 也是风险比的从属性因素。

知识点 10.3: 近似从属 (原书第 132 页)

假设每一分层的风险差 (sRD_l) 在 L 的每一分层中不尽相同。在我们可识别性假设下, 人群中的因果性风险差可以由标准化风险差给出:

$$RD_{std} = \sum_l [\Pr(Y=1 | L=l, A=1; v) - \Pr(Y=1 | L=l, A=0; v)] f(l; \rho)$$

这一结果取决于参数 $v = \{sRD_l, p_{0,l}; l=0,1\}$ 和 ρ (参见知识点 10.2)。在无条件的随机试验中, 因为互换性 $Y^a \perp\!\!\!\perp A$ 成立, 所以 RD_{std} 等于相关性 RD , 即 $\Pr[Y=1 | A=1] - \Pr[Y=1 | A=0]$ 。因为 RD_{std} 取决于 ρ , 所以 $\{A_i, L_i; i=1, \dots, n\}$ 不再是完全从属性因素, 并且纯粹的从属性因素也不存在。

考虑统计量 $\tilde{S} = \widehat{OR}_{AL} - OR_{AL}$, 其中 $OR_{AL} = OR_{AL}(\alpha) = \frac{\Pr[A=1 | L=1; \alpha]}{\Pr[A=1 | L=0; \alpha]} \frac{\Pr[A=0 | L=0; \alpha]}{\Pr[A=0 | L=1; \alpha]}$ 是超级人群中 $A-L$ 的比值比, \widehat{OR}_{AL} 是 OR_{AL} 的估计值。 \tilde{S} 渐近于均值为 0 的正态分布。另 $\hat{S} = \tilde{S} / \widehat{se}(\tilde{S})$, 其中 $\widehat{se}(\tilde{S})$ 是估计值 \tilde{S} 的标准差。在大样本中, \hat{S} 的分布收敛于标准正态分布, 因此 \hat{S} 能标准化 $A-L$ 之间的相关性。比如当 $\hat{S}=2$ 的时候, \hat{S} 就比其 (渐近) 均值 0 高出两个标准差。

当真实值 OR_{AL} 是已知的时候, \hat{S} 被称为 (大样本中的) 近似从属因素。以下例子可以帮助我们理解。考虑一个 $OR_{AL}=1$ 的随机试验, \hat{S} 就如同完全从属因素一样, 满足: 1) 能够从数据中计算 (即 $\hat{S} = (\widehat{OR}_{AL}-1) / \widehat{se}(\tilde{S})$) ; 2) $\hat{S} = \hat{S}(\alpha)$, 其取决于一个不出现在我们待估值中的参数 α ; 3) 似然函数可以分解为取决于 α 的 $f(A | L; \alpha)$ 和不取决于 α 的 $f(Y | L, A; v)f(L; \rho)$; 以及 4) 控制了 \hat{S} 后 RD_{std} 是无偏的, 而未调整的 RD_{std} 是有偏的 (知识点 10.4 会定义并比较调整和未调整的估计值)。任一量化 $A-L$ 相关性的统计量, 比如 $\frac{\widehat{\Pr}[A=1 | L=1]}{\widehat{\Pr}[A=1 | L=0]} - 1$, 都可以用来代替 \hat{S} 。

连续性准则认为, 我们的估计会因数据分布已知的任意细微变化而出现连续性变化 (Buehler, 1982)。如果一名研究者同时接受条件性准则和连续性准则, 那么他就应该控制近似从属因素。比如, 当 $OR_{AL}=1$ 是已知的, 如果遵循条件性准则, 我们会在 sRD_l 是常数的时候将未

调整的 RD_{std} 视作有偏估计, 而在 sRD_l 是近似常数的时候将未调整的 RD_{std} 视作无偏估计。如果一名研究者既调整了完全从属因素, 又调整了近似从属因素, 那他遵循的是广义条件性准则。

知识点 10.4: 调整和未调整的估计 (原书第 133 页)

知识点 10.3 中 RD_{std} 调整后的估计是最大似然估计 \widehat{RD}_{MLE} 。 RD_{std} 未调整的估计是 $\widehat{RD}_{UN} = \widehat{\Pr}[Y=1|A=1] - \widehat{\Pr}[Y=1|A=0]$ 。 \widehat{RD}_{MLE} 和 \widehat{RD}_{UN} 都无条件地渐近于正态分布和 RD_{std} 的无偏估计, 渐进方差分别为 $aVar(\widehat{RD}_{MLE})$ 和 $aVar(\widehat{RD}_{UN})$ 。

在正文中, 我们讨论了 \widehat{RD}_{UN} 在无条件情况下低效、有条件情况下有偏。我们现在来解释为什么。Robins 和 Morgenstern (1987) 证明了 \widehat{RD}_{MLE} 在控制和不控制近似从属因素 \hat{S} 的情况下都有相同的渐进分布, 这也就意味着 $aVar(\widehat{RD}_{MLE}) = aVar(\widehat{RD}_{MLE} | \hat{S})$ 。他们也证明了 $aVar(\widehat{RD}_{MLE}) = aVar(\widehat{RD}_{UN}) - [aCov(\hat{S}, \widehat{RD}_{UN})]^2$ 。因而 \widehat{RD}_{UN} 在无条件情况下低效, 当且仅当 $aCov(\hat{S}, \widehat{RD}_{UN}) \neq 0$, 也即 \widehat{RD}_{UN} 和 \hat{S} 无条件相关。更进一步, 条件渐近偏移 $aE[\widehat{RD}_{UN} | \hat{S}] - RD_{std}$ 等于 $aCov(\hat{S}, \widehat{RD}_{UN})\hat{S}$ 。因而, \widehat{RD}_{UN} 在有条件情况下有偏, 当且仅当它在无条件情况下低效。

当且仅当 $L \perp\!\!\!\perp Y|A$, $aCov(\hat{S}, \widehat{RD}_{UN}) = 0$ 。因此, 当结局数据 Y 存在的时候, \widehat{RD}_{MLE} 优于 \widehat{RD}_{UN} 。

知识点 10.5: 大多研究者都在遵循广义条件性准则 (原书第 134 页)

再思考一下表 10.2 中的数据。假设 sRD 在二分变量 L 的每一分层中都是常数, 那 sRD 的 MLE 方差估计就是 $\widehat{V}_0 \widehat{V}_1 / (\widehat{V}_0 + \widehat{V}_1)$, 其中 \widehat{V}_l 是 \widehat{RD}_l 的方差估计。

\widehat{V}_1 有两种计算方法: $\widehat{V}_1^{\text{obs}} = \frac{4}{80} \frac{76}{80} + \frac{2}{40} \frac{38}{40} = 1.78 \times 10^{-3}$ 或 $\widehat{V}_1^{\text{exp}} = \frac{4}{80} \frac{76}{60} + \frac{2}{40} \frac{38}{60} = 1.58 \times 10^{-3}$ 。唯一区别在于 $\widehat{V}_1^{\text{obs}}$ 除以的是每一层中的人数, 而 $\widehat{V}_1^{\text{exp}}$ 除以的是如果 $A \perp\!\!\!\perp L$ 成立时每一层中的期待人數。在数学上, $\widehat{V}_1^{\text{obs}}$ 是 观测信息的方差估计, 而 $\widehat{V}_1^{\text{exp}}$ 是基于期待的方差估计。

在本书作者的经历中, 几乎所有的研究者都会选用 $\widehat{V}_1^{\text{obs}}$ 作为方差估计。这意味着研究者都会潜意识里选择调整近似从属因素 \hat{S} , 也即遵循广义条件性准则。具体而言, 本书作者证明了 \widehat{RD}_l 的方差, 因而也即是 MLE, 在控制了 \hat{S} 和不控制 \hat{S} 的情况下不同。

知识点 10.6: 我们能消除维度的诅咒吗? (原书第 135 页)

在高维数据中, 不管用的是什么量度, 控制所有从属因素 $\{A_i, L_i; i = 1, \dots, n\}$ 后的估计都是毫无意义的。但这一命题对边缘随机试验中的无条件估计则不适用。无条件估计 \widehat{RD}_{UN} 就不会受到 L 维度的影响。然而仅当超级人群中 $A \perp\!\!\!\perp L$ 成立的时候, \widehat{RD}_{UN} 才是无偏估计。

接下来的问题是, 我们依然能用 L 计算估计值, 使得其无偏且比未调整的估计更高效吗? 在第十八章我们会论述这是可能的。

第十章图表

Table 10.1

	$Y = 1$	$Y = 0$
$A = 1$	24	96
$A = 0$	42	78

Table 10.2

$L = 1$	$Y = 1$	$Y = 0$
$A = 1$	4	76
$A = 0$	2	38

$L = 0$	$Y = 1$	$Y = 0$
$A = 1$	20	20
$A = 0$	40	40

Causal Inferences: What if ——第十章
作者：Miguel A. Hernan, James M. Robins;
翻译：罗家俊

第二部分 模型中的因果推断

Causal Inferences: What if ——第十一章
作者：Miguel A. Hernan, James M. Robins;
翻译：罗家俊

第十一章 统计模型

139 每章引言部分的行人抬头研究到这里就结束了，我们已经探索了这个假想试验的种种可能。在本书的第二部分，我们将涉足更多现实世界的数据。你可以从本书官网中下载这些数据。

本书第一部分更多是概念性的内容。我们的计算局限于能够手算的情形。从本书第二部分开始，我们需要使用计算机拟合回归模型，比如线性模型或 logistic 模型。本书假设读者有基本的统计学知识并了解流行病学的常用统计模型，因而不再赘述相关知识。本书网站会提供书中分析所用的代码，支持语言包括 SAS, R, Stata, Python。正文旁的代码注释仅仅指涉这部分正文所涉及的代码内容。

本章将讨论非参数估计和（基于模型的）参数估计之间的区别。本章同时也会回顾模型选择的常见概念，比如平滑、偏差方差权衡等。本章将论述在数据分析中使用统计模型的必要性——不管我们的目标是因果推断还是进行预测。在本章，我们暂时不会进行因果性考虑。请谨记，统计模型的相关文献浩如烟海，我们只能选择其中几个要点进行强调。

11.1 数据不会说话

以一个有 16 名 HIV 感染者的研究人群为例，与本书第一部分不同的是，我们不再认为其中的每个个体代表数千万人。相反，这 16 个人是从一个更大的目标人群——这可能是一个假想的超级人群——中随机抽样得到的。

在研究开始的时候，每个个体都会接受某种取值下的治疗 A ，并且这一治疗会在整个研究中保持不变。在研究结束的时候，我们会测定每一个体的连续性结局 Y (CD4 细胞数)。我们希望这 16 个人中 $A = a$ 时结局 Y 的均值，能和目标人群中 $A = a$ 时的结局一致。也就是说，我们的待估值是未知人群的参数 $E[Y | A = a]$ 。

在第十章我们将 $E[Y | A = a]$ 的估计 $\hat{E}[Y | A = a]$ 定义为已有数据的一个函数，并用来估计未知的人群参数。一般而言，一致估计 $\hat{E}[Y | A = a]$ 满足“样本量越大，离人群真实值 $E[Y | A = a]$ 越近”。以下两个例子都能被称为 $\hat{E}[Y | A = a]$: (i) 样本中 $A = a$ 时 Y 的均值；(ii) 数据中第一个 $A = a$ 的个体的结局取值。其中 (i) 是目标人群均值的一致估计，(ii) 不是。在实践中，我们要求所有的估计都是一致的，因而我们会用样本均值去估算目标人群均值。

(一致估计的严格定义参见第十章。)

140 假设 A 是一个二分变量并且有两个可能取值: 未接受治疗 ($A = 0$) 和接受了治疗 ($A = 1$)。样本中有一半的人群接受了治疗 ($A = 1$)。图 11.1 的散点图给出了这 16 个人的状态分布, 其中每一个点代表一个个体, Y 轴是结局取值。在第十章我们已经定义过, 目标人群中 $A = a$ 时 Y 的均值的估计值, 就是样本中 $A = a$ 时 Y 的均值。

因而, 目标人群中接受治疗时的均值就是样本中 $A = 1$ 的均值 146.25, 目标人群中未接受治疗时的均值就是样本中 $A = 0$ 的均值 67.50。在互换性成立时, 146.25 和 67.50 之间的差值, 就是目标人群中治疗 A 对结局 Y 的因果效应均值的估计。不过本章并不关注因果推断。目前, 我们的目标是详述使用模型对目标人群参数进行估计的必要性, 估计值是否具有因果意义并不在我们的主要考虑范围之内。

现在假设治疗 A 是有多个分类的分类变量, 并且有 4 种可能取值: 未治疗 ($A = 1$), 低剂量治疗 ($A = 2$), 中剂量治疗 ($A = 3$), 高剂量治疗 ($A = 4$)。在我们的样本中, 每个取值都有四分之一的人。图 11.2 描述了这一情形。为了估计目标人群中每种治疗取值下结局的均值, 我们只需计算样本的对应均值即可。估计值是: $A = 1$ 时为 70.0, $A = 2$ 时为 80.0, $A = 3$ 时为 117.5, $A = 4$ 时为 195.0。

(参见代码 11.1。)

图 11.1 和图 11.2 分别描绘了有 2 种和 4 种治疗取值时的情景。因为我们的样本只有固定的 16 个人, 所以随着分类的增多, 每一类的人数就会降低。每一分类下的样本均值是相应目标人群均值的无偏估计, 但是样本均值接近真实值的概率随着人数的降低而降低。图 11.2 中每一种分类下的 95% 置信区间宽度 (参见第十章) 将会比图 11.1 的更大。

最后, 让我们假设变量 A 表示的是治疗剂量, 单位是 mg/天, 此时它的取值范围是 0 到 100 之间的整数。图 11.3 展示了此时 16 个个体的治疗取值和他们的结局。因为治疗取值的数量远大于我们的样本量, 所以某些治疗取值中没有任何样本。比如, 在我们的样本中, 没有任何个体的治疗取值是 $A = 90$ 。

这也导致了一个问题: 我们如何用已有的样本去估计目标人群中 $A = 90$ 时结局 Y 均值? 我们在图 11.1 和 11.2 中每种治疗取值下的均值定义, 并不适用于图 11.3 的情形。如果治疗 A 是一个连续变量, 那也就不可能用样本均值去定义每一可能取值下的均值。(连续变量可以被视作有无限分类的分类变量。)

以上讨论告诉我们, 我们不可能总是让数据“说话”, 从而得到一个有意义的估计。因而,
141 我们要经常使用模型从而得到有意义的结果。

11.2 条件均值的参数估计

我们想从图 11.3 的数据中估计 $A = 90$ 时 Y 的均值, 也即 $E[Y | A = 90]$ 。假设这个值在 $E[Y | A = 80]$ 和 $E[Y | A = 100]$ 之间。事实上, 我们通常会假设不同治疗取值下 Y 的均值会随着 A 的取值线性变化。确切而言, 我们假设 $E[Y | A]$ 在 $A = 0$ 时有某个初始值 θ_0 , 且 A 每变化一个单位, $E[Y | A]$ 会相应变化 θ_1 个单位, 也即表示为:

$$E[Y | A] = \theta_0 + \theta_1 A$$

这个等式对条件均值函数 $E[Y | A]$ 的形状进行了假设 (或称限制)。这一假设被称为线性均值模型, θ_0 和 θ_1 被称为模型参数。以有限参数对条件均值函数进行表示的模型被称为参数条件均值模型。在我们的例子中, 参数 θ_0 和 θ_1 表示一条与纵轴交于 θ_0 、斜率为 θ_1 的直线。也就是说, 用这一模型表示的条件均值函数都是直线, 区别仅在于截距和斜率。

(一般而言, 我们会先限制或假设变量之间的关系, 这一假设被称为函数形式或者剂量反应曲线。我们不会使用剂量反应曲线这个称呼, 因为这一名称暗示了治疗剂量会因果性地影响反应, 然而在混杂存在的时候, 这可能会是错误的。)

接下来我们会用图 11.3 的数据以及参数均值模型来估计 a 从 0 到 100 的不同均值 $E[Y | A = a]$ 。第一步是计算 θ_0 和 θ_1 的估计值 $\hat{\theta}_0$ 和 $\hat{\theta}_1$ 。第二步是用这两个估计值来估计 $A = a$ 时 Y 的均值。比如, 要计算 $A = 90$ 时的均值, 我们只需用表达式 $\hat{E}[Y | A = a] = \hat{\theta}_0 + 90\hat{\theta}_1$ 进行计算即可。每一个个体的估计值 $\hat{E}[Y | A]$ 被称为预测值。

我们可以通过普通最小二乘法计算 θ_0 和 θ_1 的无偏估计, 接下来我们简单讲解这一方法。考虑图 11.3 中所有的可能直线, 每一条直线都是不同截距 θ_0 和斜率 θ_1 的组合。对每一条直线, 你都可以计算每个点到这条直线的垂直距离 (也即残差)。取这 16 个残差的平方, 然后加和, 所得和最小的直线就是“最小二乘”直线, 它的参数 $\hat{\theta}_0$ 和 $\hat{\theta}_1$ 就是“最小二乘”估计值。我们用简单的线性代数知识就能求出 $\hat{\theta}_0$ 和 $\hat{\theta}_1$, 具体方法参考其他数学或统计教材。

在我们的例子中, 参数估计值 $\hat{\theta}_0 = 24.55$, $\hat{\theta}_1 = 2.14$, 其对应的直线如图 11.4 所示。因而, $A = 90$ 时 Y 的预测均值为 $\hat{E}[Y | A = a] = 24.55 + 90 \times 2.14 = 216.9$ 。因为普通最小二乘法会用到所有的数据点, 因而 $A = a$ 时 Y 的均值, 也即 $E[Y | A = a]$, 是通过 $A \neq a$ 时的信息估计而得。

(参见代码 11.2。在残差的方差不取决于 A 的假设(即方差齐性假设)下, θ_0 的 95% Wald 置信区间是 (-21.2, 70.3), θ_1 的是 (1.28, 2.99), $E[Y|A=a]$ 的是 (172.1, 261.6)。)

所以, 模型是什么呢? 模型是我们先验地对数据联合分布进行假设(或称限制)。我们的线
142 性条件均值模型将条件均值函数 $E[Y|A=a]$ 假设为一条直线。这一假设信息表现在参数 θ_0 和 θ_1 之中。虽然是一种先验的假设, 但是参数模型能弥补数据中信息不全这一缺陷。

从参数模型中得到的参数估计能让我们对感兴趣的数值进行估计。比如在我们的例子中, 研究人群里没有 $A=90$ 的个体, 但我们却想估计目标人群中 $A=90$ 的结局均值。不过构建参数模型也需要付出一定代价。在我们使用参数模型的时候, 只有当我们的先验假设是正确的, 我们从模型中得到的推断才可能是正确的。因而, 基于模型的因果推断——本书剩下部分的主要内容——依赖于不存在模型设定错误。然而, 因为鲜有完全正确的模型设定, 所以一定程度的错误设定总是会存在。不过非参数估计可以纠正其中部分错误, 下一小节将讨论这一话题。

11.3 条件均值的非参数估计

让我们回到图 11.1 的数据中。此时治疗 A 是一个二分变量, 我们想得到 $E[Y|A=0]$ 和 $E[Y|A=1]$ 的一致估计。我们决定用以下线性模型来估计这两个数值:

$$E[Y|A] = \theta_0 + \theta_1 A$$

此时 $E[Y|A=0] = \theta_0 + 0 \times \theta_1 = \theta_0$, $E[Y|A=1] = \theta_0 + 1 \times \theta_1 = \theta_0 + \theta_1$ 。我们使用最小二乘法得到参数 θ_0 和 θ_1 的估计值, 分别是 $\hat{\theta}_0 = 67.5$ 和 $\hat{\theta}_1 = 78.75$ 。于是我们得到 $\hat{E}[Y|A=0] = 67.5$ 和 $\hat{E}[Y|A=1] = 146.25$ 。在这里, 我们用模型计算了 Y 的均值的估计值, 这里的结果和 11.1 小节的结果一样。这并不是巧合。

(参见代码 11.2。)

让我们再思考一下当 A 是二分变量时的模型 $E[Y|A] = \theta_0 + \theta_1 A$ 。如果我们将其写作 $E[Y|A=1] = E[Y|A=0] + \theta_1$, 我们会发现治疗组的均值 $E[Y|A=1]$, 只是非治疗组的均值 $E[Y|A=0]$ 再加上了一个数值 θ_1 , 这个 θ_1 可以为正, 可以为负, 也可以是零。也就是说, 这个模型并未对 $E[Y|A=0]$ 和 $E[Y|A=1]$ 做出任何限制或假设。因此, 如果 $E[Y|A] = \theta_0 + \theta_1 A$ 的 A 是一个二分变量, 那么它就不是一个模型。没有施加限制或假设的“模型”被称为“饱和模

型”。不过, 即使它们并不符合我们对模型的定义, 但它们看起来很像模型, 因此也经常被称为模型。

(在本书中, 我们将“模型”定义为对真实状况的先验假设 (Robins 和 Greenland, 1986)。本书第一部分仅讨论了饱和模型。)

一般而言, 如果一个条件均值模型的参数个数等于人群中未知的条件均值个数, 那么这个模型就是饱和模型。比如, 线性模型 $E[Y|A] = \theta_0 + \theta_1 A$ 有两个参数, 在 A 是二分变量的时候, 人群

143 中只存在两个未知条件均值: $E[Y|A=0]$ 和 $E[Y|A=1]$ 。因为这个模型没有对它的参数施加任何限制, 所以也就没有对要估计的均值做出任何限制。与之相比, 图 11.3 中 A 的取值范围是 0 到 100。线性模型 $E[Y|A] = \theta_0 + \theta_1 A$ 仅有两个参数, 却要被用来估计 101 个数值, 也即 $E[Y|A=0]$, $E[Y|A=1]$ …… $E[Y|A=100]$ 。要用两个参数来无偏地估计 101 个数值, 唯一的希望是这 101 个数值都落在同一条直线上。当一个模型仅有为数不多的几个参数, 却要用来估计人群中的大量数值时, 我们会说这个模型称是“简约的”。

(一个饱和的模型等号两边的未知数个数相等。)

没有先验假设的模型中得到的估计, 我们将其定义为非参数估计 (更严格定义参见精讲点 11.1)。 A 是二分变量时的人群均值 $E[Y|A=a]$ 就是一个例子, 此时其估计就是非参数估计, 等于样本均值, 或者说等于饱和模型的估计。当 A 是离散的, 有 100 种可能取值且样本中不存在 $A=90$ 时, 就不存在 $E[Y|A=90]$ 的非参数估计。我们在本书第一部分介绍的因果推断方法——包括标准化、逆概率加权、分层以及匹配——都是饱和模型下人群数值的非参数估计, 这是因为这些方法并没有对效应估计值施加任何先验的限制。与之相比, 本书第二部分将要介绍的方法多是参数估计方法, 它们会对部分数据的分布做出限制。当数据信息信息不全面、不能替自己“说话”的时候, 使用参数估计或者借助其他信息的方法成了我们的唯一希望。

(对因果推断而言, 可识别性假设必须时刻保持, 即使我们有无限量的数据。而模型假设则是因为我们没有无限量的数据。)

11.4 平滑

让我们再考虑一下图 11.3 的数据和线性方程 $E[Y|A] = \theta_0 + \theta_1 A$ 。参数 θ_1 表示治疗剂量 A 每上升一个单位, 结局均值的变化情况。因为 θ_1 是一个单一数值, 所以在 A 的取值范围内, A 的每

单位变化, 对应的 Y 的变化是不变的。也就是说, 这个模型将结局均值强制为一个关于治疗剂量 A 的线性函数。图 11.4 给出了最佳拟合直线。

但是在现实中, 有很多情况是单位治疗剂量在低剂量时对应的结局变化, 大于在高剂量时对应的变化。也就是说, 在治疗剂量达到一定高度后, 治疗效果会有所下降。在这种情况下, 模型 $E[Y|A] = \theta_0 + \theta_1 A$ 就是不正确的。但是我们可以灵活应用“线性”模型。

(注意: “线性模型”可能有两种不同的含义: 一种是指对参数进行线性组合的模型, 一种是指对变量函数进行线性组合的模型。后一种情形的函数不一定是线性函数。)

比如, 假设我们用图 11.3 的数据拟合模型 $E[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$, 其中 $A^2 = A \times A$ 。这个模型也被称为线性模型, 因为这里的条件均值被表述为几个变量 (A 和 A^2) 与相关参数 (θ_1 和 θ_2) 的线性组合。然而, θ_2 不等于 0 的时候, 参数 θ_0 、 θ_1 和 θ_2 将会给出一条曲线——一条抛物线——而非一条直线。

模型 $E[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$ 的参数也可以通过最小二乘法估计。图 11.5 给出了拟合曲线。参数估计值分别为 $\hat{\theta}_0 = -7.41$, $\hat{\theta}_1 = 4.11$, $\hat{\theta}_2 = -0.02$ 。 $A = 90$ 时 Y 的预测值就是 $\hat{E}[Y|A=90] = \hat{\theta}_0 + 90\hat{\theta}_1 + 90 \times 90\hat{\theta}_2 = 197.1$ 。

(参见代码 11.3。在方差齐性假设下, $\hat{E}[Y|A=90]$ 的 95% Wald 置信区间是 (142.8, 251.5)。)

我们可以再增加一个三次项 $\theta_3 A^3$ 、四次项 $\theta_4 A^4$ ……直至十五次项 $\theta_{15} A^{15}$, 此时我们的参数数目就等于我们研究人数。曲线的形状会随着参数的增加而改变。总体而言, 模型中参数越多, 拟合曲线上的拐点越多。

也就是说, 参数越多, 拟合曲线越曲折, 或者换个词, 越不平滑。如果线性模型只有 2 个参数, 那就是一条直线, 也就是最平滑的曲线。如果线性模型的参数数目等于数据点数目, 那就是最不平滑的曲线, 此时有多少数据点, 就有多少拐点, 换句话说, 我们只是将模型插入到数据当中, 也即样本的每一个数据点都等于预测值。

大多数情况下, 构建模型可以视作将含有噪声的数据转换为或多或少相对平滑的曲线。平滑处理是指借助数据的信息去预测不同变量组合下的结局, 即用不含 a 的样本信息去预测 $E[Y|A=a]$ 。所有的参数估计都包含一定程度的平滑处理。

平滑的程度取决于我们从数据点借助了多少信息。只有两个参数的模型 $E[Y|A] = \theta_0 + \theta_1 A$ 用了所有样本信息去估计 $E[Y|A=90]$ 。而参数数目和数据点数目一样多的模型，虽然借助了数据信息去估计（用插补的方法）样本中不存在的 A 对应的结局，但并没有借助数据信息去估计样本中已有的 A 对应的结局。

为了得到适度的平滑程度，我们可以选择适当的参数数目，或者限制用来估计结局的样本数据数目。比如，为估计 $E[Y|A=90]$ ，我们可以只在治疗剂量在 80 和 100 的样本中拟合 $E[Y|A] = \theta_0 + \theta_1 A$ 。也就是说，我们只借用了 $A=90$ 上下浮动 10 个单位以内的样本信息。这个范围越大，得到的结果也就越平滑。

（我们例子中的模型仅涉及连续变量。不过以上讨论对其他分类变量的模型，比如 *logistic* 模型，同样适用（参见知识点 11.1）。）

在我们上述简化例子中，所有模型都只有一个变量（不管是只有 A 及其对应参数的模型，还是有 A 和 A^2 及其对应的两个参数的模型），因此得到的曲线可以在二维空间展现出来。在现实应用中，一个模型经常包含多个参数，因此模型对应的“曲线”通常是高维曲面。无论维度怎样变化，平滑的概念总是保持不变：模型的参数越少，模型对应的曲面也就越平滑。
145

11.5 偏差方差权衡

在前几节，我们用图 11.3 的 16 个个体来估计目标人群的结局 Y 在治疗剂量 $A=90$ 时的均值 $E[Y|A=90]$ 。因为样本中没有人 $A=90$ ，所以我们需要使用模型。我们用数据拟合了一个线性模型。估计值 $\hat{E}[Y|A=90]$ 在不同的模型中有所不同。在有 2 个参数的模型 $E[Y|A] = \theta_0 + \theta_1 A$ 当中，估计值是 216.9 (95% 置信区间: 172.1, 261.6)。在有 3 个参数的模型 $E[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$ 当中，估计值是 197.1 (95% 置信区间: 142.8, 251.5)。两个不同的参数模型给出了两个不同的结果。哪一个是对的？216.9 和 197.1，哪一个更接近目标人群的真实值？

如果剂量和结局之间的关系是非线性的，那 2 个参数的模型中得到的估计就是有偏的，因为这个模型假设这个关系是线性的。另一方面，如果剂量和结局之间的关系是线性的，那两个模型的估计都是有效的，这是因为 $E[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$ 既能满足线性的情况（此时 $\theta_2 = 0$ ），又能满足抛物线的情况（此时 $\theta_2 \neq 0$ ）。一个保险的选择是使用含 3 个参数的模型

$E[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$, 而非 2 个参数的 $E[Y|A] = \theta_0 + \theta_1 A$ 。这是因为前者能满足两种不同情形, 因此有偏的可能性更小。总体而言, 模型的参数越多, 模型的先验假设就越少, 模型也就越不光滑, 而因模型错误设定导致的偏移也越少。

虽然较不光滑的模型会给出不那么有偏的估计值, 但是却会导致更大的方差, 也就是估计值的 95% 置信区间会更宽。比如, 在我们之前的例子中, 3 个参数的模型中估计值 $\hat{E}[Y|A=90]$ 的 95% 置信区间, 就宽于 2 个参数的模型的置信区间。然而, 如果 2 个参数的模型中估计值 $\hat{E}[Y|A=90]$ 是有偏的, 那么其对应的标准(名义)95% 置信区间也就不是被校准的, 也就是说, 它并没有在 95% 的时间中覆盖真实值 $E[Y|A=90]$ 。

(精讲点 11.2 讨论了频率学派和贝叶斯学派中, 模型维度对置信区间的不同含义。)

在许多数据分析中, 偏差方差权衡一直是核心考虑之一。使用模型的研究者需要衡量用来减少偏差的方法——比如增加参数——是否值得, 因为它们通常会提高方差。虽然存在一些规范能帮研究者做决策, 但是在现实中, 大多数研究还是基于常用方法、参数的阐释意义、以及可用的软件来决定使用什么模型。在本书中, 我们通常会假设我们的参数模型设定都是正确的。这个假设并不现实, 但能让我们将精力放在和因果推断密切相关的问题上。毕竟, 不管估计值是否具有因果意义, 模型设定错误是所有数据分析中都需要面对的问题。在实践中, 谨慎的研究者总是会质疑自己模型的有效性, 并总是会分析最终估计值对模型设定错误的敏感性。

接下来, 我们将会介绍如何用模型进行因果推断。

第十一章精讲点和知识点

精讲点 11.1: Fisher 一致性 (原书第 143 页)

本书正文中非参数估计的定义和统计中 Fisher 一致估计的定义相同 (Fisher, 1922)。也即, 用整个人群而非样本计算得到的数值是整个人群的真实值。在这个定义下, Fisher 一致估计不存在任何先验假设, 但许多人群数值的 Fisher 一致估计值可能并不存在。严格来说, Fisher 一致估计是饱和模型下的非参数最大似然估计。

在统计学中, 非参数估计有时被用来指代有较弱先验假设的估计, 而非 Fisher 一致估计, 比如在核回归模型中。更多讨论参见知识点 11.1。

精讲点 11.2: 模型维度以及频率学派和贝叶斯学派的置信区间 (原书第 145 页)

在频率学派看来, 概率是由频率定义的; 在贝叶斯学派看来, 概率是由信念程度定义的。两者大相径庭。第十章讲述了频率学派的置信区间。而贝叶斯学派的可信区间的阐释性更加自然: 贝叶斯学派的 95% 可信区间表示在给定的观察数据中, 有 95% 的概率待估值会在这个区间当中。然而, 因为需要具体给出研究者的信念程度, 所以在现实中贝叶斯可信区间较少使用。

有趣的是, 在简单、低维且样本量够大的参数模型中, 95% 贝叶斯可信区间和 95% 频率置信区间相同, 而在高维时, 两者可能并不相同, 此时 95% 贝叶斯可信区间能够包含待估值的概率远小于 95%。主要原因是贝叶斯方法需要给定未知参数的先验分布。在低维模型中, 数据的信息量可能远远超过先验分布的影响, 因而使用贝叶斯方法得到的结果对不同的先验分布并不敏感。然而这在高维模型中不存在。因此, 如果真实的参数值并不在给定的先验分布中, 贝叶斯可信区间的中心就会远离真实值、更趋向先验分布中最大概率的参数值。

知识点 11.1: 常用模型分类 (原书第 147 页)

正文中的线性条件均值模型可以表示为 $E[Y|X] = \theta X = \sum_{i=0}^p \theta_i X_i$, 其中 X 是一个向量, 包含所有变量 X_0, X_1, \dots, X_p , 其中, 对所有样本都有 $X_0 = 1$ 。这一类模型只是更大一类模型的一部分。更大一类模型包含两个部分: 包含自变量的线性函数部分 $\sum_{i=0}^p \theta_i X_i$, 和联系函数部分 $g\{\cdot\}$, 比如 $g\{E[Y|X]\} = \sum_{i=0}^p \theta_i X_i$ 。

正文中的线性条件均值模型用的是恒等联系函数。如果结局是必须是正数 (比如个数、新增病例等), 模型中会经常使用对数联系函数, 也即 $\log\{E[Y|X]\} = \sum_{i=0}^p \theta_i X_i$, 因此 $E[Y|X] = \exp\left(\sum_{i=0}^p \theta_i X_i\right)$ 。二分变量 (即取值只能为 0 或 1) 的条件均值模型经常使用 logit 联系函数, 即 $\log\left\{\frac{E[Y|X]}{1-E[Y|X]}\right\} = \sum_{i=0}^p \theta_i X_i$, 这样就能保证预测值在 0 和 1 之间。使用 logit 函数的条件均值模型被称为 logistic 回归模型, 在本书中经常会用到。这三个联系函数也被称为标准联系函数。我们可以在恒等联系的普通线性模型中、对数联系的泊松模型中、以及 logit 联系的 logistic 模型中使用最大似然法估计参数 θ 。只要模型是正确的, 这些模型中的估计值都是 θ 的一致估计。广义估计方程 (GEE) 模型经常被用来处理重复测量, 是条件均值模型的一个推广。

条件均值模型只估计 $E[Y|X]$, 并没有对 $Y|X$ 的分布或 X 的边缘分布作出任何限制。因此, 如果 X 或 Y 是连续的, 参数条件均值模型就是 (X, Y) 联合分布的半参数模型, 因为我们对分布的一部分进行了建模, 但对其他部分没有进行和限制。一个模型是半参数的是因为联合分布中未加限制部分的集合不能用有限个参数进行表示。

我们可以放宽 $E[Y|X]$ 必须表示为参数形式这一假设, 从而进一步延伸条件均值模型。比如, 核回归模型并没有给出 $E[Y|X]$ 的函数形式, 而是对任意 x , 由

$\sum_{i=1}^n w_h(x - X_i) Y_i / \sum_{i=1}^n w_h(x - X_i)$ 估计 $E[Y|X]$, 其中 $w_h(z)$ 是一个正函数, 也被称为核函数, 在

$z = 0$ 时取得最大值, 而随着 $|z|$ 的增大逐渐减小到 0, 减缓速率取决于 w 的下角标 h 。再比如, 在广义相加模型 (GAM) 中, 自变量的线性组合 $\sum_{i=0}^p \theta_i X_i$ 被替换为一系列平滑函数的和 $\sum_{i=0}^p f_i(X_i)$ 。

此时模型可以通过向后拟合算法进行估计, 例如从核回归的第 k 次迭代中得到的 $f(\cdot)$ 进行拟合。

在正文中, 我们讨论了参数模型的平滑概念, 而这些模型都会先验地假设 $E[Y|X=x]$ 的函数形式, 比如说假设它是一条抛物线。因此, 在估计 $E[Y|X=x]$ 的时候, 我们会从 X 远离 x 的数据中借助信息进行估计。与之相比, 核回归模型并不会先验地假设函数形式, 并且仅借助 X 在 x 附近的数据信息。因而, 核回归模型是一个典型的“非参数”回归模型。然而这里的“非参数”与我们之前的定义不同。我们之前将 $E[Y|X=x]$ 的非参数估计定义为仅借助 X 等于 x 的数据信息。在这里, 核回归的“非参数”指的是借助 X 在 x 附近的数据信息, 这个宽度取决于平滑参数 h , 有时也称带宽。

第十一章图表

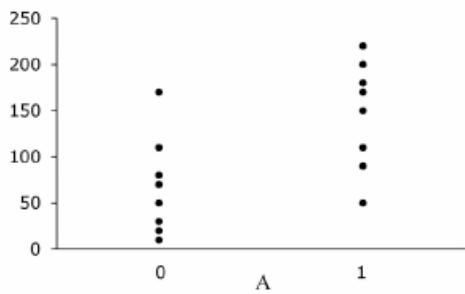


Figure 11.1

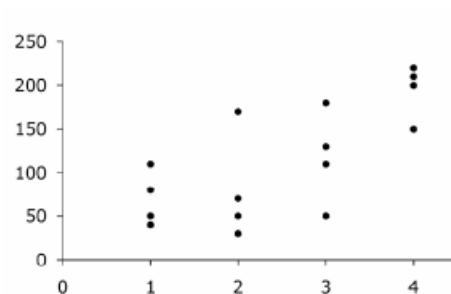


Figure 11.2

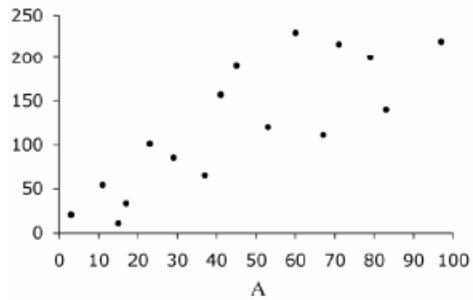


Figure 11.3

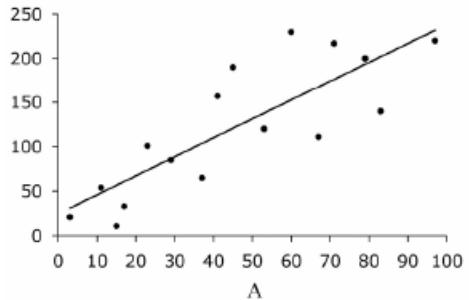


Figure 11.4

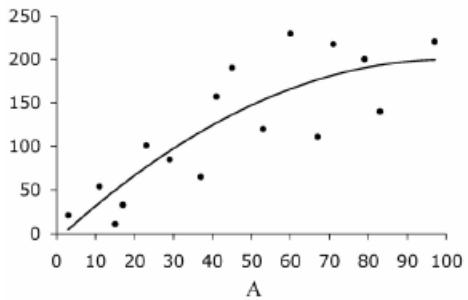


Figure 11.5

第十二章 逆概率加权和边缘结构模型

149 本书第二部分会围绕以下因果性问题展开: 戒烟对体重增加的因果效应均值是多少? 在本章, 我们会讲述如何使用逆概率加权从观察性数据中估计这一效应。虽然在第二章我们就介绍了逆概率加权, 但我们仅仅讨论了它的非参数形式。现在我们会讲述如何在模型中使用逆概率加权。在一定假设下, 这一方法能解决高维数据带来的种种问题, 并且适用于非二分的治疗变量。

我们将用现实世界中 NHEFS 的数据来估计戒烟对体重增加的效应。NHEFS 表示“美国国家健康与营养调查数据 1 期: 流行病学跟踪研究”。NHEFS 是由美国国家健康数据中心、老龄化研究所、以及其他美国公共卫生服务机构共同发起的研究。NHEFS 的详尽介绍及数据能在以下网址中找到: www.cdc.gov/nchs/nhanes/nhefs/。在本章及接下来的章节中, 我们会用到 NHEFS 数据的一部分, 这部分数据可以在本书官网上下载。

12.1 因果性问题

我们的目标是估计戒烟 A (治疗变量) 对体重增加 Y (结局变量) 的因果效应均值, 因而我们从 NHEFS 数据中选取了 1566 名 25 至 74 岁的参与人员, 他们有第一次访问时的数据, 并在 10 年后被再次随访。如果他们在随访之前戒烟了, 那就被称为治疗组 $A=1$, 如果没有则是非治疗组 $A=0$ 。在初次访问和随访中都记录了每个人的体重 (kg), 因而可以计算出体重增加了多少。大多数人体重都增加了, 但平均而言, 戒烟者增加得更多。戒烟者的平均体重增加值是 $\hat{E}[Y|A=1]=4.5$, 非戒烟者是 $\hat{E}[Y|A=0]=2.0$ 。差值是 2.5, 对应的 95% 置信区间为 1.7 到 3.4。传统的零假设统计检验——也即差值不为零的统计检验——给出的 P 值小于 0.001。

(我们将分析局限于有相关信息的人群当中, 这些信息包括两次访问时的性别、年龄、种族、体重、身高、教育程度、饮酒情况、烟龄和吸烟频率, 以及初次访问时的病史信息。参见精讲点 12.1。)

我们将 $E[Y^{a=1}]$ 定义为如果人群中所有人都戒烟所对应的体重增加均值, 将 $E[Y^{a=0}]$ 定义为如果人群中所有人都没有戒烟所对应的体重增加均值。我们将戒烟的因果效应均值定义为这两者之差, 也即 $E[Y^{a=1}]-E[Y^{a=0}]$ 。这是本章及接下来几章将会关注的因果效应。

一般而言, 我们在第一段中得到的相关性差值, 即 $E[Y|A=1]-E[Y|A=0]$, 与因果性差值 $E[Y^{a=1}]-E[Y^{a=0}]$ 有所区别。如果戒烟者和非戒烟者的特征分布有所不同, 那么相关性差值就没有因果性意义。比如, 如果平均而言戒烟者比非戒烟者大 4 岁 (戒烟者比非戒烟者多 44% 的可

150 能性会大于 50 岁), 同时不管戒不戒烟年纪更大的人体重增加值都会低于年轻的人, 那么此时相关性差值就没有因果意义。于是, 我们将年龄称为 A 对 Y 效应的混杂变量(的替代), 因而我们的分析就需要调整年龄。未调整年龄时的估计值 2.5 也许低估了真实的因果效应

$$E[Y^{a=1}] - E[Y^{a=0}]。$$

(精讲点 7.3 定义了什么是混杂变量的替代。)

如表 12.1 中所展示的, 戒烟者和非戒烟者在性别、种族、教育程度、初次访问时的体重以及吸烟频率等特征方面并不相同。如果这些变量是混杂变量, 那我们就需要在分析中调整这些变量。在本书第十八章我们将讨论混杂变量的选取。在此处, 我们假设初次访问时的 9 个变量就足以代表所有混杂, 这些变量包括: 性别 (sex¹, 0: 男性, 1: 女性), 年龄 (age, 单位: 年), 种族 (race, 0: 白人, 1: 其他), 教育程度 (education, 5 个分类), 吸烟频率 (smokeintensity, 单位: 支/天), 烟龄 (smokeyrs, 单位: 年), 每日体力情况 (active, 3 个分类), 运动量 (exercise, 3 个分类), 以及初次访问时的体重 (wt71, 单位: kg)。也即, 我们常用来表示混杂变量的 L 此时是一个含 9 个变量的向量。下一节我们将用逆概率加权的方法来调整这些变量。

(代码 12.1 给出了本节涉及变量的统计分布。)

12.2 使用模型计算逆概率权重

逆概率加权将会构建一个虚拟人群, 在这个虚拟人群中, 从混杂变量 L 到治疗变量 A 之间的箭头将会被移除。更准确一些, 这个虚拟人群将有两个性质: A 和 L 在统计上相互独立, 且虚拟人群的均值 $E_{ps}[Y | A = a]$ 将等于实际人群的标准化均值 $\sum_l E[Y | A = a, L = l] \Pr[L = l]$ 。即使在

151 实际人群中有界互换性 $Y^a \perp\!\!\!\perp A | L$ 不成立, 这两个性质也是为真的(参见知识点 2.3)。如果有界互换性 $Y^a \perp\!\!\!\perp A | L$ 在实际人群中成立, 那从这两个性质我们可以推出: (1) 虚拟人群和实际人群中 Y^a 的均值相等; (2) 无条件互换性(也即不存在混杂)在虚拟人群中成立; (3) 实际人群中的反事实均值 $E[Y^a]$ 等于虚拟人群中的 $E_{ps}[Y | A = a]$; (4) 虚拟人群中的相关性就是因果性。逆概率加权的相关概念请参见第二章。

一般而言, 我们可以用实际治疗情况的条件概率的倒数对每个人进行加权, 从而构建虚拟人群。每个人治疗情况 A 的逆概率权重被定义为 $W^A = 1 / f(A | L)$ 。对于我们的二分治疗变量 A ,

¹ 本段括号内的英文表示这些变量在代码中的名称。

作者: Miguel A. Hernan, James M. Robins;

翻译: 罗家俊

戒烟者的权重的分母是控制了变量之后的条件概率 $\Pr[A=1|L]$, 非戒烟者是 $\Pr[A=0|L]$ 。我们只需要计算 $\Pr[A=1|L]$ 就行, 因为 $\Pr[A=0|L]=1-\Pr[A=1|L]$ 。

(接受治疗的条件概率 $\Pr[A=1|L]$ 也被称为倾向性评分。更多倾向性评分的相关内容参见第十五章。)

在第 2.4 小节, 我们用非参数化的方法估计了 $\Pr[A=1|L]$: 我们在 L 的每一分层中数了数有多少人是接受治疗的 ($A=1$), 然后再用这个数目除以这一分层中的总人数。这些计算所需的信息能从有 4 个分枝 (2 (L 的两个取值) 乘以 2 (A 的两个取值)) 的因果性结构树状图中获得。但是, 在面对高维数据、且某些变量有多个分层的时候, 这一非参数的计算方法就变得不切实际。即使我们的 9 个混杂变量每个最多只有 6 个分层, 那对应的树状图也会有超过两百万个分枝。如果我们使用烟龄、吸烟频率等连续变量的实际取值范围, 那还会有更多分枝。因而, 使用非参数方法, 我们不可能估计这 1566 个人的对应条件概率。我们需要借助模型。

(在第十章我们讨论了“维度的诅咒”。)

为了得到 $\Pr[A=1|L]$ 的参数估计值, 我们需要拟合一个 logistic 回归方程, 用它来预测控制了 9 个混杂变量之后戒烟的概率。对于连续变量年龄、体重、吸烟频率、烟龄, 我们在模型中包含了它们的一次项和二次项, 同时, 我们在模型中没有包含任何两个变量的乘积项。也就是说, 我们的模型假设 $\Pr[A=1|L]$ 的比值对数和这些连续变量之间的关系是抛物线关系, 同时这些自变量的影响时相互独立的。在这个模型假设下, 我们能得到不同 L 组合所对应的估计值 $\widehat{\Pr}[A=1|L]$, 因而也就能得到样本中 1566 个人的对应概率估计值。

(代码 12.2 估计了样本中的逆概率权重 W^A , 其取值范围是 1.05 到 16.7, 平均值为 2。)

接着我们用逆概率权重构建虚拟人群, 再在虚拟人群中计算 $\widehat{E}_{ps}[Y|A=1]-\widehat{E}_{ps}[Y|A=0]$ 。如果在虚拟人群中, A 的因果效应不再受其他混杂变量的影响, 且 $\Pr[A=1|L]$ 的模型是正确设定的, 那么虚拟人群中的相关性就是因果性, 且虚拟人群的相关性差值

$E_{ps}[Y|A=1]-E_{ps}[Y|A=0]$ 的无偏估计, 也就等于实际人群的因果性差值 $E[Y^{a=1}]-E[Y^{a=0}]$ 的无偏估计。

($E[Y|A]=\theta_0+\theta_1A$ 是一个饱和模型因为它就只有 2 个参数 θ_0 和 θ_1 , 而它需要用来估计两个数值 $E[Y|A=1]$ 和 $E[Y|A=0]$ 。在这个模型中, $\theta_1=E[Y|A=1]-E[Y|A=0]$ 。)

在虚拟人群中, 我们通过加权最小二乘法拟合 (饱和) 线性均值模型 $E[Y|A] = \theta_0 + \theta_1 A$, 从而估计 $E_{ps}[Y|A=1] - E_{ps}[Y|A=0]$, 此时每个人的逆概率权重估计值 \widehat{W} 为: 戒烟者是
 152 $1/\widehat{\Pr}[A=1|L]$, 非戒烟者是 $1/(1-\widehat{\Pr}[A=1|L])$ 。参数估计值 $\widehat{\theta}_1$ 是 3.4。也即, 平均而言, 戒烟者的体重增加比非戒烟者多 3.4kg。估计的正式定义请参见知识点 12.1。

为了得到点估计 $\widehat{\theta}_1$ 的 95% 置信区间, 我们需要将逆概率权重考虑在内。一个可行方法是利用统计理论推导出方差估计。这一方法要求研究人员对估计过程进行编程, 而普通统计软件并不支持这一过程。第二种方法是通过非参数的自举法 (bootstrapping, 参见知识点 13.1) 从而得到参数方差的近似值。如果面对超大型数据, 这一方法对算力和时间都会有所要求。第三种方法是使用稳健方差估计 (比如广义估计方程模型中会加入独立的作业相关矩阵), 这一方法在大多数统计软件中都能运行。从稳健方差估计中得到的 95% 置信区间是有效的, 但与前两种方法相比, 则会更加保守, 因为理论上这种方法得到的 95% 置信区间在大于 95% 的时间中会覆盖超级人群的真实值。通过这一方法得到的 $\widehat{\theta}_1$ 的 95% 置信区间是 (2.4, 4.5)。在本章, 所有逆概率加权估计值的置信区间都是保守的。如果 $\Pr[A=1|L]$ 的模型设定错误, 那么 θ_0 和 θ_1 的估计值就会是有偏的, 并且我们在上一章讨论过, 此时 95% 的置信区间在小于 95% 的时间中会覆盖真实值。

(加权最小二乘法是要取得 $\sum_i \widehat{W}_i [Y_i - (\theta_0 + \theta_1 A_i)]^2$ 的最小值, 其中 W 是权重。如果 $\widehat{W}_i = 1$, 那这就是上一章所说的普通最小二乘法。估计值 $\widehat{E}[Y|A=a] = \widehat{\theta}_0 + \widehat{\theta}_1 a$ 等于 $\frac{\sum_{i=1}^{Y_i} \widehat{W}_i}{\sum_{i=1}^n \widehat{W}_i}$, 对样本中所有 $A=a$ 的个体求和。)

153 12.3 逆概率稳定权重

逆概率加权的目标是构建一个虚拟人群, 其中混杂变量 L 和治疗变量 A 不再相关。在逆概率权重 $W^A = 1/f(A|L)$ 构建的虚拟人群中, 所有样本人员都会被他们自身的两个拷贝取代。一个拷贝的治疗取值是 $A=1$, 另一个是 $A=0$ 。我们在第二章展示了如何将图 2.1 的原始人群, 转化为图 2.3 的虚拟人群。我们留意到, W^A 的均值是 2, 这是因为在虚拟人群中, 每个人都有两部分, 一部分在治疗组中, 另一部分在非治疗组中。

不过, 除此之外, 我们还有其他方法来构建虚拟人群, 从而让其中的 A 和 L 相互独立。比如, 我们可以构建一个无论 L 是什么, $A=1$ 的概率是 50%、 $A=0$ 的概率也是 50% 的虚拟人群, 这样一个虚拟人群的逆概率权重是 $0.5/f(A|L)$ 。这一新的虚拟人群将会和原样本的人数一样多, 并且在数学上等价于逆概率权重均值为 2 的虚拟人群。此时, $0.5/f(A|L)$ 的均值是 1, 从这一虚拟人群中得到的点估计, 等于之前逆概率权重为 $1/f(A|L)$ 的虚拟人群中得到的点估计。

(你可以用图 2.1 的数据验证这一说法。证明参见知识点 12.2。) 逆概率权重为 $p/f(A|L)$ 时同理, 其中 $0 < p \leq 1$ 。权重 $W^A = 1/f(A|L)$ 只是 $p=1$ 时的一个特例。

让我们进一步往前推理。调整混杂的目的是让治疗变量 A 的概率不取决于混杂变量 L 。于是我们可以构建一个虚拟人群来达到这个目的, 而其中每个人接受治疗 A 的概率是否相同则无关紧要, 只要这个概率不取决于 L 就行。比如, 一个常用做法是让虚拟人群中接受治疗的概率等于原本样本中治疗的概率 $\Pr[A=1]$, 对于未接受治疗的概率同理。因而, 此时的逆概率权重是: 治疗组是 $\Pr[A=1]/f(A|L)$, 非治疗组是 $\Pr[A=0]/f(A|L)$ 。或者写作 $f(A)/f(A|L)$ 。

(治疗组中的因果效应均值可以通过逆概率权重进行估计, 其中分子是 $\Pr[A=1]$ 。参见知识点 4.1。)

对图 2.1 的数据使用逆概率权重 $f(A)/f(A|L)$, 可以得到图 12.1, 其中 $\Pr[A=1] = 13/20 = 0.65$, $\Pr[A=0] = 7/20 = 0.35$ 。在第三章的可识别条件下, 虚拟人群就如一次假想的随机试验, 其中 65% 的人员被随机分到治疗组 $A=1$, 35% 的被随机分到非治疗组 $A=0$ 。注意, 为了保证 65/35 的比例, 每一分枝的人数不可能是整数。不过在数学上, 是不是整数不是一个大问题。

在我们的例子中, 逆概率权重 $f(A)/f(A|L)$ 的取值范围是 0.33 到 4.30, 而 $1/f(A|L)$ 的是 1.05 到 16.7。分子中的 $f(A)$ 被称为稳定因子, 会缩小 $f(A)/f(A|L)$ 的取值范围。

$W^A = 1/f(A|L)$ 被称为非稳定权重, 而 $SW^A = f(A)/f(A|L)$ 被称为稳定权重。稳定权重的均值是 1, 这是因为此时虚拟人群和原样本的人数一样。

154 让我们用稳定权重 SW^A 再估计一下戒烟对体重增加的因果效应。首先, 我们需要估计条件概率 $\Pr[A=1|L]$, 这将被用作分母。我们会用 12.2 小节中的 logistic 模型得到每个人的参数估计值 $\widehat{\Pr}[A=1|L]$ 。其次, 我们需要估计 $\Pr[A=1]$, 这是权重的分子。我们可以用非参数化的方

作者: Miguel A. Hernan, James M. Robins;

翻译: 罗家俊

法——即戒烟者占总样本的比例 403/1566——得到这一数值, 也可以通过拟合一个 $\Pr[A=1]$ 的饱和 logistic 模型得到, 不过这个模型只有截距, 没有其他变量。最后, 给样本中的每个人加权——戒烟者权重是 $\widehat{\Pr}[A=1]/\widehat{\Pr}[A=1|L]$, 非戒烟者权重是 $(1-\widehat{\Pr}[A=1])/(1-\widehat{\Pr}[A=1|L])$ ——得到虚拟人群, 在虚拟人群中拟合均值模型 $E[Y|A]=\theta_0+\theta_1A$, 然后估计因果效应 $E[Y^{a=1}]-E[Y^{a=0}]$ 。在我们的假设成立之下, 平均而言戒烟会提高体重 $\widehat{\theta}_1=3.4 \text{ kg}$ (95%置信区间是 2.4 至 4.5)。这一结果和我们之间用非稳定权重得到的结果相同。

(在数据分析中, 我们需要检查稳定权重 SW^A 的均值是不是 1 (参见 Hernan 和 Robins 2006 年所著论文)。如果偏离 1, 则表明模型设定错误, 或者正数性假设不成立。精讲点 12.2 讨论了如何检查正数性。)

(参见代码 12.3。稳定权重 SW^A 的取值范围是 0.33 到 4.30, 均值是 1。)

如果稳定权重和非稳定权重都能得到同样的结果, 我们为什么还要使用稳定权重? 这是因为稳定权重的通常会给出更窄的置信区间。然而, 只有当 (逆概率加权) 模型不是饱和的时候, 稳定权重的优越性才会显现出来。在我们的例子中, $E[Y|A]=\theta_0+\theta_1A$ 是饱和的, 因为治疗变量 A 只有两个可能取值。在大多数情形中 (比如时异性或连续性治疗), 加权模型就不再是饱和的, 从而我们会更常用稳定结局。下一小节我们将讨论连续性治疗。

155 12.4 边缘结构模型

让我们思考下面这个在治疗取值为 a 时的线性模型:

$$E[Y^a]=\beta_0+\beta_1a$$

这个模型和我们之前所讲的模型都不一样, 此处模型的因变量是反事实结局, 因而我们不可能观测到。因此, 这个模型不能用现实数据拟合。这一模型的因变量是反事实结局的边缘均值, 因而这一模型被称为边缘结构均值模型。

(如果治疗变量 A 是一个二分变量, 那么这个模型就是一个饱和模型。)

在结构均值模型中, 治疗变量对应的参数就是因果效应均值。对于上一段所提到的模型, 因为 $a=0$ 时 $E[Y^a]=\beta_0$, $a=1$ 时 $E[Y^a]=\beta_0+\beta_1$, 所以参数 $\beta_1=E[Y^{a=1}]-E[Y^{a=0}]$ 。在上一小节, 我们已经将戒烟 A 对体重增加 Y 的因果效应均值定义为 $E[Y^{a=1}]-E[Y^{a=0}]$ 。换句话说, 我们已经估计了边缘结构模型中的参数 β_1 。

具体而言, 我们用逆概率权重构建了一个虚拟人群, 然后再在虚拟人群中拟合模型

$E[Y|A] = \theta_0 + \theta_1 A$ 。在我们的假设下, 虚拟人群中相关性就是因果性。也就是说逆概率加权得到

156 的相关性模型 $E[Y|A] = \theta_0 + \theta_1 A$ 的 θ_1 等价于结构模型 $E[Y^a] = \beta_0 + \beta_1 a$ 的 β_1 。因而, 虚拟人群中相关性参数的一致估计 $\hat{\theta}_1$ 也是目标人群中因果效应 $\beta_1 = E[Y^{a=1}] - E[Y^{a=0}]$ 的一致估计。

因为表示戒烟的变量 A 是一个二分变量, 所以上述边缘结构模型 $E[Y^a] = \beta_0 + \beta_1 a$ 是一个饱和模型。也就是说, 这个模型的等号两侧都有 2 个未知数: 等号左边是 $E[Y^{a=1}]$ 和 $E[Y^{a=0}]$, 等号右边是 β_0 和 β_1 。因而, 虚拟人群中计算得到的均值足以用来估计我们感兴趣因果效应。

然而治疗变量经常是多分类或连续的。比如, 如果我们的治疗变量 A 是“吸烟频率的变化”, 通过随访时的吸烟频率减去初次访问时的吸烟频率计算得到, 那它就是一个连续变量, 可以有许多取值, 比如可以是 -25, 表示某人吸烟频率降低了 25 支/天, 也可以是 25, 表示某人吸烟频率增加了 25 支/天。在初次访问时, 有 1162 人每天吸烟数少于等于 25 支。假设我们的目标是估计这 1162 人中不同吸烟频率变化所对应的体重变化, 那我们想估计的是 $E[Y^a] - E[Y^{a'}]$, 此时 a 和 a' 可以是任意值。

(边缘结构模型拥有零值保留 (参见第九章) 性质, 即因果效应零假设成立时, 边缘结构模型的设定就不会有错。比如, 在边缘结构模型 $E[Y^a] = \beta_0 + \beta_1 a + \beta_2 a^2$ 中, 假设 $\beta_1 = \beta_2 = 0$ 的 Wald 检验有两个自由度, 并且是一个有效的零假设检验。)

因为治疗变量 A 可能有许多不同取值, 此时饱和模型就变得不切实际。我们需要使用非饱和模型, 并假设治疗 A 和结局 Y 的剂量反应曲线形式。如果我们觉得抛物线能恰当描述这一关系, 那我们的边缘结构模型就可以是:

$$E[Y^a] = \beta_0 + \beta_1 a + \beta_2 a^2$$

其中 $a^2 = a \times a$, 且 $E[Y^{a=0}] = \beta_0$ 是 $a = 0$ 时的体重变化均值, 也即如果吸烟频率没有变化时的体重变化。

(这里的 (非饱和) 边缘结构模型的治疗变量 A 是连续变量。)

假设我们想估计吸烟频率提高 20 支/天相比于频率没有变化所对应因果效应均值, 即

$E[Y^{a=20}] - E[Y^{a=0}]$ 。根据我们的饱和模型, $E[Y^{a=20}] = \beta_0 + 20\beta_1 + 400\beta_2$, 因而

$E[Y^{a=20}] - E[Y^{a=0}] = 20\beta_1 + 400\beta_2$ 。现在我们需要估计 β_1 和 β_2 , 因而我们需要用逆概率权重

SW^A 去构建一个虚拟人群，其中 L 不再是混杂变量，然后再在这个虚拟人群中拟合相关性模型
 $E[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$ 。

为了计算稳定权重 $SW^A = f(A)/f(A|L)$ ，我们需要先估计 $f(A|L)$ 。对于二分治疗变量 A ， $f(A|L)$ 表示概率值，因而我们可以用 logistic 模型来估计它。对于连续治疗变量 A ， $f(A|L)$ 是概率密度函数。遗憾的是，一般情况下很能正确估计一个概率密度函数，因而对连续性治疗变量使用逆概率加权方法存在风险。在我们的例子中，我们假设概率密度 $f(A|L)$ 是一个正态分布，均值为 $\mu_L = E[A|L]$ ，方差是常数 σ^2 。接下来我们用线性回归来估计不同 L 取值下的均值 $E[A|L]$ 和残差方差 σ^2 。我们假设分子中的密度 $f(A)$ 也是正态分布的。因为效应估计对条件密度 $f(A|L)$ 的模型选择非常敏感，所以我们对连续变量使用逆概率加权时需要非常小心。

(参见代码 12.4。 SW^A 的估计值在 0.19 到 5.10 之间，均值是 1.00。我们假设方差是常数 (方差齐性假设)，这一假设和方差分布图相符。其他分布假设 (比如方差异性下的截断正态分布) 得到的结果大致相同。)

157 逆概率加权下边缘结构模型的参数估计值是 $\hat{\beta}_0 = 2.005$, $\hat{\beta}_1 = -0.109$, $\hat{\beta}_2 = 0.003$ 。由这些估计值，我们得到的增重均值 (95%置信区间)：如果所有人的吸烟频率不变是 2.0 kg (1.4, 2.6)，如果所有人吸烟频率提高 20 支/天是 0.9 kg (-1.7, 3.5)。

边缘结构模型的治疗变量也可以是一个二分变量。比如，我们想估计戒烟 A (1: 是, 0: 否) 对死亡 (1: 是, 0: 否) 的因果效应，我们可以用以下边缘结构 logistic 模型：

$$\text{logit } \Pr[D^a = 1] = \alpha_0 + \alpha_1 a$$

其中 $\exp(\alpha_1)$ 是戒烟者比非戒烟者的死亡因果性比值比。在我们之前所述假设成立的情况下，我们可以在逆概率权重构建的虚拟人群中拟合 logistic 模型 $\text{logit } \Pr[D = 1 | A] = \theta_0 + \theta_1 A$ 从而一致估计上述边缘结构 logistic 模型中的参数。最后我们得到的因果性比值比是 $\exp(\hat{\theta}_1) = 1.0$ (95% 置信区间是: 0.8, 1.4)。

(这是一个饱和边缘结构 logistic 模型。对于连续性治疗，我们可以设定一个非饱和的 logistic 模型。)

(参见代码 12.5。)

12.5 效应修饰与边缘结构模型

当我们想估计的参数是目标人群中的因果效应均值时, 边缘结构模型不会包含其他协变量。然而, 当效应修饰存在时, 我们可以在边缘结构模型中放入其他协变量, 即使这些协变量不是混杂变量。假设戒烟的效应对不同性别 V (1: 女性, 0: 男性) 不一样。为了验证这一假设, 我们需要在边缘结构均值模型中加入协变量 V :

$$E[Y^a | V] = \beta_0 + \beta_1 a + \beta_2 V a + \beta_3 V$$

如果 $\beta_2 \neq 0$, 那么存在加法效应修饰。严格而言, 这不再是边缘模型, 因为存在其他协变量 V , 然而我们依然称之为边缘结构模型。

我们可以在逆概率权重构建的虚拟人群中拟合线性模型 $E[Y | A, V] = \theta_0 + \theta_1 A + \theta_2 V A + \theta_3 V$, 从而估计上述边缘模型中的参数值。此时向量 L 需要包含 V ——即使 V 不是混杂变量——以及其他可能影响互换性的变量。

(参数 β_3 可以没有因果性意义, 其不必是 V 的因果性效应。我们只是对治疗 A 假设了互换性、正数性和一致性, 而不是对性别 V 。)

因为我们现在要考虑的是 V 的每一分层中治疗的因果效应, 所以我们稳定权重的分子可以是 $f(A)$ 或 $f(A|V)$ 。基于稳定权重 $SW^A(V) = \frac{f(A|V)}{f(A|L)}$ 的逆概率加权一般而言会给出较窄的置信区间。我们可以这样理解: 因为分子分母中都含有 V , 所以分子和分母的值更加接近, 从而使得逆概率权重更加稳定, 因而 95% 置信区间更窄。此时, 在估计权重分子的时候, 我们会在 logistic 模型中加入协变量 V 。除此之外, 估计 $SW^A(V)$ 的方法和估计 SW^A 的一样。

158 V 是 L 的一个子集, 选择什么样的变量作为 V 需要反映出研究者真切关心的研究问题。比如, 只有当研究者确信 V 是一个效应修饰因子, 并且对 V 的每一分层中的因果效应非常感兴趣时, V 才能被放入边缘结构模型之中。在我们的例子中, $\hat{\theta}_2$ 的置信区间是 -2.2 到 1.9, 因而我们没有发现确切证据支持性别是一个效应修饰因子。如果研究者决定把 L 中的所有变量全部放入边缘结构模型之中, 那逆概率加权就不再是必须的, 因为此时正确设定的未加权回归模型也能完全控制 L 带来的混杂 (参见第十五章)。出于这一原因, 我们将包含了所有变量 L 的边缘结构模型称为仿制的边缘结构模型。

(如果我们要探索两个治疗变量 A 和 B 之间的交互作用, 我们需要把 A 和 B 都放入边缘结构模型之中, 在计算逆概率权重时, 分母中会包含这两个变量的联合分布。我们会对 A 和 B 都假设互换性、正数性和一致性成立。)

在本书第一部分, 我们讨论了效应修饰和混杂是两个完全不同的概念。然而, 有些同学依然难以理解这两者之间的区别, 主要是因为调整混杂和检测效应修饰都用到了同样的统计方法——分层分析(第四章)和回归(第十五章)。因而, 使用边缘结构模型进行教学就有一定优势, 这是因为, 在这一框架之下, 调整混杂的方法(逆概率加权)和检测效应修饰的方法(将治疗和协变量的乘积项放入到边缘结构模型中)截然不同。

12.6 删失和缺失值

当我们估计戒烟 A 对增重 Y 的因果效应时, 我们将分析局限于 1566 名有体重变化信息的人群中。然而, 还有 63 名人员虽然符合入组标准, 但是在随访中没有体重信息而被排除于分析之外。只选择没有结局缺失值的人——也即删失有缺失值的人——可能会引入选择偏移, 我们在第八章讨论过这个问题。

让我们用删失变量 C 表示在随访中的体重测量: 1 表示没有测量(即被删失), 0 表示有测量(即没有被删失)。显然, 我们的分析只能在没有被删失的人群中进行, 即 $C = 0$ 的人群中。也就是说, 我们在 12.2 小节和 12.4 小节中的拟合的(加权)回归模型不是 $E[Y|A] = \theta_0 + \theta_1 A$, 而是 $E[Y|A, C = 0] = \theta_0 + \theta_1 A$ 。

不过, 如果 C 是治疗 A 和结局 Y 之间某条路径上的对撞变量, 或是对撞变量的下游变量, 那么即使在零值下, 只在未删失的人群中进行分析将会引入偏移, 具体情形可参见图 8.3 至图 8.6 的因果图。我们的数据和这些因果图的结构吻合: 治疗 A 和删失 C 相关(5.8%的戒烟者被删失, 而 3.2%的非戒烟者被删失), 且至少一个 Y 的预测因素和 C 相关(初次访问时的体重均值, 被删失人员是 76.6kg, 未被删失人员是 70.8kg)。

因为失访导致的删失可能会引入选择偏移, 所以我们希望不存在删失, 并估计不存在删失时的因果效应。在我们的例子中, 我们的目标其实是估计在没有删失的情况下, 如果所有人都是戒烟者时的增重 $E[Y^{a=1,c=0}]$, 和所有人都不是戒烟者时的增重 $E[Y^{a=0,c=0}]$ 。因而因果效应是 $E[Y^{a=1,c=0}] - E[Y^{a=0,c=0}]$, 也就是我们第八章讨论的 A 和 C 的联合效应。上角标中的 $c = 0$ 明确指出我们希望估计的因果效应是在没有删失的人群中进行估计的, 这也是大多数研究者想估计的因果效应, 即使少有人使用上角标 $c = 0$ 。

此时的因果效应可以用逆概率权重 $W^{A,C} = W^A \times W^C$ 进行估计, 其中未删失的人群

$W^C = 1 / \Pr[C = 0 | L, A]$, 而删失的人群 $W^C = 0$ 。逆概率权重 $W^{A,C}$ 在联合干预 (A, C) 的可识别假设下同时调整了混杂和选择偏移, 也即 $Y^{a,c=0} \perp\!\!\!\perp (A, C) | L$ 。如果 L 中的某些变量受到治疗 A 的影响, 比如图 8.4 的情形, 此时有界互换性 $Y^{a,c=0} \perp\!\!\!\perp (A, C) | L$ 不一定成立。在本书第三部分我们会讲述存在互换性的其他形式, 从而保证在 L 受到治疗影响的情况下, 我们依然能使用逆概率加权去估计 A 和 C 的联合效应。

(删失和治疗的联合逆概率权重 $W^{A,C} = 1 / f(A, C = 0 | L)$, 其中 A 和 C 的联合密度可以分解为 $f(A, C = 0 | L) = f(A | L) \times \Pr[C = 0 | L, A]$ 。)

我们需要注意权重 $W^C = 1 / \Pr[C = 0 | L, A]$ 构建的虚拟人群, 其样本大小等于原人群删失之前的样本大小, 同时在虚拟人群中没有箭头从 L 或 A 指向 C 。在我们的例子中, 权重 W^C 构建的虚拟人群样本大小将是 (近似) $1566 + 63 = 1629$, 此时, 在我们的假设下并不存在选择偏移, 因为不存在选择。也就是说, 我们将用权重 W^C 拟合一个加权模型 $E[Y | A, C = 0] = \theta_0 + \theta_1 A$, 从而估计整个人群中边缘结构模型 $E[Y^{a,c=0}] = \beta_0 + \beta_1 a$ 的参数值。

(L 中的某些变量对应的系数可能在 $f(A | L)$ 的模型中是 0, 但在 $\Pr[C = 0 | L, A]$ 的模型中不是 0, 或者反过来。然而, 在大样本中, 我们还是会在两个模型中都包含 L 的所有变量。)

相应的, 我们也可以使用逆概率稳定权重 $SW^{A,C} = SW^A \times SW^C$ 。此时删失的权重 $SW^C = \Pr[C = 0 | A] / \Pr[C = 0 | L, A]$, 其构建的虚拟人群样本大小等于原人群删失之后的样本大小, 同时在虚拟人群中没有箭头从 L 指向 C 。在我们的例子中, 取值权重 SW^C 构建的虚拟人群样本大小将是 (近似) 1566, 也即未被删失的人群大小。也就是说, 稳定权重并没有消除虚拟人群中的删失, 而只是将删失变得随机, 不再与 L 相关。因而, 在我们对删失的有界互换性假设下, 虚拟人群中删失的比例等于研究人群中删失的比例: 虽然有选择, 但是不存在选择偏移。

(如果 $\Pr[C = 0 | L, A]$ 的模型设定正确, 那权重 SW^C 的均值是 1。参见知识点 12.2。)

为了得到 $\Pr[C = 0 | L, A]$ 的参数估计, 我们需要在原始 1629 名研究人群中拟合预测删失概率的 logistic 模型。这个模型中的协变量, 和我们之前用来估计治疗逆概率权重模型中的一样。因而在模型假设下, 我们能得到估计值 $\widehat{\Pr}[C = 0 | L, A]$ 以及 1566 名未被删失人员的权重 SW^C , 从而计算 $SW^{A,C} = SW^A \times SW^C$ 。利用这一权重, 我们会得到 $\widehat{\theta}_1 = 3.5 \text{ kg}$ (95%置信区间是 2.4 至

4.5)。这一结果基本和之前 SW^A 得到的结果一样, 可能意味着不存在选择偏移, 或者我们已有的协变量不能消除选择偏移。

(参见代码 12.7。 $SW^{A,C}$ 的取值范围是 0.35 到 4.09, 均值是 1。)

下一章我们将讨论另一种用来调整混杂和选择偏移的方法: 标准化。

第十二章精讲点和知识点

精讲点 12.1: 一个不够好的例子 (原书第 150 页)

我们戒烟的例子非常简单方便, 我们不需要更多的专业知识, 这些数据也是公开的。不过, 这个简便会有一定代价, 比如, 会存在选择偏移。

只要参与人员: (1) 在初次访问时说自己在 1971–1975 年之间有吸烟, (2) 并且在 1982 年的随访中不再吸烟, 那我们就认为他是戒烟者, $A=1$ 。而后一个条件意味着参与者在随访前没有去世, 也没有失访。也就是说, 我们是在一定条件下选择参与者的, 而这一条件是在治疗——也即戒烟或不戒烟——开始之后才出现的。如果治疗影响了我们选择参与者的概率, 就可能导致第八章所讨论的选择偏移。(因为不同参与者在不同的时间点开始戒烟, 所以 A 实际上是一个时异变量。我们在本书第二部分会忽略时异特性, 而将在第三部分讨论这一特性。)

而戒烟的随机试验就不会存在这一问题。此时, 每一个参与者都会在初访时被分入戒烟组或非戒烟组, 因而即使他们没有参加 1982 年的随访, 我们也能知道他们的分组情况。在第 12.6 小节, 我们讨论了该如何处理结局出现删失或缺失的情形——这一情形在观察性研究或随机试验中都可能遇到——但这和精讲点中所讨论的情形不太一样: 在精讲点的讨论中, 缺失数据和治疗直接相关。这种情形下的选择偏移可以通过敏感性分析来处理, 详情参见 Hernan 等人所著论文 (2008 年, 附录 3)。

我们的讨论主要是想揭示一个观察性研究中普遍存在的问题: 初访时的入选标准以及治疗分组的错位 (Hernan, 2016)。虽然我们决定暂时忽略这一问题从而保证我们教学的简便易懂, 但是在现实世界中, 我们不能忽略这一问题及其可能导致的偏移。

精讲点 12.2: 检查正数性 (原书第 155 页)

在我们的例子中, 有 4 名 66 岁以上的白人女性, 她们中没有人戒烟。也就是说, L (某一子集) 下 $A=1$ 的条件概率是 0, 正数性——逆概率加权的假设之一——不成立。有两种可能会导致正数性不成立:

- 结构性不成立: 第三章所描述的情形。 L 某些取值下的参与者不能接受治疗(或不接受治疗)。比如, 当我们想估计某种化学物暴露对死亡的影响时, 是否正常工作是一个重要混杂变量。工人只有正常工作的时候才会接触到化学物, 而不正常工作的工人更可能是因为生病了所以没上班, 也就更可能去世。也就是说, 这一结构性问题让不正常工作的工人“接受治疗”的概率为 0。所以控制这一变量之后, 我们总会发现某一分层中有 0 存在。
- 随机性不成立: 本精讲点第一段所描述的情形。我们的样本是有限的, 如果我们依据所有混杂变量进行分层, 我们将发现在某些分层中样本量为零, 而在目标人群中这一分层人群不应该是零。因而, 此时正数性假设不成立是随机性的, 而非结构性的, 此时零值出现在目标人群的样本中的某些分层里。

以上两种不同情形会导致不同的后果。如果是结构性不成立, 我们就不能用逆概率加权或标准化等方法来估计整个人群的因果效应。此时, 因果推断只能局限在正数性成立的分层当中。参见知识点 12.1。如果是随机性不成立, 我们可以用参数模型估计这些分层中治疗的概率。换句话说, 我们能用参数模型平滑掉这些零样本。比如, 12.2 小节中的 logistic 模型能够借助其他人的信息从而估计 66 岁以上白人女性戒烟的概率。分层中零样本存在的时候, 如果用逆概率加权的参数模型去估计因果效应, 都会假设某些分层中正数性不成立是随机的。

知识点 12.1: Horvitz-Thompson 估计 (原书第 152 页)

在知识点 3.1 中, 我们定义了治疗变量为 a 时的“表面”逆概率加权均值 $E\left[\frac{I(A=a)Y}{f(A|L)}\right]$,

其等于正数性和互换性成立时的反事实均值 $E[Y^a]$ 。同时, 这个值的一致估计可以由 Horvitz-

Thompson 估计 $\hat{E}\left[\frac{I(A=a)Y}{f(A|L)}\right]$ 得到。然而, 在本章我们是通过逆概率加权最小二乘法估计

$E[Y^a]$ 的, 这是一种 Horvitz-Thompson 估计的改良版, 通常被称为 Hajek 估计, 即

$$\frac{\hat{E}\left[\frac{I(A=a)Y}{f(A|L)}\right]}{\hat{E}\left[\frac{I(A=a)}{f(A|L)}\right]}.$$

Hajek 估计是 $\frac{E\left[\frac{I(A=a)Y}{f(A|L)}\right]}{E\left[\frac{I(A=a)}{f(A|L)}\right]}$ 的无偏估计, 在正数性成立的情况下等于 $E\left[\frac{I(A=a)Y}{f(A|L)}\right]$, 这

是因为 $E\left[\frac{I(A=a)}{f(A|L)}\right] = 1$ 。在实践中, 我们更偏爱 Hajek 估计, 这是因为如果 Y 是二分变量,

Hajek 估计能保证估计值在 0 和 1 之间, 而 Horvitz-Thompson 估计没有这一特性。

另一方面, 如果正数性不成立, 那么 $\frac{E\left[\frac{I(A=a)Y}{f(A|L)}\right]}{E\left[\frac{I(A=a)}{f(A|L)}\right]}$ 就等于

$\sum_l E[Y | A=a, L=l, L \in Q(a)] \Pr[L=l | L \in Q(a)]$, 如果互换性成立, 则等于

$E[Y^a | L \in Q(a)]$, 其中 $Q(a) = \{l; \Pr[A=a | L=l] > 0\}$ 是 $A=a$ 时概率不为零的 l 的集合。因此, 就如知识点 3.1 中所讨论的一样, 在正数性不成立的时候, Hajek 估计在 $a=1$ 与 $a=0$ 时的对比值没有任何因果性意义。

知识点 12.2: 稳定权重 (原书第 160 页)

稳定权重 $SW^A = \frac{f(A)}{f(A|L)}$ 是更大一类稳定权重 $\frac{g(A)}{f(A|L)}$ 的一部分, 其中 $g(A)$ 是 A 的函数, 而非 L 的函数。当我们使用不饱和结构模型的时候, 我们更喜欢使用 $\frac{g(A)}{f(A|L)}$, 而非 $\frac{1}{f(A|L)}$, 这是因为存在一个函数 $g(A)$ (通常是 $f(A)$), 能在不包含边缘模型中更高效地估计因果效应。我们接下来论述使用 $\frac{g(A)}{f(A|L)}$ 的逆概率加权均值等于反事实均值 $E[Y^a]$ 。

首先注意到逆概率均值 $E\left[\frac{I(A=a)Y}{f(A|L)}\right]$ 使用的权重是 $\frac{1}{f(A|L)}$, 且等于 $E[Y^a]$, 同时能被

表述为 $\frac{E\left[\frac{I(A=a)Y}{f(A|L)}\right]}{E\left[\frac{I(A=a)}{f(A|L)}\right]}$, 这是因为 $E\left[\frac{I(A=a)}{f(A|L)}\right]=1$ 。同理, 使用 $\frac{g(A)}{f(A|L)}$ 的均值也可以表述为

$\frac{E\left[\frac{I(A=a)Y}{f(A|L)}g(A)\right]}{E\left[\frac{I(A=a)}{f(A|L)}g(A)\right]}$, 也等于 $E[Y^a]$ 。此处证明和知识点 2.2 中的同理, 主要关键点在于分子

$E\left[\frac{I(A=a)Y}{f(A|L)}g(A)\right]=E[Y^a]g(a)$, 而分母 $E\left[\frac{I(A=a)}{f(A|L)}g(A)\right]=g(a)$ 。

第十二章图表

Table 12.1

Mean baseline characteristics	A	
	1	0
Age, years	46.2	42.8
Men, %	54.6	46.6
White, %	91.1	85.4
University, %	15.4	9.9
Weight, kg	72.4	70.3
Cigarettes/day	18.6	21.2
Years smoking	26.0	24.1
Little exercise, %	40.7	37.9
Inactive life, %	11.2	8.9

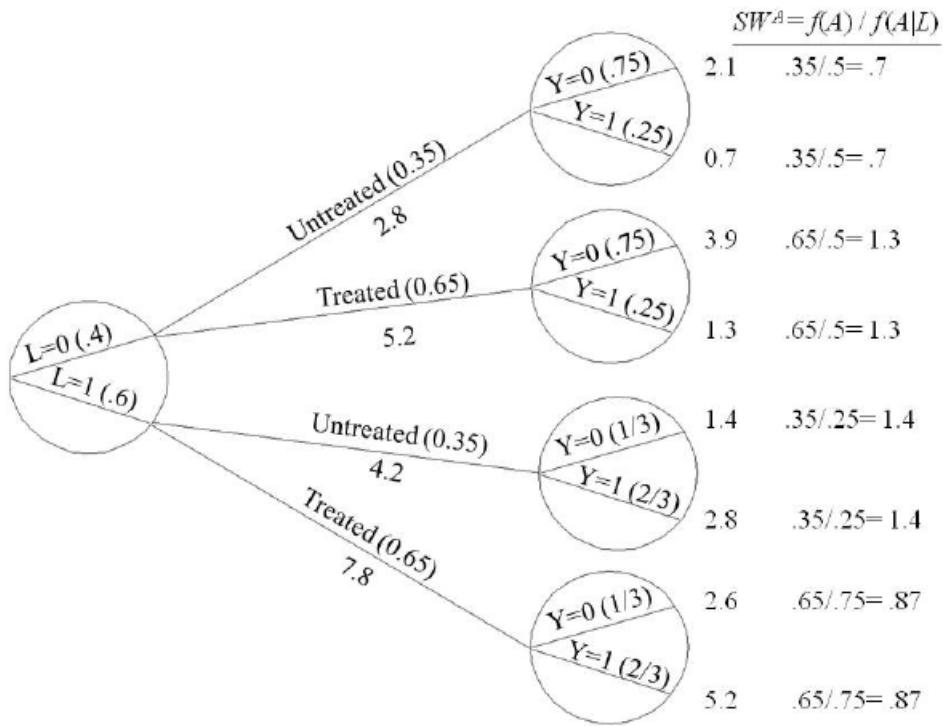


Figure 12.1

第十三章 标准化和参数 G-公式

161 在本章，我们会讨论如何使用标准化估计戒烟对增重的因果效应。我们将会使用和上一章一样的数据。虽然在第二章我们就已经介绍了标准化，但那时我们只讲述了标准化的非参数形式。本章我们将讨论模型中的标准化，从而解决高维数据与非二分治疗变量带来的问题。同上一章一样，我们会给出数据分析所用的代码。

在实践中，研究者可以在逆概率加权和标准化两种方法中选择一种用来估计因果效应。这两种方法都基于同样的可识别性条件，然而它们的模型假设却不一样。

13.1 标准化

在上一章，我们用逆概率加权估计了戒烟 A (1: 是, 0: 否) 对增重 Y 的 (单位: kg) 因果效应。本章我们会估计同样的因果效应，不过将用另一种方法，即标准化。我们的分析人群同样是 NHFES 中的 1629 名参与者，他们在 25–74 岁之间，参加了初访，并在 10 年之后参加了随访。在这些人当中，1566 名人员在初访和随访中都有体重测量，因而没有被删失 ($C = 0$)。

我们将 $E[Y^{a,c=0}]$ 定义为如果所有人接受了治疗 a 且没有人被删失时的增重均值。则戒烟的因果效应可以被表述为 $E[Y^{a=1,c=0}] - E[Y^{a=0,c=0}]$ ，也即所有人都戒烟且未被删失，同所有人都没有戒烟且未被删失的对比。

表 12.1 的数据显示戒烟者 ($A = 1$) 和非戒烟者 ($A = 0$) 在某些变量的分布上有所不同，而这些变量是增重的预测因素。观察到的相关性差值 $E[Y|A=1,C=0] - E[Y|A=0,C=0] = 2.5$ 会和因果性差值 $E[Y^{a=1,c=0}] - E[Y^{a=0,c=0}]$ 有所不同。向量 L 含括的变量包括：性别 (0: 男性，1: 女性)，年龄 (单位: 年)，种族 (0: 白人，1: 其他)，教育程度 (5 个分类)，烟龄 (单位: 年)，吸烟频率 (单位: 支/天)，每日体力情况 (3 个分类)，运动量 (3 个分类)，以及初次访问时的体重 (单位: kg)。同样，我们假设向量 L 足以调整所有的混杂和选择偏移。

我们可以通过逆概率加权来调整 L ，此时会构建一个虚拟人群，其中戒烟者和非戒烟者的 L 分布将会相同。因而，在给定 L 的互换性和正数性假设下，我们就能得到 $E[Y^{a,c=0}]$ 的估计值 $\hat{E}[Y|A=a,C=0]$ 。如果 A 是连续性治疗 (这和我们的例子不同)，我们也需要使用结构模型来估计虚拟人群中 A 可能取值下的 $E[Y|A,C=0]$ 。逆概率加权需要估计治疗和删失的联合分布。对于二分治疗戒烟来说，我们会估计 $\Pr[A=a,C=0|L]$ ，然后用它计算逆概率权重。

(同上一章一样, 我们会假设用来调整 C 的变量也是 L 中的一部分, 且不受 A 的影响。否则, 我们需要使用更广适的方法, 具体将在本书第三部分介绍。)

我们在第二章讨论过可以用标准化来替代逆概率加权。在互换性和正数性假设下, 未被删失的治疗组人群中的标准化结局均值就是 $E[Y|A=1,C=0]$ 的一致估计。精讲点 13.1 讨论了正数性不成立对逆概率加权和标准化的不同影响。

(知识点 2.3 证明了在有界互换性、正数性和一致性假设下, 治疗组中的标准化均值就是所有人被治疗时的均值。我们可以在这一结论中纳入删失, 只需在定义和证明中用 $(A=a, C=0)$ 替换 $A=a$ 就行。)

为了计算治疗组中的标准化均值, 我们需要计算 L 的每一分层 l 中的条件均值, 即 $E[Y|A=1,C=0,L=l]$ 。在我们戒烟的例子中, L 可能有上百万个不同分层。手动计算每个分层中的均值并不现实。

标准化均值是每一分层中条件均值的加权平均, 而权重是每一分层的出现概率 $\Pr[L=l]$ 。也就是说, 在计算标准化均值的时候, 人数最多的分层占有的权重最多。

我们可以把治疗为 a 且未被删失的标准化均值表述为:

$$\sum_l E[Y|A=a,C=0,L=l] \times \Pr[L=l]$$

如果 L 是连续的, 我们需要把 $\Pr[L=l]$ 替换为概率密度函数 $f_L(l)$, 同时把求和符号替换为积分符号。

(治疗组中的因果效应均值可以用知识点 4.1 中方法进行估计, 只需把 $\Pr[L=l]$ 替换为 $\Pr[L=l|A=1]$ 即可。)

13.2 通过模型估计结局均值

理想情况下, 我们可以用非参数的方法估计条件均值 $E[Y|A=1,C=0,L=l]$ 。在 2.3 小节中我们就是这么做的, 当时用的是表 2.2 的数据。

但在高维数据中——比如我们戒烟的例子——我们不可能用非参数的方法估计 $E[Y|A=1,C=0,L=l]$ 。在我们的例子中, L 有上百万个分层, 而我们只有 403 名戒烟者。因此我们需要借助模型。 $A=0$ 时同理。

为了得到上百万个 L 分层中 $E[Y | A = a, C = 0, L = l]$ 的参数估计值, 我们需要拟合一个线性模型, 其中的因变量是增重, 模型中会包含所有混杂变量。对连续性变量, 诸如年龄、初访时体重、烟龄和吸烟频率, 模型中会放入它们的一次项和二次项。也就是说, 我们假设条件均值和这些连续性协变量的关系可以用抛物线表示。我们在模型中也包括了戒烟和吸烟频率的乘积项。也就是说, 我们认为戒烟的效果会因吸烟频率的不同而不同, 并且这一关系是线性的, 同时戒烟的效果独立于其他变量。

(参见代码 13.1。)

有了这些假设后, 我们得到 A 和 L 所有可能组合之下的 $\hat{E}[Y | A = a, C = 0, L = l]$, 同时也能得到数据中 403 名戒烟者和 1163 名非戒烟者的预测值。比如, 如果某参与人员的信息是: 非戒烟者, 男性, 白人, 26 岁, 大学肄业, 12 年烟龄, 吸烟频率 15 支/天, 适当运动, 体力良好, 初访时体重 112kg, 那么我们就能得到他的增重预测值是 0.34kg (碰巧编号为 24770 的参与者满足以上条件, 你可以在看看他的预测值是多少)。总而言之, 在这 1566 名参与人员中, 增重预测均值是 2.6kg, 这与我们观测到的增重值相同, 而取值范围是 -41.3 到 48.5kg。

(Y 的标准化均值一般被表述为 $\int E[Y | A = a, C = 0, L = l] dF_L(l)$, 其中 $F_L(\cdot)$ 是 L 的累积分布函数。在本章, L 中的变量都不受治疗 (也即我们例子中的戒烟) 影响, 因而我们在不同的 L 观测值下取平均, 从而非参数地估计这个积分。)

我们的目标是估计 $A = a$ 时的 $\sum_l E[Y | A = a, C = 0, L = l] \times \Pr[L = l]$ 。这里的求和符合在更正式的情况下应该被写作积分符号, 因为 L 中的变量基本上是连续的, 不能用简单的概率表示。不过无论用什么符号表示, 我们现在已经估计了 A 和 L 所有组合下的 $E[Y | A = a, C = 0, L = l]$ 。

下一步我们需要把这些均值根据 L 中的取值 l 进行标准化。

164 13.3 根据混杂变量的分布对结局均值进行标准化

标准化均值是各条件均值 $E[Y | A = a, C = 0, L = l]$ 的加权平均。当 L 中的所有变量都是离散的时候, 条件均值对应的权重就是 $L = l$ 的人数比例, 即 $\Pr[L = l]$ 。原则上, 我们可以非参数地计算这一概率, 只需用 $L = l$ 的人数除以总人数即可, 我们在 2.3 小节中就是这么做的。然而, 在高维数据中, 我们需要用到其他方法。

幸运的是, 我们其实并不需要去估计 $\Pr[L = l]$ 。我们只需要估计每个个体 i 在 $L = l$ 时的条件均值 $E[Y | A = a, C = 0, L = l]$ 即可, 然后再计算平均值 $\frac{1}{n} \sum_{i=1}^n \hat{E}[Y | A = a, C = 0, L_i]$, 其中 n 是研究的总参与人数。这是因为 $\sum_l E[Y | A = a, C = 0, L = l] \times \Pr[L = l]$ 也可以被写作双重期望 $E[E[Y | A = a, C = 0, L = l]]$ 。

接下来我们将讲述如何在治疗组 ($A = 1$) 和非治疗组 ($A = 0$) 中估计 $\sum_l E[Y | A = a, C = 0, L = l] \times \Pr[L = l]$, 且不需要计算 $\Pr[L = l]$ 。我们将先用表 2.2 的数据作示例, 因为其中不涉及删失, 只有一个混杂变量 L 且其只有两个分层, 同时 Y 也是一个二分结局, 因而此时条件均值 $E[Y | A = a, C = 0, L = l]$ 等于概率 $\Pr[Y = 1 | A = a, L = l]$ 。之后我们再将这个方法应用到有多个混杂变量的现实数据当中。这个方法有四步: 扩充数据, 结局回归, 预测, 以及标准化。

表 2.2 只有 20 行, 每一行代表一个参与人员。我们需要把表 2.2 复制三遍, 从而创建一个新的数据集。也就是说, 新的数据集会有 60 行, 原来数据中每个参与人员在新数据集中都会出现三次, 我们把它分为三个不同板块。我们不变动第一个板块, 也即第一个板块会和表 2.2 中的原数据一样。而我们将会把第二和第三个板块变动得如同本章表中所示一样¹。在第二个板块中, 我们把这 20 个人的 A 都赋值为 0 (非治疗组); 在第三个板块中, 我们把这 20 个人的 A 都赋值为 1 (治疗组)。在这两个板块中, 我们把所有人的结局数据删去, 即赋予 Y 缺失值。下文会提到, 我们会在第二个板块中估计非治疗组的标准化均值, 在第三个板块中估计治疗组的标准化均值。

下一步我们将用这个新数据集去拟合结局均值和治疗 A 以及混杂变量 L 的回归模型。我们会在模型中加入乘积项 $A \times L$, 从而让模型饱和。注意到, 其实只有第一个板块中的数据 (也即实际数据) 会被用来拟合模型, 这是因为第二和第三个板块中的结局数据都是缺失值。

下一步我们需要用第一个板块拟合的模型去预测第二个和第三个板块的结局。(也就是, 我们将第二和第三个板块中的 L 与 A 和回归系数结合起来, 从而得到结局 Y 的预测值。) 第二个板块中的预测值是 L 与 $A = 0$ 组合下的结局均值估计, 而第三个板块中的就是 L 与 $A = 1$ 组合下的结局均值估计。

¹ 见本章图表。

最后, 我们将会计算第二个板块中的预测值均值。因为 60% 的人是 $L=1$, 40% 的是 $L=0$, 所以 $L=1$ 的均值就会有更多权重。也就是说, 第二个板块中的预测值均值就是非治疗组的标准化结局均值。同理, 第三个板块中的预测值均值是治疗组的标准化结局均值。这样一来, 我们的计算就结束了。

上一段方法得到的结果和 2.3 小节中直接计算得到的结果一样。这两个方法都是非参数的。在本章, 我们并不会直接估计 L 的分布, 而是在不同的 L 观测值下 (也即 L 的分布) 取平均。

(参见代码 13.2。)

在实践中, 用经验分布经计算标准化均值是一个可行的办法, 尤其在 L 是高维向量的情况下。我们戒烟例子中所用的方法步骤和前几段所描述的大致相同。我们需要在原数据中添加第二和第三个板块, 拟合 $E[Y | A = a, C = 0, L = l]$ 的模型, 然后生成预测值。第二个板块中预测值的均值是 1.66, 而第三个板块中是 5.18。因此, 我们的因果效应 $E[Y^{a=1,c=0}] - E[Y^{a=0,c=0}]$ 就是 $5.18 - 1.66 = 3.5 \text{ kg}$ 。我们需要使用自举法计算 95% 置信区间 (参见知识点 13.1)。总的来说, 戒烟会使体重增加 3.5kg (95% 置信区间: 2.6, 4.5)。

(参见代码 13.3 和 13.4。)

13.4 逆概率加权还是标准化?

迄今我们讲述了两种用模型估计因果效应的方法: 逆概率加权 (上一章) 和标准化 (本章)。在我们戒烟的例子中, 两种方法给出的效应估计基本一样。而我们在知识点 2.3 就证明了这两种方法是等价的。

既然一种方法就足够了, 为什么我们还要介绍两种呢? 这是因为, 逆概率加权均值和标准化均值, 只有在不使用模型的时候, 才是相等的, 否则不一定相等。为了理解这一点, 我们可以思考一个需要用逆概率加权或标准化进行估计的数值。逆概率加权需要对 $\Pr[A = a, C = 0 | L]$ 建模, 为此我们需要分别拟合 $\Pr[A = a | L]$ 和 $\Pr[C = 0 | A = a, L]$ 的模型。而标准化则是对 $E[Y | A = a, C = 0, L = l]$ 建模, 为此在本章我们使用的是线性回归模型。

在实践中, 或多或少的模型错误设定总是不可避免, 而错误设定会带来偏移。治疗模型 (逆概率加权) 和结局模型 (标准化) 的错误设定给效应估计带来的偏移影响不尽相同。因此, 逆概率加权得到的效应估计会和标准化得到的效应估计有所区别。如果逆概率加权和标准化两种方法得到结果差距很大, 会让我们警觉至少其中一种方法存在严重的模型错误设定。如果两种方法得到的结果差距不大, 虽然不能完全排除存在模型错误设定, 但可以侧面说明模型错误设定带来的

影响不是很严重——这是因为两种方法的模型都错误设定且导致的偏移影响差不多, 这一概率实在太低了。

在我们的例子中, 逆概率加权和标准化得到的结果差不多。保留一位小数时, 戒烟导致的增重都是 3.5kg。在这两种方法中, 我们都没有拟合混杂变量 L 的模型 (即 L 是因变量)。在逆概率加权中, 我们不需要用到 L 的分布, 而在标准化方法中, 我们只需要用到 L 的经验分布。

这两种方法都是 G-公式的某种估计方法。G-公式在 1986 年由 James Robins 建立, 并用于因果推断。(本书第三部分将在有时异性治疗的情境中对 G-公式进行定义。) 我们将标准化称为“插入式 G-公式”, 因为这一方法用估计值替换了 G-公式中的结局均值。如果这些估计值是从参数模型中得到的——就如本章一样——那我们就将这一方法称为“参数 G-公式”。因为我们只对因果效应均值感兴趣, 所以我们只会去估计结局均值。一般而言, 参数 G-公式可以在 A 和 L 的不同分层中, 使用结局分布的任意函数 (比如概率密度函数) 计算对应的标准化均值。如果不存在时异性混杂变量 (参见第三部分), 那么参数 G-公式并不需要混杂变量的分布情况。

(Robins 在 1986 年将标准化推广到有时异性治疗和混杂变量的场景中, 并把这个方法命名为“G-算法公式”。因为名字太长, 所以会简称为“G-公式”或“G-计算”。虽然“G-公式”和“G-计算”是同义词, 但是本书采用第一种称呼, 因为后者可能和“G-估算”混淆。我们会在第十四章介绍 G-估算。)

大多数时候, 我们没必要一定要在逆概率加权和参数 G-公式中评出个高低。两种方法都可以用来估计因果效应, 我们都可以使用。此外, 我们应尽可能使用双重稳健模型 (参见精讲点 13.2 和知识点 13.2), 从而将同一方法中治疗和结局的模型结合起来。

最后, 我们现在是用参数 G-公式估计整个人群的因果效应均值。如果我们只对人群的一个子集感兴趣, 那么我们应该将计算局限在这个子集当中。比如, 如果我们对性别的修饰效应感兴趣, 我们需要在女性和男性中分别计算标准化均值。逆概率加权和标准化两种方法都能用来计算人群子集的因果效应。

13.5 我们应该如何看待估计值?

本书第一部分都在讨论因果效应的定义, 相关假设, 以及可能的偏移。这部分讨论都是概念性的, 例子也非常简单。我们想传达的信息是: 如果观察性研究能明确类比成一个 (假想的) 随机试验——也即靶标试验——那么我们的因果推断将会更加可信。

本章和上一章的分析只是我们在现实数据中的初步尝试。在逆概率加权和标准化方法中, 我们得到同样的结果: 如果每个参与人员都戒烟, 将会平均增重 5.2kg, 而如果每个参与人员都不

作者: Miguel A. Hernan, James M. Robins;

翻译: 罗家俊

戒烟, 将会平均增重 1.7kg。两种方法得到的戒烟因果效应均值都是 3.5kg (95%置信区间: 2.5 到 4.5)。在接下来几章, 我们将用 G-估算、传统结局回归、以及倾向性评分得到同样的结果。

不同方法得到同样的结果, 这让我们的结果更加可信, 这是因为每种方法的模型假设都不一样。然而, 观测性研究得到的效应估计总是有严重缺陷。即使我们并不想将我们的结果应用到其他人群身上 (参见第四章), 且个体之间不存在干扰。我们对目标人群的效应估计依然需要其他假设条件。我们将这些条件分成三组。

第一, 可识别性条件, 即互换性、正数性和一致性 (参见第三章), 需要对观察性研究成立, 从而使得观察性研究能被类比为靶标试验。在控制了 L 中的 9 种混杂变量后, 戒烟者和非戒烟需要变得可互换 (参见精讲点 14.2)。未测混杂 (第七章) 和选择偏移 (第八章, 以及精讲点 12.2) 可能会使有界互换性不成立。正数性要求戒烟者中 L 的分布要和非戒烟者中的完全重叠, 即 L 的任一取值在戒烟者中和非戒烟者中都能找到。精讲点 13.1 讨论了正数性不成立对逆概率加权和标准化的不同影响。至于一致性, 我们注意到在这个研究中戒烟和不戒烟其实有许多种不同形式。比如戒烟可以是突然戒烟, 或者循序渐进地戒烟, 而非戒烟可以是吸烟频率变高, 或降低。因此, 我们的效应估计对应的研究问题模糊不清, 就像在一项干预试验中, 把被试随机分到不同形式的戒烟或非戒烟组中。

(基于结局回归的某些方法, 比如双重稳健, 可以在正数性不成立的情况下使用, 不过此时需要假设结局模型设定正确, 从而我们可以推断数据之外的情形。参见精讲点 13.1。)

(某一干预方案因果效应的数值估计非常重要, 这是因为它们会被用于公共政策决策或临床方案的制定。Hernan (2016) 讨论过这一问题。)

第二, 所有变量都应该被正确测量。治疗 A 、结局 Y 以及混杂变量 L 的测量误差都会导致偏移 (参见第九章)。

第三, 所有数据分析中的模型都应该是正确设定的 (第十一章)。假设在我们治疗是因变量的模型中, 如果连续变量年龄的真实函数形式不是抛物线, 而是另一种复杂曲线, 那么即使 L 中的所有变量都是正确测定的, 但因为模型设定错误, 逆概率加权依然不能调整所有混杂。模型错误设定带来的效果就如混杂变量的测量误差一样。

(因果推断的有效性依赖于以下几个条件: 互换性, 正数性, 一致性。并不涉及测量误差和模型错误设定。)

保证以上条件成立, 或者近似成立, 是研究者的主要任务。如果这些条件不一定成立, 那数据分析就是无意义的。问题是, 没有人能保证这些条件都能完美成立。未测的混杂变量、不重叠的混杂变量分布、劣定的干预措施、误测的变量、以及错误设定的模型都是我们在效应估计中常

见的问题。其中某些问题可以用经验方法解决,但是其他问题则需要专业判断,因而会被其他人批评。遗憾的是,我们的数据并不能反驳这些批评。比如,我们可以尝试使用不同的模型,但是我们不可能去调整没有测量的变量。

因果推断依赖于以上条件。这些条件看起来很美好,但是却不能实证地验证。因果推断的重要问题是“我们不可能观察到所有的反事实数据”,因此我们用上述假设去近似得到这些数据。我们越偏离这些假设,那我们的效应估计也就越偏离真实因果效应。因此,对观察性研究中得到的因果推断必须保持谨慎怀疑。实际上,我们需要依据上面提到的假设,对现实中的因果推断进行逐条讨论,从而保证因果推断的严肃性。我们需要像对待效应估计一样,严肃对待这些条件假设。

第十三章精讲点和知识点

精讲点 13.1: 正数性的结构性缺失 (原书第 162 页)

如果正数性是结构性地不成立,那我们就不能用逆概率加权去估计整个人群中的因果效应。同样,正数性也是标准化的必要条件,这是因为如果 $\Pr[A = a | L = l] = 0$ 且 $\Pr[L = l] \neq 0$,那么条件结局均值 $E[Y | A = a, L = l]$ 就不是良定的。

不过,正数性不成立对逆概率加权和标准化两种方法的影响不尽相同。当我们使用标准化的时候,如果我们愿意使用基于参数模型的外推,那就可以忽略正数性。也就是说,我们可以使用 $E[Y | A, L]$ 的模型去平滑插补概率是零的分层中的数值。而平滑插补会给估计带来偏移,因而 95% 名义置信区间会在少于 95% 的时间中覆盖真实值。同时,我们需要注意此时我们使用模型的目的是什么:我们使用模型并不是因为缺少数据,而是因为我们想估计即使有无限数据也不能识别的数值,这一数值无法识别是因为正数性的结构性缺少。这是一个非常重要的区别。

总而言之,在正数性不成立或基本不成立的情况下,标准化得到的效应估计的标准差,会小于逆概率加权得到的结果。而这并不意味着标准化优于逆概率加权,偏移带来的误差可能完全淹没这两种方法标准差的差值。

精讲点 13.2: 双重稳健估计 (原书第 167 页)

上一章我们讲述了逆概率加权,这一方法需要我们正确设定治疗 A 的模型,而该模型控制了混杂变量 L 。本章我们讲述了标准化,这一方法需要我们正确设定结局 Y 的模型,而该模型控制了治疗 A 和混杂变量 L 。有没有方法只需要正确设定 A 的模型或者 Y 的模型就足够了呢?这就是双重稳健估计要做的事。在可识别假设下,只要这两个模型其中有一个是正确设定的(且我们不

需要知道哪一个是正确设定的), 那么双重稳健估计就能给出因果效应的一致估计。也就是说, 双重稳健估计会给我们两次机会。

有很多种不同形式的双重稳健估计。接下来我们会讲述其中一种。在我们的例子中, 我们需要估计一个二分治疗 A 对结局 Y 的因果效应, 这个例子在 Bang 和 Robins (2005) 的论文中有提到。为了简便, 我们假设不存在删失。

为了得到因果效应均值的双重稳健估计, 我们需要先按上一章的方法估计逆概率权重 $W^A = 1/f(A|L)$ 。接下来我们再拟合一个结局回归模型, 它与本章描述的模型相似, 可以是一个搭配标准联系函数的广义线性模型。此时, 因变量是 $E[Y|A=a, L=l, R]$ 。其中, 如果 $A=1$ 则 $R=W^A$, 如果 $A=0$ 则 $R=-W^A$ 。最后, 用这个回归模型去预测 $A=1$ 和 $A=0$ 时的标准化结局均值。这两个标准化均值的差就是双重稳健下的估计值。也就是说, 在给定 L 的互换性和正数性下, 如果治疗或结局的模型中有一个是正确设定的, 那么上述方法就能给出因果效应的一致估计值, 而我们不必知道哪一个模型是正确设定的。

知识点 13.1: 自举法 (原书第 166 页)

在给出效应估计的时候, 我们通常也会给出一个衡量随机变异性的量度, 比如标准误差或 95% 置信区间。我们在第十章已经讨论过了变异性的来源。在实践中, 基于渐进理论, 我们利用大样本近似来估计标准误差。然而, 在某些情况下, 大样本近似过于复杂, 从而难以运行。此时, 我们可以考虑使用自举法来估计标准误差或 95% 置信区间。接下来, 我们会叙述如何用自举法去估计戒烟效应的 95% 置信区间。

以研究中 1629 名参与人员为例, 从其中可重复地随机抽样取出 1629 个观测, 因而某些参与人员会在新样本中出现不止一次, 有些则不会出现。这个含 1629 个观测的新样本被称为“自举样本”, 然后我们在自举样本中估计戒烟的因果效应。接下来我们再用可重复抽样创建第二个自举样本。这两个自举样本所包含的观察大体相似, 但却不尽相同, 因而两个自举样本中的效应估计不尽相同。我们如此重复多次——比如 1000 次)——每个自举样本中都会得到一个效应估计, 再计算这 1000 个效应估计的标准差。这个标准差, 就是研究人群效应估计的标准误差。因而我们就能计算得到 95% 置信区间。自举法的证明及其统计理论, 请参见其他统计教材。

在我们的代码中, 我们用 1000 个自举样本计算了效应估计的 95% 的置信区间。虽然自举法很简单, 但是对于较大的数据而言, 它可能需要很高的算力。因而, 通常情况下只会用 200–500 个

自举样本, 并且结果会和 1000 个自举样本的基本相同。最后要补充的是, 自举法是大样本的通用方法, 我们也可以用自举法去计算上一章边缘结构模型中估计值的置信区间。

知识点 13.2: 双重稳健估计的偏移 (原书第 170 页)

假设我们有一个二分治疗变量 A 、结局变量 Y 、以及保证互换性和正数性成立的混杂变量向量 L (我们假设一致性已经成立)。为了简化, 我们仅估计 $E[Y^{a=1}]$, 而不是因果效应均值。

$E[Y^{a=1}]$ 可以被写作 $E[b(L)]$, 其中 $b(L) = E[Y | A=1, L]$; 或者写作 $E\left[\frac{AY}{\pi(L)}\right]$, 其中

$\pi(L) = \Pr[A=1 | L]$ 。在本章, 我们讲述了被称为插入式估计 $\frac{1}{n} \sum_{i=1}^n \hat{b}(L_i)$: 我们从 (线性) 回归

模型中得到 $b(L)$ 的估计值, 并对所有参与人员取平均; 然后我们用这个估计值替代了结局均

值。在上一章, 我们讨论了 Horvitz-Thompson 逆概率加权估计 $\frac{1}{n} \sum_{i=1}^n \frac{A_i Y_i}{\hat{\pi}(L_i)}$, : 我们从

(logistic) 回归模型中得到 $\pi(L)$ 的估计值, 并对所有参与人员取平均; 然后我们用这个估计值替代了治疗概率。如果 $\hat{b}(L)$ 和 $b(L)$ 的差距较大, 那插入式估计的偏移就会较大。同理, 如果 $\hat{\pi}(L)$ 和 $\pi(L)$ 的差距较大, 那么逆概率加权估计的偏移也就会较大。

$E[Y^{a=1}]$ 的双重稳健估计将结局模型中的 $\hat{b}(L)$ 和治疗模型中的 $\hat{\pi}(L)$ 结合了起来。实践中有许多不同形式的双重稳健估计方法, 精讲点 13.2 仅描述了其中一种。所有双重稳健估计的核心, 都是用一个涉及治疗模型的函数 (一阶影响函数) 去校正结局模型, 也可以视作用涉及结局模型的函数去校正 Horvitz-Thompson 估计。比如, 考虑以下 $E[Y^{a=1}]$ 的双重稳健估计:

$$\hat{E}[Y^{a=1}]_{DR} = \frac{1}{n} \sum_{i=1}^n \left[\hat{b}(L_i) + \frac{A_i}{\hat{\pi}(L_i)} (Y_i - \hat{b}(L_i)) \right]$$

其也可以被写作 $\frac{1}{n} \sum_{i=1}^n \left[\frac{A_i Y_i}{\hat{\pi}(L_i)} + \left(\frac{A_i}{\hat{\pi}(L_i)} - 1 \right) \hat{b}(L_i) \right]$

在互换性和正数性条件下, 如果 $\hat{b}(L)$ 和 $b(L)$ 的差距不大, 或 $\hat{\pi}(L)$ 和 $\pi(L)$ 的差距不大, 那么双重稳健估计的偏移就会很小。具体而言, 偏移 $E\left[\hat{E}\left[Y^{a=1}\right]_{DR} - E\left[Y^{a=1}\right]\right]$ 在大样本中的形式是:

$$E\left[\pi(L)\left(\frac{1}{\pi(L)} - \frac{1}{\pi^*(L)}\right)(b(L) - b^*(L))\right]$$

其中 $\pi^*(L)$ 和 $b^*(L)$ 分别是 $\hat{\pi}(L)$ 和 $\hat{b}(L)$ 的概率极限。如果治疗模型设定正确, 则 $\pi^*(L) = \pi(L)$; 如果结局模型设定正确, 则 $b^*(L) = b(L)$ 。因此, 当结局模型或治疗模型有一个是正确设定的时候(我们不必知道是哪一个), 大样本(即渐近)偏移是 0。当然, 如果 L 是一个非常高维的向量, 就鲜有参数模型是正确设定的, 因此即使是双重稳健估计中的偏移, 也将非常大。

不过, 双重稳健估计还有一个性质: 偏移取决于 $\frac{1}{\pi(L)} - \frac{1}{\hat{\pi}(L)}$ 和 $b(L) - \hat{b}(L)$ 这两个误差的乘积。我们将在第十八章讨论到, 这一性质——也即二阶偏移——能在机器学习中加以应用。我们用机器学习得到的 $\pi(L)$ 和 $b(L)$ 的估计, 会比常规参数模型得到的估计偏移更小。这是因为, 在拥有大量数据的高维情境中, 机器学习是基于更复杂的算法, 因而能产生更准确的估计。而此时的常规参数模型中的参数个数, 相比于样本量实在太少了。

Causal Inferences: What if ——第十三章
作者: Miguel A. Hernan, James M. Robins;
翻译: 罗家俊

第十三章图表

Second block: All untreated			Third block: All treated			
	L	A	Y	L	A	
Rheia	0	0	.	0	1	.
Kronos	0	0	.	0	1	.
Demeter	0	0	.	0	1	.
Hades	0	0	.	0	1	.
Hestia	0	0	.	0	1	.
Poseidon	0	0	.	0	1	.
Hera	0	0	.	0	1	.
Zeus	0	0	.	0	1	.
Artemis	1	0	.	1	1	.
Apollo	1	0	.	1	1	.
Leto	1	0	.	1	1	.
Ares	1	0	.	1	1	.
Athena	1	0	.	1	1	.
Hephaestus	1	0	.	1	1	.
Aphrodite	1	0	.	1	1	.
Cyclope	1	0	.	1	1	.
Persephone	1	0	.	1	1	.
Hermes	1	0	.	1	1	.
Hebe	1	0	.	1	1	.
Dionysus	1	0	.	1	1	.

第十四章 G-估算和结构嵌入模型

171 在前两章, 我们讨论了如何使用逆概率加权和标准化两种方法去估计戒烟对增重的因果效应均值。在本章, 我们将介绍第三种估计因果效应均值的方法: G-估算。我们将使用同样的数据, 并且也会给出数据分析所用的代码。

逆概率加权、标准化和 G-估算经常被统称为 G-方法, 因为它们被用于广义 (Generalized) 的治疗效果比较, 且可以涉及不同时间下的治疗。本书第二部分的例子没有涉及时异性治疗, 因而 G-方法显得不是那么实用。不过, 在简单情境中介绍 G-方法将有助于我们对它的理解。我们将在本书第三部分介绍更复杂的情境。

本书在第一部分介绍了逆概率加权和标准化 (第二章), 并在第二部分论述了如何在模型中使用这两种方法 (第十二章和第十三章)。与之相比, 我们直到此时才正式介绍 G-估算。这是因为介绍 G-估算之前, 我们需要先讲解什么是结构模型。如果一个模型的参数是通过 G-估算进行估计的, 那么这个模型就被称为结构嵌入模型。这三种不同的 G-方法所需的模型假设不同。

14.1 再谈因果性问题

前两章我们用逆概率加权和标准化估计了戒烟 A (治疗变量) 对增重 Y (结局变量) 的因果效应。我们的数据有 1566 名 25 至 74 岁的参与人员, 他们被分为治疗组 $A=1$ (戒烟者) 和非治疗组 $A=0$ (非戒烟者)。我们假设在控制了混杂变量 L 之后治疗组和非治疗组可互换。我们将因果效应均值定义为 $E[Y^{a=1,c=0}] - E[Y^{a=0,c=0}]$, 也即所有人都戒烟且未被删失与所有人都没有戒烟且未被删失的对比。

(本章所用数据和前两章一样。)

$E[Y^{a=1,c=0}] - E[Y^{a=0,c=0}]$ 衡量的是整个人群中的因果效应均值。但有时我们只对人群中特定群体的因果效应感兴趣。比如, 我们只想计算 45 岁以上女性当中戒烟的因果效应。此时, 为了估计这个因果效应, 我们可以在边缘结构模型中添加乘积项 (参见第十二章), 或者只在这个特定人群中使用标准化估计因果效应 (参见第十三章)。

172 假设研究者想在 L 的每一分层中估计戒烟 A 对增重 Y 的因果效应。在我们的例子中, 我们有上百万个分层, 因而只能使用参数模型去估计。此时, 我们可以在边缘结构模型中加入 L 中的所有变量, 以及这些变量和 A 的乘积项, 而不必再使用逆概率权重。这是因为如果模型设定正确, 这个未加权的回归模型也能完全调整 L 中的所有混杂 (参见第十五章)。

在本章, 我们将用 G-估算来估计 L 的每一分层中戒烟 A 对增重 Y 的因果均值。这一条件效应被表示为 $E[Y^{a=1,c=0} | L] - E[Y^{a=0,c=0} | L]$ 。在我们介绍 G-估算之前, 我们会先讨论结构嵌入模型和保序性。在下一小节, 我们将从一个新角度论述互换性。

14.2 再谈互换性

在第二章我们介绍了互换性。有界互换性意味着, 在 L 取值相同的子人群中, 非戒烟者 ($A = 0$) 如果戒烟了, 那么他们的增重均值会和戒烟者 ($A = 1$) 一样。换句话说, 有界互换性意味着现实中的戒烟者和非戒烟者如果戒烟状态一样, 那么这两个组的结局分布相同。治疗取值为 a 的反事实结局 Y^a 在治疗组和非治疗组中一样, 也即 Y^a 和观察到的实际治疗取值 A 无关。我们记为 $Y^a \perp\!\!\!\perp A | L$ 。

(如果你对一段关于有界互换性的叙述感到重复多余, 那我们将会感到非常开心——这证明你已经理解了什么是互换性。)

让我们思考一下没有治疗时的反事实结局 $Y^{a=0}$ 。如果有界互换性成立, 在给定 L 的情况下, 知道 $Y^{a=0}$ 的值是多少并不能帮助我们区分戒烟者和非戒烟者。也就是说, 不管 $Y^{a=0}$ 的值是多少, 给定 L 的情况下, 成为戒烟者的条件概率都是一样的。在数学上表述为:

$$\Pr[A = 1 | Y^{a=0}, L] = \Pr[A = 1 | L]$$

这和二分治疗 A 的有界互换性的定义等价。

把有界互换性表述为概率形式将有助于我们理解接下来要说的 G-估算。具体而言, 假设我们要拟合下述参数模型来预测接受治疗的概率:

$$\text{logit } \Pr[A = 1 | Y^{a=0}, L] = \alpha_0 + \alpha_1 Y^{a=0} + \alpha_2 L$$

其中 α_2 是一个参数向量, 和 L 的维度一样。如果 L 中有 p 个变量, 那么 $\alpha_2 L = \sum_{j=1}^p \alpha_{2j} L_j$ 。除了多

了一项反事实结局 $Y^{a=0}$ 和其对应参数, 这个模型其余部分和我们第十二章用来估计逆概率权重分母的模型一样。

(为了简便, 本书在数学符号中并不区分向量和标量, 但我们相信这不会带来太多混淆。)

当然, 我们在现实世界中不能拟合这个模型, 因为我们不知道所有人的反事实结局 $Y^{a=0}$ 。但假设我们知道了、拟合了上述模型, 并且有界互换性成立、模型也设定正确, 那么参数 α_1 会是什么样呢? 请停下来思考一下这个问题, 因为我们在 G-估算中会估计 α_1 的取值。如果你已经猜到是

作者: Miguel A. Hernan, James M. Robins;

翻译: 罗家俊

什么了, 那你就理解了 G-估算的一半。正确答案是 0, 这是因为 $Y^{a=0}$ 并不能用来预测参与人员是否接受了治疗。接下来, 我们将介绍 G-估算的另外一半: 结构模型。

14.3 结局均值的结构嵌入模型

我们想估计在 L 的每一分层中, 治疗 A 的因果效应均值, 也即 $E[Y^{a=1} | L] - E[Y^{a=0} | L]$ 。

(为了简便, 我们在这一小节假设不存在删失。) 我们也把它写作 $E[Y^{a=1} - Y^{a=0} | L]$, 这是因为均值的差等于差的均值。如果 L 没有效应修饰作用, 那这个均值将在所有 L 的分层中一样, 也即 $E[Y^{a=1} - Y^{a=0} | L] = \beta_1$, 其中 β_1 是每个分层以及整个人群中的因果效应均值。对因果效应构建结构模型, 就有 $E[Y^{a=1} - Y^{a=0} | L] = \beta_1 a$ 。

一般而言, L 可能存在效应修饰作用。比如, 戒烟的效应可能在烟瘾更大的参与者中更明显。为了容纳 L 的效应修饰作用, 我们可以在结构模型中增加一个乘积项, 模型也就变成 $E[Y^{a=1} - Y^{a=0} | L] = \beta_1 a + \beta_2 aL$, 其中 β_2 也是一个参数向量。在有界互换性 $Y^a \perp\!\!\!\perp A | L$ 成立的情况下, L 的每一分层中治疗组和非治疗组的效应均值应该相等。因而, 在有界互换性下, 结构模型也能被表述成:

$$E[Y^a - Y^{a=0} | A = a, L] = \beta_1 a + \beta_2 aL$$

这被称为结构嵌入均值模型, 其中的参数 β_1 和 β_2 将由 G-估算进行计算 (β_2 是一个向量)。 β_1 和 β_2 将表示 A 和 L 的不同分层中, 戒烟的因果效应。

(Robins (1991) 首次介绍了结构嵌入模型。当治疗是时异的, 这些模型就是“嵌入式”的。详见本书第三部分。)

在第十三章, 我们介绍了结局 Y 的参数模型, 这些模型控制了治疗 A 和 L , 就如同结构嵌入模型一样。这些模型是标准化的基础, 我们会用它们来估计参数化的 G-公式。与这些参数模型相比, 结构嵌入模型是半参数化的, 因为这个模型中没有截距项和 L 的效应项, 也即没有 β_0 和 $\beta_3 L$ 两项。因而结构嵌入模型所需的假设更少, 相比于之前的参数 G-公式, 也更不会受到模型设定错误的影响。精讲点 14.1 讨论了结构嵌入模型和第十二章边缘结构模型的关系。

在删失存在的情况下, 我们关心的因果效应不再是 $E[Y^{a=1} - Y^{a=0} | A, L]$, 而是 $E[Y^{a=1,c=0} - Y^{a=0,c=0} | A, L]$, 也即所有人都没有被删失时的因果效应均值。估计这一因果效应需

174 要我们同时调整治疗 A 的混杂和选择偏移。在前两章已经讨论过, 我们可以用逆概率加权和标准化来调整这两种偏移。然而, G-估算只能调整混杂, 不能调整选择偏移。

因而, 当我们使用 G-估算的时候, 我们需要先使用逆概率加权调整删失。在实践中, 这意味着我们需要先使用逆概率加权去构建一个虚拟人群, 其中没有人被删失, 然后再在虚拟人群中使用 G-估算。在我们戒烟的例子中, 我们将使用第十二章的非稳定权重 $W^C = 1 / \Pr[C = 0 | L, A]$ 。

再强调一下, 我们假设 L 中的所有变量足以调整所有混杂和选择偏移。

(严格而言, 当治疗不是时异的且不影响 L 中的任何变量时, 我们就不必利用逆概率加权调整选择偏移。也就是说, 如果我们假设 $Y^a \perp\!\!\!\perp (A, C) | L$, 我们就能直接在未被删失的人群中使用 G-估算, 而不必涉及逆概率权重。)

在本章, 我们讨论的 G-估算都会事先使用逆概率权重调整选择偏移。我们假设控制了 L 之后, 被删失的和未被删失的人群是可互换的。在这一假设之下, 虚拟人群中的结构模型将是:

$$E[Y^{a,c=0} - Y^{a=0,c=0} | A = a, L] = \beta_1 a + \beta_2 aL$$

为了简便, 在本章我们都会忽略上角标 $c = 0$ 。

175 在本章, 我们将使用结构嵌入模型的 G-估算来估计二分治疗“戒烟”的因果效应, 不过结构嵌入模型也可以用于连续性治疗变量, 比如吸烟频率的变化(参见第十二章)。对于连续性变量, 模型需要假设治疗 A 对结局 Y 的剂量反应函数。比如

$$E[Y^{a,c=0} - Y^{a=0,c=0} | A = a, L] = \beta_1 a + \beta_2 a^2 + \beta_3 aL + \beta_4 a^2 L, \text{ 或者其他任意光滑函数, 比如样条。}$$

(与逆概率加权不同, G-估算不能随意延伸用于二分结局的 logistic 模型。详情请参见知识点 14.1。)

接下来我们要介绍一个重要概念: 保序性。这将有助于我们理解结构嵌入模型。

14.4 保序性

在我们戒烟的例子中, 所有人员都能根据他们结局 Y 的观测值进行排序。编号为 23522 的人员增重 48.5kg, 排在第一位, 编号为 6928 的人员增重为 47.5kg, 排在第二位, 编号为 23321 的人员增重 -41.3kg, 排在最后一位。假设我们知道所有人的 $Y^{a=1}$ 和 $Y^{a=0}$, 我们可以就根据 $Y^{a=1}$ 和 $Y^{a=0}$ 进行排序, 并得到两份排序名单。如果两份名单的顺序是一样的, 那我们就把这称为“保序性”。

如果在加法尺度上, 治疗 A 对结局 Y 的效应在所有个体中都一样, 那我们会说加成保序性成立。比如, 如果戒烟让每个人都增重 3kg, 那根据 $Y^{a=0}$ 的排序也就等于根据 $Y^{a=1}$ 的排序。极端零

假设(参见第一章)成立的时候, 就会出现加成保序性。在结构嵌入模型中, 我们只会在意 L 分层中的加成保序性。如果在 L 的所有分层中, A 对 Y 的效应都相等, 那么我们会说条件加成保序性成立。

一个(条件加成)保序的结构模型如下所示:

$$Y_i^a - Y_i^{a=0} = \psi_1 a + \psi_2 a L_i$$

其中 $\psi_1 + \psi_2 l$ 是 $L = l$ 时因果效应的大小。也就是说, 对个体 i , $Y_i^{a=1}$ 等于 $Y_i^{a=0} + \psi_1 + \psi_2 l$ 。个体在没有治疗时的反事实结局 $Y_i^{a=0}$ 移动 $\psi_1 + \psi_2 l$ 个单位就能得到有治疗时的反事实结局。

图 14.1 展示了 $L = l$ 分层的一个加成保序性例子。图中的钟形曲线表示反事实结局 $Y^{a=0}$ (左侧曲线) 和 $Y^{a=1}$ (右侧曲线) 的分布。上部的两个点表示个体 i 的反事实结局, 下部的两个点表示个体 j 的反事实结局。箭头表示从 $Y^{a=0}$ 移动到 $Y^{a=1}$, 对这个分层的所有人来说, 移动距离都是 $\psi_1 + \psi_2 l$ 。图 14.2 展示了另一个分层 $L = l'$ 当中的保序性。我们可以看到, 这两个分层中反事实结局的分布不一样。在图 14.2 的分层中, 所有人的移动距离都是 $\psi_1 + \psi_2 l'$, 就如图中 p 和 q 两个个体所展示的一样。

对大多数治疗和结局而言, 个体的因果效应并不是一个常数, 甚至不可能接近常数。因而(条件加成)保序性几乎不可能成立。在我们的例子中, 戒烟对每个人的因果效应不太可能都是一样的。有些人可能因基因或其他方面因素, 更容易受到戒烟效果的影响。戒烟的因果效应因人而异: 有的人戒烟后增重多一些, 有的人增重少一些, 有的人体重甚至会降低。因而现实更可能是如图 14.3 中一样, 不同人从 $Y^{a=0}$ 移动到 $Y^{a=1}$ 的距离并不一样, 于是, 保序性就不可能成立。

因为保序性不可能成立, 所以我们的因果推断方法不能依赖于保序性。实际上, 本书所讨论的方法都不需要保序性。比如, 在第十二章的边缘结构模型中, 模型的因变量是因果效应均值, 而非个体因果效应, 因而不需要保序性。我们估计得到的戒烟的因果效应均值是 3.5kg (95%置信区间: 2.5, 4.5)。这一结果不需要个体因果效应的保序性成立。同理, 上一小节介绍的结构嵌入模型也不需要保序性。

177 基于保序性的模型需要一个非常强的假设: 在 L 的同一分层中, 个体的治疗因果效应是一个常数。而我们在现实中并不希望使用这么一个不切实际的模型。不过在下一小节我们将用保序性介绍 G-估算, 这只是因为用保序性能更好地解释 G-估算, 并且不管保序性是否成立, G-估算的具体步骤都是一样的。也就是说 G-估算不一定需要保序性假设。同时要注意到, 保序性结构模型也是一个结构均值模型——一个体从 $Y^{a=0}$ 移动到 $Y^{a=1}$ 距离的均值等于个体移动的距离。

(在没有保序性假设时, 结构嵌入模型是良定的。比如, 我们可以提出一个模型用以描述图 14.3 的情形, 从而估计因果效应均值。这一因果效应均值模型和个体因果效应模型不尽相同。)

14.5 G-估算

这一小节将把前面 3 个小节的内容整合在一起。假设我们的目标是估计结构嵌入模型 $E[Y^{a=1} - Y^{a=0} | L] = \beta_1 a$ 当中的参数。为了简便, 我们仅考虑只有一个参数 β_1 的模型。因为这个模型没有乘积项 $\beta_2 aL$, 所以我们假设戒烟的因果效应均值在 L 的每个分层中都一样, 也即不存在 L 的效应修饰。

我们同时也假设加成保序性成立, 也即对所有个体 i 有 $Y_i^a - Y_i^{a=0} = \psi_1 a$ 。因而个体的因果效应 ψ_1 就等于因果效应均值 β_1 , 这也是我们想估计的值。我们进一步将保序性模型中的下角标 i 去掉, 写作 $Y^a - Y^{a=0} = \psi_1 a$, 这是因为这一模型对所有个体都适用。接下来我们通过移项得到 (很快你就知道为什么要这么做) :

$$Y^{a=0} = Y^a - \psi_1 a$$

G-估算的第一步是将模型和观测数据联系起来。根据一致性, 个体的观测结局 Y 等于个体治疗状态所对应的反事实结局, 也即如果治疗 $A = 1$, 那么 $Y = Y^{a=1}$; 如果治疗 $A = 0$, 那么 $Y = Y^{a=0}$ 。因此, 如果我们把结构模型中的固定值 a 替换为每个个体的观测值 A , 我们相应也就将反事实结局 Y^a 替换为观测结局 $Y^A = Y$ 。因而, 保序结构模型就意味着每个个体的反事实结局 $Y^{a=0}$ 可以表示为观测数据和未知参数 ψ_1 的一个函数:

$$Y^{a=0} = Y - \psi_1 A$$

如果这个模型是正确的, 并且我们知道 ψ_1 的值, 那我们就能计算每个个体没有治疗时的反事实结局 $Y^{a=0}$ 。现在我们还不知道 ψ_1 的值, 我们的目标就是估计 ψ_1 。

让我们再思考一个问题。某人告诉你他知道 ψ_1 的值, 但他不直接告诉你是多少, 而是说 ψ_1 是以下三个值中的一个: $\psi^\dagger = -20$ 、 $\psi^\dagger = 0$ 、或 $\psi^\dagger = 10$, 并让你猜一下哪个是正确的。为了回答这个问题, 我们在每一种可能下计算

$$H(\psi^\dagger) = Y - \psi^\dagger A$$

- 178 现在我们得到 3 种不同的反事实结局取值: $H(-20)$ 、 $H(0)$ 以及 $H(10)$, 其中只有一个正确的是 $Y^{a=0}$ 。或者说, 当 $\psi^\dagger = \psi_1$ 的时候, $H(\psi^\dagger) = Y^{a=0}$ 。在这个问题中, 选择正确的 ψ_1 的值, 就等价于选择正确的反事实结局 $Y^{a=0} = H(\psi_1)$ 。因而, 我们应该怎样选择正确的 $H(\psi^\dagger)$?

在 14.2 小节, 我们论述了互换性假设能用概率进行表述, 也就是可以在因变量是治疗且包含了反事实结局和混杂变量 L 的 logistic 模型中进行表示。当有界互换性成立的时候, 反事实结局的参数 α_1 应该等于 0。因而, 我们可以有一种很简便的方法, 在三个 $H(\psi^\dagger)$ 中选出正确的反事实结局。我们用这三个值分别对下述 logistic 模型进行拟合:

$$\text{logit Pr}[A = 1 | H(\psi^\dagger), L] = \alpha_0 + \alpha_1 H(\psi^\dagger) + \alpha_2 L$$

$\alpha_1 = 0$ 时对应的 $H(\psi^\dagger)$ 就是真正的反事实结局 $Y^{a=0}$, 此时的 ψ^\dagger 就是 ψ_1 的真实值。比如, 假设 $H(\psi^\dagger = 10)$ 在控制了 L 后和治疗 A 无关, 那么 ψ_1 的估计值 $\hat{\psi}_1$ 就是 10, 我们的估计就结束了。这就是 G-估算。

(Rosenbaum (1987) 在非时异治疗的情形中提出了这一估计过程。)

(谨记: G-估算并不意味着有界互换性是否成立, 而是假设有界互换性成立。)

然而在实践中, 没有人会给出备选选项, 因而我们需要自己去找到 ψ_1 的估计值。因此, 我们需要遍历所有 ψ^\dagger 的可能取值, 直到 $\alpha_1 = 0$ 。因为不可能去检验所有可能取值——毕竟, 任意一个区间, 都有无数个取值——所以我们只能尽可能精细地去遍历 ψ^\dagger 的可能取值。比如, 在 -20 到 20 这个区间范围内, 我们以 0.01 为单位, 遍历所有可能取值。我们的遍历越精细, 我们也就能越接近真实估计值 $\hat{\psi}_1$, 然而这对算力的要求也就越高。

(参见代码 14.2。)

在我们戒烟的例子中, 我们选取区间为 2.0 到 5.0, 遍历单位是 0.1。首先要计算出每个个体在 31 个不同 ψ^\dagger 取值下的 $H(2.0)$, $H(2.1)$ …… $H(5.0)$, 然后再根据这些值拟合 31 个不同的预测戒烟概率的 logistic 模型, 这些模型和第十二章用来估计逆概率权重分母的模型基本一样, 唯一区别是每个模型中多了 $H(\psi^\dagger)$ 一项。 $H(\psi^\dagger)$ 的参数估计 $\hat{\alpha}_1$ 在 $H(3.4)$ 和 $H(3.5)$ 的时候最接近 0。我们在 3.4 到 3.5 这个区间更精细地遍历一遍, 得到 $H(3.446)$ 时 $\hat{\alpha}_1 = 0$ 。因而, 我们的 G-估算得到戒烟对增重的因果效应均值 $\psi_1 = \beta_1$ 是 3.4kg。

为了计算 G-估算的 95% 置信区间，我们可以使用上述 logistic 模型中 $\alpha_1 = 0$ 的 P 值检验。按照设想， $\psi^\dagger = 3.446$ 时， $\widehat{\alpha}_1 = 0$ ，此时的 P 值应该是 1（实际上是 0.998）。而根据我们上面拟合的 31 个 logistic 模型的结果，我们知道 ψ^\dagger 在 2.5 到 4.5 之间时，P 值大于 0.05。因而，95% 置信区间是 2.5 到 4.5。另一种方法是用自举法计算 95% 置信区间。

（我们也可以用 Wald 检验计算 P 值，也可以使用其他有效检验。比如，我们可以用 Score 检验，这将简化我们的计算（它不需要拟合多个模型），并且在大样本中等价于 Wald 检验。）

一般而言，计算 G-估算的 95% 置信区间，需要找到 $\alpha_1 = 0$ 的检验 P 值大于 0.05 所对应的 ψ^\dagger 的取值。这可以通过逆向检验 $\alpha_1 = 0$ 得到：95% 置信区间的边界值就是使 $P > 0.05$ 的值。在我们的例子中，统计检验是在稳健方差估计的基础上进行的，因为我们使用了逆概率权重去调整删失。因此，我们的 95% 置信区间在大样本中，会至少在 95% 的时间中包含真实值。在大样本中，自举法给出的置信区间并不保守，因而也就更窄。

（在有删失存在的情况下，logistic 模型需要在未被删失的人群中拟合。而每个个体的贡献需要用逆概率权重进行加权。参见知识点 14.2。）

让我们回到非保序的模型中。假设模型是正确设定的（也就是说， L 的不同分层中效应均值相等），那 G-估算的算法（也即执行这一过程的代码）会给出均值模型中参数 β_1 的一致估计，而不管每个个体的因果效应是否为常数，也即不管（条件加成）保序性是否成立。换句话说，G-估算算法的有效性并不需要 $H(\beta_1) = Y^{a=0}$ 对每个个体都成立，其中 β_1 是均值模型的参数值。与之相反，这一算法值要求在控制了 L 之后， $H(\beta_1)$ 和 $Y^{a=0}$ 的条件均值相等。

有趣的是，上述 G-估算步骤能够进行未知混杂的敏感性分析，详见精讲点 14.2。

14.6 两个或多个参数的结构嵌入模型

迄今我们的结构嵌入模型中只有一个参数 β_1 。我们没有包含乘积项 $\beta_2 \alpha L$ 是因为我们假设戒烟的因果效应在 L 的分层中都一样。不过，如果 L 中的部分变量 V 有效应修饰作用的话，那我们的模型就是错误设定的，我们的因果推断也就是错的。这一点与边缘结构模型 $E[Y^a] = \beta_0 + \beta_1 a$ 有所区别。在边缘结构模型中，即使 V 有效应修饰作用，没有在模型中增加 $\beta_2 \alpha L$ 和 $\beta_3 V$ 两项并不会使模型设定错误。在没有控制 V 的时候，边缘结构模型估计的是在人群中的因果效应均值；在控制了 V 的时候，边缘结构模型估计的是 V 不同分层中的因果效应均值。而根据定义，结构嵌

入模型估计的是 L 不同分层中的效应均值, 而不是人群中的因果效应均值。如果存在效应修饰作用, 未包括乘积项的结构嵌入模型是错误设定的, 给出的结果将是有偏的。

(我们在第十二章讨论过, 边缘结构模型的一个特性是零值保留: 当因果效应均值为 0 的时候, 模型就不会错误设定。结构嵌入模型也具备这一特性。不过, 虽然参数 G-公式对时间固定的治疗具备零值保留特性, 但对时异治疗并不具备这一特性。参见本书第三部分。)

幸运的是, 我们可以在上一小节介绍的 G-估算中加入乘积项。比如, 假设我们认为吸烟频率 V 能够修饰戒烟的因果效应。此时的结构嵌入模型是 $E[Y - Y^{a=0} | A = a, L] = \beta_1 a + \beta_2 aV$, 其对应的保序模型是 $Y_i^a - Y_i^{a=0} = \psi_1 a + \psi_2 aV_i$ 。因为这个结构模型有两个参数 ψ_1 和 ψ_2 , 所以我们要在 $\Pr[A = 1 | H(\psi^\dagger), L]$ 的逆概率加权模型中包含两个参数, 此时 $\psi^\dagger = (\psi_1^\dagger, \psi_2^\dagger)$ 。比如, 我们就可以拟合以下 logistic 模型:

$$\text{logit } \Pr[A = 1 | H(\psi^\dagger), L] = \alpha_0 + \alpha_1 H(\psi^\dagger) + \alpha_2 H(\psi^\dagger)V + \alpha_3 L$$

在这个模型中, 我们需要找到可能的 ψ_1^\dagger 和 ψ_2^\dagger 组合, 使得参数 α_1 和 α_2 都等于 0。

因为这个模型有两个参数, 所以需要在一个二维空间中遍历所有可能选择, 因而也需要更多的时间和算力。不过, 现在的统计软件大多都能提供需要较少算力的近似方法。对于线性均值模型 (比如我们此处讨论的模型) 来说, 其参数估计式存在一个闭合公式, 因而我们可以直接用这一公式进行计算, 也就不需要遍历所有可能取值 (参见知识点 14.2)。但对于某些模型, 比如生存模型, 则不存在这一闭合公式。在我们的这个例子中, 我们能得到 $\widehat{\psi}_1 = 2.86$, $\widehat{\psi}_2 = 0.03$ 。对应的 95% 置信区间能够用联合检验 $\alpha_1 = \alpha_2 = 0$ 的 P 值进行计算, 或者用自举法计算。

(Nelder-Mead 单纯形法是其中一种方法。)

(参见代码 14.3。)

在更一般情况下, 戒烟的因果效应可能在 L 的所有分层中都不一样。对于二分变量, 对应的保序模型是 $Y_i^a - Y_i^{a=0} = \psi_1 a + a \sum_{j=1}^p \psi_{2j} L_{ij}$, 其中有 $p+1$ 个参数, $\psi_1, \psi_{21}, \dots, \psi_{2p}$ 。 ψ_{2j} 是乘积项 aL_j 的系数, 且 L_j 表示 L 中的一个变量, 而 L 总共有 p 个变量。整个研究人群中的因果效应均值可以写成 $\psi_1 + \frac{1}{n} a \sum_{j=1}^p \psi_{2j} L_j$, 其中 n 是研究人群的总人数。在实践中, 很少使用有多个参数的结果嵌入模型。

(你可能会觉得没必要使用有多个参数的结构嵌入模型。如果 L 中的变量都是离散的, 且我们的研究人群足够大, 难研究者就能在 L 的每一分层中拟合仅含一个参数的结构嵌入模型。)

实际上, 不管什么形式的结构嵌入模型都很少有人使用, 一方面是因为没有统计软件能便捷地直接运行结构嵌入模型, 另一方面是因为如果把结构嵌入模型推广到生存分析, 我们需要考虑其他更多因素(参见第十七章)。下一章我们将会讨论两种调整混杂最常用的方法: 结局回归和倾向性评分。

第十四章精讲点和知识点

精讲点 14.1: 边缘结构模型与结果嵌入模型的关系(原书第 174 页)

V 是 L 中的一个二分变量, 在 V 的每一分层中考虑治疗取值为 a 时的边缘结构模型:

$$E[Y^a | V] = \beta_0 + \beta_1 a + \beta_2 aV + \beta_3 V$$

$\beta_1 + \beta_2 v$ 是 $V = v$ 分层中的因果效应均值 $E[Y^{a=1} - Y^{a=0} | V = v]$, 而 $\beta_0 + \beta_3 v$ 是 $V=v$ 分层中没有治疗时的反事实结局 $E[Y^{a=0} | V = v]$ 。假设我们的目标是估计因果效应均值 $\beta_1 + \beta_2 v$, 也即我们对 $\beta_0 + \beta_3 v = E[Y^{a=0} | V = v]$ 不感兴趣, 那模型可以写作 $E[Y^a | V] = E[Y^{a=0} | V] + \beta_1 a + \beta_2 aV$, 或者, 写作:

$$E[Y^a - Y^{a=0} | V] = \beta_1 a + \beta_2 aV$$

这被称为半参数边缘结构模型, 因为这个模型并没有设定没有治疗时的反事实结局 $E[Y^{a=0} | V]$ 。

在没有删失的时候, 我们要估计这个半参数边缘结构模型中的参数。首先, 我们需要用逆概率权重 $SW^A = f(A|v)/f(A|L)$ 构建一个虚拟人群。在虚拟人群中, 只存在一个混杂变量, 即 V , 因而这个半参数边缘结构模型是一个饱和嵌入模型, 我们可以用 G-估算对参数进行估计, 值需要把 L 替换成 V , 把权重替换成 SW^A 即可。因此, 在没有时异治疗的情形中, 结构嵌入模型和半参数边缘结构模型是一样的。因为边缘结构模型中的参数比结构嵌入模型中的参数多, 所以结构嵌入模型可能更不易受模型错误设定的影响。

再考虑一个特殊情形: L 所有分层中的半参数边缘结构模型, 也就是仿制的半参数边缘结构模型 $E[Y^a - Y^{a=0} | L] = \beta_1 a + \beta_2 aL$ 。在有界互换性假设下, 这个模型就是我们本章讨论的结果嵌入模型。

精讲点 14.2: 未测混杂的敏感性分析 (原书第 179 页)

G-估算依赖于有界互换性成立的情形下 $\alpha_1 = 0$ 。现在让我们考虑有界互换性不成立的情形。

比如, 如果一个人的伴侣也抽烟, 那么这个人戒烟的概率就会降低, 同时伴侣的吸烟状态和增重 Y 的某个决定因素密切相关, 而这个因素不包含在 L 中。这样一来, 就存在未测混杂, 从而 L 就不足以达成互换性, 控制了 L 之后戒烟 A 和反事实结局 $Y^{a=0}$ 依然相关。也就是说, 在我们正文的 G-估算中, $\alpha_1 \neq 0$ 。

而 G-估算并不要求 $\alpha_1 = 0$ 。假设因为存在未测混杂, 所以 α_1 应该是 0.1 而不是 0, 那我们依然可以使用正文中描述的 G-估算方法, 只不过此时我们需要找的估计值是让 $\alpha_1 = 0.1$, 而不是 $\alpha_1 = 0$ 。G-估算并不要求有界互换性成立, 而是要求有界互换性不成立的程度——也即 α_1 的值——是已知的。这一特性能够让我们用 G-估算去进行未测混杂的敏感性分析。

如果我们相信 L 不足以调整所有混杂, 那我们可以在不同的未测混杂情形下——也即不同的 α_1 取值下——重复我们的 G-估算分析, 然后再将不同情形下的效应估计画成图展示出来。这幅图将会展现我们的效应估计对不同强度、不同方向的未测混杂的敏感性。不过, 一个问题是, 我们应该如何量化未测混杂? $\alpha_1 = 0.1$ 究竟代表多少未测混杂? Robins, Rotnitzky 和 Scharfstein (1999) 详细讨论了 G-估算中未测混杂敏感性分析的具体细节。

知识点 14.1: 乘性结构嵌入模型 (原书第 175 页)

在正文中, 我们指考虑了加性结构嵌入模型。当结局变量 Y 只能取正值的时候, 我们可能需要使用乘性结构嵌入模型。一个例子是:

$$\log\left(\frac{\mathbb{E}[Y^a | A = a, L]}{\mathbb{E}[Y^{a=0} | A = a, L]}\right) = \beta_1 a + \beta_2 aL$$

我们可以通过 G-估算拟合这个模型, 其中 $H(\psi^\dagger) = Y \exp(-\psi_1^\dagger a - \psi_2^\dagger aL)$ 。

上述乘性模型可以用于二分结局变量, 只需要在 L 的所有分层中, $Y = 1$ 的概率都不是太大。否则, 这个模型给出的概率估计可能会大于 1。如果概率较大, 我们可以考虑以下二分变量 Y 的结构嵌入 logistic 模型:

$$\text{logit Pr}[Y^a = 1 | A = a, L] - \text{logit Pr}[Y^{a=0} = 1 | A = a, L] = \beta_1 a + \beta_2 aL$$

不过, 结构嵌入 logistic 模型不适用于时异性治疗, 并且其中的参数不能用正文中的算法个计算。详情请参阅 Robins (1999) 和 Tchetgen (2011) 等人所著论文。

知识点 14.2: 结构嵌入均值模型的 G-估算 (原书第 181 页)

考虑结构嵌入模型 $E[Y - Y^{a=0} | A = a, L] = \beta_1 a$ 。在正文给出的假设中, 我们可以用 G-估算得到 β_1 的一致估计。具体而言, β_1 的估计值能使 $H(\psi^\dagger)$ 和 A 之间的相关性最小。当我们把 G-估算和 Score 检验结合在一起时, 这一过程就相当于求解下列关于 ψ^\dagger 的方程:

$$\sum_{i=1}^n I[C_i = 0] W_i^C H_i(\psi^\dagger)(A_i - E[A | L_i]) = 0$$

其中, $I[C_i = 0]$ 在 $C_i = 0$ 时取 1, 其余时候为 0; 逆概率权重 W_i^C 和期望 $E[A | L_i] = \Pr[A = 1 | L_i]$ 可以用估计值替代。 $E[A | L_i]$ 可以用因变量是治疗、且控制了协变量 L 的 logistic 模型进行估计, 这个模型中如果个体 i 未被删失 ($C_i = 0$), 则权重为 W_i^C 。(如果所有个体都未被删失, 且对每个个体我们都观察到了 A 和 Y , 那我们就可以用未加权的 logistic 回归计算 $E[A | L_i]$ 。)

这一方程有解析解, 因而能够直接计算, 也即不用遍历参数空间。具体而言, 因为

$H_i(\psi^\dagger) = Y_i - \psi^\dagger A_i$, 我们可以得到 $\widehat{\psi}_1$ 等于:

$$\sum_{i=1}^n I[C_i = 0] W_i^C Y_i (A_i - E[A | L_i]) / \sum_{i=1}^n I[C_i = 0] W_i^C A_i (A_i - E[A | L_i])$$

如果 ψ 是多维的, 我们需要在方程左侧乘以多维向量 L 的一个函数。函数的选择将影响我们的统计效率, 但并不会影响一致性。也就是说, 虽然函数地选择会生成有效的置信区间, 但是置信区间的宽度却取决于函数。Robins (1994) 给出了结构嵌入模型的正式描述, 并且推导了能最小化置信区间的函数。

如果我们将上述方程中的 $H_i(\psi^\dagger)$ 替换为一个非线性函数, 比如 $[H_i(\psi^\dagger)]^3$, 我们依然可以提高统计效率, 且保证估计的一致性, 那么这个方程就是一个自然方程。不过, 在本章的结构嵌入模型中, 我们不能在这个方程中使用非线性函数, 这是因为本章的模型仅假设了给定 L 的均值独立性, 即 $E[H(\beta_1) | A, L] = E[H(\beta_1) | L]$ 。而使用 $H_i(\psi^\dagger)$ 的非线性函数需要假设分布独立性, 即 $H(\beta_1) \perp\!\!\!\perp A | L$ 。某些 (本章未涉及的) 结构嵌入模型描述的是 Y^a 的分位数分布, 此时可以使用 $H_i(\psi^\dagger)$ 的非线性函数。

只有用来估计 $E[A|L]$ 和 $\Pr[C=1|A,L]$ 的模型都是正确的时候, ψ 的估计才是一致的。我们可以将方程中的 $H(\psi^\dagger)$ 替换为 $H(\psi^\dagger) - E[H(\psi^\dagger)|L]$, 从而得到一个更稳健的估计。为了估计 $H(\psi^\dagger) - E[H(\psi^\dagger)|L]$, 我们可以拟合一个因变量是 $E[H(\psi^\dagger)|L] = E[Y^{a=0}|L]$ 的未加权线性模型。如果这个模型是正确的, 那么 ψ 的估计也是一致的, 即使 $E[A|L]$ 和 $\Pr[C=1|A,L]$ 的模型都是设定错误的。因此, 如果我们满足下述条件之一, 那么 ψ 的估计就是一致的: (1) $E[H(\psi^\dagger)|L]$ 的模型是正确的; (2) $E[A|L]$ 和 $\Pr[C=1|A,L]$ 的模型都是正确的。我们将这一估计称为双重稳健估计。Robins (2000) 给出了线性结构嵌入模型双重稳健估计的闭合形式。

第十四章图表

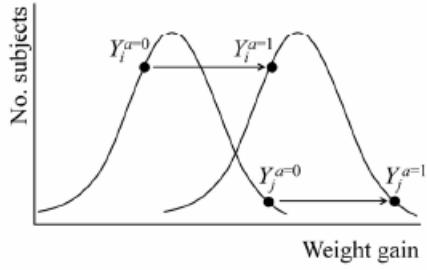


Figure 14.1

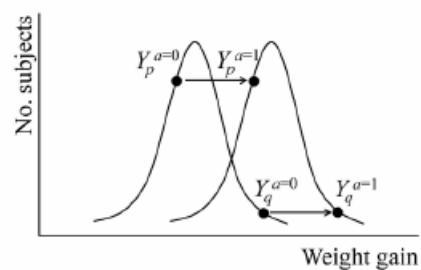


Figure 14.2

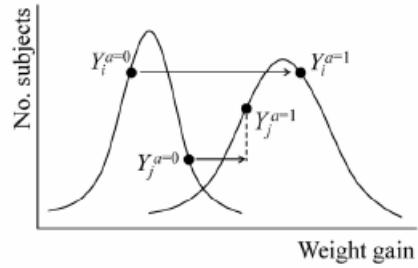


Figure 14.3

第十五章 结局回归与倾向性评分

183 结局回归和不同形式的倾向性评分分析方法是因果推断中最常用的参数化方法。你可能会好奇,为什么我们现在才介绍这些方法。迄今,我们讨论了逆概率加权,标准化,以及G-估算,这些方法被统称为G-方法。可能先介绍不常用的方法,再介绍常用方法,这一顺序看起来很奇怪。那为什么我们不直接介绍结局回归和倾向性评分呢?原因很简单,因为这两种方法并不是在所有情形中都有用。

已经有无数著作详细介绍并运用了这两种方法。然而,这两种方法只在较简单的情形中适用。当情形更复杂一些,比如涉及时异性治疗的时候,这两种方法就不再适用。在本书第三部分我们将主要讨论G-方法。本章将介绍结局回归和倾向性评分这两种方法。不过谨记,这两种方法不适用于复杂的纵向数据。

15.1 结局回归

在前三章,我们讨论了逆概率加权,标准化,以及G-估算三种方法,并用它们估计了戒烟A(治疗)对增重Y(结局)的因果效应。我们也介绍了在边缘结构模型或结构嵌入模型中增添乘积项,从而估计子人群中的因果效应。以结构嵌入模型为例,这一模型可以包含治疗A和变量L的乘积项,不过模型中可以不需要添加单独的L项。这是结构嵌入模型的一个重要特性,这是因为我们只想估计不同L分层中A对Y的因果效应,而不是L和Y的关系。使用结构嵌入模型,我们不需要知道L-Y之间的函数关系,因而就不容易受到模型错误设定的影响。

(谨记: 我们将效应均值定义为 $E[Y^{a=1,c=0}] - E[Y^{a=0,c=0}]$ 。我们假设在控制了L之后,治疗组和非治疗组可互换。)

另一方面,如果我们愿意在A的不同分组中加入L-Y之间的关系,那我们可以考虑下述结构模型:

$$E[Y^{a,c=0} | L] = \beta_0 + \beta_1 a + \beta_2 aL + \beta_3 L$$

其中 β_2 和 β_3 是参数向量。 L 不同分层中A对Y的因果效应可以用 β_1 和 β_2 表示,而 L 每一分层中没有治疗的反事实结局可以用 β_0 和 β_3 表示。 β_3 经常被认为是L的“主要效应”,但是使用“效应”这个词汇很有误导性,因为 β_3 可能在阐释上并没有因果意义(可能存在对L的混杂)。

184 参数 β_3 仅仅表示结构模型中反事实结局 $Y^{a=0,c=0}$ 随L的变化情况。精讲点15.1详细讨论了为什么 β_0 和 β_3 没有因果意义。

(在第十二章, 我们将这个模型称为仿制的边缘结构模型, 因为它的形式和边缘结构模型一样, 但没必要使用逆概率加权。所有个体的稳定逆概率权重 $SW^A(L)$ 都等于 1, 这是因为我们控制了整个 L , 而非 L 的一个子集 V 。)

在互换性、正数性成立, 以及治疗良定的情况下, L 的每个分层 l 中如果所有人都接受了治疗的反事实结局, 即 $E[Y^{a=1,c=0} | L=l]$, 等于治疗组的实际结局, 即 $E[Y | A=1, C=0, L=l]$ 。对未治疗组同理。因此上述结构模型中的参数就能通过拟合常规结局回归得到, 即:

$$E[Y | A, C = 0, L] = \alpha_0 + \alpha_1 A + \alpha_2 AL + \alpha_3 L$$

我们在 13.2 小节中介绍过这个模型。如同第三章的分层一样, 在 L 的每一分层中进行一次结局回归, 能够调整可能的混杂, 从而估计治疗的因果效应。如果变量 L 足以调整所有的混杂 (以及选择偏移), 并且结局模型是正确设定的, 那么我们就不再需要进行更多调整。就是说, 这个模型中的 α 等于结构模型中的 β 。

(β_0 和 β_3 的存在表明 $Y^{a=0,c=0}$ 取决于 L 。如果我们想估计反事实结局均值, 或者乘法尺度 (而非加法尺度) 上的条件效应 (也即 L 分层中的效应), 那这两个参数是必需的。)

在 13.2 小节, 结局回归只是估计标准化结局均值的中间步骤。在这里, 结局回归则是最后一步。与标准化方法不同的是, 我们不再对条件均值进行标准化进而估计边缘均值, 而只是简单比较条件均值的估计值。在 13.2 小节中, 我们的回归模型中只有一个乘积项 (戒烟和吸烟频率), 也就是说, 我们先验地假设其他乘积项的系数是 0。使用同一模型, 我们得到参数估计值为:

$\hat{\beta}_1 = 2.6$, $\hat{\beta}_2 = 0.05$ 。此时, $\hat{E}[Y | A=1, C=0, L] - \hat{E}[Y | A=0, C=0, L]$ 在吸烟频率 5 支/天时是 2.8 (95%置信区间: 1.5, 4.1), 在吸烟频率 40 支/天时是 4.4 (95%置信区间: 2.8, 6.1)。使用结局回归的时候, 一般会假设 L 中的变量不存在效应修饰作用, 于是拟合的模型中就没有任何乘积项, $\hat{\beta}_1$ 就是效应均值的估计值。在我们的例子中, 如果没有任何乘积项, 得到的估计值是 3.5 (95%置信区间: 2.6, 4.3)。

在本章, 我们不再解释如何去拟合一个结局回归模型。(在第十三章已经解释过一次。) 对于离散型结局 Y 来说, 我们也可以使用常规结局回归, 此时我们需要拟合一个 logistic 模型, 其因变量是 $\Pr[Y=1 | A=a, C=0, L]$ 。

15.2 倾向性评分

当使用逆概率加权（第十二章）和 G-估算（第十四章）的时候，我们需要先估计给定变量 L 时治疗的概率。我们把这一条件概率记为 $\pi(L)$ 。 $\pi(L)$ 越接近 0，表示个体接受治疗的概率越低； $\pi(L)$ 越接近 1，表示个体接受治疗的概率越高。也就是说，在给定 L 的情况下， $\pi(L)$ 衡量的是个体接受治疗的倾向。因此 $\pi(L)$ 被称为倾向性评分。

在一个假想实验中，有一半的人被分到治疗组 $A = 1$ 。因而对所有人而言， $\pi(L) = 0.5$ 。同时也要注意，这里不管 L 是什么，都有 $\pi(L) = 0.5$ 。与之相比，在观察性研究中，某些个体接受治疗的概率要比另一些个体更高。此时，因为治疗分配不是研究者所能控制的，所以真实的倾向性评分 $\pi(L)$ 是未知的，因而我们需要使用数据来估计这一概率。

（参见代码 15.2。此处我们仅考虑二分治疗的倾向性评分。使用倾向性评分的分析方法很难推广到非二分治疗的情形。）

在我们的例子中，我们拟合了一个 logistic 模型，用来估计给定变量 L 时戒烟 A 的概率，从而得到倾向性评分。这个模型和逆概率加权以及 G-估算中的模型一样。在这个模型之中，编号 22941 的倾向性评分是 0.053，编号 24949 的则是 0.793。图 15.1 展示了倾向性评分在戒烟组 $A = 1$ （下侧）和非戒烟组 $A = 0$ （上侧）中的分布情形。正如我们所设想的一样，平均而言，戒烟组中的倾向性评分均值（0.312）大于非戒烟组中的均值（0.245）。如果 $\pi(L)$ 在治疗组 $A = 1$ 和非治疗组 $A = 0$ 中的分布一样，那么就不存在 L 带来的混杂，也即在因果图中，没有从 L 到 A 的开放路径。

一般而言，拥有相同倾向性评分 $\pi(L)$ 的个体，不一定有相同的 L 。比如，两个 $\pi(L) = 0.2$ 的个体，可能在吸烟频率和锻炼程度上有所不同，但是综合考虑所有 L 的时候，他们的戒烟 ($A = 1$) 概率却是相同的。在超级人群中，对倾向性评分 $\pi(L)$ 等于某一特定值的所有人而言，他们 L 的分布可以不尽相同，但是在他们的戒烟者和非戒烟者两组中， L 的分布是相同的，也就是说， $A \perp\!\!\!\perp L | \pi(L)$ 。因而，我们会说倾向性评分平衡了治疗组和非治疗组中的变量分布。

（在研究人群中，因为抽样变异性，倾向性评分只能“近似地”平衡 L 的分布。正确的倾向性评分模型才能给出更好的平衡效果。）

当然，倾向性评分只能平衡已测的变量 L ，而不能阻止未测变量带来的混杂。随机分配则能 186 同时平衡已测的和未测的变量，因而被认为是消除混杂的最好方法。知识点 15.1 给出了平衡分数的正式定义。

与其他因果推断方法相同, 使用倾向性评分需要互换性、正数性、以及一致性假设。治疗组和非治疗组可互换也就意味着倾向性评分 $\pi(L)$ 可互换。也就是说, 有界互换性 $Y^a \perp\!\!\!\perp A|L$ 意味着 $A \perp\!\!\!\perp L|\pi(L)$ 。正数性意味着没有任何个体的倾向性评分等于 1 或 0。当且仅当 L 每一分层中的正数性成立时, $\pi(L)$ 每一分层中的正数性才会成立。

(如果 L 不足以调整所有的混杂和选择偏移, 那么 $\pi(L)$ 也不能。)

在互换性和正数性假设下, 我们可以使用倾向性评分, 通过分层 (包括结局回归)、标准化、和匹配等方式估计因果效应。下面两个小节会讨论如何使用这些方法。首先第一步, 我们需要从数据中估计倾向性评分 $\pi(L)$, 然后在分层、标准化、以及匹配中用 $\pi(L)$ 替代 L 。

(在随机试验中, $\pi(L)$ 的估计值比真实值效果更好, 因为估计值能同时调整变量的系统性和随机性不均衡, 而真实值会忽略随机性不均衡。)

15.3 倾向性分层和标准化

在人群中, 如果倾向性评分 $\pi(L)$ 等于某个值 s , 那么因果效应均值可以表述为

$E[Y^{a=1,c=0} | \pi(L)=s] - E[Y^{a=0,c=0} | \pi(L)=s]$, 在可识别假设下, 也就等于

$E[Y | A=1, C=0, \pi(L)=s] - E[Y | A=0, C=0, \pi(L)=s]$ 。我们可以只在 $\pi(L)=s$ 的人群中进行数据分析, 从而得到这一效应估计值。然而, 一般而言, $\pi(L)$ 是一个连续性变量, 取值可以是 0 和 1 之间的任意值, 因而不太可能会有两个人的 $\pi(L)$ 值完全相同。比如, 只有编号 22005 的 $\pi(L)$ 估计值是 0.6563, 也就意味着我们不能通过比较两个 $\pi(L)$ 估计值为 0.6563 的个体, 从而得到因果效应均值。

解决这一问题的一个方法是分层, 每一层中的个体虽然 $\pi(L)$ 估计值并不完全一样, 但是却大致相同。用 $\pi(L)$ 估计值的十分位数进行分层是一种常见做法。我们可以根据 $\pi(L)$ 估计值的十分位数, 把研究人群分为十个子群体, 然后再在每个子群体中估计因果效应均值。在我们的例子中, 每个子群体中有 162 名个体。在这十个子群体中, 戒烟对增重的因果效应均值从 0 到 6.6kg 不等, 但是每个效应估计值的 95% 置信区间都很宽。

(参见代码 15.3。)

我们也可以通过拟合 $E[Y | A, C = 0, \pi(L)]$ 的回归模型得到效应估计。在模型中, 自变量包括治疗 A , $\pi(L)$ 十个子群体的 9 个指示变量 (十个子群体中的一个会用作参考组, 因而其信息已经被包含在模型截距中, 因此只需要 9 个指示变量), 以及指示变量和治疗 A 的 9 个乘积项。在现实中, 大多数研究并不会包含乘积项, 也即研究者假设 $\pi(L)$ 不具有效应修饰作用。在我们的例子中, 没有乘积项的模型给出的效应估计是 3.5kg (95%置信区间: 2.6, 4.4)。精讲点 15.2 讨论了倾向性评分效应修饰的更多细节。

根据十分位数或者倾向性评分的其他函数进行分层可能会导致一个潜在问题: 一般而言, 在这些分层中, 治疗组和非治疗组的 $\pi(L)$ 分布会不一样, 因而治疗组和非治疗组在这些分层中不可互换。这一问题在前几章并不存在。当时我们用倾向性评分的某一函数——也即逆概率权重——在结构模型和 G-估算中估计效应均值, 而这些方法用的是倾向性评分的连续值, 而非几个分类。同理, 如果我们在 $E[Y | A, C = 0, \pi(L)]$ 的回归模型中将 $\pi(L)$ 视作连续变量, 也就不会存在这一问题, 此时我们得到的效应估计值是 3.6kg (95%置信区间: 2.7, 4.5)。

(注意: 二分治疗 A 的逆概率权重的分母并不是倾向性评分 $\pi(L)$, 而是 $\pi(L)$ 的某一函数。
对治疗组 ($A = 1$) 来说, 是 $\pi(L)$, 而对非治疗组 ($A = 0$) 来说, 是 $1 - \pi(L)$ 。)

这一方法的有效性取决于我们是否正确设定了 $\pi(L)$ 和结局 Y 之间的关系 (在我们的例子中, 我们假设是线性关系)。然而, 因为倾向性评分是多维变量 L 的一个一维概括, 所以我们可以通过拟合更灵活的模型来避免这一关系的错误设定, 比如, 我们可以拟合倾向性评分的三次样条曲线模型。不过我们需要注意, 逆概率加权和 G-估算并不需要知道倾向性评分和结局之间的关系。

(尽管倾向性评分是一维的, 我们仍然需要通过高维变量 L 来估计它。)

如果 $E[Y | A, C = 0, \pi(L)]$ 的参数假设是正确的, 且互换性和正数性成立, 那么模型就能估计所有分层中的因果效应均值。如果我们对整个人群的因果效应均值 $E[Y^{a=1,c=0}] - E[Y^{a=0,c=0}]$ 感兴趣, 那我们可以用倾向性评分的分布把条件均值 $E[Y | A, C = 0, \pi(L)]$ 标准化。这一步骤和第十三章对连续变量的做法一样, 只是我们将 L 换成了 $\pi(L)$ 。注意到, 这一步骤会自动在模型中包含一个治疗 A 和倾向性评分 $\pi(L)$ 的乘积项。在我们的例子中, 这一方法得到的标准化效应估计是 3.6kg (95%置信区间: 2.7, 4.6)。

(参见代码 15.4。)

188 15.4 倾向性评分匹配

根据倾向性评分 $\pi(L)$ 进行匹配和根据某一连续变量进行匹配是一样的，我们在第四章简要讨论过这一方法。我们有许多不同的匹配方法，而所有这些方法都是为了构建一个匹配人群，其中治疗组和非治疗组 $\pi(L)$ 分布相同，从而这两组可互换。比如，每个治疗组的个体都会匹配一个（或多个）倾向性评分相同的非治疗组个体。因而匹配人群就是原人群的一个子集，包含了多个治疗组-非治疗组匹配对。在给定 $\pi(L)$ 的互换性和正数性假设下，匹配人群中的相关性量度就是效应量度的一致估计。

（在匹配完成后，匹配人群中 $\pi(L)$ 的分布可以是治疗组中的分布，可以是非治疗组中的分布，也可以是某一任意分布。）

（在过去，匹配的一个缺陷是没人知道怎么计算效应估计的方差。不过 Abadie 和 Imbens 在 2006 年的一篇论文中解决了这个问题。）

不过，我们要再一次强调，几乎不可能存在倾向性评分 $\pi(L)$ 相等的两个个体。在我们的例子中，我们将给治疗组中的每个个体匹配一个和其倾向性评分 $\pi(L)$ 最接近的非治疗组个体，且这两者 $\pi(L)$ 的差值在 0.05 以内。比如，编号 1089 ($\pi(L)$ 的估计值为 0.6563) 可能就会和编号 1088 ($\pi(L)$ 的估计值为 0.6579) 进行匹配。现在有许多种算法可以用来进行匹配，而这些匹配算法不在本书的讨论范畴之内。

匹配过程也是一个偏差方差权衡过程。如果匹配条件太过宽松，那么 $\pi(L)$ 差距较大的两个个体可能就会被匹配成一对，从而在我们的匹配人群中，治疗组和非治疗组 $\pi(L)$ 的分布会有较大差距。而如果匹配条件太过严苛，许多个体又可能找不到匹配对象从而被舍弃，因而即使互换性成立，效应估计的 95% 置信区间也会变宽。

而匹配过程也和正数性相关。在我们戒烟的例子中，治疗组和非治疗组 $\pi(L)$ 的分布在大部分区间是有重合的（参见图 15.1）。只有 2 个治疗组个体（研究人群的 0.01%）的 $\pi(L)$ 估计值大于所有非治疗组的个体。我们在上一小节使用结局回归并把 $\pi(L)$ 放进模型中的时候，其实是假设了只是因为随机性，非治疗组才没有这么大的取值，因而在理论上把所有个体都包含在了我们的分析当中。然而，在匹配的时候，可能因为非治疗组中不存在相近的取值，所以这两个治疗

组的个体就会被排除于匹配分析之外。在这一点上，匹配分析并不会区分正数性是随机不成立，还是结构性地不成立。

(我们要注意，此时正数性是根据倾向性评分进行定义，也即对任意满足 $\Pr[\pi(L)=s] > 0$ 的 s ， $\Pr[A=a | \pi(L)=s] > 0$ 。)

189

以上讨论说明了匹配人群为什么可能会和我们的目标人群（或超级人群）不同。理论上，只要一个目标人群的特征被清晰描述，倾向性评分匹配就能用来估计这个人群中的因果效应。比如，如果我们给每个治疗组个体都匹配了一个非治疗组个体，并且排除了未被匹配的非治疗组个体，那我们就能估计治疗组中的因果效应（详见精讲点 15.2）。然而，在实践中，倾向性评分匹配后得到的效应估计却很难说明适用于哪个人群。这是因为，在某个给定的匹配标准之下，不是所有治疗组的个体都能成功匹配到一个非治疗组的个体。因而，我们得到的效应估计所对应的人群，是根据倾向性评分匹配成功与否所定义的人群。

而匹配使得研究者只能将数据分析局限于倾向性评分重叠的人群中，这通常被视为匹配的一个优点。没有研究者愿意看到正数性不成立导致的有偏估计。然而，抛开随机变异性不谈，根据倾向性评分从而把某些个体排除在外是有代价的。假设在看到图 15.1 之后，我们就下结论说我们只能在倾向性评分估计值小于 0.67 的人群中估计因果效应。但问题是，这个人群有什么样的特征？我们并不知道。匹配人群的各种特征并不清晰，因而也就难以评估我们的研究结果是否适用于其他人群。

当正数性问题出现的时候，基于现实世界的变量（比如年龄，抽烟频率等）进行人群限制仍然能使我们得到一个较为自然的估计，因为这使我们知道我们的估计能应用于什么特征的人群当中。在我们戒烟的例子中，治疗组那两个 $\pi(L)$ 估计值大于 0.67 的个体是仅有的两个年龄大于 50 岁且烟龄小于 10 年的人。因而，我们可以说匹配后的结果适用于小于 50 岁或烟龄小于 10 年的人。这样的定义会比 $\pi(L) < 0.67$ 更自然、更清晰。

(就算每个人的倾向性评分都是直接出现在他们的脑门上，这个人群特征依然可能并不清楚，因为不同的特征组合依然可能产生同样的倾向性评分。)

利用倾向性评分去检验治疗组和非治疗组的重合区间是很有用的，但是简单地将研究人群限制于这个重合区间从而保证正数性则是不负责任的做法。当我们只使用倾向性评分来选择我们的研究人群时，我们需要认真思考我们的结论是否适用于其他人群。

15.5 倾向性评分模型，结构模型，以及预测模型

在本书第二部分, 我们讲述了两种不同的因果推断模型: 倾向性评分模型和结构模型。现在我们对这两者进行比较。

倾向性评分模型是在给定变量 L 的情况下, 对治疗 A 的概率进行建模, 其中 L 能保证有界互换性成立。在本章, 我们将倾向性评分模型用于匹配和分层, 在第十二章, 我们将其用于逆概率权重, 而在第十四章, 则是用于 G-估算。倾向性评分模型中的参数是冗余参数 (参见精讲点 15.1), 并没有因果性意义, 这是因为变量 L 可能由于许多原因从而与治疗 A 相关, 而不是单纯因为 L 是 A 的诱因。比如, 图 7.1 和 7.2 中, L 和 A 之间的相关性就来自于不同的因素。然而倾向性评分对因果推断却非常有用, 经常被视作结构模型参数估计的基础, 我们在前几章已经论述过。

结构模型描述了治疗 A 和反事实结局 Y^a 的关系, 其可以是边缘性的, 也可以是在 L 的不同分层中。对于连续性治疗而言, 结构模型经常被称为剂量-反应模型。而结构模型中的参数不再是冗余参数, 这是因为它们可以直接被阐述为不同治疗取值 a 下的结局差值。我们讨论过两类结构模型: 边缘结构模型和结构嵌入模型。边缘结构模型可以包含治疗、效应修饰因子 V 、以及治疗和 V 的乘积项。 V 的选择反映了研究者对哪个子群体的效应修饰感兴趣 (参见 12.5 小节)。如果没有包含任何 V , 那么就是一个真正的边缘模型。如果 L 中的所有变量都被认为是效应修饰因子并被包含在模型当中, 那么这个边缘结构模型就会成为一个仿制的边缘结构模型。结构嵌入模型只包含治疗和治疗与效应修饰因子的乘积项。

(精讲点 14.1 讨论了结构嵌入模型和仿制的半参数边缘结构模型之间的关系。)

我们可以把结局回归视作仿制的边缘结构模型, 从而估计因果效应。然而, 结局回归也经常被用于预测, 而非因果推断。比如, 聪明的淘宝店主就会用很复杂的回归模型去预测什么样的顾客更可能会买他们的商品。此时, 他们的目的并不是想知道年龄、性别、收入等因素是否对购买有因果效应, 而是只想知道什么样的人群更愿意购买他们的商品, 从而他们能够更精准地投放广告。此时, 淘宝店主们关心的是相关性, 而非因果性。同样, 医生们会用结局回归开发不同的算法, 从而是预测什么样的病人会有死亡风险。在这些预测模型中, 参数不一定要有因果性意义, 并且此时模型中的所有变量都是同样的地位, 也即不再区分治疗 A 和混杂 L 。

(研究发现微博点赞能够预测性取向和人格特质等信息。而购买哈雷摩托车也能预测智商高低。不过, 这些都是预测性的, 而非因果性的。)

191 结局回归在因果推断和预测性研究中都可以使用, 从而引起了许多误解。其中最大的一个误解是关于模型中的参数选择。当我们只想预测结局的时候, 研究者需要在模型中放入任何能提高模型预测能力的变量。许多变量选择方法都能提高模型的预测能力, 这些方法包括前向选择、后

向消元、逐步筛选、以及最近机器学习中发展出来的种种新方法。对于只关心预测的研究者而言, 这些方法非常有用, 尤其是在面对高维数据的时候。

然而遗憾的是, 许多统计课程及相关教科书并没有严格区分预测和因果推断。因而, 这些变量选择方法被大量应用于因果推断模型之中。这可能会导致我们在倾向性评分模型和结构模型中放入冗余、甚至是有害的其他变量。具体而言, 将预测模型的算法应用到因果推断模型之中可能会导致方差被夸大。

这一问题主要源于一个常见、但却是错误的想法: 倾向性评分模型应该尽可能地预测我们的治疗概率 A 。然而实际上, 倾向性评分模型并不需要非常精确地预测治疗 A , 它只需要包括能满足互换性的变量 L 就可以了。与治疗强烈相关、但对互换性无益的其他变量, 并不能帮助我们减少偏移。如果这些变量也被包含在模型当中, 那会使得我们的误差变大。

(在现实中, 许多使用倾向性评分的研究者都会报告一个预测能力的度量参数: $Mallow's Cp$ 。然而, 这个参数和因果推断之间可能并没有什么关系。)

我们可以思考以下例子。假设某研究的所有参与人员都来自于甲医院或乙医院。甲医院 99% 的参与人员都接受治疗, 即 $A = 1$; 乙医院 99% 的参与人员都没有接受治疗, 即 $A = 0$ 。而医院这个变量其实对结局应没有因果效应 (除了通过治疗 A 体现的效应), 因而互换性并不需要医院这个变量。如果我们在倾向性评分模型中加入了医院这个变量, 那么倾向性评分 $\pi(L)$ 在甲医院中是 0.99, 在乙医院中是 0.01。因而, 在变量 L 的某些分层中, 甲医院只有 $A = 1$ 的研究人群, 而乙医院只有 $A = 0$ 的研究人群。也就是说, 在这些分层中, 我们效应估计的方差近乎无限, 而同时我们并没有降低任何混杂。就算此时我们能够完美地对治疗进行预测, 但这和我们的因果推断一点关系都没有。

(如果我们能完美地预测治疗, 那在治疗组中 $\pi(L) = 1$, 在非治疗组中 $\pi(L) = 0$ 。此时治疗组和非治疗组就没有重叠范围, 因而也就不可能进行接下来的分析。)

除去误差增大, 在因果推断模型中使用预测模型的变量选择算法也会导致更多的偏移。比如, 在模型中加入对撞变量会造成系统性偏移, 而即使我们只是希望提高我们的预测能力。我们将在第十八章再次讨论这一话题。

所有基于模型的因果推断方法——不管是倾向性评分模型还是结构模型——都需要模型设定正确。为了减小模型设定错误的可能性, 我们经常会使用有弹性的设定, 比如放入多次项。此外, 这些因果推断方法也要求互换性、正数性成立, 以及一个良定的治疗。在下一章, 我们会讲述一种非常不同的因果推断方法, 而它不需要互换性成立。

第十五章精讲点和知识点

精讲点 15.1: 冗余参数 (原书第 184 页)

假设我们的目标是估计参数 β_1 和 β_2 。我们需要拟合一个结局回归模型

$E[Y^{a,c=0} | L] = \beta_0 + \beta_1 a + \beta_2 aL + \beta_3 L$ 。而在这个模型中, 只有当 $\beta_0 + \beta_3 L$ 准确表示

$E[Y^{a,c=0} | L]$ 取决于 L 的时候, 我们 β_1 和 β_2 的估计值才是一致的。此时, 我们将 β_0 和 β_3 称为冗余参数, 因为它们并不是我们关心的对象。

另一方面, 如果我们是在结构嵌入模型 $E[Y^{a,c=0} - Y^{a=0,c=0} | L] = \beta_1 a + \beta_2 aL$ 中估计 β_1 和 β_2 , 那只有当 $\Pr[A=1 | L]$ 的模型是正确的时候, β_1 和 β_2 的估计值才是一致的。也就是说, 此时模型 $\text{logit } \Pr[A=1 | L] = \alpha_0 + \alpha_1 L$ 中的参数就是冗余参数。

在一个结局回归中, 如果 L 的正确函数形式应该包含二次项, 即 $\beta_3 L + \beta_4 L^2$, 那么只包含线性项 $\beta_3 L$ 就会导致偏移, 此时结构嵌入模型并不会有这方面的困扰, 因为我们并不需要知道 $L-Y$ 之间的关系。然而, 如果 $L-A$ 的关系设定错误, 那么结构嵌入模型的 G-估算就是有偏的, 不过此时结局回归不会有这方面的困扰。对于不随时间变化的治疗而言, 选择什么方法就被转化为我们能无偏地估计哪些冗余参数。因而, 我们应该尽可能地使用双重稳健方法 (参见精讲点 13.2)。

精讲点 15.2: 效应修饰与倾向性评分 (原书第 190 页)

因为效应修饰的存在, 匹配人群和未匹配人群中的效应估计可能并不一样。比如, 在某个研究中非治疗组的人数远大于治疗组的人数, 那倾向性评分匹配就会排除许多非治疗组人员。因而, 匹配人群中的各变量分布 (包括效应修饰因子的分布) 更接近于治疗组中的分布, 所以匹配人群中的效应估计也就更接近于治疗组的效应估计, 而非整个人群的效应估计。知识点 14.1 讨论了如何使用逆概率加权和标准化估计治疗组中的治疗效应。

倾向性评分的效应修饰作用将有助于决策者做出决定, 比如医生知道哪些病人能最大程度受益于治疗后, 会积极给这些病人进行治疗。然而, 倾向性评分的效应修饰作用也会让我们的效应估计更加复杂。如果存在质的效应修饰作用——比如你发现治疗对倾向性评分在 0.11 至 0.93 之间的人有用, 却对其他人有害——那么此时这一结论就没有什么现实意义, 因为它们脱离了实际的变量 L , 正如我们在正文中所讨论的一样。

最后, 除了效应修饰, 还有其他原因可以使得匹配人群的效应估计和总体的效应估计不一样, 这些原因包括未匹配人群中正数性不成立、在匹配人群中未测混杂更多等等。我们在第四章讨论过, 效应修饰有时也能用不同分层中的残余混杂进行解释。

知识点 15.1: 均衡评分与预后评分 (原书第 186 页)

我们在正文中讨论过, 倾向性评分 $\pi(L)$ 能均衡治疗组和非治疗组之间变量的分布。实际上, 倾向性评分 $\pi(L)$ 是最简单的一种均衡评分。广义而言, 均衡评分 $b(L)$ 是变量 L 一个函数, 使得 $A \perp\!\!\!\perp L|b(L)$ 。Rosenbaum 和 Rubin (1983) 证明了基于变量 L 的互换性和正数性也就是基于均衡分数 $b(L)$ 的互换性和正数性。如果调整 L 就足够了, 那么调整 $b(L)$ 也就足够了。图 15.2 展示了如何在因果图中包含倾向性评分: $\pi(L)$ 可以被视作 L 和 A 之间的一个中介变量, 并且有一个命定的箭头从 L 指向 $\pi(L)$ 。因而, 调整 $\pi(L)$ 就能阻断所有从 A 到 L 的后门路径。

均衡评分 $b(L)$ 的另一个替代形式是预后评分 $s(L)$ 。预后评分 $s(L)$ 也是 L 的一个函数, 且使得 $Y^{a=0} \perp\!\!\!\perp L|s(L)$ 。然而, 调整预后评分的方法需要更强的假设, 并且不能轻易扩展到时异治疗。Hansen (2008) 和 Abadie (2013) 在他们的论文中详细讨论了预后评分。

第十五章图表

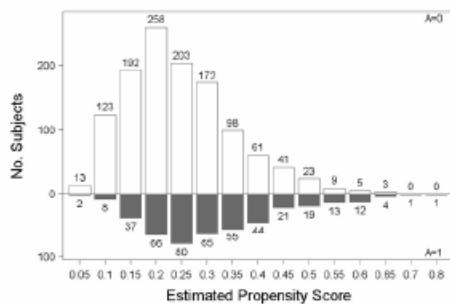


Figure 15.1

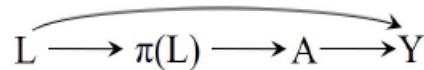


Figure 15.2

第十六章 工具变量

193 迄今, 本书所讨论的因果分析方法都依赖于一个很重要、但又不能验证的假设: 所有用以调整混杂和选择偏移的变量都能被识别, 且已经被完美测量。如果这一假设是不正确的——很大程度上确实是这样——那我们的效应估计就会有残余偏移。

不过, 我们有其他方法, 在不需要测量所有调整因素的情况下仍然能有效地估计因果效应。工具变量是其中一种方法。工具变量在经济学以及其他社会科学中已经被广泛使用。本章我们将讨论如何使用工具变量进行效应估计。

16.1 工具变量的三个条件

图 16.1 描绘了一项双盲随机试验, Z 是被试的分组情况 (1: 治疗组, 0: 安慰剂组), A 是被试的实际接受治疗情况 (1: 实际上接受治疗, 0: 实际上未接受治疗), Y 是结局, U 是所有能影响结局和被试配合程度的未测变量。

如果我们想一致地估计 A 对 Y 的因果效应均值, 那不管使用什么方法, 我们都需要测量 U , 并在我们的模型中调整 U , 这是因为存在一个途经 U 的后门路径: $A \leftarrow U \rightarrow Y$ 。不过 U 是未测的, 所以这些方法只能得到有偏的估计。

但工具变量方法却并不一样, 即使我们不能测量并调整 U , 我们依然可以用这一方法去估计 A 对 Y 的效应均值。首先, 我们需要一个工具变量 Z , 它需要满足以下三个条件:

- (1) Z 和 A 相关;
- (2) Z 仅通过 A 影响 Y , 而不能直接影响 Y ;
- (3) Z 和 Y 没有共同诱因。

知识点 16.1 给出了这三个条件的严格定义。

在这个双盲随机试验中, Z 就是一个工具变量。 Z 满足条件 (1), 因为试验中被分配到治疗组的被试更可能接受治疗; 也可能满足 (2), 因为这是一个双盲设计; 同时也满足 (3), 因为治疗组的分配是随机的, 不受任何其他因素影响。

(在随机试验中, 条件 (2) 不一定成立, 这是因为有时被试需要知晓治疗的副作用。)

在图 16.1 中, 工具变量 Z 对实际治疗 A 有因果效应, 我们将 Z 称作因果性工具变量。某些工具变量不一定对 A 有因果效应。比如, 图 16.2 中的 Z 通过 U_z 和 A 相关, 也满足条件 (1)。

194 此时, U_z 是未测的因果性工具变量, Z 则是已测的工具变量的替代变量。 Z 和 A 有共同诱因 U_z 并不违反条件 (3), 因为此时 U_z 是一个因果性工具变量 (参见知识点 16.1)。图 16.3 描绘了

人群中工具变量的一个替代变量 Z : Z 和未测的因果性工具变量 U_z 有一个共同后果 S , 并且我们控制了 S , 因而此时 Z 和 A 相关。只要满足 16.4 小节提到的几个注意点, 我们就可以用因果性工具变量或其替代变量进行因果效应估计。

在前几章, 我们用不同的方法在观察性数据中估计了戒烟对增重的因果效应。为了使用工具变量估计因果效应, 我们需要一个工具变量 Z 。然而观察性研究并不像随机试验一样有一个表示随机分组的变量, 所以我们需要使用其他变量作为我们的工具变量, 其中一个可能选项是香烟的价格。香烟价格似乎能满足工具变量的三个条件: (1) 香烟价格能影响一个人是否戒烟; (2) 香烟价格只通过戒烟与否对体重产生影响; (3) 香烟价格和增重之间没有共同诱因。精讲点 16.1 讨论了观察性研究中常用的工具变量。

现在我们有了香烟价格 Z 这个变量, 并把它用在我们接下来的讨论当中。假设 $Z=1$ 表示研究参与人员所在州的香烟均价高于 1.5 美元, $Z=0$ 表示其他情况。不过, 我们依然不能判定 Z 是否是一个工具变量。在工具变量的三个条件中, 只有 (1) 是可验证的, 此时我们只需证明 Z 和 A 相关, 也即 $\Pr[A=1|Z=1]-\Pr[A=1|Z=0]>0$ 。 $Z=1$ 时有 25.8% 的人戒烟; $Z=0$ 时有 19.5% 的人戒烟。因而 $\Pr[A=1|Z=1]-\Pr[A=1|Z=0]=6.3\%$ 。在我们的例子中, Z 和 A 微弱相关, 此时 Z 被称为弱工具变量 (16.5 小节讨论了弱工具变量)。

(只要 Z 通过某一未知变量 U_z 和 A 相关, 那就能满足条件 (1)。)

然而, 我们不能验证条件 (2) 和 (3)。为了验证条件 (2), 我们需要证明 Z 只通过 A 影响 Y 。但如果存在 U_z , 就可能存在一条对撞路径 $Z \leftarrow U_z \rightarrow A \leftarrow U \rightarrow Y$, 此时控制 A 就会使得 Z 和 Y 相关, 也就不能用这一方法验证条件 (2)。同样, 我们也不能验证条件 (3), 因为我们没有办法知道效应估计中是否存在混杂。我们只能假设 (2) 和 (3) 成立。因此, 工具变量这一方法和其他方法一样, 需要依赖于一些不可验证的假设。

(有些时候, 我们能够利用数据证伪 (2) 和 (3)。然而, 证伪只是因为假设中的一小部分不成立从而拒绝这一假设。然而即使有再多的数据, 证伪也不具备效力说明假设中的大部分是否不成立。参见 Bonet (2001) 和 Glymour 等人 (2012) 的讨论。)

在观察性研究中, 我们不能证明我们认为的工具变量 Z 是否是真正的工具变量。因为我们不能保证变量之间的结构就如图 16.1 和 16.2 中的一样, 所以我们将 Z 称为候选工具变量。我们能做的, 就是利用各专业知识说明候选变量 Z 为什么能满足条件 (2) 和 (3)。这就如同前几章我们用专业知识为我们的模型假设辩护一样。

不过暂时让我们假设 Z 就是真正的工具变量。那接下来会发生什么? 我们能在不涉及任何混杂的情况下用工具变量一致地估计 A 对 Y 的因果效应吗? 遗憾的是, 如果没有其他假设, 那么这些问题的答案依然是否定的。只有工具变量并不足以让我们估计戒烟 A 对增重 Y 的因果效应, 而只能帮我们估计这一效应的上下限。一般而言, 这一界限非常宽广, 且通常会包含零值 (参见知识点 16.2)。

- 196 在我们的例子中, 这些界限值基本没什么用, 它们能提供的信息我们早已知晓。如果我们不再增添其他不可验证的假设, 那么这就是工具变量所能给我们的信息。16.3 和 16.4 小节讨论了这些额外假设。在此之前, 我们先讲述怎么用工具变量估计因果效应。

16.2 工具变量的效应估计¹

如果 Z 满足工具变量的三个条件 (以及一个额外条件, 我们将在下一小节讨论), 并且是一个二分变量, 那么在加法尺度上, 工具变量的效应估计就可以表示为:

$$E[Y^{a=1}] - E[Y^{a=0}] = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[A|Z=1] - E[A|Z=0]}$$

这是二分工具变量的效应估计表达式 (对二分治疗有 $E[A|Z=1] = \Pr[A=1|Z=1]$)。知识点

- 197 16.3 在一个加成结构模型中给出了这一结果的证明。最好在看完下一小节之后再看这一证明。

为了直观地理解这一表达式, 我们可以再思考一下上一小节中的随机试验。这个表达式中的分子是 Z 对 Y 的效应, 也被称为治疗意向效应; 而分母是 Z 对 A 的效应, 表示被试的配合程度。如果被试完全配合研究人员, 那么分母就等于 1, A 对 Y 的效应就等于 Z 对 Y 的效应。如果配合程度不佳, 那么分母就会趋近于 0, 从而有 A 对 Y 的效应大于 Z 对 Y 的效应。不配合的被试越多, 这两个效应之间的差距也就越大。

(在随机试验中, 工具变量得到的效应估计是两个效应的比值: Z 对 Y 的效应比上 Z 对 A 的效应。因为 Z 是随机分配的, 所以在估计这两个效应时我们不需要调整任何变量。)

这一效应估计不需要调整任何混杂, 而是通过给治疗意向效应乘以一个膨胀系数得到。被试的配合程度越低, 也即 $Z-A$ 之间的关系越接近于 0, 那膨胀系数也就越大。在观察性研究中同理, 其中的分母可以是 Z 对 A 的直接因果效应 (图 16.1), 或者是 Z 和 A 之间的非因果性相关性 (图 16.2 或 16.3)。

¹ 译者注: 在本章, 当我们说“工具变量的效应估计”的时候, 我们指的是用工具变量方法得到的治疗对结局的因果效应估计。

所以, 使用工具变量时, 我们需要对这一表达式分子和分母中的效应进行估计。在我们戒烟的例子中, 工具变量 Z 是一个二分变量 (1: 参与人员所在州的香烟均价高于 1.5 美元, 0: 其他)。分子的估计值是 $\hat{E}[Y|Z=1] - \hat{E}[Y|Z=0] = 2.686 - 2.536 = 0.153$, 而分母的估计值为 $\hat{E}[A|Z=1] - \hat{E}[A|Z=0] = 0.2578 - 0.1951 = 0.0627$, 因而可以得到治疗效应的估计值是 $0.153 / 0.0627 = 2.4 \text{ kg}$ 。在工具变量的三个条件以及第四个附加条件之下, 这就是戒烟对增重的因果效应均值。

(这一方法也被称为 *Wald 估计法*。)

(参见代码 16.1。为了简便, 我们将工具变量缺失的个体排除于数据分析之外。在实践中, 我们可以在使用工具变量之前先用逆概率加权调整选择偏移。)

在上述估计过程中, 我们只是简单计算了四个样本均值: $\hat{E}[Y|Z=1]$, $\hat{E}[Y|Z=0]$, $\hat{E}[A|Z=1]$, 以及 $\hat{E}[A|Z=0]$ 。我们也可以拟合两个(饱和)线性模型从而估计表达式中的分子和分母。对于分母, 需要拟合模型 $E[A|Z] = \alpha_0 + \alpha_1 Z$; 对于分子, 需要拟合模型 $E[Y|Z] = \beta_0 + \beta_1 Z$ 。

我们也可以用双阶最小二乘法拟合线性模型, 从而估计效应。做法如下: 首先拟合第一阶段的模型 $E[A|Z] = \alpha_0 + \alpha_1 Z$, 并给出每个个体的预测值 $\hat{E}[A|Z]$; 然后拟合第二阶段的模型 $E[Y|Z] = \beta_0 + \beta_1 \hat{E}[A|Z]$, 模型中的参数估计 $\hat{\beta}_1$ 就等于使用工具变量得到的效应估计。在我们的例子中, 使用双阶最小二乘法得到的效应估计也是 2.4 kg 。

(参见代码 16.2。)

不过, 这个效应估计所对应的 95% 置信区间却十分宽广: 从 -36.5 到 41.3。而这主要是因为我们的 Z 和 A 弱相关, 所以在第一阶段的模型中存在大量不确定性。在实践中, 如果第一阶段模型的 F 值小于 10, 那么就可以认定这是一个弱工具变量(在我们的例子中, F 值仅为 0.8)。在 16.5 小节, 我们将讨论弱工具变量所带来的问题。

二阶最小二乘法以及其对应的方差使得研究者需要作出很强的参数假设。我们可以使用加成性或乘积性结构模型避开某些假设, 就如知识点 16.3 和 16.4 中所讨论的一样。而结构模型中的某些参数可以通过 G-估算得到。而在工具变量中到底是使用二阶最小二乘法还是使用结构模型, 就类似于在没有工具变量的时候到底是使用结局回归还是结构嵌入模型一样(参见第十四和十五章), 各有利弊。

(参见代码 16.3。)

不过无论如何, 本小节介绍的表达式如果要被认为是治疗 A 对结局 Y 的因果效应均值, 那我们需要工具变量的第四个条件成立。接下来, 我们将讨论第四个条件。

16.3 工具变量的第四个条件: 同质性

16.1 小节中工具变量的三个条件 (1) – (3) 不足以保证工具变量的效应估计就是治疗 A 对 Y 的因果效应。除此之外, 我们需要第四个条件: (4) 效应的同质性。本小节, 我们将讨论 (4) 的不同形式以及它在学术历史中的发展变化。

最极端也是最古老的同质性被表述为: 治疗 A 对结局 Y 的效应对每个人来说都是相等且不变的。在我们的例子中, 如果戒烟让每个人都增加同样的体重 (比如 2.4kg), 那么这个假设就是正确的。因果效应是一常数就等价于我们 14.4 小节所讨论的加成保序性, 而这对我们的结局和治疗来说是不可能的。在我们的例子中, 我们假设的是戒烟会让有的人增重许多, 让有的人增重一点, 而甚至让有的人体重变轻。因此, 我们不会接受这一形式的同质性条件。

(实际上, 在许多早年工具变量研究中, 即使使用了双阶最小二乘, 也很少假设加成保序性。)

条件 (4) 后来被放宽为第二种形式: 对二分变量 Z , 治疗组或非治疗组中 A 对 Y 的效应在 Z 的每一个分层中相等, 即 $E[Y^{a=1} - Y^{a=0} | Z = 1, A = a] = E[Y^{a=1} - Y^{a=0} | Z = 0, A = a]$ 。我们在知识点 16.3 的证明中使用了这一条件。不过这是加法尺度上的同质性。知识点 16.4 讨论了乘法尺度上的同质性。(乘法尺度上的同质性所对应的效应估计表达式不同于之前我们介绍的常见表达式。)

(即使当条件 (3) 成立, 也即 $Y^a \perp\!\!\!\perp Z$ 成立, $Y^a \perp\!\!\!\perp Z|A$ 不一定成立。因此治疗的因果效应大小可能取决于工具变量 Z , 也即同质性的第二种形式也不一定成立。)

上述第二种形式其实也不是非常直观。我们能用什么专业知识去判断治疗 A 的效应在 Z 和 A 的分层中是不变的呢? 因此, 更直观的同质性表述应该包含一个未测混杂 U , 且对 U 的效应修饰作用进行描述 (然而这个假设依然是不可验证的)。在这个更直观的表述中, 也即同质性的第三种形式中, U 不是一个加性效应修饰因子, 也即 A 对 Y 的因果效应在未测混杂 U 的每一分层中都相等, 即 $E[Y^{a=1} | U] - E[Y^{a=0} | U] = E[Y^{a=1}] - E[Y^{a=0}]$ 。然而, 第三种形式的同质性依然不太可信, 因为未测混杂很可能就是效应修饰因子。比如戒烟对增重的影响很可能就因吸烟频率的不同而不同, 同时吸烟频率自身也是一个混杂因素。

(Hernan 和 Robins (2006b) 在论文中讨论了如果 U 是加性效应修饰因子, 那么前两种不同形式的同质性也不成立。)

199 另一种形式的同质性是加法尺度上 Z - A 的相关性在混杂因素 U 的不同分层中不变, 也即 $E[A|Z=1,U] - E[A|Z=0,U] = E[A|Z=1] - E[A|Z=0]$ 。与前三种形式不同的是, 这一形式在极端零假设情况下不一定成立。不过我们可以部分验证这一形式的同质性。对于二分变量 A , 如果某些混杂是已测的, 那我们可以在已测混杂中验证这一形式的同质性。对于连续性变量 A , 如果我们假设了线性关系, 那 A 的方差在 Z 的不同分层中必须相等, 否则这一同质性不成立。

(Wang 和 Tchetgen-Tchetgen (2018) 在他们的论文中提出了第三种和第四种同质性。)

不过, 因为同质性在许多场合中都显得不合理, 所以许多研究者对能否使用工具变量得到有效的效应估计抱持怀疑态度。不过我们还有其他两种方法能让我们避开同质性条件。

其中一种方法是在变量工具的模型中包含研究起始时的变量。如此一来, 我们就能放宽双阶最小二乘法中的参数假设, 也就能更放心地使用工具变量。在模型中包含初始变量会限制治疗在 200 协变量各分层中的变化情况, 同时允许治疗组中的因果效应随 Z 而变化。16.5 小节和知识点 16.5 更详细讨论了结构模型中的协变量。

(在工具变量方法中, 模型有很多用处。我们可以在模型中包含多个工具变量、处理连续性治疗、以及在结局是二分变量的情况下计算风险比。)

另一种方法是放弃同质性, 使用另一个条件 (4)。新的条件 (4) 虽然不足以让我们估计人群中的因果效应均值, 但将会让我们工具变量的效应估计有因果性意义。我们将在下一小节讨论另一种条件 (4)。

16.4 另一种第四个条件: 单调性

让我们再回到最开始的双盲随机试验, 其中有表示治疗分组的变量 Z 、表示实际治疗情况的变量 A 、以及表示结局的变量 Y 。对于试验中的每一名被试, 我们将反事实变量 $A^{z=1}$ 定义为这名被试被分配到治疗组 ($z=1$) 时他的实际治疗情况。 $A^{z=0}$ 同理。

如果我们知道每个被试的 $A^{z=1}$ 和 $A^{z=0}$, 那我们就可以将所有被试分为互不重叠的四组:

1. 都会接受治疗。这组中的被试不管有没有被分配到治疗组, 都会接受治疗, 即

201 $A^{z=1} = A^{z=0} = 1$ 。

2. 都不会接受治疗。这组中的被试不管有没有被分配到治疗组, 都不会接受治疗, 即

$A^{z=1} = A^{z=0} = 0$ 。

3. 配合者。这组中的被试实际治疗情况总会和他们的分组相同, 即 $A^{z=1} = 1$ 且 $A^{z=0} = 0$ 。

4. 对抗者。这组中的被试实际治疗情况总会和他们的分组相反, 即 $A^{z=1} = 0$ 且 $A^{z=0} = 1$ 。

这种分组方式也被称为配合情形分组。然而一般而言, 我们在现实中并不能识别这四组。比如, 如果我们观察到一名被试 $Z = 1$ 且 $A = 1$, 我们并不知道他是配合者还是总会接受治疗者。

如果不存在对抗者, 那单调性成立, 这是因为工具变量 Z 的取值变大, 要么未改变 A 的值 (也即总是接受治疗或总是不接受治疗, 如图 16.4 和 16.5 所示), 要么也导致 A 的取值变大 (配合者, 如图 16.6 所示)。只有对对抗者来说, Z 取值变大会导致 A 取值变小 (如图 16.7 所示)。或者换句话说, 如果对所有人有 $A^{z=1} \geq A^{z=0}$, 那么单调性成立。

现在我们将上一小节中的同质性替换为单调性, 让单调性成为我们的第四个条件。那么工具变量的效应估计就不再是 $E[Y^{a=1}] - E[Y^{a=0}]$ 。在单调性下, 工具变量的效应估计等于配合者中 202 的因果效应均值, 也即:

$$E[Y^{a=1} - Y^{a=0} | A^{z=1} = 1, A^{z=0} = 0]$$

知识点 16.6 对这一相等关系给出了证明。简单而言, 工具变量的效应估计的分子, 也即 Z 对 Y 的因果效应, 就等于本小节四个不同分组中 Z 的效应的加权平均。然而, 在都会接受治疗与都不会接受治疗这两组中, Z 对 Y 的效应为 0, 这是因为 Z 对 Y 的因果效应需要通过 A , 而 A 的取值在这两组中是固定的。同时, 在单调性条件下, 不存在对抗者。因此, 工具变量的效应估计的分子, 就是配合者中 Z 对 Y 的因果效应——也就是配合者中 A 对 Y 的因果效应——再乘以配合者所占的比例。而配合者所占的比例, 就等价于工具变量的效应估计中的分母。

(配合者中的因果效应均值是一个子群体中的因果效应, 可能与整个人群的因果效应不尽相同。在有的文献中, 配合者也被称为合作者。)

在观察性研究中, 如果不存在对抗者, 我们就可以通过工具变量估计配合者中的治疗效应。然而严格而言, 在观察性研究中不存在配合者或对抗者, 这是因为观察性研究中没有治疗分配, 也就不存在配合或不配合。在我们戒烟的例子中, 配合者指的是住在香烟价格高的州就戒烟、而住在香烟价格低的州就不戒烟的人。与之相反, 对抗者指的是住在香烟价格高的州就不戒烟、而住在香烟价格低的州就戒烟的人。如果不存在对抗者且因果性工具变量是二分变量 (参见知识点 16.6), 那么 2.4kg 就是配合者中的效应估计。

在 1990 年代, 用单调性替代同质性被视为工具变量方法的救星。主要是因为同质性在大多数情形中并不可信, 而单调性则更可信。不过, 单调性假设下的工具变量不能估计整个人群的因果

效应, 而只能估计配合者中的因果效应, 这似乎是单调性的一个代价。然而就算如此, 仍然有许多研究者从其他不同角度批评单调性。

(Deaton (2010) 在论文中这样评价配合者中的效应: “这偏离了我们的初衷。原本光线很强, 能照耀各个角落。但现在我们控制了光线能照射的地方, 然后宣称这就是我们一直以来希望看到的东西。”)

首先, 配合者中的因果效应到底和我们的研究目的有什么关系。一般而言, 哪些人是配合者, 这是不可识别的。即使我们能计算出配合者在整个人群中所占的比例 (这一比例是工具变量效应估计的分母, 参见知识点 16.6), 这一比例也因工具变量的不同而不同, 在不同研究中也不同。因此, 我们的结论并不能给决策者带来什么信息。假设在我们的研究中只有 6% 的人是配合者, 可能在现实人群中会更多一点。我们的结论告诉人们, 治疗 $A=1$ 在配合者中是有益的。那基于只适用于配合者的结论, 决策者就要向所有人群推荐治疗 $A=1$ 吗? 如果治疗在都会接受和都不会接受的人中没有用怎么办? 因而, 我们也许需要承认, 这一结论可能与实践没有什么关系。我们估计了这一效应只是因为这一效应更方便计算, 而并非更有用。

(不过, 令人稍感安慰的是, 在很强的额外假设之下, 我们可以描述配合者的人群特征, 从而将结论推广到更大的人群。参见 Angrist, Pischke, Baiocchi 等人的论文。)

其次, 在观察性研究者, 单调性假设也并非总是成立。在随机试验中, 每一个被试都会签署 203 知情同意书, 因而不太可能会出现事事都与研究者对着干的被试。同时, 随机试验也能保证被分配到非治疗组的被试不会接受治疗, 从而保证了单调性。然而, 在观察性研究中, 我们很难保证某些工具变量的单调性。比如在某一诊所中有两名风格完全不同的医生, 其中一位偏爱直接给予治疗, 除非病人患有糖尿病。另外一位则一般不会给予治疗, 除非病人精神状况良好。既有糖尿病又精神状态良好的病人, 其实际治疗状况可能和两位医生的偏好都相反, 因此这位病人可以被视为一名对抗者。也就是说, 如果我们从多个维度考察治疗状态时, 单调性就不太可能成立。在这一情形中, 对抗者的比例就不能被忽略。

(这一例子由 Swanson 和 Hernan 给出, 同时 Swanson 在观察性研究中实证地证明了对抗者的存在。)

如果是图 16.2 和 16.3 中一样的 Z , 也即工具变量的替代变量, 那么情况会更加复杂。如果实际的因果性工具变量 U_z 是连续的, 而 Z 是二分的, 那由 Z 得到的效应估计并不是配合者中的因果效应。这一效应估计将是人群中的某个加权均值, 因而我们难以阐释这一估计的意义。因此, 当我们使用的二分工具变量并不是因果性工具变量的时候, 即使单调性成立, 我们依然不能给我们的效应估计一个合理的解释。

(连续性工具变量 U_Z 单调性的定义: A^{u_Z} 是 u_Z 的一个非减函数。)

最后, 将人群根据配合程度分成四组这一做法可能并不合理。在许多现实情景中, 配合者是劣定的。比如, 在研究中, 我们也许会假设工具变量是“医生偏好”, 并假设有同样偏好的医生会用同样的方式治疗病人。然而这一假设并不现实且不可能成立, 其中一个原因是我们并不知道看病的医生是什么样的偏好。将病人依据配合程度分为四组, 需要一个命定的反事实情形(我们估计效应均值的时候并不需要这一假设)、不存在个体间的干扰、不存在治疗的多种形式以及其他形式的异质性(比如, 在某一工具变量下是配合者, 在另一工具变量下又不是配合者)。

(Swanson (2015) 在论文中讨论了定义单调性的困难, 同时在观察性研究中引入了全局单调性和局部单调性。)

总而言之, 如果我们只关心配合者中的因果效应, 那在只涉及两个分组的双盲随机试验中, 单调性就是一个合适的假设。然而, 当情形更加复杂或涉及观察性研究, 即使我们的工具变量是真正的因果性工具变量, 我们也需要倍加小心。

(Sommer 和 Zeger (1991), Imbens 和 Rubin (1997), 以及 Greenland (2000) 等人在论文中讨论了对照组的完全配合情形。)

204 16.5 再谈工具变量的三个条件

前两个小节我们讨论了同质性或单调性作为工具变量第四个条件的优劣。我们的讨论都假设了 Z 是一个有效的工具变量。然而在观察性研究中, Z 可能并不满足工具变量三个条件中的(2) 和 (3), 因而就不是一个有效的工具变量; 也可能 Z 仅仅勉强满足 (1), 那么 Z 就只是一个弱工具变量。在这两种情况中, 就算条件 (4) 完美成立, 使用工具变量也会导致很强的偏移。现在, 我们再详细讨论一下每个条件。

条件 (1): Z 和 A 相关。这一条件可以实证地验证。研究者在使用工具变量之前, 需要先验证 Z 和 A 是否相关。然而, 就像我们戒烟例子中一样, 如果 Z 和 A 之间的相关性很弱, 那就是一个弱工具变量(参见精讲点 16.2), 这可能带来三个严重的问题。

第一, 弱工具变量会导致 95% 置信区间变宽。第二, 弱工具变量会放大因条件 (2) 和 (3) 不成立带来的偏移。 Z 和 A 的弱相关性会使得工具变量效应估计中的分母变小, 而条件 (2) 和 (3) 影响的是分子, 如果分子有偏移, 那么这一偏移将会被放大。在我们戒烟的例子中, 任何分子中的偏移都会被乘以 15.9 ($1/0.0627$)。第三, 即使样本够大, 弱工具变量也会带来偏移, 并导致效应估计方差的低估。也即, 效应估计是错误的, 且置信区间太窄了。

(在线性模型中, 如果存在强混杂, 那么工具变量就会变弱, 这是因为强 $A-U$ 关系会使得强 $A-U_z$ 或 $A-Z$ 关系的变化减弱。)

(*Bound, Jaeger 和 Baker (1995)* 在他们的论文中详细讨论了弱工具变量带来的偏移。许多使用弱工具变量的论文大量引用了这篇文章。)

为了理解第三个问题, 也即弱工具变量带来的偏移, 我们可以思考一个随机生成的二分变量 Z 。在一个无限人群中, 治疗 A 和 Z 没有任何相关性, 因而使用 Z 作为工具变量得到的效应估计是无意义的。然而, 在无限人群中, 有小概率会使得 Z 和某未测混杂 U 产生相关性, 因而 Z 和 A 之间也就存在相关性——即使很弱, 但也不为零。如果我们把 Z 作为工具变量, 效应估计的分母就会非常小, 因而分子中的值就会被不当地扩大, 产生更大的偏移。实际上, 我们例子中的工具变量“香烟均价高于 1.5 美元”就像一个随机生成的变量。如果我们用 1.6 美元、1.7 美元、1.8 美元、以及 1.9 美元, 那我们得到的估计值会分别是 41.3、-40.9、-21.1、以及-12.8kg。在每一种情形中, 95%置信区间都很宽广, 但依然低估了真实的不确定性。因为弱工具变量可能会造成很大的偏移, 所以一个稍微违反条件 (2) 和 (3) 的强工具变量, 依然优于一个满足 (2) 和 (3) 的弱工具变量。

(参见代码 16.4。)

条件 (2) : Z 仅通过 A 影响 Y , 而不能直接影响 Y 。在因果图中, 如果有箭头从 Z 指向 Y , 那就违反了条件 (2), 就如图 16.8 所示。这一箭头并不经过治疗 A , 因而将会直接作用于工具变量效应估计的分子, 而这一额外部分也会被视为 A 的效应的一部分, 从而被分母扩大。如果连续性的或者有多个取值的治疗 A 被不那么精确的 A^* (比如一个二分变量) 取代, 那么条件 (2) 可能就不成立。在图 16.9 中, 条件 (2) 对原变量 A 成立, 但对于 A^* 并不成立, 这是因为路径 $Z \rightarrow A \rightarrow Y$ 所表示的 Z 的效应并没有经过 A^* , 而我们估计的却是 A^* 的效应。在实践中, 为了简便, 很多时候我们只能用近似的 A^* 替代真实的 A 。这种近似替代是工具变量的一个主要问题, 但在前几章的方法中却不构成大问题。

条件 (3) : Z 和 Y 没有共同诱因。这一条件同样无法验证。图 16.10 描绘了 Z 和 Y 存在共同诱因的情形, 其中 U_1 不仅是 Z 和 Y 的共同诱因, 也是 A 的诱因。在观察性研究中, Z 的混杂总是存在 (对于其他研究者不能控制的变量也同理)。而混杂会影响效应估计里面的分子, 同时也会被视为 A 的效应的一部分, 从而被分母扩大。

某些时候, 条件 (3) 和其他条件在某些变量的分层中更可能成立。相比于直接假设 Z 和 Y 之间没有混杂, 加上“在变量 V 的分层中”这一限制可能会更好一些, 也即假设“在某些变量 V 的分层中, Z 和 Y 之间没有混杂”。从而我们就可以在 V 的分层中利用工具变量估计因果效应, 然

后再假设治疗的因果效应在整个人群（同质性）或者配合者（单调性）中是不变的，进而汇总这些分层中的效应估计。另一种方法是，我们可以在双阶模型中放入变量 V 。在我们的例子中，这一方法减小了我们的估计值大小，同时增加了 95% 置信区间的宽度。

（参见代码 16.5。）

研究者也经常检验工具变量 Z 的不同分层中已测混杂的分布，从而为条件（3）提供支撑。
这一做法是基于这样一种想法：如果已测的变量已经分布均衡了，那未测变量同样分布均衡的可能性会高一些。然而，这一想法可能会造成致命的错误，这是因为即使再小的不均衡，经过（前文讨论的）放大之后，也会造成很大的偏移。
206

即使 Z 和 Y 之间没有混杂，条件（3）依然可能不成立。条件（3）的正式定义需要工具变量不同分层中的个体可互换。这一互换性可能因为其他混杂或者选择偏移而不成立。在使用工具变量时，产生选择偏移的一种常见原因是把治疗 A 某个取值的所有人排除在分析之外。比如，治疗 A 的可能取值是 0、1、2，但工具变量仅将取值为 1 或 2 的个体包含在分析之中。然而这一做法在其他方法中并不会造成选择偏移。

（Swanson (2015b) 在论文中详细讨论了这一选择偏移。）

有些研究者会同时使用多个工具变量，从而缓和只有一个工具变量的不足。然而使用多个工具变量会加剧我们上述讨论的种种问题。工具变量的数目越多，它们中的某些也就越可能违反工具变量的三个基本条件。

16.6 工具变量与其他方法比较

工具变量和我们前几章讨论的方法至少在三个方面不同。

第一，就算数据是无限的，工具变量也需要模型假设，而此时逆概率加权与标准化不需要。如果我们有超级人群中每个人的治疗、结局、以及混杂数据，我们就能像本书第一部分那样直接用非参数化的逆概率加权和标准化两种方法估计治疗的因果效应均值。而在工具变量中，我们仍然需要模型去估计治疗的因果效应。在数学上，同质性等价于将结构模型中的所有乘积项参数设定为 0（参见知识点 16.1）。也就是说，工具变量方法不存非参数的形式。

（工具变量不是唯一没有非参数形式的方法，断点回归分析等方法也没有非参数形式。）

第二，稍微违反条件（1）至（4）会造成不可预测的极大偏移。工具变量的一个理论基础是效应估计中的分母会扩大分子中的效应。因此，当这些条件不成立的时候，或者是一个弱工具变量的时候，就可能导致方向不可预测的极大偏移。因此，工具变量的效应估计可能比未调整的效应估计偏差更大。与之相比，可识别条件稍微不成立的时候，前几章介绍的方法只会造成轻微的

偏移, 而调整混杂或偏移并不会引入更多的偏移。工具变量的效应估计对假设条件的变化非常敏感, 这一特质使得工具变量对于圈外人来说是一种非常危险的方法, 同时也再一次强调了敏感性分析的重要性。此外, 专业知识更有助于我们思考 A 和 Y 之间的可能混杂、以及这些混杂如何影响我们的结果, 而非 Z 和 Y 之间的混杂、以及存在对抗者或者效应的异质性会如何影响我们的效应估计。

(*Baiocchi 和 Small (2014)* 在他们的论文中讨论了量化工具变量敏感性的方法。)

207

第三, 能使用工具变量的理想情形相较于其他方法来说更不常见。我们讨论过, 工具变量主要用于有大量未测混杂、有一个二分且时间固定的治疗 A 、以及有一个很强的因果性工具变量 Z 的情形之中, 同时还需要同质性或单调性成立。这些限制使得工具变量只能用于简单的因果推断情形当中, 比如比较 $A=1$ 与 $A=0$ 的效应。本书第三部分主要涉及时异治疗, 以及多时段复杂治疗策略的比较, 因而工具变量不再适用。

因果推断依赖于假设的明确性, 以及对各种假设的详细剖析。工具变量所需的假设和其他方法不同, 因而成为一种吸引人的新方法。然而, 因为工具变量效应估计的 95% 置信区间都太过宽广, 所以这一方法带来的实际价值就显得微不足道。同时, 使用工具变量的时候, 我们需要对它的种种限制保持警惕——虽然这一提醒对所有因果推断方法都适用, 但是工具变量不可预测的较大偏差需要我们的格外注意。

(透明性要求我们恰当地描述我们工具变量的分析方法。*Brookhart 等学者的论文 (2010)* 给出描述工具变量的指导参考。)

第十六章精讲点和知识点

精讲点 16.1: 观察性研究中的常用工具变量 (原书第 195 页)

观察性研究也可以有许多工具变量。我们不可能对所有工具变量都一一讲述。接下来, 我们将讲述常见的三个工具变量。

- 基因。这一工具变量会假设基因只通过某些生理特征从而影响疾病结局。比如, 在酒精与冠心病的研究中, 基因 Z 和酒精代谢 A (比如 ALDH2 酶) 有关, 而 A 又与冠心病 Y 有关, 并假设 Z 只通过 A 影响 Y 。在观测性研究中, 利用基因作为工具变量是“孟德尔随机化”框架下的重要一部分。
- 医生偏好。这一工具变量指医生对治疗的偏好, 而这一偏好并不直接影响疾病结局。比如, 当我们想比较选择性 COX-2 抑制剂和非选择性甾体抗炎药时, U_z 可以表示医生对这两种

药物的偏好。因为 U_z 通常是未知的，所以研究人员一般会用某个已测的变量 Z 进行替代，比如“医生上一次开的治疗”。

- 医疗资源。这一工具变量的想法是治疗资源会影响病人是否接受了治疗，同时治疗资源又不会直接影响疾病结局。比如，研究人员可以用距离医院的距离作为工具变量。在我们的例子中，我们用的是价格，也是治疗资源的一种体现。

精讲点 16.2: 弱工具变量 (原书第 204 页)

在研究文献中，弱工具变量有两种不同但是相关的定义：

1. 如果 $Z-A$ 的相关性——也即工具变量效应估计中的分母——很小，那这就是一个弱工具变量。
2. 如果 $Z-A$ 相关性的 F 值很小（一般是小于 10），那这就是一个弱工具变量。

在我们戒烟的例子中，我们的工具变量满足这两个定义：效应估计的分母是 6%，而 F 值是 0.8。

第一个定义基于 $Z-A$ 相关性的真值，体现出即使我们有无限的数据，弱工具变量依然会放大分子中的偏差。第二个定义低于 $Z-A$ 相关性的统计特性，体现出即使我们有一个完美的工具变量，我们的估计值在无限数据中依然可能是有偏的（这是我们正文中讨论的工具变量的第三个问题）。

知识点 16.1: 工具变量条件的正式定义 (原书第 194 页)

工具变量的条件 (1) 有时也被称为相关性条件，也即 $Z \perp\!\!\!\perp A$ 不成立。

条件 (2) 经常被称为排他性限制，指“没有从 Z 到 Y 的直接效应”。在个体层面上，条件 (2) 可以表示为 $Y_i^{z,a} = Y_i^{z',a} = Y_i^a$ 。然而对于本章的讨论，我们只需要人群层面的条件 (2) 成立即可，也即 $E[Y^{z,a}] = E[Y^{z',a}]$ 。对工具变量的替代变量而言，哪种形式为真并不重要。

条件 (3) 可以表示为边缘互换性： $Y^{a,z} \perp\!\!\!\perp Z$ ，在图 16.1、16.2、以及 16.3 所对应的单一世界干涉图中，这一条件为真。结合个体层面的条件 (2)，我们可以推出 $Y^a \perp\!\!\!\perp Z$ 。另一种较强形式的条件 (3) 是联合互换性，即对于二分治疗和工具变量而言，有 $\{Y^{z,a}; a \in [0,1], z \in [0,1]\} \perp\!\!\!\perp Z$ 。知识点 2.1 讨论了不同形式的互换性，知识点 16.2 不同形式互换性下的不同结果。因为随机试验中的工具变量 Z 是随机分配的，所以条件 (3) 的两种形式在随机试验中都为真。

知识点 16.2: 效应的界限——可部分识别的因果效应 (原书第 196 页)

对一个二分结局 Y , 治疗的因果效应 $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$ 的取值范围是 -1 至 1。这一效应的界限是 $(-1, 1)$ 。因为我们知道每个人 $Y^{a=1}$ 或 $Y^{a=0}$ 中的一个, 所以我们可以用已知的数据去缩短界限的范围, 但是零值还是在这个范围之中。对于连续性的结局, 我们需要知道结局的最大值和最小值, 才能确定界限的范围。

如果存在一个工具变量, 能满足个体层面的条件 (2) 以及边缘互换性的条件 (3), 那么我们就行进一步缩短 $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$ 的界限。此时的界限宽度可以表示为

$\Pr[A = 1 | Z = 0] + \Pr[A = 0 | Z = 1]$, 并被称为自然界限。如果用联合互换性替代边缘互换性, 那么这一界限可以进一步缩短, 这称称为极端界限。

我们也可以用联合干预 z 和 a 的单一世界干涉图来表示极端界限所需的条件。研究者证明了 $Y^{a,z} \perp\!\!\!\perp (Z, A) | U$, $Z \perp\!\!\!\perp U$, 以及 $E[Y^{z,a} | U] = E[Y^{z',a} | U]$ 成立的情况下就足以得到极端界限。具体而言, 这些条件可以推出 $Y \perp\!\!\!\perp Z | U, A$ 以及 $E[Y^{z,a}] = \int E[Y | A = a, U = u] dF(u)$ (因为在 U 的分层中 Z 对 Y 并没有直接效应, 所以这一等式略去了 Z)。通过这一等式我们即可得到极端界限。在更进一步的假设之下, 我们可以得到更窄的界限, 详情参考 Richardson, Evan 以及 Robins 在 2011 年所著论文。

确定界限也通常被称为部分识别因果效应。遗憾的是, 这些界限通常给不了太多的信息, 因为一般而言这些界限都很宽广。Swanson 在论文中 (2015c) 讲述了现实世界中这些界限识别方法的一些应用, 并讨论了它们的优缺点。

还有一种方法可以缩短界限的宽度, 也即对 A 和 Y 之间的关系做出一些参数假设。在 16.2 小节的足够强假设下, 这个上下界限会收敛到一个数, 也就是因果效应的点估计。

知识点 16.3: 加成结构模型和工具变量 (原书第 199 页)

下述式子是图 16.1、16.2、以及 16.3 所对应的饱和加成结构模型 (治疗是二分的) :

$$E[Y^{a=1} - Y^{a=0} | A = 1, Z] = \beta_0 + \beta_1 Z$$

这一模型可以写作 $E[Y - Y^{a=0} | A, Z] = A(\beta_0 + \beta_1 Z)$ 。其中参数 β_0 是 $A = 1$ 且 $Z = 0$ 时治疗的效应均值, $\beta_0 + \beta_1$ 是 $A = 1$ 且 $Z = 1$ 时治疗的效应均值。因而, β_1 量化了 Z 在加法尺度上的效应修饰作用。

如果我们先验地假设 Z 没有效应修饰作用, 那么 $\beta_1 = 0$ 。 β_0 就是工具变量的效应估计。也就是说, 工具变量的效应估计, 就是 Z 不存在效应修饰时结构模型中治疗所对应的参数。

这一证明很简单。当 Z 是工具变量的时候, 条件 (2) 成立, 因此有

$$E[Y^{a=0} | Z=1] = E[Y^{a=0} | Z=0]。而条件均值的独立性可以被写作$$

$$E[Y - A(\beta_0 + \beta_1) | Z=1] = E[Y - A\beta_0 | Z=0]。 \beta_1 = 0 \text{ 时, 就有}$$

$$\beta_0 = \frac{E[Y | Z=1] - E[Y | Z=0]}{E[A | Z=1] - E[A | Z=0]}$$

可能有人会问为什么我们需要先验地假设 $\beta_1 = 0$ 。这是因为方程中有两个未知数 (β_0 和 β_1) , 因此我们需要对其中一个进行假设从而解出这个方程, 于是我们选择了 $\beta_1 = 0$ (而不是 $\beta_1 = 2$)。这就是为什么我们会说工具变量本身不足以估计治疗的因果效应均值。

然而, 为了得到上述结论, 我们必须假设治疗的效应在接受了治疗的人群和未接受治疗的人群中都是一样的, 而这是一个不可验证的假设。

知识点 16.4: 乘积结构模型和工具变量 (原书第 200 页)

考虑下述饱和乘积结构模型:

$$\frac{E[Y^{a=1} | A=1, Z]}{E[Y^{a=0} | A=1, Z]} = \exp(\beta_0 + \beta_1 Z)$$

这也可以被写作 $E[Y | A, Z] = E[Y^{a=0} | A, Z] \exp[A(\beta_0 + \beta_1 Z)]$ 。对于二分结局 Y , $\exp(\beta_0)$ 是 $A=1$ 且 $Z=0$ 时的因果性风险比, $\exp(\beta_0 + \beta_1)$ 是 $A=1$ 且 $Z=1$ 时的因果性风险比。因而, β_1 量化了 Z 在乘法尺度上的效应修饰作用。如果我们先验地假设 $\beta_1 = 0$, 那么乘法尺度上的因果效应就是 $E[Y^{a=1}] / E[Y^{a=0}] = \exp(\beta_0)$, 而在加法尺度上则是

$E[Y^{a=1}] - E[Y^{a=0}] = E[Y | A=0](1 - E[A])[\exp(\beta_0) - 1] + E[Y | A=1]E[A][1 - \exp(-\beta_0)]$ 。这一证明将会基于工具变量的三个基本条件, 感兴趣的读者可以参考 Robins 在 1989 年的论文。

也就是说, 如果我们假设 Z 没有效应修饰作用, 那么 $E[Y^{a=1}] - E[Y^{a=0}]$ 就是可识别的, 但不再等于工具变量的效应估计。因此, 我们的估计就将取决于是否假设 Z 在乘法或加法尺度上有效应修饰作用。遗憾的是, 即使有无限的数据, 我们也不能验证哪一个假设为真, 这是因为当我

们使用饱和加成或乘积结构模型的时候, 我们的未知数与方程数多。这也是为什么我们需要同质性假设的原因。

知识点 16.5: 广义结构模型 (原书第 201 页)

下述加成结构模型可以包含连续性治疗 A 、工具变量 Z 、以及协变量 V : (上述变量均可以是多个变量)

$$E[Y - Y^{a=0} | Z, A, Z] = \gamma(Z, A, V; \psi)$$

$\gamma(Z, A, V; \psi)$ 表示某个已知的函数形式。 ψ 是未知参数, 并且 $\gamma(Z, A=0, V; \psi) = 0$ 。也就是说, 这一模型表示治疗取值为 A 与取值为 0 时相比而得到的因果效应均值, 并且这一效应仅适用于工具变量取值为 Z 、协变量取值为 V 、同时治疗取值为 A 的人群中。我们可以在条件均值独立假设之下, 也即 $E[Y^{a=0} | Z=1, V] = E[Y^{a=0} | Z=0, V]$ 时, 利用 G-估算确定这个模型中的未知参数。

同理, 类似的乘积结构模型可以表示为:

$$E[Y | Z, A, Z] = E[Y^{a=0} | Z, A, Z] \exp[\gamma(Z, A, V; \psi)]$$

其中各参数、函数的定义与加成结构模型中的相同, 我们也可以在相似假设下利用 G-估算确定未知参数。因而, 我们可以使用 G-估算, 将工具变量推广到涉及时异治疗和时异变量的情形。

知识点 16.6: 单调性与配合者中的效应 (原书第 208 页)

考虑一个二分的因果性工具变量 Z 和治疗 A 。前人已经证明了在单调性假设下 (也即不存在对抗者), 工具变量的效应估计等于配合者中的因果效应均值 $E[Y^{a=1} - Y^{a=0} | A^{z=1} - A^{z=0} = 1]$ 。以下我们给出简易证明。

治疗意向效应可以写作四个配合分组的加权均值:

$$\begin{aligned} E[Y^{z=1} - Y^{z=0}] &= E[Y^{z=1} - Y^{z=0} | A^{z=1} = 1, A^{z=0} = 1] \Pr[A^{z=1} = 1, A^{z=0} = 1] \\ &\quad + E[Y^{z=1} - Y^{z=0} | A^{z=1} = 0, A^{z=0} = 0] \Pr[A^{z=1} = 0, A^{z=0} = 0] \\ &\quad + E[Y^{z=1} - Y^{z=0} | A^{z=1} = 1, A^{z=0} = 0] \Pr[A^{z=1} = 1, A^{z=0} = 0] \\ &\quad + E[Y^{z=1} - Y^{z=0} | A^{z=1} = 0, A^{z=0} = 1] \Pr[A^{z=1} = 0, A^{z=0} = 1] \end{aligned}$$

然而, 根据我们分组的定义以及工具变量的条件 (2), 治疗意向的效应在都会接受以及都不接受治疗的人群中是零。如果不存在对抗者, 那上述等式就可以简化为:

$$E[Y^{z=1} - Y^{z=0}] = E[Y^{z=1} - Y^{z=0} | A^{z=1} = 1, A^{z=0} = 0] \Pr[A^{z=1} = 1, A^{z=0} = 0]$$

但是在配合者中, Z 对 Y 的效应就等于 A 对 Y 的效应 (因为 $Z = A$), 所以配合者中的治疗效应将会是:

$$E[Y^{z=1} - Y^{z=0} | A^{z=1} = 1, A^{z=0} = 0] = \frac{E[Y^{z=1} - Y^{z=0}]}{\Pr[A^{z=1} = 1, A^{z=0} = 0]}$$

如果 Z 是随机分配的, 也即 $\{Y^{z,a}, A^z; a \in [0,1], z \in [0,1]\} \perp\!\!\!\perp Z$, 这一等式就是工具变量的效应估计。在联合独立以及一致性假设下, 治疗意向效应 $E[Y^{z=1} - Y^{z=0}]$ ——也就是分子——等于 $E[Y|Z=1] - E[Y|Z=0]$; 而配合者的人数占比——也就是分母——等于 $\Pr[A=1|Z=1] - \Pr[A=1|Z=0]$ 。后一等式成立时因为在单调性假设下, 不存在对抗者。所以配合者的比例就等于 1 减去都会接受治疗和都不会接受治疗所占的比例。而都接受治疗的人数占比 $\Pr[A^{z=0} = 1] = \Pr[A=1|Z=0]$, 都不接受治疗的人数占比 $\Pr[A^{z=1} = 1] = \Pr[A=0|Z=1]$ 。

这一简略证明只适用于 Z 是因果性工具变量的简单情形, 也即图 16.1 中的情形。在图 16.2 和 16.3 中, Z 不是因果性工具变量, 研究者也证明了此时配合者 (根据 U_Z 进行定义) 中的因果效应均值等于工具变量的效应估计。他们的证明需要两个假设: 在控制了 U_Z 的时候, Z 独立于 A 和 Y , 以及 U_Z 是一个二分变量。然而, 如果 U_Z 是连续的, 其中的独立性假设就有任何可信度。这就使得工具变量在观察性研究中的应用成为问题。

第十六章图表

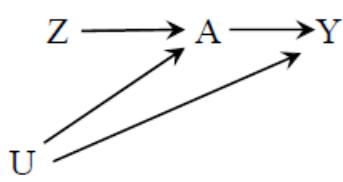


Figure 16.1

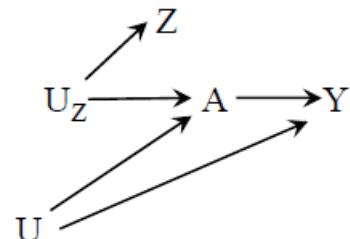


Figure 16.2

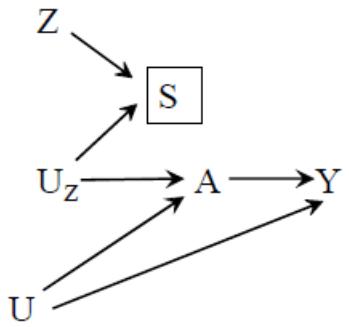


Figure 16.3

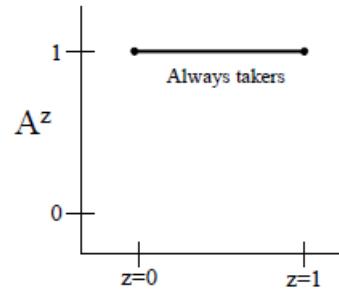


Figure 16.4

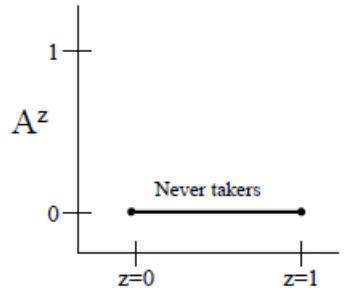


Figure 16.5

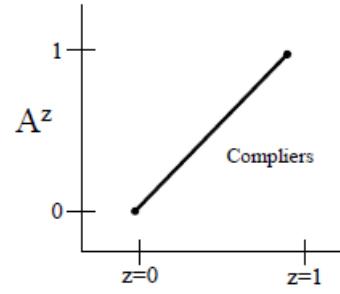


Figure 16.6

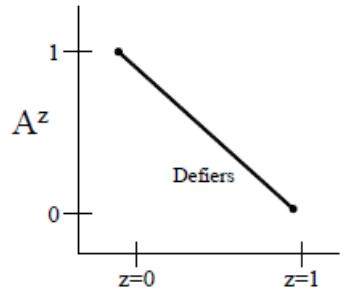


Figure 16.7

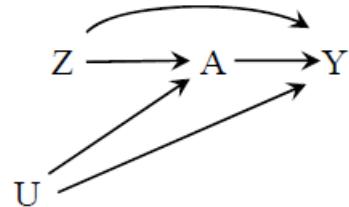


Figure 16.8

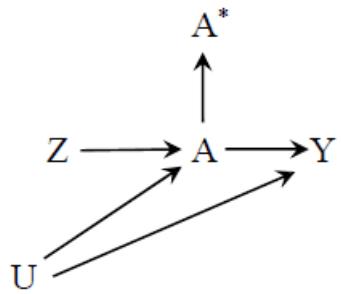


Figure 16.9

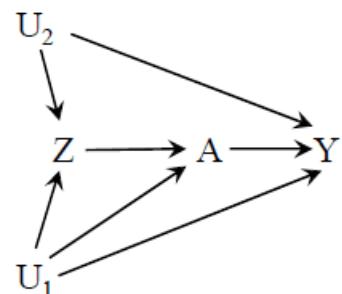


Figure 16.10

第十七章 因果推断中的生存分析

209 在前几章, 我们只考虑治疗在某个特定时间点的因果效应。比如, 在戒烟的例子中, 我们只考虑截止到 1982 年的体重增加。然而许多因果问题却关乎某个结局出现的时间。比如, 我们想估计戒烟对寿命的因果效应, 也即我们关心死亡出现的时间。这些研究被称为生存分析。

“生存分析”这一术语并不是说我们只能关注死亡, 我们也可以称之为“失效时间分析”, 并将其应用于各种各样的结局, 包括死亡、结婚、坐牢、癌症、感染等等。因为有些个体的结局会在研究结束之后才出现, 所以生存分析需要一些额外的考量以及特殊的分析方法。本章将利用时间固定的治疗, 在简单情形中讨论生存分析的基本概念和分析方法。

17.1 危害与风险

如果我们想估计戒烟对存活时间 T 的因果效应, 那我们就是在进行一项生存分析研究。我们的结局是某个事件出现的时间, 从研究开始之时算起, 而事件可能发生在研究开始之后的任意时间点。在大多数跟踪研究中, 不是所有个体的事件都会发生在研究阶段 (也即跟踪阶段), 这是因为大多数研究都有一个结束时间, 而之后的数据就不再收集, 这被称为跟踪的强制结束。

在强制结束之后, 研究人员不再收集任何数据。对于在强制结束之前没有发生事件的个体来说, 他们的存活时间也就被强制删失, 也就是说, 我们知道他们在强制结束之前没有发生事件, 但我们不知道在强制结束之后事件会在什么时候发生。比如, 在我们戒烟的例子中, 有 1629 名人员活到了 1982 年。我们新一轮的跟踪从 1983 年 1 月 1 日开始, 在 1992 年 12 月 31 日强制结束。我们将强制删失时间定义为进行跟踪的时间, 也就是 120 个月。在我们的数据中, 只有 318 人在 1992 年底之前去世, 所以对于剩下 1311 名人员, 他们的存活时间被强制删失了。在我们的研究中, 他们的存活时间等于强制删失时间。

(对于存在分批入组的研究 (也就是有多个跟踪开始日期的研究), 即使最后的研究结束日期都是相同的, 不同的个体依然会有不同的强制删失时间。)

210 强制删失是生存分析的一个根本性问题。我们戒烟和死亡的研究不太可能跟踪每一个人直到所有人都去世。此外, 生存分析也有许多种不同形式的删失, 强制删失只是其中的一种。与其他因果推断分析类似, 生存分析也需要处理其他形式的删失, 比如失访和矛盾事件 (参见精讲点 17.1)。在前几章, 我们用标准化或逆概率加权调整了非强制删失导致的偏移。我们也可以在生存分析中使用同样的方法。因此, 本章将把重点放在强制删失上。我们将在本书第三部分更详尽地讨论非强制删失, 这是因为非强制删失一般而言是一个时异性问题, 而强制删失时间则是在研究一开始就基本确定了。

(为了简便, 我们假设没有明确死亡证明的人在研究期间都是活着的。在现实中, 有些人可能已经死了, 但研究人员得不到死亡证明, 因而依然认为他们还活着。同样为了简便, 我们会忽略精讲点 12.1 所描述的问题。)

在我们的例子中, 存活时间 T 的单位是月, 取值可以是 1 (1983 年 1 月) 到 120 (1993 年 12 月)。治疗组 ($A=1$) 有 102 人在研究阶段去世, 非治疗组 ($A=0$) 则是 216 人, 因而这 318 人的 T 是已知的。而剩下 1311 人则被强制删失, 我们只知道他们的存活时间大于 120 个月。因此, 我们不能像前几章一样计算结局的均值, 也即平均存活时间 $\hat{E}[T]$ 。我们需要将强制删失考虑在内, 因而我们需要使用其他的量度。生存分析中的常见量度有存活概率、风险、危害等。它们都是存活时间 T 的一个函数。我们首先需要定义这些量度。

存活概率 $\Pr[T > k]$ 指的是在第 k 个月, 存活人数占总人数的比例。如果我们在每个月都计算一次存活概率, 直到研究强制结束 (也即 $k_{end} = 120$ 的时候), 然后在一幅坐标图中, 用 Y 轴表示存活概率, X 轴表示时间, 我们就可以得到一幅被称为“生存曲线”的图。在生存曲线的起点, $k = 0$ 的时候, $\Pr[T > 0] = 1$, 然后从此单调递减。我们将 k 时的风险, 也即事件的累积发生率, 定义为 1 减去当前的存活概率, 即 $1 - \Pr[T > k] = \Pr[T \leq k]$ 。累积发生曲线从 $\Pr[T \leq 0] = 0$ 开始, 然后从此单调递增。

在生存分析中, 量化治疗效应的一个直观方法是比较部分时间或所有时间内不同治疗下的风险或存活概率。当然, 在我们戒烟的例子中, 直接比较戒烟组 ($A=1$) 和非戒烟组 ($A=0$) 的生存曲线或风险曲线并没有因果性意义, 这是因为戒烟者和非戒烟组并不能互换。不过, 让我们暂时假设这两组可以直接互换 (实际上, 直到 17.4 小节之前, 我们都会这样假设), 那我们就能在所有时间 k 当中比较 $\Pr[T > 0 | A=1]$ 和 $\Pr[T > 0 | A=0]$ 。比如, 在第 120 个月, 戒烟者的存活概率是 76.2%, 非戒烟者的存活概率则是 82.0%。换一个说法, 在第 120 个月, 戒烟者的风险是 23.8%, 非戒烟者的风险是 18.0%。

(我们还能从生存曲线中得到其他的效用量度, 包括生命损失年, 以及受限存活时间均值等。)

(比较两个生存曲线的常用方法是 log-rank 检验。在我们戒烟的例子中, log-rank 检验的 P 值是 0.005。参见代码 17.1。)

在任何时间 k , 我们都可以计算这段时间内事件发生的人数, 占 k 之前还未发生事件人数的比例, 这就是危害, 也即 $\Pr[T = k | T > k-1]$ 。严格来说, 这是离散时间段当中的危害, 也就是

说, 时间是离散地测量的, 而非连续地测量。在现实世界中, 时间都是离散地测量的(比如年、月、周、日等单位), 因此我们将离散时间段中的危害直接称为危害。

风险和危害是不同的量度。风险的分母是参加研究的所有人, 在所有时间段内都是一个不变的常数; 分子则是研究起始到时间 k 的所有事件数目, 是一个随时间逐渐增加的数。所以, 随着时间的推移, 风险要么保持不变, 要么变大。而危害的分母是到了时间 k 还没有发生事件的人数, 随着时间的增加在逐渐减小, 从而在每段时间内可能都不一样; 分子则是这段时间内的事件数目。也就是说, 随着时间的推移, 危害可能变大, 也可能变小。在我们的例子中, 第 120 个月的时候, 戒烟者的危害是 0% (因为戒烟者中的最后一个死亡出现在第 113 个月), 而非戒烟者的危害是 $1/986 = 0.10\%$ 。从 0 到 120 个月, 危害曲线的形状大致是 M 型的。

生存分析中常用来估计治疗效应的另一个方法是计算治疗组和非治疗组的危害比。不过我们在精讲点 17.2 讨论了危害比存在的问题。因为这些原因, 所以本书的生存分析会将重点放在存活概率或者风险之上, 而非大家常用的危害比。然而, 这并不意味着我们完全放弃危害这个概念。许多时候, 估计危害是估计风险或存活概率的中间步骤。

17.2 从危害到风险

在生存分析中, 有两种整理数据的方式。第一种, 数据中每一行代表一个人, 本书迄今用的 212 都是这种形式。在上一小节, 我们的数据有 1629 行, 每一行代表一个人。

第二种, 数据中的每一行代表某一时间段内的一个人。比如, 第一行是编号 1 的个体在 $k=0$ 时的状态, 第二行是编号 1 的个体在 $k=1$ 时的状态, 第三行是编号 1 的个体在 $k=2$ 时, 以此类推, 直到所有时间段结束, 然后开始下一个人。我们将这种格式称为“人时”格式。在本章和本书的第三部分, 我们将大量使用人时格式。在我们戒烟的例子中, 人时格式共有 176764 行, 每一行代表某个个体在某个月的信息。

为了将各时间段的生存信息包含在人时格式中, 我们需要一个新的指示变量 D_k 。这是一个时异性变量。在第 k 个月, 如果 $T \leq k$, 那么 $D_k = 1$; 如果 $T > k$, 那么 $D_k = 0$ 。图 17.2 表示了治疗 A 和 D_1 以及 D_2 的关系。变量 U 表示能影响事件发生的未测变量。有时 U 也是一个时异变量, 这个时候, 我们可以用 U_0 、 U_1 等形式表示。本书第三部分将讨论这种情况。

在人时格式中, 如果某一行代表时间 k , 那这一行就会包括变量 D_{k+1} 。在我们戒烟的例子中, 第一行的时间是 $k=0$, 因此这一行会包含指示变量 D_1 , 如果这个人在第 1 个月内死亡, 那么 $D_1 = 1$, 否则 $D_1 = 0$ 。第二行的时间是 $k=1$, 因此这一行包含指示变量 D_2 , 如果这个人在第

2个月内死亡, 那么 $D_2 = 1$, 否则 $D_2 = 0$ 。以此类推, 直到数据的最后一行。在我们的数据中, 各个体的最后一行的时间要么是 $D_{k+1} = 1$ 时的第 k 个月, 要么是第 119 个月。

(根据定义, 所有人在第 0 个月都必然存活, 也即对所有人有 $D_0 = 0$ 。)

使用时异变量 D_k , 我们可以把 k 时的存活概率定义为 $\Pr[D_k = 0]$, 等于 $\Pr[T > k]$; 将 k 时的风险定义为 $\Pr[D_k = 1]$, 等于 $\Pr[T \leq k]$ 。 k 时的危害定义为 $\Pr[D_k = 1 | D_{k-1} = 0]$ 。根据定义, 所有人在 $k = 0$ 时都是存活状态, 因此 $k = 1$ 时的危害等于此时的风险。

k 时的存活概率, 等于 0 到 k 之间每个时间段所有条件存活概率的乘积。比如, $k = 2$ 时的存活概率 $\Pr[D_2 = 0]$, 等于 $k = 1$ 时的存活概率 $\Pr[D_1 = 0]$, 乘以 $k = 1$ 阶段还活着的人在 $k = 2$ 时依然活着的概率 $\Pr[D_2 = 0 | D_1 = 0]$ 。因此, k 时的存活概率可以表示为:

$$\Pr[D_k = 0] = \prod_{m=1}^k \Pr[D_m = 0 | D_{m-1} = 0]$$

也就是说, k 时的存活概率, 等于之前各时间段 1 减去危害的乘积。如果我们知道 k 及之前所有时间段的危害, 我们就能计算出 k 时的存活概率以及风险 (1 减去存活概率)。

我们可以使用非参数方法估计 k 时的危害 $\Pr[D_k = 1 | D_{k-1} = 0]$ 。我们只需要在每个时间段, 用这个时间段事件发生的人数, 除以上个时间段仍未发生事件的人数, 就能得到危害。把上述存活概率表达式中的危害替换为这一非参数的估计, 就叫做存活概率的 Kaplan-Meier 估计, 或者乘积极限估计。²¹³ 图 17.1 就是使用 Kaplan-Meier 估计画出来的。鉴于整个跟踪研究期间事件总数会非常多, 这一方法是绘制生存曲线的最佳方法之一。一般而言, 每一个时间段内的事件数都不是很多 (甚至是 0), 因此危害的非参数估计会很不稳定, 得到的生存曲线也就崎岖不平。如果我们希望在估计危害的同时, 还能将曲线变得更加平滑, 那么我们可能需要使用参数模型 (第十一章和精讲点 17.3 讨论了平滑问题)。

一个简便的参数方法是在 k 时还未发生事件的人群中拟合 $\Pr[D_{k+1} = 1 | D_k = 0]$ 的 logistic 回归模型。如果我们使用人时格式的数据, 那么拟合这一模型就会十分方便。在我们的例子中, 我们可以在治疗组和非治疗组中拟合下述 logistic 模型:

$$\text{logit } \Pr[D_{k+1} = 1 | D_k = 0, A] = \theta_{0,k} + \theta_1 A + \theta_2 A \times k + \theta_3 A \times k^2$$

其中 $\theta_{0,k}$ 是时异性结局, 可以表示为时间的函数, 比如 $\theta_{0,k} = \theta_0 + \theta_4 k + \theta_5 k^2$ 。这一灵活的函数形式可以让我们在模型中加入时异危害, 以及治疗 A 与时间的乘积项 (比如 $\theta_2 A \times k + \theta_3 A \times k^2$)。

知识点 17.1 详细讨论了这一 logistic 模型为何可以近似看作一个危害模型。

(其他能用于二分结局的联系函数也可以用于危害。)

(参见代码 17.2。)

(即使在人时格式中每个个体都有许多行, 但是只要模型设定正确, 危害的 logistic 模型依然能够给出正确的标准误差。)

我们能从 logistic 模型中得到 $\Pr[D_{k+1} = 1 | D_k = 0, A = a]$ 的估计值, 我们用 1 减去这个值, 然后再把它们乘起来, 就得到存活概率 $\Pr[D_{k+1} = 0 | A = a]$ 。图 17.4 描绘了使用参数方法得到的生存曲线, 这一曲线比图 17.1 更加平滑。

214 这一方法需要我们的模型设定正确。在我们的例子中, 我们用参数法和非参数法得到的生存曲线几乎是一样的, 因此我们的模型假设应该是正确的。而生存曲线的 95% 置信区间可以用自举法计算得到。

17.3 删失

在我们戒烟的例子中, 只存在强制删失, 并且强制删失时间 $k_{end} = 120$ 对所有人都是一样的。在这种简易情形下, 上一小节用来估计存活概率的方法就显得牛刀小用。我们只需要知道治疗取值为 a 且存活到 $k+1$ 的人数所占比例, 就能估计存活概率 $\Pr[D_{k+1} = 0 | A = a]$, 或者在每个时间段 $k = 0, 1, \dots, k_{end}$ 中分别拟合 $\Pr[D_{k+1} = 0 | A]$ 的 logistic 模型。

假设我们在不同的时间点对研究中的个体开始跟踪研究, 而研究结束日期对所有人都是一样的, 那么此时所有个体的强制删失时间不再相同。在这种情况下, 定义一个新的时异变量 C_k 就会很有用。在第 k 个月, 如果强制删失时间大于 k , 那么 C_k 值为 0, 否则为 1。在人时格式的数据中, 对应时间 k 的每一行都会包含变量 C_{k+1} 。不过在我们戒烟的例子中, 我们没有包含 C_{k+1} 这个变量, 这是因为数据中的所有人在 k 小于 120 的时候, 都有 $C_{k+1} = 0$ 。不过在更普遍的情形中 (也即各个体的强制删失时间不尽相同时), 每个个体的 C_{k+1} 会在不同时间 k 处从 0 变成 1。

最理想的情况是沒有人在 k_{end} (也即整个研究的最大强制删失时间) 之前被删失, 然后我们 215 在这种情况下估计生存曲线。也就是说, 我们的目标是在知道 D_k 的情况下 (即使被删失了) 估计

存活概率 $\Pr[D_k = 0 | A = a]$ 。在更严格意义上, 我们会将这一值写作 $\Pr[D_{\bar{c}=\bar{0}} = 0 | A = a]$, 其中

$\bar{c} = (c_1, c_2 \dots c_{k_{end}})$ 。我们在第十二章讨论过, 使用上角标 $\bar{c} = \bar{0}$ 能让我们更清楚地理解这一数值的含义。在很多情况下, 只要不存在歧义, 我们就会忽略上角标 $\bar{c} = \bar{0}$ 。为了简便, 假设各个体的研究开始时间是随机的, (如果没有任何变量显示出某种有迹可循的长期趋势, 那这一假设是合理的。) 那么强制删失时间以及指示变量 \bar{C} 就会独立于治疗和死亡时间。

在 k 时, 如果我们只是简单地计算治疗取值为 a 且未被删失的存活人数在总人数中所占比例, 那我们得到的是 $\Pr[C_{k+1} = 0, D_{k+1} = 0 | A = a]$, 而非我们想要的 $\Pr[D_k = 0 | A = a]$ 。比如下列的一个简化例子就描绘了这一情形:

- 研究的最大强制删失时间 $k_{end} = 2$ 。
- $\Pr[C_1 = 0] = 1$, 也即没有人在 $k = 1$ 时被删失。
- $\Pr[D_1 = 0 | C_0 = 0] = 0.9$, 也即 90% 的个体在 $k = 1$ 时存活。
- $\Pr[C_2 = 0 | D_1 = 0, C_1 = 0] = 0.5$, 也即有一半的存活者在 $k = 2$ 时被删失。
- $\Pr[D_2 = 0 | C_2 = 0, D_1 = 0, C_1 = 0] = 0.9$, 也即剩下的人中有 90% 在 $k = 2$ 时存活。

在 $k = 2$ 时, 未被删失的存活概率是 $1 \times 0.9 \times 0.5 \times 0.9 = 0.405$ 。然而, 如果没有人被删失, 也即 $\Pr[C_2 = 0 | D_1 = 0, C_1 = 0] = 1$, 那么存活概率是 $1 \times 0.9 \times 1 \times 0.9 = 0.81$ 。这一例子说明了为什么

216 我们要用上一小节的办法去估计存活概率。具体而言, 假设删失是随机的, 那在时间 k 的存活概率 $\Pr[D_k = 0 | A = a]$ 可以表示为:

$$\prod_{m=1}^k \Pr[D_m = 0 | D_{m-1} = 0, C_m = 0, A = a]$$

其中 $k < k_{end}$ 。其实我们在前面已经讨论了如何估计这一数值, 区别只是使用非参数方法, 还是使用参数方法。

不过有时删失并不是随机的。如果存在分批入组的情形, 那么各个体的强制删失时间就取决于入组的日期, 而这个日期可能和结局相关。因此, 上述方法需要调整日期。此外, 某些基线预后因素的分布可能在治疗组 ($A = 1$) 和非治疗组 ($A = 0$) 中并不均衡。下一小节将会讨论如何使用 G-方法调整这些混杂。在本书第三部分, 我们将把这一方法延伸到有时异治疗的情形中。

17.4 生存分析中的逆概率加权

当治疗组和非治疗组不可互换时, 直接比较这两组的生存曲线并不能得到有意义的结果。在我们戒烟的例子中, 戒烟者和非戒烟者 120 个月的存活概率分别是 76.2% 和 82.0%, 但这并不是说戒烟会提升死亡率。在我们的人群中, 戒烟者的平均年龄更大。年老一些的个体更可能会戒烟, 也更可能死亡。年龄的混杂影响使得戒烟看起来似乎不好。

我们用 $D_k^{a,\bar{c}=\bar{0}} = 0$ 表示在时间 k , 治疗取值为 a 且没有删失时的死亡情况, 这是一个反事实异变量。为了简化, 我们省略上角标 $\bar{c} = \bar{0}$, 仅写作 $D_k^a = 0$ 。更进一步, 我们在本章剩余部分将会忽略 $C_k = 0$, 从而将 k 时的危害表示为 $\Pr[D_{k+1} = 0 | D_k = 0, L = l, A]$ 。也就是说, 我们假设所有人的强制删失时间相同, 就如同我们戒烟的例子一样。

我们想比较的是 $k = 0, 1, 2, \dots, k_{end} - 1$ 时的反事实结局 $\Pr[D_{k+1}^{a=1} = 0]$ (也即所有人都接受治疗) 和 $\Pr[D_{k+1}^{a=0} = 0]$ (也即所有人都没有接受治疗)。然而因为混杂的存在, 我们上一小节中的 $\Pr[D_{k+1} = 0 | A = 1]$ 和 $\Pr[D_{k+1} = 0 | A = 0]$ 并不能直接用来估计这两个反事实结局。因此, 我们还需要调整混杂变量。我们有许多不同的方法, 这一小节将介绍逆概率加权。

217 假设在变量 L 的每一分层中, 治疗组 ($A = 1$) 和非治疗组 ($A = 0$) 是可互换的, 就如同图 17.5 所描绘的一样。在本章, 变量 L 和前几章一样, 包括性别、年龄、种族、教育程度、吸烟频率、烟龄、每日体力情况、运动量、以及体重。我们依然假设正数性和一致性成立。用逆概率加权估计生存曲线有两个步骤。

第一步, 我们先估计人群中每一个人的稳定逆概率权重 SW^A 。这一步和第十二章中的做法完全一样。我们先拟合一个 logistic 模型, 其中因变量是治疗的条件概率 $\Pr[A = 1 | L]$, 而自变量是 L 。从这一模型可得到估计值 $\hat{\Pr}[A = 1 | L]$ 。 SW^A 的分母在治疗组中是 $\hat{\Pr}[A = 1 | L]$, 在非治疗组中是 $1 - \hat{\Pr}[A = 1 | L]$ 。我们可以通过非参数的方法得到 $\hat{\Pr}[A = 1]$, 也即治疗组在总人群中所占的比例, 也可以通过参数方法得到这一估计值。 SW^A 的分子在治疗组中是 $\hat{\Pr}[A = 1]$, 在非治疗组中是 $1 - \hat{\Pr}[A = 1]$ 。

(参见代码 17.3。)

利用 SW^A , 我们可以构建一个虚拟人群, 人群中的变量 L 独立于治疗 A , 从而消除了 L 带来的混杂。在我们的例子中, 权重的均值是 1, 取值范围从 0.33 到 4.21。

第二步, 如同上一小节一样, 利用人时格式的数据拟合每个时间段危害的模型, 区别在于此时我们会给每个人进行加权。加权后的结构模型可以表示为:

$$\text{logit } \Pr[D_{k+1}^a = 0 | D_k^a = 0] = \beta_{0,k} + \beta_1 a + \beta_2 a \times k + \beta_3 a \times k^2$$

理论上, 这一模型估计的是时异性的反事实危害。

我们可以将模型中得到的各时间段的存活概率 $\Pr[D_{k+1}^a = 0 | D_k^a = 0]$ 相乘, 从而得到治疗组和非治疗组的整体存活概率 $\Pr[D_{k+1}^a = 0]$ 。图 17.6 描绘了利用这一方法得到存活曲线。

在我们的例子中, 戒烟者 120 个月的存活概率是 80.7%, 非戒烟者则是 80.5%。两者差值是 0.2% (利用 500 次自举可以得到 95% 置信区间为 -4.1% 至 3.7%)。尽管在大多数时间内, 戒烟者的存活概率都低于非戒烟者, 但是最大的差值也没有超过 -1.4% (95% 置信区间: -3.4%, 0.7%)。也就是说, 使用逆概率加权调整了 L 之后, 我们没有发现戒烟会导致更大的死亡率。这一方法需要我们的治疗模型和危害模型都设定正确。

17.5 生存分析中的 G-公式

在上一小节, 我们用逆概率加权估计了整个研究中治疗组和非治疗组的生存曲线。在这一方法中, 我们调整了 L 并假设互换性、正数性、以及一致性成立。在这些假设下, 我们也可以使用标准化方法估计生存曲线, 也即使用参数 G-公式估计生存曲线。

$k=1$ 时的存活概率 $\Pr[D_{k+1}^a = 0]$, 是 L 各分层中条件存活概率的加权总和, 其权重是每一分层在总人数中所占比例。也就是说, 在互换性、正数性、一致性假设下, 存活概率可以表示为:

$$\Pr[D_{k+1}^a = 0] = \sum_l \Pr[D_{k+1}^a = 0 | L = l, A = a] \Pr[L = l]$$

证明在 2.3 小节已经给出。

因此, 利用参数 G-公式估计存活概率也需要两步。第一步, 估计每一分层中的条件存活概率 $\Pr[D_{k+1}^a = 0 | D_k^a = 0, L, A]$ 。第二步, 计算加权均值。我们在第十三章已经介绍过这两步。

对于第一步, 我们像 17.2 小节中一样, 需要拟合一个危害的参数模型, 区别在于此时 L 作为协变量会被放入模型中。如果模型设定正确, 那就能在 L 的每一分层中有效地估计时异危害 $\Pr[D_{k+1}^a = 1 | D_k^a = 0, L, A]$, 1 减去该时异危害就能得到这一时间段的条件存活概率, 而将各时间段的存活概率相乘就能得到总的条件概率, 也即:

$$\Pr[D_{k+1}^a = 0 | L, A] = \prod_{m=0}^k \Pr[D_{m+1}^a = 0 | D_m^a = 0, L = l, A = a]$$

因为 L 之下的有界互换性成立, 所以给定 $L = l$ 和 $A = a$ 的条件存活概率就能被阐释为治疗取值为 a 时的反事实结局, 也即:

$$\Pr[D_{k+1} = 0 | L = l, A = a] = \Pr[D_{k+1}^a = 0 | L = l]$$

因此, 我们就能估计出 L 不同分层 l 下治疗组和非治疗组的生存曲线。不过我们的目标是估计整个人群的边缘生存曲线。

(在第十二章, 我们将控制了 L 中所有变量的模型称为仿制的边缘结构模型。)

对于第二步, 我们需要计算 L 所有分层中存活概率的加权均值, 也就是根据混杂的分布, 将存活概率标准化。我们将使用 13.3 小节中的方法: 先将数据扩大, 然而进行结局回归, 最后再估计预测值。就算 L 中的某些变量是连续性变量, 这一方法依然有效, 此时的求和就会变成积分。用这一方法得到生存曲线如图 17.7 所示。

(参见代码 17.4。这一方法和第十三章的同理。)

在我们的例子中, 戒烟者在大部分时间的存活概率都与非戒烟者不同, 但是最大的差值也没有超过 -2.0% (95% 置信区间: -5.6%, 1.8%)。戒烟者 120 个月的存活概率是 80.4%, 非戒烟者则是 80.6%。两者差值是 0.2% (利用 500 次自举可以得到 95% 置信区间为 -4.6% 至 4.1%)。也就是说, 使用标准化调整了 L 之后, 我们没有发现戒烟会导致更大的死亡率。注意到, 本小节使用标准化得到的结果, 和上一小节使用逆概率加权得到的结果虽然相似, 但是有所不一样。这是因为两种方法的参数假设不一样: 逆概率加权需要治疗的模型和没有协变量的危害模型设定正确, 而参数 G-公式要求危害模型中的协变量设定正确。

17.6 生存分析中的 G-估算

前面几个小节我们讲述了如何比较不同治疗取值下的存活概率或风险。我们用 logistic 回归模型估计危害, 再用危害计算存活概率。我们可以用逆概率加权或者 G-公式实现这一过程。然而, 这个过程却不能使用结构嵌入模型及其 G-估算。在第十四章我们解释过, 结构嵌入模型适用于条件性的比较 (比如, 协变量等于某个值时, 不同治疗取值下结局的差值或者比值), 但不适用于构成比较的部分 (比如, 不同治疗取值下的结局均值)。因此, 我们不能用结构嵌入模型估计存活概率或者危害。

(实际上, 因为结构嵌入模型并不能轻易扩展到时异治疗, 所以我们甚至不能用这一方法得到危害比的近似值 (参见知识点 14.1)。)

不过, 我们可以用联系函数为对数函数的结构嵌入模型, 对不同治疗取值下的累积发生率 (比如风险) 进行建模。“结构嵌入累积失效时间模型”就是用来做这个的 (参见知识点

17.2)。然而, 因为对数联系函数并不会将风险的上界设定为 1, 所以这一方法最好仅用于罕见事件。对于不罕见的事件, 我们可以用联系函数是对数的结构嵌入模型估计不同治疗取值下累积存活概率 (也即 1 减去风险), “结构嵌入累积存活时间模型” 就是用来做这个的 (参见知识点 17.2)。然而, 因为对数联系函数并不会将存活概率的上界设定为 1, 所以这一方法最好仅用于存活是一件罕见事件的时候。另一种更普适的方法是用结构嵌入模型直接对不同治疗下存活时间的比值进行建模。我们把这一方法称为“加速失效时间”模型 (英文简写: AFT 模型)。

(研究者发现在生存分析中使用工具变量, 会遇到和结构嵌入模型一样的问题。)

用 T_i^a 表示个体 i 在治疗取值为 a 时的反事实存活时间。治疗 A 对个体 i 的因果效应就可以表示为 $T_i^{a=1} / T_i^{a=0}$ 。如果这一比值大于 1, 那么治疗就是有益的, 因为有治疗时的存活时间更长了。如果这一比值小于 1, 那么治疗就是有害的; 等于 1, 就是无效的。让我们暂时假设治疗的效应对于人群中的所有个体来说都是一样。

接下来我们就可以构建“加速失效时间模型”: $T_i^a / T_i^{a=0} = \exp(-\psi_1 a)$, 其中 ψ_1 衡量的是治疗给存活时间带来的效果 (可能增大可能减小)。如果 $\psi_1 < 0$, 那就是增大存活时间; 如果 $\psi_1 > 0$, 那就是减小存活时间; 如果 $\psi_1 = 0$, 那就是没有效果。不过一般而言, 治疗的效应也会取决于其余协变量 L , 因此一个更普适的 AFT 模型会是: $T_i^a / T_i^{a=0} = \exp(-\psi_1 a - \psi_2 a L_i)$, 其中 ψ_2 是一个向量, ψ_1 和 ψ_2 对所有个体而言都是一样的。通过移项变形, 我们可以得到:

$$T_i^{a=0} = T_i^a \exp(\psi_1 a + \psi_2 a L_i)$$

(当治疗是时异性的时候, “结构嵌入模型” 中“嵌入”一词的意义才会更加明显。参见第二十一章。)

(ψ 前面的负号只是为了让我们的阐释与大多研究相同, 也即正的参数表示有害, 负的参数表示有益。)

在第十四章我们已经论述过, 根据一致性我们可以用观测值 T_i^A 替代 T_i^a 。参数 ψ_1 和 ψ_2 也可以通过 G-估算得到, 不过因为强制删失, 所以我们对 G-估算进行了改良。在本小节我们将讲述改良版的 G-估算。

上述 AFT 模型没有现实意义, 因为这是一个命定模型, 且需要保序性。这个模型是命定的, 因为这个模型假设对于所有个体, 我们都可以用观测到的存活时间 T 、治疗 A 、以及协变量 L 准确无误地算出没有治疗时的反事实结局 $T^{a=0}$ 。这个模型需要保序性, 因为在这个模型中, 如果 $T_i^{a=0} < T_j^{a=0}$, 那么个体 i 和 j 都接受治疗时, i 必须死在 j 的前面, 也即 $T_i^{a=1} < T_j^{a=1}$ 。

221 因为保序性不可能成立, 所以我们不能依赖于这一假设, 这一点在第十四章已经讨论过。不过我们将用保序性介绍 AFT 模型的 G-估算, 因为保序性有助于我们理解 G-估算, 同时也因为不管保序性成不成了, G-估算的步骤都是一样的。

(Robins (1997b) 讨论过非命定、保序性不成立的结构嵌入 AFT 模型。)

为了简便, 我们先考虑没有治疗和协变量乘积项的简易保序性模型 $T_i^{a=0} = T_i \exp(\psi A_i)$ 。如果不存在强制删失, 也即我们能观测到每个人的存活时间 T , 那么参数 ψ 的 G-估算就会非常直接方便, 此时 G-估算的步骤就和 14.5 小节中的一样。第一步, 计算不同 ψ 的可能取值 ψ^\dagger 之下的 $H_i(\psi^\dagger) = T_i \exp(\psi^\dagger A_i)$ 。第二步, 在治疗的 logistic 模型中, 找到使得 $H_i(\psi^\dagger)$ 独立于治疗 A 的 ψ^\dagger 。

不过, 如果在 K 时存在强制删失, 对于 T_i 未知的个体, 我们无法计算 $H_i(\psi^\dagger)$, 因而也就不能使用上述方法。此时, 可能有的人就想把 G-估算局限于 $T_i \leq K$ 的人当中。然而这样的话, 会造成选择偏移。下面一段中的简易例子说明了为什么会有选择偏移。

我们要进行一项 60 个月的随机试验, 治疗 A 是一个二分变量, 我们想估计 A 对存活时间 T 的因果效应。参加试验的被试可以分成三种不同类型。第一种是如果没有治疗, 那就会在第 36 个月死亡 ($T^{a=0} = 36$)。第二种是如果没有治疗, 那就会在第 72 个月死亡 ($T^{a=0} = 72$)。第三种是如果没有治疗, 那就会在第 108 个月死亡 ($T^{a=0} = 108$)。也就是说, 第一种类型的被试预后最差, 而第三种的最好。因为是随机分组, 所以每一种类型的被试在治疗组 ($A = 1$) 和非治疗组 ($A = 0$) 中的人一样多, 也就是说治疗组和非治疗组可互换。

假设治疗会减少存活时间。表 17.1 显示了每种类型的被试在治疗组和非治疗组中的存活时间。因为强制删失时间是 60 个月, 所以第一种类型的被试不管在治疗组还是在非治疗组, 我们都能观测到他们的存活时间。而第三种类型的被试则不管是在治疗组还是在非治疗组, 都会被强制删失。但对于第二种类型的被试而言, 在治疗组中我们能观测到他们的存活时间, 但在非治疗组中则不能。如果我们把分析局限在没有被强制删失的人群之中, 那我们的治疗组和非治疗组就不可互换。因此, 平均而言, 非治疗组中的个体预后更差, 从而使得我们的治疗似乎是有益的。此时只要治疗的效应不为零, 那就可能会产生选择偏移 (第八章)。

为了避免选择偏移, 我们需要有治疗和没治疗时的反事实结局都小于强制删失时间的个体, 也即 $T_i^{a=0} \leq K$ 且 $T_i^{a=1} \leq K$ 的个体。在我们的例子中, 我们需要把所有第二种类型的个体全都排除于分析之外, 从而保证互换性。也就是说, 我们不仅需要排除被强制删失 ($T_i > K$) 的个体, 还

要排除某些没有被强制删失 ($T_i \leq K$) 的个体, 因为这些没有被强制是的个体可能在改变治疗取值后会被强制删失。

(排除没有被强制删失的个体也被称为人工删失。)

我们定义一个指示变量 $\Delta(\psi)$, 如果某个体被排除于分析之外, 那 $\Delta(\psi)$ 取值为 1, 否则为 0。改良后的 G-估算, 将会用 $\Delta(\psi^\dagger)$ 替代 $H(\psi^\dagger)$, 具体细节参考知识点 17.3。在我们的例子中, 利用改良后的 G-估算, 从保序性成立的 AFT 模型 $T_i^{a=0} = T_i \exp(\psi A_i)$ 当中, 我们可以得到估计值为 -0.047 (95% 置信区间: -0.223, 0.333)。 $\exp(-\hat{\psi}) = 1.05$ 可以被阐释为 $a=1$ 与 $a=0$ 的反事实存活时间均值的比值。这一存活时间比值表示戒烟 A 对死亡时间基本没有影响。

(参见代码 17.5。)

(ψ 的点估计值能使知识点 17.3 中的估计函数取得最小值。95% 置信区间的上下界对应的是使一个自由度下的卡方数为 3.84 的估计函数的取值。)

我们在第十四章已经讨论过, 结构嵌入模型 (包括 AFT 模型) 很少用于实际研究中。一个主要原因是尚无软件能简便地运行这一模型。另一个更重要的原因是生存分析需要使用搜索算法估计 AFT 模型的参数值, 而这一方法并不能保证我们一定能得到唯一解。当我们有两个或多个参数 ψ 的时候, 这一问题更加凸显。然而使用了简易版的 AFT 模型, 研究者就不能在模型中包含其他协变量的效应修饰作用。

相较于危害模型, 我们能更方便地在 AFT 模型中反映研究问题的背景知识 (比如生物机理), 尤其是当我们使用的是非命定的、不需要保序性的 AFT 模型的时候。遗憾的是, 因为上述提到的困难, AFT 模型很少被人使用。

第十七章精讲点和知识点

精讲点 17.1: 矛盾事件 (原书第 210 页)

我们在第 8.5 小节讨论过, 矛盾事件 (比如死亡) 会阻碍我们关注事件的发生 (比如中风): 因其他死因 (比如癌症) 去世之后就不能再观测到中风。在生存分析中, 我们需要考虑是否将矛盾事件算作一种非强制删失。

- 如果把矛盾事件算作删失, 那么我们的分析就是假设死于其他原因的人, 要么直接被排除了, 要么被认为不受中风的风险因素影响。这会导致我们得到的估计值基本没有实际意义

(参见第八章)。除此之外, 删失也会导致零值下的选择偏移, 从而我们需要调整我们关注事件的已知风险因素。

- 如果不把矛盾事件算作删失, 那么我们的分析就是假设事件出现的时间是无限的。也就是说, 死亡的个体出现我们关注事件的概率为 0。这会导致我们的估计值也基本没有意义, 因为我们可能会得到治疗对死亡的非零效应估计, 但死亡阻止了中风发生。

处理矛盾事件的一个方法是构建一个复合事件, 其同时包含了矛盾事件以及我们关注的事件, 并对复合事件进行生产分析。不过, 这一方法虽然能有效地解决矛盾事件带来的干扰, 却会改变我们的研究问题。并且, 这一方法得到的效应估计也难以阐释, 因为其中的效应可能是对死亡的, 也可能是对中风的。另一种方法是将分析局限于没有死亡的人群当中。这一方法只能得到局部的效应均值(参见第十六章), 会给这个估计的有效性以及我们的阐释打上问号。

上述没有一种方法能完美地解决矛盾事件带来的问题。实际上, 矛盾事件的存在会让我们思考统计模型中得到的估计值有什么意义。Young (2019) 等人就这个问题有更深入的探讨。

精讲点 17.2: 危害比的“危害”(原书第 213 页)

当我们使用危害比作为因果效应的量度的时候, 我们需要考虑到危害比的两个特点。

第一, 因为危害随时间的不同而不同, 危害比也随时间的不同而不同。也就是说 k 时的危害比可能和 $k+1$ 时的危害比不一样。然而, 许多研究都只报告一个危害比, 这是因为 Cox 模型假设危害比在所有时间都是一样的。因此, 这个单一危害比是各时间段危害比的加权平均, 从而它的实际意义很难说明。如果事件发生概率很小, 并且只有强制删失, 那么 k 时的危害比权重正比于这一时间段内非治疗组的事件总数。(严格而言, 这一权重等于 $A=0$ 且 $T < k_{end}$ 时, k 在 T 中的条件密度。) 因为这是一个加权平均, 所以即使生存曲线不太一样, 危害比也可能等于 1。与危害比不同的是, 存活概率和风险的表述总是需要和时间一起出现, 比如 5 年存活率等。

第二, 即使我们报告的是每个时间段的危害比, 这些危害比的意义依然不是十分明朗。假设在 k 时治疗会让高风险人群死亡, 但对其他人没有影响, 那么 $k+1$ 时的治疗组人群就会都是低风险人群, 而非治疗组则同时有低风险和高风险人群。因此, 即使治疗无益, $k+1$ 时的危害比依然会小于 1。

这一矛盾是选择偏移的一个绝佳例子, 此时我们控制了某个治疗后的变量(也即 k 时的存活情况), 而这一变量受到治疗的影响。比如, 在时间 2 的危害比, 是活过时间 1 的人群中事件发生的概率, 也即 $\Pr[D_2 = 1 | D_1 = 0, A]$ 。如图 17.3 所描绘的一样, 控制了对撞变量 D_1 会打开路径 $A \rightarrow D_1 \leftarrow U \rightarrow D_2$, 因而就会使得治疗 A 和 D_2 相关。如果治疗组和非治疗组的生存曲线是一样

的话, 那么这种选择偏移就不会出现, 也就是说此时没有从 A 指向后续事件 D 的箭头。Hernan (2010) 在他的论文中讨论了这一问题。

精讲点 17.3: 生存分析的模型 (原书第 214 页)

生存分析的模型需要考虑强制删失。

生存分析的非参数方法——比如 Kaplan-Meier 曲线——不需要对未观测到的存活时间做出任何假设。而生存分析的参数模型需要对此假设一个分布 (比如指数分布, Weibull 分布)。正文中的用来估计危害的 logistic 模型就是其中一个例子。

其他模型, 比如 Cox 模型和 AFT 模型, 则不需要对存活时间进行假设。即使所有协变量都是零, 这些模型也不会给出危害的分布 (此时的危害称为初始危害)。不过这些模型先验地假设了初始危害和不同协变量下的危害之间的关系。因此, 这些模型也被称为半参数模型。

知识点 17.1: 通过 logistic 模型近似估计危害比 (原书第 215 页)

危害模型 $\Pr[D_{k+1} = 1 | D_k = 0, A] = \Pr[D_{k+1} = 1 | D_k = 0, A = 0] \times \exp(\alpha_1 A)$ 中的 $\exp(\alpha_1)$ 就是危害比。如果我们把模型两侧都取对数, 我们可以得到 $\log \Pr[D_{k+1} = 1 | D_k = 0, A] = \alpha_{0,k} + \alpha_1 A$, 其中 $\alpha_{0,k} = \log \Pr[D_{k+1} = 1 | D_k = 0, A = 0]$ 。

假设 $k+1$ 时的危害很小, 也即 $\Pr[D_{k+1} = 1 | D_k = 0, A = 0] \approx 0$, 那么 1 减去 $k+1$ 时的危害就近似于 1, 因此危害就近似等于发生比: $\Pr[D_{k+1} = 1 | D_k = 0, A = 0] \approx \frac{\Pr[D_{k+1} = 1 | D_k = 0, A]}{\Pr[D_{k+1} = 0 | D_k = 0, A]}$ 。

因此我们有:

$$\log \frac{\Pr[D_{k+1} = 1 | D_k = 0, A]}{\Pr[D_{k+1} = 0 | D_k = 0, A]} = \text{logit } \Pr[D_{k+1} = 1 | D_k = 0, A] \approx \alpha_{0,k} + \alpha_1 A$$

也就是说, 如果 $k+1$ 时的危害接近于 0, 那我们就能用 logistic 模型 $\text{logit } \Pr[D_{k+1} = 1 | D_k = 0, A] = \theta_{0,k} + \theta_1 A$ 当中的 θ_1 近似估计危害比的对数 α_1 , 就如我们在正文中所作的一样。如果 $\Pr[D_{k+1} = 1 | D_k = 0, A] < 0.1$, 那这一近似做法就被认为是合理的。

而这一罕见事件假设基本上都能成立, 我们只需要把时间单位 k 定义得足够段就能保证 $\Pr[D_{k+1} = 1 | D_k = 0, A] < 0.1$ 。比如, 如果 D_k 表示肺癌, k 可以以年为单位。如果 D_k 表示普通流感, k 可以以天为单位。单位越小, 人时格式数据的行数就越多。

知识点 17.2: 累积失效时间 (CFT) 和累积存活时间 (CST) 的结构嵌入模型 (原书第 220 页)

对于时间固定的治疗, CFT 的 (非嵌入式) 结构模型是对不同治疗取值下反事实风险的比值进行建模, 并控制协变量 L 。其一般形式是:

$$\frac{\Pr[D_k^a = 1 | L, A]}{\Pr[D_k^{a=0} = 1 | L, A]} = \exp[\gamma_k(L, A; \psi)]$$

其中 $\gamma_k(L, A; \psi)$ 表示治疗和协变量的某个函数, 其对应的参数 (向量) 是 ψ 。根据一致性, $A = 0$ 时, $\exp[\gamma_k(L, A; \psi)] = 1$, 这是因为此时你要比较的两个值是同样治疗下的值, 所以应该一样。其中一种可能的函数形式是 $\gamma_k(L, A; \psi) = \psi A$ 。

同理, 对于时间固定的治疗, CST 的 (非嵌入式) 结构模型是对不同治疗取值下反事实存活时间的比值进行建模, 并控制协变量 L 。其一般形式是:

$$\frac{\Pr[D_k^a = 0 | L, A]}{\Pr[D_k^{a=0} = 0 | L, A]} = \exp[\gamma_k(L, A; \psi)]$$

虽然 CFT 和 CST 模型的唯一区别是模型中的 $\Pr[D_k^a = 1 | L, A]$ 和 $\Pr[D_k^a = 0 | L, A]$, 但是 $\gamma_k(L, A; \psi)$ 的含义却迥然不同。如果 k 是连续的而非离散的, 那么 CST 的结构模型就等价于危害的加成结构模型, 这是因为任何对 $\frac{\Pr[D_k^a = 0 | L, A]}{\Pr[D_k^{a=0} = 0 | L, A]}$ 建模的模型衡量的都是时间点 T^a 和 $T^{a=0}$ 中危害的差值, 反之亦然。

CFT 的结构模型要求在 L 的所有分层中, 任何治疗取值下的累积失效概率需要满足一个特定的罕见事件假设。此时, CFT 模型会优于 AFT 模型, 因为 CFT 模型认为模型参数中的无偏估计方程可微, 因此能轻易解出这个方程。

知识点 17.3: 人工删失 (原书第 222 页)

用 $K(\psi)$ 表示没有治疗时的最小存活时间, 某些个体可能正好在强制删失时间 K 时死亡。对于一个二分治疗 A , $K(\psi) = \inf\{K \exp(\psi A)\}$, 其表示如果治疗缩短了存活时间 (即 $\psi > 0$), 那么有 $K(\psi) = K \exp(\psi \times 0) = K$; 如果治疗延长了存活时间 (即 $\psi < 0$), 那么有

$K(\psi) = K \exp(\psi \times 1) = K \exp(\psi)$; 如果治疗对存活时间没有效应 (即 $\psi = 0$) , 那么有
 $K(\psi) = K \exp(0) = K$ 。

所有被强制删失的人 (即 $T > K$) 都有 $\Delta(\psi) = 0$, 因为他们的实际存活时间大于强制删失时间, 即 $H(\psi) > K(\psi)$ 。某些人没有被强制删失 (即 $T \leq K$) 也有 $\Delta(\psi) = 0$, 这些人被人工删失——为了减少选择偏移——而被排除于分析之外。

人工删失的指示变量 $\Delta(\psi)$ 是 $H(\psi)$ 和 K 的一个函数。在给定 L 的有界互换性下, 如果用 ψ 的真实值衡量, 所有这些函数在控制了协变量 L 之后都会独立于 A 。也就是说, AFT 模型的 G-估算并不需要 $H(\psi)$, 而可以在 $\Delta(\psi)$ 的基础上进行。或者说, 我们用 $\Delta(\psi)$ 替代了知识点 14.2 的估计方程中的 $H(\psi)$ 。更多讨论请参见 Hernan (2005) 等人的论文及其附录。

第十七章图表

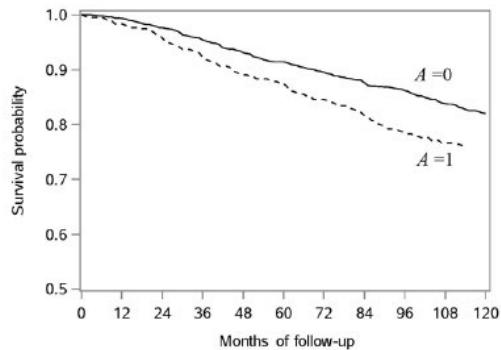


Figure 17.1

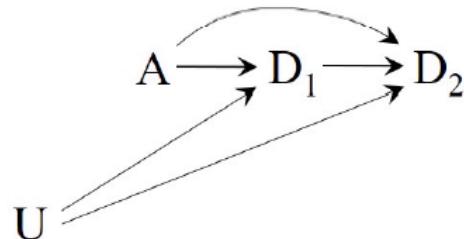


Figure 17.2

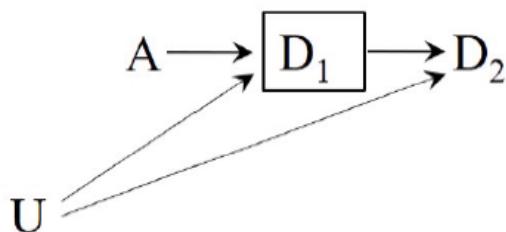


Figure 17.3

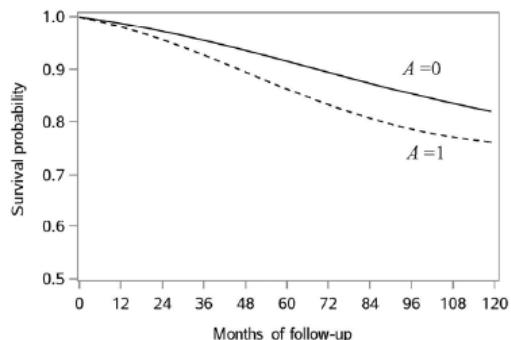


Figure 17.4

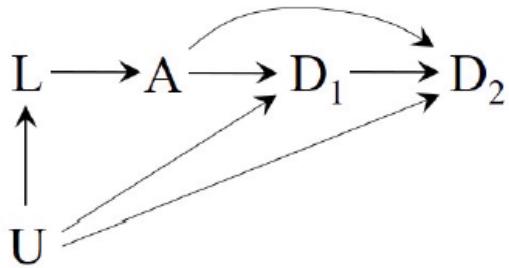


FIGURE 17.5

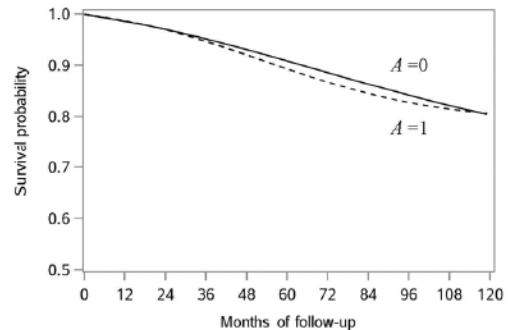


Figure 17.6

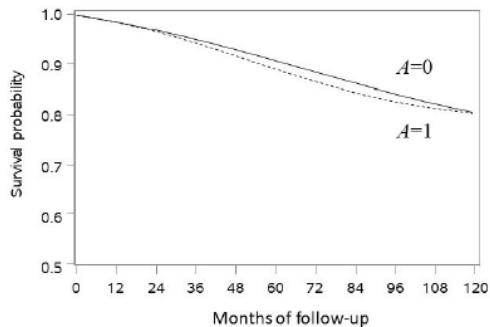


Figure 17.7

Type	1	2	3
$T^{a=0}$	36	72	108
$T^{a=1}$	24	48	72

Table 17.1

第十八章 因果推断中的变量选择

223 在前几章, 我们讲述了几种用来估计治疗效应的不同方法, 包括结局回归、逆概率加权、G-公式、以及 G-估算。每一种方法都有自己的特点, 但所有这些方法都会遇到同一个问题: 我们调整的变量 L 是否足以保证互换性。

在实践中, 这一问题表现为我们应该选择什么样的变量 L 进行调整。本章将在因果推断的框架下, 讨论变量选取的标准与注意事项。因为因果推断和预测有着本质区别, 所以用于预测模型的变量选择方法——也是现在常用的方法——不一定适用于因果推断。本章将讨论在因果推断中什么样的变量选择方法是错误, 并给出变量选择的指导意见。

18.1 变量选择的不同目的

我们在本书反复强调过, 有效的因果推断需要我们尽可能调整混杂和其他偏移。如果治疗 A 和结局 Y 之间的相关性可以部分 (或完全) 被混杂 L 解释, 我们就需要在数据分析中调整这些混杂变量。否则, 我们得到的相关性量度就不能被阐释为因果效应。

(即使模型调整了所有 A 和 Y 之间的混杂, 我们在模型中得到的混杂与结局之间的相关性也不一定有因果意义, 因为我们没有调整混杂和结局之间的混杂。)

但如果我们的目标是纯粹性的预测, 那就不一定需要调整混杂。如果我们只是想量化戒烟 A 和增重 Y 之间的相关性, 我们只需要比较戒烟和不戒烟者的体重变化即可。或者, 如果我们想用模型预测增重大小, 我们只需要在模型中放入能预测体重变化的协变量即可。我们不必知道这些变量是否是混杂变量。在预测模型中, 我们不需要对参数估计给出因果性阐释, 因此我们也就不再需要调整混杂, 甚至可以说, 此时不存在混杂这一概念。

(谨记: 混杂是因果推断中的概念, 只适用于因果效应而非相关性。)

我们在第 15.5 小节已经讨论过预测模型 (或相关性模型) 与因果模型之间的区别。比如, 临床研究者能用结局回归确定什么样的病人是高风险病人。此时研究者的目标是分类, 是预测的一种形式。预测模型的参数不需要有因果性意义, 所有参数的地位都是一样的, 也即不存在治疗 A 与协变量 L 的区别。再比如。过去是否住院可能是预测心衰的有效变量, 但是没有人会说为了防止心衰, 我们应该尽可能让病人不住院。如何判断一个病人预后是否良好 (预测问题) 和如何判断预防疾病的最有效方法 (因果问题) 是两个截然不同的问题。

对纯粹的预测而言, 研究者会选择能提高预测能力的任何变量。某些自动选择变量的算法——比如 lasso——适用于预测性的回归模型 (参见精讲点 18.1), 而某些则用于回归以外的方法 (比如神经网络)。所有这些算法都需要用到交叉验证 (或称交互效度分析)。参见精讲点

18.2），从而提高预测的精确性。因为大多选择算法其实是一个“黑箱”过程，所以我们并不能解释它们是怎么选择变量的，也不知道它们为什么选择了这一变量。然而这并不重要。对于预测性模型来说，只要能提高预测能力，那就是好的。

与之相比，在因果推断中，我们需要认真选择混杂变量，这样才能对治疗的效应估计进行因果性阐释。自动选择变量的算法也许适用于预测模型，但基本上不适用于因果推断的模型。已有的变量选择算法可能会造成偏移，本书之前已经讨论了其中的部分原因。下一小节，将会对所有原因进行汇总并讨论。

225 18.2 造成偏移或放大偏移的变量

假设我们的算力无限，数据中的研究个体（行数）近似无限，并且数据中各个体都有许多已测变量（列数），包含治疗 A 、结局 Y 、以及一大堆其余变量 X 。 X 中的部分可能是 A 和 Y 之间的混杂变量。此时我们没有算力或数据的限制，想调整多少变量就调整多少变量。

如果我们想无偏地估计 A 对 Y 的因果效应，也即 $E[Y|A=1] - E[Y|A=0]$ ，那么我们调整变量的目标就是尽可能消除混杂。我们能用回归、逆概率加权、G-公式等方法调整所有变量 X 。不过我们有必要调整所有的 X 吗？答案是否定的。在某些情形下，我们并不希望在模型中放入能造成偏移的变量。下面的例子将会解释这个问题。

（如果用分层调整变量，那即使不存在混杂，每个分层中的效应估计依然可能和未分层的效应估计不一样。此外这还涉及我们讨论过的可伸缩性。参见精讲点 4.3。）

L 是 X 的一个子集。假设我们问题中的因果结构和图 18.1 一致（也即和图 7.7 一致），其中 L 是对撞变量。因为 L 不是混杂变量，所以 $E[Y|A=1] - E[Y|A=0]$ 是 $E[Y|A=1] - E[Y|A=0]$ 的无偏估计。如果我们用 G-公式调整 L ，得到

$$\sum_l E[Y|A=1, L=l] \times \Pr[L=l] - \sum_l E[Y|A=0, L=l] \times \Pr[L=l],$$
 那这一估计就不等于

$E[Y|A=1] - E[Y|A=0]$ ，而是一个有偏的估计。这是因为 L 不仅在给定 A 的情况下与 Y 相关，同时也边缘性地和 A 相关，所以有 $\Pr[L=l] \neq \Pr[L=l|A]$ 。此时，即使 A 对 Y 没有因果效应，调整了 L 得到的 $A-Y$ 相关性也不会是零，我们称为零值下的选择偏移。如果因果结构如图 18.2 所示，那调整 L 会导致同样的偏移，因为此时 L 是对撞变量的下游变量。我们在第八章详细讨论过对撞变量以及它们的下游变量。

如果 L 不是对撞变量但也受到治疗影响, 如图 18.3 所示, 那调整 L 也可能导致选择偏移。此时 $E[Y|A=1]-E[Y|A=0]$ 依然是 $E[Y^{a=1}]-E[Y^{a=0}]$ 的无偏估计。然而, 如果我们调整了 L , 那我们得到的是一个有偏估计, 理由和上一段一样。如果没有从 A 到 Y 的箭头, 也即 $E[Y^{a=1}]-E[Y^{a=0}]=0$, 那不管是否控制 L , A 和 Y 都会相互独立。此时调整 L 得到的估计就是无偏的。也就是说, 与图 18.1 与 18.2 不同在于, 图 18.3 中, 只有当 A 对 Y 的因果效应不为零, 调整 L 才会得到有偏的估计。我们称之为脱离零值(参见第 6.5 小节)。

如果我们调整的变量受到 A 影响, 但这个变量会进一步影响 Y , 那我们把这个变量称为中介变量。在图 18.4 中, L 就是 A 和 Y 之间的中介变量。如果我们调整了 L , 或者 L 的下游变量, 那得到的 $A-Y$ 相关性就不是 A 对 Y 的总效应, 因为调整 L 会阻断途经 L 的效应。如果我们想估计的是 A 对 Y 的总效应, 那我们就应该过度调整中介变量。

(在图 18.4 中, 调整 L 会阻断路径 $A \rightarrow L \rightarrow Y$, 但不会阻断 $A \rightarrow Y$ 。因而, 我们调整 L 得到的相关性是 A 对 Y 直接效应的无偏估计, 而这一直接效应不经过 L 。)

上述讨论的情形都是调整一个变量可能会造成偏移。这些变量都有一个共同特征: 它们都受到治疗的影响, 因而它们都是治疗之后的变量。也许有人就会因此认为我们不应该调整出现在治疗之后变量。因为在实践中我们通常都知道各变量和治疗的时间先后顺序, 所以这一想法能够很容易落实。然而, 这一想法会遗漏一些原本应该被调整的变量, 我们在精讲点 7.4 中讨论过这一问题。以图 18.5 为例, 变量 L 在出现在治疗之后, 但我们可以用它阻断 A 和 Y 之间的后门路径。因此, 调整 L 我们就能得到无偏的 $A-Y$ 关系。而遵循上述想法不调整 L 的话, 只能得到有偏的估计。因此, 当 A 不影响 L 的时候, 不管 L 在时序上出现在 A 之前或之后, 我们的分析方法都应该是一样的。

不过问题是, 即使我们知道变量之间的时序关系。我们依然不能从数据中知晓 A 是否影响 L 。²²⁷ 实际上, 就算 A 、 L 、 Y 之间是某一特定时序关系, 不存在独立性的 (A, L, Y) 联合分布也依然满足多个因果关系图。因此, 我们需要借助数据之外的信息去决定是否调整 L 。也就是说, 我们不能用自动选择变量的方法决定是否调整 L , 因为自动选择变量的方法只依赖于数据之间统计上的关系。比如, 我们在第七章讨论过, 从数据相关性中我们不能区分混杂变量和对撞变量。因此, 我们只能凭借专业知识(如果有的话)判断调整一个变量是否会引入混杂。

接下来我们会讨论出现在治疗之前的变量 L , 也即此时我们的时序顺序是 LAY 。为了简便, 假设我们的样本足够大, 远大于协变量 X 的数目。因此, 误差就可以被忽略, 我们只需关心是否

存在偏移。此时, 所多人认为调整治疗之前的变量会减小偏移。然而这一想法也是错误的, 主要有两个原因。

第一个原因, 图 18.6 中, 变量 L 出现在治疗之前。在一条连接 A 和 Y 的路径中, L 是一个对撞变量, 因此调整 L 会引入选择偏移, 我们在第七章将这种情况称为 M 偏移。再一次注意, 我们不能从数据中区分混杂变量和对撞变量, 因此我们只能依赖数据之外的信息决定是否调整 L 。事实上, 如果有一个从 L 指向 A 的箭头 (如图 18.6 所示), 那么 L 既是混杂变量, 也是对撞变量, 也就意味着不管调不调整 L , 我们都不能无偏地估计因果效应均值。

另一个原因是无差别调整出现在治疗之前的变量可能会导致偏移放大, 我们在本书中尚未讨论这一问题。在图 18.7 (与图 16.1 相同) 中, A 和 Y 之间有一个未测的混杂变量 U 。因为数据中没有 U , 所以我们就不能调整 U , 也就不能消除所有混杂。而调整图中的 Z 不仅不能消除混杂 (因为 Z 并不在 A 和 Y 的后门路径上。其实 Z 是一个工具变量, 我们在第十六章讨论过), 还可能放大因 U 导致的混杂。也就是说, 调整 Z 后的 $A-Y$ 相关性, 可能比不调整 Z 的 $A-Y$ 相关性更偏离真实值。这是因为我们调整了一个工具变量, 这也被称为 Z 偏移。然而并不是调整 Z 都会导致偏移放大, 有时调整 Z 也会使得偏移减小。一般而言, 我们并不知道调整工具变量是会放大偏移, 还是减小偏移。

(在因果图对应的结构方程模型中, 如果所有方程都是线性的, 那么就一定会出现放大偏移。)

总而言之, 即使我们没有算力和样本大小上的限制, 我们也不应该调整所有的变量 X 。理想情况是, 我们调整的变量不会造成偏移或放大偏移。因为我们不能用数据实证地找出会造成偏移的变量, 所以我们需要与研究课题相关的专业知识指导我们选择变量。
228

(Hernan (2002) 等人在论文中举例说明了如何用专业知识指导变量选择。)

18.3 因果推断与机器学习

(本章接下来三个小节是 James Robins 在 2018 与 2019 年于波士顿、柏林、鹿特丹等地的学术演讲的一个大致概括。)

在本章剩下部分, 我们假设 X 不包括会造成偏移、放大偏移的变量, 同时还包括了所有的混杂变量 L 。那我们接下来的问题是, 在 X 是一个高维向量的时候, 如何在实际中无偏地估计因果效应 $E[Y^{a=1}] - E[Y^{a=0}]$ 。

如何利用 X , 将会因不同的调整方法而不同。当使用插入式 G-公式 (也即标准化) 的时候, 我们在控制了 X 之后估计结局 Y 的均值, 我们把它记为 $b(X)$; 当使用逆概率权重的时候, 我们是在控制了 X 之后估计治疗 A 的概率, 我们把它记为 $\pi(X)$ 。我们可以通过参数模型 (比如线性回归和 logistic 回归) 得到估计值 $\hat{b}(X)$ 和 $\hat{\pi}(X)$ 。为了减少模型错误设定的可能性, 我们会在模型中加入大量的乘积项, 并使用非线性的函数形式 (比如立方样条)。

在实践中, 使用传统的参数模型会遇到一个严重的问题: 模型中的参数数目可能远大于我们的样本量。此时, 我们得到的估计值就会非常的不精确, 甚至模型不能收敛, 从而得不到任何估计值。我们在 15.5 小节中也讨论过, X 可能会包含不是混杂变量的变量, 但这些变量与治疗 A 强烈相关。因此, 把这些变量全部放入 $\pi(X)$ 的模型中, 可能会导致正数性不成立。

(X 中的某些变量不是混杂变量, 我们不需要在模型中放入这些变量。)

一个可行的方法是用 lasso (参见精讲点 18.1) 拟合参数模型, 而 lasso 是在预测性模型中选择变量的一种算法。另一个可行的方法是用机器学习算法估计 $b(X)$ 和 $\pi(X)$, 可选的机器学习算法包括随机森林、神经网络等。在高维数据中, 这些机器学习算法能够同时考虑上千个参数, 从而相较于传统的参数模型, 机器学习更能精确地预测结局。不过使用机器学习可能会导致两个问题。

(在机器学习中, 我们可以使用交叉验证 (参见精讲点 18.2) 优化预测的准确性。)

第一个问题, 当仅估计 $b(X)$ 或 $\pi(X)$ 的时候, 机器学习不能保证消除所有混杂。一个改良做法是使用双重稳健估计。双重稳健估计会同时考虑 $b(X)$ 和 $\pi(X)$, 并将这两个估计值综合起来。
229 $b(x)$ 的误差是 $b(x) - \hat{b}(x)$, $\frac{1}{\pi(x)}$ 的误差是 $\frac{1}{\pi(x)} - \frac{1}{\hat{\pi}(x)}$, 而双重稳健估计的误差是这两个的乘积。因此, 如果机器学习能精确地估计 $b(X)$ 和 $\pi(X)$, 那么双重稳健估计就能给出更小的误差。

(双重稳健估计的相关性质被称为二阶偏移。参见知识点 13.2。)

第二个问题, 机器学习算法是一个黑箱, 我们并不知道其中的统计原理。也就是说, 即使双重稳健估计是无偏的, 估计值的方差也可能是错的。因此, 我们得到的置信区间在传统频率学派看来没有任何意义。尤其是机器学习给出的 95% 置信区间一般都会较窄, 并不能在至少 95% 的时间中涵盖真实值, 因此是无效的。

(如果超级人群中存在一定的混杂, 那么置信区间就更加不可信。参见第十章。)

因此, 将因果推断和机器学习结合起来的关键是双重稳健估计, 但这还不够。下一小节将讨论如何两种改良双重稳健估计的方法: 样本分割和交叉拟合。

18.4 机器学习中的双重稳健估计

让我们假设使用双重稳健估计的机器学习算法能够给出较小的偏移。较小的偏移意味着估计值的偏移小于它的标准误差。或者说, 这一偏移要小于 $1/\sqrt{n}$ 。我们简单说过为什么双重稳健估计一般会给出更小的偏移, 因为双重稳健的偏移是两个误差 $b(x) - \hat{b}(x)$ 和 $\pi(x) - \hat{\pi}(x)$ 的乘积。就算其中一个误差大于 $1/\sqrt{n}$, 通过双重稳健估计得到的偏移也可能足够小, 从而用来得到有效的置信区间。

在随机大样本中, 某些条件下, 我们可以假设一个一致的双重稳健估计遵循正态分布, 其均值是参数的真实值。也就是说, 我们能用很小的偏移给双重稳健估计构建一个有效的置信区间。不过只有当双重稳健估计包含样本分割和交叉拟合的时候, 这一假设才为真。

我们先说样本分割。首先, 我们将大小为 n 的样本分为两部分, 一部分用来估计, 另一部分用来训练, 每部分的样本大小都是 $n/2$ 。其次, 我们在训练样本中使用预测性算法, 从而得到结局的估计值 $\hat{b}(x)$ 和治疗的估计值 $\hat{\pi}(x)$ 。接下来, 我们在估计样本中用之前得到的 $\hat{b}(x)$ 和 $\hat{\pi}(x)$ 进行双重稳健估计, 从而得到因果效应的估计值。如此一来, 我们就可以在一半研究人群中, 用机器学习和双重稳健估计得到因果效应的估计值。

(我们也可以把训练样本称为冗余样本, 因为我们用它做 $b(x)$ 和 $\pi(x)$ 的冗余回归。冗余参数的概念参见精讲点 15.1。)

样本分割能让我们在一半的研究人群中用标准统计方法得到有效的估计值和有效的置信区间。但是, 我们失去了另一半研究人群。因此, 我们的置信区间比起我们用整个人群而言就会宽上许多。解决这一问题的办法是交叉拟合。

230 接下来我们讨论如何用交叉拟合恢复因样品分割而丢失的统计效力。首先, 我们重复上述步骤, 只不过将估计和训练两部分互换。也就是说, 原先用来训练的现在用来估计, 原先用来估计的现在用来训练。这样一来, 我们就可以在另一半研究人群中, 用机器学习和双重稳健分析得到因果效应的估计值。

(样本分割和交叉拟合并不是新鲜事。不过直到近年, 这两部分才在机器学习中被广泛应用。)

接下来是计算研究人群不同两部分双重稳健估计的均值。这一均值就是整个人群的因果效应均值的估计值。它的 95% 置信区间可以用自举法得到。

这就是我们要做的全部步骤。通过样本分割和交叉拟合，我们能将双重稳健估计和机器学习结合起来，从而得到因果效应的估计值。这一过程的统计性质简明易懂，并且用到了所有的数据。现在的一个研究前沿是设计新方法检验双重稳健估计的偏移是否过大，如果过大，如何得到偏移更小的估计值。

18.5 变量选择永远是一个难题

上一小节的方法告诉我们，并不是所有基于数据的变量选择方法都会导致不可信的置信区间。如果我们能将机器学习和因果推断结合起来，在某些条件下，我们依然能得到正确的估计值。然而，结合了双重稳健估计的机器学习依然不能解决我们的所有问题。主要有以下四个方面的原因。

第一，在许多实际应用中，我们的专业知识不足以找出所有的混杂变量，也不足以排除所有能造成偏移或放大偏移的变量。因此，我们不能 100% 保证机器学习中的双重稳健分析总能够给出较小的偏移。

第二，虽然我们已经有了许多机器学习算法，但是没有一个算法适用于所有情境。在数学上，我们不能证明某个算法就一定优于另一个算法。对于算法的选择需要借助于变量之间的因果结构，然而真实的因果结构可能是不可知的，或者难以适用于实际应用。

第三，在高维变量与时异治疗的情形中，使用双重稳健估计非常麻烦，而结合机器学习则会进一步增加对算力的要求，尤其是对生存分析而言。因此，许多复杂纵向数据的因果推断不会用双重稳健估计，只会用单一稳健估计。我们将在本书第三部分讨论。

第四，虽然结合双重稳健估计的机器学习给出的误差可能等于知道真实值时的误差，但是这一方法不能保证误差足够小从而我们能从中得到有意义的因果推断。

230 可能我们用双重稳健机器学习得到了一个估计值，但会发现估计值的误差太大以致于我们得不到有效的结论。在 X 的某些变量和治疗 A 强烈相关的时候，这种情形时常发生，即使我们所有步骤都是正确的。此时，治疗概率 $\pi(X)$ 的估计值，可能在某个 X 值下非常接近 0 或 1。因此，我们的估计值就会有一个很大的误差，而它的 95% 置信区间虽然正确，但却很宽以致于我们得不到任何有用信息。因为我们并不喜欢一个很宽的置信区间，所以即使它是正确的，我们也会认为导致这一结果的变量有问题，从而将这些变量排除于模型之外。如果我们这么做了，我们将彻底改变变量选择的规则——我们抛弃了某些和治疗相关的变量，从而我们不再能保证我们的 95% 置

信区间是有效的。把所有变量都放入模型中从而消除偏移，与排除某些变量从而减小误差，这两者之间有不可调和的矛盾。

(这一结果会引起另一费解的哲学问题：如果我们排除几个（比如 5 个）变量后置信区间就是无效的，那如果一开始就没有这 5 个变量，我们的置信区间还是有效的吗？实际上，我们的数据不可能包含所有的混杂变量，我们应该如何阐释观察性研究中的置信区间？)

因为以上所有问题，给出一个普适的变量选择方法基本上是不可能的。实际上，某些研究者正致力于方法论的开发，从而解决本章提到的种种问题。我们能给出的最科学建议是多进行敏感性分析：使用不同的分析方法，然后检视不同方法得到的效应估计。如果这些效应估计并不冲突，那我们就会对我们的结果有所信心。如果这些效应估计相互冲突，那我们就需要去探索为什么它们相互冲突。

第十八章精讲点和知识点

精讲点 18.1：回归模型中不同的变量选择方法（原书第 224 页）

假设我们想拟合一个用于预测的回归模型，但是我们的数据中又有许多变量，变量数甚至可能超过样本人数。如果把这些变量都放进模型里，势必会使得模型不稳定。统计学者设计出多种用于选择变量的方法。许多书中都有这些方法的详细介绍。接下来我们会简述其中几种。

一种思路是选择这些变量中的部分。其中一个简单方法是事先确定模型中应该有多少个变量，然后在这个固定数目之下，尝试不同的组合，直到选出使得我们的选择准则（比如赤池信息准则，AIC）最优的组合。不过，当有大量变量的时候，这一方法在算力上几乎不可能实现。另其他方法包括前向选择法（从没有任何变量开始，在接下来的每一步中，往模型中放入一个能使得模型最优的变量），后向消元法（从所有变量都放入模型开始，在接下来的每一步中，去掉一个对模型最没有影响的变量），以及逐步筛选法（前向选择法和后向消元法的综合）。这些变量选择算法在不能再改进模型时结束，而改进模型的定义将由我们事先确定的准则决定。这些算法实施起来很简单，不过也不能穷尽所有变量的组合。

另一种思路是缩减。这一思路是在估计方法中加入一项“惩罚”，使得除了截距项的参数估计更接近于零。也就是说，参数估计被缩减。收缩会使得方差减小且预测更加稳定。最常用的缩减方法是岭回归（也译为嵴回归）和 lasso（英文全称：least absolute shrinkage and selection regression）。和岭回归不同，lasso 允许某些参数的值取零。因此，lasso 既是一个缩减的方法，又是一个选择部分变量的方法。在现实应用中，lasso 被大量用于回归模型的变量选择。就预测准确性而言，lasso 由于逐步筛选法。

精讲点 18.2: 过拟合与交叉验证 (原书第 225 页)

过拟合是变量选择的一个常见问题: 我们用变量尽可能精确地预测观测数据, 而不考虑某些观测数据的误差纯粹是随机的。从而会使得我们的模型能够非常精确地预测用来拟合模型的数据, 但是却基本不能预测未用来拟合模型的数据。这一问题在随机森林、神经网络、以及其他机器学习算法中都可能出现。

解决过拟合的一个直接方法是把样本分割成两部分, 一部分被称为训练样本, 用来拟合模型, 另一部分被称为验证样本, 用来衡量模型的准确性。比如我们的样本大小为 n , 我们用 v 作为验证样本, 那么其中的 $n - v$ 就会被用作训练样本。当我们使用 lasso 的时候, 训练样本中的缩减程度也应该根据验证样本中模型的准确性进行调整。

分割样本的一个显然缺陷是我们只在一部分样本中拟合模型, 从而增加了误差。一个解决办法是多次重复样本分割, 从而增加用于模型拟合的人数。然后我们就可以将所有验证样本中的均值视作我们模型的准确度。这一过程被称为交叉验证, 或样本外检验。交叉验证也有不同的形式与实现方法。

一种叫做“样本为 v 的交叉验证”是从总样本中分出大小为 v 的验证样本, 找出所有可能的样本组合, 然后再分析这些所有可能组合。不过样本总量 n 以及 v 过大时, 这一方法在算力上很难实现。因此, 一种常见方法使 v 的取值为 1, 这就被称为“样本为 1 的交叉验证”。如果依然需要大量计算, 那我们可以考虑不使用所有组合。比如, “ k 重交叉验证”是指我们将样本分为大小相等的 k 组, 其中一组用作验证, 另外 $k - 1$ 组用作训练。然后这一过程重复 k 次, 每一次的验证样本都不相同。

第十八章图表

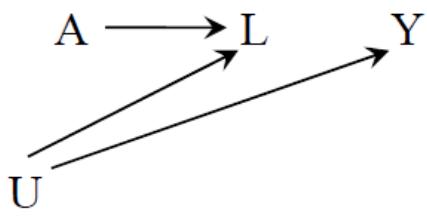


Figure 18.1

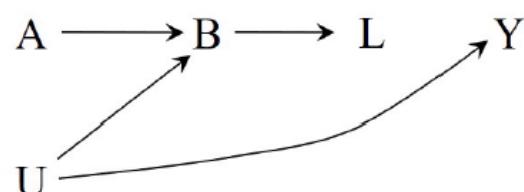


Figure 18.2



Figure 18.3

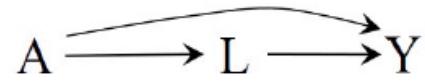


Figure 18.4

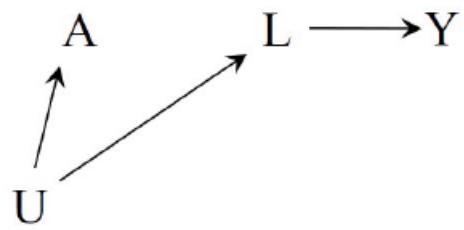


Figure 18.5

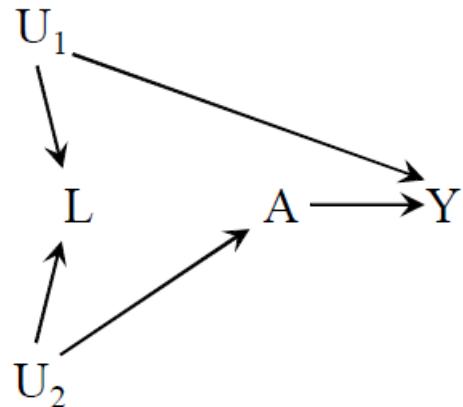


Figure 18.6

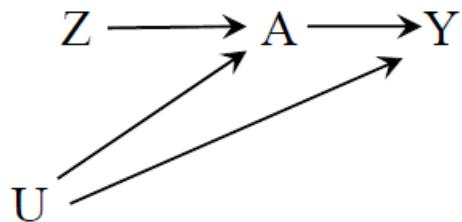


Figure 18.7

Causal Inferences: What if ——第十八章
作者：Miguel A. Hernan, James M. Robins;
翻译：罗家俊

第三部分

复杂纵向数据的因果推断

Causal Inferences: What if ——第十九章
作者：Miguel A. Hernan, James M. Robins;
翻译：罗家俊

第十九章 时异治疗

235 迄今, 本书只讨论了不随时间变化的固定治疗(或暴露、干预)。然而现实研究中, 许多变量都是因时而异的, 比如生活习惯、工作内容、婚姻状态等都会因时而异。因此, 治疗变量可能在不同时间有不一样的取值, 我们将此称为时异治疗。

在本书第一和第二部分, 我们只讨论了不随时而动的(也即非时异)治疗变量, 这有助于我们理解因果推断的基本概念和分析方法。现在, 我们需要思考更加现实的因果推断问题了, 也即时异治疗的因果效应。本书第三部分会将前两部分的内容延伸到时异治疗中。本章将会介绍时异治疗的相关术语及概念。虽然我们尽可能简化了这些概念(如果你不信, 你可以亲自看一看因果推断的相关文献), 但是本章仍然是本书最晦涩的几章之一。遗憾的是, 再进一步的简化可能会得不偿失。不过既然你都读到本章了, 那你肯定能理解本章的种种概念。

19.1 时异治疗的因果效应

假设在研究开始之时, 我们有一个不因时而异的治疗变量 A (1: 接受治疗, 0: 未接受治疗), 在 60 个月后, 我们有一个结局变量 Y 。我们将 A 对 Y 的因果效应定义为所有人都接受治疗时的反事实结局 $Y^{a=1}$, 以及所有人都未接受治疗时的反事实结局 $Y^{a=0}$ 两者之间的对比, 也即 $E[Y^{a=1}] - E[Y^{a=0}]$ 。因为所有人的治疗状态在唯一一个时间点(研究开始时)就确定了, 所以我们就没必要考虑治疗究竟出现在什么时候。而与此不同的是, 涉及时异治疗的因果对比需要我们将时间点纳入考虑的范围之内。

为了明白这一点, 让我们假设有一个时异性的二分治疗变量 A_k , 它出现在第 k 个月的每次访问中, k 的取值从 0 到 59。比如, 在一项长达 5 年的 HIV 感染人群研究中, 如果某人在第 k 个月接受了抗逆转录病毒疗法, 那么 A_k 取值为 1, 否则为 0。因为在研究开始之前没有人接受这一疗法, 所以对所有人都有 $A_{-1} = 0$ 。

(为了简化, 我们会暂时假设这项研究没有失访或者人员死亡, 我们同时也会假设所有变量都是完美测量的。)

我们在字母上用一横杆表示既往历史, 也即 $\bar{A}_k = (A_0, A_1, \dots, A_k)$ 表示从第 0 个月到第 k 个月的治疗情况。当我们表示整个研究阶段(共 K 个月)的治疗史时, 我们通常会忽略下角标, 从而用 \bar{A} 代替 \bar{A}_K 。因此, 如果某人在整个研究阶段都没有接受过治疗, 那么他的治疗史就可以表示为

$\bar{A} = (0, 0, \dots, 0) = \bar{0}$ 。大部分人都会在研究中的某个阶段接受治疗, 因此每个人的治疗史中总有一些 1 或 0, 从而我们也就不能用 $\bar{0}$ 或 $\bar{1}$ 表示他们的治疗史。

(为了兼容其他相关文献, 我们用 0 表示起始时间。也就是说, k 的第一个取值是 0, 而不是 1。)

236 假设 Y 衡量的是研究结束时, 第 $K+1=60$ 个月的健康状态, 取值越大表示越健康。那么我们想估计的因果效应是时异治疗 \bar{A} 对结局 Y 的因果效应均值。此时, 我们就不能再把因果效应定义为时异变量在某单一时间点 k 的对比, 因为这一对比, 也即 $E[Y^{a_k=1}] - E[Y^{a_k=0}]$, 表示的是在某个时间点 k 治疗 A_k 的效应, 而不是时异治疗 A_k 在所有时间点 (0 到 59) 的效应。

(为了简化, 我们只考虑在某个固定时间点测量的结局。不过本章所讨论的概念适用于时异结局和失效时间结局。)

实际上, 我们只能将因果效应定义为整个研究阶段内不同治疗策略下反事实结局的对比。因而, 时异治疗的因果效应的定义并不是唯一的。在下一小节, 我们将讨论时异变量的因果效应的不同定义。

(谨记: 我们用小写表示随机变量的现实情况。比如, a_k 是 A_k 的现实情况。)

19.2 治疗策略

治疗策略, 有时也被称为治疗计划、治疗方案等, 指的是在时间点 k 应该进行治疗还是不进行治疗的一系列规则。比如, 某研究有两种治疗策略, 一种是“肯定会治疗”, 另一种是“从不会治疗”。肯定会治疗被表示为 $\bar{a} = (1, 1, \dots, 1) = \bar{1}$, 从不会治疗则是 $\bar{a} = (0, 0, \dots, 0) = \bar{0}$ 。因而, 我们可以将 \bar{A} 对 Y 的因果效应定义为“肯定会治疗”与“从不会治疗”这两种策略所对应的反事实结局的对比, 也就是 $E[Y^{\bar{a}=\bar{1}}] - E[Y^{\bar{a}=\bar{0}}]$ 。

(Robins (1896, 1987, 1997a) 首次在理论上讨论如何对比不同治疗策略的反事实结局。)

但对于时异治疗 \bar{A} 而言, 还有许多不同治疗策略, 也就还有许多不同策略下的对比。比如, “隔一个月治疗一次”策略 $\bar{a} = (1, 0, 1, 0, \dots)$, 以及“仅第一个月不治疗”策略 $\bar{a}' = (0, 1, 1, 1, \dots)$ 之间的对比, 也即 $E[Y^{\bar{a}}] - E[Y^{\bar{a}'}]$ 。不过这样的对比实在是太多了, 对于二分治疗 a_k 而言, 我们有 2^K 种不同组合, 也就有 2^K 种不同策略, 而我们接下来会说明, 这 2^K 种策略还不是所有的可能治疗策略。

假设我们有一个 HIV 的研究, 时异协变量 L_k 表示第 k 个月的 CD4 细胞数 (单位: 个/ μL)。

CD4 细胞数很低时, L_k 取值为 1, 表示较差预后, 反之为 0。在最开始, 所有人的 CD4 细胞数都很高, 也即 $L_0 = 0$ 。我们可以设计一种治疗策略, 在 $L_k = 0$ 时不进行治疗, 而在 $L_k = 1$ 时, 将在之后的时间内都给予治疗。这一治疗策略和上一段的有所不同, 因为此时我们不能把这一治疗策略用简单的 $\bar{a} = (a_0, a_1, a_2, \dots, a_K)$ 表示, 因为其中的 a_k 表示所有人在 k 时都会接受同样的治疗 a_k 。此时, 在每一个时间点, 治疗与否取决于不断变化的 L_k 。如果 k 时的治疗 a_k 取决于时异协变量 L_k , 那就被称为动态治疗策略。而不取决于 L_k 的策略 \bar{a} 则被称为非动态或静态治疗策略。

237

时异治疗的因果推断会涉及两个或多个治疗策略的反事实结局的对比。只有治疗策略是良定的, 时异治疗的因果效应才可能是良定的。在我们 HIV 的例子中, 我们可以将因果效应定义为策略 \bar{a} (比如“肯定会治疗”) 和策略 \bar{a}' (比如“从不会治疗”) 之间的对比 $E[Y^{\bar{a}}] - E[Y^{\bar{a}'}]$, 或者策略 \bar{a} (“肯定会治疗”) 和策略 g (比如“在 CD4 细胞数降低后给予治疗”) 之间的对比 $E[Y^{\bar{a}}] - E[Y^g]$ 。我们经常会用 g 表示各种不同的策略, 可能是静态的, 也可能是动态的。当我们用它表示静态策略的时候, 我们一般会写作 $Y^{g=\bar{a}}$, 而非 $Y^{\bar{a}}$ 或 Y^g 。

也就是说, 对于时异治疗而言, 没有一个唯一的因果效应定义。即使在每个时间点只存在两种治疗情况 (治疗或不治疗), 我们依然可以根据不同治疗策略定义出许许多多不同的因果效应。在下一小节, 我们将介绍时序性随机试验, 一种能有效估计这些因果效应的研究设计。

19.3 时序性随机试验

238

图 19.1、19.2、以及 19.3 汇总了涉及时异治疗的研究所对应的三种因果结构。在这三幅图中, A_k 表示时异治疗, L_k 表示已测的变量, Y 表示结局, U_k 表示 k 时的未测变量, 并且是图中至少两个其他变量的共同诱因。因为 U_k 是未测的, 所以它的取值是未知的, 也就不能用于数据分析。在我们 HIV 的例子中, L_k 表示时异变量 CD4 数目, 它是免疫系统状况 U_k 的一个直接结果, 而 U_k 是未测的。如果一个人的免疫系统状况 U_k 越糟糕, 那么他的 L_k 也就越低, 结局 Y 的取值也会越低。为了简化, 我们的因果图只有 $k = 0$ 和 $k = 1$ 两个时间点, 同时我们也假设所有的研究参与者都遵循了分配给他们的治疗方案。

(根据定义, 一幅因果图需要包含图中任意两个变量的共同的诱因。)

在图 19.1 中, 没有从 \bar{L}_k 或 \bar{U}_k 指向治疗 A_k 的箭头。在图 19.2 中, 存在从 \bar{L}_k 指向 A_k 的箭头, 但不存在从 \bar{U}_k 指向 A_k 的箭头。在图 19.3 中, 存在从 \bar{L}_k 或 \bar{U}_k 指向 A_k 的箭头。

图 19.1 可以用来描述一项随机试验, 其中在每个时间点接受治疗 A_k 的概率只取决于上一个时间点的治疗情况。在我们的 HIV 研究中, 如果某人在前一个时间点没有接受治疗 ($A_{k-1} = 0$), 那么他在这个时间点接受治疗的概率是 0.5; 反之, 如果在前一个时间点接受了治疗 ($A_{k-1} = 1$), 那么在这个时间点接受治疗的概率是 1, 那此时这项随机试验就能用图 19.1 表示。图 19.1 表示的是一种静态治疗策略, 其他变量——不管是已测的还是未测的——都不会对时异治疗造成混杂。在这一情况下, 每个人都遵循治疗策略 \bar{a} 时的反事实结局 $E[Y^{\bar{a}}]$, 就等于实际人群中遵循了这一策略 \bar{a} 的所有人的结局均值 $E[Y | \bar{A} = \bar{a}]$ 。这一结论对动态治疗策略并不成立。受到 L 影响的动态治疗策略 g 所对应的反事实结局 $E[Y^g]$, 只有在 $A_k = 1$ 的概率是 0.5 时 (注意, 此时 A_k 受到 \bar{L}_k 的影响), 才会等于实际遵循策略 g 的人群的结局均值。否则, 我们需要在图 19.1 或 19.2 中使用 G-方法, 以及 \bar{L} 、 \bar{A} 和 Y 的相关数据, 才能识别 $E[Y^g]$ 。

图 19.2 也可以用来表示一项随机试验, 其中在每个时间点接受治疗 A_k 的概率取决于上一个时间点的治疗情况, 以及已测变量。在我们的 HIV 研究中, 假设我们根据三种不同情况对治疗概率赋值: 如果上一次未接受治疗且本次 CD4 数目较高 ($A_{k-1} = 0$, $L_k = 1$), 那概率是 0.4; 如果上一次未接受治疗且本次 CD4 数目较低 ($A_{k-1} = 0$, $L_k = 0$), 那概率是 0.8; 如果上一次接受了治疗 ($A_{k-1} = 1$), 那不管 CD4 数目多少, 给予治疗的概率都是 0.5。此时这项随机试验就可以用图 19.2 表示。在图 19.2 中, 已测的变量会对时异治疗造成混杂, 而未测的变量不会。

一项随机试验如果在每个时间点 k 都是随机地分配被试, 那这项随机试验被称为时序性随机试验。因此图 19.1 与 19.2 都可以用来表示一项时序性随机试验。而图 19.3 却不能用来表示时序性随机试验, 这是因为图 19.3 中每个时间点 A_k 的概率都会受到未测变量 U 的影响, 而 U 又是 L_k 和 Y 的诱因, 但是随机试验不可能用未测变量对治疗概率进行赋值。也就是说, 在时序性随机试验的因果图中, 任何时间点都不存在从未测变量 U 指向治疗 A_k 的箭头。

而在观察性研究中, 是否接受治疗常常取决于结局的某些预测因素。因此, 我们一般用图 19.2 或 19.3 表示观察性研究, 而不会用图 19.1。比如, 假设我们的 HIV 研究是一项长期跟踪研究, 那研究参与者的 CD4 数目越低, 就越可能接受治疗。因此, 如果这项研究的参与者仅仅用之

前的治疗史以及 CD4 数目 (\bar{A}_{k-1} , \bar{L}_k) 决定自己的治疗情况, 而不涉及其他未测变量 \bar{U}_k , 那么这项研究就能用图 19.2 表示。此时, 用图 19.2 表示的观察性研究, 和同样用图 19.2 表示的时序性随机试验, 唯一的区别在于接受治疗的概率是否是已知的 (不过我们依然可以用数据进行估算)。遗憾的是, 我们不能实证地证明一项观察性研究到底应该用图 19.2 表示, 还是用图 19.3 表示。用图 19.3 表示的观察性研究存在未测混杂, 我们接下来会讨论这一问题。

在实践中, 时序性随机试验并不常见。然而, 时序性随机试验这一概念有助于我们理解时异治疗因果效应估计所需的假设。下一小节将讨论这些假设。

240 19.4 时序互换性

在本书第一和第二部分我们已经讨论了因果推断中的有界互换性 $Y^a \perp\!\!\!\perp A|L$, 不过当时我们只讨论了非时异治疗的情形。当有界互换性 $Y^a \perp\!\!\!\perp A|L$ 成立的时候, 我们只需适当调整协变量 L , 就能得到治疗 A 对结局 Y 的无偏估计。我们可以用许多方法调整变量 L , 包括标准化、逆概率加权、G-估算、以及其他方法。我们认为在有界随机试验中, 有界互换性能够成立, 这是因为在一项随机试验中, 接受治疗的概率只取决于已知的协变量 L 。在观察性研究中, 如果控制了已测协变量 L 之后, 接受治疗的条件概率不再受其他未知变量的影响, 那么有界互换性成立。

同理, 涉及时异治疗的因果推断也需要调整每个时间点的时异协变量 \bar{L}_k , 从而保证有界互换性, 也即时序有界互换性。比如, 在只有两个时间点的研究中, 时序有界互换性就是在第一个和第二个时间点的有界互换性的总和, 也就是 $Y^g \perp\!\!\!\perp A_0|L_0$ 和 $Y^g \perp\!\!\!\perp A_1|(A_0 = g(L_0), L_0, L_1)$ 。(为了简便, 在本书我们会省去“有界”两字, 从而简单称为时序互换性。) 我们会将这一系列独立性表达式的集合称为 Y^g 的时序互换性, 其中的 g 可以是静态策略, 也可以是动态策略, 并且包含了时异治疗的所有成分 (在我们的例子中是 A_0 与 A_1 两个变量)。

(对于治疗策略为 g 的个体而言, 他们的实际治疗史 ($A_0 = g(L_0)$, $A_1 = g(A_0, L_0, L_1)$) 下的结局 Y , 不仅对应反事实结局 Y^g , 同时也对应 $a_0 = A_0$ 与 $a_1 = A_1$ 下的反事实结局。)

在时序性随机试验中, 时间点 k 处的治疗 A_k 只受到之前的治疗史 \bar{A}_{k-1} 与协变量史 \bar{L}_k 的影响, 这也就意味着对结局 Y^g 而言, 时序互换性成立。也就是说, 给定任意治疗策略 g , 我们只要控制了协变量史 \bar{L}_k 以及之前观察到的治疗史 $\bar{A}_{k-1} = g(\bar{A}_{k-2}, \bar{L}_{k-1})$, 那么在时间点 k 处, 对于 Y^g 而言, 接受治疗和未接受治疗的人群是可互换的。利用数学符号, Y^g 的时序互换性被表述为:

$$Y^g \perp\!\!\!\perp A_k | (\bar{A}_{k-1} = g(\bar{A}_{k-2}, \bar{L}_{k-1}), \bar{L}_k), \text{ 其中 } k = 0, 1, \dots, K$$

只要在因果图中没有从未测变量 U 指向治疗变量 A 的箭头 (如图 19.2)，那么这一形式的时序互换性 (当然还有其他形式，我们之后会介绍) 就会成立。因此，在时序性随机试验，以及治疗概率在控制治疗史和已测协变量史 (\bar{A}_{k-1} , \bar{L}_k) 后独立于其他结局预测因素的观察性研究当中，时序互换性成立。

(在图 19.1 中，结局 Y 的时序无界互换性成立，也即 $Y^{\bar{a}} \perp\!\!\!\perp A_k | \bar{A}_{k-1} = \bar{a}_{k-1}$ 对所有治疗策略 \bar{a} 都成立。无界互换性意味着相关性就是因果性，也即 $E[Y^{\bar{a}}] = E[Y | \bar{A} = \bar{a}]$ 。)

也就是说，如果我们能用图 19.2 表示一项观察性研究，那么我们就能识别并估计任一治疗策略 g 的反事实结局 $E[Y^g]$ 。但如果这项观察性研究只能用图 19.3 表示，那我们就不能识别反事实结局 $E[Y^g]$ 。如果这项观察性研究是用其他因果图表示的，那某些策略——而非所有策略——的反事实结局可以识别并估计。

(当我们谈论因果效应的识别时，我们用来识别因果效应的方法是 G -公式。只有在极少且与我们当前讨论不相关的情况下，我们会用与 G -公式相关，但又不完全等同于 G -公式的方法。)

比如，假设一项观察性研究如图 19.4 所示，其中有一个未测变量 W_0 。在我们 HIV 的例子中， W_0 可能是研究刚开始时的门诊情况，但没有被记录在数据库中。因而 W_0 就是最开始的治疗 241 A_0 以及第一个月的 CD4 测量数目 L_1 的一个共同诱因，而图中的 U_1 代表 CD4 的真实数目，但却是未知的。此时，虽然 W_0 是未测的，对于任意静态策略 $g = \bar{a}$ 而言，反事实结局 $E[Y^g]$ 是可识别的；但对于受到 L_1 影响的动态策略 g 而言，反事实结局 $E[Y^g]$ 却是不可识别的。为了解释为什么有的可识别有的不可识别，我们需要借助单一世界干涉图进行讲解，我们将在下一小节讨论。

除了不同时序互换性之外，时异治疗的因果推断还需要其他两个假设：扩展到时间序列上的正数性和一致性。在时序性随机试验中，时序正数性和时序一致性都会成立 (参见知识点 19.2)。在我们接下来的讨论中，我们会假设时序正数性和时序一致性成立。在这三个可识别假设下，只要我们能用适当的方法 (比如标准化、逆概率加权、和 G -估算) 调整治疗史和协变量史 (\bar{A}_{k-1} , \bar{L}_k)，我们就能识别治疗策略 g 的反事实结局 $E[Y^g]$ 。

19.5 部分治疗策略下的可识别性

在第七章，我们讨论了如何在一幅因果图中判断互换性是否成立，使用到了后门准则，但是当时只讨论了非时异治疗。后门准则完全可以扩展到时异治疗当中。比如，对于静态策略而言，

其因果效应可识别性的充分条件是: 在每个时间点 k , 指向 A_k , 且不涉及 k 之后任意治疗变量的

242 后门路径被阻断。

然而, 这一扩展版的后门准则没有直接将阻断后门路径和时序互换性联系起来, 这是因为此时的后门准则建立在不涉及反事实结局的有向无环图之上。不过我们在第七章讨论过, 我们可以用单一世界干涉图判断因果效应的可识别性, 而单一世界干涉图对于时异治疗而言特别有效。

图 19.5 和 19.6 分别是图 19.2 和 19.4 的简化版本, 此处我们省略了 U_0 , L_0 , 以及从 A_0 和 U_1 出发的箭头。此外, 没有从 L_1 指向 U 的箭头, 因而 L_1 不再是 Y 的直接诱因。图 19.5 和 19.6 的区别在于 A_k 和之后时间段的 L_t ($t > k$) 是否有共同诱因 W_k (这也是图 19.2 和 19.4 的区别)。

我们在本书第一部分讨论过, 单一世界干涉图表示的是某一干涉下的反事实世界。图 19.7 就是一幅单一世界干涉图, 其表示如果图 19.5 中的所有人都接受静态策略 (a_0 , a_1) 时的反事实世界 (a_0 和 a_1 的取值是 0 或 1)。比如, 图 19.7 可以用来表示“肯定会治疗”策略 ($a_0 = 1$, $a_1 = 1$), 或“从不会治疗”策略 ($a_0 = 0$, $a_1 = 0$) 所对应的反事实世界。要构建一幅单一世界干涉图, 我们需要把治疗节点 A_0 和 A_1 分裂成左右两部分。右边表示特定干涉下的治疗取值, 左边表示对过往治疗进行干涉后你能观测到的治疗取值。因此, 左半部分的 A_0 就是实际上的 A_0 , 因为这是第一个时间点, 在这之前没有其他治疗变量。而左半部分的 A_1 则是我们把 A_0 设定为 a_0 之后的反事实情形 $A_1^{a_0}$ 。所有指向治疗节点的箭头都是指向左半部分, 而所有从治疗节点出发的箭头都是从右半部分出发。结局变量是反事实结局 Y^{a_0, a_1} , 而协变量 L 也将会由对应的反事实变量替代。注意, 我们将治疗策略 (a_0 , a_1) 下 L_1 的反事实结局写作 $L_1^{a_0}$, 而非 $L_1^{a_0, a_1}$, 这是因为对 A_1 的干涉不会影响到 L_1 。

与图 19.5 不同的是, 图 19.7 包含了反事实结局, 这也意味着我们能够直观地判断互换性是否成立, 比如用我们在第一部分介绍的有向分离方法。

在图 19.7 中, 使用有向分离, 我们可知对于任何静态策略 (a_0 , a_1) 而言, $Y^{a_0, a_1} \perp\!\!\!\perp A_0$ 和 $Y^{a_0, a_1} \perp\!\!\!\perp A_1^{a_0} | A_0, L_1^{a_0}$ 成立。注意到, 图中似乎存在一条开放路径 $A_1^{a_0} \leftarrow a_0 \rightarrow L_1^{a_0} \leftarrow U_1 \rightarrow Y^{a_0, a_1}$, 使得上述第二个独立性不成立。然而这条路径实际上是被阻断的, 这是因为在反事实世界中, a_0 是一个常数, 而在概率论中我们默认常数就代表了我们已经对其进行控制。因此, 在使用有向分离检视互换性的时候, 我们需要谨记常数就代表了被控制, 也就阻断了上述路径。

根据定义, 上一段中的第二个条件独立 $Y^{a_0, a_1} \perp\!\!\!\perp A_1^{a_0} | A_0, L_1^{a_0}$ 也意味着在接受了治疗 $A_0 = a_0$ 的人群中 $Y^{a_0, a_1} \perp\!\!\!\perp A_1^{a_0} | A_0 = a_0, L_1^{a_0}$ 。因此, 根据一致性, 我们可知在图 19.7 对应的因果图 19.5 中, $Y^{a_0, a_1} \perp\!\!\!\perp A_0$ 和 $Y^{a_0, a_1} \perp\!\!\!\perp A_1 | A_0 = a_0, L_1$ 成立。如果存在多个时间点, 那么互换性可表示为:

$$Y^{\bar{a}} \perp\!\!\!\perp A_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_k, \text{ 其中 } k = 0, 1, \dots K$$

(因为根据一致性, 在 $A_0 = a_0$ 时有 $L_1^{a_0} = L_1$ 以及 $A_1^{a_0} = A_1$, 所以 $Y^{a_0, a_1} \perp\!\!\!\perp A_1^{a_0} | A_0 = a_0, L_1^{a_0}$ 等于 $Y^{a_0, a_1} \perp\!\!\!\perp A_1 | A_0 = a_0, L_1$ 。)

243 我们将这一表述称为 $Y^{\bar{a}}$ 的静态时序互换性, 其弱于 Y^g 的时序互换性, 因为其只需要静态策略 $g = \bar{a}$ 的反事实结局 $Y^{\bar{a}}$ 和治疗 A_k 之间的独立性。静态时序互换性足以用来识别任意静态策略 $g = \bar{a}$ 的反事实结局, 详情参见知识点 19.3。

静态时序互换性在图 19.6 中依然成立, 我们可以在图 19.8 (图 19.6 所对应的单一世界干涉图) 中用同样的方法检验互换性。因此, 如果一项观察性研究能用图 19.8 表示, 那么我们就能识别任意静态策略 (a_0, a_1) 下的反事实结局。

244 让我们回到图 19.5 当中, 假设没有箭头从 L_1 指向 A_1 , 那在图 19.7 中也就没有箭头从 $L_1^{a_0}$ 指向 $A_1^{a_0}$ 。此时根据有向分离可知, 时序互换性对于 A_0 无条件成立, 控制了 A_0 之后对 $A_1^{a_0}$ 成立, 因此即使没有 L_1 的数据, 静态策略下的反事实结局依然是可识别的。现在让我们假设在图 19.5 中, 有一个箭头从 U_1 指向 A_1 。于是在图 19.7 中, 就会有一个箭头从 U_1 指向 $A_1^{a_0}$, 那此时任何形式的时序互换性都不成立, 也就意味着任何策略下的反事实结局都不可被识别。

接下来我们讨论动态策略下单一世界干涉图的使用。如果图 19.5 中的治疗策略是动态的, 那么其对应的单一世界干涉图如图 19.9 所示, 其中动态策略 $g = [g_0, g_1(L_1)]$, A_0 被赋予一个固定值 g_0 (0 或 1), 而 $k=1$ 时 A_1 的赋值 $g_1(L_1^g)$ 受到 L_1^g 的影响, 而 L_1^g 又受到 g_0 的影响。比如, g 可能代表的策略是: 在最开始时不治疗, $k=1$ 时如果 CD4 过低 (即 $L_1^g=1$) 就治疗。在这一策略下, 所有人都有 $g_0=0$, 而 $L_1^g=1$ 时有 $g_1(L_1^g)=1$, $L_1^g=0$ 时有 $g_1(L_1^g)=0$ 。因此在单一世界干涉图中, 存在从 L_1^g 指向 $g_1(L_1^g)$ 的箭头。这一箭头不存在于因果图中, 而是干预的结果。因此, 为了将这一箭头和其他箭头区别开来, 我们用虚线表示这一箭头。不过在使用有向分离时, 我们需要同等考虑所有箭头。在图 19.9 中, Y^g 就是动态治疗策略 g 对应的反事实结局。

在图 19.9 中, 由有向分离可知, $Y^g \perp\!\!\!\perp A_0$ 和 $Y^g \perp\!\!\!\perp A_1 | A_0 = g_0, L_1^g$ 对任意策略 g 成立, 也就意味着我们可以识别任意策略 g 所对应的反事实结局 (参见精讲点 19.3)。不过这一结论不适用于图 19.6。

图 19.10 是动态策略 $g = [g_0, g_1(L_1)]$ 下图 19.6 所对应的单一世界干涉图。根据有向分离可知 $Y^g \perp\!\!\!\perp A_0$ 不成立, 因为此时存在一条开放路径 $A_0 \leftarrow W_0 \rightarrow L_1^g \rightarrow g_1(L_1^g) \rightarrow Y^g$ 。也就是说, Y^g 的时序互换性不成立, 也就意味着我们不能识别反事实结局。

(严格而言, 我们从单一世界干涉图中得到的是 $Y^g \perp\!\!\!\perp A_1^g | A_0, L_1^g$, 而根据一致性, 我们进一步有 $Y^g \perp\!\!\!\perp A_1 | A_0 = g_0, L_1$ 。)

总而言之, 在图 19.5 表示的观察性研究 (或时序性随机试验) 之中, Y^g 的时序互换性成立, 因此我们就能用数据有效地估计静态或动态治疗策略的因果效应。而在图 19.6 表示的观察性研究中, 只有静态策略的弱互换性成立, 因此我们只能有效估计静态治疗策略的因果效应。我们可以用图 19.10 帮助我们理解这一结论, 此时 Y^g 的分布受到 $g_1(L_1^g)$ 的影响, 也就受到 L_1^g 的影响。然而, L_1^g 的分布是不可确定的, 因为存在开放路径 $A_0 \leftarrow W_0 \rightarrow L_1^g$ 。

245 最后, 让我们思考一下图 19.11。除去一个从 L_1 到 Y 的箭头, 图 19.11 和图 19.6 相同, 而其对应的单一世界干涉图是图 19.12。由有向分离可知, 此时不管是时序互换性还是静态时序互换性都不成立。因此, 在图 19.11 表示的观察性研究中, 我们不能用数据有效估计任何治疗策略的因果效应。

19.6 时异混杂

在观察性研究中, 没有哪一种形式的时序互换性能保证一定成立。我们需要借助专业知识从而近似得到互换性, 这也意味着在研究设计阶段, 研究者需要尽可能地测量相关变量 \bar{L}_k 。比如, 在我们的 HIV 研究中, 研究者可能需要收集的时异变量包括 CD4 数量、病毒载量、以及部分症状, 从而在分析中调整这些变量。

我们永远无法肯定地回答我们收集的变量是否足以保证时序互换性。但我们可以用因果图展现我们的专业知识以及收集到的变量。图 19.1 到 19.4 就展现了不同的情形。注意, 在本章我们的因果图都没有包含可能的选择偏移 (比如删失和失访), 而只是将重心放在混杂之上。

让我们再思考一下图 19.5。就像在本书第一部分一样, 假设我们只想知道 A_1 (此时这是一个非时异变量) 对结局 Y 的因果效应。我们认为 A_1 和 Y 之间有混杂, 这是因为 U 是两者的共同诱

因, 也即 A_1 和 Y 之间存在一条途经 U 的开放路径。为了无偏地估计 A_1 对 Y 的因果效应, 我们需要调整所有的混杂变量, 阻断所有的开放路径, 就如第七章所述。然而路径 $A_1 \leftarrow L_1 \leftarrow U \rightarrow Y$ 中的 U 是一个未测变量, 所以我们也就不能调整它。不过我们可以控制已测变量 L_1 从而阻断这条路径。因此, 只要研究者收集了 L_1 的信息, 那就不存在未测混杂, 虽然 U 才是 A_1 和 Y 真正的共同诱因 (如果你对这部分感到陌生, 可以再回顾一下第七章)。

我们在第七章讨论过, 混杂变量不一定是结局的直接诱因。在图 19.5 中, 不存在从混杂变量 L_1 到结局 Y 的箭头。因此混杂的真正来源 (也即因果性混杂变量) 是未测变量 U 。然而, 因为 L_1 足以阻断 A_1 和 Y 之间的后门路径, 所以我们也将 L_1 称为混杂变量。

(不过, L_1 在另一条路径中是对撞变量: $A_1 \leftarrow A_0 \rightarrow L_1 \leftarrow U \rightarrow Y$ 。控制 L_1 会打开这条路径, 但我们可以通过控制 A_0 从而阻断这条路径。)

假设有一个很长的因果图, 包含了 $k = 0, 1, 2, \dots$ 多个时间点的信息, 其中的 L_k 能够影响之后时间段的治疗变量 A_k, A_{k+1}, \dots , 并且 L_k 和结局 Y 之间还有未测的共同诱因 U_k 。假设我们想估计某个治疗策略的因果效应, 那我们就需要每个时间点的协变量史 \bar{L}_k , 以及治疗史 \bar{A}_{k-1} , 这样我们才能阻断治疗 A_k 和结局 Y 之间的所有后门路径。因此要无偏地估计 \bar{A} 的因果效应, 我们需要收集所有 246 人 \bar{L}_k 的信息。我们将 \bar{L}_k 称为时异混杂, 精讲点 19.4 给出了时异混杂更精确的定义。

(时异混杂有时也被称为时依混杂。)

遗憾的是, 我们不能实证地证明所有混杂——无论时异的还是非时异——是否已被完全测量。也就是说, 我们不能实证地区分图 19.2 (不存在未测混杂) 和图 19.3 (存在未测混杂)。不过有意思的是, 即使所有的混杂都被测量了, 大多数模型依然不能无偏地估计治疗策略的因果效应。下一章将解释为什么 G-方式是用来调整时异混杂的最佳方法之一。

第十九章精讲点和知识点

精讲点 19.1: 命定的治疗策略与随机的治疗策略 (原书第 237 页)

动态治疗策略 $g = [g_0(\bar{a}_{-1}, l_0), \dots, g_K(\bar{a}_{K-1}, \bar{l}_K)]$, 其中 $g_k(\bar{a}_{k-1}, \bar{l}_k)$ 表示在时间点 k 的治疗概率由过往历史 $(\bar{a}_{k-1}, \bar{l}_k)$ 决定。比如在我们 HIV 的例子中: 如果一个人的 CD4 数目 (也即 \bar{l}_k) 的一个函

数) 在 k 时或之前都很低, 那 $g_k(\bar{a}_{k-1}, \bar{l}_k)$ 等于 1, 否则 $g_k(\bar{a}_{k-1}, \bar{l}_k)$ 等于 0。静态治疗策略

$g = [g_0(\bar{a}_0), \dots, g_K(\bar{a}_{K-1})]$ 则不依赖于 \bar{l}_k 。我们经常会把 $g_k(\bar{a}_{k-1}, \bar{l}_k)$ 简写作 $g(\bar{a}_{k-1}, \bar{l}_k)$ 。

很多时候, 我们所说的动态策略或者静态策略都是命定的, 也即每一个人在每一个时间点的治疗取值都是某个特定的值 (0 或 1)。而更一般的情况, 也即随机治疗策略, 指的是每个人在每个时间点都有一个接受治疗的概率, 而非治疗取值本身。随机治疗策略可以是静态的 (比如: 每个人在每个月都有 0.3 的概率接受治疗, 也即 0.7 的概率不接受治疗), 也可以是动态的 (比如: 在每个月, 如果 CD4 数目过低, 那就有 0.3 的概率接受治疗, 如果 CD4 数目正常, 那就不会接受治疗)。

如果结局的取值越高表示越好, 那我们将能够把反事实结局均值 $E[Y^g]$ 最大化的策略 g 称为最佳治疗策略。对于某种药物治疗而言, 最佳策略基本上都是动态的, 因为必须留出足够长的用药间隙从而保证不会出现药物中毒。同理, 随机策略不可能优于命定的策略。然而, 随机策略 (实际上, 普通的随机试验和时序性随机试验都是随机策略) 也是必要的, 因为在一项试验之前, 我们并不知道哪一个命定的测量时最优的。在正文中, 除非特别说明, 否则我们的符号 g 表示的都是命定的治疗策略。

精讲点 19.2: 不同动态策略的依方案效应 (原书第 239 页)

许多随机试验会在一开始就把被试分配到不一样的治疗方案中, 并希望被试一直遵循这一方案直到研究结束, 除非在研究过程中发现这一治疗有害或者有其他不可抗力。也就是说, 随机试验的目标旨在对比不同治疗策略的因果效应, 而依方案效应是如果每个被试都严格遵循治疗策略我们所能观察到的效应。

比如, 某项随机试验的目标是对比“在研究一开始就服用他汀类药物, 直到研究结束或出现横纹肌肉瘤”和“在胆固醇过高或出现冠心病之前, 不服用他汀类药物”这两种动态策略的因果效应。不过, 在这项随机试验中对比这两种治疗策略的依方案效应, 会遇到和观察性研究一样的问题。这是因为估计依方案效应需要研究者在随机分组后继续收集被试的依从情况, 以及会影响被试依从情况的协变量。随机试验的分组, 只能保证治疗策略分配的互换性, 而不能保证实际治疗策略的时序互换性。

精讲点 19.3: 依赖于初始协变量的动态策略 (原书第 244 页)

为了简便, 本章的因果图都不含初始时的协变量 L_0 。如果我们在图 19.9 中加入 L_0 , 那在我们的治疗策略中, 初始时的变量就不再是 g_0 , 而是 $g_0(L_0)$ 。在图中, 根据有向分离, 我们可以得到 $Y^g \perp\!\!\!\perp A_1^g | A_0, g_0(L_0), L_0, L_1^g$ 。因此, 我们需要控制所有的既往史, 包括 $g_0(L_0)$ 。如果我们用 $A_0 = g_0(L_0)$ 进行替换, 并根据一致性, 我们可以得到 $Y^g \perp\!\!\!\perp A_1 | A_0 = g_0(L_0), L_0, L_1$, 这就是存在初始协变量时的时序互换性。

精讲点 19.4: 时异混杂的定义 (原书第 246 页)

在没有选择偏移的情况下, 如果 $E[Y^{\bar{a}}] \neq E[Y | A = \bar{a}]$, 也即所有人都遵循策略 \bar{a} 得到的反事实结局不等于实际上遵循策略 \bar{a} 的人群的观测结局, 那么我们就说存在 $E[Y^{\bar{a}}]$ 的混杂。

如果 $E[Y^{\bar{a}} | L_0] = E[Y | A = \bar{a}, L_0]$, 也即在图 19.2 中没有从 L_1 指向 A_1 的箭头, 那么混杂就是非时异的。反之, 如果 $E[Y^{\bar{a}} | L_0] \neq E[Y | A = \bar{a}, L_0]$, 而三个可识别性假设成立, 那我们说存在时异混杂。如果可识别性假设不成立, 如图 19.3 所示, 那我们说存在未测混杂。

不存在时异混杂的一个充分条件是 $Y^{\bar{a}}$ 的无条件时序互换性, 即 $Y^{\bar{a}} \perp\!\!\!\perp A_k | \bar{A}_{k-1} = \bar{a}_{k-1}$ 。这一互换性在图 19.1 所示的时序性随机试验中成立, 其中每个时间点的治疗 A_k 都是随机给予的, 概率仅取决于既往治疗史 \bar{A}_{k-1} 。事实上, 图 19.1 可以被进一步简化。首先, 注意到 L_1 并不是任意两个变量的共同诱因, 因此我们可以省略 L_1 。其次, 省去 L_1 后, L_0 和 U_1 不再是两个变量的共同诱因, 因此也可以省去。没了 L_0 、 L_1 、以及 U_1 之后, U_0 也可以进一步省略。最后, 图 19.1 可以被简化为只含 A_0 、 A_1 、以及 Y 。

知识点 19.1: 动态策略的定义 (原书第 238 页)

每一个动态策略 $g = [g_0(\bar{a}_{-1}, l_0), \dots, g_K(\bar{a}_{K-1}, \bar{l}_K)]$ 都会受到过往治疗史和协变量史的影响, 并且有趣的是, 这一策略还和仅取决于协变量史的动态策略 $g' = [g'_0(l_0), \dots, g'_K(\bar{l}_K)]$ 相关联。根据一致性 (参见知识点 19.2), 遵循策略 g 的个体, 会和遵循策略 g' 的个体有同样的治疗、协变量和结局, 我们会得到 $Y^g = Y^{g'}$ 以及 $\bar{L}^g(K) = \bar{L}^{g'}(K)$ 。这是因为在最开始有 $g'_0(l_0) = g_0(\bar{a}_{-1} = 0, l_0)$ (没有特殊说明时, \bar{a}_{-1} 取值为零, 也即不会对函数造成任何影响), 再进一步递推, 我们可以得

作者: Miguel A. Hernan, James M. Robins;

翻译: 罗家俊

到 $g'_k(\bar{l}_k) = g_k[g'_k(\bar{l}_{k-1}), \bar{l}_k]$ 。因此, 当且仅当遵循了策略 g' (即 $A_k = g'_k(\bar{L}_k)$) , 才可能会遵循动态策略 g ($A_k = g_k(\bar{A}_{k-1}, \bar{L}_k)$)。

知识点 19.2: 时异治疗的正数性和一致性 (原书第 241 页)

正数性的一般形式为: 如果 $f_L(l) \neq 0$, 那么 $f_{A|L}(a|l) > 0$ 。把它推广到时序形式可得: 如果 $f_{\bar{A}_{k-1}, \bar{L}_k}(\bar{a}_{k-1}, \bar{l}_k) \neq 0$, 那么对于所有 $(\bar{a}_{k-1}, \bar{l}_k)$ 有 $f_{A_k|\bar{A}_{k-1}, \bar{L}_k}(a_k | \bar{a}_{k-1}, \bar{l}_k) > 0$ 。

在时序性随机试验中, 如果每个时间点随机分配治疗的概率不是 0 或 1, 那么不管既往治疗史或协变量史如何, 正数性都会成立。如果我们只是对某个特定测量 g 感兴趣, 那么正数性只需要对与 g 相关的治疗史成立即可, 也即对于每个时间点 k , 有 $a_k = g(\bar{a}_{k-1}, \bar{l}_k)$ 。

一致性的一般形式为: 如果 $A = a$, 那么 $Y^a = Y$ 。把它推广到时序形式可得: 如果 $\bar{a}^* = \bar{a}$, 那么 $Y^{\bar{a}} = Y^{\bar{a}^*}$; 如果 $\bar{A} = \bar{a}$, 那么 $Y^{\bar{a}} = Y$; 如果 $\bar{a}_{k-1}^* = \bar{a}_{k-1}$, 那么 $\bar{L}_k^{\bar{a}} = \bar{L}_k^{\bar{a}^*}$; 如果 $\bar{A}_{k-1} = \bar{a}_{k-1}$, 那么 $\bar{L}_k^{\bar{a}} = \bar{L}_k$ 。其中 $\bar{L}_k^{\bar{a}}$ 是治疗策略 \bar{a} 之下的反事实协变量 L 史。严格而言, 静态和动态治疗策略所需的一致性不同, 但都比上述一致性弱。静态治疗策略只需要: 如果 $\bar{A} = \bar{a}$, 那么 $Y^{\bar{a}} = Y$ 。动态治疗策略则需要: 对于任意策略 g , 如果在每一个时间点有 $A_k = g_k(\bar{A}_{k-1}, \bar{L}_k)$, 那么 $Y^g = Y$ 。不过我们总是欢迎更强的一致性成立。

最后, 如果我们将“时间点 k 时接受治疗”定义为 $A_k = 1$, 将“时间点 k 时未接受治疗”定义为 $A_k = 0$, 那涉及 A_k 的静态或动态策略都是良定的。

知识点 19.3: 不同形式的时序互换性 (原书第 243 页)

假设有一项涉及多个时间点 $k = 0, 1, \dots, K$ 以及时异治疗 A_k 的时序性随机试验。图 19.7 描述了这项研究。从图中, 我们可以直接得到:

$$(Y^{\bar{a}}, \underline{L}_{k+1}^{\bar{a}}) \perp\!\!\!\perp A_k^{\bar{a}_{k-1}} | \bar{A}_{k-1}^{\bar{a}_{k-2}}, \bar{L}_k^{\bar{a}_{k-1}}$$

其中 $\underline{L}_{k+1}^{\bar{a}}$ 是从 $k+1$ 直到研究结束的反事实协变量史。这一独立性也就意味着根据一致性, 对于任意 $\bar{A}_{k-1}^{\bar{a}_{k-2}} = \bar{a}_{k-1}$ (其中 \bar{a}_{k-1} 是 \bar{a} 的一个成分), 都有:

$$(Y^{\bar{a}}, \underline{L}_{k+1}^{\bar{a}}) \perp\!\!\!\perp A_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_k$$

当这一表述对所有 \bar{a} 成立时, 我们说时序互换性成立。如果 $g = \bar{a}$ 是静态策略, 这一表述等价于一个更强的互换性表述:

$$(Y^{\bar{a}}, \underline{L}_{k+1}^{\bar{a}}) \perp\!\!\!\perp A_k | \bar{A}_{k-1} = g(\bar{A}_{k-1}, \bar{L}_k), \bar{L}_k$$

在上述最后一个互换性表述成立的情况下, 如果正数性也成立, 那不管是静态策略还是动态策略, 我们都能识别结局和协变量的分布, 这是因为此时 $(Y^{\bar{a}}, \underline{L}_{k+1}^{\bar{a}})$ 独立于 A_k 。但是注意, 对于动态策略而言, 这一联合独立性并不等于各自独立性, 也即我们得不到 $Y^{\bar{a}} \perp\!\!\!\perp A_k | \bar{A}_{k-1} = g(\bar{A}_{k-1}, \bar{L}_k), \bar{L}_k$, 以及 $\underline{L}_{k+1}^{\bar{a}} \perp\!\!\!\perp A_k | \bar{A}_{k-1} = g(\bar{A}_{k-1}, \bar{L}_k), \bar{L}_k$ 。

在时序性随机试验中, 上述最强的互换性必然成立, 但注意以下几点: (1) 这一互换性不能从单一世界干涉图中得到; (2) 不是识别治疗策略因果效应的必要条件。

在时序性随机试验中, 还能有一个更强的互换性表述:

$$(Y^{\bar{A}}, \bar{L}^{\bar{A}}) \perp\!\!\!\perp A_k | \bar{A}_{k-1}, \bar{L}_k$$

其中, 如果 A_k 是一个二分治疗变量 (只有 0 或 1), 那么 \bar{A} 表示所有静态策略 \bar{a} (共 2^K 个), $Y^{\bar{A}}$ 所有表示反事实结局 $Y^{\bar{a}}$ 的集合, $\bar{L}^{\bar{A}}$ 表示所有反事实协变量史的集合。与知识点 2.1 类似, 我们将这一联合独立性称为完全时序互换性。

第十九章图表

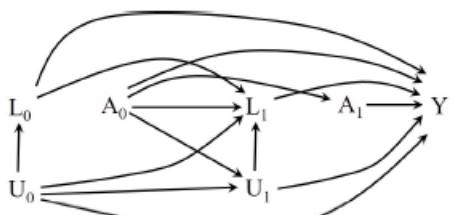


Figure 19.1

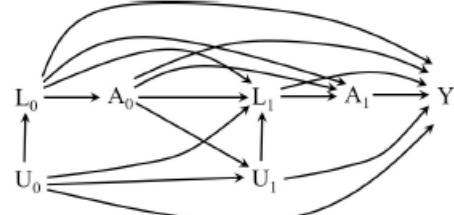


Figure 19.2

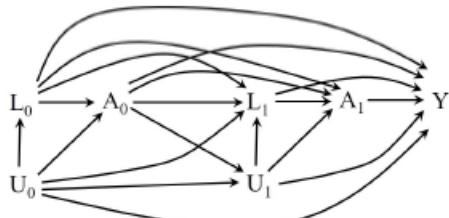


Figure 19.3

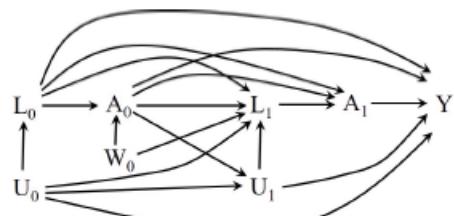


Figure 19.4

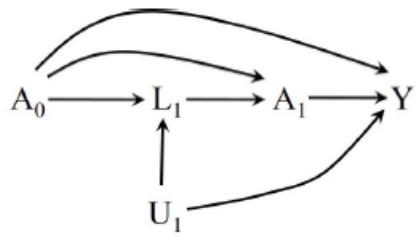


Figure 19.5

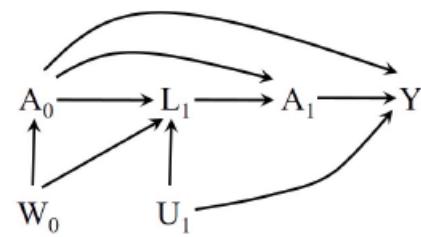


Figure 19.6

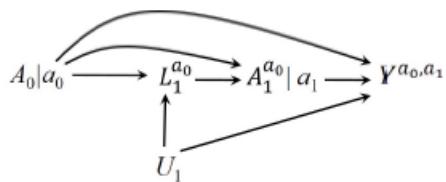


Figure 19.7

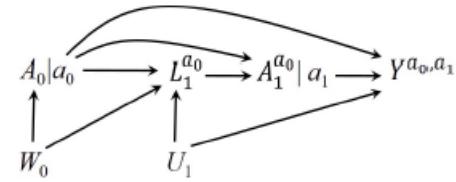


Figure 19.8

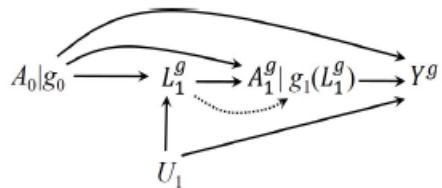


Figure 19.9

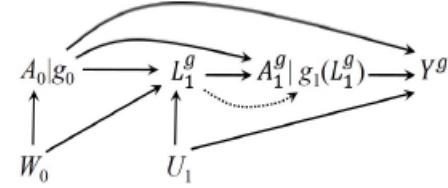


Figure 19.10

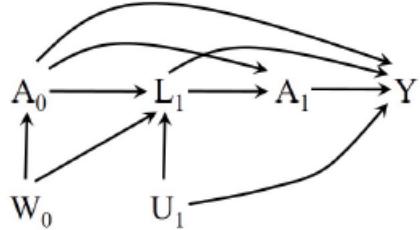


Figure 19.11

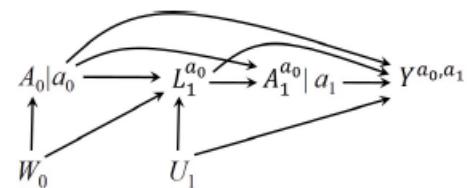


Figure 19.12

第二十章 治疗-混杂反馈

247 在上一章, 我们讨论了时序互换性, 这也是时异治疗最重要的概念。如果在一项研究中最强形式的时序互换性成立, 那么已测的时异变量足以有效估计任意治疗策略的因果效应。那接下来的问题是, 我们应该用什么方法调整这些变量? 这一问题关平时异治疗的一个重要特征: 治疗-混杂反馈。

当治疗-混杂反馈存在的时候, 传统方法会对估计值造成偏移。也就是说, 即使我们有了不同时间点全部变量的数据, 但如果使用传统方法, 我们依然不能有效地估计时异治疗的因果效应。本章将讨论治疗-混杂反馈的结构特征, 同时论述为什么传统方法不适用于时异治疗。

20.1 治疗-混杂反馈的要素

再思考一下我们上一章讨论的 HIV 时序性随机试验。假设我们拥有所有被试在每个时间点的治疗数据 A_k (1: 接受治疗; 0: 未接受治疗)、协变量数据 L_k 、以及最后的结局数据 Y 。图 20.1 (与图 19.2 相同) 描述了这一研究的前两个月。时异协变量 L_k 是时异混杂变量。和上一章一样, 在本章我们主要考虑混杂因素, 不考虑删失等其他因素。

现在我们来分析图 20.1 中的路径。此时不仅有从 L_k 指向 A_k 的箭头, 还有从 A_{k-1} 指向 L_k 的箭头 (比如, 在上一个时间段接受治疗会影响接下来的 CD4 数目)。也就是说, 不仅混杂会影响治疗, 同时治疗也会影响混杂。我们称之为治疗-混杂反馈 (参见精讲点 20.1)。

注意, 没有治疗-混杂反馈的时候, 依然可能存在时异混杂。在图 20.2 中, 没有从 A_{k-1} 指向 L_k 和 U_k 的箭头, 除此之外其与图 20.1 一样。此时, L_k 依然是时异混杂, 但不再受上一个时间段治疗的影响。因此, 图 20.2 之中存在时异混杂, 但是不存在治疗-混杂反馈。

治疗-混杂反馈会给因果推断带来意想不到的难题。我们先将图 20.1 简化成图 20.3, 也即能用来讲解治疗-混杂反馈的最简形式:

- 248
- 因为我们只关心 L_1 带来的混杂, 所以我们就没必要再纳入 L_0 。
 - 移除未测变量 U_0 。
 - 移除从 A_0 到 A_1 的箭头, 从而 A_1 只受到 L_1 的影响。
 - 移除从 A_0 、 A_1 、或 L_1 指向 Y 的箭头, 也即极端零假设成立。

上述简化并不会影响我们下面的论述。我们也可以用更复杂的因果图讲解治疗-混杂反馈, 但除了增加阅读难度, 并不会提升大家对这一概念的理解。

假设真实情况是治疗对结局 Y 没有效应, 也即如图 20.3 所示 (此时图中没有箭头从 A_0 或 A_1 指向 Y) , 但研究者并不知道这一情况。同样假设我们有图 20.3 中所有变量的数据。我们希望用这些数据比较 “肯定会治疗” ($a_0 = 1, a_1 = 1$) 和 “从不会治疗” ($a_0 = 0, a_1 = 0$) 两种不同策略对结局 Y 的影响, 也即估计 $E[Y^{a_0=1, a_1=1}] - E[Y^{a_0=0, a_1=0}]$ 。

根据上一章的讨论, 我们知道图 20.3 可以表示一项时序性随机试验, 因此理论上研究者可以用已有数据得到 $E[Y^{a_0=1, a_1=1}] - E[Y^{a_0=0, a_1=0}] = 0$ 。然而, 我们在下一小节会论述为什么使用传统方法——比如分层分析、结局回归、以及匹配等——并不能得到正确的估计值, 也即, 在这个例子中, 使用传统方法得到的估计值会偏离零, 因此是无效的。

249 换句话说, 当存在时异混杂与治疗-混杂反馈的时候, 我们不能用传统方法得到无偏的估计值。即使我们有足够的数据保证时序互换性成立, 传统方法依然不能给出有效、无偏的估计值。与之相反, G-方法能够在治疗-混杂反馈存在的情况下给出正确的估计值。

(图 20.3 可以表示一项时序性随机试验或者一项没有未测混杂的观察性研究。图 20.4 只能表示一项观察性研究。)

传统方法不仅在图 20.3 中存在局限性, 同时也不适用于图 20.4, 也即时异混杂和前一时间段的治疗有一个共同诱因 W 的情形 (图 20.4 是图 19.4 的简化情形)。我们将图 20.4 和 20.4 (以及图 19.2 和 19.4) 用作治疗-混杂反馈的讲解用例。下一小节将论述为什么传统方法不足以处理治疗-混杂反馈。

20.2 传统方法的不足

为了解释传统方法的不足, 我们先假设有一项时序性随机试验, 其中有 32000 名 HIV 阳性的被试, 并且有两个时间点 $k = 0$ 和 $k = 1$ 。在 $k = 0$ 时, 每名被试接受治疗 $A_0 = 1$ 的概率都是 0.5。而在 $k = 1$ 时, 接受治疗 $A_1 = 1$ 的概率则取决于当前 CD4 数目 L_1 。如果 $L_1 = 0$ (也即 CD4 数目高), 那么概率是 0.4。如果 $L_1 = 1$ (也即 CD4 数目低), 那么概率是 0.8。在研究结束时测量结局 Y , Y 的值越大表示被试越健康。

(这是一个假想试验, 每名被试都完美依从治疗方案, 并且没有失访发生。)

表 20.1 是这个假想试验的数据。为了节约空间, 这张表每行代表的是不同 A_0 、 A_1 、以及 L_1 组合下的人数 N 和结局 $E[Y | A_0, A_1, L_1]$ 的均值, 而非每名被试的情况。比如, 第一行就代表有 2400 名被试的变量取值是 $A_0 = 0, A_1 = 0, L_1 = 0$; 而他们结局均值 $E[Y | A_0, A_1, L_1] = 84$ 。在这

项假想的时序性随机试验中, 三个可识别性假设——时序互换性、正数性、一致性——成立。同时由试验的特质可知, A_0 和 Y 之间没有混杂, 因此 (控制了 L_1 之后) 时序互换性成立。同时, 从表 20.1 中可知每一行的被试数目都不为零, 所以正数性成立。

(如果还有更多的时间节点 k , 并且治疗 A_k 受到 L_k 的影响, 那么 L_k 就是时异混杂。)

图 20.3 描绘了在这项时序性随机试验, 其中极端零假设成立。为了验证表 20.1 中的数据是否符合和图 20.3 一致, 我们可以在不同协变量和治疗取值下分别估计非时异治疗 A_0 和 A_1 对结局的因果效应, 理论上结果应该都是零。在下面的计算中暂时忽略随机变异性。

(图 20.3 中, 不存在从 L_1 到 Y 的箭头, 因此 A_1 对 Y 的因果效应也为零。)

利用表 20.1 的第一和第二行的数据, 我们可以得到在 $A_0 = 0$ 且 $L_1 = 0$ 时, A_1 对 Y 的因果效
250 应为: $E[Y | A_0 = 0, A_1 = 1, L_1 = 0] - E[Y | A_0 = 0, A_1 = 0, L_1 = 0] = 0$ 。因为可识别性假设成立, 所以这一相关性等于因果性, 有: $E[Y^{A_1=1} | A_0 = 0, L_1 = 0] - E[Y^{A_1=0} | A_0 = 0, L_1 = 0] = 0$ 。同理, 我们可以得到在剩下三个 A_0 和 L_1 的分层中, A_1 对 Y 的因果效应为零。

接下来我们要验证 A_0 对 Y 的因果效应 $E[Y^{A_0=1}] - E[Y^{A_0=0}] = 0$ 。首先我们要计算相关性 $E[Y | A_0 = 1] - E[Y | A_0 = 0]$ 。其中, $E[Y | A_0 = 0]$ 是第一、二、三、四行结局的加权均值, 即 60。同理可得 $E[Y | A_0 = 1] = 60$ 。因此, A_0 的因果效应为零。

$$(加权均值的计算: \frac{2400}{16000} \times 84 + \frac{1600}{16000} \times 84 + \frac{2400}{16000} \times 52 + \frac{9600}{16000} \times 52 = 60)$$

因此, 如果我们分别计算 A_0 和 A_1 的因果效应, 那么它们对结局的因果效应均值都为零。但如果我们将两者同时考虑并视为时异治疗, 然后对比两种不同治疗策略, 会发生什么呢? 比如, 我们想比较的是“肯定会治疗” ($a_0 = 1, a_1 = 1$) 和“从不会治疗” ($a_0 = 0, a_1 = 0$) 两种策略的因果效应。因为可识别性的三个假设成立, 所以表 20.1 中的数据足以用来估计这一效应。

因为我们已经知道 a_0 和 a_1 各自的因果效应为零, 所以联合因果效应应该也为零, 也即: $E[Y^{a_0=1, a_1=1}] - E[Y^{a_0=0, a_1=0}] = 0$ 。但我们能从数据中得到这个结论吗? 接下来我们用两种不同的方式, 对表 20.1 中的数据进行分析。在第一种方式中, 我们不调整 L_1 , 因此我们应该得到一个有偏的效应估计。在第二种方式中, 我们用分层分析的方法调整 L_1 。

1. 我们直接比较表 20.1 中在两个时间点都治疗的 9600 名被试 (第六和第八行), 和两个时间点都没治疗的 4800 名被试 (第一和第三行), 因此有:

251 $E[Y | A_0 = 1, A_1 = 1] - E[Y | A_0 = 0, A_1 = 0] = 54.7 - 68 = -13.3$, 这一结果似乎说明两个时间点都未接受治疗的被试更健康。不过这个结论是错误的, 因为我们没有调整混杂 L_1 , 所以相关性 $E[Y | A_0 = a_0, A_1 = a_1]$ 也就不等于因果性 $E[Y^{a_0, a_1}]$ 。

$$(E[Y | A_0 = 1, A_1 = 1] = \frac{3200}{9600} \times 76 + \frac{6400}{9600} \times 44 = 54.7,$$

$$E[Y | A_0 = 0, A_1 = 0] = \frac{2400}{4800} \times 84 + \frac{2400}{4800} \times 52 = 68.0)$$

2. 接下来我们在 L_1 不同取值下比较两种不同策略。在 $L_1 = 0$ 时,

$E[Y | A_0 = 1, A_1 = 1, L_1 = 0] = 76$ (第六行), $E[Y | A_0 = 0, A_1 = 0, L_1 = 0] = 84$ (第一行)。两者之差是-8, 似乎说明在 $L_1 = 0$ 时, 在两个时间点都不治疗比都接受治疗要好一些。同理, 在 $L_1 = 1$ 时,

$$E[Y | A_0 = 1, A_1 = 1, L_1 = 1] - E[Y | A_0 = 0, A_1 = 0, L_1 = 1] = -8.$$

(注意, 此时在 L_1 的两个分层中结果都是-8, 所以最后加权得到的总效应不可能是 0。)

我们知道正确的结果应该是 0, 而不是-8。为什么会这样? 为什么我们调整了混杂依然得不到正确的答案? 这是因为, 当治疗-混杂反馈存在的时候, 利用传统方法调整混杂只会引入新的偏移。下一小节将解释这个偏移从何而来。

20.3 为什么传统方法失效了

表 20.1 中的数据来自一项时序性随机试验, 其可以用图 20.3 表示。我们有所有变量的数据。根据第十九章的讨论, 我们应该能够正确估计不同治疗策略的因果效应, 不管是动态的还是静态的。不过我们上一小节的计算——不管有没有调整 L_1 ——似乎表示这个说法是不对的。

此时的问题在于我们没有用正确的方法调整混杂。分层分析是常用来调整混杂的方法之一, 但是它不能用来处理治疗-混杂反馈。分层分析指的是在混杂变量的每一分层中估计治疗和结局的相关性, 我们曾在本书第四章详细介绍过分层分析。

不过此时的 L_1 受到上一时间点的治疗 A_0 的影响, 因此控制 L_1 会产生没有预料到的结果。图 20.5 描绘了控制 L_1 会带来的后果。因为 L_1 是一个对撞变量, 控制 L_1 也就打开了原本被阻断的路径 $A_0 \rightarrow L_1 \leftarrow U_1 \rightarrow Y$ 。也就是说, 分层分析会引入新的偏移。比如, 在 CD4 数目较低 ($L_1 = 1$) 的被试中, 一开始接受治疗 ($A_0 = 1$) 就成了较严重免疫抑制 (U_1 值较大) 的标志, 也就预示了

252 更差的结局; 而在 CD4 数目较高 ($L_1 = 0$) 的被试中, 一开始并未接受治疗 ($A_0 = 0$) 就成了较良好免疫抑制 (U_1 值较小) 的标志, 也就预示了更好的结局。因此, 分层分析得到的结果就是有偏的。

换句话说, 分层分析消除了对 A_1 的混杂, 不过代价是引入了对 A_0 的选择偏移。因此, 即使各时间点治疗的单独因果效应都为零, 相关性 $E[Y | A_0 = 1, A_1 = 1, L_1 = l] - E[Y | A_0 = 0, A_1 = 0, L_1 = l]$ 也不等于 0。这一偏移源于我们控制了受到时异治疗影响的变量 L_1 。而净偏移的大小取决于我们消除的混杂和引入的选择偏移两者的总和。

严格来说, 传统方法不只是在混杂受到过往治疗影响时会造成偏移, 同时, 如果混杂和过往治疗之间有一个未测的共同诱因 W (常见于观察性研究), 传统方法也会造成偏移。在图 20.6 中, 控制对撞变量 L_1 会开启路径 $A_0 \leftarrow W_0 \rightarrow L_1 \leftarrow U_1 \rightarrow Y$ 。因此, 我们将图 20.3 和 20.4 中的情境都称为治疗-混杂反馈, 而我们在现实观察性研究中并不能被区分这两者。

我们以上讨论所用的因果图都非常简单, 其中治疗对结局的因果效应都为零。如果治疗对结局的因果效应不为零, 使用传统方法控制混杂依然会造成偏移, 就如图 20.7 所示: 存在一个从 L_1 指向 Y 的箭头并不会改变以上讨论 (参见精讲点 20.2)。同样, 我们用以讨论的因果图都只有两个时间点, 而在更高维、有更多时间点的数据中, 我们的讨论依然有效: 用传统方法控制混杂依然会造成偏移。此时, 因为混杂受到多个——而不再是一个——过往治疗的影响, 传统方法造成的偏移会进一步增强。

253 总而言之, 估计治疗策略的因果效应, 需要我们无偏地同时估计策略中各成分 A_k 的联合效应。即使我们有混杂的所有数据, 分层分析也依然不能给出正确的估计值。

20.4 我们能改进传统方法吗?

当治疗-混杂存在的时候, 我们论述了为何分层分析不能正确地估计治疗策略的因果效应。那其他传统方法效果如何呢, 比如结局回归?

在高维数据中, 简单的分层分析不再现实, 而结局回归也就成了最常用的分析方法。在表 20.1 中, 我们只有两个时间点, 那对于静态治疗策略 \bar{a} 而言, 也就只有 2^2 个分层。但是当我们有 100 个时间点的时候, 那对静态策略而言就有 2^{100} 个分层, 大大超过了我们的样本量。如果再加上动态策略, 分层数只会更多。

(如果考虑上每个时间点的混杂变量 L_k , 分层数只会更多。)

此时, 就如我们在第十一章讨论的一样, 我们需要借助模型估计治疗策略的因果效应。于是, 我们需要假设治疗策略 \bar{A} 和结局 Y 之间的函数关系。其中一种是假设治疗策略 \bar{A} 的因果效应是整个策略中累计治疗次数的线性函数。在这一假设下, 不同时间点治疗的因果效应都是一样的。不过, 如果我们对于治疗策略和结局两者间关系的假设是错误的, 那么我们用模型得到的估计值也就不是有效的, 这是使用模型的代价之一。

遗憾的是, 治疗-混杂反馈存在的时候, 传统回归模型依然会造成新的偏移。在表 20.1 中, 我们定义治疗的累计次数 $cum(\bar{A}) = A_0 + A_1$, 因而其有 3 个可能取值: 0 (在两个时间点都没有接受治疗), 1 (在某一个时间点接受了治疗), 2 (在两个时间点都接受了治疗)。我们想比较的是“肯定会治疗”策略 ($cum(\bar{A})=2$) 和“从不会治疗”策略 ($cum(\bar{A})=0$), 也即估计 $E[Y|cum(\bar{A})=2] - E[Y|cum(\bar{A})=0]$ 。

在我们之前提到的线性假设中, 结局均值 $E[Y|\bar{A}, L_1]$ 是 $cum(\bar{A})$ 的线性函数, 因此我们可以拟合以下回归模型:

$$E[Y|\bar{A}, L_1] = \theta_0 + \theta_1 cum(\bar{A}) + \theta_2 L_1$$

- 此时, 相关性差值 $E[Y|cum(\bar{A})=2, L_1] - E[Y|cum(\bar{A})=0, L_1]$ 就等于 $\theta_1 \times 2$ 。(这一模型假设因果效应在 $L_1=1$ 和 $L_1=0$ 两个分层中相等。)因此, 就有人会将 $\theta_1 \times 2$ 阐释为控制了 L_1 之后“肯定会治疗”策略相较于“从不会治疗”策略的因果效应。但这一阐释是错误的。因为在图 20.5 中我们已经说明了控制 L_1 会对 A_0 引入新的偏移, 而 A_0 是 $cum(\bar{A})$ 的一部分。这也就意味着即使真实因果效应为零, 模型中 θ_1 的估计值也不会为零。同样的道理适用于匹配等传统方法。

不过, 当治疗-混杂反馈存在的时候, 我们可以使用 G-方法调整时异混杂, 进而得到治疗策略因果效应的无偏估计。

(读者可以自行用表 20.1 中的数据估计上述模型中的 θ_1 。)

20.5 过往治疗

在论述如何用 G-方法调整时异混杂之前, 我们还需要再讨论一个话题。迄今为止, 为了方便我们的讲解, 我们所用的因果图都非常简单, 在这些图中, 过往时间点的治疗都不会对之后的治疗造成影响。也就是如图 20.3 和 20.4 一样, 不存在从 A_0 指向 A_1 的箭头。如果过往治疗会影响之后的治疗, 那会发生什么?

比如, 在我们 HIV 的例子中, 假设医生会用过往的治疗史 \bar{A}_{k-1} 决定是否在 $k=1$ 时给予治疗 A_k , 那此时过往治疗就会影响之后的治疗。为了表示过往治疗的影响, 我们在图 20.3 和 20.4 中添加从 A_0 指向 A_1 的箭头, 得到图 20.8 和 20.9。

在图 20.8 和 20.9 中, 治疗-混杂反馈依然存在, 不过此时控制 L_1 不足以阻断所有 A_1 和 Y 之间的后门路径。并且, 控制 L_1 会开放图 20.8 中的路径 $A_1 \leftarrow A_0 \rightarrow L_1 \leftarrow U_1 \rightarrow Y$ 和图 20.9 中的路径 $A_1 \leftarrow A_0 \leftarrow W_0 \rightarrow L_1 \leftarrow U_1 \rightarrow Y$ 。当然, 在过往治疗对结局的因果效应不为零时 (比如图 20.10), 无论是否存在治疗-混杂反馈, 我们都需要控制过往治疗, 因为此时 A_0 是 A_1 因果效应的混杂变量。

因此, 在时间点 k 的时序互换性需要控制过往治疗史 \bar{A}_{k-1} , 而只控制协变量 L 是不够的。这也是为什么在这一章和上一章, 所有的时序互换性表达式都会控制治疗史。

在我们估计非时异治疗的因果效应时, 过往治疗也非常重要。假设我们只想估计非时异治疗 A_1 (而非涉及 A_0 和 A_1 的治疗策略) 对 Y 的因果效应 (有时 A_1 的因果效应也被称为时异治疗 \bar{A} 的短期效应), 那么不调整 A_0 就可能会造成选择偏移 (如果存在治疗-混杂反馈) 和混杂 (如果 A_0 直接影响 Y)。也就是说, 在图 20.8 至 20.10 中, 如果 A_1 对 Y 的因果效应为零, 那么我们的估 255 计值 $E[Y | A_1 = 1, L_1] - E[Y | A_1 = 0, L_1]$ 不会等于 0。在实际中, 如果我们只关心当前治疗情况, 那我们可以调整过往治疗史, 从而消除偏移与混杂。这就是为什么在有的研究中, 研究者只分析没有过往治疗的参与者。

不过, 当过往治疗存在测量误差的时候, 调整过往治疗可能会带来额外的偏移。在第 9.3 小节我们讨论过, 调整有误差的混杂会给估计值带来不可预料的偏移。在我们 HIV 的例子中, 假设研究者使用问卷收集参与者的过往治疗情况, 那就不可避免的存在测量误差。此时, 研究者得到的是测量后的变量 A_0^* , 而非真实的 A_0 。为了表示这一情形, 我们可以在图 20.8 至 20.10 中增加一个 A_0^* , 以及从 A_0 指向 A_0^* 的箭头, 从而我们可以清楚看到控制 A_0^* 并不能完全阻断 A_1 和 Y 之间途经 A_0 的后门路径。因此, 控制了 A_0^* 和 L_1 之后, 研究者依然会观察到 A_1 和 Y 之间的相关性, 即使这两者之间没有因果关系。也因此, 当治疗是时异性的時候, 我们会发现, 治疗的误差——即使是独立无差型——就算在零值下也会造成偏移, 这是因为我们的过往治疗是当前治疗的混杂变量。另外, 如果过往治疗对结局有直接因果效应, 那么调整有误差的过往治疗会夸大当前治疗的实际因果效应。

第二十章精讲点和知识点

精讲点 20.1: 有向无环图中的反馈 (原书第 248 页)

我们常用来表示因果关系的有向“无环”图 (如图 20.1) 也可以用来表示治疗-混杂反馈中的反馈环, 只需要我们把反馈在时间上的顺序表示出来即可, 也即类似 $A_{k-1} \rightarrow L_k \rightarrow A_k \rightarrow L_{k+1}$ 。

用无环图表示反馈环需要我们将时间视作一个离散的变量。也就是说, 我们会认为治疗和变量在每个区间 $[k, k+1]$ 都可能发生变化, 但我们不知道在具体哪一点会发生变化。在实际中, 将时间离散化不是坏事, 我们可以根据数据的要求适当地选取每个区间的长度。比如, 在 HIV 研究中, 患者可能一个月或者更长一段时间才去见一次医生, 那么时间区间就可以用月为单位。在其他例子中, 也可以用年或者天。我们在第十七章也讨论过, 在现实中, 研究者都是用离散的单位 (年、月、日等) 衡量时间。因此, 将时间离散化有时不是一个选择, 而是不得不做的事。

精讲点 20.2: 因果路径上的混杂变量 (原书第 252 页)

如果混杂变量 L_1 被过往治疗影响, 那么即使 L_1 不在治疗和结局之间的因果路径上, 控制 L_1 依然会造成选择偏移。而在我们的图 20.5 和 20.6 中, 也不存在这样的因果路径。

另一方面, 在图 20.7 中, A_1 的混杂变量 L_1 在 A_0 和 Y 之间的因果路径上, 也即存在 $A_0 \rightarrow L_1 \rightarrow Y$ 。如果此时的 U_1 不是 L_1 和 Y 的共同诱因, 那么 L_1 各分层中 A_0 对 Y 直接效应 (也即不通过 L_1 的效应) 的估计值就是无偏的, 但是 \bar{A}_1 对 Y 总效应的估计值是有偏的, 因为 A_0 对 Y 的部分因果效应是通过 L_1 中介的。

我们很多时候不会将治疗和结局因果路径上的变量视作混杂变量, 因为调整这些变量会给我们的效应估计造成偏移。然而, 这一经验在时异治疗中不再适用。不过, 不是所有的方法都会在这种情况下造成偏移。使用传统方法——比如分层分析——调整这些变量会造成偏移, 但是使用 G-方法调整这些变量则不会。

知识点 20.1: G-零值检验 (原书第 250 页)

假设极端零假设成立。此时我们观测到的结局 Y 就是反事实结局 Y^g 。在这种情况下, 如果只有两个时间点, Y^g 的时序互换性能够写成 $Y \perp\!\!\!\perp A_0 | L_0$ 以及 $Y \perp\!\!\!\perp A_1 | (A_0 = g(L_0), L_0, L_1)$ 。其中, 第一个独立性意味着在 L_0 的分层中, A_1 对 Y 没有因果效应; 而第二个独立性意味着在 L_1 和 A_0 的任

意分层中, A_1 对 Y 没有因果效应。因此, 在时序互换性成立的时候, 对这两个独立性的检验也就是对极端零假设的检验, 这被称为 G-零值检验。

与之相反的是, G-零值定理认为, 如果这两个独立性成立, 那么 Y^g 和 $E[Y^g]$ 的分布对任意 g 而言都是一样的, 也就等于我们观测到的 Y 。注意, 上述相等关系只在忠实性成立的情况下才会成立, 也即不会出现多个因果效应相互抵消从而为零的情况。我们在精讲点 6.2 中讨论过忠实性, 并在本书都假设忠实性成立。

第二十章图表

Table 20.1

N	A_0	L_1	A_1	Mean Y
2400	0	0	0	84
1600	0	0	1	84
2400	0	1	0	52
9600	0	1	1	52
4800	1	0	0	76
3200	1	0	1	76
1600	1	1	0	44
6400	1	1	1	44

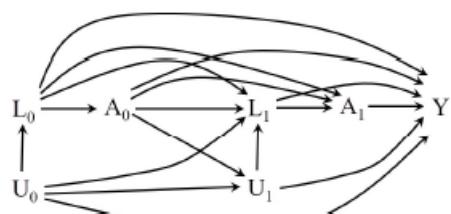


Figure 20.1

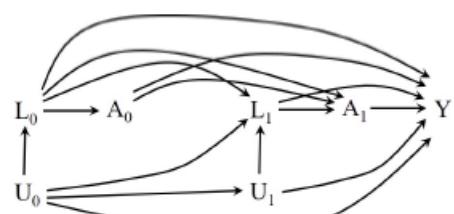


Figure 20.2

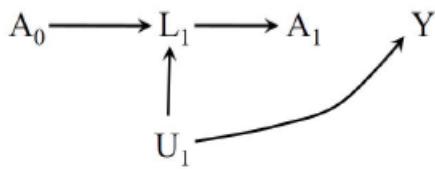


Figure 20.3

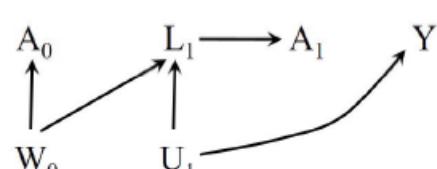


Figure 20.4

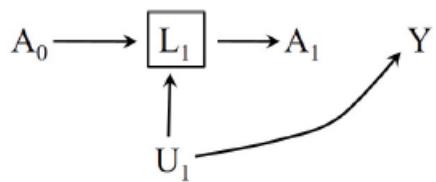


Figure 20.5

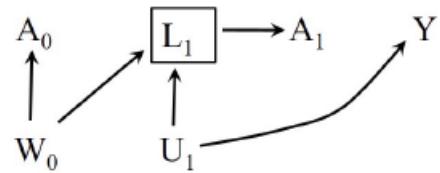


Figure 20.6

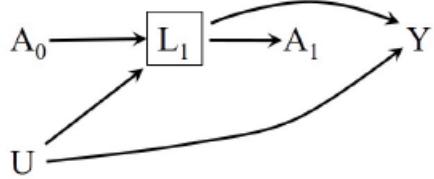


Figure 20.7

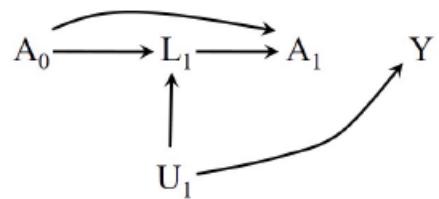


Figure 20.8

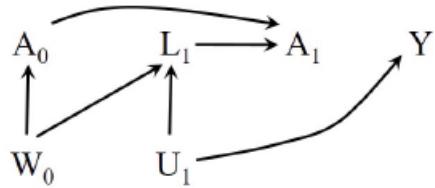


Figure 20.9

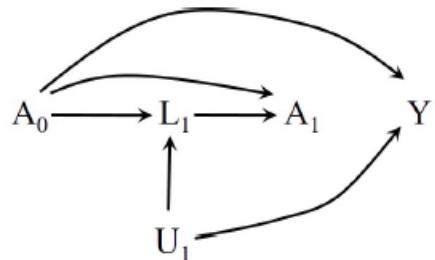


Figure 20.10

第二十一章 时异治疗的 G-方法

257 在上一章, 我们讨论了含时异治疗的数据以及治疗-混杂反馈, 并论证了为什么传统方法不适用于处理涉及时异治疗的情形——即使时异治疗对结局的因果效应为零, 传统方法也会造成偏移, 从而使我们的估计值偏离零值。

本章我们将讨论存在治疗-混杂反馈的时候, 如何使用 G-方法——也即 G-公式、逆概率加权、G-估算、以及它们的双重稳健估计形式——估计时异治疗的因果效应。我们将会使用和上一章一样的数据, 并用 G-方法给出正确的效应估计值 (也即零)。我们在非时异变量中已经介绍过这几种方法, G-公式参见第十三章, 逆概率加权和边缘结构模型参见第十二章, G-估算和结构嵌入模型参见第十五章。在本章, 我们默认第十九章讨论的可识别性假设成立 (也即时序互换性、正数性、和一致性), 并在这些假设下用上述 G-方法比较不同静态策略的因果效应。

21.1 时异治疗的 G-公式

我们再考察一下表 20.1 中的数据, 为了方便, 我们把它移到了本章, 并记为表 21.1。假设我们只对非时异治疗 A_1 的因果效应感兴趣, 也即, 假设我们只想比较 $E[Y^{a_1=1}]$ 和 $E[Y^{a_1=0}]$ 。在本书第一和第二部分我们已经知道, 如果可识别性假设成立, 那 $E[Y^{a_1}]$ 就是

$E[Y | A_1 = a_1, L_1 = l_1]$ 的加权均值。具体而言, 可以表述为:

$$E[Y^{a_1}] = \sum_{l_1} E[Y | A_1 = a_1, L_1 = l_1] f(l_1), \text{ 其中 } f(l_1) = \Pr[L_1 = l_1]$$

这一表述就是 G-公式。在有界互换性成立的情况下, G-公式就是根据研究人群的混杂变量对结局进行标准化处理。

表 21.1 中的数据来自于一项时序性随机试验, 其中 $\bar{A} = (A_0, A_1)$ 是时异性的, 并且存在治疗-混杂反馈。接下来我们将上述 G-公式扩展到时异治疗。

258 $E[Y^{a_0, a_1}]$ 的 G-公式就是 $E[Y | A_0 = a_0, A_1 = a_1, L_1 = l_1]$ 的加权平均。在 (静态) 时序互换性成立的情况下, 有:

$$E[Y^{a_0, a_1}] = \sum_{l_1} E[Y | A_0 = a_0, A_1 = a_1, L_1 = l_1] f(l_1 | a_0)$$

也就是说, 在可识别性假设成立的情况下, 根据混杂变量进行标准化后的结局均值就是反事实结局, 而这一表达式中的每一个因子都需要控制过往治疗变量和协变量。对非时异变量而言, 我们不需要控制过往治疗史和协变量史, 因为我们只有一个时间点。

注意, 上述表达式只有在 $f(l_1 | a_0) \neq 0$ 、且 $(A_0 = a_0, A_1 = a_1, L_1 = l_1)$ 的人数不为零的时候才是良定的(也即可以计算的), 这一条件等同于我们在知识点 19.2 中所说的正数性。

让我们将 G-公式应用到表 21.1 的数据中, 可得 $E[Y^{a_0=0, a_1=0}] = 84 \times 0.25 + 52 \times 0.75 = 60$, $E[Y^{a_0=1, a_1=1}] = 76 \times 0.50 + 44 \times 0.50 = 60$, 因而可知 $E[Y^{a_0=1, a_1=1}] - E[Y^{a_0=0, a_1=0}] = 0$ 。此时, 传统方法不能得到因果效应的正确估计值, 而 G-公式可以。

另一种理解 G-公式的方法是把它理解成一种模拟仿真。在时序互换性成立的情况下, G-公式模拟了如果每个人的治疗策略都是 \bar{a} 时的结局 $Y^{\bar{a}}$ 和协变量史 $\bar{L}^{\bar{a}}$ 。换句话说, G-公式模拟了治疗策略为 \bar{a} 时 ($Y^{\bar{a}}, \bar{L}^{\bar{a}}$) 的联合分布。我们可以用图来理解 G-公式。图 21.1 的树状图对应的是表 21.1 数据的分布情况。在可识别性假设成立的情况下, G-公式可以被视为构建了一个新的树状图, 如图 21.2 所示, 其表示所有人的治疗策略都是“总是会治疗”($a_0 = 0, a_1 = 1$) 时我们观测到的反事实结局。

为了模拟这一反事实结局, 我们需要: (1) 在 $k = 0$ 和 $k = 1$ 时把接受治疗的概率固定为 1; (2) 计算原人群中的 $\Pr[L_1 = l_1 | A_0 = a_0]$ 和 $E[Y | A_0 = a_0, A_1 = a_1, L_1 = l_1]$, 并把它们代入到模拟人群中。

(在时序互换性假设下, $\Pr[L_1 = l_1 | A_0 = a_0] = \Pr[L_1^{a_0} = l_1]$, 且
 $E[Y | A_0 = a_0, A_1 = a_1, L_1 = l_1] = E[Y^{a_0, a_1} | L_1^{a_0} = l_1]$, 因而 G-公式就是
 $E[Y^{a_0, a_1}] = \sum_l E[Y^{a_0, a_1} | L_1^{a_0} = l_1] \Pr[L_1^{a_0} = l_1]$ 。)

不过要注意两点。第一, G-公式的结果取决于 L 。比如, 如果我们没有 L_1 的数据, 那 G-公式中就不再有 L_1 , 于是变为 $E[Y | A_0 = a_0, A_1 = a_1]$, 此时 G-公式的结果就不再有因果性意义。

(对于本书中所有因果图而言, 包含了未测变量(比如 U 和 W) 的 G-公式也是正确的。然而在现实中, 未测变量意味着研究者不可能有这些数据。)

第二, 即使 G-公式的结果有因果性意义, G-公式的每一个因子也不一定有因果性意义。比如, 在图 20.9 中, 仅有静态时序互换性成立。此时, 无论是否存在从 A_0 或 A_1 指向 Y 的箭头, 整个 G-公式的结果依然可以被视为 $E[Y^{\bar{a}}]$, 而组成 G-公式的两个因子——

$E[Y | A_0 = a_0, A_1 = a_1, L_1 = l_1]$ 和 $\Pr[L_1 = l_1 | A_0 = a_0]$ ——并没有任何因果性意义。也就是说,

$E[Y | A_0 = a_0, A_1 = a_1, L_1 = l_1] \neq E[Y^{a_0, a_1} | L_1^{a_0} = l_1]$, $\Pr[L_1 = l_1 | A_0 = a_0] \neq \Pr[L_1^{a_0} = l_1]$ 。不过, 在图 20.1 和 20.2 所代表的时序性随机试验中, 最后两个等式成立。

将 G-公式扩展到高维情形中可得:

$$\sum_{\bar{a}} E[Y | \bar{A} = \bar{a}, \bar{L} = \bar{l}] \prod_{k=0}^K f(l_k | \bar{a}_{k-1}, \bar{l}_{k-1})$$

260 这一公式表示对所有的 \bar{l} 历史求和 (\bar{l}_{k-1} 表示从 0 到 $k-1$ 时的历史)。在时序互换性成立的情况下, 这一表达式等于治疗策略 \bar{a} 的反事实结局 $E[Y^{\bar{a}}]$ 。精讲点 21.1 对过往史给出了更细致的定义。知识点 21.1 给出了更广义的 G-公式。

在实践中, 面对高维数据, 我们不可能直接计算出 G-公式中的每一个因子。因而, 我们只能估计 $E[Y | \bar{A} = \bar{a}, \bar{L} = \bar{l}]$ 和 $f(l_k | \bar{a}_{k-1}, \bar{l}_{k-1})$ 。比如, 我们可以用线性模型估计结局变量的均值 $E[Y | \bar{A} = \bar{a}, \bar{L} = \bar{l}]$, 用 logistic 回归模型估计离散变量 L_k 的分布 (我们可以不使用模型得到 L_0 的分布, 参见第 13.3 小节)。从这些模型中得到的估计值 $\hat{E}[Y | \bar{A} = \bar{a}, \bar{L} = \bar{l}]$ 和 $\hat{f}(l_k | \bar{a}_{k-1}, \bar{l}_{k-1})$ 会被代入到 G-公式。在第十三章, 我们将这一方法称为“插入式 G-公式”, 如果插入的估计值是从参数模型中得到的, 那我们就将这一方法称为“参数 G-公式”。

为了简便, 本章的 G-公式只涉及命定的静态治疗策略。不过 G-公式可以适用于任意策略, 不管是命定的还是随机的、静态的还是动态的。如果我们在时间点 k 对所有人都施加治疗 a_k , 并把 $f^{\text{int}}(a_k | \bar{a}_{k-1}, \bar{l}_k)$ 定义为 a_k 的条件概率, 那么 G-公式的一般形式为:

$$\sum_{\bar{a}} E[Y | \bar{A} = \bar{a}, \bar{L} = \bar{l}] \prod_{k=0}^K f(l_k | \bar{a}_{k-1}, \bar{l}_{k-1}) \prod_{k=0}^K f^{\text{int}}(a_k | \bar{a}_{k-1}, \bar{l}_k)$$

如果治疗策略是命定的, 那么 $f^{\text{int}}(a_k | \bar{a}_{k-1}, \bar{l}_k) = 1$, 因此就不需写出。比如, 在“总是会治疗”策略之下, 也即 $\bar{a} = (1, 1, \dots, 1)$, 我们在任意时间点 k 都有 $f^{\text{int}}(1 | \bar{a}_{k-1}, \bar{l}_k) = 1$ 。不过, 在其他类型的治疗策略之下, $f^{\text{int}}(a_k | \bar{a}_{k-1}, \bar{l}_k)$ 不一定等于 1。举个例子, 如果我们的策略是“在每个时间点, 被试接受治疗的概率是 0.3”, 那么就有 $f^{\text{int}}(1 | \bar{a}_{k-1}, \bar{l}_k) = 0.3$ 。目前已有代码能运行 G-公式, 并且适用于任意治疗策略。

(代码: 运行 G-公式的 R 包 *gformula* 能从 CRAN 上下载。G-公式的 SAS 宏程序 *GFORMULA* 也能从本书网站下载。)

21.2 时异治疗的逆概率加权

假设我们只对非时异治疗 A_1 的因果效应感兴趣。也即，假设我们只想比较 $E[Y^{a_1=1}]$ 和 $E[Y^{a_1=0}]$ 。利用第十二章的逆概率加权法，我们首先需要计算稳定权重 $SW^{A_1} = f(A_1) / f(A_1 | L_1)$ 或非稳定权重 $W^{A_1} = 1 / f(A_1 | L_1)$ ，然后利用权重构建虚拟人群。虚拟人群中的结局均值 $E_{ps}[Y | A_1 = a_1]$ 就是反事实结局 $E[Y^{a_1}]$ 。我们可以将权重的分母视为该个体在控制了混杂变量之后接受治疗的概率。

262

如果治疗和混杂是时异的，那么我们就需要扩展权重的表达式。对于时异治疗 $\bar{A} = (A_0, A_1)$ 和时异混杂 $\bar{L} = (L_0, L_1)$ 而言，非稳定权重是：

$$W^{\bar{A}} = \frac{1}{f(A_0 | L_0)} \times \frac{1}{f(A_1 | A_0, L_0, L_1)} = \prod_{k=0}^1 \frac{1}{f(A_k | \bar{A}_{k-1}, \bar{L}_k)}$$

稳定权重是：

$$SW^{\bar{A}} = \frac{f(A_0)}{f(A_0 | L_0)} \times \frac{f(A_1 | A_0)}{f(A_1 | A_0, L_0, L_1)} = \prod_{k=0}^1 \frac{f(A_k | \bar{A}_{k-1})}{f(A_k | \bar{A}_{k-1}, \bar{L}_k)}$$

根据定义， $A_{-1} = 0$ 。时异治疗逆概率权重的分母是控制了协变量史之后接受这一治疗策略的概率。

假设我们想比较反事实结局 $E[Y^{a_0=1, a_1=1}]$ 和 $E[Y^{a_0=0, a_1=0}]$ 。在静态策略的可识别性假设下，反事实结局 $E[Y^{a_0, a_1}]$ 等于虚拟人群中相应治疗策略的结局均值 $E_{ps}[Y | A_0 = a_0, A_1 = a_1]$ 。

现在我们对表 21.1 中的数据使用逆概率加权法。图 21.3 给出了非稳定权重、稳定权重、以及它们各自对应的虚拟人群。非稳定权重构建的虚拟人群共有 128000 人，是原人群数 32000 乘以 4，也即静态策略的个数。因为研究中没有 L_0 ，所以分母是 $f(A_0)f(A_1 | A_0, L_1)$ 。

在虚拟人群中，共有 32000 人治疗策略为 $(A_0 = 0, A_1 = 0)$ ，从图 21.3 中我们可以得到

$$E_{ps}[Y | A_0 = a_0, A_1 = a_1] = 84 \times \frac{8000}{32000} + 52 \times \frac{24000}{32000} = 60，\text{ 因为 } E[Y^{a_0=0, a_1=0}] = 60。同理有}$$

$E[Y^{a_0=1, a_1=1}] = 60$ ，因此 $E[Y^{a_0=0, a_1=0}] - E[Y^{a_0=1, a_1=1}] = 0$ 。同 G-公式一样，逆概率加权能够正确估计治疗策略的因果效应。

(使用稳定权重我们依然可以得到因果效应为零。区别在于, 在非稳定权重构建的虚拟人群中,

中, $\Pr_{ps}[A_k = 1 | \bar{A}_{k-1}, \bar{L}_k] = \frac{1}{2}$, 而在稳定权重构建的虚拟人群中,

$$\Pr_{ps}[A_k = 1 | \bar{A}_{k-1}, \bar{L}_k] = \Pr[A_k = 1 | \bar{A}_{k-1}]。)$$

注意, 就算可识别性假设不成立, 使用 G-公式和逆概率加权得到的结果依然是一样的, 只是此时两者都没有因果性意义。

- 263 接下来我们将逆概率加权推广到有多个时间点 $k = 0, 1, \dots, K$ 的高维情形。此时, 非稳定逆概率权重是:

$$W^{\bar{A}} = \prod_{k=0}^K \frac{1}{f(A_k | \bar{A}_{k-1}, \bar{L}_k)}$$

稳定逆概率权重是:

$$SW^{\bar{A}} = \prod_{k=0}^K \frac{f(A_k | \bar{A}_{k-1})}{f(A_k | \bar{A}_{k-1}, \bar{L}_k)}$$

如果可识别性假设成立, 这两个权重构建的虚拟人群满足: (1) 虚拟人群和原人群的 $E[Y^{\bar{a}}]$ 相等; (2) 在每个时间点接受治疗的概率是常数 (稳定权重), 或只取决于过往治疗 (非稳定权重)。因此有 $E[Y^{\bar{a}}] - E[Y^{\bar{a}'}] = E_{ps}[Y | \bar{A} = \bar{a}] - E_{ps}[Y | \bar{A} = \bar{a}']$ 。

在时序性随机试验中, $f(A_k | \bar{A}_{k-1}, \bar{L}_k)$ 的值是事先设定的, 因此是已知的。所以我们就能直接估计 $E[Y^{\bar{a}}] - E[Y^{\bar{a}'}]$, 并且能保证结果是无偏的。但是在观察性研究中, 我们需要用观测数据估计 $f(A_k | \bar{A}_{k-1}, \bar{L}_k)$ 。面对高维数据时, 我们可以用 logistic 回归估计每个时间点治疗的概率 $\Pr[A_k = 1 | \bar{A}_{k-1}, \bar{L}_k]$, 然后用估计值计算权重以及 $E[Y^{\bar{a}}] - E[Y^{\bar{a}'}]$ 。如果我们的模型设定并不正确, 那么我们的最终结果可能是有偏的。对于稳定权重 $SW^{\bar{A}}$, 我们还需要估计分子 $f(A_k | \bar{A}_{k-1})$ 的值, 不过对这个分子而言, 就算模型是不正确的, $E[Y^{\bar{a}}] - E[Y^{\bar{a}'}]$ 依然可能是无偏的。

(在实践中, 最常用的方法是只拟合一个 $\Pr[A_k = 1 | \bar{A}_{k-1}, \bar{L}_k]$ 的模型, 而不是在每个时间点都拟合一个模型。这个模型会包含一个时间 k 的函数作为协变量。)

- 264 假设我们分别用参数 G-公式和逆概率加权得到两个 $E[Y^{\bar{a}}]$ 的估计值, 不过这两个估计值有所不同, 并且这个差异不能用随机变异性解释 (我们可以用自举法量化随机变异性造成的差

异), 那此时我们就知道参数 G-公式所用的模型或者逆概率加权所用的模型, 至少有一个是不正确的。不管可识别性假设是否成立, 这一结论都是正确的。因此, 我们应该尽可能用两种方法都估计一次。如果两种方法得到的估计值差异过大, 那我们就需要再检视我们的模型。在下一小节, 我们将讲述如何用双重稳健法解决模型的错误设定问题。

(不过, 就算用这两个方法得到的估计值相同, 也不能保证我们的模型设定是正确的, 因为它们可能往同一方向偏移同样的大小。)

同样, 就如我们上一小节讨论的一样, 治疗策略的数量可能会大大超过我们的样本量。因此我们需要使用其他方法估计 $E[Y^{\bar{a}}]$ 。比如, 我们可以假设治疗史 \bar{a} 对结局的因果效应是累计治疗

次数 $cum(\bar{a}) = \sum_{k=0}^K a_k$ 的线性函数, 所以我们可以拟合边缘结构模型:

$$E[Y^{\bar{a}}] = \beta_0 + \beta_1 cum(\bar{a})$$

这就类似我们第十二章讨论的非时异治疗的边缘结构模型。在模型左侧有 2^K 个未知量, 但我们在右侧只有两个参数: β_0 和 β_1 。 β_1 衡量的是时异治疗 \bar{A} 的因果效应均值。 $E[Y^{\bar{a}}] - E[Y^{\bar{a}=0}]$ 可以表示为 $\beta_1 \times cum(\bar{a})$ 。

(这是一个未饱和的边缘结构模型。)

为了估计边缘结构模型的这两个参数值, 我们需要在虚拟人群中用加权最小二乘法拟合以下线性模型:

$$E[Y | \bar{A}] = \theta_0 + \theta_1 cum(\bar{A})$$

在可识别性假设成立下, 相关性系数 θ_1 就是因果性系数 β_1 , 而估计值 $\hat{\beta}_1$ 的方差或置信区间可以用自举法计算(参见第十二章), 也可以用稳健方差法计算。对于未饱和的边缘结构模型而言, 使用稳定权重 $SW^{\bar{A}}$ 会比非稳定权重 $W^{\bar{A}}$ 得到更窄的置信区间, 因此我们会更经常使用 $SW^{\bar{A}}$ 。

当然, 如果我们的模型不正确, 那 $E[Y^{\bar{a}}]$ 的估计值也就不正确。也就是说, 如果治疗策略的因果效应不是累计治疗次数 $cum(\bar{a})$ 的线性函数, 而是其他形式, 那么此时我们的估计值就是无效的。比如, 真实的函数形式的是累计治疗次数的二次函数, 并且治疗策略的因果效应还额外取决于最后 5 次治疗的累计次数, 那我们可以拟合以下模型:

$$E[Y | \bar{A}] = \theta_0 + \theta_1 cum(\bar{A}) + \theta_2 cum_{-5}(\bar{A}) + \theta_3 cum(\bar{A})^2$$

我们依然会用 $SW^{\bar{A}}$ 或 $W^{\bar{A}}$ 对这个模型进行加权。我们可以通过检验 $\theta_2 = \theta_3 = 0$ 判断模型是否正确。知识点 21.2 论述了 G-零值悖论, 而逆概率加权的边缘结构模型并不会受到这一悖论的影响。在实践中, 我们可以使用治疗史 \bar{A} 的不同函数形式, 比如三次函数、样条等等。

(如果使用稳定权重 $SW^{\bar{A}}$, 上述提到的检验会有更强的统计效力。)

最后, 我们可以用边缘结构模型探索部分变量 V 的效应修饰作用, 我们在 12.5 小节讨论过这一话题。比如, 对于一个研究初始的二分变量 V , 我们可以用以下模型探索它是否有效应修饰作用:

$$E[Y^{\bar{a}} | V] = \beta_0 + \beta_1 cum(\bar{a}) + \beta_2 V + \beta_3 cum(\bar{a})V$$

模型中的系数可以通过在数据中拟合常规线性模型 $E[Y | \bar{A}] = \theta_0 + \theta_1 cum(\bar{A}) + \theta_2 V + \theta_3 cum(\bar{A})V$

得到, 并且我们依然会用 $SW^{\bar{A}}$ 或 $W^{\bar{A}}$ 对模型进行加权。如果存在治疗-混杂反馈, 那么 V 就只能是基线变量, 也即研究初始时的变量。否则, 如果 V 当中包含了变量 L_k ($k > 0$), 那么即使治疗的因果效应在各时间点为零, θ_1 和 θ_3 的估计值也可能不为零。

266

接下来我们将讲述边缘结构模型的双重稳健估计。

21.3 时异治疗的双重稳健估计

本书第二部分简要提到了同时结合逆概率加权和 G-公式的双重稳健估计。我们已经知道, 逆概率加权是对治疗建模, 而 G-公式是对结局建模。双重稳健估计则会对治疗和结局同时建模, 并且只需其中一个模型正确就能给出正确的估计值。知识点 13.2 论述了非时异治疗的双重稳健估计。在本小节, 我们会把双重稳健估计推广到时异治疗当中。

(双重稳健估计的优势在于我们会有两次得到正确结果的机会。)

对非时异治疗 A 而言, 双重稳健估计共有三步。第一步是估计治疗的概率 $\widehat{Pr}[A=1 | L]$, 并通过这一估计值计算逆概率权重 \hat{W}^A 。第二步是估计结局的预测值 $\hat{E}[Y | A=a, L=l, R]$, 其中如果 $A=1$, 那么 $R=\hat{W}^A$; 如果 $A=0$, 那么 $R=-\hat{W}^A$ 。第三步是分别在 $A=1$ 和 $A=0$ 之下对 $\hat{E}[Y | A=a, L=l, R]$ 进行标准化。而标准化结局的差值也就是因果效应 $E[Y^{a=1}] - E[Y^{a=0}]$ 。只要我们估计治疗概率的模型是正确的, 或者估计结局的模型是正确, 那么双重稳健估计就能给出有效的估计值。

(我们也可以使用“聪明变量”得到时异治疗与非时异治疗的双重稳健。参见 Bang 和 Robins 在 2005 年所著论文。)

接下来我们将双重稳健估计扩展到时异治疗, 此时我们想比较的是两个不同治疗策略 \bar{a} 和 \bar{a}' 的反事实结局 $E[Y^{\bar{a}}]$ 和 $E[Y^{\bar{a}'}]$ 。同非时异治疗一样, 时异治疗的双重稳健估计也分为三个步骤, 不过每一步的具体做法会和非时异治疗有所差别。接下来我们会以“总是会治疗”策略的反事实结局 $E[Y^{\bar{a}}]$ 为例。

第一步, 拟合 $\Pr[A_k = 1 | \bar{A}_{k-1}, \bar{L}_k]$ 的回归模型, 并用模型的估计值计算时异治疗在每一个时间点 m 的逆概率权重 $W^{\bar{A}_m} = \prod_{k=0}^m \frac{1}{f(A_k | \bar{A}_{k-1}, \bar{L}_k)}$ 。其中, 如果 $A_k = 1$, 那么

$$f(A_k | \bar{A}_{k-1}, \bar{L}_k) = \Pr[A_k = 1 | \bar{A}_{k-1}, \bar{L}_k]; \text{ 如果 } A_k = 0, \text{ 那么}$$

267 $f(A_k | \bar{A}_{k-1}, \bar{L}_k) = \Pr[A_k = 0 | \bar{A}_{k-1}, \bar{L}_k]$ 。此时与上一小节不同的是, 我们会在每一个时间点都估计一个逆概率权重, 而非只有一个总的权重。比如, 如果我们拟合的参数模型是

$\Pr[A_k = 1 | \bar{A}_{k-1}, \bar{L}_k] = \alpha_{0,k} + \alpha_1 A_{k-1} + \alpha_2 L_k$, 那么在表 21.1 的数据中, 两个时间点处的估计值就是 $\hat{\Pr}[A_1 = 1 | A_0, \bar{L}_1] = \hat{\alpha}_{0,1} + \hat{\alpha}_1 A_0 + \hat{\alpha}_2 L_1$ 和 $\hat{\Pr}[A_0 = 1 | \bar{L}_0] = \hat{\alpha}_{0,0} + \hat{\alpha}_2 L_0$ (此时 $A_{-1} \equiv 0$), 然后利

用估计值我们可以计算权重 $\hat{W}^{\bar{A}_m} = \prod_{k=0}^m \frac{1}{\hat{f}(A_k | \bar{A}_{k-1}, \bar{L}_k)}$ 。此外, 我们还需要计算一个改进后的权

重 $\hat{W}^{\bar{A}_{m-1}, a_m=1} = \hat{W}^{\bar{A}_{m-1}} \times \frac{1}{\hat{f}(a_m | \bar{A}_{m-1}, \bar{L}_m)}$, 这表示在时间点 m 的治疗状态被设定为“总是会治疗”

策略。

第二步, 我们需要在每个时间点 m 拟合一个回归模型, 并且拟合的顺序是从最后一个时间点 K 到 $m=0$, 也即倒序。这一系列模型有两个特点。其一, 时异治疗的逆概率权重 $\hat{W}^{\bar{A}_m}$ 会作为协变量放在模型中。其二, 只有在最后一个时间点 K , 模型的因变量是结局 Y 。在其余时间点, 模型的因变量是上一个模型 (也即后一个时间点的模型) 所给出的预测值 \hat{T}_{m+1} 。

(因为时异治疗的双重稳健估计依赖于时序性的回归模型, 所以我们需要按顺序在每个时间点拟合一次模型, 而不能同时拟合这些模型。)

比如, 我们会拟合以下模型:

$$E[\hat{T}_{m+1} | \bar{A}_m, \bar{L}_m] = \theta_{0,m} + \theta_1 cum(\bar{A}_m) + \theta_2 L_m + \theta_3 \hat{W}_m^4$$

其中 $cum(\bar{A}_m)$ 表示累计治疗次数, 同上一小节一样。接下来我们以两个时间点的简单情形 (也即 $K=1$) 讨论变量 \hat{T}_{m+1} 的定义。(知识点 21.3 给出了多个时间点的广义定义。)

我们先拟合 $E[\hat{T}_2 | \bar{A}_1, \bar{L}_1] = E[Y | \bar{A}_1, \bar{L}_1] = \theta_{0,1} + \theta_1 cum(\bar{A}_1) + \theta_2 L_1 + \theta_3 \hat{W}^{\bar{A}_1}$, 此时 $\hat{T}_2 = Y$ 。利用模型中的参数估计值 $\hat{\theta}$, 我们可以计算 $A_1=1$ 时的预测值, 并且此时 $\hat{W}^{\bar{A}_1} = \hat{W}^{A_0, a_1=1}$ 。于是, 对每一个个体 i , 我们有 $\hat{T}_{1i} = \hat{\theta}_{0,1} + \hat{\theta}_1 \times 2 + \hat{\theta}_2 L_{1i} + \hat{\theta}_3 \hat{W}_i^{A_0, a_1=1}$ 。这一预测值 \hat{T}_1 就是下一个模型的因变量。接下来我们再拟合模型 $E[\hat{T}_1 | A_0, L_0] = \theta_{0,0} + \theta_1 A_0 + \theta_2 L_0 + \theta_3 \hat{W}^{\bar{A}_0}$, 并且计算 $A_0=1$ 的预测值, 也即 $\hat{T}_{0i} = \hat{\theta}_{0,0} + \hat{\theta}_1 \times 1 + \hat{\theta}_2 L_{0i} + \hat{\theta}_3 \hat{W}_i^{a_0=1}$ 。

第三步, 对 \hat{T}_0 进行标准化, 我们只需求其在人群中的均值即可。均值 $\hat{E}[\hat{T}_0]$ 就是反事实结局 $E[Y^{a_0=1, a_1=1}]$ 的双重稳健估计值。也即, 在互换性、正数性、和一致性成立的情况下, 只要以下三个条件中的一个成立, 那么这一估计值就是有效的: (1) 各个时间点治疗的模型是正确的; (2) 各个时间点结局的模型是正确的; (3) 治疗的模型在时间 0 到 k 是正确的, 且结局的模型在时间 $k+1$ 到 K 是正确的, 其中 $k < K$ 。最后一个条件也被称为 $k+1$ 稳健性。

($k+1$ 稳健性的更多细节请参考 Molina 等人 2007 年所著论文。)

重复上述步骤, 我们可以估计“从不会治疗”策略下的反事实结局 $E[Y^{a_0=0, a_1=0}]$ 。两次 \hat{T}_0 的差就是因果效应均值 $E[Y^{a_0=1, a_1=1}] - E[Y^{a_0=0, a_1=0}]$ 。

268 双重稳健估计的理论基础已经很完备, 但是在现实中, 因为缺少能简易运行的软件, 所以尚未被大多数研究者接受。不过, 我们预计, 在不久的将来, 我们会在大多数研究中看到双重稳健分析的应用。精讲点 21.2 论述了 G-公式的不同形式, 以及其和双重稳健估计之间的联系。

(van der Laan 和 Gruber 提出了一种新的双重稳健估计形式, 其中包含了数据自适应过程。详情参见他们 2012 年所著论文。)

21.4 时异治疗的 G-估算

在第十四章, 我们论述了如何使用结构嵌入模型估计非时异治疗的因果效应。彼时, 模型只有一个方程, 只有一个时间点。如果我们将结构嵌入模型扩展到时异治疗, 那么数据中有多少个

时间点, 我们就需要多少个方程。对表 21.2 中有两个时间点的时异治疗 $\bar{A} = (A_0, A_1)$ 而言, 我们

269 的结构嵌入模型会有两个方程。

在 $k=0$ 时, 有 $E[Y^{a_0, a_1=0} - Y^{a_0=0, a_1=0}] = \beta_0 a_0$ 。

在 $k=1$ 时, 有 $E[Y^{a_0, a_1} - Y^{a_0=0, a_1=0} | L_1^{a_0} = l_1, A_0 = a_0] = a_1 (\beta_{11} + \beta_{12} l_1 + \beta_{13} a_0 + \beta_{14} a_0 l_1)$ 。

第二个方程是对 4 种不同治疗和协变量历史下的治疗效应进行建模。第二个方程是饱和的, 因为其中的 4 个参数 β_1 对应的是 4 种不同的治疗和协变量历史。第一个方程是对 $k=0$ 时的治疗效应进行建模, 并且 $k=1$ 时的治疗被设定为零。第一个方程也是饱和的, 因为它只有一个参数 β_0 , 也就只对应 $k=0$ 时的治疗 (此时不存在过往治疗或者过往协变量)。

(如果 $a_0 = 0$, a_1 的因果效应: (1) 若 $L_1^{a_0=0} = 0$, 则为 β_{11} ; (2) 若 $L_1^{a_0=0} = 1$, 则为 $\beta_{11} + \beta_{12}$ 。如果 $a_0 = 1$, a_1 的因果效应: (1) 若 $L_1^{a_0=0} = 0$, 则为 $\beta_{11} + \beta_{13}$; (2) 若 $L_1^{a_0=0} = 1$, 则为 $\beta_{11} + \beta_{13} + \beta_{14}$ 。根据一致性, 有 $L_1^{a_0=0} = L_1$ 。)

上述两个方程也说明了为什么这一方法会被称为“嵌入”模型。第一个方程表示在最初时间点 $k=0$ 接受治疗、但之后不再接受治疗的因果效应。第二个方程表示在时间点 $k=1$ 接受治疗、但之后不再接受治疗的因果效应。

接下来我们在 $K=1$ (也即只有两个时间点) 的情况下使用结构嵌入模型估计因果效应。同 270 第十四章一样, 我们先考察加成保序性。对每个个体 i , 有: $Y_i^{a_0, 0} = Y_i^{0, 0} + \psi_0 a_0$,

$$Y_i^{a_0, a_1} = Y_i^{a_0, 0} + \psi_{11} a_1 + \psi_{12} a_1 L_{1,i}^{a_0} + \psi_{13} a_1 a_0 + \psi_{14} a_1 a_0 L_{1,i}^{a_0} \quad (\text{为了简便, 我们用 } Y_i^{0, 0} \text{ 表示 } Y_i^{a_0=0, a_1=0})$$

第一个方程是保序性模型, 其中 ψ_0 对每个个体而言都是一样的。如果个体 i 的 $Y_i^{0, 0}$ 大于个体 j 的 $Y_j^{0, 0}$, 那么也自然有 $Y_i^{1, 0}$ 大于 $Y_j^{1, 0}$ 。在第二个方程中, 如果个体 i 和 j 的 $L_1^{a_0=1}$ 相同, 那么 $Y_i^{1, 0}$ 大于 $Y_j^{1, 0}$ 时有 $Y_i^{1, 1}$ 大于 $Y_j^{1, 1}$ 。因为这一保序性需要取决于其他因素 (比如 $L_1^{a_0=1}$ 的取值), 所以我们称之为条件保序性。

我们在第十四章讨论过, 因为个体之间的差异, 所以保序性基本上是不成立的。这也是为什么我们只关注结构嵌入模型给出的均值, 这样一来就没涉及个体之间是否存在差异。不过, 如果加强版的可识别性假设成立, 那么即使保序性不成立, 由 G-估算得到的 ψ 的估计值也会和普通回归模型得到的参数 β 一致。

(Robins 给出了上述命题的证明 (1994)。注意, 用 G-估算拟合未饱和结构嵌入模型的时候, 正数性不是必要条件。)

G-估算的第一步是将模型和观测数据相结合, 就如第十四章所做的一样。首先, 根据一致性, 反事实结局 Y^{a_0, a_1} 等于人群中治疗取值分别为 a_0 和 a_1 的观测结局 Y , 即在 ($A_0 = a_0, A_1 = a_1$) 的人群中, $Y^{a_0, a_1} = Y^{A_0, A_1} = Y$ 。同理, 如果 ($A_0 = a_0, A_1 = 0$), 有 $Y^{a_0, 0} = Y^{A_0, 0}$; 如果 $A_0 = a_0$, 有 $L_1^{a_0} = L_1$ 。因此, 上面提到的结构嵌入模型可以写作:

$$Y_i^{A_0, 0} = Y - (\psi_{11} A_1 + \psi_{12} A_1 L_1 + \psi_{13} A_1 A_0 + \psi_{14} A_1 A_0 L_1)$$

$$Y_i^{0, 0} = Y_i^{A_0, 0} + \psi_0 A_0$$

(为了简便, 我们忽略了表示特定个体的下角标 i 。)

第二步, 使用观测数据计算反事实变量 $H_0(\psi^\dagger)$ 和 $H_1(\psi^\dagger)$ 。我们用某特定值 ψ^\dagger 替代真实值 ψ , 从而有:

$$H_1(\psi^\dagger) = Y - (\psi_{11}^\dagger A_1 + \psi_{12}^\dagger A_1 L_1 + \psi_{13}^\dagger A_1 A_0 + \psi_{14}^\dagger A_1 A_0 L_1)$$

$$H_0(\psi^\dagger) = H_1(\psi^\dagger) - \psi_0^\dagger A_0$$

同第十四章一样, 我们的目标是找到等于真实值的 ψ^\dagger 。 ψ^\dagger 等于真实值时, 反事实变量 $H_k(\psi^\dagger)$ 就等于反事实结局 $Y^{\bar{a}_{k-1}, 0_k}$ 。接下来我们就可以用 G-估算计算参数估计值。精讲点 21.3 论述了如何在饱和结构嵌入模型中用 G-估算计算 ψ 的估计值。最后得到的结果是结构嵌入模型中的所有参数都是 0, 也就意味着在任何静态或动态策略下, $E[Y^g]$ 都等于 60。这一结果和我们 G-公式以及逆概率加权得到的结果一样。

271 在实际中, 我们的数据会有多个时间点, 每个时间点也会有多了协变量 L_k 。有多少个时间点, 结构嵌入模型就有多少个方程。在每个时间点, 结构嵌入模型的一般形式为:

$$E[Y^{\bar{a}_{k-1}, a_k} - Y^{\bar{a}_{k-1}, 0_k} | \bar{L}_k^{\bar{a}_{k-1}} = \bar{l}_k, \bar{A}_{k-1} = \bar{a}_{k-1}] = a_k \gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta)$$

其中 (\bar{a}_{k-1}, a_k) 是一个静态策略。这个静态策略表示时间点 0 到 $k-1$ 的治疗取值为 \bar{a}_{k-1} , 时间点 k 的治疗取值为 a_k 。 (\bar{a}_{k-1}, a_k) 和 $(\bar{a}_{k-1}, 0_k)$ 的区别在于, 前者在时间点 k 的治疗取值是 a_k , 后者则是 0。

(函数 $\gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta)$ 满足 $\gamma_k(\bar{a}_{k-1}, \bar{l}_k, 0) = 0$, 因而在零假设成立的情况下 $\beta = 0$ 。)

也就是说, 结构嵌入模型估计的是治疗策略在最后一个时间点 a_k 对结局均值的影响, 并且假设其与之前变量史 $(\bar{a}_{k-1}, \bar{l}_k)$ 之间的关系是 $\gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta)$ 。知识点 21.4 论述了结构嵌入模型和边缘结构模型之间的关系。

在我们的例子中, $K=1$, 因而 $\gamma_0(\bar{a}_{-1}, \bar{l}_0, \beta)$ 就是 β_0 (因为 \bar{l}_0 和 \bar{a}_{-1} 都可以视作 0), 而 $\gamma_1(\bar{a}_0, \bar{l}_1, \beta)$ 就等于 $\beta_{11} + \beta_{12}l_1 + \beta_{13}a_0 + \beta_{14}a_0l_1$ 。在每个时间点 k 我们关注的反事实变量, 可以表示为:

$$H_k(\psi^\dagger) = Y - \sum_{j=k}^K A_j \gamma_j(\bar{A}_{j-1}, \bar{L}_j, \psi^\dagger)$$

如果有多个时间点或协变量, 则我们需要拟合不饱和的结构嵌入模型。比如, 我们可以假设函数 $\gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta)$ 在所有时间点 k 都是一样的, 而最简单的形式就是 $\gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta) = \beta_1$, 也即假设最后一次治疗的效应和其余时间一样。其他假设包括 $\beta_1 + \beta_2 k$, 也即治疗的效应随着时间 k 线性变化; 以及 $\beta_1 + \beta_2 k + \beta_3 a_{k-1} + \beta_4 l_k + \beta_5 l_k a_{k-1}$, 也即过往治疗和协变量有效应修饰作用。

接下来我们说明如何在多个时间点的结构嵌入模型中使用 G-估算。首先假设我们的不饱和模型是 $\gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta) = \beta_1$ 。然后从观测数据中可以计算不同 ψ^\dagger 取值下的 $H_k(\psi^\dagger) = Y - \sum_{j=k}^K A_j \psi^\dagger$ 。接下来我们要选取一个区间, 上下界是 ψ_{low} 和 ψ_{up} , 这个区间需要包含真实值 ψ 。然后遍历区间中的所有可能取值, 并计算相应的 $H_k(\psi^\dagger)$ 。比如, 我们可以以 0.1 为单位, 分别计算 ψ_{low} , $\psi_{low} + 0.1$, $\psi_{low} + 0.2$ ……直到 ψ_{up} 所对应的 $H_k(\psi^\dagger)$ 。

接下来我们需要在每个时间点拟合 logistic 回归模型:

$$\text{logit Pr}[A_k = 1 | H_k(\psi^\dagger), \bar{L}_k, \bar{A}_{k-1}] = \alpha_0 + \alpha_1 H_k(\psi^\dagger) + \alpha_2 W_k$$

其中 $W_k = w_k(\bar{L}_k, \bar{A}_{k-1})$ 是一个根据治疗史和协变量史 $(\bar{L}_k, \bar{A}_{k-1})$ 计算得到的向量, α_2 是未知参数的向量, 而每一个个体都会有 $K+1$ 次观测数据。在时序互换性和一致性假设下, 如果某个 ψ^\dagger 能使模型中参数 α_1 的估计值为零, 那么这个 ψ^\dagger 就是 ψ 的估计值, 也即 β 的估计值。

上述步骤就是时异治疗的 G-估算。为了简便, 我们的结构嵌入模型只有一个参数 β_1 , 也即我们假设治疗的因果效应不随时间 k 、治疗史或协变量史改变。如果参数 β 是一个向量, 更具体一

作者: Miguel A. Hernan, James M. Robins;

翻译: 罗家俊

些, 我们假设 $\gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta) = \beta_1 + \beta_2 k + \beta_3 a_{k-1} + \beta_4 l_k + \beta_5 l_k a_{k-1}$, 那此时 β 就是一个五维的向量。

要估计 5 个参数, 在最后一步 logistic 模型中我们就需要 5 个额外的协变量。比如, 我们可以拟合下述模型:

$$\text{logit Pr}[A_k = 1 | H_k(\psi^\dagger), \bar{L}_k, \bar{A}_{k-1}] = \alpha_0 + H_k(\psi^\dagger)(\alpha_1 + \alpha_2 k + \alpha_3 a_{k-1} + \alpha_4 l_k + \alpha_5 l_k a_{k-1}) + \alpha_6 W_k$$

不同的协变量形式并不会影响 β 估计值的一致性, 只会影响置信区间的宽度。

(在 $\alpha_1 = 0$ 的统计检验中, 使得 $P > 0.05$ 的 ψ^\dagger 构成的区间, 就是 ψ 的 95% 置信区间。)

此时我们就需要在一个五维空间中遍历 β 的可能取值。如果每一个参数有 20 个候选取值, 我们就要计算 20^5 次。 β 的估计值需要使 $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ 五个变量的估计值都为零。当 β 的维度大于 2 的时候, 遍历法需要大量计算。不过如果结构嵌入模型是线性的, 就如本小节介绍的一样, 那么 β 的估计值就有闭合形式, 更多细节参见知识点 21.5。

(5 个自由度的 score 检验在 5% 水平上不拒绝原假设时的取值, 就是 β 每一部分 β_j 的 95% 联合置信区间。为了计算方便, 我们也可以计算 95% Wald 置信区间, 也即点估计加减 1.96 倍标准差。)

通过 G-估算得到结构嵌入模型的参数估计值后, 我们的最后一步是估计特定策略 g 的反事实

274 结局 $E[Y^g]$, 如果不存在其他变量的效应修饰作用, 那么静态策略 \bar{a} 的反事实结局就是:

$$\hat{E}[Y^{\bar{a}}] = \hat{E}[Y^{\bar{0}_K}] + \sum_{k=0}^K a_k \gamma_k(\bar{a}_{k-1}, \tilde{\beta})$$

此外, 如果结构嵌入模型有含 L_k 的项, 或者我们想估计动态策略的反事实结局, 那么我们就需要用知识点 21.6 介绍的算法模拟 L_k 。

275 21.5 删失与时异变量

在本章, 我们所讨论的例子都没有删失, 也即表 21.1 中所有人的结局我们都知道。但在实际中, 我们经常遇到失访或者结局缺失的情况。我们在本书第二部分讨论了删失及其分析方法。在第八章, 我们论述了即使在零假设成立的时候, 删失也可能造成选择偏移。

(删失的讨论参见第 12.6 小节。如果删失 C 是一个对撞变量或者对撞变量的下游变量, 那么控制 C 就会引入选择偏移。)

然而在本书第二部分的讨论中, 我们只是讨论了简单的删失情形, 并没有明确删失出现在“什么时候”。也即, 我们将删失 C 视为非时异变量。但在实际中, 删失是一个时异变量, 比如

C_1, C_2, \dots, C_{K+1} 。如果在时间点 m 没有被删失, 那么 $C_m = 0$, 反之 $C_m = 1$ 。删失是一种单调性的数据缺失, 也即如果有 $C_m = 0$, 那么一定有 $C_1 = C_2 = \dots = C_m = 0$ 。同样, 根据定义, $C_0 = 0$ 。

如果有人在时间点 m 被删失, 也即 $C_m = 1$, 那么在 m 之后, 治疗、混杂、以及结局都是缺失的。因此, 我们的分析只能在未被删失的时间段进行, 也即 $C_m = 0$ 的时候。比如, 如果考虑删失, 21.1 小节的 G-公式就需要改写为:

$$\sum_{\bar{l}} E[Y | \bar{A} = \bar{a}, \bar{C} = \bar{0}, \bar{L} = \bar{l}] \prod_{k=0}^K f(l_k | \bar{a}_{k-1}, c_{k-1} = 0, \bar{l}_{k-1})$$

如果我们把 C_m 也视为治疗, 并且可识别性假设对 C_m 成立, 那么上述 G-公式就相当于联合治疗 $(\bar{a}, \bar{c} = \bar{0})$ 的反事实结局 $E[Y^{\bar{a}, \bar{c} = \bar{0}}]$, 也即所有人的治疗策略都是 \bar{a} 且没有删失的结局。

我们也可以通过逆概率加权估计反事实结局 $E[Y^{\bar{a}, \bar{c} = \bar{0}}]$ 。我们可以在逆概率权重 $W^{\bar{A}} \times W^{\bar{C}}$ 构建的虚拟人群中选择拟合以下模型:

$$E[Y | \bar{A}, \bar{C} = \bar{0}] = \theta_0 + \theta_1 cum(\bar{A})$$

其中权重 $W^{\bar{C}} = \prod_{k=1}^{K+1} \frac{1}{\Pr[C_k = 0 | \bar{A}_{k-1}, C_{k-1} = 0, \bar{L}_k]}$, 其分母可以通过 $\Pr[C_k = 0 | \bar{A}_{k-1}, C_{k-1} = 0, \bar{L}_k]$

的 logistic 模型估计。

(我们使用上角标 $\bar{c} = \bar{0}$ 只是想表明是在未删失的人群中估计 \bar{A} 的因果效应。)

在我们构建的虚拟人群中, 我们会用未被删失者的拷贝代替被删失者, 而这些拷贝和被删失者有一样的治疗史和协变量史, 从而保证我们的虚拟人群大小和原人群一样, 并且变量分布相同。此处的权重——具体而言, 非稳定权重——抵消了删失。

不过上面的权重是非稳定权重, 我们也可以用稳定权重构建虚拟人群。此时:

$SW^{\bar{C}} = \prod_{k=1}^{K+1} \frac{\Pr[C_k = 0 | \bar{A}_{k-1}, C_{k-1} = 0]}{\Pr[C_k = 0 | \bar{A}_{k-1}, C_{k-1} = 0, \bar{L}_k]}$, 分子可以通过 $\Pr[C_k = 0 | \bar{A}_{k-1}, C_{k-1} = 0]$ 的 logistic

模型估计。

用稳定权重构建的虚拟人群, 其样本大小等于删失后的研究人群, 也就是说, 虚拟人群的删失比例等于原人群的删失比例。因而, 稳定权重并不是抵消删失, 而是让删失随机出现, 从而不受到其余变量的影响。也就是说, 使用稳定权重依然存在选择, 但是不存在选择偏移。不过, 无论是使用稳定权重还是非稳定权重, 在虚拟人群中, 都不存在从 L_k 指向 C_m 的箭头 ($m > k$)。

同时, 在可识别性假设下, 即使 \bar{L} 中的部分变量受到过往治疗的影响, 逆概率加权依然能无偏地估计联合治疗 (\bar{A}, \bar{C}) 的因果效应。

(谨记, 如果 $\Pr[C_k = 0 | \bar{A}_{k-1}, C_{k-1} = 0, \bar{L}_k]$ 的模型正确, 那么稳定权重 $SW^{\bar{C}}$ 的均值为 1。)

最后, 当我们用结构嵌入模型估计治疗策略的因果效应时, 我们首先应该用逆概率权重调整删失造成的选择偏移。在实践中, 我们需要先估计逆概率权重, 然后构建虚拟人群, 最后再在虚拟人群中使用结构嵌入模型和 G-估算。

第二十一章精讲点和知识点

精讲点 21.1: 治疗史和协变量史 (原书第 260 页)

估计时异治疗的因果效应时, 我们需要借助治疗史, 以及能用以满足时序互换性的协变量史。G-公式的每一部分都控制了过往的治疗史和协变量史。比如, $k = 2$ 时的混杂变量 L_2 在 G-公式中的对应部分 $f(l_2 | \bar{A}_1 = \bar{a}_1, \bar{L}_1 = \bar{l}_1) = \Pr[L_2 = l_2 | A_0 = a_0, A_1 = a_1, L_0 = l_0, L_1 = l_1]$ 就控制了之前时间点 (0 和 1) 的治疗和混杂。同理, $k = 3$ 时的对应部分也会控制之前时间点 (0、1、2) 的治疗和混杂变量。

然而, “史”这个字不是精确无误的。我们在精讲点 7.2 中讨论过, 混杂变量理论上也可以出现在治疗之后。与此同时, 调整治疗之前的某些变量可能会导致选择偏移 (参见图 7.4, 也称之为 M 偏移)。因此, 在时间点 k , 我们应该将“史”理解为能让 A_k 互换性成立的所有变量的集合。一般而言, 这些混杂变量 \bar{L}_k 都出现在 A_k 之前。所以为了简便, 我们在本书中都统称为“史”。

精讲点 21.2: G-公式的数学表述 (原书第 269 页)

在数学上, G-公式可以表述为不同形式。虽然这些不同形式是等价的, 但它们通常会给出不一样的估计值。在本书, 我们所用的表述方法着重于标准化 (这是一个流行病学术语) 的一般形式。也就是说, 对非时异治疗, G-公式是 $\sum_l E[Y | A = a, L = l] f(l)$; 对时异治疗, G-公式是

$\sum_{\bar{l}} E[Y | \bar{A} = \bar{a}, \bar{L} = \bar{l}] \prod_{k=0}^K f(l_k | \bar{a}_{k-1}, \bar{l}_{k-1})$ 。而在这一表述方法中, 我们需要先估计混杂变量在时

间上的联合密度 $\prod_{k=0}^K f(l_k | \bar{a}_{k-1}, \bar{l}_{k-1})$, 因此我们把这一方法称为 G-公式的联合密度形式。

G-公式的另一表述方式是条件期望。对非时异治疗, G-公式是 $E[E[Y | A=a, L=l]]$, 我们在 13.3 小节介绍过。对时异治疗, 则是递归形式下的迭代期望。这需要我们拟合一系列时序模型。我们在 21.3 小节和知识点 21.3 介绍了这一迭代期望形式, 其也可以和逆概率权重一起用于双重稳健估计。

还有一种表述方式就是逆概率加权。我们在知识点 2.3 中论证了非时异治疗的标准化和逆概率加权是等价的, 这一结论对时异治疗依然成立。G-公式的逆概率加权形式需要我们先估计时间上治疗的条件密度。不过我们把这一方式称为逆概率加权, 而非 G-公式。

精讲点 21.3: 饱和结构嵌入模型的 G-估算 (原书第 271 页)

在时间点 $k=1$, 时序互换性成立也就意味着只要 (A_0, L_1) 相同, $A_1=1$ 的人群和 $A_1=0$ 的人群都会有一样的反事实结局 $Y^{A_0, 0}$ 。因此, 当 $\psi^\dagger = \psi$ 的时候, 这两个人群的 $H_1(\psi^\dagger)$ 也会相等。

在表 21.1 的第一和第二行 $(A_0, L_1)=(0, 0)$ 的人群中, 我们会发现 $A_1=0$ 时 $H_1(\psi)=84$, 而 $A_1=1$ 时 $H_1(\psi)=84-\psi_{11}$, 因此可知 $\psi_{11}=0$ 。同理, 在第三和第四行 $(A_0, L_1)=(0, 1)$ 的人群中, 我们有 $52=52-\psi_{11}-\psi_{12}$, 由 $\psi_{11}=0$ 可得 $\psi_{12}=0$ 。接下来, 在第五和第六行有 $\psi_{13}=0$, 在第七和第八行有 $\psi_{14}=0$ 。

要估计 ψ_0 , 我们需要将 $\psi_{11}、\psi_{12}、\psi_{13}、\psi_{14}$ 的值代入到 $H_0(\psi)$ 的表达式中。在表 21.2 的例子中, 所有参数的值都是 0, 所以 $H_0(\psi)$ 也就等于我们观测到的结局 Y 。接下来我们用结构嵌入模型中的第一个方程计算每一行的 $H_0(\psi)$, 计算方法是用 $\psi_0 A_0$ 减去 $H_1(\psi)$, 如表 21.3 所示。因为时序互换性 $Y^{0,0} \perp\!\!\!\perp A_0$ 成立, 所以在 $k=0$, $A_0=1$ 和 $A_0=0$ 两个人群的 $H_0(\psi)$ 相同。在 $A_0=0$ 人群中, $H_0(\psi)=84 \times 0.25 + 52 \times 0.75 = 60$; 在 $A_0=1$ 人群中, $H_0(\psi)=(76-\psi_0) \times 0.5 + (44-\psi_0) \times 0.5 = 60-\psi_0$ 。因此有 $\psi_0=0$, G-估算结束。

知识点 21.1: 静态策略的 G-公式密度 (原书第 261 页)

治疗策略为 \bar{a} 时, 变量 (Y, \bar{L}) 在 (y, \bar{l}) 处的 G-公式密度为:

$$f(y | \bar{a}, \bar{l}) \prod_{k=0}^K f(l_k | \bar{a}_{k-1}, \bar{l}_{k-1})$$

Y 的密度就是上述联合密度中 Y 的边缘密度:

$$\int f(y | \bar{a}, \bar{l}) \prod_{k=0}^K dF(l_k | \bar{a}_{k-1}, \bar{l}_{k-1})$$

考虑到 L_k 中可能存在连续变量, 我们在这里使用更为广义的积分符号。

观测数据记为 $O = (\bar{A}, \bar{V}, Y)$, 其中 \bar{V} 表除了治疗 \bar{A} 和结局 Y 之外的其他已测数据集合。要使用 G-公式, 我们就需要知道: (1) 治疗策略 \bar{a} ; (2) 表示各变量之间关系的 DAG 图; (3) \bar{V} 当中需要调整的变量 \bar{L} ; (4) \bar{L} , \bar{A} , 以及 Y 之间的正确顺序。 L_k 表示 A_{k-1} 之后、 A_k 之前 L 中所有的变量。虽然我们经常按照时间顺序确定这一顺序, 但这些变量之间的正确顺序不一定就是时间顺序, 不过这一话题已经超出了本书范畴, 参见 Pearl 和 Robins 在 1995 年所著论文。确定顺序后, 如果 $Y^{\bar{a}}$ 的时序互换性以及正数性成立, 那么 Y 的 G-公式密度就等于所有人都遵循治疗策略 \bar{a} 时我们观测到的结局。如果这些假设不成立, 我们依然可以计算 Y 的 G-公式密度, 但是结果不再有因果性意义。

不过注意, (Y, \bar{L}) 在治疗策略为 \bar{a} 时的 G-公式密度和这两个变量的联合分布不同, 其联合分布为:

$$f(y | \bar{A}_k = \bar{a}, \bar{L}_k = \bar{l}) \prod_{k=0}^K f(l_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_{k-1} = \bar{l}_{k-1}) \prod_{k=0}^K f(a_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_k = \bar{l}_k)$$

差别在于 $f(a_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_k = \bar{l}_k)$ 一项被消除了, 其余一样。

知识点 21.2: G-零值悖论 (原书第 265 页)

当我们使用参数 G-公式的时候, 错误的模型会导致有偏的 $E[Y^{\bar{a}}]$ 估计值。假设存在治疗-混杂反馈, 并且极端零假设为真, 也即对所有 \bar{a}' 和 \bar{a} , $Y^{\bar{a}} - Y^{\bar{a}'} = 0$ 的概率为 1。此时, 虽然 $E[Y | \bar{A} = \bar{a}, \bar{L} = \bar{l}]$ 和 $f(l_k | \bar{a}_{k-1}, \bar{l}_{k-1})$ 依赖于治疗策略 \bar{a} , 但是 G-公式得到的 $E[Y^{\bar{a}}]$ 估计值对所有策略都是一样的, 也即为零。假设 $E[Y | \bar{A} = \bar{a}, \bar{L} = \bar{l}]$ 和 $f(l_k | \bar{a}_{k-1}, \bar{l}_{k-1})$ 的模型是基于两个方差相互独立的参数 θ 和 φ (记为 $E[Y | \bar{A} = \bar{a}, \bar{L} = \bar{l}; \theta]$ 和 $f(l_k | \bar{a}_{k-1}, \bar{l}_{k-1}; \varphi)$), 如果此时 L_k 中有离散型变量, 那么这两个模型不可能都是正确设定的, 因为此时 $E[Y^{\bar{a}}]$ 的 G-公式估计值取决于 \bar{a} (参见 Robins 和 Wasserman 在 1997 年所著论文)。因此, 根据 G-公式的结果, 我们可能会错误地拒绝极端零假设, 即使这是一个时序性随机试验。这一现象被称为时异治疗 G-公式的零值悖

论。不过, G-零值悖论并不会妨碍我们在实际中运用 G-公式估计零效应, 因为和随机变异性相比, 这一悖论导致的偏移太小以至于可以忽略不计。

与 G-公式不同的是, 逆概率加权和 G-估算不会受到这一悖论的影响。因为不管我们选择什么样的函数形式, 我们都能保证在因果效应是零值的时候, 模型总是正确的。比如, 在逆概率加权的边缘结构模型中, 如果模型是 $E[Y^{\bar{a}}] = \beta_0 + \beta_1 cum(\bar{a})$, 那么 $\beta_1 = 0$ 的时候, $E[Y^{\bar{a}}]$ 就和 \bar{a} 无关。而在 G-估算的结构嵌入模型中, 因果效应为零也即 $\beta = 0$, 就有 $\gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta) = \beta = 0$ 。

知识点 21.3: 时异治疗的双重稳健估计 (原书第 268 页)

为了估计治疗策略 $\bar{a} = (a_0, a_1, \dots, a_K)$ 的反事实结局 $E[Y^{\bar{a}}]$, Bang 和 Robins 提出了一种递归法。对于二分治疗和连续性结局, 这一方法的步骤如下:

1. 使用所有时间点和所有人员的数据, 拟合一个 logistic 模型 $s(\bar{A}_{m-1}, \bar{L}_m; \alpha)$ 估计 $\Pr[A_m = 1 | \bar{A}_{m-1}, \bar{L}_m]$, 并得到参数 α 的最大似然估计 $\hat{\alpha}$ 。对每个时间点的每个人, 计算常规的时异治疗逆概率权重 $\hat{W}^{\bar{A}_m} = \prod_{k=0}^m \frac{1}{\hat{f}(A_k | \bar{A}_{k-1}, \bar{L}_k; \hat{\alpha})}$ 和策略为 a_m (我们关注的策略) 时的改进权重 $\hat{W}^{\bar{A}_{m-1}, a_m} = \frac{\hat{W}^{\bar{A}_{m-1}}}{\hat{f}(a_m | \bar{A}_{m-1}, \bar{L}_m; \hat{\alpha})}$ 。根据定义, $A_{-1} \equiv 0$ 。

2. 令 $\hat{T}_{K+1} = Y^{\bar{A}_K} = Y$, 然后从 $m = K$ 到 $m = 0$ 依次:

- (a) 拟合线性回归模型 $h(\bar{A}_{m-1}, \bar{L}_m; \theta)$ 估计条件期望 $E[\hat{T}_{m+1} | \bar{A}_m, \bar{L}_m]$, $\hat{W}^{\bar{A}_m}$ 是模型中的一个协变量。

- (b) 在拟合后的模型 $\hat{h}(\bar{A}_{m-1}, \bar{L}_m; \theta)$ 中, 将 $\hat{W}^{\bar{A}_m}$ 替换为 $\hat{W}^{\bar{A}_{m-1}, a_m}$, 并计算估计值 $\hat{T}_m^{\bar{A}_{m-1}, a_m, \dots, a_K} \equiv \hat{T}_m$ 。

3. 在最后 $m = 0$ 时, 有 $\hat{E}[Y^{\bar{a}}] = E[\hat{T}_0]$ 。

只要 $s(\bar{A}_{m-1}, \bar{L}_m; \alpha)$ 或 $h(\bar{A}_{m-1}, \bar{L}_m; \theta)$ 有一个是正确的, 那么 $E[\hat{T}_0]$ 就是 $E[Y^{\bar{a}}]$ 的一致估计。置信区间可以用自举法计算。注意, 如果模型 $h(\bar{A}_{m-1}, \bar{L}_m; \theta)$ 中没有 $\hat{W}^{\bar{A}_m}$, 那么这一方法不再是双重稳健估计, 而是参数 G-公式的另一种替代方法。

知识点 21.4: 边缘结构模型和结构嵌入模型的关系 (第二部分) (原书第 272 页)

现在我们可以将精讲点 14.1 中的论述扩展到时异治疗。当且仅当下列条件成立时, 结构嵌入模型是半参数化的边缘结构模型: 对于所有 $(\bar{a}_{k-1}, \bar{l}_k, \beta)$, 有 $\gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta) = \gamma_k(\bar{a}_{k-1}, \beta)$ 。具体而言, 结构嵌入模型是一个有以下函数形式的半参数边缘结构模型:

$E[Y^{\bar{a}}] = E[Y^{\bar{0}_k}] + \sum_{k=0}^K \alpha_k \gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta)$, 其中 $E[Y^{\bar{0}_k}]$ 是未定的。不过, 因为这个结构嵌入模型假设了过往变量不存在效应修饰作用, 而边缘结构模型是不需要这一假设的, 所以这个结构嵌入模型不仅仅是一个边缘结构模型。

如果我们指明了一个结构嵌入模型 $\gamma_k(\bar{a}_{k-1}, \beta)$, 那我们就可以用 G-估算或逆概率加权得到 β 的估计值。此时, 如果模型设定正确, G-估算应该在统计计算上更高效。

不过, 如果边缘结构模型是正确的, 而结构嵌入模型是错误的 (因为 $\gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta) \neq \gamma_k(\bar{a}_{k-1}, \beta)$), 那么 G-估算的结果就会是诱骗的, 而逆概率加权的结果会是正确的。于是我们就会遇到经典的偏差方差权衡——此时 G-估算能减小方差, 但会引入偏差。

知识点 21.5: 线性结构嵌入模型估计值的闭合形式 (原书第 273 页)

在我们的所有例子中, $\gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta) = \beta^T R_k$ 对于 β 而言都是线性的 (这里 $R_k = r_k(\bar{L}_k, \bar{A}_{k-1})$ 表示一个向量, 其中每一个元素都是已知函数)。于是, 在模型 $\text{logit } \Pr[A_k = 1 | \bar{L}_k, \bar{A}_{k-1}] = \alpha^T W_k$ 当中, 估计值 $\hat{\beta}$ 有一个闭合表达式:

$$\hat{\beta} = \left\{ \sum_{i=1, k=0}^{i=N, k=K} A_{i,k} X_{i,k}(\hat{\alpha}) Q_{i,k} S_{i,k}^T \right\}^{-1} \left\{ \sum_{i=1, k=0}^{i=N, k=K} Y_i X_{i,k}(\hat{\alpha}) Q_{i,k} \right\}$$

其中 $X_{i,k}(\hat{\alpha}) = [A_{i,k} - \text{expit}(\hat{\alpha}^T W_{i,k})]$, $S_{i,k} = \sum_{i=1, j=k}^{i=N, j=K} R_{i,j}$, 而 $Q_{i,k} = q_k(\bar{L}_{i,k}, \bar{A}_{i,k-1})$ 是我们选取的函

数形式, 只影响误差, 不影响一致性。Robins 在 1994 年论文中讨论 Q_k 的最佳形式。

实际上, 如果 $\gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta) = \beta^T R_k$ 对于 β 而言是线性的, 我们也可以推导出 β 的双重稳健估计闭合表达式 $\tilde{\beta}$ 。首先, 我们需要一个工作模型 $\zeta^T D_k = \zeta^T d_k(\bar{L}_k, \bar{A}_{k-1})$ 对 $E[H_k(\beta) | \bar{L}_k, \bar{A}_{k-1}] = E[Y^{\bar{A}_{k-1}, \bar{0}_k} | \bar{L}_k, \bar{A}_{k-1}]$ 建模, 并且定义

$$\begin{pmatrix} \tilde{\beta} \\ \tilde{\zeta} \end{pmatrix} = \left\{ \sum_{i=1, k=0}^{i=N, k=K} \begin{pmatrix} A_{i,k} X_{i,k} (\hat{\alpha}) Q_{i,k} \\ D_{i,k} \end{pmatrix} \left(S_{i,k}^T, D_{i,k}^T \right) \right\}^{-1} \left\{ \sum_{i=1, k=0}^{i=N, k=K} Y_i \begin{pmatrix} X_{i,k} (\hat{\alpha}) Q_{i,k} \\ D_{i,k} \end{pmatrix} \right\}。如果模型 \zeta^T D_k 是正确的，$$

的, 或者模型 $\text{logit Pr}[A_k = 1 | \bar{L}_k, \bar{A}_{k-1}] = \alpha^T W_k$ 是正确的, 那么 $\tilde{\beta}$ 就会是 ψ 的一致估计。

知识点 21.6: G-估算之后 $E[Y^g]$ 的估计值 (原书第 274 页)

假设可识别性假设成立, 我们可以得到结构嵌入模型 $\gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta)$ 的双重稳健估计值 $\tilde{\beta}$ 。同时, 我们也可能需要估计某治疗策略 g 的反事实结局 $E[Y^g]$ 。我们可以用以下 Monte Carlo 算法进行估计:

1. 利用 $H_0(\tilde{\beta})$ 估计所有人都未接受治疗时的反事实结局 $E[Y^{\bar{0}_k}]$, 记为 $\hat{E}[Y^{\bar{0}_k}]$ 。
2. 利用所有时间点、所有人的数据拟合 $f(l_k | \bar{a}_{k-1}, \bar{l}_{k-1})$ 的模型, 并从模型中计算估计值 $\hat{f}(l_k | \bar{a}_{k-1}, \bar{l}_{k-1})$ 。
3. 从 $v=1, 2, \dots, V$, 进行以下步骤:
 - (a) 从 $\hat{f}(l_0)$ 中计算 $l_{v,0}$ 。
 - (b) 在 $k=1, 2, \dots, K$ 时, 令 $\bar{a}_{v,k-1} = \bar{g}_{k-1}(\bar{l}_{v,k-1})$, 从 $\hat{f}(l_k | \bar{a}_{v,k-1}, \bar{l}_{v,k-1})$ 中计算 $l_{v,k}$ 。
 - (c) $Y^g - Y^{\bar{0}_k}$ 的第 v 个 Monte Carlo 估计值是 $\hat{\Delta}_{g,v} = \sum_{j=0}^{j=K} a_{v,j} \gamma_j(\bar{a}_{v,k-1}, \bar{l}_{v,j}, \tilde{\beta})$ 。
4. $E[Y^g]$ 的估计值是 $\hat{E}[Y^g] = \hat{E}[Y^{\bar{0}_k}] + \sum_{v=1}^{v=V} \hat{\Delta}_{g,v} / V$ 。

如果 $f(l_k | \bar{a}_{k-1}, \bar{l}_{k-1})$ 的模型和结构嵌入模型 $\gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta)$, 以及 $\text{Pr}[A_k = 1 | \bar{L}_k, \bar{A}_{k-1}]$ 或 $E[Y^{\bar{A}_{k-1}, 0_k} | \bar{L}_k, \bar{A}_{k-1}]$ 两个模型中的一个是正确的, 那么估计值 $\hat{E}[Y^g]$ 就是一致的。置信区间可以用自举法计算。

注意, 如果 $\beta = 0$, 那 $\gamma_k(\bar{a}_{k-1}, \bar{l}_k, \beta)$ 会收敛到 0。因此 $\hat{\Delta}_{g,v}$ 会收敛到 0, 而 $\hat{E}[Y^g]$ 会收敛到 $\hat{E}[Y^{\bar{0}_k}]$, 即使 $f(l_k | \bar{a}_{k-1}, \bar{l}_{k-1})$ 的模型是不正确的。换句话说, 在可识别性假设成立的情况下, 我

们如果知道 $\Pr[A_k = 1 | \bar{L}_k, \bar{A}_{k-1}]$ (比如在时序性随机试验中), 抑或 $\Pr[A_k = 1 | \bar{L}_k, \bar{A}_{k-1}]$ 或

$E[Y^{\bar{A}_{k-1}, 0_k} | \bar{L}_k, \bar{A}_{k-1}]$ 的模型有一个是正确的, 那么结构嵌入模型就能正确给出零值估计。

第二十一章图表

Table 21.1

N	A ₀	L ₁	A ₁	Mean Y
2400	0	0	0	84
1600	0	0	1	84
2400	0	1	0	52
9600	0	1	1	52
4800	1	0	0	76
3200	1	0	1	76
1600	1	1	0	44
6400	1	1	1	44

Table 21.2

A ₀	L ₁	A ₁	Mean H ₁ (ψ)
0	0	0	84
0	0	1	$84 - \psi_{1,1}$
0	1	0	52
0	1	1	$52 - \psi_{11} - \psi_{12}$
1	0	0	76
1	0	1	$76 - \psi_{11} - \psi_{13}$
1	1	0	44
1	1	1	$44 - \psi_{11} - \psi_{12}$ $- \psi_{13} - \psi_{14}$

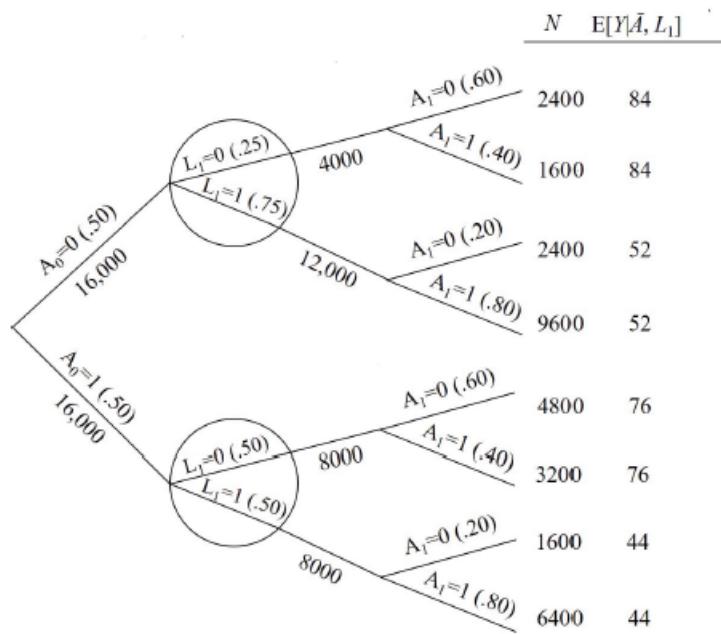


Figure 21.1

Causal Inferences: What if ——第二十一章

作者: Miguel A. Hernan, James M. Robins;

翻译: 罗家俊

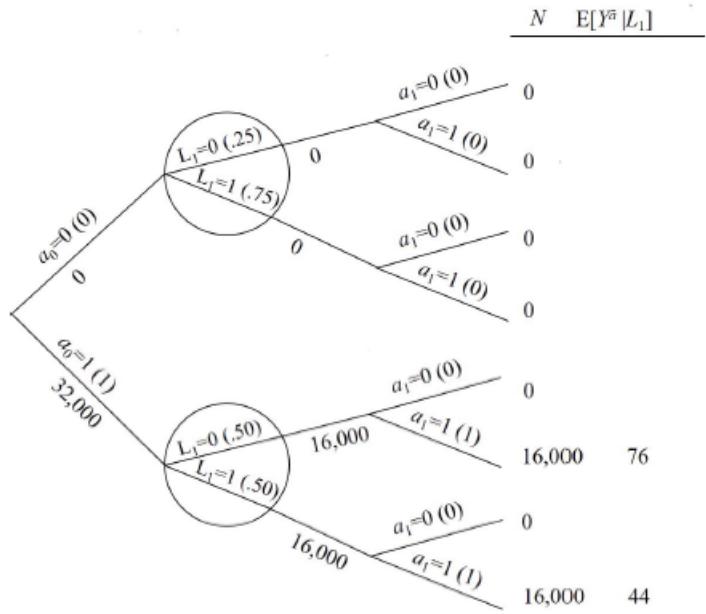


Figure 21.2

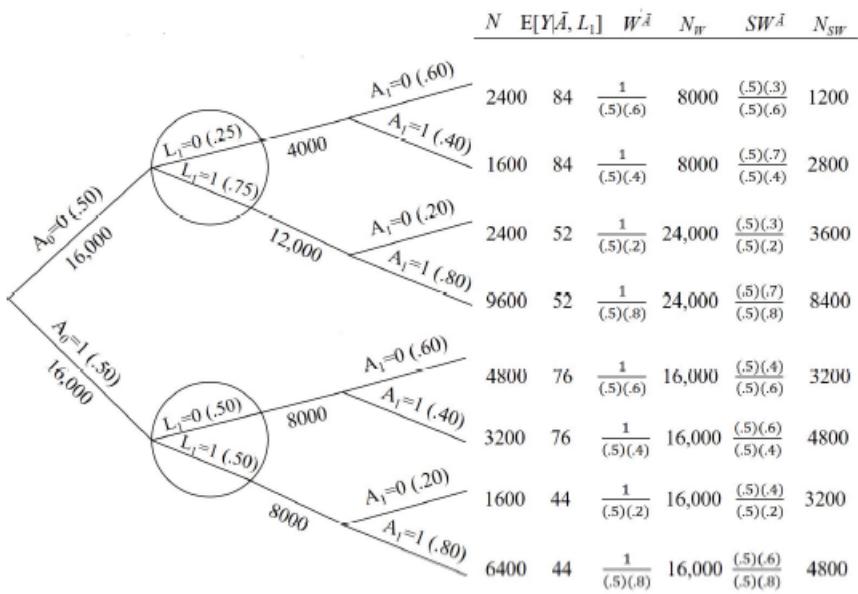


Figure 21.3

第二十二章 靶标试验

277 在本书第一部分,为了得到有效的因果推断,我们尽可能地将观察性研究视为一个假想的随机试验。这一假想的随机试验被称为靶标试验。在前几章我们讨论了时异治疗的分析方法,而本章会将靶标试验概念推广到所有治疗策略中,期望以此建立一个统一的因果分析框架。如此一来,不管是随机试验还是观察性研究,我们都能在这一框架下进行分析。

本章还会介绍一些随机试验的相关概念,比如治疗意向效应、依方案消息等。要正确估计这些因果效应,我们需要有时异治疗以及时异预后因素的相关数据。G-方法的发展使得我们现在的讨论不再是纸上的练习,而是一件可能完成的事情:只要有数据,我们就能有效地估计治疗的因果效应。

22.1 再论靶标试验

一项随机试验要估计HIV阳性患者中抗逆转录病毒疗法对5年生存率的因果效应,被试都是18岁以上,没有艾滋病,未曾接受过抗逆转录病毒疗法。在研究初始($k=0$),所有被试被随机分到治疗策略 g 和 g' 两个小组。随访从分组的时候开始,直到去世、失访、或者满60个月。当然,和其他试验一样,不是所有被试都会遵循给他们分配的治疗策略。也就是说,部分被试偏离了原治疗方案。

在这项随机试验中,研究人员和被试都知道实际的分组情况(也即没有单盲或双盲设计),没有安慰剂组(治疗策略 g 和 g' 都涉及接受治疗),同时每名被试都和正常病人一样被随访观察。因而,从这一试验中得到的因果效应估计值可以视为大多数现实情形中抗逆转录病毒疗法的因果效应。

如果因种种原因我们不能在实际中开展这项随机试验,那么我们也可以用观察性研究模拟这项随机试验。我们将这一假想的随机试验称为观察性研究的靶标试验。

(靶标试验的精确定义,参见Hernan和Robins在2016年所著论文。)

详细说明靶标试验的研究方案有助于阐明我们的研究问题。至少,我们需要说明研究方案中的几个关键部分:适用标准,随访的起始与结束时间,治疗策略,结局,以及数据分析方法。注意,要详细说明一项假想靶标试验的研究方案,我们需要先充分了解已有的观测数据。比如,只有确定我们的观测数据中有HIV诊断数据,我们才能有理由说这项靶标试验是在HIV阳性患者中进行的。

278 我们在第三章介绍过靶标试验,不过本书第一和第二部分只讨论了简单的非时异情形。现在我们将讨论更符合实际情况的靶标试验,其中的治疗都是时异性的,比如“一直接受治疗,直到

产生毒副作用”策略 g_1 , 或者“尽可能不接受治疗”策略 g_0 。下一小节我们将定义随机试验中时异治疗不同类型的因果效应。同时, 知识点 22.1 论述了在随机试验中如何比较不同的治疗策略, 这一对比也被称为直接效应。

(简写 PICO 被用来表示靶标试验的几个重要组成部分, 分别是人群 (population)、干涉 (intervention)、比较 (comparator)、以及结局 (outcome)。)

22.2 随机试验中的因果效应

我们先回顾一下随机试验涉及的三种不同类型因果效应。首先, 我们先定义数学符号。 A_k 表示在时间点 k 的治疗状态, 接受治疗则取值为 1, 反之为 0。 C_k 表示在时间点 k 是否被删失, 未被删失则取值为 0, 反之为 1。每名被试被随机分到“接受治疗 $A_k = 1$, 直到出现毒副作用”策略 g_1 , 或者“整个研究阶段不接受治疗, 即 $A_k = 0$ ”策略 g_0 。变量 Z 表示分组情况, 如果被分到 g_1 则取值为 1, 被分到 g_0 则取值为 0。

在第三部分的前几章, 我们于随访结束之后测量结局变量 Y , 并且我们只关心治疗策略对 Y 的因果效应。现在我们需要讨论治疗策略对失效时间的因果效应。也就是说, 我们的目标是估计 279 治疗策略对生存情况 (参见知识点 22.3) 的因果效应。我们用 D_k 表示在第 k 个月的死亡情况。1 表示死亡, 反之为 0。

首先, 我们先思考一下治疗策略的分组情况 (而非实际治疗策略) 对生存状态的因果效应。这一效应通常被称为治疗意向效应, 并被定义如下两种干涉方式的比较: (1) 在研究初始被分到策略 g_1 , 并一直参与随访直到研究结束; (2) 在研究初始被分到策略 g_0 , 并一直参与随访直到研究结束。在时间点 k 治疗意向的因果效应可以表述为: $\Pr[D_k^{z=1, \bar{c}_k=\bar{0}} = 1] - \Pr[D_k^{z=0, \bar{c}_k=\bar{0}} = 1]$, 其中 $\bar{c}_k = \bar{0}$ 表示没有人失访, 所有人都参与研究直到结束。

(第九章简要介绍了治疗意向效应和依方案效应。)

在某些随机试验中, 分组和开始执行治疗策略是同时发生的。也即被分到策略 g_1 的被试在 $k=0$ 就接受了治疗, 而被分到策略 g_0 的被试则不会在 $k=0$ 接受治疗。而这些被试之后是否接受治疗则与分组无关。在这种情况下, 治疗意向效应就不仅是分组的因果效应, 同时还是初始治疗的因果效应, 比如 $\Pr[D_k^{a_0=1, \bar{c}_k=\bar{0}} = 1] - \Pr[D_k^{a_0=0, \bar{c}_k=\bar{0}} = 1]$ 。

治疗意向效应对初始之后是否接受治疗持不关心态度, 也即研究中是否继续治疗、是否用了其他治疗方式等行为都不会影响治疗意向的定义。这种不关心的态度意味着治疗意向效应的大小取决于偏离原策略的程度。因此, 有同样研究方案的两个随机试验可能会给出不同的治疗意向效应, 但这两个效应可能都是无偏的, 因为效应的不同只是反映了这两个试验中不同的偏离程度。

其次, 让我们思考一下如果实际执行的策略就是分组策略, 我们将观测到的因果效应。这一效应被称为依方案效应, 并被定义如下两种干涉方式的比较: (1) 在研究初始被分到策略 g_1 , 并一直遵循策略 g_1 , 直到研究结束; (2) 在研究初始被分到策略 g_0 , 并一直遵循策略 g_0 , 直到研究结束。在时间点 k 依方案效应可以表述为: $\Pr[D_k^{g_1, \bar{c}_k=0} = 1] - \Pr[D_k^{g_0, \bar{c}_k=0} = 1]$, 其中 $\bar{c}_k = 0$ 表示没有人失访, 所有人都参与研究直到结束。

一份合理的研究方案会说明什么时候应该停止治疗。比如, 我们的治疗策略 g_1 就会说明, 如果出现毒副作用, 治疗将会停止。因而, 依方案效应可能是动态策略之间的比较。我们在精讲点 19.2 中提到过这一问题。

有时研究方案未能清晰描述治疗策略的动态特点。比如, 研究方案 g_1 可能只是说“在各时间点都接受治疗, 即 $A_k = 1$ ”, 但未说明在产生毒副作用时应停止治疗。这样的简化描述可能会让人产生误解。具体而言, 一名被分配到 g_1 的被试, 如果是因为毒副作用没有继续接受治疗, 那么其应该被认为是遵循了策略 g_1 , 而非相反。在现实研究中, 如果被试因种种不可抗原因不能继续接受治疗, 那么这些被试不应该被算作偏离了研究方案。因为依方案效应是两种现实策略的对比, 所以这一效应和我们现实世界的因果推断最为相关。

(理想情况下, 研究方案应该尽可能给出治疗策略的详尽说明, 从而避免各种模糊不清。这样一来, 依方案效应就是良定的。)

研究者在潜意识中更在意依方案效应。比如, 研究者经常会讨论现实中治疗的执行情况和研究方案中治疗的描述两者之间的差异, 进而会说存在“偏移”。这一用语表明研究者更在意研究方案在现实中的执行情况(也即依方案效应), 而非初始阶段的分组情况(也即治疗意向效应)。不过我们已经知道, 初始分组之后偏离研究方案其实不会对分组的因果效应造成偏移。

最后, 让我们思考一下如果被试接受了不在研究方案中的治疗会发生什么。假设“在各时间点都不接受治疗, 即 $A_k = 0$ ”策略 g_0 实际上劣于策略 g_1 。而医生们会在被试身体情况恶化时(比如 CD4 数目 L_k 小于 200 个/ μL) 推荐治疗。因而许多被分到 g_0 的被试实际上接受的是策略 g'_0 , 也即“不接受治疗, 直到 $L_k < 200$, 之后接受治疗”。此时我们比较的是以下两种干涉的结

局: (1) 在研究阶段执行策略 g_1 , 直到研究结束; (2) 在研究阶段执行策略 g'_0 , 直到研究结束。这一效应既不是治疗意向效应, 也不是最初设想的依方案效应, 而是被试被随机分配到 g'_0 或 g_1 的依方案效应。

这一例子说明了我们估计的因果效应, 可能并不是最初设想的依方案效应, 而是另一个假想的靶标试验的依方案效应。有趣的是, 如果此时我们关注的治疗策略和原方案不同, 那么被试遵循原方案就变成了一件不利的事。也即, 完美遵循原方案也就意味着我们不能用其他靶标试验思考我们的研究。比如, 在一个抗逆转录病毒疗法随机试验中, 如果所有人都严格根据 CD4 数目决定接受治疗与否, 那就没必要把这个试验想象成被试被随机分配到治疗或不治疗策略的靶标试验了。事实上, 正是因为现实世界中的随机试验会出现被试不配合, 所以我们才有机会将一个现实中的随机试验想象成另一个靶标试验, 从而得出更符合实际的结论。

22.3 观察性研究中的因果效应

如果一项观察性研究能被类比成随机试验, 那么我们也可以在这项观察性研究中估计因果效应, 就像我们在随机试验中做的一样。

治疗意向效应在观察性研究中可以类比为以下两种干涉的对比: (1) 在研究初始治疗 $A_0 = 1$, 并在整个研究阶段一直保持; (2) 在研究初始治疗 $A_0 = 0$, 并在整个研究阶段一直保持。
282 因而在观察性研究中, 治疗意向效应可以表述为: $\Pr[D_k^{a_0=1, \bar{c}_k=\bar{0}} = 1] - \Pr[D_k^{a_0=0, \bar{c}_k=\bar{0}} = 1]$ 。可以理解为在一个靶标试验中, 分组和执行治疗策略同时出现。

观察性研究类比的治疗意向效应和实际随机试验的治疗意向效应有些许差别, 因为在随机试验中, 被分配到某一策略的被试并不一定会执行该策略; 而在观察性研究中, 治疗意向效应被类比为在研究初始执行两种不同策略的人群的对比。不过就算这样, 观察性研究类比的治疗意向效应依然保留了一个重要特征: 我们只根据研究初始的情形定义因果效应。

观察性研究类比的依方案效应和靶标试验的依方案效应定义一样。在随机试验中, 我们会区分最初设想的依方案效应以及实际靶标试验的依方案效应。而在观察性研究中我们没有必要进行区分, 因为观察性研究并没有事先预设好的研究方案, 所以我们可以将每一种不同的依方案效应对应成一个特定的靶标试验。总而言之, 如果我们想将观察性研究类比为一个靶标试验, 那么这个靶标试验中的干涉必须存在于观察性研究中。然而, 在一些情况下, 研究者可能会用模型去外推不存在于现实数据中的情况。
283

根据靶标试验定义观察性研究的因果效应还有一个优点, 也即我们必须详细说明治疗策略是什么。一旦我们接纳了这一看法, 我们就会发现, 某些效应估计是不能类比成两种假想干涉的比较的, 而我们应该尽量避免这种情形出现。之前 3.5 和 3.6 小节讨论了这一问题。

(比如, 在绝经期激素治疗对心脏病影响的研究中, 大量观察性研究比较的是“现在接受激素治疗”和“从前接受激素治疗”。然而很少有随机试验会用这两者的比较定义因果效应, 因为治疗状态会随时间变化, 同时没有相对应的干涉方式, 还可能存在选择偏移的问题。)

在观察性研究中详细定义因果效应还能避免模糊性。我们在精讲点 9.4 中讨论过, 一种观点认为治疗意向效应衡量的是治疗的效力(在现实情形下能观测到的治疗的效应), 而依方案效应衡量的是治疗的效益(在理想情况下能观测到的效应)。如果治疗方式是长时间的治疗策略, 那么这一观点就存在显而易见的问题, 因为此时很难论证现实中的依方案效应衡量的就是效益, 或者在治疗的收益(或危害)没有明确的时候, 治疗意向效应衡量的就是治疗的效力。因而, 在长时间的治疗策略中, “效力”与“效益”这两个概念就暧昧不清。而清晰明确地定义我们的因果效应是什么将会提让我们的研究问题更明晰, 也更有利决策者进行参考。

22.4 初始时间

模拟靶标试验的一个重要部分就是确定研究开始的时间, 也称初始时间, 或者时间零点。研究参与者需要在这之前而非之后符合入选标准, 而研究结局要在这之后而非之前测量。

在随机试验中, 时间零点是入选的被试被分配到某一个治疗组的时间点。比如, 在我们抗逆转录病毒疗法的随机试验中, 初始时间——时间零点——被定义为被试分组的时间点, 治疗策略可能在这一时间点开始执行, 或者在这一时间点之后不久执行。我们不可能在随机分组的两年前或两年之后才进行研究。在分组之前就开始研究没有什么意义。而在随机分组之后许久才开始研究可能会引入偏移, 比如有的被试会在这期间死亡, 从而导致选择偏移。

284 观察性研究也需要考虑初始时间, 其理由与随机试验相同。一般而言, 观察性研究的随访开始时间就是靶标试验的初始时间, 否则就难以给效应估计赋予因果意义, 且可能存在选择偏移。然而, 如何模拟靶标试验的初始时间依然是一个难题。在以下两种情形中, 参与人员可能在不同的时间点满足我们的研究入选标准, 而这会影响对初始时间的判定:

(精讲点 22.1 讲述了在研究开始之后一段宽限期内执行治疗策略的具体操作。)

1. 入选标准可能在一个时间点就被满足。这是最简单的情形, 而我们的初始时间从入选标准被满足时算起。比如, 在我们抗逆转录病毒疗法的研究中, 入选标准定为 CD4 数目在 500 以下。那只要参与人员的 CD4 数目第一次降到 500 以下, 随访就算开始了。

2. 入选标准可能在多个时间点被满足, 而这种情形常常会造成许多混淆。比如, 我们想在绝经期妇女中比较使用激素疗法和不使用激素疗法的差别。入选标准是没有慢性病史且在过去两年没有用过激素治疗。如果一名妇女在 51 到 65 岁之间都满足这个要求, 那我们的随访应该从什么时候算起? 在靶标试验中, 这名妇女可以在多个时间点加入试验, 也即她有多个满足入选标准的时间点。

如果存在多个满足入选标准的时间点, 我们有几种不同的方法判定初始时间。我们可以:

(1) 使用第一次满足入选标准的时间点; (2) 从满足要求的时间点中随机选择一个; (3) 使用所有时间点。此外我们还可以有其他方法判定初始时间。上述最后一种方法需要我们模拟多个靶标试验, 具体有多少个则取决于以下两个因素:

1. 如果收集数据的方案和时间点是事先定好的 (比如每两年收集一次), 那就在每个预定的时间点模拟一次靶标试验。

2. 如果收集数据的时间因人而异 (比如使用参与者的电子病历), 那就选一个固定的时间单位 (比如日、星期、月), 然后每隔一个时间单位模拟一次靶标试验。

从统计的角度来说, 第 (3) 种做法是最有效率的, 因为这种做法用到了更多的数据。不过, 因为这种方法会将每名参与者置于多个靶标试验中, 所以我们也需要调整相应的误差。有很多方法可以解决这一问题, 比如自举法。

22.5 因果分析的统一框架

285

将随机试验和观察性研究的因果推断统一起来, 需要我们用一套通用的术语描述这两种不同的研究, 而靶标试验就可以提供这一套通用术语。除了初始的随机分组之外, 用以模拟靶标试验的观察性研究和真正的随机试验在后续分析上没有什么不同。也就是说, 随机试验可以被视为有初始随机分组的跟踪随访研究, 而观察性研究可以被视为没有初始随机分组的跟踪随访研究。

实际上, 随机试验和观察性研究的区别只有三点。在随机试验中, (1) 因为有随机分组, 所以不存在初始混杂; (2) 随机分组的概率是已知的; (3) 每名被试被分到哪一组是已知的。一项观察性研究, 如果测量并调整了足够多的变量, 那就可以用来模拟第一点; 如果治疗分组的模型是正确的, 那就可以用来模拟第二点; 而如果我们想估计的是依方案效应, 那么我们可以不用理会第三点, 这是因为依方案效应的估计并不涉及这一点。

(Robins 在 1986 年的论文里证明了只要混杂变量都被测量且调整, 那么我们完全可以忽略随机分组。)

有初始分组和没有初始分组之间的相似点被越来越多的研究者注意到，并被用来估计现实中一段时间内治疗策略的因果效应。这些研究是对早期临床研究范式的一个挑战。因为在早期范式 286 中，受到严格控制的随机试验被认为是最佳的研究方案。而一个时间跨度较长的随机试验非常可能出现偏离研究方案的情形，也就会产生混杂与选择偏移，这与观察性研究类似。尤其是当我们需要估计依方案效应的时候，随机试验和观察性研究都需要调整会影响到失访和接受治疗情况的变量。

(观察性研究的时异混杂和随机试验的不配合情形在因果图中有一样的结构。)

这样看来，观察性研究和随机试验的唯一不同点在于观察性研究有初始混杂，其余方面非常相似，所以有研究者认为这两种不同的研究方式应该有相似的分析方法。然而在实际中，随机试验和观察性研究的分析方法大相径庭，这造成了许多困扰，并且是站不住脚的。

首先我们需要辨析第一个问题：在随机试验中，是否“治疗意向分析”和“依方案分析”就能有效地估计治疗意向效应和依方案效应？答案是否定的。典型的“治疗意向分析”不会调整任何混杂或选择偏移，只会比较结局变量在不同治疗分组中的分布情况。不调整初始混杂是可以理解的，因为随机分组保证了没有初始混杂。不调整分组后的混杂也是可以理解的，因为分组后的混杂不可能影响初始分组情况。

不过，治疗意向分析假设了所有被试都参与到了研究当中（无论他们是否遵循研究方案），并且直到研究结束。因而治疗意向分析会受到分组后的选择偏移影响，比如失访。因此，正确的治疗意向分析应该调整一些分组后的变量，从而消除选择偏移。当其中某些时异变量受到过往治疗的影响，我们应该用 G-方法调整这些变量。

(因为初始分组并不能保证失访和未失访人群的互换性，所以治疗意向分析需要调整相关变量，从而保证互换性。)

除了“治疗意向分析”，许多随机试验还会用到所谓的“依方案分析”。典型的依方案分析只包括遵循治疗分组的被试。如果治疗只有一个单一时间点，那就简单比较不同治疗分组的被试即可。如果是时异治疗，那么被试在第一次脱离治疗方案时被删失。因而每个时间点剩下的依方案人群就是我们的分析人群。

287 不过，这样的“依方案分析”有两个方面的问题。首先，同“治疗意向分析”一样，依方案分析也应该考虑随机分组后会影响选择偏移的变量。其次，依方案分析只包括遵循治疗分组的被试，也就抛弃了随机分组带来的好处。因此，依方案分析更像是观察性研究中的分析方法。同许多观察性分析一样，依方案分析需要考虑影响治疗情况的混杂变量，因而需要用 G-方法调整混

杂。我们也可以用工具变量估计依方案效应，从而不需调整任何混杂（第十六章），然而这一方法存在局限性，并且依赖于大量模型假设。我们在第十六章已经讨论过这些问题。

（我们将不调整任何混杂的依方案分析称为简单依方案分析。）

随机试验需要调整混杂和选择偏移，而观察性研究也需要调整混杂和偏移。因而，我们都可以用 G-方法调整这些混杂和偏移，从而用观察性研究模拟靶标试验，估计靶标试验的治疗意向效应和依方案效应。

总而言之，随机试验和观察性研究的分析方法是相似的。如果我们认为观察性研究需要调整混杂和选择偏移，那么随机试验也需要调整初始分组后的混杂和选择偏移。这两种研究设计的唯一区别只在于初始的随机分组。而有了靶标试验概念和 G-方法，我们就能在一个统一的因果推断框架下分析这两种不同的研究设计。

在历史上，随机试验一直被认为能提供更可靠的结论，因而优于观察性研究。不过因为种种原因，不是所有研究问题都可以进行随机试验，因而研究者只能求助于观察性研究。而观察性研究也需要深思熟虑的设计和分析。把观察性研究比拟为靶标试验，并用随机试验的方式进行评判，使我们能够得到可靠的结论，为决策提供重要信息。

第二十二章精讲点和知识点

精讲点 22.1：宽限期（原书第 285 页）

让我们再思考一下正文中提到随机试验。在这个试验中，如果被试的 CD4 数目小于 500 就接受治疗，不过我们想比较立即执行治疗策略和稍过一段时间再执行治疗策略两者之间的差异。在现实生活中，因为医疗系统的差异，被试可能要几周甚至几个月之后才能正式接受治疗。因此研究者通常会给出一段宽限期（比如 3 个月），只要在这个期间开始执行治疗策略都算作立即执行治疗策略。

不过使用宽限期会导致被试的观测数据对应的治疗策略不止一种。举例来说，如果我们有 3 个月的宽限期，只要被试在 CD4 数目降至 500 以下的 3 个月内执行治疗策略，那么就算他立即执行了治疗策略。如果这名被试是在第三个月才执行策略，那么根据我们的定义，他的数据应该和在第一个月或第二个月执行策略的其他被试一致。假如这名被试在前两个月去世了，那么我们应该把他归入到哪一组，立即执行策略，还是稍后再执行策略？很可能我们会把他随机分配到任意一组。

另一种可能是把这名被试克隆成两份, 一组一份。只要这名被试的数据和治疗策略不再一致, 那就在不一致开始的那一时间点删失这名被试。举例来说, 如果这名被试是在第三个月才执行策略, 那么被分到“3个月后执行策略”的那份克隆就在第三个月被删失。不过, 这种删失可能会导致选择偏移, 因此我们需要用逆概率加权调整选择偏移。更重要的是, 如果这名被试在第二个月去世, 那么他在两个分组中的克隆也算作死亡。通过双重分组, 研究者可以避免把宽限期都算作某一分组所带来的偏移。

当我们只用宽限期以及上述提到的克隆与删失的时候, 我们不能估计治疗意向效应, 因为几乎所有人都会在每一分组有一个克隆。因为每名被试都是在研究初始被分组的, 所以只基于初始分组的比较此时比较的是两组一模一样的人群。因此, 在研究初始引入宽限期只是为了估计某种形式的依方案效应。

知识点 22.1: 受控直接效应 (原书第 278 页)

假设治疗 A 对结局 Y 的效应之间有一个中介变量 B , 那我们将 A 对 Y 因果效应中不通过 B 的那部分称为直接效应。如果 B 有两个取值 (0 或 1), 那么 $B=1$ 的时候, 我们有一个直接效应, $B=0$ 的时候, 我们又有一个直接效应。在加法尺度上, 这两个直接效应可以定义为 $E[Y^{a=1,b=1}] - E[Y^{a=0,b=1}]$ 以及 $E[Y^{a=1,b=0}] - E[Y^{a=0,b=0}]$ 。这两个直接效应也被称为受控直接效应, 我们也可以在靶标试验中估计这两个效应。在靶标试验中, 我们只需依次对治疗 A 和中介变量 B 随机分组即可。(知识点 22.2 讨论了不能用靶标试验模拟的受控直接效应。)

假设我们进行了一项随机试验, 其中被试在研究初始被随机分配到 $A=1$ 和 $A=0$ 两个组中, 并在一个月后被随机分到 $B=1$ 或 $B=0$ 。因而我们总共就有四个分组。3 个月之后测量每名被试的结局 Y (假设没有人失访)。因为随机组保证了互换性, 所以我们就可以用这项随机试验估计受控直接效应。

在观察性研究中, 如果 A 和 B 的可识别性假设都成立, 那么我们也可以估计受控直接效应。因为 A 和 B 可以看作时异治疗的组成部分, 所以此时的可识别性假设同第十九章中的一样。不过, 治疗 A 和中介变量 B 自己本身也可以是时异变量。

知识点 22.2: 自然直接效应和分层主要直接效应 (原书第 279 页)

中介变量 B 存在的时候, 除去知识点 22.1 讨论的受控直接效应, 还存在不同定义的直接效应。

自然直接效应指的是 B 等于 $A=0$ 时 B 应有的取值时 (记为 $B^{a=0}$) , A 对 Y 不经过 B 的直接效应, 也即 $E[Y^{a=1, B^{a=0}}] - E[Y^{a=0, B^{a=0}}]$, 其中 $E[Y^{a=1, B^{a=0}}]$ 是一个不同反事实世界下的交叉数值, 因为同时涉及 $a=1$ 和 $a=0$ 两个反事实世界。因此, 我们不能用随机试验定义自然直接效应, 而在观察性研究中估计的自然直接效应也不可验证。即使如此, 自然直接效应是许多中介分析的目标之一。这是因为在某些假设下, 治疗的总效应可以分解为自然直接效应和自然非直接效应。自然直接效应由 Robins 和 Greenland 在 1992 年的论文中引入, 并被称为纯粹直接效应。Pearl 在 2001 年的论文中将其重命名为自然直接效应。更多详情, 参考 VanderWeele 在 2015 年的著作。

主层直接效应是当 B 的取值为 b 时, A 对 Y 的直接因果效应。此时 B 的取值不受 A 的影响, 也即 $B^{a=0} = B^{a=1} = b$ 。主层直接效应可以表述为

$$E[Y^{a=1, b} | B^{a=0} = B^{a=1} = b] - E[Y^{a=0, b} | B^{a=0} = B^{a=1} = b], \text{ 也可以记为}$$

$E[Y^{a=1} | B^{a=0} = B^{a=1} = b] - E[Y^{a=0} | B^{a=0} = B^{a=1} = b]$ 。注意, 和其他类型直接效应不同的是, 主层直接效应并不涉及联合反事实结局 $Y^{a,b}$, 而只涉及 Y^a 。因此, 某种意义上来说, 主层直接效应是子人群中的总效应, 而非直接效应, 因而没有太大的现实意义。在现实中, 如果 B 的取值过多, 那么我们的分层也就越多, 统计效力也就越低。Robins 在 1986 年引入了主层直接效应概念, Rubin 在 2004 年推广了这一概念。Frangakis 和 Rubin 在 2002 年用主层概念分析了矛盾事件。

知识点 22.3: 时异治疗的生存分析 (原书第 282 页)

第十七章介绍了如何用 G-方法估计算单一时间点治疗对失效时间结局的因果效应。第二十一章介绍了如何用 G-方法估计时异治疗的因果效应。综合这两章, 我们也可以用 G-方法估计时异治疗对失效时间结局的因果效应。以下我们会介绍两种方法, 分别基于 G-公式和逆概率加权, 从而估计治疗策略 \bar{a} 下的反事实风险 $\Pr[D_{k+1}^{\bar{a}} = 1]$ 。

$\Pr[D_{k+1}^{\bar{a}} = 1]$ 在 G-公式中可以定义为:

$$\sum_{\bar{l}_k} \sum_{j=0}^k \Pr[D_{j+1} = 0 | \bar{A}_j = \bar{a}_j, \bar{L}_j = \bar{l}_j, D_j = 0] \prod_{s=0}^j \left\{ \Pr[D_s = 0 | \bar{A}_{s-1} = \bar{a}_{s-1}, \bar{L}_{s-1} = \bar{l}_{s-1}, D_{s-1} = 0] f(l_s | \bar{a}_{s-1}, \bar{l}_{s-1}, D_s = 0) \right\}$$

我们通过对各时间点的危害 $\Pr[D_{k+1} = 0 | \bar{A}_k = \bar{a}_k, \bar{L}_k = \bar{l}_k, D_k = 0]$ 和 $f(l_k | \bar{a}_{k-1}, \bar{l}_{k-1}, D_k = 0)$ 建

模, 从而得到这个插入式 G-公式。我们在第十七章介绍过用来估计危害的 logistic 模型。更多细节, 请参考 Young 在 2011 年发表的论文。

我们还可以用逆概率加权代替上述 logistic 模型, 此时的非稳定权重为

$$W_k^{\bar{A}} = \prod_{m=0}^k \frac{1}{f(A_m | \bar{A}_{m-1}, \bar{L}_m)}。更多细节, 参见 Hernan 在 2001 年发表的论文。$$

中英词汇对照表

Accelerated failure time model - 加速失效时间模型

Additive - 加性, 加成

Additive scale - 加法尺度

Administrative end of follow-up - 跟踪的强制结束

Ancestor - 上游变量

Antagonistic - 拮抗的

Artificial censoring - 人工删失

Ascertainment bias - 确认偏移

Associational measure - 相关量度

Attributable fraction - 归因比例

Average causal effect - 因果效应的均值

Backdoor criterion - 后门准则

Backdoor path - 后门路径

Backward elimination - 后向消元

Bias - 偏移

Bias-variance tradeoff - 偏差方差权衡

Bias under the null - 零值下的偏移

Bootstrapping - 自举法

Calibrated - 被校准的

Cause - 诱因

Causal effect - 因果效应

Causal effect modifier - 因果性修饰因子

Causal diagram - 因果图

Causal inference - 因果推断

Causal Markov assumption - 因果性马尔科夫假设

Causal null hypothesis - 因果零假设

Censor - 删失

Channeling - 引流效应

Child - 下游变量

Clever covariate - 聪明变量

Cohort study - 队列研究

Collapsibility - 伸缩性

Collider - 对撞变量

Common cause - 共同诱因

Common effect - 共同后果

Compatibility interval - 相容性区间。

Competing event - 矛盾事件

Complier - 配合者。

Component - 成分

Compositional epistasis - 组合上位性

Compound treatment - 复合治疗

Conditional effect measure - 条件效应量度

Conditional exchangeability - 有界互换性, 条件互换性, $Y^a \perp\!\!\!\perp A|L$

Conditional independence - 条件独立

Conditionality principle - 条件性准则

Conditionally randomized experiment - 条件随机试验

Confounder - 混杂变量

Confounding - 混杂(因素)

Confounding by indication - 指征混杂

Consistency - 一致性

Consistent estimator - 一致估计量

Counterfactual outcome - 反事实结局

Controlled directed effect - 受控直接效应

Credible interval - 可信区间

Cross fitting - 交叉拟合

Cross validation - 交叉验证; 交互效度分

Cumulative incidence - 累积发生率

D-separation - 有向分离, D-分离

Decision nodes - 决策节点

Defier - 对抗者。

Degree of belief - 信念度

Descendant - 下游变量

Deterministic - 命定(的)

Difference-in-difference - 双重差分

Differential - 有差, 有差异

Directed acyclic graph - 有向无环图

Do-calculus - 介入算法

Doomed - 注定

Double-blind placebo-controlled - 双盲安慰剂对照

Doubly robust model - 双重稳健模型

Dynamic treatment strategy - 动态治疗策略

Effect estimate - 效应估计

Effect measure - 效应量度

Effect modification - 效应修饰

Effectiveness - 效

Efficacy - 效益

Estimand - 待估值

Etiological fraction - 病因分数

Excess fraction - 超出比例

Exchangeability - 互换性, $Y^a \perp\!\!\!\perp A$

Exclusion restriction - 排他性限制

Experimental treatment assumption - 试验-治疗假设

Failure time - 失效时间

Faux - 仿制的

Faithfulness - 忠实性

Fine point - 精讲点

Forward selection - 前向选择

Frailty - 脆弱

Front door criterion - 前门准则

G-computation - G-计算

G-estimation - G-估算

G-formula - G-公式

G-method - G 方法

G-null - G-零值

Generalized additive model - 广义相加模型

Hazard ratio - 危害比

Head-to-head trial - 头对头研究

Healthy worker bias - 健康工人偏移

Helped - 受益

Heterogeneity of causal effect - 因果效应异质性

Homoscedasticity - 方差齐性

Hurt - 受害

Identifiable - 可识别的

Identification - 可识别(性)

Ignorability - 可忽略性

Ill-defined - 劣定的

Immune - 免疫

Incidence-prevalence bias - 发病率-流行率偏移

Identity - 恒等

Individual causal effect - 个体因果效应

Influence diagram - 影响图

Informative censoring - 不对称删失

Intention-to-treat - 治疗意向

Interaction - 交互作用

Instrumental variable - 工具变量

Interference - 干扰

Inverse probability weighting - 逆概率加权

Iterated conditional expectation - 迭代期望

Joint intervention - 联合干预

Kernel regression - 核回归

Linear mean model - 线性均值模型

Link function - 联系函数

Linkage disequilibrium - 连锁不平衡

Longitudinal data - 纵向数据

Marginal structural model - 边缘结构模型。

Marginally randomized experiment - 边缘随机试验

Mendelian randomization - 孟德尔随机化

Misclassification - 错分类, 误分类

Missing at random - 随机缺失

Missing completely at random - 完全随机缺失

Model misspecification - 模型设定错误

Modifier - 修饰因子

Moralization - 教化准则

Multiplicative - 乘性

Multiplicative scale - 乘法尺度

Multiplicative survival model - 乘积生存模型

Natural bounds - 自然界限。

Natural directed effect - 自然直接效应

Natural value of treatment - 治疗的自然取值

Negative outcome control - 阴性结局对照

Nested structural model - 嵌入式结构模型

Neural network - 神经网络

No unmeasured confounding - 不存在未知混杂

Non-inferiority trial - 非劣效性试验

Nondifferential - 无差, 无差异。

Nondifferentiality - 差异性

Nonparametric structural equation model - 非参数化结构方程

Not missing at random - 非随机缺失

Nuisance parameter - 冗余参数

Nuisance sample - 冗余样本
Null preservation - 零值保留
Number needed to treat - 须治数
Observational study - 观察性研究
Odds ratio - 比值比
Outcome regression - 结局回归
Out-of-sample testing - 样本外检验
Parent - 母变量
Partial exchangeability - 部分互换性
Parsimonious - 简约的
Person-time - 人时
Per-protocol - 依方案
Placebo test - 安慰剂检验
Population stratification - 人群分层
Positive function - 正函数
Positivity - 正数性
Potential outcome - 潜在结局
Principal stratum direct effect - 主层直接效应
Propensity score - 倾向性评分
Prospective study - 前瞻性研究
Pseudo-population - 虚拟人群
Qualitative effect modification - 质的效应修饰
Random variability - 随机变异性
Randomization-based inference - 基于随机化的推
Rank preservation - 保序性
Regression discontinuity analysis - 断点回归分析
Residual - 残差
Residual confounding - 残余混杂
Response type - 回应类型
Restricted mean survival time - 受限生存时间均

Restriction - 限制

Reverse causation - 逆向因果关系

Ridge regression - 岭回归；嵴回归

Risk periods - 风险期

Sample splitting - 样本分割

Saturated model - 饱和模型

Sequentially randomized experiments - 时序性随机试验。

Sharp bounds - 极端界限

Sharp causal null hypothesis - 极端因果零假设

Single world intervention graph - 单一世界干涉图

Smoothing - 平滑

Stable unit treatment value assumption - “稳重治疗”假设

Standardization - 标准化

Standardized morbidity ratio - 标准化发病率

Static treatment strategy - 静态治疗策略

Stepwise selection - 逐步筛选

Stratification - 分层计算

Structural nested cumulative failure time model - 结构嵌入累积失效时间模型

Structural nested cumulative survival time model - 结构嵌入累积生存时间模型

Structural nested model - 结构嵌入模型

Super-population - 超级人群

Subadditive - 劣加性

Submultiplicative - 劣乘性

Sufficient component cause model - 充分成因模型

Sufficient set for confounding adjustment - 混杂调整的充分集合

Superadditive - 超加性

Supermultiplicative - 超乘性

Surrogate confounder - 混杂变量的替代

Surrogate effect modifier - 修饰因子替代物

Survival analysis - 生存分析

Survival curve - 生存曲线

Synergistic - 协同的

Target trial - 靶标试验

Technical point - 知识点

Time-dependent - 时依

Time-fixed - 固定时间的

Time-varying - 时异(性)

Transportability - 可移植性

Treatment-confounder feedback - 治疗-混杂反馈

Treatment variation irrelevance - 治疗差异无关紧要

Truncated normal distribution - 截断正态分布

Two-stage-least-squares - 双阶最小二乘法

Validation sample - 验证样本。

Well-defined - 良定的

Years of life lost - 生命损失年