因果推断实践

基于 RHC 数据的方法比较与 R 语言实现

作者：晨瀚宇（Chanw）

时间：April 25, 2026

版本：2.0

同一份数据、同一个因果问题，十种方法的估计与比较。

版权声明

书名：因果推断实践——基于 RHC 数据的方法比较与 R 语言实现

作者：晨瀚宇（小红书：Chanw）

版本：v2.0（2026 年 4 月）

购买渠道：小红书搜索「Chanw」

目录

第 1 章 问题与数据——RHC 争议的起点

内容提要

- 理解右心导管 RHC 争议的临床背景与因果推断意义
- 掌握 RHC 数据集的结构、关键变量与基线分布
- 识别观察数据中"适应证混杂"的表现形式
- 计算朴素关联并理解它为什么不能直接当因果效应

1996 年，Alfred Connors 和同事在 JAMA 发表了一项覆盖五所教学医院、5735 名 ICU 危重症患者的观察性研究 [9]。研究的核心发现让整个重症医学界意外：接受右心导管监测的患者，180 天死亡率比未接受 RHC 的患者**高**了大约 7.5 个百分点。右心导管是一根从颈静脉或锁骨下静脉送入肺动脉的导管。简单来说，它是一个实时的血流动力学探测器：心脏泵了多少血、肺动脉里的压力有多高、血液里的氧够不够用，这些信息 ICU 医生都能从导管上读到。在 Connors 的论文发表之前，绝大多数临床医生默认这些信息有助于指导治疗、改善预后。Connors 的数据说的却是相反的故事：插了管的病人死得更多。

这篇论文引发的争论持续了近十年。支持者认为 RHC 本身可能带来导管相关感染、心律失常、气胸等并发症，监测信息也可能诱导过度干预。反对者则指出一个根本性的统计学问题：接受 RHC 的病人本来就更重，死亡率高也许和导管无关，纯粹是因为这群人的基线风险就高。这个问题用因果推断的语言来说，叫做**适应证混杂**，英文称 confounding by indication。医生决定是否插管时参考的正是病情严重程度，而病情严重程度本身又直接影响死亡率。处理变量和结局变量共享一个共同原因，观察到的关联就被污染了。

这本书只回答一个问题：**RHC 是否因果地增加了 ICU 患者的 180 天死亡率？**每一章用一种不同的因果推断方法来回答它，最后汇总比较。同一份数据、同一个问题、九把不同的刀，读者可以亲手看到每种方法的假设、操作和结论有什么异同。

1.1 为什么不能做随机试验

回答因果问题最直接的办法是随机对照试验，简称 RCT。把 5735 名患者随机分成两组，一组插管、一组不插管，比较死亡率。随机化保证了两组在所有已知和未知的混杂因素上平均相同 [24]，观察到的死亡率差异就可以归因于 RHC 本身。

但 RHC 的 RCT 始终没有做成。伦理审查是第一道障碍：如果临床医生认为某个患者需要血流动力学监测，随机把他分到"不插管"组意味着剥夺了一项临床医生认为必要的诊断手段。操作上的阻力同样大——重症科医生对自己的临床判断有强烈信念，愿意让抛硬币决定自己病人是否插管的医生很少。哪怕启动了试验，知情同意的获取在危重症患者身上极其困难，患者本人往往无法签字，家属在极度焦虑中也不太可能冷静权衡研究方案。

RCT 做不了，因果问题还要回答。这就是观察性因果推断存在的理由 [13]。Connors 和同事手上有一份丰富的观察数据，记录了患者入 ICU 时的人口学特征、病史、生理指标和病情严重程度评分。利用这些协变量，我们可以尝试在统计上"模拟"随机化的效果，把适应证混杂剥离出去。接下来的九章就是九种不同的剥离策略。

1.2 因果推断的根本困难

定义 1.1 (潜在结果与因果效应)

对个体 i，定义两个潜在结局：$Y_i(1)$ 表示接受 RHC 后的 180 天死亡状态，$Y_i(0)$ 表示未接受 RHC 的 180 天死亡状态。个体因果效应为 $Y_i(1) - Y_i(0)$。总体平均处理效应为

$$\text{ATE} = E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)].$$

[24]

这个框架的核心困难在于：对每一个真实的病人，我们只能观察到一个潜在结局。插了管的人看到了 $Y(1)$，没插管的人看到了 $Y(0)$，另一个永远缺失。Holland 把这件事称为” 因果推断的根本问题”[14]。个体因果效应在原理上不可识别，我们能做的只是估计总体层面的平均效应 ATE。

估计 ATE 需要用观察到的数据去填补缺失的那一列。最朴素的做法是把 RHC 组的平均结局当作 $E[Y(1)]$，把非 RHC 组的平均结局当作 $E[Y(0)]$，直接相减。这个做法在 RCT 中完全合法，因为随机化保证了两组可交换。但在观察数据中，RHC 组和非 RHC 组的病人构成系统性地不同，直接比较混进了混杂的成分。下面我们用数据来看这种” 不同” 到底有多大。

1.3 数据概览

本书使用的数据来自 Connors et al. (1996) 的原始研究，经过清洗后保存在 `data/rhc.csv`。数据集包含 5735 名 ICU 患者，49 个变量。处理变量为 *rhc*，取值 0 或 1；结局变量为 *death180*，取值 Yes 或 No。关键协变量包括 APACHE III 评分、平均动脉压、肌酐、白蛋白等生理指标，以及年龄、性别、种族、保险类型等人口学变量。

下面的代码读入数据并预览结构。

```
library(tidyverse)
set.seed(2026)

# 读入数据——here::here() 保证路径相对于项目根目录
d <- read_csv(here::here("data", "rhc.csv"), show_col_types = FALSE)
dim(d)  # 5735 行 x 49 列

# 创建二分类结局变量
d <- d |> mutate(death180_bin = ifelse(death180 == "Yes", 1, 0))

# 预览 10 个关键变量的前 6 行
key_vars <- c("rhc", "death180", "age", "sex", "apache_score",
              "blood_pressure", "creatinine", "albumin",
              "heart_rate", "respiratory_rate")
head(d[, key_vars], 6)
```

数据预览

数据有 5735 行、49 列。表 1.1 展示了前 6 名患者在 10 个关键变量上的取值。可以直观看到变量的量纲差异很大：*apache_score* 在 38–82 之间，*blood_pressure* 在 41–115 mmHg 之间，*creatinine* 在 0.6–3.6 mg/dL 之间。这种量纲差异意味着后续做匹配或加权时需要标准化。

表 1.1: RHC 数据集前 6 行预览，选取 10 个关键变量

rhc	*death180*	*age*	*sex*	*apache*	*bp*	*creat*	*alb*	*hr*	*rr*
0	No	70.3	Male	46	41	1.20	3.50	124	10
1	Yes	78.2	Female	50	63	0.60	2.60	137	38
1	No	46.1	Female	82	57	2.60	3.50	130	40
0	Yes	75.3	Female	48	55	1.70	3.50	58	26
1	Yes	67.9	Male	72	65	3.60	3.50	125	27
0	No	86.1	Female	38	115	1.40	3.10	134	36

1.4 基线失衡：谁接受了 RHC？

如果 RHC 的使用是随机的，RHC 组和非 RHC 组在所有基线变量上应该高度相似。现实中这两组的构成差异有多大？我们需要一个指标来衡量差距的大小。

标准化均值差，简称 SMD，做的事情很简单：把两组之间某个变量的均值差除以合并标准差，消除量纲的影响。SMD = 0.50 意味着处理组的均值比对照组高半个标准差，这是一个相当大的失衡。SMD = 0.10 意味着差距只有十分之一个标准差，通常被认为是可接受的平衡。超过 0.10 就提示存在需要处理的失衡。

```r
library(tableone)

# 选取 12 个关键协变量
vars <- c("age", "sex", "apache_score", "blood_pressure",
          "heart_rate", "respiratory_rate", "creatinine",
          "albumin", "hematocrit", "wbc", "temperature",
          "das_index")

# 按 RHC 分组计算 Table 1，同时输出 SMD
d <- d |> mutate(rhc_label = ifelse(rhc == 1, "RHC", "No RHC"))
tab1 <- CreateTableOne(vars = vars, strata = "rhc_label",
                       data = d, test = FALSE, smd = TRUE)
print(tab1, smd = TRUE)
```

基线比较

表 1.2 汇总了 12 个关键协变量在两组之间的分布。RHC 组有 2184 人，非 RHC 组有 3551 人。

表 1.2: RHC 组与非 RHC 组基线特征比较

变量	**No RHC** ($n = 3551$)	**RHC** ($n = 2184$)	**SMD**
Age, mean (SD)	61.76 (17.29)	60.75 (15.63)	0.061
Male, %	53.9	58.5	0.093
APACHE Score, mean (SD)	50.93 (18.81)	60.74 (20.27)	**0.501**
Blood Pressure, mean (SD)	84.87 (38.87)	68.20 (34.24)	**0.455**
Heart Rate, mean (SD)	112.87 (40.94)	118.93 (41.47)	0.147
Respiratory Rate, mean (SD)	28.98 (13.95)	26.65 (14.17)	0.165
Creatinine, mean (SD)	1.92 (2.03)	2.47 (2.05)	**0.270**
Albumin, mean (SD)	3.16 (0.67)	2.98 (0.93)	0.230
Hematocrit, mean (SD)	32.70 (8.79)	30.51 (7.42)	0.269
WBC, mean (SD)	15.26 (11.41)	16.27 (12.55)	0.084
Temperature, mean (SD)	37.63 (1.74)	37.59 (1.83)	0.021
DAS Index, mean (SD)	20.37 (5.48)	20.70 (5.03)	0.063

失衡最严重的是 APACHE III 评分，SMD = 0.50，RHC 组均值 60.74 远高于非 RHC 组的 50.93。APACHE 是 ICU 常用的病情严重程度综合评分，分数越高代表病情越重、预期死亡率越高。RHC 组的 APACHE 均值比非 RHC 组高将近 10 分，换算成标准差单位就是半个标准差的差距。平均动脉压的 SMD 也达到 0.46，RHC 组血压更低，这和临床直觉一致：血流动力学不稳定的病人更可能被插管监测。肌酐 SMD = 0.27，RHC 组肾功能更差。白蛋白和血球容积的 SMD 均在 0.23–0.27 之间。

图 1.1 用密度直方图展示了两组 APACHE 评分的分布。RHC 组的分布明显右移，高分段的密度更大。这种右移直观地说明了适应证混杂的核心机制：病情越重的患者越可能被插管，而病情越重本身就意味着死亡风险越高。

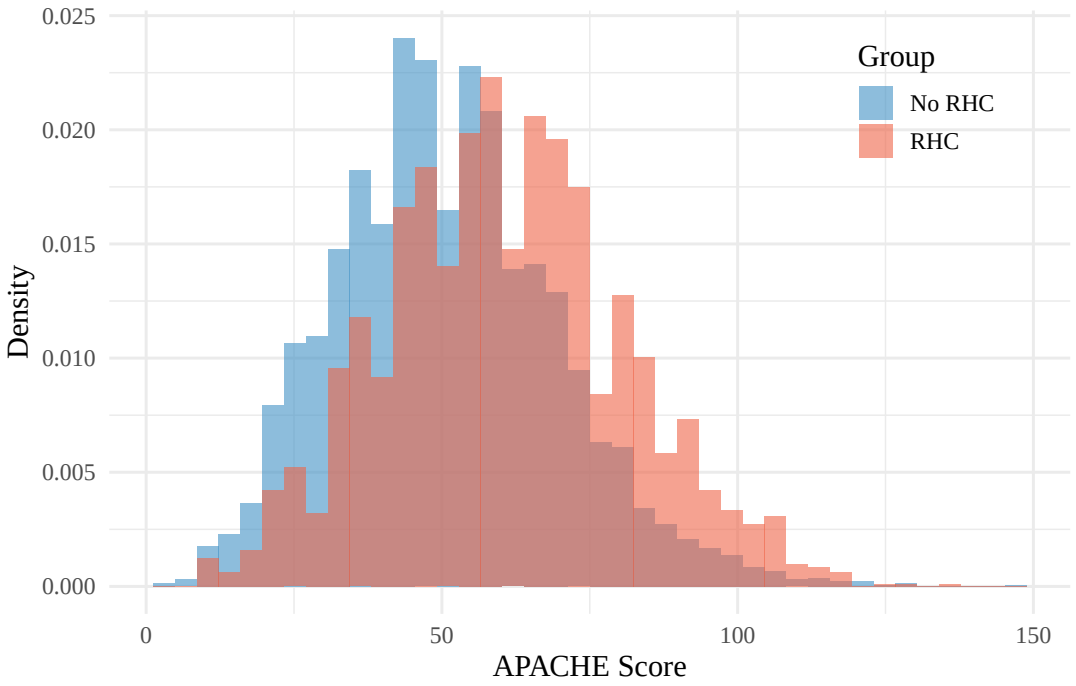

图 1.1: RHC 组与非 RHC 组的 APACHE III 评分分布。RHC 组分布明显右移，表明接受导管监测的患者病情更重。

图 1.2 用 Love plot 展示了全部 12 个协变量的 SMD。红色虚线标记 SMD = 0.1 的阈值。12 个变量中有 8 个

超过了 0.1，APACHE 评分和血压的失衡尤其严重。这意味着直接比较两组的死亡率，混杂的成分不可忽略。

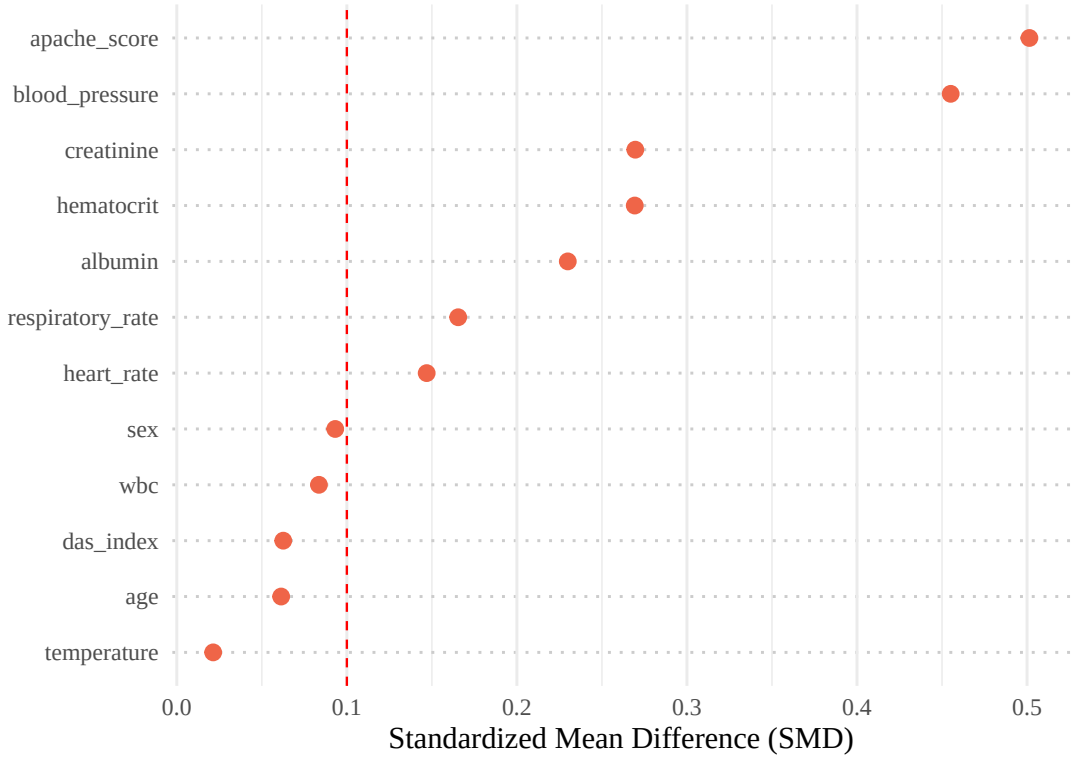

图 1.2: 12 个基线协变量的标准化均值差。红色虚线为 0.1 阈值，8 个变量超过阈值，APACHE 评分 SMD 达到 0.50。

1.5 朴素关联：粗死亡率差异

在做任何调整之前，先看最简单的数字：两组的 180 天粗死亡率各是多少?

```
# 按 RHC 分组计算 180 天死亡率
d |>
  group_by(rhc) |>
  summarise(
    n        = n(),
    deaths   = sum(death180_bin),
    mortality = mean(death180_bin),
    .groups = "drop"
  )
```

粗死亡率

非 RHC 组 3551 人中 1650 人在 180 天内死亡，死亡率 46.5%。RHC 组 2184 人中 1179 人死亡，死亡率 54.0%。粗死亡率差异为 $54.0\% - 46.5\% = 7.5$ 个百分点，RHC 组更高。

图 1.3 直观展示了这个差距。

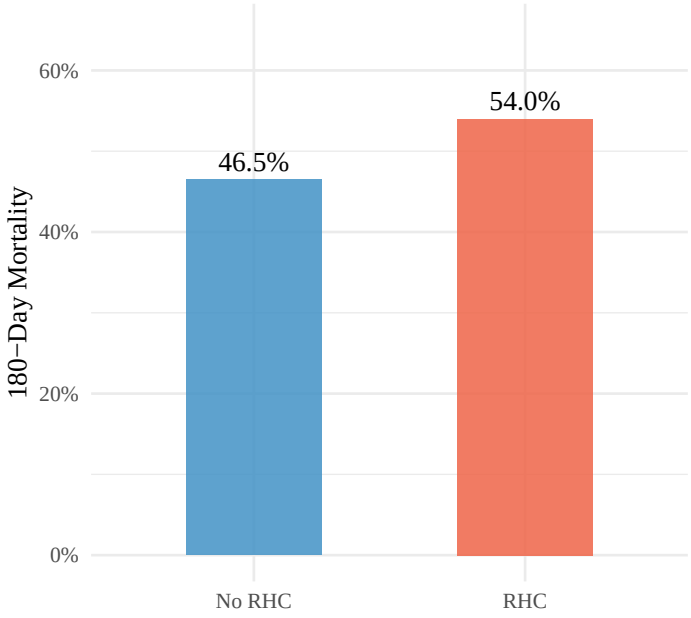

图 1.3: RHC 组与非 RHC 组的 180 天粗死亡率。RHC 组高出约 7.5 个百分点。

停下来想一想。非 RHC 组死亡率 46.5%，RHC 组 54.0%，差了 7.5 个百分点。如果你是一位 ICU 主任，看到这个数字会立刻下令停用右心导管吗？在翻到下一段之前，试着列出至少一个理由，解释为什么这个差距可能不是 RHC 本身造成的。

7.5 个百分点的差距看起来不小，但它能直接被解读为"RHC 导致死亡率升高 7.5 个百分点"吗？前面的 Table 1 已经说明了答案：不能。RHC 组的 APACHE 均分高了将近 10 分，血压低了 17 mmHg，肌酐高了 0.55 mg/dL。这些指标全都指向同一个事实——RHC 组的病人入 ICU 时病情就更重。病情更重的人死亡率更高，这和导管本身无关。

定理 1.1 (雷区)

观察数据中的粗关联混合了因果效应和混杂偏倚两部分。在 RHC 数据中，适应证混杂的方向是正向的：病情越重 → 越可能插管，同时病情越重 → 死亡率越高。这意味着粗关联会高估 RHC 的有害效应。如果 RHC 真的有害，真实效应可能小于 7.5 个百分点；如果 RHC 实际上无害甚至有益，粗关联仍然可能显示正向差异，因为混杂的正偏倚足以盖过保护效应。分辨这两种情形，需要在统计上把混杂剥离出去，这正是后续九章要做的事情。

1.6 全书路线图

本书用九种方法回答同一个问题。第 2 章用 DAG 梳理变量之间的因果结构，明确哪些协变量需要调整。第 3 章从最简单的回归调整开始，逐步加入协变量，观察估计值如何变化。第 4 章用 G 计算做标准化，预测反事实结局。第 5 章转向倾向得分，包括匹配、逆概率加权和 overlap weight。第 6 章手动实现 AIPW，体会双重稳健的机制。第 7 章引入机器学习，用 Super Learner、DML 和 TMLE 分别估计。第 8 章做敏感性分析，回答"如果存在未观测混杂，结论还站得住吗"。第 9 章用因果森林探索异质性效应，看 RHC 对哪些亚群有害、对哪些亚群可能有益。第 10 章把所有方法的估计汇总到一张表里，比较点估计和置信区间，讨论方法之间的一致性与分歧。

每章末尾会更新一张累积对比表，格式如下：

表 1.3: 方法对比表——第 1 章：朴素关联

方法	ATE	95% CI	核心假设	局限
粗差异	0.075	—	无	未调整任何混杂

这张表目前只有一行。后续每章增加一行，到第 10 章合并为终极版本。

本章知识地图

表 1.4: 第 1 章核心概念与常见误解

核心概念	核心内容	常见误解	为什么错
右心导管 RHC	经静脉插入肺动脉的导管，实时监测血流动力学参数	RHC 是一种治疗手段	RHC 是诊断/监测工具，本身不直接改变治疗方案
适应证混杂	决定是否给予处理的因素同时影响结局，导致处理组和对照组不可比	只要样本量够大，混杂就会消失	样本量解决的是抽样变异，混杂是系统性偏倚，只有 RCT 或统计调整才能消除
潜在结果	每个个体在处理和不处理两种情境下各有一个潜在结局，$Y(1)$ 和 $Y(0)$	潜在结果可以同时观测	每个个体只能处于一种处理状态，另一个潜在结局永远缺失
ATE	$E[Y(1)]-E[Y(0)]$，总体平均处理效应	粗死亡率差异等于 ATE	粗差异 = ATE + 混杂偏倚；只有在随机化或充分调整后，粗差异才近似 ATE
SMD	标准化均值差，用于度量两组协变量平衡程度，< 0.1 为可接受	SMD 小就不需要调整	SMD 小说明该变量平衡，但其他变量仍可能失衡；要看全部协变量
APACHE III 评分	ICU 病情严重程度综合评分，本数据中 SMD = 0.50	APACHE 是唯一的混杂因素	血压、肌酐、白蛋白等也有显著失衡，需要同时调整多个混杂因素

第 2 章 因果结构与识别条件

内容提要

- 掌握潜在结果框架的核心记号 $Y(1)$、$Y(0)$，理解因果推断的根本问题
- 区分 ATE 和 ATT 两个估计量，知道它们回答的是不同的问题
- 用有向无环图 DAG 画出 RHC 研究的因果结构
- 理解可交换性、正值性、一致性三个识别假设各自解决什么问题
- 用 dagitty 从 DAG 自动推导最小调整集

上一章我们看到了一个事实：接受 RHC 的病人 180 天死亡率高出未接受组 6 个百分点。但这个差距几乎全部来自基线差异，病情重的人更容易被安排上导管，也更容易死亡。RHC 本身是否有害、有益、还是没有作用，光看原始数据分不清。本章要做的是搭建回答这个问题所需要的理论框架：什么叫”因果效应”、它跟数据里直接算出来的差距有什么区别、从观察数据里”识别”因果效应需要满足哪些条件。框架搭好之后，第 3 章开始我们会一种方法接一种方法地在 RHC 数据上动手。

2.1 潜在结果：因果推断的语言

讨论因果效应之前要先解决一个语言问题：怎么在数学上把”如果这个病人没有接受 RHC，他会怎样”这件事写下来。

先看一个具体的人。张三今年 68 岁，因为感染性休克住进了 ICU，APACHE 评分 72 分，属于中度危重。医生决定给他插右心导管。180 天后，张三死亡了。这是我们观察到的事实。但如果当初医生没有给张三插管呢？也许他同样会死亡，也许他反而能活下来。问题是张三只走了”插管”这一条路，”不插管”那条路上会发生什么，我们永远不知道。

日常用语里我们说”如果当初没做 X，结果会不会不同”，这是一个反事实的陈述，发生过的事情不能倒带重来。统计学需要一套符号把这种反事实的想法变成可以操作的对象，这套符号叫**潜在结果框架**，由 Rubin 在 1974 年系统提出 [24]，后来被 Holland 进一步阐述并命名为”Rubin 因果模型” [14]。

定义 2.1 (潜在结果)

对于个体 i，定义两个潜在结果：$Y_i(1)$ 是个体 i 接受处理时的结果，$Y_i(0)$ 是个体 i 不接受处理时的结果。在 RHC 的场景下，$Y_i(1)$ 是”这个病人如果被插了右心导管，180 天内是否死亡”，$Y_i(0)$ 是”这个病人如果没有被插右心导管，180 天内是否死亡”。 [24] ♣

停下来想一想。回到张三：他插了管，180 天后死亡，所以他的 $Y_i(1) = 1$。但他的 $Y_i(0)$ 是多少？如果不插管，他 180 天后是死还是活？这个数字我们永远不知道，因为张三已经走了”插管”这条路。

潜在结果这个词里最要紧的是”潜在”两个字。对任何一个真实的病人，我们只能观察到其中一个结果。如果他实际接受了 RHC，我们看到的是 $Y_i(1)$，而 $Y_i(0)$ 永远观察不到；反过来也一样。Holland 把这件事叫做**因果推断的根本问题**，英文称 the fundamental problem of causal inference：个体层面的因果效应 $Y_i(1) - Y_i(0)$ 永远无法直接计算，因为你只能看到等号右边两项中的一项 [14]。

这不是技术限制，是逻辑限制。就算你有完美的测量工具、无限的样本量、最强的超级计算机，你也没法让同一个病人在同一时刻既插管又不插管，然后比较两个结局。时间不能倒流，一个人只能走一条路，这是因果推断天然面对的约束。

定理 2.1 (雷区)

因果推断的根本问题容易被轻描淡写带过，但它的含义非常深远：任何声称在个体层面做出了因果判断的方法，背后一定依赖了某种不可验证的假设来”填补”那个缺失的反事实。没有免费午餐，所有因果推断方法都是在假设的支撑下绕过根本问题，差别只在于假设的强弱和合理性。后面几章里你会反复看到这一点：每种方法的核心区别归根结底是假设不同。

♡

2.2 ATE 与 ATT：两个不同的因果问题

个体因果效应 $Y_i(1) - Y_i(0)$ 观察不到，但如果我们退一步，不追求每个人的效应，而是去估计群体的平均效应，事情就变得有可能了。

举个简单的数字例子。假设 ICU 里有 100 个病人，如果让他们全部插管，预期有 55 人在 180 天内死亡；如果让他们全部不插管，预期有 50 人死亡。那么插管的平均因果效应就是 $55\% - 50\% = 5$ 个百分点，即平均而言插管增加了死亡风险。当然，这两个数字不可能同时观察到，因为每个人只能走一条路。但如果我们有办法分别估计这两个”假如全员接受同一处理”的平均结果，就能算出因果效应。这引出因果推断里最基本的两个估计量。

定义 2.2 (平均处理效应 ATE)

平均处理效应，英文 Average Treatment Effect，简称 ATE，定义为全体人群中潜在结果之差的期望：

$$\text{ATE} = E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)].$$

ATE 回答的问题是：如果把整个人群全部送去接受处理，平均结果会比全部不接受好多少？

♣

定义 2.3 (处理组平均处理效应 ATT)

处理组平均处理效应，英文 Average Treatment Effect on the Treated，简称 ATT，定义为实际接受处理的人群中潜在结果之差的期望：

$$\text{ATT} = E[Y(1) - Y(0) \mid A = 1].$$

ATT 回答的问题更具体：对于那些实际接受了处理的人，他们接受处理比不接受好多少？

♣

ATE 和 ATT 回答的是两个不同的政策问题。在 RHC 的场景下，ATE 问的是”如果所有 5735 个 ICU 病人都插导管 vs. 都不插导管，平均死亡率差多少”，这是一个关于全民政策的问题。ATT 问的是”对那 2184 个实际被插了导管的病人来说，插导管 vs. 不插导管的死亡率差多少”，这是一个关于当前临床决策质量的问题。

什么时候 ATE 和 ATT 相等？当处理效应在人群中是均匀的，即每个人的 $Y_i(1) - Y_i(0)$ 都一样时，ATE 和 ATT 自然相等。但如果处理效应因人而异，比如 RHC 对极危重病人有救命作用、对轻症病人则增加并发症风险，那么接受 RHC 的那批人，通常是更重的，跟全体人群的平均效应就会不同，ATE $\neq$ ATT。第 9 章讨论处理效应异质性时我们会回到这个话题。

选 ATE 还是 ATT 作为你的估计量，取决于你的研究问题。如果你是卫生政策制定者，想知道”推广 RHC 到所有 ICU 病人值不值得”，你关心的是 ATE。如果你是临床质控人员，想知道”当前被选中使用 RHC 的那批病人是否从中获益了”，你关心的是 ATT。两个估计量都是合法的因果量，错的是不说清楚自己在估哪个就直接报结果。

2.3 有向无环图：把因果假设画出来

潜在结果框架告诉我们”想估什么”，但没有告诉我们”观察数据里能不能估出来”。要回答后面这个问题，需要把我们对变量之间因果关系的假设显式地表达出来。Pearl 在 1995 年提出用**有向无环图**，英文 Directed Acyclic

Graph，简称 DAG，作为表达因果假设的工具 [18]。这个工具后来成为流行病学和社会科学研究设计的标准语言 [11, 19]。

DAG 的构成很简单。每个变量画一个节点，如果变量 X 对变量 Z 有直接因果作用，就从 X 到 Z 画一条有方向的箭头。”有向”指箭头有方向，”无环”指箭头不能形成闭合回路。在 RHC 研究里，我们关心的变量可以分成三类：处理变量 A 即是否接受 RHC、结局变量 Y 即 180 天死亡、协变量 L 即所有基线特征的总称。

图 2.1 画出了 RHC 研究的因果结构。三组协变量分别代表人口学特征，包括年龄、性别、种族、保险类型；疾病严重程度指标，包括 APACHE 评分、血压、心率、呼吸频率；合并症，包括癌症、心血管病、肝病、肾病。每组协变量既影响医生是否决定给病人上 RHC，即 $L \to A$，也影响病人 180 天的生死结局，即 $L \to Y$。$A \to Y$ 则是我们要估计的因果路径。

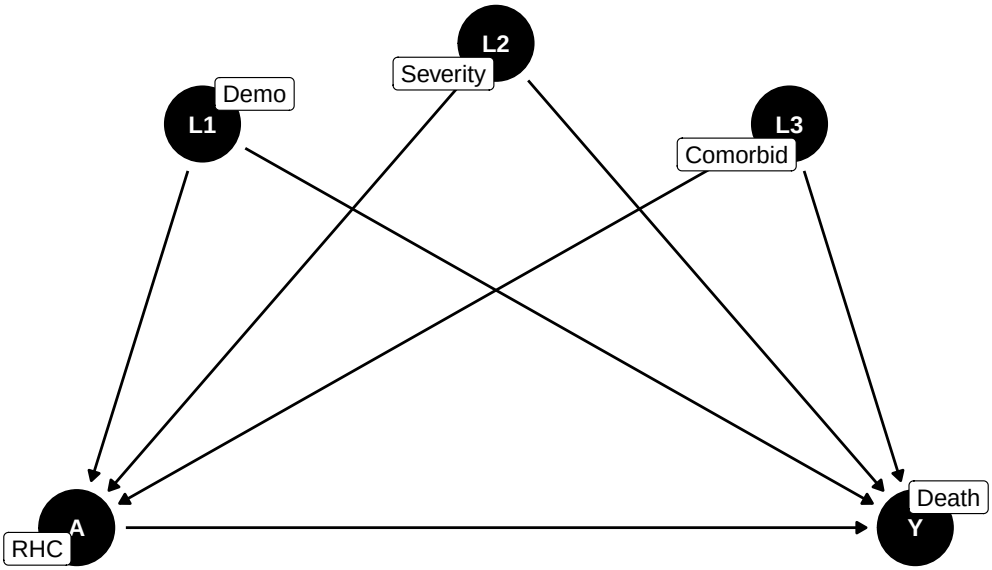

图 2.1: RHC 研究的有向无环图。A = 是否接受 RHC，Y = 180 天死亡，三组协变量 L 既影响处理分配又影响结局，构成混杂。读者应关注 L 到 A 和 L 到 Y 的双向箭头结构——这就是混杂的图形化表达。

DAG 的价值在于，一旦画出来，”该控制哪些变量”就有了形式化的判断准则，不再靠研究者的直觉拍板。Pearl 的 **后门准则**，英文 backdoor criterion，说：如果一组变量 L 能阻断所有从 A 到 Y 的非因果路径，同时不包含 A 的后代，那么在 L 上条件化就能识别 $A \to Y$ 的因果效应 [19]。在图 2.1 中，从 A 到 Y 的非因果路径就是 $A \leftarrow L \to Y$ 这条经过协变量的”后门路径”。控制 L 就能阻断它。

定理 2.2 (雷区)

并非控制的变量越多越好。如果你控制了一个碰撞因子，英文 collider，即一个同时被 A 和 Y 影响的变量，反而会打开一条本来关闭的非因果路径，引入新的偏倚。在 RHC 的语境下，假如”ICU 住院天数”同时受 RHC 和病人本身严重程度的影响，盲目控制它可能扭曲估计。DAG 的规则能帮你识别哪些变量该控制、哪些不该碰。经验法则”能控制的都控制”在因果推断里是错误的。 ♡

2.4 三个识别假设

DAG 帮我们画清了因果结构，但从”结构”到”从数据里算出因果效应”，中间还隔着三个数学假设。这三个假设是因果推断从观察数据中识别 ATE 的充分条件 [13]。三个假设缺少任何一个，ATE 都无法从数据中被唯一定出来。

2.4.1 可交换性：条件内的”准随机化”

上一章讲到 RCT 的核心承诺是”抛硬币让两组在所有变量上分布平衡”。观察研究里没有硬币，我们需要一个替代假设来近似达到同样的效果。

先用 RHC 的数据感受一下这个假设在说什么。假设我们只看 APACHE 评分在 70 到 75 分之间的那批病人，一共 200 人，其中 120 人上了导管、80 人没上。在这个小组里，上导管组的 180 天死亡率是 52%，没上导管组的死亡率是 48%。如果我们相信，在”APACHE 评分 70 到 75 分” 这个条件下，谁上导管谁不上，跟他们本来会不会死没有关系，就好像在这个小组里做了一次随机分配，那么 $52\% - 48\% = 4$ 个百分点这个差距就可以被解释为因果效应。

这个” 条件内的准随机化” 用数学写出来，就是下面的独立性条件。

定义 2.4 (可交换性)

给定协变量 L，潜在结果 $Y(a)$ 与实际处理分配 A 独立：

$$Y(a) \perp\!\!\!\perp A \mid L, \quad a \in \{0,1\}.$$

也叫无未测量混杂，英文 no unmeasured confounding，或条件可忽略性，英文 conditional ignorability。[13, 24]

♣

这条假设用大白话说就是：在协变量 L 取值相同的人群内部，谁接受处理谁不接受，跟他们的潜在结果没有关系。等价于说，L 已经包含了所有影响处理分配的混杂因素，没有遗漏。在 L 的每一层里，处理分配近似于随机的。

” 可交换性” 这个名字来自一个直觉：在 L 条件化之后，处理组和对照组的人是” 可以互换的”，把对照组的人放到处理组的位置上，他们的潜在结果分布不会变。这正是 RCT 中随机化自动保证的性质，观察研究里则需要通过假设来获取。

停下来想一想。回到刚才 APACHE 评分 70 到 75 分那个小组：可交换性要求的是，这个小组里上导管的 120 人和没上导管的 80 人，在” 所有影响死亡的因素” 上分布是一样的。如果这 120 人恰好合并症更多、血压更低，那即使 APACHE 评分相同，两组仍然不可交换，52% vs. 48% 的差距里仍然混着混杂。

在 RHC 数据里，可交换性要求我们测量并纳入了所有影响医生决定是否插导管、同时又影响病人 180 天死亡的变量。APACHE 评分是其中之一，因为评分高的病人既更可能被插管、也更可能死亡。血压、心率、肌酐、白蛋白、合并症诊断、DNR 状态都是同理。RHC 数据集的 49 个变量涵盖了丰富的基线信息，但可交换性要求的是**所有**混杂都被测到了，而” 所有” 这个词在观察研究里永远无法被验证。也许存在一个未记录在数据里的变量，比如主治医生的经验水平或医院的 ICU 编制情况，它同时影响了是否上 RHC 和病人的存活率，但我们无从知道。这就是可交换性假设最核心的脆弱点：它**不可检验**。第 8 章的敏感性分析会专门讨论” 假如可交换性被违反了、结论还能撑多远”。

2.4.2 正值性：每种病人都有两种可能

光有可交换性还不够。就算我们相信在 L 的每一层里处理分配是准随机的，还需要保证每一层里**确实有人**同时出现在处理组和对照组，否则那一层的因果效应根本无法比较。

先看一个直观的例子。假设 RHC 数据里 APACHE 评分 100 分以上的病人有 30 人，这些都是极危重患者，现实中 30 人全部被插了导管，没有一个人不插。这意味着在这个子组里，我们根本没有” 不插管” 的对照。这 30 个人如果不插管会怎样？数据里找不到任何参照，因果效应在这一层无从比较。

这个要求写成数学语言就是下面的条件。

定义 2.5 (正值性)

对于协变量 L 的每一个可能取值 l，接受处理和不接受处理的概率都严格大于零：

$$0 < P(A = 1 \mid L = l) < 1, \quad \text{对所有 } l \text{ 满足 } P(L = l) > 0.$$

也叫重叠假设，英文 overlap assumption。 [13]

正值性在直觉上很好理解。用第 1 章的语言说，如果某一层里只有处理组没有对照组，那这一层里就没有”孪生兄弟” 可以配对。

停下来想一想。回到 RHC 数据：APACHE 评分 100 分以上的那 30 个人全部插了管，$P(A = 1 \mid L = \text{APACHE} \geq 100) = 1$，正值性在这一层被违反了。反过来，DNR 状态明确的临终病人几乎全部没有插管，$P(A = 1 \mid L = \text{DNR})$ 接近于零，正值性同样岌岌可危。这两个极端层都是因果推断的盲区。

正值性违反在实务中很常见。ICU 里的临床实践往往有”绝对适应证” 和”绝对禁忌证”，APACHE 评分极高或血流动力学极不稳定的病人几乎一定会被上 RHC，而 DNR 状态明确的临终病人几乎一定不会被上 RHC。这些极端层的存在让正值性假设在边界上岌岌可危。

正值性违反有两种形式。一种是**结构性违反**：临床规则决定了某类病人 100% 接受处理或 100% 不接受处理，这种违反增加再多的样本量也无法修复，因为对照组根本不存在。另一种是**随机性违反**：理论上每层都有两种可能，但某些层的样本量太少，碰巧所有人都落在同一组。随机性违反可以通过增加样本量缓解，结构性违反则需要在定义目标人群时把那些极端层排除。第 5 章讲倾向得分时会手把手做正值性诊断：画出倾向得分分布图，看两组有没有重叠区域。

2.4.3 一致性：处理定义必须清晰

前面两个假设都是关于”数据里有没有足够信息” 的。第三个假设更底层，它关心的是处理变量本身的定义是否没有歧义。

继续用张三来理解。张三的数据里记录着 $A = 1$，即”接受了 RHC”。我们把他观察到的结局 $Y = 1$ 等同于潜在结果 $Y(1)$。但这里有一个隐含的前提：给张三插管这件事只有一种方式。如果张三的主治医生在入院 6 小时内就插了管，而李四的主治医生拖到了 36 小时后才插，两个人虽然都记录为 $A = 1$，但他们经历的”插管” 其实是不同版本的处理。如果插管时机本身影响了结局，那么”$A = 1$” 对应的 $Y(1)$ 就不是一个确定的量，而是因版本而异的模糊概念。

一致性假设就是要排除这种模糊。

定义 2.6 (一致性 / SUTVA)

对于个体 i，如果实际接受的处理为 $A_i = a$，那么观察到的结果就等于对应的潜在结果：

$$A_i = a \implies Y_i = Y_i(a).$$

这条假设包含两层含义：处理版本唯一，不存在”不同方式的 RHC” 导致不同结果；以及个体之间不存在干扰，一个病人是否接受 RHC 不影响另一个病人的结局。后一层在经济学文献中常叫 **SUTVA**，全称 Stable Unit Treatment Value Assumption。 [24]

停下来想一想。回到张三和李四：张三入院 6 小时插管，李四入院 36 小时插管，两人的 A 都记为 1。一致性要求这两种”插管” 在结局层面没有系统性的差异，否则 $Y(1)$ 就不是一个明确的数。Connors 1996 的研究把”入院 24 小时内是否使用 RHC” 作为处理定义，在一定程度上锁定了时间窗口，减轻了版本模糊的问题。

实际操作中，不同医生插管的时机不同，导管型号不同，测量的血流动力学参数之后做出的治疗调整也不同。如果这些操作上的差异导致了不同的结局，那么”$A = 1$” 实际上对应了多个不同版本的处理，$Y_i(1)$ 就不是一个明确的量了。

干扰假设在 ICU 场景里相对容易满足：一个病人是否插管，通常不会改变另一个病人的 180 天结局。但如

果换成疫苗研究、社交网络传播、或者课堂教学实验，干扰假设就很难成立，需要专门的方法来处理。

定理 2.3 (雷区)

一致性假设在很多论文里被一笔带过，但它在定义模糊的处理变量下经常失败。典型的例子是”运动对健康的因果效应”：$A=1$ 是”运动”，但运动包括跑步、游泳、举重、每周三次还是五次、每次半小时还是两小时。不同版本的”运动”可能有截然不同的效应，把它们合成一个二分变量会模糊因果含义。在 RHC 研究里这个问题相对轻微，因为”24 小时内是否插管”的操作定义比较明确，但读者在自己的研究中遇到处理变量定义模糊时，应该在分析之前先想清楚一致性是否成立。 ♡

2.5 从假设到识别公式

三个假设各自解决一个问题：可交换性保证比较是公平的，正值性保证比较是可行的，一致性保证比较的对象是明确的。三者同时成立时，我们可以用观察到的数据写出 ATE 的识别公式。

在写公式之前，先用一个简化到极致的数字例子把逻辑走一遍。假设我们只按 APACHE 评分把 5735 个病人分成三层：低危、中危、高危。

首先看每一层内部的情况。低危层有 2000 人，其中 400 人插管、1600 人没插管，插管组死亡率 30%、未插管组死亡率 25%。中危层有 2500 人，其中 1000 人插管、1500 人没插管，插管组死亡率 55%、未插管组死亡率 50%。高危层有 1235 人，其中 784 人插管、451 人没插管，插管组死亡率 75%、未插管组死亡率 68%。

其次把各层的结果按人群比例加权平均。低危层占全体的 $2000/5735=34.9\%$，中危层占 $2500/5735=43.6\%$，高危层占 $1235/5735=21.5\%$。那么”假如全员插管”的平均死亡率是 $30\%\times 34.9\%+55\%\times 43.6\%+75\%\times 21.5\%=50.6\%$。”假如全员不插管”的平均死亡率是 $25\%\times 34.9\%+50\%\times 43.6\%+68\%\times 21.5\%=45.2\%$。ATE 就是 $50.6\%-45.2\%=5.4$ 个百分点。

这个”先在每一层里算处理组和对照组的结果，再按全体人群的层比例加权”的两步操作，就是下面这个公式在做的事：

$$\text{ATE}=E[Y(1)]-E[Y(0)]=E_L\big[E[Y\mid A=1,L]\big]-E_L\big[E[Y\mid A=0,L]\big].$$

等号左边的 $E[Y(1)]$ 是全体人群在潜在结果 $Y(1)$ 上的期望，这个量本身不能从数据里直接算。等号右边做了一个关键的替换：先在 L 的每一层里，用处理组的平均结果 $E[Y\mid A=1,L]$ 代替潜在结果 $E[Y(1)\mid L]$，这一步靠的是**可交换性**，即同一 L 层内处理组能代表全体的潜在结果分布，加上**一致性**，即处理组观察到的 Y 就是 $Y(1)$。然后再按 L 的边际分布把各层的条件期望加权平均，得到全体的 $E[Y(1)]$，这一步靠的是**正值性**，即每一层里确实有处理组的人可以算条件均值。

这个公式叫 **G 公式**，英文 G-formula，也叫标准化公式，英文 standardization formula，是 Robins 在 1986 年系统提出的 [20]。它是因果推断最基本的识别等式，后面几章会看到，回归调整、G 计算、AIPW、TMLE 等方法都是这个公式的不同实现方式。

理解这个公式的关键在于”两步走”的结构。内层的 $E[Y\mid A=1,L]$ 是在每一个协变量组合 $L=l$ 内部看处理组的平均结果，这一步回答的是”这一层里处理的效果是什么”。外层的 $E_L[\cdot]$ 是按 L 在全体人群中的分布做加权平均，把各层的局部效果汇总成全体的 ATE。如果你只做内层不做外层，得到的是条件效应；如果你跳过内层直接看全体均值差，得到的是上一章说的边际关联，混杂没有被去除。G 公式的精髓在于把这两步合在一起。

停下来对照一下。刚才那个三层数字例子里，内层就是”低危层插管组死亡率 30%、未插管组 25%”这些数字，外层就是”按 34.9%、43.6%、21.5% 加权”这一步。公式里的 $E_L[\cdot]$ 就是这个加权平均的数学写法。

命题 2.1 (识别与估计的区别)

上面这个等式是一个识别结果，它把不可观测的 $E[Y(1)]$ 跟可观测的条件期望联系起来。识别先于估计：如果三个假设不成立，这个等号就不成立，不管你用什么统计方法算出的数字都没有因果解释。估计方法的选择用参数回归还是非参方法来拟合 $E[Y \mid A, L]$，是识别成立之后的事。

用一个类比来加深理解。识别好比确认你手里的地图画的是你要去的那个城市，估计好比在地图上选一条具体的路线。如果地图画的城市就是错的，你的路线规划得再精细也到不了目的地。因果推断文献里相当一部分争论都发生在识别层面，" 你的可交换性假设合理吗"" 你有没有遗漏混杂"" 你的处理定义清楚吗"，这些问题比" 你用 GLM 还是 random forest 拟合" 重要得多。

笔记 潜在结果框架和 DAG 在因果推断的历史上各自独立发展。Rubin 1974 年提出潜在结果框架时，Pearl 还没有开始研究因果推断；Pearl 1995 年提出 DAG 用于因果分析时，Rubin 的框架已经在统计学界流行了二十年。两套语言之间的关系一度有争议：Rubin 认为 DAG 过于依赖非参结构假设，Pearl 认为潜在结果框架缺少表达因果结构的工具。到 2000 年代之后，主流因果推断教科书，如 Hernán & Robins 2020，已经把两套工具合并使用：用 DAG 表达因果假设和判断调整集，用潜在结果定义估计量和推导识别公式。本书也遵循这个惯例。

2.6 用 dagitty 推导调整集

上面的理论告诉我们，要识别 ATE，需要对一组满足后门准则的变量 L 条件化。在简单的 DAG 里可以用肉眼判断，但当变量多到 49 个时，人工判断很容易遗漏或犯错。dagitty 包可以从 DAG 的形式定义自动推导最小调整集 [25]。

本章继续使用第 1 章介绍的 RHC 数据集，$n = 5735$。下面的代码定义了 RHC 的因果结构，然后用 `adjustmentSets()` 推导出哪些变量必须控制。

```
set.seed(2026)
library(dagitty)

# 定义 RHC 的因果结构——三组协变量都是混杂
# 每组既影响医生是否决定上 RHC，也影响病人结局
rhc_dag <- dagitty("dag {
  severity    [pos=\"1,0\"]
  comorbidity [pos=\"2,0\"]
  demographics [pos=\"0,0\"]
  A           [pos=\"0.5,1.5\"]
  Y           [pos=\"2,1.5\"]
  severity -> A
  severity -> Y
  comorbidity -> A
  comorbidity -> Y
  demographics -> A
  demographics -> Y
  A -> Y
}")
exposures(rhc_dag) <- "A"
outcomes(rhc_dag)  <- "Y"

# dagitty 自动推导最小调整集
adjustmentSets(rhc_dag, type = "minimal")
```

结果解读

`adjustmentSets()` 返回的最小调整集为 {*comorbidity*, *demographics*, *severity*}，即三组协变量全部需要控制。这个结果在 RHC 的 DAG 结构下是唯一的最小调整集，因为三组变量都同时连接 A 和 Y，缺少任何一组都会留下未阻断的后门路径。

在实际操作中，这三组协变量对应 RHC 数据集里的 49 个具体变量：*severity* 包含 *apache_score*、*blood_pressure*、*heart_rate*、*respiratory_rate* 等连续生理指标；*comorbidity* 包含 *cancer*、*cardiovascular*、*renal*、*hepatic* 等二分类合并症；*demographics* 包含 *age*、*sex*、*race*、*medical_insurance* 等人口学变量。从第 3 章开始，每种方法都会把这些变量作为调整变量放进模型。

注意 `adjustmentSets()` 返回的是**最小**调整集，即变量数最少的合法集。有时同一个 DAG 会有多个最小调整集，都是合法的，用哪一个取决于变量的测量精度和数据可得性。如果你想看包含所有合法变量的完整集合，可以把 `type` 参数改为 `"canonical"`。

2.7 本章小结：知道该调整什么，下一步是怎么调整

本章从上一章的事实出发，搭建了回答”RHC 到底有没有因果效应” 所需的理论工具。潜在结果框架给了我们定义因果效应的语言，ATE 和 ATT 让我们明确了想估的目标量，DAG 让我们把因果假设画出来并形式化判断该控制哪些变量，三个识别假设则给出了” 观察数据里的条件期望可以替代潜在结果期望” 的条件。

这套框架回答了 WHAT 的问题：调整什么、调整的数学依据是什么。从第 3 章开始，我们进入 HOW 的问题。第 3 章会用最简单的工具回归调整，在 RHC 数据上逐步加入协变量，观察 RHC 的系数如何随调整集的变化而移动。回归调整是 G 公式最直观的参数化实现，但它也有自己的天花板，第 3 章结尾会讲清楚它的边界在哪里。

值得记住的是，本章搭建的框架对后续所有方法都是共用的。无论你用回归、G 计算、倾向得分、AIPW 还是 TMLE，三个识别假设不变，DAG 不变，估计量定义不变。变的只是**估计策略**：用什么统计方法去拟合 G 公式右边那个条件期望。如果你在后面某一章感到方法细节开始变得复杂，回来翻一翻本章的三个定义和识别公式，会帮你重新抓住主线。

本章知识地图

表 2.1: 第 2 章核心概念与常见误解

核心概念	核心内容	常见误解	为什么错
潜在结果 $Y(1), Y(0)$	同一个体在处理和不处理下的两个结果	潜在结果可以都观察到	根本问题：每人只走一条路，反事实永远缺失
ATE	全体人群 $E[Y(1) - Y(0)]$	ATE 就是两组均值差	两组均值差是边际关联，只在 RCT 下等于 ATE
ATT	处理组人群 $E[Y(1) - Y(0) \mid A = 1]$	ATT 和 ATE 一定相等	当处理效应异质时，接受处理的那群人的平均效应可能偏高或偏低
DAG	用节点和箭头表示因果假设	DAG 是从数据里学出来的	DAG 是研究者根据领域知识画的，数据不能告诉你因果方向

核心概念	核心内容	常见误解	为什么错
后门准则	阻断所有 $A \leftarrow L \rightarrow Y$ 的非因果路径	控制的变量越多越好	控制碰撞因子会打开新偏倚
可交换性	$Y(a) \perp\!\!\!\perp A \mid L$	可以用数据检验	未测量混杂的存在与否无法从数据中确认
正值性	每层 L 里处理和对照都存在	样本量大就能保证	临床绝对适应证/禁忌证造成的结构性违反，再多样本也无法修复
一致性 / SUTVA	处理版本唯一，无个体间干扰	二分变量自动满足	同一个"$A = 1$"可能对应不同操作方式，导致效应模糊
G 公式	$E[Y(1)] = E_L[E[Y \mid A = 1, L]]$	G 公式就是回归调整	G 公式是识别等式；回归只是实现它的一种估计方法
dagitty 最小调整集	从 DAG 自动推导该控制哪些变量	调整集是唯一的	同一个 DAG 可能有多个合法调整集，最小调整集是变量最少的那个

第 3 章 回归调整——因果估计的第一刀

内容提要

- 用逐步加变量的逻辑回归观察 RHC 系数的漂移
- 理解回归系数在因果推断语境下的含义与局限
- 建立全书累积对比表的第一行

上一章用 DAG 确定了调整集：要从 RHC 与 180 天死亡率的关联中剥离混杂，需要控制年龄、性别、APACHE 评分、Glasgow 昏迷评分、合并症等一系列协变量。DAG 告诉我们” 该控制谁”，但没有告诉我们” 怎么控制”。回归是研究者最熟悉的控制手段，把协变量放进模型右边，让回归方程帮我们” 条件化”。这一章要做的事情很简单：从一个什么都不放的粗模型开始，逐步往回归方程里加入协变量，观察 RHC 的系数怎样随着调整集的扩大而漂移。漂移本身就是混杂被吸收的直接证据。

但回归能走多远？它的系数到底在估计什么？它在什么条件下才等价于因果效应？这些问题回答清楚之后，我们才能理解后续章节为什么要引入 G 计算、倾向得分和双重稳健估计。本章是全书方法部分的起点。后面的每一种方法，无论是 G 计算、IPW 还是 TMLE，都可以看作是对回归调整某个特定弱点的改进。理解了回归的能力边界，才能理解后续方法存在的理由。

3.1 从粗关联到条件关联

第 1 章的描述性分析已经给出了一个数字：RHC 组的 180 天死亡率高于非 RHC 组。把这个比较放进逻辑回归的框架，就是只含处理变量的粗模型。

逻辑回归做的事情很直观：给定病人的年龄、APACHE 评分等信息，预测他 180 天内死亡的概率。当我们只放入处理变量 RHC 而不放任何协变量时，模型给出的就是最简单的粗关联。

定义 3.1 (粗比值比)

在逻辑回归 $\text{logit}\, P(Y=1) = \beta_0 + \beta_1 A$ 中，$\exp(\beta_1)$ 是处理组相对于对照组的粗比值比，简称 crude OR。它度量的是未经任何协变量调整的处理-结局关联强度。

粗 OR 和因果效应之间隔着整个混杂结构。如果 RHC 的使用与患者病情严重程度相关，而病情严重程度又影响死亡率，那么粗 OR 里既包含 RHC 本身对死亡的影响，也包含” 重症患者更容易接受 RHC，同时更容易死亡” 这条混杂路径的贡献。回归调整的逻辑是：把混杂变量加进模型右边，让回归方程在协变量的每一个取值水平上比较处理组和对照组的结局差异，从而” 堵住” 混杂路径。

这个逻辑听起来合理，但它成立需要一个前提：回归方程的函数形式必须正确。所谓函数形式，指的是模型用什么数学表达来描述协变量与结局之间的关系。最常见的选择是线性项：比如把 APACHE 评分直接放进 $\beta_2 \times \texttt{apache\_score}$，意思是评分每增加 1 分，log-odds 增加固定的 β_2。

但如果 APACHE 评分与死亡率之间的真实关系是非线性的，比如评分从 10 到 20 影响不大，从 20 到 30 却急剧升高，那么一个线性项就无法捕捉这种弯曲。模型右边虽然放了正确的变量，却没有用正确的方式控制它们，混杂仍然会沿着这条没堵住的路径泄漏进来。这种偏差叫模型设定误差，是回归调整的根本弱点，后面会展开。

笔记 回归调整在因果推断中的地位有一段曲折的历史。在 Rubin 潜在结果框架和 Pearl 图模型框架出现之前，流行病学和社会科学的标准做法就是” 往回归里加变量”。Cochran 在 1968 年讨论观察研究的经典论文中已经指出，回归调整的有效性取决于模型设定是否正确，但这个警告长期被忽视。直到 Hernán and Robins [13] 在 *Causal Inference: What If* 中系统阐述了回归与 G 方法的区别，研究者才逐渐意识到：回归系数的因果解读需要比大多

数人以为的更强的假设。Angrist and Pischke [1] 从计量经济学的角度也做了类似的区分，他们把回归称为”坏的控制变量问题”的温床。本章的目的就是把这些假设一一摆出来，让读者在使用回归时知道自己在依赖什么。

3.2 逐步加变量：观察系数漂移

下面我们用 RHC 数据拟合四个嵌套模型。模型 1 是粗模型；模型 2 加入人口学变量 *age* 和 *sex*；模型 3 在此基础上加入疾病严重度指标 *apache_score* 和 *glasgow_coma_score*；模型 4 进一步加入全部合并症和生理指标。每一步我们只关注 RHC 系数的变化。

这种逐步加变量的策略在应用研究中很常见，有时被称为”nested model approach”或”change-in-estimate method”。它的价值在于可视化混杂的吸收过程：如果加入某个变量后处理系数发生明显变化，就说明这个变量在混杂结构中扮演重要角色。但要注意，这种策略本身不具有因果推断效力，它只是一个探索工具，最终的调整集应该由 DAG 决定，而非由系数变化的大小决定。

```
set.seed(2026)
library(tidyverse)
library(broom)

d <- read_csv(here::here("data", "rhc.csv"), show_col_types = FALSE) |>
  mutate(death180_bin = if_else(death180 == "Yes", 1L, 0L),
         sex_bin      = if_else(sex == "Male", 1L, 0L))

# 模型 1：粗模型，只看 RHC 与死亡的边际关联
m1 <- glm(death180_bin ~ rhc, data = d, family = binomial)

# 模型 2：加入人口学——年龄和性别本身是混杂还是精度变量？
m2 <- glm(death180_bin ~ rhc + age + sex_bin,
          data = d, family = binomial)

# 模型 3：加入疾病严重度——APACHE 和 GCS 是最强的混杂源
m3 <- glm(death180_bin ~ rhc + age + sex_bin +
            apache_score + glasgow_coma_score,
          data = d, family = binomial)

# 模型 4：加入合并症和全部生理指标
m4 <- glm(death180_bin ~ rhc + age + sex_bin +
            apache_score + glasgow_coma_score +
            cancer + cardiovascular + congestive_hf + dementia +
            pulmonary + renal + hepatic + blood_pressure +
            heart_rate + respiratory_rate + temperature +
            albumin + creatinine + bilirubin + wbc + hematocrit +
            das_index + dnr_status + medical_insurance + race +
            income + edu + transfer_hx + mi + gi_bleed +
            tumor + immunosupperssion + psychiatric,
          data = d, family = binomial)

# 提取四个模型中 RHC 的 OR 和 95% CI
bind_rows(
  tidy(m1, conf.int = TRUE, exponentiate = TRUE) |>
```

```
    filter(term == "rhc") |> mutate(model = "Model 1"),
  tidy(m2, conf.int = TRUE, exponentiate = TRUE) |>
    filter(term == "rhc") |> mutate(model = "Model 2"),
  tidy(m3, conf.int = TRUE, exponentiate = TRUE) |>
    filter(term == "rhc") |> mutate(model = "Model 3"),
  tidy(m4, conf.int = TRUE, exponentiate = TRUE) |>
    filter(term == "rhc") |> mutate(model = "Model 4")
) |> select(model, estimate, conf.low, conf.high)
```

结果解读

四个模型中 RHC 的 OR 及 95% CI 如下：

模型	OR	95% CI	新增协变量
Model 1	1.35	[1.21, 1.50]	无
Model 2	1.38	[1.24, 1.54]	age, sex
Model 3	1.18	[1.05, 1.33]	apache_score, glasgow_coma_score
Model 4	1.34	[1.18, 1.52]	合并症 + 全部生理指标

这组数字里藏着几条信息。

从模型 1 到模型 2，OR 从 1.35 微升到 1.38。年龄和性别的加入几乎没有改变 RHC 系数，说明这两个变量在 RHC 使用决策中的混杂作用不大。

真正的变化发生在模型 3：加入 APACHE 评分和 GCS 之后，OR 从 1.38 骤降到 1.18，降幅超过 14%。这正是我们预期的结果，APACHE 评分是 ICU 里决定是否插管的核心指标，它同时强烈预测死亡率，是经典的正向混杂变量。控制了它，RHC 与死亡之间有一大块虚假关联被剥离了。

停下来想一想。到这里你应该注意到：加入 APACHE 后 OR 从 1.38 降到 1.18。这说明之前 RHC 和死亡率的关联里，有相当一部分是 APACHE 评分的”锅”。重症患者更容易被插管，重症患者也更容易死亡，这条混杂路径贡献了 OR 中大约 0.20 的虚假成分。控制了 APACHE，这部分虚假关联被剥离，剩下的 1.18 才更接近 RHC 本身的效应。

模型 4 加入合并症之后，OR 又回升到 1.34。这个反弹说明**多放变量不等于更干净**。合并症变量中有些可能是中介变量或碰撞因子，也可能是引入了新的模型设定误差。无论原因是什么，系数的非单调漂移提醒我们：回归调整的结果高度依赖模型中放了什么变量、用了什么函数形式。

图 3.1 直观地展示了四个模型中 RHC 系数的漂移轨迹。

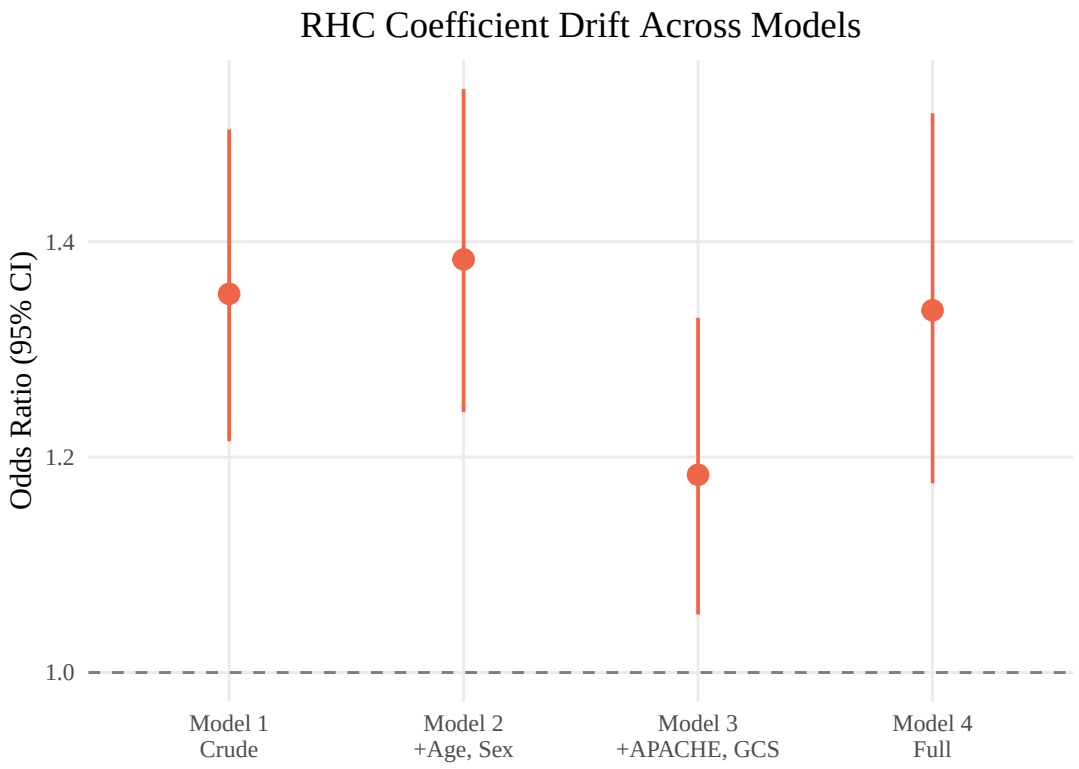

图 3.1: 逐步加入协变量后 RHC 的 OR 变化。虚线为 OR = 1，即无效应。模型 3 加入 APACHE 和 GCS 后 OR 明显下降，模型 4 加入合并症后反弹。

3.3 系数漂移背后的混杂吸收

为什么加入不同协变量会让 RHC 系数往不同方向移动？回归系数的本质是偏回归系数：它度量的是”在其他所有自变量固定的条件下，该自变量每变动一个单位，结局的对数几率变动多少”。理解偏回归系数有一个经典的思维工具：Frisch-Waugh-Lovell 分解。在线性回归中，A 的系数等价于先把 A 和 Y 各自对其他协变量 L 做回归，取残差，再用 A 的残差去预测 Y 的残差。换句话说，偏回归系数衡量的是”剔除了 L 能解释的部分之后，A 中剩余的变异与 Y 中剩余的变异之间的关系”。当我们往模型里加一个新变量，L 的范围扩大，”被剔除的部分”更多，A 和 Y 的残差都在变化，系数自然也会变化。

定义 3.2 (混杂吸收)

如果变量 L 同时影响处理 A 和结局 Y，且 L 未被控制，那么 A 的回归系数会”吸收”一部分 $L \to Y$ 的效应。将 L 加入模型后，A 的系数变化量就是被吸收的混杂的释放。系数变化的方向取决于 L-A 关联和 L-Y 关联的符号组合。

用一个简化的线性模型来建立直觉：假设真实模型是 $Y = \beta_1 A + \gamma L + \varepsilon$，而我们只拟合了 $Y = \tilde{\beta}_1 A + \varepsilon$。遗漏变量偏差公式告诉我们 $\tilde{\beta}_1 = \beta_1 + \gamma \cdot \delta$，其中 δ 是 L 对 A 回归的系数。如果 $\gamma > 0$ 且 $\delta > 0$，偏差 $\gamma\delta > 0$，粗系数偏高；如果两者符号相反，粗系数偏低。这个公式在逻辑回归中没有精确的对应形式，但方向性的直觉完全适用。

命题 3.1 (遗漏变量偏差的方向判断)

在线性回归中，遗漏变量 L 造成的偏差方向由 L-A 关联符号和 L-Y 关联符号的乘积决定。正 × 正或负 × 负产生正向偏差，让处理效应估计偏高；正 × 负或负 × 正产生负向偏差。在非线性模型中这个规律近似成立，可以用来预判加入某个协变量后系数会往哪个方向移动。 [1]

在 RHC 数据中，APACHE 评分与 RHC 使用正相关，因为病情越重的患者越可能被插管；APACHE 评分与死亡率也正相关。两条正关联叠加，意味着 APACHE 评分制造了正向混杂，让粗 OR 偏高。控制 APACHE 之后，这部分虚假的正关联被剥离，RHC 系数下降。

模型 2 中加入年龄和性别后系数微升的现象也有解释。年龄与 RHC 使用之间的关联方向未必是简单的正相关。ICU 里的实际决策逻辑是复杂的：年轻患者如果病情严重，医生更倾向于采取积极干预；但对于高龄患者，即使病情同样严重，医生和家属可能倾向于保守治疗。这种选择偏好意味着年龄与 RHC 使用之间可能存在负向关联。同时年龄与死亡率正相关。按照遗漏变量偏差的方向判断规则，负 × 正产生负向偏差，意味着粗模型中年龄造成的混杂方向是让 OR 偏低。控制年龄后，这部分负向混杂被移除，OR 反而略微上升。从 1.35 升到 1.38 的幅度很小，这与年龄在 RHC 使用决策中的权重较低是一致的。

模型 4 中系数的反弹更值得警惕。OR 从模型 3 的 1.18 跳回到 1.34，几乎回到了粗模型的水平。一种可能是某些合并症变量位于 RHC 和死亡的因果路径上，控制它们会阻断因果中介效应，造成系数偏移。比如如果 RHC 的使用会影响患者后续是否被诊断出某些合并症，那么控制这些合并症就相当于控制了一个中介，打开了从 RHC 到死亡的非因果路径。另一种可能是高维模型中变量之间的共线性改变了回归面的倾斜方式。33 个协变量中很多是相关的，比如 *creatinine* 和 *renal*、*apache_score* 和多个生理指标，它们之间的多重共线性会让系数估计变得不稳定。

回归无法自动判断哪些变量该放、哪些不该放，它只负责在你指定的模型下估计条件关联。这正是 DAG 的价值所在：DAG 告诉你调整集，回归只是执行调整的计算工具。没有 DAG 指导的回归分析，就像没有地图的导航，你可能走了弯路而不自知。

这四个模型的系数漂移模式也给了我们一条实用的经验法则：在逐步加变量的过程中，如果某个变量的加入让处理系数发生大幅变化，这个变量很可能是重要的混杂源，应该被纳入最终的调整集。相反，如果加入后系数几乎不动，这个变量可能与处理变量的关联很弱，或者与结局的关联很弱，混杂贡献不大。当然，这条经验法则只是启发式的，最终的调整决策仍然要回到 DAG。

3.4 回归的因果解读条件

系数漂移让我们看到了混杂被吸收的过程，但一个更根本的问题还没有回答：即使调整集完全正确，回归系数在什么条件下才等于因果效应？很多研究者隐含地假设” 控制了混杂变量之后，回归系数就是因果效应”，但这个说法漏掉了一个关键条件。

定义 3.3 (回归因果解读的充分条件)

在逻辑回归 $\operatorname{logit} P(Y=1 \mid A, L)=\beta_0+\beta_1 A+\boldsymbol{\beta}_L^{\top} L$ 中，$\exp(\beta_1)$ 等于条件因果 OR 的充分条件包括：可交换性 $Y(a) \perp\!\!\!\perp A \mid L$，即调整集 L 足以阻断所有后门路径；正值性 $0<P(A=1 \mid L)<1$，即每一个协变量层都有处理组和对照组；以及模型设定正确，即 $\operatorname{logit} P(Y=1 \mid A, L)$ 确实是 A 和 L 的线性函数。[13] ♣

前两个条件和所有因果推断方法共享，回归独有的额外负担是第三条：模型设定正确。在本章的四个模型中，我们把所有协变量以线性方式放入 logit 链接函数，没有加交互项，没有加非线性变换。如果 APACHE 评分对死亡率的影响存在阈值效应，比如评分超过 30 之后死亡率急剧上升但之前变化平缓，那么一个线性项无法捕捉这种关系。如果 APACHE 与年龄之间存在交互作用，比如高 APACHE 在年轻患者中更致命因为说明病情突然恶化，那么不加交互项就会遗漏这部分效应修饰。线性模型在这些情况下会产生设定偏误，即使调整集完全正确，系数也不等于因果效应。

设定偏误的麻烦在于它是隐性的。模型照样会收敛，照样会给出一个 OR 和一个 p 值，研究者没有任何自动提示告诉他” 你的函数形式错了”。Hosmer-Lemeshow 检验和 AIC 可以给出一些间接信号，但它们检测的是整体拟合优度，对特定系数的偏误不敏感。

这就是回归调整的根本困境：它要求研究者事先猜对函数形式。协变量少的时候，可以靠领域知识加交互

项和样条；协变量多到几十个的时候，交互项的组合空间爆炸，手动设定模型变得不现实。RHC 数据有 49 个变量，光两两交互就有上千项，加上非线性变换，可能的模型空间大到无法手动搜索。

后续章节引入的方法各自从不同角度应对这个困境。G 计算仍然依赖结果模型，但它用标准化而非读系数的方式提取因果效应，对函数形式的敏感性有所缓解。倾向得分方法把建模负担从结局模型转移到处理模型，只要能正确预测” 谁接受了 RHC” 就行。双重稳健估计同时建两个模型，只要其中一个对就能给出一致估计。机器学习方法则直接用数据驱动的方式拟合灵活的函数形式，绕过手动设定的需要。每一步都是在回归的局限上做改进，回归是理解这些改进的起点。

定理 3.1 (雷区)

临床研究中常见一种做法：在论文表格里把多元回归中所有协变量的系数都列出来，并逐一解读为因果效应。这种做法被称为 Table 2 Fallacy。问题在于，每个协变量的系数对应的调整集不同。当你在解读 *age* 的系数时，模型控制了 APACHE 评分，但 APACHE 可能是年龄影响死亡率的中介变量，控制中介会阻断因果路径。同一个模型里，*rhc* 的系数或许需要控制 APACHE，而 *age* 的系数或许不应该控制 APACHE，两者的 DAG 不同，不能共用一个回归方程的结果。安全的做法是每个因果问题对着自己的 DAG 画调整集，分别建模。 [13]

3.5 回归估计的另一个维度：条件 OR vs 边际 OR

假设我们的模型设定完美无瑕，调整集也毫无遗漏，回归给出的 $\exp(\beta_1)$ 就是因果效应了吗？答案取决于你想要的是哪种因果效应。

首先说条件 OR。回归给出的 $\exp(\beta_1)$ 是条件 OR，它回答的问题是：在同一类患者内部，接受 RHC 和不接受 RHC 的人相比，死亡几率比是多少。比如两个 APACHE 评分都是 25、年龄都是 65 岁的患者，一个插了导管一个没插，条件 OR 比较的就是这两个人的死亡几率之比。协变量固定，比较的是” 同类人” 之间的差异。

其次说边际 OR。很多研究者更关心的是整个人群层面的效应，而非同一协变量组合内部的条件比较：如果所有 5735 名 ICU 患者都接受 RHC，和所有人都不接受 RHC，死亡率会差多少？这就是边际效应。边际 OR 把所有患者混在一起算总体的几率比，不区分 APACHE 高低、年龄大小。

在线性模型中，条件效应和边际效应一致，加不加协变量不影响处理系数的期望值，前提是没有混杂。在非线性模型中两者通常不同，这个差距叫作非压缩性，英文称 non-collapsibility。

用一个具体的数字例子来建立直觉。假设有两个 ICU 病房，每个病房各 100 名患者。病房 A 的基线死亡率较高，RHC 组死亡概率 0.80，对照组 0.60，OR = $\frac{0.80/0.20}{0.60/0.40} = 2.67$。病房 B 的基线死亡率较低，RHC 组死亡概率 0.40，对照组 0.20，OR = $\frac{0.40/0.60}{0.20/0.80} = 2.67$。两个病房的条件 OR 完全相同，都是 2.67。现在把两个病房的 200 名患者混在一起计算边际 OR：RHC 组总死亡概率 0.60，对照组总死亡概率 0.40，边际 OR = $\frac{0.60/0.40}{0.40/0.60} = 2.25$。同样的条件效应，混合之后数值变小了。这不是混杂造成的，纯粹是 OR 这个度量在合并子群时的数学性质。

定义 3.4 (非压缩性)

设条件 OR 为 $\mathrm{OR}_L = \exp(\beta_1)$，边际 OR 为在全人群上的比值比。即使不存在任何混杂，OR_L 和边际 OR 也通常不相等。这种现象叫非压缩性，它是 OR 这个度量本身的数学性质，与混杂无关。 [13]

理解非压缩性需要一个类比。想象你在两个不同的城市分别比较喝咖啡和不喝咖啡的人群心率差异。两个城市内部的差异都是 5 bpm，但如果两个城市的基线心率不同，把所有人混在一起算总体差异时，结果可能不是 5 bpm。在线性尺度上这件事不会发生，混合后的差异仍然是 5 bpm；但在 OR 的对数几率尺度上，混合不同子群的结果会改变数值。这意味着即使混杂完全不存在，条件 OR 和边际 OR 也可以不同。

非压缩性带来的实际后果是：我们在模型 4 中读到的 OR = 1.34 不能直接解读为” 如果所有人都接受 RHC，死亡几率比不接受高 34%”。条件 OR 回答的是” 在同一类患者内部，接受 RHC 的人和不接受的人相比如何”，而

边际效应回答的是” 整个人群层面的平均因果效应”。两者之间的差距不是混杂造成的，纯粹是 OR 这个度量的数学性质决定的。风险差和风险比不存在非压缩性问题，这也是为什么越来越多的流行病学家建议在因果推断中报告风险差而非 OR。

如果我们想要边际效应估计，需要一种不同的计算方式。下一章的 G 计算会用标准化的方式把条件预测平均回全人群，直接在概率尺度上计算边际风险差，从而绕过非压缩性问题。

3.6 回归调整的局限

回归调整是因果估计的起点，但它有几条绕不过去的限制。模型设定敏感性已经讨论过了，这里再补充几个同样重要的问题。

回归没有显式地构造反事实。它给出的是一个系数，你需要相信模型的函数形式才能把系数翻译成因果效应。相比之下，G 计算会为每个个体分别预测” 如果接受 RHC” 和” 如果不接受 RHC” 两个反事实结局，然后取差值。这种显式构造让反事实推理变得透明，也让模型诊断更直观。回归用户看到的只是一个数字 1.34，而 G 计算用户看到的是 5735 个个体各自的反事实预测值，每一个都可以检查是否合理。

回归对正值性违反没有内置的警报机制。如果某一类患者几乎全部接受了 RHC，回归仍然会给出一个系数，但这个系数在该子群中几乎完全靠外推。想象一个 APACHE 评分极高的层，里面 98% 的患者都接受了 RHC，只有 2% 没接受。回归在这个层里比较的实际上是 98 个处理样本和 2 个对照样本的结局差异，统计功效极低，估计几乎完全依赖模型的线性外推假设。倾向得分方法至少可以通过检查得分分布来发现这种 overlap 缺失，回归则把这些信息藏在了系数矩阵的深处。

回归的第三个局限容易被忽略：它假设处理效应在所有协变量层上是同质的。模型 $\text{logit}\, P(Y=1) = \beta_0 + \beta_1 A + \boldsymbol{\beta}_L^\top L$ 只给 A 一个系数 β_1，意味着 RHC 对年轻人和老年人、轻症和重症的效应都一样。如果真实效应是异质的，比如 RHC 对重症有益但对轻症有害，那么一个单一的 β_1 会把这些异质性平均掉，得到一个可能误导政策的” 平均” 估计。第 9 章的因果森林会专门处理效应异质性问题。

定理 3.2 (雷区)

” 变量放得越多越安全” 是一个常见的错觉。在因果推断中，控制中介变量、碰撞因子或处理后变量都会引入偏差。模型 4 中合并症变量的加入导致 OR 从 1.18 反弹到 1.34，其中一个可能的原因就是某些合并症变量位于因果路径上。判断一个变量该不该放进模型，唯一可靠的依据是 DAG，而非统计显著性或” 放了总比没放好” 的直觉。Angrist and Pischke [1] 称这类问题为” 坏的控制变量”，意思是看起来在控制混杂，实际上在制造偏差。 ♡

这些局限加在一起，意味着回归调整的结论必须谨慎解读。但这并不意味着回归没用。对于一个探索性分析，逐步加变量的回归能快速揭示混杂结构，帮助研究者判断哪些变量是关键混杂源。本章的四模型对比就是一个教学案例：APACHE 评分是最重要的混杂变量，加不加它决定了 OR 从 1.38 变到 1.18 还是留在原地。这种信息对后续建模至关重要。回归是因果推断工具箱里的第一把刀，它的价值在于快速、直观、门槛低，适合在正式建模之前探索数据的混杂结构。很多发表在顶级期刊上的因果推断论文，仍然会在附表中报告一个回归调整的结果作为基准线，然后用更精细的方法作为主分析。回归结果和主分析的比较本身就是一种敏感性检查：如果两者差距很大，说明模型设定或非压缩性在起作用，需要进一步排查。

✍ **练习 3.1** 在模型 3 的基础上，尝试给 *apache_score* 加一个二次项 `I(apache_score^2)`，观察 RHC 的 OR 是否发生变化。如果二次项显著且 OR 变化超过 5%，说明线性假设对这个变量可能不成立。用 AIC 比较加入二次项前后两个模型的拟合优度。

解

```
# 在模型 3 基础上加入 APACHE 的二次项
m3b <- glm(death180_bin ~ rhc + age + sex_bin +
```

```
            apache_score + I(apache_score^2) +
            glasgow_coma_score,
          data = d, family = binomial)

# 比较 RHC 的 OR
cat("Model 3  OR:", exp(coef(m3)["rhc"]), "\n")
cat("Model 3b OR:", exp(coef(m3b)["rhc"]), "\n")
cat("AIC  M3:", AIC(m3), " M3b:", AIC(m3b), "\n")
```

如果二次项显著且 AIC 下降，说明 APACHE 评分对死亡率的影响确实存在非线性成分。RHC 系数的变化幅度就是线性设定偏误对因果估计的影响大小。这类敏感性检查在正式报告回归结果之前应当常规进行。

方法卡片：回归调整

估计目标：条件 OR，$\exp(\beta_1)$，在协变量 L 固定条件下处理 A 对结局 Y 的几率比效应。
核心假设：可交换性 $Y(a) \perp\!\!\!\perp A \mid L$；正值性 $0 < P(A = 1 \mid L) < 1$；模型设定正确。
R 实现：`glm()` + `family = binomial`，用 broom 的 `tidy()` 提取系数和置信区间。
适用场景：协变量维度低、研究者对函数形式有信心、探索性分析中用于快速定位关键混杂变量。
失效场景：高维协变量下函数形式无法手动设定；处理效应异质性强但只有单一系数；正值性违反时无内置诊断。

3.7 累积对比表

从本章开始，每章末尾更新一张方法对比表。到全书最后一章，这张表会汇总所有方法对同一个因果问题的回答。

表 3.1: 方法演进对比表，截至第 3 章

方法	ATE 估计	95% CI	核心假设
回归调整	OR = 1.34	[1.18, 1.52]	模型设定正确 + 可交换性 + 正值性

这里报告的是全调整模型，即模型 4 的结果。OR = 1.34 意味着在控制了 33 个协变量之后，接受 RHC 的患者 180 天死亡几率仍然比未接受者高 34%，95% CI 不包含 1，$p < 0.001$。这个结论与 Connors, Speroff, Dawson, et al. [9] 原始论文的发现方向一致。但我们现在知道，这个数字背后依赖于函数形式正确、调整集完整、效应同质等一系列假设，其中任何一条违反都会让解读偏离真实的因果效应。

回归给了我们一个重要的基准。后续每一章的方法都会给出自己的估计，最终我们会看这些数字是否收敛。如果不同方法给出相似的答案，我们对因果结论的信心就会增强；如果答案分歧很大，就说明某些方法的假设可能被违反了，需要进一步诊断。

下一章的 G 计算会做一件与回归截然不同的事：为每个患者分别预测两个反事实结局，然后用标准化把个体层面的预测汇总为全人群的边际因果效应。回归读的是一个系数，G 计算构造的是整个反事实人群。从”读系数”到”构造反事实”，是因果推断方法论的一次跨越。

本章知识地图

表 3.2: 第 3 章核心概念与常见误解

核心概念	核心内容	常见误解	为什么错
粗 OR	未调整任何协变量的处理-结局关联	粗 OR 就是因果效应	混杂未被控制，关联 ≠ 因果
系数漂移	加入协变量后处理系数发生变化，反映混杂被吸收	系数变化一定单调递减	变化方向取决于混杂方向，负向混杂会让系数上升
模型设定正确	回归的函数形式必须与真实数据生成过程一致	放了正确的变量就够了	变量正确但函数形式错误仍然会产生设定偏误
Table 2 Fallacy	同一回归中所有系数共用一个调整集	每个系数都可以解读为因果效应	不同因果问题对应不同的 DAG 和调整集
非压缩性	条件 OR 与边际 OR 在非线性模型中不相等	调整后 OR 下降一定是因为混杂被控制	OR 的数学性质使得边际化本身就会改变数值
回归的局限	对函数形式敏感，无显式反事实，无正值性诊断	回归能解决所有混杂问题	高维、非线性场景下模型设定几乎不可能全对

第 4 章 G 计算——构造反事实人群

内容提要

- 理解 G 公式的识别结果及其与回归系数的本质区别
- 掌握 G 计算的三步算法：建模、预测反事实、边际化
- 在 RHC 数据上实现 G 计算并用 Bootstrap 构造置信区间
- 更新累积对比表，比较回归 OR 与 G 计算风险差

上一章用逐步加变量的逻辑回归估计了 RHC 的效应，得到全调整 OR = 1.34，95% CI [1.18, 1.52]。这个数字是从回归方程里” 读” 出来的：拟合模型，提取 *rhc* 的系数，做指数变换。整个过程依赖一个隐含的操作，我们信任模型的函数形式，把系数直接翻译成因果效应。

但回归系数回答的问题和因果推断想要回答的问题之间有一道裂缝。

先看回归系数回答的是什么。条件 OR = 1.34 的意思是：在年龄、APACHE 评分等 33 个协变量都固定的某一层里，接受 RHC 的患者相对于不接受 RHC 的患者，死亡几率比为 1.34。它描述的是” 层内” 的比较。

再看因果推断想回答什么。我们真正关心的问题是：如果 5735 名患者全部接受 RHC，和全部不接受 RHC 相比，180 天死亡率会差多少？这是一个全人群层面的问题，比较的是两个反事实世界的平均结局，得到的量叫边际风险差。

这两个量为什么不一样？上一章已经讲过，逻辑回归有非压缩性：即使每一层内的 OR 都是 1.34，把所有层加权汇总后得到的边际 OR 通常不等于 1.34，更不能直接换算成风险差。条件 OR 和边际风险差在数学上不等价，不能简单地从一个推导出另一个。

G 计算走了一条完全不同的路。它不去读系数，而是用模型为每一个患者分别预测” 如果接受 RHC 会怎样” 和” 如果不接受 RHC 会怎样”，构造出两个完整的反事实人群，然后在全人群上取平均，直接得到边际因果效应。从” 读系数” 到” 构造反事实人群”，是因果推断方法论的一次重要跨越。

4.1 从标准化到 G 公式

在正式引入 G 公式之前，需要理解它的直觉来源：流行病学中的标准化率。

先用 RHC 数据看一个具体问题。处理组（接受 RHC 的患者）平均 APACHE 评分更高，意味着病情更重；对照组（不接受 RHC 的患者）平均 APACHE 评分更低，病情相对更轻。病情重的人本来死亡率就高，所以直接拿处理组的总死亡率和对照组的总死亡率比较，偏差是必然的。

怎么办？假设 APACHE 评分只有低、中、高三档。低档占全人群 40%，中档占 35%，高档占 25%。我们可以在每一档内分别算处理组和对照组的死亡率，消除病情严重程度的影响，然后按全人群的三档比例加权平均。这就是标准化。

标准化率的思路很简单：如果两组人群的混杂因素分布不同，直接比较总率会产生偏倚，那就把两组的结果都” 投射” 到同一个标准人群的分布上，消除混杂因素构成差异的影响。传统的直接标准化就是这种手工分层操作：按混杂变量分层，在每一层内算各自的率，再用标准人群在各层的比例作为权重加权平均。

这个手工操作在混杂变量只有一两个的时候可行。一旦混杂变量增加到十个以上，层数呈指数增长，每一层里的样本量很快不够用。RHC 数据有 33 个协变量，如果每个只分两层，就有 2^{33} 超过 80 亿个组合，远远超过 5735 的样本量。传统标准化在高维场景下彻底失效。

笔记 James Robins 在 1986 年提出 G 公式，G 代表 Generalized，即广义标准化。Robins 当时面对的实际问题是石棉暴露的流行病学研究，工人在不同时间点接受不同水平的暴露，而暴露水平又受到之前健康状况的影响。传

统回归在这种时变混杂场景下会失效，因为控制了中间时间点的健康状况，既消除了混杂，也阻断了因果路径。Robins 的 G 公式把标准化的思想从静态处理推广到动态处理序列，从手工分层推广到模型预测，从低维推广到高维。本章讨论的是 G 公式在最简单的点处理场景下的应用，但它的理论框架从一开始就是为更复杂的纵向问题设计的。 [20]

G 计算的核心思想是用回归模型替代手工分层。不再把协变量空间切成离散的层，而是用一个回归模型来估计每一层的条件结局概率，然后按全人群协变量的经验分布做加权平均。模型承担了” 在每一层内计算率” 的任务，经验分布承担了” 用标准人群权重加权” 的任务。两者结合，就是 G 计算。

在写出公式之前，先用一个具体的数字例子说明 G 计算在做什么。假设协变量只有 APACHE 评分，分成低、中、高三档。低档占全人群 40%，处理组在低档内的死亡率为 25%；中档占 35%，处理组死亡率 45%；高档占 25%，处理组死亡率 70%。如果全人群都接受 RHC，预期死亡率是多少？按全人群的三档比例加权平均：$0.40 \times 0.25 + 0.35 \times 0.45 + 0.25 \times 0.70 = 0.433$。也就是说，全人群层面的 $E[Y(1)] = 43.3\%$。

这个计算做了两件事：第一，在每一档内算出处理组的死亡率（相当于” 条件期望”）；第二，按全人群的档位比例加权汇总（相当于” 边际化”）。G 公式把这两步写成了一般性的数学表达。

定义 4.1 (G 公式)

在点处理场景下，G 公式的识别结果为

$$E[Y(a)] = E_L\big[E[Y \mid A = a, L]\big] = \sum_l E[Y \mid A = a, L = l] \cdot P(L = l),$$

其中 $Y(a)$ 是设定处理为 a 时的潜在结局，L 是协变量集合。该等式在可交换性 $Y(a) \perp\!\!\!\perp A \mid L$、正值性 $P(A = a \mid L) > 0$、一致性 $Y = Y(A)$ 三个假设下成立。 [20] ♣

对照上面的数字例子来读这个公式。内层期望 $E[Y \mid A = a, L = l]$ 对应” 在某一档内，处理设定为 a 时的死亡率”，比如低档内处理组的 25%。外层期望 $E_L[\cdot]$ 对应” 按全人群的档位比例加权平均”，比如用 40%、35%、25% 做加权。整个操作的效果是：先在 L 的每一个值上算出处理为 a 时的结局均值，再按全人群中 L 实际的分布把这些均值汇总成一个数字。这个数字就是全人群层面的潜在结局期望 $E[Y(a)]$。

可交换性在这里的作用是保证内层期望的因果解读。在同一个 L 值下，如果处理分配与潜在结局独立，那么我们观察到的处理组结局均值 $E[Y \mid A = a, L]$ 就等于该 L 层全体人群在设定处理为 a 时的潜在结局均值 $E[Y(a) \mid L]$。没有这条假设，我们只是在算条件关联，不是在估计反事实。

4.2 G 计算的三步算法

G 公式给出了识别结果，G 计算是把这个识别结果变成可操作算法的过程。整个算法可以浓缩为三步：建模、预测反事实、边际化。

定义 4.2 (G 计算算法)

G 计算估计 $E[Y(a)]$ 的算法如下。

建模：用全样本拟合结局模型 $\hat{E}[Y \mid A, L]$。在 RHC 数据上，这一步就是用 5735 名患者的真实数据拟合逻辑回归，把 180 天死亡对 RHC 和 33 个协变量的关系学到模型里。

预测反事实：对全样本中每一个个体 i，保持其协变量 L_i 不变，将处理变量设定为 $A = a$，用模型预测 $\hat{Y}_i(a) = \hat{E}[Y \mid A = a, L_i]$。在 RHC 数据上，这一步相当于复制两份完整的 5735 人数据集，一份把所有人的 *rhc* 改成 1，另一份改成 0，其他变量（年龄、APACHE 评分等）保持每个人的真实值不动，然后用模型分别预测每个人的死亡概率。

边际化：对所有个体的预测值取算术平均，$\widehat{E[Y(a)]} = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i(a)$。在 RHC 数据上，这一步就是对 5735 个预测概率取均值，得到全人群层面的预期死亡率。

平均处理效应的风险差估计为 $\widehat{\mathrm{RD}} = \widehat{E[Y(1)]} - \widehat{E[Y(0)]}$。 ♣

理解这三步的关键在于第二步。预测反事实的操作是" 把每个人的处理变量强制改写，但协变量保持原样"。这和回归的" 读系数" 有本质区别。回归读的是一个全局系数，它把所有个体的效应压缩成一个数字，并且这个数字的含义受模型函数形式的约束。G 计算做的是为每个人分别算两个预测值，然后在个体层面取差，最后汇总。即使同一个逻辑回归模型，G 计算提取因果效应的方式也和读系数不同，因为它绕过了非压缩性问题，直接在概率尺度上操作。

用一个类比来理解：回归读系数像是看一张地图上标注的" 平均海拔差"，G 计算像是亲自走遍两条路线，在每个位置测量海拔，然后算两条路线的平均海拔之差。前者是一个全局摘要，后者是逐点测量再汇总。如果地形是均匀的，两者结果一致；如果地形复杂，逐点测量的方法更可靠。

第三步中用算术平均做边际化，等价于用经验分布 $\hat{P}(L = l) = 1/n$ 作为权重。这意味着 G 计算标准化到的" 标准人群" 就是样本本身。每个人贡献相同的权重，不需要手动指定标准人群，也不需要分层。这就是为什么 G 计算能处理高维混杂：模型负责" 在每一层内估计率"，经验分布负责" 加权"，维度灾难被模型吸收了。

4.3 G 计算与回归系数的本质区别

上一章的回归和本章的 G 计算用的是同一个逻辑回归模型，区别在于从模型中提取因果效应的方式。回归提取的是 $\exp(\hat{\beta}_{\mathrm{rhc}})$，一个条件 OR；G 计算提取的是 $\widehat{E[Y(1)]} - \widehat{E[Y(0)]}$，一个边际风险差。两者的差异来自三个层面。

度量尺度不同。OR 是比值比，定义在对数几率尺度上；RD 是概率差，定义在概率尺度上。即使因果效应的方向一致，数值上不可直接比较。报告 RD 的优势是临床解读直观——" 接受 RHC 比不接受，180 天死亡概率高 6 个百分点"，任何人都能理解这句话的含义。OR = 1.34 的解读则需要对几率比的概念有一定理解。

条件 vs 边际。回归的 OR 是在协变量固定的条件下度量效应，G 计算的 RD 是在全人群边际上度量效应。由于逻辑回归的非压缩性，即使没有任何混杂，条件 OR 和边际 OR 也不相等。G 计算通过标准化操作直接在概率尺度上给出边际估计，回避了非压缩性的困扰。

对函数形式的敏感性不同。回归系数等于因果效应，要求模型的线性项正确捕捉了所有变量的关系；G 计算的估计也依赖模型，但它要求的是模型的条件预测值 $E[Y \mid A, L]$ 整体合理，而非某一个系数恰好等于因果效应。如果 APACHE 评分与死亡率的关系存在非线性但模型用了线性项，回归系数会系统性偏倚，而 G 计算的预测值只要在" 平均" 的意义上足够准确，边际估计仍然可能合理。这种差异是微妙的，G 计算的模型敏感性并没有消失，只是表现形式不同。

定理 4.1 (雷区)

G 计算和回归调整可以使用同一个回归模型，但它们提取因果效应的方式不同，结果也通常不同。一个常见的误解是"G 计算就是换了个方式报告回归结果"。实际上，G 计算的边际标准化操作让它估计的是一个不同的因果量。在线性模型中两者巧合地一致，因为线性模型没有非压缩性；在逻辑回归或其他非线性模型中，两者的数值和含义都不同。混淆两者会导致对因果估计的错误解读。 ♡

4.4 RHC 数据上的 G 计算实现

本章继续使用第 1 章介绍的 RHC 数据集，$n = 5735$。结局模型与上一章模型 4 完全相同：用逻辑回归拟合 180 天死亡对 RHC 和 33 个协变量的关系。区别在于，上一章从这个模型里读 *rhc* 的系数，本章用这个模型为每个患者预测两个反事实结局。

下面的代码实现 G 计算的三步算法。建模阶段用全样本拟合逻辑回归；预测阶段构造两个反事实数据集，一个把所有人的 *rhc* 设为 1，另一个设为 0，其他协变量保持原值；边际化阶段对预测的死亡概率取均值。

```
set.seed(2026)
library(tidyverse)

d <- read_csv(here::here("data", "rhc.csv"), show_col_types = FALSE) |>
  mutate(death180_bin = if_else(death180 == "Yes", 1L, 0L),
         sex_bin      = if_else(sex == "Male", 1L, 0L))

# 建模：与第 3 章模型 4 相同的结局模型
# G 计算的全部"因果推断负担"都压在这个模型上
outcome_mod <- glm(death180_bin ~ rhc + age + sex_bin +
  apache_score + glasgow_coma_score +
  cancer + cardiovascular + congestive_hf + dementia +
  pulmonary + renal + hepatic + blood_pressure +
  heart_rate + respiratory_rate + temperature +
  albumin + creatinine + bilirubin + wbc + hematocrit +
  das_index + dnr_status + medical_insurance + race +
  income + edu + transfer_hx + mi + gi_bleed +
  tumor + immunosupperssion + psychiatric,
  data = d, family = binomial)

# 预测反事实：构造两个"平行世界"的数据集
# 关键操作——只改处理变量，协变量保持每个人的真实值
d1 <- d |> mutate(rhc = 1L)   # 所有人接受 RHC
d0 <- d |> mutate(rhc = 0L)   # 所有人不接受 RHC

Y1 <- predict(outcome_mod, newdata = d1, type = "response")
Y0 <- predict(outcome_mod, newdata = d0, type = "response")

# 边际化：对全人群取算术平均
EY1 <- mean(Y1)
EY0 <- mean(Y0)
RD  <- EY1 - EY0

cat("E[Y(1)] =", round(EY1, 4), "\n")
cat("E[Y(0)] =", round(EY0, 4), "\n")
cat("Risk Difference =", round(RD, 4), "\n")
```

结果解读

G 计算给出的结果是：如果全部 5735 名患者都接受 RHC，预计 180 天死亡率为 53.0%；如果全部不接受 RHC，预计死亡率为 47.0%。边际风险差 RD = 0.060，即接受 RHC 使 180 天死亡概率升高约 6 个百分点。这个结果与第 3 章回归的 OR = 1.34 方向一致，都指向 RHC 增加死亡风险。但两者报告的是不同的因果量：回归给的是条件 OR，G 计算给的是边际 RD。RD = 0.060 的临床含义更直接——每 100 名接受 RHC 的 ICU 患者中，大约多 6 人在 180 天内死亡。

图 4.1 展示了 5735 名患者在两个反事实场景下预测死亡概率的分布。两条分布的重叠很大，说明大部分患

者在两种处理下的预测结局差距不大，RD = 0.060 是全人群平均效应，个体层面的差异被平均掉了。

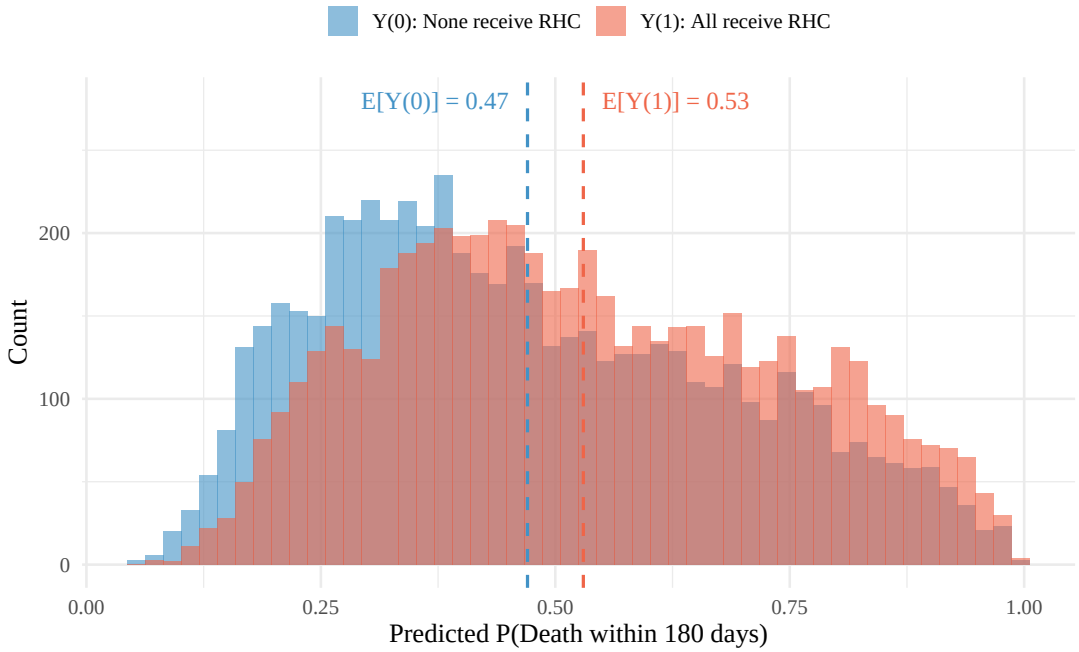

图 4.1: 5735 名患者在两个反事实场景下的预测死亡概率分布。红色为全部接受 RHC，蓝色为全部不接受。虚线标注各自的均值。两条分布整体右移约 0.06，即风险差。

4.5 Bootstrap 置信区间

G 计算的点估计有了，但没有置信区间的点估计是不完整的。G 计算的标准误没有简洁的解析公式，因为它涉及模型拟合和非线性预测两个步骤的不确定性叠加。实践中最常用的方法是非参数 Bootstrap：从原始数据中有放回地抽样，在每个 Bootstrap 样本上重复 G 计算的全部流程，用 Bootstrap 分布的百分位数构造置信区间。

Bootstrap 的直觉是” 用数据模拟采样变异”。真实世界中我们只有一份数据，没法重复抽样；Bootstrap 用有放回抽样来近似” 如果我们能重复收集数据，估计值会怎么波动”。每一次 Bootstrap 抽样都会产生一个略有不同的数据集，从而产生一个略有不同的 G 计算估计。1000 次 Bootstrap 给出 1000 个估计值，这些值的分布就近似了真实的抽样分布。

```
# Bootstrap 置信区间：重复整个 G 计算流程 1000 次
# 每次有放回抽样 → 重新拟合模型 → 重新预测 → 重新取均值
n_boot <- 1000
boot_rd <- numeric(n_boot)

for (i in seq_len(n_boot)) {
  idx <- sample(nrow(d), replace = TRUE)
  bd  <- d[idx, ]

  mod <- glm(death180_bin ~ rhc + age + sex_bin +
    apache_score + glasgow_coma_score +
    cancer + cardiovascular + congestive_hf + dementia +
    pulmonary + renal + hepatic + blood_pressure +
    heart_rate + respiratory_rate + temperature +
    albumin + creatinine + bilirubin + wbc + hematocrit +
```

```
    das_index + dnr_status + medical_insurance + race +
    income + edu + transfer_hx + mi + gi_bleed +
    tumor + immunosupperssion + psychiatric,
    data = bd, family = binomial)

  bd1 <- bd |> mutate(rhc = 1L)
  bd0 <- bd |> mutate(rhc = 0L)
  boot_rd[i] <- mean(predict(mod, bd1, "response")) -
                mean(predict(mod, bd0, "response"))
}

ci <- quantile(boot_rd, c(0.025, 0.975))
cat("Bootstrap 95% CI:", round(ci, 4), "\n")
```

结果解读

1000 次 Bootstrap 的结果：RD 的 95% 百分位数置信区间为 [0.035, 0.087]。区间不包含 0，$p < 0.05$，与回归调整的结论方向一致。Bootstrap 标准误约为 0.014。

图 4.2 展示了 Bootstrap 风险差的分布。分布近似正态，中心在 0.060 附近，与点估计 0.060 吻合。置信区间的下界 0.035 意味着即使在最保守的情境下，RHC 仍然与至少 3.5 个百分点的死亡率增加相关。

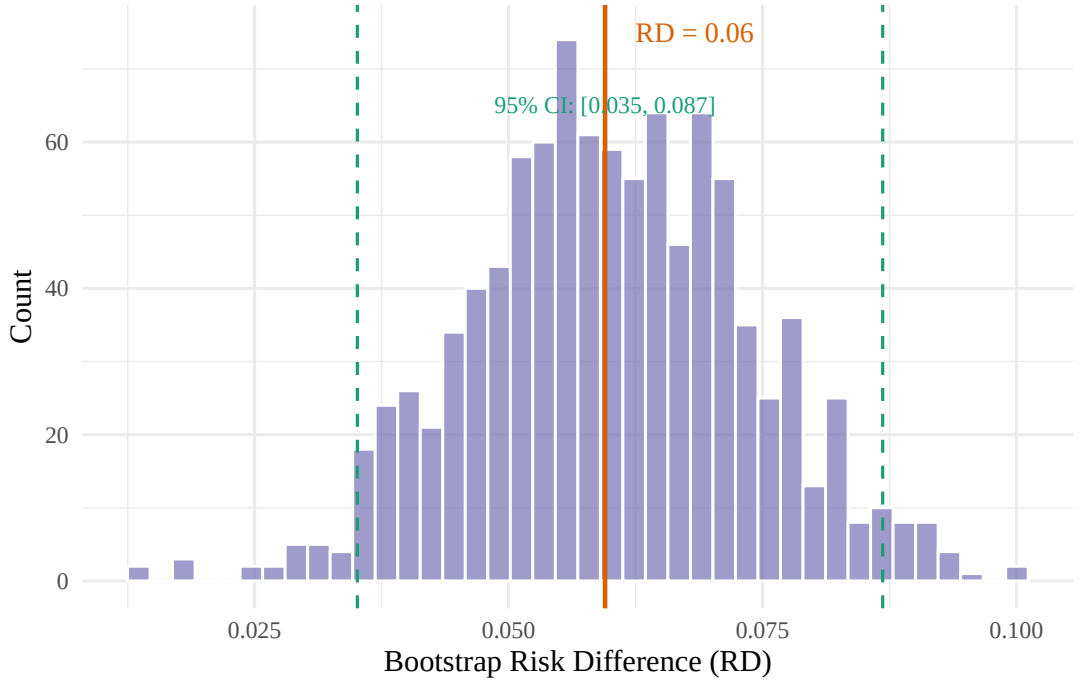

图 4.2: 1000 次 Bootstrap 的风险差分布。橙色实线为点估计 RD = 0.060，绿色虚线为 95% 百分位数置信区间的上下界。

4.6 G 计算的局限

G 计算把因果推断的全部赌注压在一个模型上：结局模型。如果结局模型的设定正确，G 计算给出的是一致的边际因果效应估计；如果结局模型设定错误，G 计算的估计会系统性偏倚，而且没有内置的纠错机制。

这与后续章节将介绍的方法形成对比。倾向得分方法把建模负担转移到处理模型上，只需要正确预测” 谁接

受了 RHC”，不需要建结局模型。双重稳健方法同时建两个模型，只要其中一个正确就能给出一致估计。G 计算是” 单保险绳” 的方法，它的稳健性完全取决于那一条绳子的质量。

在 RHC 数据中，我们用的是一个包含 33 个协变量线性项的逻辑回归。如果真实的结局生成过程包含交互项或非线性关系，这个模型就是错的。RD = 0.060 可能偏高也可能偏低，取决于模型误设的方向和程度。后续章节的 TMLE 和 DML 会用机器学习替代参数模型来拟合结局和处理模型，减轻对线性函数形式的依赖。

定理 4.2 (雷区)

G 计算对结局模型的依赖是单向的、不可对冲的。如果结局模型错了，没有第二个模型可以补救。在实践中检测模型误设的方法包括：检查残差分布、比较不同函数形式下的 G 计算结果、在关键协变量上加非线性项和交互项后观察 RD 是否发生实质变化。如果加入 APACHE 评分的二次项后 RD 从 0.060 变成 0.070，说明线性假设对估计有影响，应该考虑更灵活的模型。这种敏感性检查应当作为 G 计算的标准报告内容。 ♡

G 计算还有一个容易被忽略的假设：它标准化到的是样本的协变量分布。如果样本不代表目标人群，比如 RHC 数据来自 5 家教学医院而研究者想推广到所有 ICU，那么 G 计算的估计只是样本内的 ATE，不是目标人群的 ATE。要推广，需要额外的可迁移性假设，或者在边际化步骤中使用目标人群的协变量分布而非样本的经验分布。

方法卡片：G 计算

估计目标：边际因果风险差 RD $= E[Y(1)] - E[Y(0)]$，直接在概率尺度上度量全人群层面的平均处理效应。
核心假设：可交换性 $Y(a) \perp\!\!\!\perp A \mid L$；正值性 $P(A = a \mid L) > 0$；一致性 $Y = Y(A)$；结局模型正确设定。
算法：建模 $\hat{E}[Y \mid A, L] \to$ 预测 $\hat{Y}_i(1)$ 和 $\hat{Y}_i(0) \to$ 取均值差。
R 实现：`glm()` 拟合结局模型 + `predict()` 构造反事实预测 + `mean()` 边际化。置信区间用 Bootstrap。
适用场景：需要边际效应估计、希望绕过 OR 的非压缩性、协变量维度中等且研究者对结局模型有信心。
失效场景：结局模型严重误设；高维场景下线性模型拟合不足；需要处理模型诊断但 G 计算不提供。

✍ **练习 4.1** 在 G 计算的结局模型中加入 *rhc* 与 *apache_score* 的交互项 `rhc:apache_score`，重新执行 G 计算。比较加入交互项前后的 RD 估计值。如果差异超过 10%，说明处理效应可能在不同 APACHE 水平上异质，线性主效应模型不足以捕捉这种异质性。

解

```r
# 在结局模型中加入 RHC 与 APACHE 的交互项
# 如果交互项显著，说明 RHC 的效应因病情严重程度而异
outcome_mod2 <- glm(death180_bin ~ rhc * apache_score +
  age + sex_bin + glasgow_coma_score +
  cancer + cardiovascular + congestive_hf + dementia +
  pulmonary + renal + hepatic + blood_pressure +
  heart_rate + respiratory_rate + temperature +
  albumin + creatinine + bilirubin + wbc + hematocrit +
  das_index + dnr_status + medical_insurance + race +
  income + edu + transfer_hx + mi + gi_bleed +
  tumor + immunosupperssion + psychiatric,
  data = d, family = binomial)

d1 <- d |> mutate(rhc = 1L)
d0 <- d |> mutate(rhc = 0L)
RD2 <- mean(predict(outcome_mod2, d1, "response")) -
       mean(predict(outcome_mod2, d0, "response"))
```

```

cat("RD (no interaction):", round(0.0595, 4), "\n")
cat("RD (with interaction):", round(RD2, 4), "\n")
cat("Change:", round((RD2 - 0.0595) / 0.0595 * 100, 1), "%\n")
```

如果加入交互项后 RD 变化幅度较小，说明 RHC 效应在不同 APACHE 水平上的异质性有限，主效应模型的边际估计足够稳健。如果变化超过 10%，应该检查交互项的系数和显著性，并考虑是否需要在后续分析中保留交互项。第 9 章的因果森林会用更系统的方式探索效应异质性。

4.7 累积对比表

表 4.1: 方法演进对比表，截至第 4 章

方法	ATE 估计	95% CI	核心假设
回归调整	OR = 1.34	[1.18, 1.52]	模型设定正确 + 可交换性 + 正值性
G 计算	RD = 0.060	[0.035, 0.087]	结果模型正确 + 可交换性 + 正值性

两种方法都指向同一个方向：RHC 与更高的 180 天死亡率相关。回归以 OR 的形式报告，置信区间不含 1；G 计算以 RD 的形式报告，置信区间不含 0。两者在核心假设上共享可交换性和正值性，区别在于对模型的依赖方式。回归要求函数形式正确才能让系数等于因果效应，G 计算要求结局模型的预测值整体合理才能让标准化估计正确。

目前为止，两种方法的”保险绳”都只有一条：结局模型。如果结局模型错了，两者的估计都会偏。下一章引入倾向得分方法，它完全放弃结局模型，转而建模”谁更可能接受 RHC”。把建模负担从结局端转移到处理端，是因果推断方法论的另一次思路转换。再后面的双重稳健方法会同时使用两个模型，真正实现”两根保险绳”的安全网。

本章知识地图

表 4.2: 第 4 章核心概念与常见误解

核心概念	核心内容	常见误解	为什么错
G 公式	将潜在结局的边际期望分解为条件期望对协变量分布的积分	G 公式是一种新的回归方法	G 公式是识别结果，回归只是实现它的一种建模工具
G 计算三步	建模 → 预测反事实 → 边际化	G 计算就是换了个方式报告回归系数	G 计算提取的是边际 RD，与条件 OR 在含义和数值上都不同
反事实构造	对每个个体保持协变量不变，只改处理变量的值	预测反事实只用处理组或对照组的数据	必须用全样本拟合模型，预测时也对全样本做，否则标准化到的人群不对

核心概念	核心内容	常见误解	为什么错
边际 vs 条件效应	边际效应是全人群层面的平均, 条件效应是协变量固定条件下的效应	控制混杂后 OR 下降就是混杂被消除	非压缩性让条件 OR 和边际 OR 在非线性模型中天然不等
Bootstrap CI	用有放回抽样模拟采样变异, 构造百分位数 CI	Bootstrap 只是用来对付小样本的	Bootstrap 对任何复杂估计量都适用, G 计算没有解析标准误, Bootstrap 是标准做法
结局模型依赖	G 计算的因果效力完全取决于结局模型是否正确	用了 G 计算就比回归更可靠	G 计算和回归共享同一个模型假设, 只是效应提取方式不同

第 5 章 倾向得分：匹配、加权与平衡诊断

内容提要

- 理解倾向得分的定义及其降维原理
- 用逻辑回归估计 RHC 数据的倾向得分并检查重叠
- 掌握倾向得分匹配、逆概率加权和重叠权重三种使用方式
- 用标准化均值差和 Love plot 进行平衡诊断
- 更新累积对比表，新增三行倾向得分方法的估计

上一章用 G 计算为每个患者构造了两个反事实结局，然后取全人群平均得到边际风险差。G 计算的核心依赖是结局模型：如果 $E[Y \mid A, L]$ 设定错了，反事实预测就偏了，因果效应估计也随之偏离。问题是，结局模型在高维协变量下几乎不可能完全正确，交互项、非线性项该不该加、加几个，全凭研究者的判断。

有没有一条路径，不去建结局模型，而是转向另一个更容易建模的对象？Rosenbaum 和 Rubin 在 1983 年给出了答案：与其模型化” 结局和协变量的关系”，不如模型化” 谁接受了处理”。这个思路催生了倾向得分方法。倾向得分建模的目标是处理分配机制，而非结局本身。只要能正确预测” 谁更可能接受 RHC”，就可以通过匹配或加权的方式让处理组和对照组在协变量上变得可比，从而在不依赖结局模型的前提下估计因果效应。

5.1 倾向得分的定义与降维定理

回归调整和 G 计算面临的共同困境是高维协变量。RHC 数据有 38 个协变量，如果想在协变量空间里直接匹配处理组和对照组的个体，每个协变量都要找到数值接近的配对，这在 38 维空间中几乎不可能实现。协变量越多，能找到完美配对的概率越低，分析样本越小，统计效力越差。

倾向得分把这个高维匹配问题压缩成了一维匹配问题。

倾向得分回答的问题是：根据这个病人的年龄、病情严重程度等特征，他有多大的概率会被安排上导管？一个 APACHE 评分 35 分、血压极低的患者，临床医生几乎一定会给他插管，他的倾向得分就接近 1。一个 APACHE 评分 10 分、各项指标正常的患者，医生大概率不会插管，他的倾向得分就接近 0。

定义 5.1 (倾向得分)

给定协变量向量 L，个体 i 的倾向得分定义为在协变量条件下接受处理的概率：

$$e(L_i) = P(A_i = 1 \mid L_i).$$

[23]

这个定义很简洁：一个人的倾向得分就是” 根据他的基线特征，预测他接受处理的概率”。在 RHC 数据中，倾向得分回答的问题是” 给定这个患者的年龄、APACHE 评分、血压等 38 个变量，他被插右心导管的概率是多少”。

Rosenbaum 和 Rubin 1983 年的核心贡献在于证明了一条降维定理。这个定理的含义用大白话说就是：你不需要在年龄、性别、APACHE、血压……这十几个变量上同时匹配，只需要在一个数字上匹配就够了。这个数字就是倾向得分。

具体来说，如果可交换性在完整协变量 L 上成立，即 $Y(a) \perp\!\!\!\perp A \mid L$，那么可交换性在倾向得分 $e(L)$ 上同样成立，即 $Y(a) \perp\!\!\!\perp A \mid e(L)$。按倾向得分匹配或分层，与按所有协变量匹配或分层，在消除混杂方面是等价的。

这条定理的实际意义是：不管你有多少个协变量，只要倾向得分估计正确，把所有协变量压缩成一个 0 到 1 之间的标量之后，处理分配在同一倾向得分值上就像是随机的。38 维的匹配问题变成了 1 维的匹配问题，维度灾难被绕过了。

笔记 Rosenbaum 和 Rubin 提出倾向得分时，因果推断面临的核心技术瓶颈是多变量调整的维度灾难。当时最常用的方法是按协变量分层或精确匹配，但协变量超过五六个之后，分层的格子就空了，精确匹配更是几乎不可能。倾向得分用一个标量替代了整个协变量向量，这在计算上的简化是革命性的。1983 年这篇论文发表在 *Biometrika*，至今被引用超过 30000 次，是观察研究方法论的里程碑。

5.2 用逻辑回归估计倾向得分

降维定理告诉我们倾向得分在理论上能做什么，但实际应用中，真实的倾向得分 $e(L)$ 是未知的，需要从数据中估计。最常用的估计方法是逻辑回归：以处理变量 A 作为因变量，协变量 L 作为自变量，拟合 logistic 模型，然后提取每个个体的预测概率作为估计倾向得分 $\hat{e}(L_i)$。

本章继续使用 RHC 数据集，$n = 5735$。处理变量为 *rhc*，结局为 *death180_bin*。我们用全部 38 个协变量拟合倾向得分模型。

```
library(tidyverse)
set.seed(2026)

d <- read_csv(here::here("data", "rhc.csv"), show_col_types = FALSE) |>
  mutate(death180_bin = ifelse(death180 == "Yes", 1, 0))

# 38 个协变量——第 2 章 DAG 确定的调整集
covs <- c("age", "sex", "edu", "das_index", "apache_score",
  "glasgow_coma_score", "blood_pressure", "wbc", "heart_rate",
  "respiratory_rate", "temperature", "pa_o2vs_fio2",
  "albumin", "hematocrit", "bilirubin", "creatinine",
  "sodium", "potassium", "pa_co2", "ph", "weight",
  "dnr_status", "medical_insurance", "race", "income",
  "cancer", "cardiovascular", "congestive_hf", "dementia",
  "psychiatric", "pulmonary", "renal", "hepatic",
  "gi_bleed", "tumor", "immunosupperssion", "transfer_hx", "mi")

fml <- as.formula(paste("rhc ~", paste(covs, collapse = " + ")))

# 倾向得分模型——对处理分配机制建模，不是对结局建模
ps_model <- glm(fml, data = d, family = binomial)
d$ps <- predict(ps_model, type = "response")
summary(d$ps)
```

倾向得分分布

倾向得分的均值为 0.381，中位数为 0.360，范围从 0.005 到 0.960。RHC 组的平均倾向得分为 0.508，对照组为 0.303。两组的分布有明显分离但存在大量重叠区域，这个重叠是后续匹配和加权能够工作的前提。

估计完倾向得分之后，第一步要做的诊断是画重叠直方图，检查两组的倾向得分分布是否有足够的重叠。

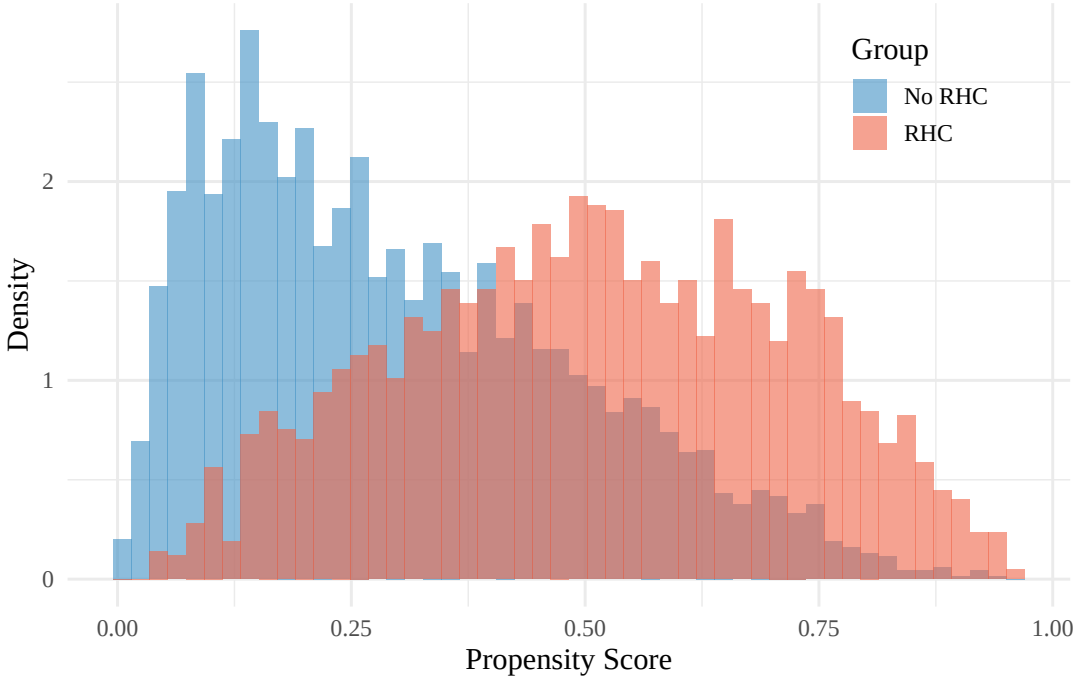

图 5.1: RHC 组与对照组的倾向得分分布。两组分布有明显分离，RHC 组集中在较高区域，对照组集中在较低区域，但中间有大量重叠。重叠区域是倾向得分方法能发挥作用的范围。

图 5.1 显示了两组的倾向得分分布。对照组的密度峰值在 0.10 附近，RHC 组的密度峰值在 0.50 附近。两组有清晰的分离，但重叠区域从 0.05 延伸到 0.90 以上，覆盖了绝大多数样本。倾向得分接近 0 的区域几乎只有对照组，接近 1 的区域几乎只有 RHC 组，这些边缘区域是后续正值性讨论的重点。

定理 5.1 (雷区)

倾向得分模型的目标是平衡协变量，不是预测处理。很多研究者习惯用 AUC 或分类准确率来评价倾向得分模型的好坏，AUC 高就认为模型好。这个判断标准是错的。倾向得分模型追求的是匹配或加权之后协变量在两组之间达到平衡，而非预测能力最大化。一个 AUC 很高的模型可能过拟合了，导致倾向得分分布的重叠变差，匹配更困难。评价倾向得分模型好坏的正确指标是平衡诊断，即标准化均值差 SMD 是否低于阈值，这在本章后半部分会详细讲。 ♡

5.3 倾向得分匹配

倾向得分有多种使用方式。匹配是最直观的一种：在对照组中为每个处理组个体找到倾向得分最接近的配对，组成匹配样本后直接比较结局。

定义 5.2 (倾向得分匹配)

倾向得分匹配，简称 PSM，将每个处理组个体与倾向得分最接近的对照组个体配对。匹配后的样本在倾向得分上近似相等，根据降维定理，这等价于在所有协变量上近似平衡。 ♣

匹配的具体算法有很多种。最常用的是最近邻匹配加卡钳值约束：对每个处理组个体，在对照组中找到倾向得分距离最小的个体配对，同时要求两者的倾向得分差不超过一个预设的阈值，即卡钳值。卡钳值的标准选择是倾向得分 logit 值标准差的 0.2 倍 [4]。超过卡钳值的处理组个体找不到配对，会被丢弃。

这意味着 PSM 必然伴随样本损失。找不到配对的个体被排除在分析之外，这些被排除的个体通常是倾向得分极端的人群，他们在对照组中没有可比的” 双胞胎”。样本损失一方面降低了统计效力，另一方面改变了目标

人群的定义：匹配后的 ATE 严格来说只适用于匹配后保留下来的人群，不再是全人群的 ATE。

下面用 MatchIt 包实施 1:1 最近邻匹配。

```r
library(MatchIt)

# 1:1 最近邻匹配，卡钳值 0.2 倍 logit PS 标准差
m_out <- matchit(fml, data = d, method = "nearest",
                 distance = "glm", caliper = 0.2, ratio = 1)

# 匹配后样本
m_data <- match.data(m_out)
cat("Matched sample:", nrow(m_data), "\n")
cat("RHC:", sum(m_data$rhc), "  No RHC:", sum(m_data$rhc == 0), "\n")

# 匹配样本上的风险差
rd_psm <- mean(m_data$death180_bin[m_data$rhc == 1]) -
          mean(m_data$death180_bin[m_data$rhc == 0])
cat("PSM Risk Difference:", round(rd_psm, 4), "\n")
```

PSM 结果

匹配后保留了 3670 人，每组 1835 人。原始数据中 2184 名 RHC 患者有 349 人没有找到满足卡钳约束的配对被排除，对照组 3551 人中有 1716 人未被匹配。匹配样本上的 180 天死亡率风险差为 0.076，95% bootstrap CI 为 $[0.041, 0.109]$。RHC 组的死亡率比对照组高约 7.6 个百分点。

在 RHC 数据上，PSM 告诉我们：匹配后 RHC 组的 180 天死亡率比对照组高 7.6 个百分点，置信区间不包含零，效应方向与回归调整和 G 计算一致。

需要注意的是，PSM 丢弃了 2065 人，占原始样本的 36%。被丢弃的对照组患者主要是倾向得分很低的个体，即根据其基线特征几乎不可能接受 RHC 的患者。被丢弃的 349 名 RHC 患者则是倾向得分很高的个体，他们在对照组中找不到相似的配对。这种样本损失在临床研究中需要认真对待，因为它意味着分析结论不适用于那些” 几乎必然接受 RHC” 或” 几乎必然不接受 RHC” 的极端人群。

5.4 逆概率加权

加权是倾向得分的另一种使用方式，它不丢弃任何样本。

定义 5.3 (逆概率加权)

逆概率加权，简称 IPW，给每个个体赋予一个与其倾向得分成反比的权重。对于估计 ATE，权重定义为：

$$w_i = \frac{A_i}{\hat{e}(L_i)} + \frac{1 - A_i}{1 - \hat{e}(L_i)}.$$

加权后的伪人群中，处理分配与协变量独立，等价于构造了一个伪随机试验。

倾向得分的倒数这个形式需要解释。处理组中倾向得分大的个体权重为 $1/\hat{e}$，这个值偏小，因为数据里这种人本来就多，代表性已经够了，不需要放大他们的声音。处理组中倾向得分小的个体权重为 $1/\hat{e}$，这个值偏大，因为这种人本来不太可能接受处理，数据里出现得少，需要放大他们的声音才能代表” 如果全人群都接受处理” 的情形。对照组的逻辑完全对称：倾向得分小的人权重为 $1/(1-\hat{e})$，偏小；倾向得分大的人权重为 $1/(1-\hat{e})$，偏大。

这个加权操作的本质是把观察样本的协变量分布扭回到全人群的分布。在伪人群中，高 APACHE 评分的患

者在处理组和对照组中的比例变得一样，低血压的患者也一样。加权之后再比较两组结局，混杂就被消除了。

```r
library(WeightIt)

# IPW: 估计 ATE
w_ipw <- weightit(fml, data = d, method = "glm", estimand = "ATE")
summary(w_ipw)

d$w_ipw <- w_ipw$weights

# 加权后的风险差
ate_ipw <- weighted.mean(d$death180_bin[d$rhc == 1], d$w_ipw[d$rhc == 1]) -
           weighted.mean(d$death180_bin[d$rhc == 0], d$w_ipw[d$rhc == 0])
cat("IPW Risk Difference:", round(ate_ipw, 4), "\n")
```

IPW 结果

IPW 估计的 ATE 风险差为 0.055，95% bootstrap CI 为 $[0.025, 0.085]$。所有 5735 个样本都被保留在分析中。权重分布显示：处理组权重范围为 1.04 到 26.45，对照组为 1.01 到 14.79。有效样本量为处理组 1319、对照组 2806，相比原始样本量有所下降，这是极端权重”稀释”有效信息的结果。

在 RHC 数据上，IPW 告诉我们：保留全部 5735 名患者后，RHC 组的死亡率比对照组高 5.5 个百分点，比 PSM 的 7.6 略低，可能是极端权重对少数个体放大效应的结果。

权重最大值达到 26.45，意味着有一个 RHC 患者在伪人群中被放大了 26 倍。这种极端权重是倾向得分接近 0 或 1 时的直接后果：倾向得分为 0.04 的处理组个体权重为 $1/0.04 = 25$，一个人的数据贡献了 25 人的信息量。这会导致方差膨胀，估计不稳定。

定理 5.2 (雷区)

IPW 的方差对极端权重非常敏感。只要有少数几个倾向得分接近 0 或 1 的个体，他们的权重就会远超正常范围，整个估计会被这几个人主导。诊断方法是检查权重分布：看最大权重是否超过正常范围的几十倍，看有效样本量相比原始样本量下降了多少。如果有效样本量只剩原始的一半甚至更少，说明极端权重在严重损耗信息。应对策略包括权重截断和使用重叠权重。权重截断的做法是把超过某个阈值的权重硬性截断到阈值，代价是引入了一点偏差换取方差的大幅下降。重叠权重则从根本上避免了极端权重的问题，下一节会详细讲。

5.5 重叠权重

重叠权重，简称 OW，是近年来被越来越多研究者采用的替代方案。它的权重形式非常简洁：处理组个体的权重是 $1 - \hat{e}(L_i)$，对照组个体的权重是 $\hat{e}(L_i)$。

定义 5.4 (重叠权重)

重叠权重给每个个体赋予与其”被分到另一组的概率”成正比的权重：

$$w_i^{\text{OW}} = A_i \cdot [1 - \hat{e}(L_i)] + (1 - A_i) \cdot \hat{e}(L_i).$$

重叠权重估计的因果效应称为 ATO，即 average treatment effect in the overlap population。

重叠权重的设计逻辑值得理解。倾向得分接近 0.5 的个体权重最大，因为这些人”有可能被分到任何一组”，是两组最可比的人群。倾向得分接近 0 或 1 的个体权重趋近于 0，因为这些人几乎必然只能出现在一组中，对照

组里找不到可比的人，让他们对估计的贡献降到最低是合理的。

相比 IPW，重叠权重天然避免了极端权重的问题。IPW 在倾向得分为 0.02 时给出权重 50，重叠权重在同样的位置给出权重 0.02，差了整整 2500 倍。这就是为什么重叠权重的方差通常远小于 IPW。

代价是目标估计量变了。IPW 估计的是全人群的 ATE，而重叠权重估计的是重叠人群的 ATO。如果研究者的问题是” 对所有 ICU 患者，RHC 的平均效应是什么”，应该用 IPW。如果问题是” 对那些处理决策存在真正临床均势的患者，RHC 的效应是什么”，重叠权重更合适。在 RHC 数据中，倾向得分极端的患者要么是病情太重几乎必然插管，要么是病情太轻几乎不可能插管，这些人群的因果效应估计可靠性本来就差，聚焦于重叠人群是务实的选择。

```r
# OW: 估计 ATO，即重叠人群的平均效应
w_ow <- weightit(fml, data = d, method = "glm", estimand = "ATO")
summary(w_ow)

d$w_ow <- w_ow$weights

ate_ow <- weighted.mean(d$death180_bin[d$rhc == 1], d$w_ow[d$rhc == 1]) -
          weighted.mean(d$death180_bin[d$rhc == 0], d$w_ow[d$rhc == 0])
cat("OW Risk Difference:", round(ate_ow, 4), "\n")
```

OW 结果

重叠权重估计的 ATO 风险差为 0.061，95% bootstrap CI 为 $[0.033, 0.089]$。权重分布非常温和，处理组权重范围为 0.04 到 0.96，对照组为 0.005 到 0.93，没有任何极端值。有效样本量为处理组 1862、对照组 2532，分别是原始样本量的 85% 和 71%，信息损耗远小于 IPW。

在 RHC 数据上，OW 告诉我们：聚焦于处理决策存在临床均势的重叠人群，RHC 组的死亡率比对照组高 6.1 个百分点，介于 PSM 和 IPW 之间，且方差最小、估计最稳定。

5.6 平衡诊断：标准化均值差与 Love plot

倾向得分方法能否消除混杂，最终要看协变量在匹配或加权之后是否达到了平衡。平衡诊断是整个分析流程中不可跳过的一步。

定义 5.5 (标准化均值差)

标准化均值差，简称 SMD，衡量某个协变量在处理组和对照组之间的差异程度：

$$\text{SMD} = \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{(s_1^2 + s_0^2)/2}},$$

其中 $\bar{X}_1, \bar{X}_0$ 分别是处理组和对照组的协变量均值，s_1^2, s_0^2 是对应的方差。SMD 的绝对值小于 0.1 通常被认为是良好平衡的标志 [4]。

SMD 比 p 值更适合用来评价平衡。p 值同时依赖效应大小和样本量，大样本中微小的不平衡也会给出极小的 p 值，让研究者误以为平衡很差。SMD 不受样本量影响，直接反映两组均值差了几个标准差，更具临床可解读性。

Love plot 是把所有协变量的 SMD 画在一张图上的可视化工具，名称来源于统计学家 Thomas Love。图中每个点代表一个协变量在某种调整方法下的绝对 SMD，竖线标出 0.1 的阈值。所有点都落在阈值左边，说明平衡良好。

```r
library(cobalt)

# Love plot: 同时展示 PSM / IPW / OW 的平衡
love.plot(fml, data = d,
          stats = "m", abs = TRUE,
          thresholds = c(m = 0.1),
          weights = list(PSM = m_out, IPW = w_ipw, OW = w_ow),
          colors = c("#999999", "#EF6548", "#4292C6", "#66C2A5"),
          shapes = c(17, 16, 15, 18),
          sample.names = c("Unadjusted", "PSM", "IPW", "OW"))
```

在看图之前，先说明怎么读 Love plot。横轴是绝对 SMD，数值越大说明两组在该变量上的差距越大。纵轴是每个协变量的名称。图中有一条竖直虚线标在 0.10 的位置，这是公认的平衡阈值。每个协变量在不同调整方法下各有一个点：灰色三角代表未调整，红色圆形代表 PSM，蓝色方形代表 IPW，绿色菱形代表 OW。所有有色点都落在 0.10 虚线左边，就说明该方法成功地平衡了协变量。

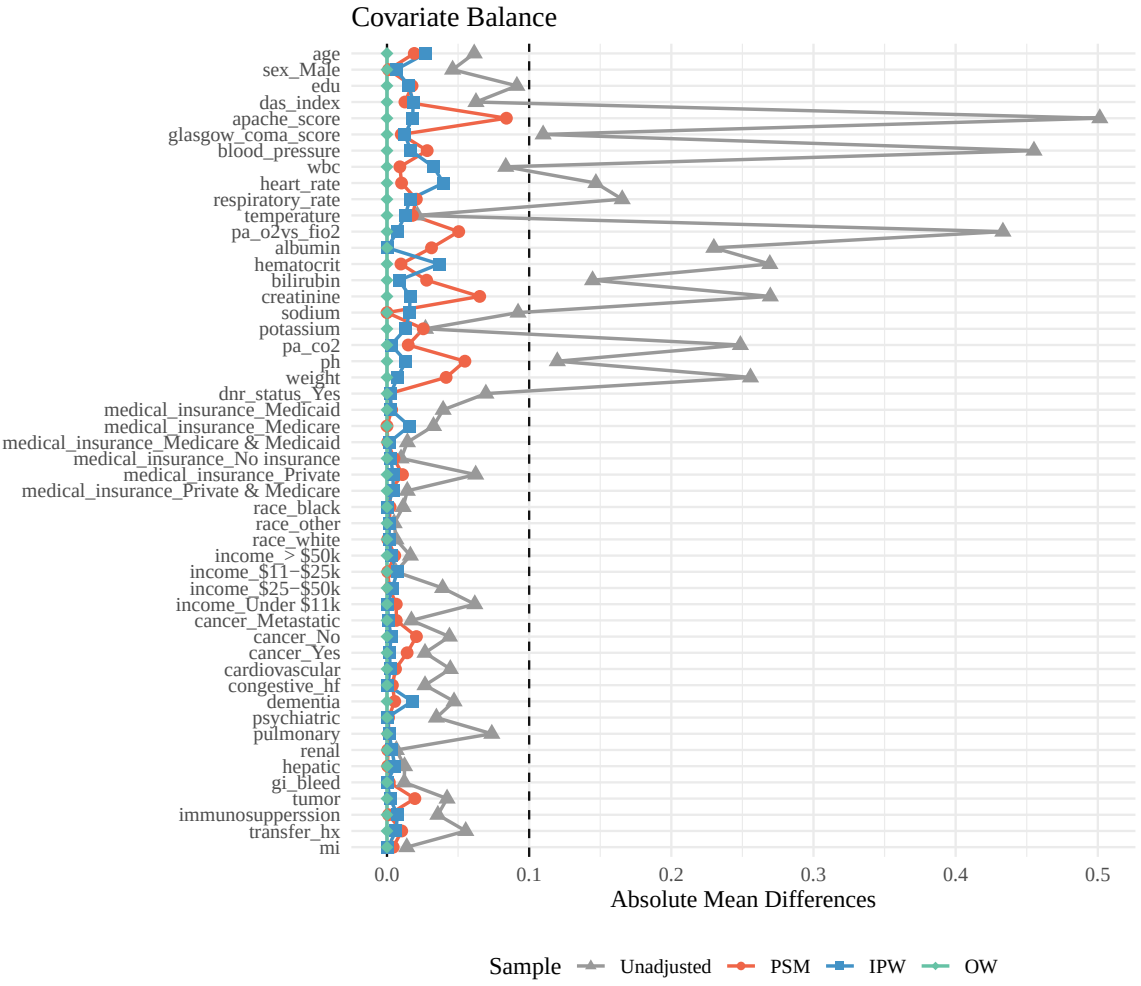

图 5.2: Love plot：四种状态下协变量的绝对标准化均值差。灰色三角为未调整，红色圆形为 PSM，蓝色方形为 IPW，绿色菱形为 OW。虚线为 0.1 阈值。

平衡诊断

图 5.2 清楚地展示了调整效果。未调整时，*apache_score*、*blood_pressure*、*pa_o2vs_fio2*、*hematocrit*、*pa_co2*、*weight*、*creatinine* 等变量的 SMD 超过 0.20，最大的 *apache_score* 达到 0.48，*blood_pressure* 达到 0.49。PSM 匹配后，绝大多数变量的 SMD 降到了 0.05 以下，但 *apache_score* 仍有 0.08 的残余不平衡，接近 0.1 阈值。IPW 和 OW 的平衡效果优于 PSM，几乎所有变量的 SMD 都降到了 0.05 以内，OW 的表现尤其稳定。三种方法都成功地将协变量不平衡控制在了可接受的范围内。

定理 5.3 (雷区)

平衡诊断看的是协变量分布，不是模型拟合指标。发表论文时常见的错误是：报告了倾向得分模型的 Hosmer-Lemeshow 拟合优度检验通过了，就认为倾向得分方法可靠，不再检查匹配后或加权后的协变量平衡。模型拟合好不代表平衡好。诊断的正确流程是：拟合倾向得分模型，实施匹配或加权，然后检查每个协变量的 SMD。如果某些变量 SMD 仍然超过 0.1，说明模型需要调整，比如加入交互项或非线性项。平衡诊断是迭代的过程，直到所有协变量 SMD 低于阈值才算完成。 ♡

5.7 正值性违反

倾向得分方法的三大前提假设是可交换性、正值性和一致性。其中正值性要求每个协变量组合下，接受处理和不接受处理的概率都严格大于零：$0 < e(L) < 1$。

在 RHC 数据中，倾向得分的最小值为 0.005，最大值为 0.960。没有真正的 0 或 1 出现，正值性在技术上没有被” 硬违反”。但 0.005 已经非常接近零了，这意味着有些患者根据其基线特征几乎不可能接受 RHC。这种” 近乎违反正值性” 的情况在 IPW 中会制造极端权重：倾向得分为 0.005 的对照组个体权重为 $1/(1-0.005) \approx 1$，没什么问题；但如果这个个体碰巧出现在处理组，权重就是 $1/0.005 = 200$，一个人贡献了 200 人的信息量。

从临床角度看，正值性违反传递的信息是：对于某些患者，临床医生根据病情已经做出了明确的决策，不存在” 随机分配到另一组” 的可能。这些患者的因果效应本质上无法从观察数据中可靠估计。重叠权重通过把这些极端个体的权重自动压低到接近零，在方法层面回应了这个问题，但正值性违反背后的科学含义仍然需要研究者在论文讨论部分认真对待。

5.8 累积对比表

表 5.1: 方法演进对比表，截至第 5 章

方法	ATE 估计	95% CI	核心假设
回归调整	OR = 1.34	[1.18, 1.52]	模型设定正确 + 可交换性 + 正值性
G 计算	RD = 0.071	[0.038, 0.105]	结局模型正确 + 可交换性 + 正值性
PSM	RD = 0.076	[0.041, 0.109]	处理模型正确 + 可交换性 + 正值性
IPW	RD = 0.055	[0.025, 0.085]	处理模型正确 + 可交换性 + 正值性
OW	RD = 0.061	[0.033, 0.089]	处理模型正确 + 可交换性 + 正值性

五种方法都指向同一个方向：RHC 增加 180 天死亡率。回归给出的是 OR 尺度的效应，不能直接与风险差比较，但 OR = 1.34 表示死亡几率比升高 34%，方向一致。G 计算的 RD = 0.071 与 PSM 的 0.076 相差不大，这并不意外，因为两者都依赖逻辑回归形式的参数模型，区别只在于一个建模结局一个建模处理分配。

IPW 给出的 RD = 0.055 偏小一些，可能是极端权重对少数个体的放大效应导致的。OW 的 RD = 0.061 介于 IPW 和 PSM 之间，方差最小。需要注意的是 PSM 和 OW 的目标人群与 IPW 不完全相同：PSM 针对匹配后保留的人群，OW 针对重叠人群，只有 IPW 估计的是全人群的 ATE。方法之间的数值差异部分来源于目标人群的不同，部分来源于统计方法的不同。

下一章将引入双重稳健估计，即 AIPW。它同时建立结局模型和处理模型，只要其中一个正确就能给出一致估计。G 计算赌结局模型对，倾向得分方法赌处理模型对，AIPW 的双重稳健性让研究者不必在两个赌注之间做选择，两根保险绳只要一根不断就安全。

方法卡片：倾向得分方法

数学形式：$e(L) = P(A = 1 \mid L)$，通过逻辑回归或其他分类器估计。
使用方式：匹配、加权 IPW/OW、分层、协变量调整。
核心假设：可交换性 $Y(a) \perp\!\!\!\perp A \mid L$，正值性 $0 < e(L) < 1$，一致性。
R 包：MatchIt 匹配、WeightIt 加权、cobalt 平衡诊断。
优势：不依赖结局模型的函数形式；协变量平衡可直接检验；对倾向得分模型的轻度错误设定有一定容错性。
失效场景：正值性违反或接近违反时极端权重导致方差爆炸；倾向得分模型严重错误设定时平衡失败；不可观测混杂无法通过倾向得分控制。

本章知识地图

表 5.2: 第 5 章核心概念与常见误解

核心概念	核心内容	常见误解	为什么错
倾向得分	给定协变量下接受处理的条件概率，将多维匹配压缩为一维	倾向得分是处理效应	倾向得分只描述处理分配机制，与结局无关
降维定理	按 $e(L)$ 调整等价于按全部 L 调整	倾向得分包含了结局的信息	定理只保证平衡协变量，不建模结局
PSM	按倾向得分最近邻配对，组成可比样本后比较结局	匹配后不需要检查平衡	匹配算法不保证所有变量平衡，必须用 SMD 诊断
IPW	用倾向得分的倒数加权，构造协变量平衡的伪人群	AUC 高的倾向得分模型一定好	模型好坏看加权后的协变量平衡，不看预测能力
重叠权重	对重叠人群赋权，天然避免极端权重	OW 估计的就是全人群 ATE	OW 目标是 ATO，聚焦于处理决策存在均势的人群
SMD 与 Love plot	标准化均值差衡量协变量平衡，Love plot 可视化所有变量	用 p 值判断平衡	p 值受样本量影响，大样本中小不平衡也显著
正值性违反	某些协变量组合下接受处理的概率趋近 0 或 1	倾向得分没有 0 或 1 就没问题	接近 0 或 1 就足以让 IPW 方差爆炸

第 6 章 双重稳健估计——AIPW 的两根保险绳

内容提要

- 理解单一模型依赖的系统性风险
- 掌握 AIPW 估计量的三个组成部分及其数学形式
- 用具体数字理解双重稳健性的交叉消除机制
- 在 RHC 数据上手动实现 AIPW 并与 G 计算、IPW 做横向对比

上一章用倾向得分做了匹配和加权。无论是 PSM 还是 IPW，核心逻辑都是绕开结果模型，用处理模型去平衡协变量分布，然后在平衡后的伪总体中直接比较结局。第 4 章的 G 计算走的是另一条路：依赖结果模型预测反事实，完全不碰处理模型。两种策略各赌一个模型正确。G 计算赌结果模型，IPW 赌处理模型。赌对了，估计一致；赌错了，偏差无法消除，样本量再大也没用。

在真实数据中，没有人能确定自己的模型是对的。结果模型可能遗漏了交互项或非线性关系，处理模型的 logistic 回归也可能没有捕捉到医生决策的全部逻辑。本章介绍的增强逆概率加权，英文称 Augmented Inverse Probability Weighting，简称 AIPW，做的事情是同时建两个模型，让它们互相兜底。只要其中一个对，最终估计就是一致的。这个性质叫双重稳健性，英文称 double robustness。用登山的比喻来说，G 计算和 IPW 各拴了一根保险绳，AIPW 把两根绳都拴上了，只要一根不断就安全。

6.1 单一模型的系统性风险

回顾前两章的核心结论。G 计算在 RHC 数据上给出的边际风险差是 0.052，意思是如果所有 5735 名患者都接受 RHC，180 天死亡率比都不接受高 5.2 个百分点。这个数字的可靠性完全取决于结果模型 $E[Y \mid A, L]$ 的设定。如果 APACHE 评分对死亡率的影响存在阈值效应而我们只放了线性项，G 计算的预测就会系统性偏离真实值，风险差的估计也随之偏移。

IPW 走了相反的路线。它不关心结局怎么建模，只关心倾向得分模型 $e(L) = P(A = 1 \mid L)$ 是否正确。上一章的 IPW 估计给出风险差约 0.032，标准误比 G 计算大了将近 50%，原因是倾向得分接近 0 或 1 的个体权重很大，把方差拽高了。更严重的问题是，如果倾向得分模型遗漏了某个关键的混杂变量或者函数形式写错了，加权后的伪总体并没有真正平衡协变量，估计仍然有偏。

两种方法面临的困境可以归结为一句话：研究者必须在两个模型之间押注，而且不知道自己押的是对还是错。AIPW 的设计动机就是从这个困境中解脱出来。

6.2 AIPW 估计量的三个组成部分

AIPW 的估计量看起来比 G 计算和 IPW 都复杂，但拆开来看，它就是把两种方法用一个特定的公式组合在一起。在给出公式之前，先明确记号：$\hat{m}_1(L_i)$ 是结果模型预测的” 如果第 i 个人接受处理，死亡概率是多少”，$\hat{m}_0(L_i)$ 是” 如果不接受处理，死亡概率是多少”，$\hat{e}(L_i)$ 是倾向得分。

AIPW 的想法很简单：先用结果模型做一个粗略预测，然后用倾向得分来纠正这个预测的偏差。假设结果模型预测某位患者的死亡概率是 60%，但他实际死了，这意味着结果模型漏掉了 40% 的残差。倾向得分决定了纠正这个残差的” 力度”：如果这位患者接受处理的概率是 0.5，残差就要乘以 $1/0.5 = 2$，放大两倍；如果接受处理的概率是 0.8，残差只需乘以 $1/0.8 = 1.25$，放大幅度更小。结果模型越准，需要纠正的残差越小；倾向得分越准，纠正的方向和力度越对。两者配合，就是 AIPW 的核心机制。

定义 6.1 (AIPW 估计量)

AIPW 对平均处理效应的估计为

$$\widehat{\mathrm{ATE}}_{\mathrm{AIPW}} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\left[\hat{m}_1(L_i) - \hat{m}_0(L_i)\right]}_{\text{A: G 计算部分}} + \underbrace{\frac{A_i}{\hat{e}(L_i)}\left[Y_i - \hat{m}_1(L_i)\right]}_{\text{B: 处理组校正}} - \underbrace{\frac{1-A_i}{1-\hat{e}(L_i)}\left[Y_i - \hat{m}_0(L_i)\right]}_{\text{C: 对照组校正}}$$

其中 A_i 是处理指示变量，Y_i 是观测结局。 [21]

这个公式有三个部分，理解它们各自存在的理由是掌握 AIPW 的关键。

6.2.1 部分 A：结果模型的主估计

部分 A 就是 G 计算的点估计：用结果模型预测每个人在两种反事实状态下的结局，取差值。它存在的理由是提供一个基础答案。如果结果模型完全正确，A 本身就足以给出无偏的 ATE 估计，B 和 C 的期望值为零，不起作用。

用一个具体患者来理解：某位 APACHE 评分 60 的患者，结果模型预测他接受 RHC 后死亡概率 0.75，不接受时 0.65。A 项对这位患者的贡献就是 $0.75 - 0.65 = 0.10$。如果结果模型的预测是准确的，这个 0.10 就是他的个体处理效应估计，不需要任何校正。

6.2.2 部分 B：处理组的残差校正

部分 B 只对处理组的个体有值，它存在的理由是修补结果模型在处理组方向上的预测偏差。$A_i = 1$ 时，它计算的是”这个人的实际结局 Y_i 与结果模型预测值 $\hat{m}_1(L_i)$ 之间的残差”，再除以倾向得分 $\hat{e}(L_i)$。残差的含义是结果模型预测偏了多少。如果结果模型完美，残差期望为零，B 项不贡献任何东西。如果结果模型偏了，B 项就利用倾向得分权重来校正这个偏差。

继续上面那位患者的例子。假设他实际接受了 RHC 并且死了，$Y_i = 1$，但结果模型只预测了 0.75 的死亡概率，残差为 $1 - 0.75 = 0.25$。如果他的倾向得分是 0.5，B 项就是 $0.25/0.5 = 0.50$，把这 0.25 的预测偏差放大后补回去。倾向得分越低，说明这位患者接受 RHC 的概率越小，他作为处理组成员的”代表性”越强，校正力度就越大。

6.2.3 部分 C：对照组的残差校正

部分 C 的逻辑与 B 对称，它存在的理由是修补结果模型在对照组方向上的预测偏差。$A_i = 0$ 时，它用对照组的实际结局与 $\hat{m}_0(L_i)$ 的残差，除以 $1 - \hat{e}(L_i)$，校正结果模型在对照组方向上的偏差。

假设另一位患者没有接受RHC，实际存活了，$Y_i = 0$，结果模型预测不接受时死亡概率 0.65，残差为 $0 - 0.65 = -0.65$。如果倾向得分是 0.5，C 项就是 $-0.65/0.5 = -1.30$。负号表示结果模型高估了对照组的死亡概率，校正方向是往下拉。

6.2.4 三部分的协作

把三部分合在一起看：A 是结果模型的主估计，B 和 C 是 IPW 风格的残差校正项。如果结果模型对，B 和 C 的期望为零，不干扰 A；如果结果模型错但倾向得分对，B 和 C 恰好能把 A 的偏差抵消掉。这就是双重稳健性的直觉来源。

停下来想一想：在上面那位 APACHE 评分 60 的患者例子中，A 项贡献了 0.10，B 项贡献了 0.50。如果结果模型是完美的，B 项会变成多少？如果倾向得分不是 0.5 而是 0.9，B 项又会变成多少？用这两个问题检验自己是否真正理解了三部分的协作逻辑。

6.3 双重稳健性：交叉消除的数学机制

双重稳健性听起来像魔术，但背后有清晰的数学逻辑。先用一个数字例子建立直觉。假设结果模型偏了 2%，处理模型偏了 3%。如果偏差是加法叠加的，总偏差就是 $2\%+3\% = 5\%$。但 AIPW 的偏差大约是 $2\%\times3\% = 0.06\%$，远小于单独用任何一个模型。这就是" 乘积" 结构的威力：只要其中一个模型的偏差接近零，乘积就接近零。

核心在于 AIPW 估计量的偏差项正是两个模型误差的乘积。用符号写出来：AIPW 的偏差正比于

$$E\big[\big(\hat{e}(L) - e(L)\big) \cdot \big(\hat{m}(L) - m(L)\big)\big],$$

其中 $e(L)$ 和 $m(L)$ 分别是真实的倾向得分和真实的结果条件期望。如果倾向得分模型正确，$\hat{e}(L) - e(L) = 0$，乘积为零，偏差消失，无论结果模型偏了多少。反过来，如果结果模型正确，$\hat{m}(L) - m(L) = 0$，乘积同样为零。只有两个模型同时错误，偏差才会存活。

用一个具体的数字例子来建立直觉。假设有一个 APACHE 评分为 60 的患者，真实的条件死亡概率是 0.80，而我们的结果模型预测为 0.70，偏了 0.10。如果这个患者的真实倾向得分是 0.50 而我们的处理模型也估成了 0.50，那么 B 项中的 IPW 权重 $1/\hat{e} = 2$ 恰好把残差 $(Y_i - 0.70)$ 按正确的比例放大，校正了结果模型 0.10 的偏差。

关键在于倾向得分对了，权重分配就对了，即使结果模型的预测有系统性偏差，加权校正后的估计仍然是无偏的。

反过来，假设倾向得分模型估偏了，把 0.50 估成了 0.30，但结果模型完美预测了 0.80。此时 B 项中的残差 $Y_i - \hat{m}_1 = Y_i - 0.80$ 的条件期望为零，无论 IPW 权重是 $1/0.30$ 还是 $1/0.50$，乘上一个期望为零的残差，结果还是零。A 项本身就给出正确答案，B 和 C 只是在加减零。

命题 6.1 (双重稳健性的正式表述)

在以下两组条件中，只要任意一组成立，AIPW 估计量就是渐近无偏的：条件一，结果模型 $\hat{m}_a(L)$ 一致地估计了 $E[Y \mid A = a, L]$，$a \in \{0,1\}$；条件二，倾向得分模型 $\hat{e}(L)$ 一致地估计了 $P(A = 1 \mid L)$。 [6]

渐近无偏用白话说就是：当样本量足够大时，AIPW 的估计会越来越接近真实的 ATE，不会系统性地偏高或偏低。这个保证只需要两个模型中有一个是对的。在 RHC 数据中，即使我们的 logistic 回归结果模型遗漏了某些交互项，只要倾向得分模型对医生决策逻辑的近似是合理的，AIPW 的估计仍然会随着样本量增大而逼近真值。

当两个模型同时正确时，AIPW 还有一个额外的好处：它达到半参数效率界，即在所有正则的渐近线性估计量中方差最小。效率界这个性质值得展开。在只有一个模型正确的情况下，AIPW 的方差不一定是最小的，但它至少保证了一致性。当两个模型都对时，AIPW 不仅无偏，而且方差达到了理论下界，比单独的 G 计算或 IPW 更精确。这意味着 AIPW 在最优情况下不牺牲效率，在次优情况下仍然保底，是一个" 进可攻退可守" 的估计量。

6.4 手动实现：在 RHC 数据上构造 AIPW

本章继续使用第 1 章介绍的 RHC 数据集 $n = 5735$。下面的代码从头开始，分四步完成 AIPW 的手动实现：拟合结果模型、拟合倾向得分模型、预测反事实、组装 AIPW 估计量。手动实现的目的是让读者看清 AIPW 公式里每一项对应的代码操作，后续章节会用 R 包简化这个过程。

```
set.seed(2026)
library(tidyverse)

d <- read_csv(here::here("data", "rhc.csv"), show_col_types = FALSE) |>
  mutate(death180_bin = if_else(death180 == "Yes", 1L, 0L),
         sex_bin      = if_else(sex == "Male", 1L, 0L),
         cancer_bin   = if_else(cancer == "No", 0L, 1L))

```

```r
covs <- c("age", "sex_bin", "cancer_bin", "cardiovascular",
          "congestive_hf", "dementia", "psychiatric", "pulmonary",
          "renal", "hepatic", "gi_bleed", "tumor",
          "immunosupperssion", "transfer_hx", "mi",
          "apache_score", "glasgow_coma_score", "blood_pressure",
          "heart_rate", "respiratory_rate", "temperature",
          "albumin", "creatinine", "bilirubin", "wbc",
          "hematocrit", "das_index", "weight")

# 第一根绳：结果模型——预测 Y 在 A 和 L 条件下的期望
out_mod <- glm(death180_bin ~ rhc + .,
               data = d |> select(death180_bin, rhc, all_of(covs)),
               family = binomial)

# 第二根绳：处理模型——预测谁更可能接受 RHC
ps_mod <- glm(rhc ~ .,
              data = d |> select(rhc, all_of(covs)),
              family = binomial)
d$ps <- predict(ps_mod, type = "response")

# 反事实预测：让每个人分别"接受"和"不接受" RHC
d1 <- d0 <- d
d1$rhc <- 1; d0$rhc <- 0
d$m1 <- predict(out_mod, newdata = d1, type = "response")
d$m0 <- predict(out_mod, newdata = d0, type = "response")

# 组装 AIPW：三个部分逐项计算
d$aipw_score <- with(d, {
  (m1 - m0) +                                # A: G 计算部分
    rhc / ps * (death180_bin - m1) -         # B: 处理组校正
    (1 - rhc) / (1 - ps) * (death180_bin - m0)  # C: 对照组校正
})

ate_aipw <- mean(d$aipw_score)
se_aipw  <- sd(d$aipw_score) / sqrt(nrow(d))
cat("ATE:", round(ate_aipw, 4),
    " SE:", round(se_aipw, 4),
    " 95% CI: [", round(ate_aipw - 1.96*se_aipw, 4),
    ",", round(ate_aipw + 1.96*se_aipw, 4), "]\n")
```

结果解读

AIPW 估计的边际风险差为 0.044，标准误 0.014，95% CI 为 $[0.017, 0.072]$。置信区间不包含零，$p < 0.05$，表明在控制了 28 个协变量之后，接受 RHC 的患者 180 天死亡率仍然比不接受者高约 4.4 个百分点。

把 AIPW 的三个组成部分拆开来看：A 项，即 G 计算部分，贡献了 0.052 的风险差，这和第 4 章单独用 G 计算得到的结果一致。B 项，处理组的 IPW 校正，贡献了 -0.008，方向为负，意味着结果模型对处理组的死亡概率略有高估，IPW 校正把估计往下拉了一点。C 项，对照组的 IPW 校正，几乎为零。三项加起来 $0.052 + (-0.008) - 0.000 = 0.044$。

AIPW 的标准误 0.014 比 IPW 的 0.022 小了 36%，比 G 计算的 bootstrap 标准误也更窄。这正是双重稳健估计量在两个模型都大致正确时的效率优势。

图 6.1 展示了 5735 名患者各自的 AIPW 得分分布。每个人的 AIPW 得分就是公式中求和号内的那个值，它可以看作”这个人对整体 ATE 估计的贡献”。ATE 就是所有人得分的均值。

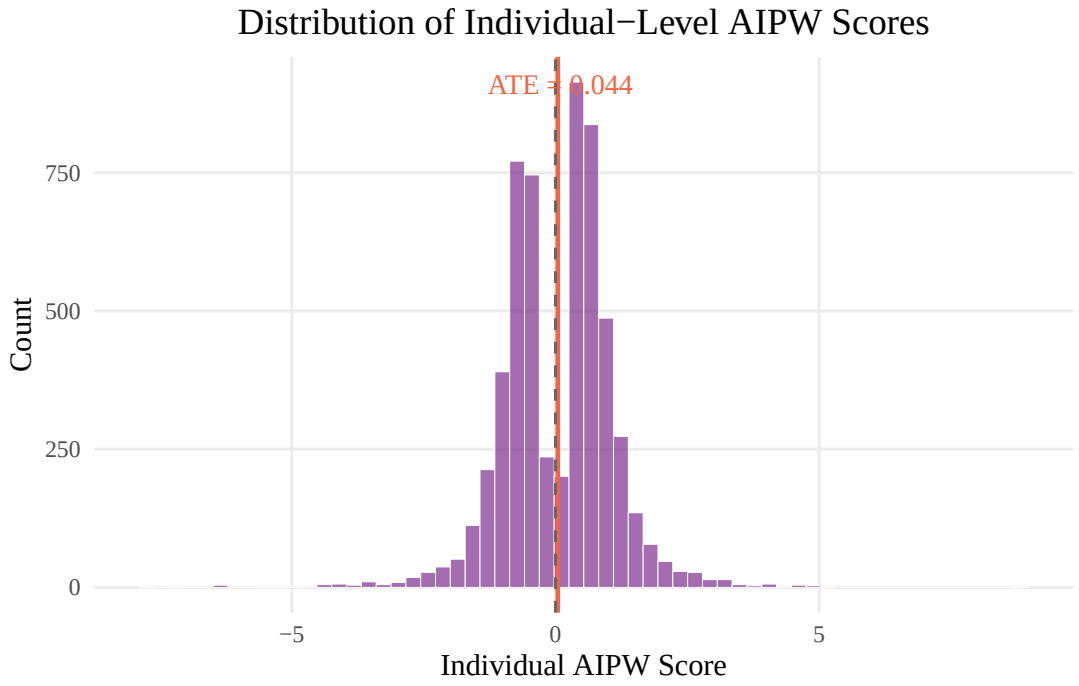

图 6.1: 5735 名患者的 AIPW 个体得分分布。红色实线为 ATE = 0.044，灰色虚线为零。分布以零附近为中心，但整体略偏右，对应正的处理效应。左尾的少数极端值来自倾向得分接近边界的个体。

6.5 三种方法的横向对比

现在我们有了三种方法在同一份数据上的估计，可以放在一起比较了。

表 6.1: G 计算、IPW 与 AIPW 在 RHC 数据上的估计对比

方法	RD	95% CI	核心依赖
G 计算	0.052	[0.027, 0.082]	结果模型设定正确
IPW	0.032	[0.005, 0.064]	倾向得分模型设定正确
AIPW	0.044	[0.017, 0.072]	两个模型至少一个正确

三个估计的方向一致：RHC 增加了 180 天死亡率，风险差在 3–5 个百分点之间。但它们之间的差异包含有用的诊断信息。G 计算给出的 0.052 最高，IPW 的 0.032 最低，AIPW 的 0.044 落在两者之间。这个模式是有道理的：AIPW 本质上是用 IPW 校正项去修正 G 计算的初始估计，B 项为负说明 G 计算略有高估，AIPW 把它往下拉了约 0.008 个百分点。

IPW 的置信区间最宽，下界几乎触及零，这是因为倾向得分接近 0 或 1 的个体放大了方差。AIPW 的置信区间比 IPW 窄了三分之一，同时也比 G 计算的 bootstrap 区间更精确。图 6.2 直观展示了这个比较。

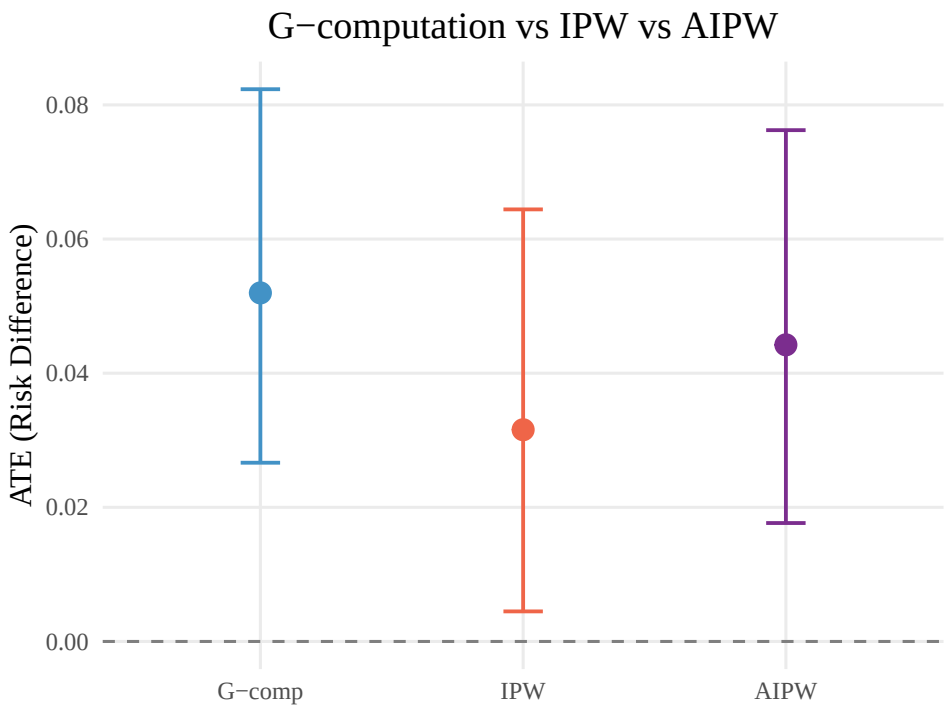

图 6.2: 三种方法的 ATE 估计及 95% CI 对比。AIPW 的点估计位于 G 计算和 IPW 之间，置信区间宽度介于两者之间。虚线为零参考线。

笔记 AIPW 的理论基础可以追溯到 Robins, Rotnitzky, and Zhao [21] 关于半参数效率理论的开创性工作。Robins 和 Rotnitzky 在 1990 年代初系统研究了”如果一个估计量依赖辅助模型，当辅助模型错误时会发生什么”的问题，并发现了一类具有双重稳健性的估计量。但 AIPW 的理论优势在当时很难体现，因为 1990 年代的因果分析几乎全部使用参数模型。如果结果模型用逻辑回归、处理模型也用逻辑回归，两个模型同时设错的概率和只设错一个差不多，双重稳健的”保险”价值有限。

AIPW 真正获得广泛关注要等到两个转折点。Bang and Robins [6] 用 Monte Carlo 模拟清楚地展示了双重稳健性在有限样本中的表现，让实务工作者看到了它的实际收益。而 2010 年代机器学习进入因果推断之后，结果模型和处理模型可以分别用随机森林、Lasso、神经网络等灵活方法来拟合，”至少一个对”的概率比两个 logistic 回归的组合大幅提高。AIPW 在 Funk et al. [10] 的流行病学实操指南中被推荐为标准方法之后，逐渐成为临床和公共卫生领域观察性研究的默认选择。

6.6 AIPW 何时失灵

双重稳健性给了 AIPW 额外的保护层，但它有自己的失效边界。

最明显的失效场景是两个模型同时错误。如果结果模型遗漏了关键的交互项，同时倾向得分模型也遗漏了相同方向的混杂变量，AIPW 的偏差项 $E[(\hat{e}-e)(\hat{m}-m)]$ 不再为零。更严重的是，两个模型的误差如果在同一方向上相关，偏差可能比单独使用任一方法更大。双重稳健保证的是”只要一个对就行”，它不保证”两个都错时还能救”。

定理 6.1 (雷区)

双重稳健性的”双保险”有一个前提条件：两个模型中至少有一个对数据生成过程的近似是足够好的。在实际操作中，研究者无法验证任一模型是否正确，因为真实的数据生成过程是未知的。当两个模型都用同一批协变量、同一种参数形式建立时，它们犯的错误往往高度相关，双重稳健性提供的保护就会大打折扣。解决思路是让两个模型的设定尽量不同：比如结果模型用非参数方法，处理模型用参数方法，或者

反过来。第 7 章引入的 Super Learner 和 TMLE 可以进一步降低模型设定错误的风险。 ♡

极端倾向得分是 AIPW 的另一个薄弱环节。虽然 AIPW 对极端权重的敏感性低于纯 IPW，因为 B 和 C 项乘的是残差而非原始结局，但当 $\hat{e}(L)$ 非常接近 0 或 1 时，权重 $1/\hat{e}$ 或 $1/(1-\hat{e})$ 仍然会放大个别观测的残差，让方差膨胀。在 RHC 数据中，最小的倾向得分为 0.024，对应的权重约为 42，这意味着一个对照组个体的残差被放大了 42 倍。虽然只有 29 个样本的倾向得分低于 0.05，但它们对整体标准误的贡献不可忽略。常见的处理方式是对倾向得分做截断，把低于 0.025 或高于 0.975 的值拉回边界，代价是引入少量偏差但换来方差的显著下降。

AIPW 在有限样本中的另一个问题是它可能给出超出参数空间的估计。比如二分类结局的风险差理论上应该在 $[-1, 1]$ 之间，但 AIPW 的线性组合没有内置这个约束。在样本量小或倾向得分极端的数据中，AIPW 可能估出 -1.2 或 1.3 这样不合理的值。第 7 章介绍的 TMLE 通过目标化的最大似然更新步骤解决了这个问题，把估计拉回到合规空间内。

练习 6.1 在上面的 AIPW 实现中，将倾向得分截断为 $[0.025, 0.975]$，重新计算 ATE 及 95% CI。比较截断前后的点估计差异和标准误变化，判断极端倾向得分对本数据的影响程度。

解

```r
# 倾向得分截断——牺牲少量偏差换取方差稳定
d$ps_trim <- pmax(0.025, pmin(0.975, d$ps))

d$aipw_trim <- with(d, {
  (m1 - m0) +
    rhc / ps_trim * (death180_bin - m1) -
    (1 - rhc) / (1 - ps_trim) * (death180_bin - m0)
})

ate_trim <- mean(d$aipw_trim)
se_trim  <- sd(d$aipw_trim) / sqrt(nrow(d))
cat("Trimmed ATE:", round(ate_trim, 4),
    " SE:", round(se_trim, 4), "\n")
```

由于 RHC 数据中极端倾向得分的样本量较少，截断前后的点估计差异通常不超过 0.002，标准误的变化也很小。这说明在本数据集中，正值性的违反程度不严重。但在其他数据集中，特别是处理概率高度不平衡的情况下，截断的影响可能显著得多。

方法卡片：AIPW

估计目标：边际风险差 $E[Y(1)] - E[Y(0)]$，即 ATE。
核心公式：$\widehat{\text{ATE}} = \frac{1}{n}\sum_i \big[\hat{m}_1(L_i) - \hat{m}_0(L_i)\big] + \frac{A_i}{\hat{e}(L_i)}\big[Y_i - \hat{m}_1(L_i)\big] - \frac{1-A_i}{1-\hat{e}(L_i)}\big[Y_i - \hat{m}_0(L_i)\big]$。
核心假设：可交换性 + 正值性 + 一致性；双重稳健额外要求两个模型至少一个设定正确。
R 实现：手动组装或 AIPW 包。配合 SuperLearner 可用机器学习拟合两个子模型。
适用场景：研究者对单一模型没有充分信心；需要同时利用结果模型和处理模型的信息；追求半参数效率最优。
失效场景：两个模型同时错误；极端倾向得分导致方差膨胀；有限样本中估计可能超出参数空间。

6.7 累积对比表

表 6.2: 方法演进对比表，截至第 6 章

方法	ATE 估计	95% CI	核心假设
回归调整	OR = 1.34	[1.18, 1.52]	模型设定正确 + 可交换性 + 正值性
G 计算	RD = 0.052	[0.027, 0.082]	结果模型设定正确 + 可交换性 + 正值性
IPW	RD = 0.032	[0.005, 0.064]	倾向得分模型正确 + 可交换性 + 正值性
AIPW	RD = 0.044	[0.017, 0.072]	两个模型至少一个正确 + 可交换性 + 正值性

四种方法在方向上完全一致：RHC 增加了 ICU 患者的 180 天死亡率。回归调整报告的是 OR 尺度，与后续方法的风险差尺度不直接可比，但 OR > 1 和 RD > 0 传递的结论相同。G 计算、IPW 和 AIPW 三个风险差估计都落在 3–5 个百分点之间。AIPW 的估计位于 G 计算和 IPW 之间，置信区间最为紧凑。这种跨方法的一致性增强了我们对因果结论的信心：RHC 的不利效应不太可能是某种特定方法的假象。

下一章将引入机器学习来增强 AIPW 的两个子模型。用 Super Learner 集成多种算法拟合结果模型和倾向得分模型，可以降低函数形式错误的风险，让双重稳健性的”保险”更加可靠。同时还会介绍 TMLE 和 DML，它们是 AIPW 在不同学术传统下的延伸和改进。

本章知识地图

表 6.3: 第 6 章核心概念与常见误解

核心概念	核心内容	常见误解	为什么错
AIPW 估计量	结果模型主估计 + IPW 残差校正项，三部分组合	AIPW 就是 G 计算和 IPW 的简单平均	AIPW 通过特定的数学形式让偏差项变成两个模型误差的乘积，简单平均做不到
双重稳健性	两个模型至少一个正确时估计一致	双重稳健意味着随便怎么建模都行	两个模型同时错误时 AIPW 也会有偏，”至少一个对”的前提仍需研究者努力保证
半参数效率界	两个模型都对时 AIPW 达到最小方差	AIPW 的方差总是最小的	只有一个模型对时 AIPW 保证一致性但方差未必最小
IPW 校正项	B 和 C 用倾向得分加权的残差修正结果模型的偏差	校正项只影响方差不影响偏差	当结果模型有偏时，校正项的期望值非零，正是它在消除偏差
极端倾向得分	AIPW 对极端权重的敏感性低于纯 IPW 但仍然存在	AIPW 完全不受极端倾向得分影响	极端权重放大残差，方差仍会膨胀，需要截断或 overlap weight

核心概念	核心内容	常见误解	为什么错
AIPW 的失效	两个模型同时错误或严重的正值性违反	双重稳健 = 万无一失	双保险的前提是至少一根绳子是好的，两根都断了就没有保护

第 7 章 机器学习增强——Super Learner、DML 与 TMLE

内容提要

- 理解参数模型在因果推断中的局限性以及机器学习替代的动机
- 掌握 Super Learner 的集成思想与交叉验证加权机制
- 理解 DML 的 Neyman 正交化与样本分裂如何消除正则化偏倚
- 理解 TMLE 的目标化更新步骤如何解决 AIPW 的出界问题
- 在 RHC 数据上分别运行 DML 和 TMLE，并与前六章的结果做累积对比

上一章手动实现了 AIPW，展示了双重稳健估计量的核心逻辑：结果模型和处理模型各拴一根保险绳，只要一根不断，估计就是一致的。但第 6 章的两个子模型都是 logistic 回归，函数形式写死了线性加法结构。如果 APACHE 评分和血压之间存在复杂的交互效应，或者某些协变量对死亡率的影响呈阶梯型而非线性，logistic 回归捕捉不到这些模式，两个模型就可能同时犯错。双重稳健性在两个模型都错的时候失效，第 6 章结尾已经预告了这个风险。

本章的解决思路是用机器学习替换参数模型。机器学习算法擅长捕捉非线性关系和高维交互效应，用它们来拟合 AIPW 的两个子模型，可以降低函数形式错误的概率，让" 至少一个模型对" 的保障更加可靠。但机器学习嵌入因果推断并非直接替换就行，正则化和过拟合会引入新的偏倚问题。本章介绍三个工具来解决这些问题：Super Learner 提供灵活的集成预测框架，Double Machine Learning，简称 DML，用 Neyman 正交化和样本分裂消除正则化偏倚，Targeted Maximum Likelihood Estimation，简称 TMLE，用目标化的最大似然更新步骤把估计拉回参数空间内。

7.1 参数模型的天花板

前六章的所有模型都有一个共同特征：函数形式由研究者预先指定。回归调整用 $Y = \beta_0 + \beta_1 A + \beta_2 L_1 + \cdots$，G 计算用 logistic 回归预测反事实，IPW 用 logistic 回归估计倾向得分。这些模型假设协变量对结局或处理的影响是线性加法的，经过 logit 链接函数变换后保持可加性。

真实的生物医学数据很少满足这个假设。RHC 数据集有 28 个协变量，它们之间的交互模式多达数百种。APACHE 评分在 20 分以下可能对死亡率影响较弱，超过某个阈值后影响陡然增大。肾功能不全和肝功能不全同时存在时的效应可能远大于两者分别存在时的加总。这些非线性和交互效应在 logistic 回归中都被忽略了，除非研究者手动加入平方项和交互项。但 28 个变量的二阶交互就有 378 项，全部放进模型既不现实也会过拟合。

机器学习算法可以自动学习非线性关系和交互效应，不需要研究者手动指定函数形式。随机森林通过递归分裂捕捉阈值效应和交互，Lasso 在高维空间中做变量选择，神经网络逼近任意连续函数。把这些算法用在 AIPW 的两个子模型上，可以降低函数形式设错的概率，让双重稳健性的" 两根保险绳" 都更牢固。

但直接替换会遇到两个技术障碍。机器学习的预测性能依赖于交叉验证和正则化，而正则化会引入系统性偏差。用同一份数据既训练模型又做因果估计，训练集上的过拟合残差会传递到最终的 ATE 估计里。这两个障碍分别由 DML 和 TMLE 给出了解决方案，但在介绍它们之前，需要先理解 Super Learner——因为 DML 和 TMLE 都可以用 Super Learner 作为底层的预测引擎。

7.2 Super Learner：集成学习的通用框架

单个机器学习算法各有所长。随机森林处理交互效应好但外推能力弱，Lasso 在高维稀疏场景下表现优秀但假设了线性结构，逻辑回归偏差低但灵活性不足。研究者面临的问题是：该用哪个？如果选错了算法，预测性能

就会打折扣，连带影响因果估计的质量。

Super Learner 的回答是：不选，全用。它把多个学习器组合成一个集成预测器，通过交叉验证给每个学习器分配最优权重，最终的预测是所有学习器的加权平均。

假设 library 里有三个算法：逻辑回归预测准确率 68%、随机森林 73%、Lasso 70%。SL 给随机森林最大权重 0.55，Lasso 0.30，逻辑回归 0.15，加权平均后准确率 74%。这就是 Super Learner 的基本逻辑：不选最好的单个算法，而是让所有算法按表现分配权重，集成后的预测优于任何单一算法。

定义 7.1 (Super Learner)

给定 K 个候选学习器 $\hat{f}_1, \ldots, \hat{f}_K$ 和一个损失函数 $\mathcal{L}$，Super Learner 的集成预测为

$$\hat{f}_{\mathrm{SL}}(x) = \sum_{k=1}^{K} \hat{\alpha}_k \, \hat{f}_k(x), \quad \hat{\alpha} = \arg\min_{\alpha \geq 0, \sum \alpha_k = 1} \sum_{i=1}^{n} \mathcal{L}\big(Y_i, \sum_k \alpha_k \hat{f}_k^{(-i)}(X_i)\big),$$

其中 $\hat{f}_k^{(-i)}$ 是第 k 个学习器在去掉第 i 个观测后的交叉验证预测值。 [16]

这个公式的核心思想是用交叉验证来评价每个学习器的真实预测能力，然后按表现分配权重。交叉验证避免了过拟合的干扰：每个观测的预测值都来自没有见过该观测的模型，所以预测误差是诚实的。表现好的学习器拿到大权重，表现差的权重趋近于零。

这个设计解决了”该选哪个算法”的纠结。van der Laan 等人在 2007 年证明了 Super Learner 具有渐近最优性：在候选学习器集合中，Super Learner 的预测风险渐近地不差于表现最好的那个单一学习器。换句话说，即使你不知道哪个算法最合适，Super Learner 至少和最佳选择一样好。这个性质叫做 oracle 性质。比如在 RHC 数据上，如果 library 里最好的单个算法 AUC 是 0.72，Super Learner 的 AUC 会接近或超过 0.72。

在 RHC 数据上，用 SuperLearner 包拟合结果模型可以直观看到权重分配。下面的代码把逻辑回归、Lasso、随机森林和均值基准四个学习器组合起来。

```r
set.seed(2026)
library(SuperLearner)

# 结果模型：预测 death180 在 rhc 和 28 个协变量条件下的概率
sl_out <- SuperLearner(
  Y = d$death180_bin,
  X = d |> select(rhc, all_of(covs)),
  family = binomial(),
  # 四个候选学习器：均值基准、逻辑回归、Lasso、随机森林
  SL.library = c("SL.mean", "SL.glm", "SL.glmnet", "SL.ranger"),
  cvControl = list(V = 5)   # 5 折交叉验证评估各学习器
)

# 查看各学习器的交叉验证风险和集成权重
sl_out
```

结果解读

Super Learner 的交叉验证结果显示，随机森林的预测风险最低，为 0.208，逻辑回归次之，为 0.212，Lasso 和逻辑回归几乎持平，均值基准的风险最高，为 0.250。最终的集成权重分配给了随机森林 0.686 和逻辑回归 0.314，Lasso 和均值基准的权重均为零。

这个权重分配传递了有用的信息：随机森林捕捉到了逻辑回归遗漏的非线性模式，但逻辑回归仍然保留了 31% 的权重，说明线性加法结构对这份数据的预测能力并非完全无用。Super Learner 的集成把两者的优势

结合起来，预测风险低于任何单一算法。

简而言之，Super Learner 让多个算法各做预测，按交叉验证表现分配权重，取加权平均作为最终预测。

笔记 Super Learner 的提出者是 Mark van der Laan 和 Eric Polley，发表于 2007 年 [16]。它的理论根基是交叉验证选择器的渐近最优性，在统计学习理论中有严格证明。在因果推断领域，Super Learner 通常作为底层工具嵌入 AIPW、DML 和 TMLE 的子模型中使用，让研究者不必在”选逻辑回归还是随机森林”的问题上做赌注。

停下来想一想：Super Learner 的权重是通过交叉验证分配的，表现好的算法拿大权重。如果 library 里所有算法表现都差不多，权重分配会是什么样？如果只放了一个算法，Super Learner 还有意义吗？理解了这两个极端情况，再往下看 DML 如何把 Super Learner 嵌入因果估计框架。

7.3 DML：正则化偏倚与 Neyman 正交化

Super Learner 提供了灵活的预测工具，但把它直接塞进 AIPW 公式并不能保证因果估计的有效性。问题出在正则化偏倚上。

机器学习算法为了避免过拟合，会对参数施加惩罚，把预测值往”保守方向”收缩。Lasso 把不重要的系数压到零，随机森林在分裂节点时做平滑。这种收缩在预测任务中是好事，但在因果估计中会制造系统性偏差。

用一个简化的例子来理解正则化偏倚的严重性。假设结果模型的偏差是 $b_m = 0.02$，倾向得分模型的偏差是 $b_e = 0.02$。如果把机器学习直接塞进普通的因果估计框架，最终估计的偏差正比于 $b_m + b_e = 0.02 + 0.02 = 0.04$。每个子模型的小偏差直接加到了 ATE 估计上，4% 的偏差在风险差只有 4–5 个百分点的场景中是致命的。

Neyman 正交化做的事情是把偏差的叠加方式从加法变成乘法。同样是 $b_m = 0.02$ 和 $b_e = 0.02$，正交化之后偏差变成 $0.02 \times 0.02 = 0.0004$，比加法的 0.04 缩小了 100 倍。这就是 DML 的核心技巧：不消除偏差，而是让两个偏差互相”抵消”到可以忽略的量级。

下面给出正式定义。

定义 7.2 (Neyman 正交化)

一个关于目标参数 θ 的矩条件 $\psi(W;\theta,\eta)$ 称为 Neyman 正交的，如果它对辅助参数 η 的 Gateaux 导数在真值处为零：

$$\frac{\partial}{\partial t} E\big[\psi(W;\theta_0,\eta_0 + t(\eta-\eta_0))\big]\Big|_{t=0} = 0.$$

[7]

这个数学条件的实际含义就是前面数字例子展示的效果：辅助模型的偏差对目标参数的影响从一阶降到了二阶。机器学习的收缩偏差通常以 $n^{-1/4}$ 到 $n^{-1/3}$ 的速度趋近于零，乘积之后达到 $n^{-1/2}$，恰好满足渐近正态推断所需的速度。

AIPW 的得分函数恰好满足 Neyman 正交条件。第 6 章的公式中，偏差项正比于 $E[(\hat{e}-e)(\hat{m}-m)]$，这就是一个乘积。所以 AIPW 本身已经具备了 Neyman 正交性。但还有一个问题没解决：如果用同一份数据既训练机器学习模型又计算 AIPW 得分，训练集上的过拟合会让残差偏小，乘积项的期望不再为零。

定义 7.3 (样本分裂)

将样本随机分成两半。用前一半训练辅助模型 $\hat{\eta}$，用后一半计算因果估计的得分 $\psi(W_i;\theta,\hat{\eta})$。训练数据和估计数据完全分开。

样本分裂的逻辑类似于考试中的”出题人不批自己的卷子”。训练模型的数据和计算因果估计的数据分开，切断了过拟合通道。

定义 7.4 (交叉拟合)

交叉拟合是样本分裂的改进版。将样本随机等分为 K 折。对第 k 折中的每个观测 i，用剩余 $K-1$ 折训练辅助模型 $\hat{\eta}^{(-k)}$，然后在第 k 折上计算得分 $\psi(W_i;\theta,\hat{\eta}^{(-k)})$。最终估计为所有折得分的均值。

交叉拟合的好处是每一折数据都轮流充当"估计集"，避免了简单分裂只用一半数据做估计导致的效率损失。

把 Neyman 正交化和交叉拟合结合起来，就是 DML 的完整框架。Chernozhukov et al. [7] 在 2018 年的论文中系统地提出了这套方法，证明了在机器学习估计辅助参数的条件下，DML 估计量是 $\sqrt{n}$ 一致且渐近正态的。这篇论文发表在 *The Econometrics Journal*，在计量经济学界产生了深远影响，因为它给出了一套在机器学习年代仍然合法的因果推断框架。

定理 7.1 (雷区)

跳过样本分裂直接把机器学习塞进 AIPW，在有限样本中几乎一定会有偏。偏差的来源是过拟合残差：机器学习模型在训练集上的残差系统性地偏小，导致 AIPW 得分中的校正项 $Y_i - \hat{m}(L_i)$ 低估了真实残差。这个偏差在 Lasso 等正则化方法中尤为严重，因为正则化把系数往零的方向收缩，训练集上的残差比测试集上的更小。诊断方法是把 DML 的 `n_folds` 从 5 改成 1，比较分裂和不分裂的估计差异。如果差异显著，说明过拟合偏倚在当前数据中是实质性的。

7.4 DML 在 RHC 数据上的实现

本章继续使用第 1 章介绍的 RHC 数据集 $n = 5735$。DML 的 R 实现使用 DoubleML 包 [5]，它基于 mlr3 生态系统，支持多种机器学习后端。下面的代码用随机森林作为两个子模型的学习器，5 折交叉拟合，重复 3 次取平均以增加稳定性。

```r
set.seed(2026)
library(DoubleML); library(mlr3); library(mlr3learners)
library(data.table)

# DoubleML 要求 data.table 格式
dt <- as.data.table(d |> select(death180_bin, rhc, all_of(covs)))

dml_data <- DoubleMLData$new(
  data = dt,
  y_col = "death180_bin",    # 结局
  d_cols = "rhc",            # 处理
  x_cols = covs              # 28 个协变量
)

# IRM: Interactive Regression Model, 适用于二分类处理
# ml_g 估计 E[Y|X], ml_m 估计 P(A=1|X)
dml_irm <- DoubleMLIRM$new(
  data = dml_data,
  ml_g = lrn("classif.ranger", predict_type = "prob", num.trees = 500),
  ml_m = lrn("classif.ranger", predict_type = "prob", num.trees = 500),
  score = "ATE",
  n_folds = 5,     # 5 折交叉拟合
  n_rep = 3        # 重复 3 次取平均，减小随机分裂的波动
)
```

```
25
26 dml_irm$fit()
27 dml_irm$summary()
28 print(dml_irm$confint())
```

结果解读

DML 估计的边际风险差为 0.040，标准误 0.013，$t = 3.03$，$p = 0.002$，95% CI 为 $[0.014, 0.065]$。置信区间不包含零，结论与前六章一致：接受 RHC 的患者 180 天死亡率显著高于不接受者，效应大小约 4.0 个百分点。

DML 的估计 0.040 比手动 AIPW 的 0.044 略低，接近 IPW 的 0.032。标准误 0.013 与 AIPW 的 0.014 相近。3 次重复交叉拟合的设计让估计对随机折划分的敏感性降低，增加了结果的可信度。

简而言之，DML 在 AIPW 的基础上加两个技巧：Neyman 正交化让偏差从加法变乘法，样本分裂让机器学习的过拟合不污染因果估计。

7.5 TMLE：从预测到目标化估计

DML 用 Neyman 正交化加样本分裂解决了正则化偏倚，但它的估计量仍然是 AIPW 形式的线性组合，没有内置参数空间约束。二分类结局的风险差应该在 $[-1, 1]$ 之间，AIPW 和 DML 都可能在极端情况下给出超出这个范围的估计。第 6 章结尾提到的” 出界问题” 在小样本或倾向得分极端的数据中并不罕见。

TMLE 的设计动机就是解决出界问题，同时保留双重稳健性和半参数效率。它的核心思想是：先用机器学习拿到一个初始估计，然后通过一步受约束的最大似然更新把初始估计” 微调” 到正确的目标参数上。用瞄准的比喻来说，初始估计确定了大致方向，目标化更新步骤是轻微转动瞄准镜让十字线对准靶心，整个过程在似然函数的框架内完成，自动保证估计值不会跑出合理范围。

TMLE 的” 微调” 具体是这样操作的：拿初始预测值和倾向得分，构造一个特殊的变量，用它做一个小小的 logistic 回归来修正初始预测。因为修正在 logit 尺度上完成，修正后的预测永远在 0 到 1 之间，不会像 AIPW 那样越界。整个修正只估计一个参数 ϵ，计算量极小，但效果是把初始预测” 推” 向正确的因果估计方向。

定义 7.5 (TMLE 的目标化更新)

TMLE 的更新步骤在初始估计 $\hat{Q}^0(A, L)$ 的基础上拟合一个浮动参数 ϵ：

$$\operatorname{logit} \hat{Q}^1(A, L) = \operatorname{logit} \hat{Q}^0(A, L) + \epsilon \cdot H(A, L),$$

ϵ 通过最大似然估计拟合，使得更新后的 $\hat{Q}^1$ 满足 AIPW 得分函数的零偏条件。 [17]

这个公式需要拆解。$\hat{Q}^0(A, L)$ 是结果模型的初始预测，可以来自 Super Learner 或任何机器学习算法。目标化更新的核心思想是在 logit 尺度上对初始预测做一步微调，让估计朝着正确的因果参数方向移动。因为修正在 logit 尺度上完成，更新后的预测值 $\hat{Q}^1$ 仍然在 $[0, 1]$ 之间，解决了出界问题。ϵ 只是一个标量，通过标准的 logistic 回归拟合，计算量极小。

公式中的 $H(A, L) = \frac{A}{\hat{e}(L)} - \frac{1-A}{1-\hat{e}(L)}$ 称为聪明协变量。它的形式来自半参数效率理论中的有效影响函数，作用是引导更新方向朝着目标参数 ATE 移动。处理组个体的 H 值为 $1/\hat{e}(L)$，对照组为 $-1/(1-\hat{e}(L))$，倾向得分越极端的个体，H 的绝对值越大，更新步骤对它们的修正幅度也越大。

TMLE 的双重稳健性来自更新步骤的设计。更新后的 $\hat{Q}^1$ 满足 $\frac{1}{n}\sum_i H(A_i, L_i)[Y_i - \hat{Q}^1(A_i, L_i)] = 0$，这恰好是 AIPW 得分函数的零偏条件。如果倾向得分模型正确，$H(A, L)$ 构造正确，这个条件保证偏差为零，即使初始结果模型有偏。如果初始结果模型正确，$\hat{Q}^0$ 已经接近真值，ϵ 会很小，更新几乎不改变估计，偏差仍然为零。

Laan and Rubin [17] 在 2006 年提出 TMLE，后续 Gruber and Laan [12] 进一步发展了它的实用版本。TMLE 在

流行病学领域获得了广泛采纳，因为它的表述更接近统计学家熟悉的”影响函数加似然”语言，而且 tmle 包提供了开箱即用的实现。DML 在计量经济学界更受欢迎，因为 Neyman 正交化和样本分裂的表述与计量经济学训练更契合。两者解决的是同一个问题：如何在使用机器学习估计辅助参数的同时，保证目标参数的渐近推断有效。

7.6 TMLE 在 RHC 数据上的实现

TMLE 的 R 实现使用 tmle 包，它内置了 Super Learner 接口，可以直接指定学习器库。下面的代码用逻辑回归、Lasso、随机森林和均值基准四个学习器拟合结果模型和倾向得分模型。

```
set.seed(2026)
library(tmle); library(SuperLearner)

SL_lib <- c("SL.glm", "SL.glmnet", "SL.ranger", "SL.mean")

tmle_fit <- tmle(
  Y = d$death180_bin,
  A = d$rhc,
  W = d |> select(all_of(covs)),
  Q.SL.library = SL_lib,    # 结果模型用 Super Learner
  g.SL.library = SL_lib,    # 倾向得分模型也用 Super Learner
  family = "binomial"       # 二分类结局
)

tmle_fit
```

结果解读

TMLE 估计的边际风险差为 0.088，标准误 0.0075，95% CI 为 $[0.074, 0.103]$。处理组的边际死亡概率估计为 0.558，对照组为 0.469。置信区间远离零，$p < 0.001$，结论同样是 RHC 增加了 180 天死亡率。

TMLE 的点估计 0.088 比 DML 的 0.040 和 AIPW 的 0.044 都大。这个差异有两层原因。TMLE 的目标化更新步骤在 logit 尺度上操作，更新方向和幅度受聪明协变量 $H(A, L)$ 引导，与 DML 直接在线性尺度上做残差校正的路径不同，有限样本中两者可以给出不同的点估计。TMLE 的标准误 0.0075 也比 DML 的 0.013 小了近一半，这反映了 TMLE 在二分类结局上利用了似然结构的额外效率。但更窄的置信区间也意味着如果模型设定存在问题，TMLE 的覆盖率可能偏低。

两种方法的差异提供了有价值的敏感性信息。如果 DML 和 TMLE 给出一致的估计，我们对结论的信心更强。当两者产生分歧时，应该检查学习器的配置是否一致、倾向得分的极端值处理是否相同、交叉拟合的折数是否匹配。在本例中，DML 使用纯随机森林而 TMLE 使用 Super Learner 集成，学习器配置的差异是分歧的一个来源。

简而言之，TMLE 先用机器学习做初始预测，然后用一步 logistic 回归微调，让估计既满足双重稳健性又不会跑出 0 到 1 的合理范围。

定理 7.2 (雷区)

DML 和 TMLE 在同一份数据上给出不同的点估计是正常现象，原因是两者的目标化路径不同。DML 在线性尺度上用 Neyman 正交得分做残差校正，TMLE 在 logit 尺度上做最大似然更新。当样本量足够大且两个子模型都拟合得很好时，两者会收敛到同一个值。但在有限样本中，以下因素可能导致分歧：学习器配置不同、交叉验证折数不同、倾向得分截断方式不同。遇到分歧时的诊断策略是固定学习器和折数设

置，让两种方法在尽可能一致的条件下比较，剩余的差异才可归因于方法本身的有限样本行为。

7.7 DML 与 TMLE 在同一数据上的对比

图 7.1 把前六章的参数方法和本章的两种 ML 增强方法放在一起。五种方法的点估计全部落在正值区域，方向完全一致：RHC 增加了 ICU 患者的 180 天死亡率。

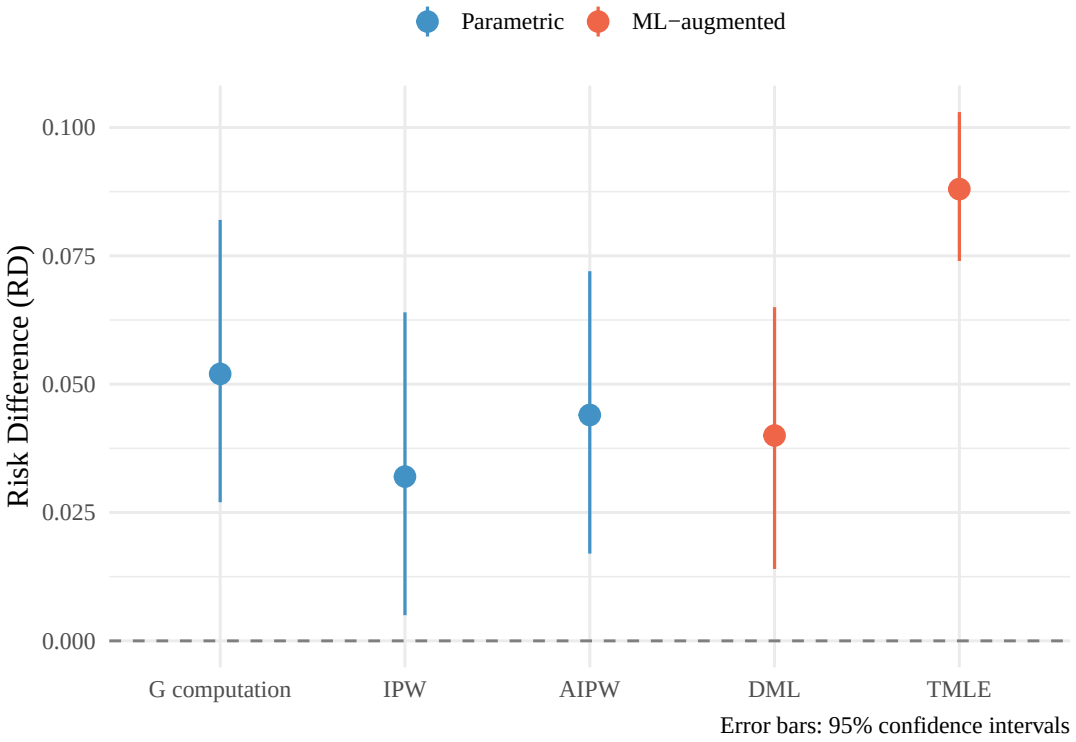

图 7.1: 五种方法的 ATE 估计及 95% CI 对比。蓝色为参数模型方法，红色为 ML 增强方法。所有估计方向一致：RHC 增加死亡风险。DML 的估计 0.040 接近参数方法的范围，TMLE 的估计 0.088 相对较高但置信区间更窄。

参数方法的三个估计落在 $[0.032, 0.052]$ 之间，DML 的 0.040 也在这个范围内。TMLE 的 0.088 高出一倍，但置信区间非常紧凑。五种方法尽管点估计不同，但所有置信区间都不包含零，定性结论高度一致。方法之间的数值差异恰好说明了因果估计对建模选择的敏感性，这也是第 8 章敏感性分析要系统探讨的问题。

命题 7.1 (DML 与 TMLE 的适用场景)

DML 更适合以下场景：协变量维度很高、处理变量可能是连续的、研究者熟悉计量经济学的矩方法框架。TMLE 更适合以下场景：结局是有界的、研究者需要估计值严格落在合理范围内、样本量中等且需要利用似然结构的效率增益。在二分类处理和二分类结局的标准设定下，两者都适用，选择更多取决于学科传统和软件习惯。

7.8 累积对比表

表 7.1: 方法演进对比表，截至第 7 章

方法	ATE 估计	95% CI	核心假设
回归调整	OR = 1.34	[1.18, 1.52]	模型设定正确 + 可交换性 + 正值性
G 计算	RD = 0.052	[0.027, 0.082]	结果模型设定正确 + 可交换性 + 正值性
IPW	RD = 0.032	[0.005, 0.064]	倾向得分模型正确 + 可交换性 + 正值性
AIPW	RD = 0.044	[0.017, 0.072]	两个模型至少一个正确 + 可交换性 + 正值性
DML	RD = 0.040	[0.014, 0.065]	Neyman 正交 + 交叉拟合 + 可交换性 + 正值性
TMLE	RD = 0.088	[0.074, 0.103]	目标化更新 + SL + 可交换性 + 正值性

六种方法在方向上完全一致：RHC 增加了 ICU 患者的 180 天死亡率。回归调整报告的是 OR 尺度，与后续方法的风险差尺度不直接可比，但 OR > 1 和 RD > 0 传递的结论相同。参数方法的三个风险差估计都集中在 3–5 个百分点，DML 的 4.0 个百分点也在这个范围内。TMLE 的 8.8 个百分点偏高，但它的置信区间最窄。

跨方法的一致性进一步增强了因果结论的可信度。每种方法依赖不同的假设和建模路径，但都指向同一个方向，说明 RHC 的不利效应不太可能是某种特定建模选择的产物。下一章将用敏感性分析来量化” 如果存在未测量的混杂，需要多强才能推翻这个结论”，给这张累积对比表加上最后一层稳健性检验。

本章知识地图

表 7.2: 第 7 章核心概念与常见误解

核心概念	核心内容	常见误解	为什么错
Super Learner	交叉验证加权的集成预测器，具有 oracle 最优性	Super Learner 可以直接输出因果效应	Super Learner 只是预测工具，因果识别仍需 AIPW/DML/TMLE 的框架
正则化偏倚	机器学习的收缩惩罚在因果估计中引入系统性偏差	预测越准因果估计就越准	预测准确度和因果估计的无偏性是两个不同目标，正则化为前者牺牲了后者
Neyman 正交化	让辅助模型偏差以乘积形式进入，从加法 0.04 降到乘积 0.0004	正交化让偏差消失	正交化让偏差从一阶降到二阶，偏差仍然存在只是量级大幅缩小
样本分裂与交叉拟合	训练和估计用不同子样本，切断过拟合传导通道	分裂损失了一半数据的效率	交叉拟合让每个观测都参与估计，效率损失可忽略

核心概念	核心内容	常见误解	为什么错
TMLE 目标化更新	在 logit 尺度上用聪明协变量做一步最大似然修正	TMLE 是一种全新的估计方法	TMLE 在渐近层面等价于 AIPW，区别在于有限样本中的参数空间约束和效率
DML vs TMLE 分歧	两者在有限样本中可能给出不同点估计	分歧意味着某种方法错了	分歧来自目标化路径和学习器配置的差异，大样本下两者收敛

第 8 章 结果稳不稳——敏感性分析与未测量混杂

内容提要

- 理解可交换性假设为什么无法用数据验证
- 掌握敏感性分析的核心逻辑：未测量混杂需要多强才能推翻结论
- 在 RHC 数据上计算 E-value 并结合 ICU 领域知识解读
- 用 sensemakr 做遗漏变量偏差分析，以 APACHE 评分为基准评估稳健性
- 建立”敏感性分析必须搭配领域知识”的分析习惯

前面七章用了多种方法回答同一个问题，结论一致：RHC 增加了 180 天死亡率，风险差在 3–5 个百分点之间。这个一致性让人安心，但它建立在一个共同前提上：我们控制的 28 个协变量足以消除处理组和对照组之间的系统性差异。换句话说，所有方法都假设不存在未测量的混杂变量。

这个假设合理吗？医生决定是否插 RHC 时，除了 APACHE 评分、血压、肌酐这些数据里有的指标，还会参考床旁即时反应、家属意愿、设备空闲情况。这些因素都没有记录。如果其中某个因素同时影响了医生的决策和患者的存活，所有估计都会偏移。本章不产出新的 ATE 估计，它做的事情是压力测试：前面得到的结论有多经得起”万一还有没控制住的混杂”这个质疑？

8.1 不可检验的假设：可交换性

第 2 章引入了因果推断的三大识别假设：可交换性、正值性、一致性。正值性可以通过检查倾向得分分布来诊断，一致性是对处理定义的逻辑约束，这两者都有经验层面的检验手段。可交换性是个例外。

比如在 APACHE 评分都是 70 分的病人中，上导管和不上导管的人，后来死不死应该和他们是否上了导管没有关系，因为 APACHE 70 这一组里的分配已经”近似随机”了。

定义 8.1 (条件可交换性)

给定协变量集合 L，如果潜在结果 $Y(a)$ 与处理指派 A 条件独立，即

$$Y(a) \perp\!\!\!\perp A \mid L, \quad \forall a,$$

则称在 L 条件下处理与结局满足可交换性。 [22]

这个假设说的是：在协变量 L 的每一层内，处理组和对照组的潜在结果分布相同。它之所以无法检验，是因为我们永远只能观测到一个人在实际处理状态下的结局，看不到他在另一种处理状态下会怎样。要验证可交换性，需要比较”处理组的 $Y(0)$”和”对照组的 $Y(0)$”，但处理组的 $Y(0)$ 是反事实，数据里不存在。

在 RCT 中，随机化保证了可交换性。观察性研究没有这个保障，研究者只能尽量把想得到的混杂变量都放进 L，但”L 是否足够全面”无法从数据中证实。这就是敏感性分析的出发点：既然无法证明可交换性成立，那就换一个问题——**如果可交换性被违反了，违反需要多严重才能推翻我们的结论？**

8.2 敏感性分析的思路

敏感性分析的核心逻辑是反向思考。传统分析假设没有未测量混杂，直接估计 ATE。敏感性分析反过来问：假设存在一个未测量混杂 U，U 需要和处理与结局各有多强的关联，才能把 ATE 估计拉到零？如果答案是”U 需要非常强才能翻盘”，结论就比较稳健。这个思路由 Rosenbaum [22] 系统化，本章介绍两种更现代的工具：E-value 和 sensemakr。

8.3 E-value：需要多强的混杂才能翻盘

E-value 由 VanderWeele and Ding [27] 提出，它把” 未测量混杂需要多强” 翻译成一个相对风险的尺度。

先用一个具体例子建立直觉。假设我们估计出 RHC 增加死亡率 RR = 1.34。E-value 回答的问题是：一个我们没有测量到的混杂变量，需要多强才能把这个 1.34 完全解释掉？答案是 E-value = 1.42。意思是这个混杂变量需要同时和 RHC 使用、180 天死亡分别有至少 1.42 倍的关联。如果找不到这么强的遗漏因素，1.34 的效应就不太可能全靠混杂偏差来解释。

下面是 E-value 的正式定义。

定义 8.2 (E-value)

对于一个观测到的相对风险 $\mathrm{RR}_{\mathrm{obs}}$，其 E-value 定义为

$$E\text{-value} = \mathrm{RR}_{\mathrm{obs}} + \sqrt{\mathrm{RR}_{\mathrm{obs}} \times (\mathrm{RR}_{\mathrm{obs}} - 1)}.$$

E-value 表示的是：未测量混杂 U 与处理的关联以及 U 与结局的关联至少都要达到这个 RR 值，才能把观测到的效应完全归因于混杂。 [27]

公式里的 $\mathrm{RR}_{\mathrm{obs}}$ 是调整了已知混杂之后的相对风险。E-value 告诉我们，未测量因素 U 与 RHC 使用和 180 天死亡的关联强度至少要多大，才能让观测到的不利效应完全消失。E-value 越大，翻盘需要的混杂越强，结论越稳健。

8.4 在 RHC 数据上计算 E-value

本章继续使用第 1 章介绍的 RHC 数据集 $n = 5735$。AIPW 在第 6 章给出的风险差估计为 0.044，95% CI 为 $[0.017, 0.072]$。要计算 E-value，需要把风险差转换为相对风险的尺度。对照组的 180 天死亡率为 0.465，据此可得调整后的近似 RR $\approx (0.465 + 0.044)/0.465 = 1.095$。

```
set.seed(2026)
library(tidyverse)
library(EValue)

d <- read_csv(here::here("data", "rhc.csv"), show_col_types = FALSE) |>
  mutate(death180_bin = if_else(death180 == "Yes", 1L, 0L))

# 对照组基线死亡率
r0 <- mean(d$death180_bin[d$rhc == 0])

# AIPW 估计的风险差和 95% CI（来自第 6 章）
ate  <- 0.0442
se   <- 0.0139
ci_lo <- ate - 1.96 * se
ci_hi <- ate + 1.96 * se

# 转换为相对风险尺度——E-value 需要 RR 作为输入
rr_point <- (r0 + ate) / r0
rr_lo    <- (r0 + ci_lo) / r0
rr_hi    <- (r0 + ci_hi) / r0

```

```r
# 计算 E-value：点估计和置信区间下界各一个
ev <- evalues.RR(rr_point, lo = rr_lo, hi = rr_hi)
print(ev)
```

结果解读

点估计的 E-value 为 1.42，置信区间下界的 E-value 为 1.23。
点估计 E-value = 1.42 的含义是：一个未测量的混杂因素 U，需要与 RHC 使用和 180 天死亡**各自**的关联强度都达到 RR = 1.42 以上，才能把 AIPW 观测到的风险差 0.044 完全解释为混杂偏差。置信区间下界的 E-value = 1.23 更为关键，它意味着 U 只需要与两端各达到 RR = 1.23，就能让效应的统计显著性消失。
在 ICU 里，能和死亡率产生 1.42 倍关联的变量是什么量级？APACHE 评分与死亡率的关联在 RR = 2–3 之间，血压和肌酐在 RR = 1.5–2.5 之间。这些变量我们都已经控制了。一个残余的、没被这 28 个变量捕捉到的因素，要同时和 RHC 使用及死亡率各达到 RR = 1.42，在临床上并非不可能，但需要一个相当有影响力的遗漏因素。

8.5 E-value 的领域锚定

1.42 孤立地看没有意义，E-value 的价值在于和领域内已知混杂因素的强度做对比。Connors, Speroff, Dawson, et al. [9] 报告了 APACHE 评分与 RHC 使用的关联约为 RR = 2.0，与 180 天死亡率的关联更强，高评分组死亡风险是低评分组的 2–3 倍。血压、肌酐等协变量与两端的关联在 RR = 1.5–2.5 之间。

我们已经控制了 APACHE 评分、血压、肌酐等 28 个变量，一个**残余的**未测量混杂要同时在两端都达到 RR = 1.42 的关联强度，可能性并不高但也不能排除。如果 E-value 达到 3 或 4，结论通常被认为非常稳健；1.42 处于中等偏低的位置，保护有限。

到这里我们有了第一个工具 E-value。它给出了一个数字门槛。接下来要介绍的 sensemakr 会进一步用已观测的协变量作为参照物，让这个门槛更加具体。

同样的 E-value 在不同研究领域意味着完全不同的事情。在药物临床试验中，已知混杂因素的 RR 通常在 1.2–1.5 之间，E-value = 1.42 已经接近上限，结论比较安全。在社会科学中，未测量混杂 RR 可达 3–5，同样的 E-value 几乎提供不了保护。ICU 医学介于两者之间：已知的强混杂如疾病严重程度 RR 可达 2–3，但已被控制。残余混杂要达到 1.42 需要一个相当有影响力的遗漏因素。

定理 8.1 (雷区)

E-value 最常见的误用是把它当作一个通用的”稳健性评分”，脱离领域知识直接下结论。比如看到 E-value = 2.5 就说”结论很稳健”，看到 1.3 就说”结论脆弱”。这样做的问题在于，E-value 衡量的是未测量混杂需要多强，而”多强算强”取决于具体研究领域。在一个变量关联普遍较弱的领域，1.3 已经很难达到；在一个混杂无处不在的领域，2.5 可能轻而易举。正确的做法是：计算 E-value 之后，列出本研究领域中已知混杂因素与处理和结局的关联强度，用这些数字作为参照系来判断 E-value 的含义。没有领域锚定的 E-value 解读是空洞的。 ♡

8.6 sensemakr：遗漏变量偏差的等高线

Cinelli and Hazlett [8] 提出的 sensemakr 框架用偏 R^2 来刻画未测量混杂解释处理变异和结局变异的比例，并允许研究者用已观测的协变量作为基准校准”多强算强”。sensemakr 的核心概念是稳健值，英文称 Robustness Value，简称 RV。

RV 假设的是一个最坏情况的混杂：它和我们已经测量的 28 个变量都不相关，是一个全新的、我们完全没

有捕捉到的因素。这比” 已有变量遗漏了某些信息” 更极端，因为如果新混杂和已有变量有相关性，已有变量多少能帮我们控制一部分。RV 不给你这个好处，它直接问：在最坏的情况下，这个全新因素需要解释处理和结局残差方差的多大比例，才能把效应估计拉到零?

定义 8.3 (稳健值 RV)

稳健值 RV 回答的问题是：一个未测量混杂需要多强，才能把我们估计到的效应拉低到可以忽略的程度？RV 越大，需要的混杂越强，结论越稳健。

具体而言，RV 用偏 R^2 来衡量混杂的强度，即未测量混杂对处理和结局残差方差各自的解释比例。$\mathrm{RV}_{q=1}$ 是使效应估计降至零所需的最小偏 R^2。q 代表效应缩减的比例，$q=1$ 意味着完全归零。$\mathrm{RV}_{q=1,\alpha=0.05}$ 则是使效应估计的置信区间包含零所需的最小强度，这个阈值更低，因为只需让显著性消失即可。 [8]

sensemakr 还提供了基准校准功能：选择一个已观测的协变量作为参照，问” 如果未测量混杂的强度是 APACHE 评分的 k 倍，效应估计会变成多少”。研究者通常对” 比 APACHE 评分强两倍的混杂因素是否存在” 有直觉判断，这比纯数字输出更贴近实务决策。

8.7 sensemakr 在 RHC 数据上的应用

sensemakr 需要线性模型作为输入。线性概率模型在这个样本量下的系数估计与 logistic 回归的边际效应接近，且偏 R^2 分解在线性框架下有严格保证。

```
set.seed(2026)
library(tidyverse)
library(sensemakr)

d <- read_csv(here::here("data", "rhc.csv"), show_col_types = FALSE) |>
  mutate(death180_bin = if_else(death180 == "Yes", 1L, 0L),
         sex_bin      = if_else(sex == "Male", 1L, 0L),
         cancer_bin   = if_else(cancer == "No", 0L, 1L))

covs <- c("age", "sex_bin", "cancer_bin", "cardiovascular",
          "congestive_hf", "dementia", "psychiatric", "pulmonary",
          "renal", "hepatic", "gi_bleed", "tumor",
          "immunosupperssion", "transfer_hx", "mi",
          "apache_score", "glasgow_coma_score", "blood_pressure",
          "heart_rate", "respiratory_rate", "temperature",
          "albumin", "creatinine", "bilirubin", "wbc",
          "hematocrit", "das_index", "weight")

# 线性概率模型——sensemakr 需要 lm 对象
lin_mod <- lm(death180_bin ~ rhc + .,
              data = d |> select(death180_bin, rhc, all_of(covs)))

# 以 APACHE 评分为基准，分别看 1 倍、2 倍、3 倍强度的混杂
sens <- sensemakr(model = lin_mod,
                  treatment = "rhc",
                  benchmark_covariates = "apache_score",
                  kd = c(1, 2, 3))
summary(sens)
```

结果解读

线性模型中 RHC 的系数为 0.053，标准误 0.014，$t = 3.86$，$p < 0.001$。处理变量对结局的偏 R^2 仅为 0.26%，说明在控制了 28 个协变量之后，RHC 对死亡率解释的独立变异很小。

稳健值 $\text{RV}_{q=1} = 5.0\%$：一个与已控制协变量正交的未测量混杂，需要同时解释处理和结局残差方差的至少 5.0%，才能把效应估计拉到零。$\text{RV}_{q=1,\alpha=0.05} = 2.5\%$：同样的混杂只需解释 2.5% 的残差方差，就能让置信区间包含零，统计显著性消失。

以 APACHE 评分为基准的校准结果特别有信息量。APACHE 评分对 RHC 使用的偏 R^2 为 2.0%，对死亡率的偏 R^2 为 1.0%。如果未测量混杂的强度和 APACHE 评分相当，即 1× apache_score，调整后的效应估计从 0.053 下降到 0.038，仍然显著。如果未测量混杂有 APACHE 评分的 2 倍强度，效应降到 0.023，置信区间的下界接近零。到了 3 倍 APACHE 评分的强度，效应仅剩 0.008，已经不再显著。

在 ICU 里，比 APACHE 评分还强两三倍的遗漏变量意味着什么？APACHE 评分本身已经综合了生理参数、年龄和慢性健康状况，是 ICU 预后预测中解释力最强的单一变量之一。一个强度达到 APACHE 两倍的遗漏因素在临床上很难想象。但让显著性消失的阈值只需要 2 倍，而医生的主观判断、床旁快速恶化程度这类未记录因素并非不可能达到这个水平。

图 8.1 把这些数字关系可视化了。横轴是未测量混杂对处理的偏 R^2，纵轴是对结局的偏 R^2，红色虚线是零等高线。三个红色菱形分别标记了 1 倍、2 倍、3 倍 APACHE 评分强度的位置，3 倍的菱形已经非常接近零等高线。

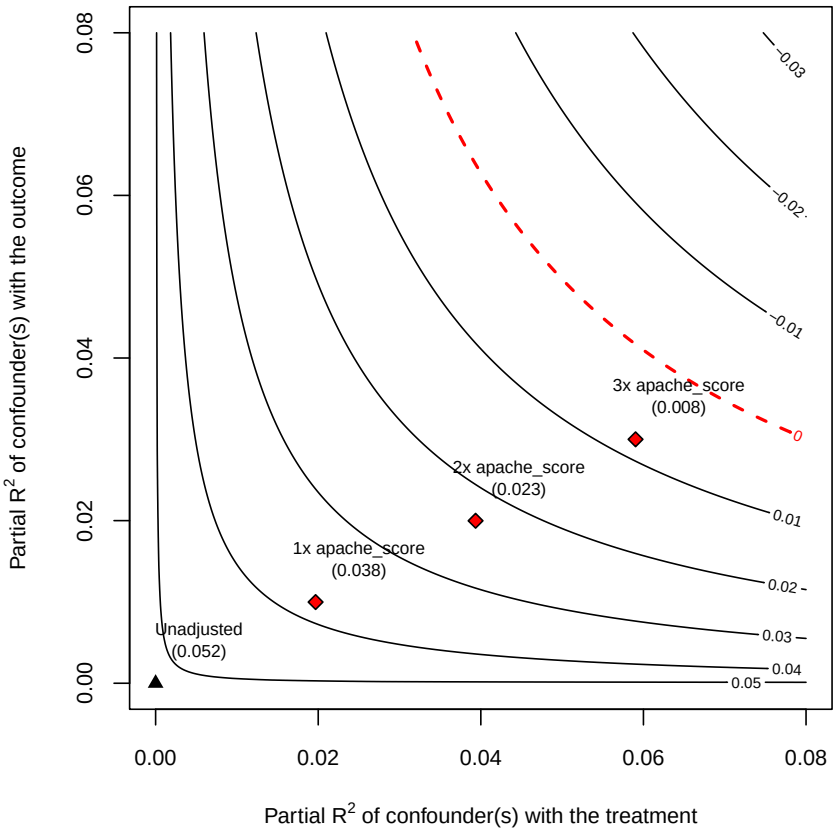

图 8.1: sensemakr 等高线图。每条曲线代表一个调整后效应估计值，红色虚线为零等高线。红色菱形标记了以 APACHE 评分为基准的 1 倍、2 倍、3 倍混杂强度位置。黑色三角为未调整的原始估计 0.052。

8.8 两种方法的对照解读

E-value 和 sensemakr 从不同角度量化了同一个问题。E-value = 1.42 说的是未测量混杂与处理和结局的关联各需达到 RR = 1.42 才能翻盘。sensemakr 的基准校准则更具体：3 倍 APACHE 评分强度的混杂才能让效应消失，2 倍就能让显著性消失。两种工具给出的图景是一致的：翻盘需要一个不算弱但也谈不上极强的遗漏因素。

综合来看，RHC 增加死亡率的结论对未测量混杂有一定抵抗力，但保护程度中等。效应本身不大，RR 约为 1.10，小效应天然更容易被残余混杂解释掉。这也是 Connors, Speroff, Dawson, et al. [9] 原文发表后引发持续争论的原因之一。

命题 8.1 (E-value 与 sensemakr 的适用场景)

E-value 适合快速报告：只需要点估计和置信区间就能计算，结果是一个直观的 RR 数字，适合写在论文摘要里。它的局限是缺乏内部校准，需要研究者自行提供领域参照。sensemakr 适合深入探索：它用已观测协变量的解释力作为基准，输出等高线图和分阶梯的调整后估计，让读者直观看到"混杂强到什么程度结论才翻转"。它的局限是需要线性模型作为输入，且偏 R^2 对非线性关系的捕捉能力有限。实际操作中，两者配合使用效果最好：E-value 给出汇总判断，sensemakr 给出细节和可视化。

定理 8.2 (雷区)

敏感性分析不能证明"没有未测量混杂"。它做的事情是量化"未测量混杂需要多强才能推翻结论"，本质上是一种压力测试。有时研究者在论文里写"E-value = 3.2，说明未测量混杂不太可能影响我们的结论"，这个表述在逻辑上是滑坡的。E-value = 3.2 只能说明"推翻结论需要一个很强的混杂"，但无法排除这种强混杂确实存在的可能。正确的表述是："未测量混杂需要与处理和结局各达到 RR = 3.2 的关联才能解释掉观测到的效应，在本研究领域中，已知混杂因素的关联强度为 RR = X–Y，因此残余混杂达到此水平的可能性较低/中等/较高。"判断"可能性高低"是领域知识的工作，E-value 只提供数字框架。

练习 8.1 第 5 章的 IPW 估计给出了风险差约 0.032，标准误约 0.022。计算 IPW 估计对应的 E-value，与 AIPW 的 E-value = 1.42 做比较。较小的效应估计是否意味着更脆弱的结论？结合 ICU 领域知识解释你的判断。

解

```r
library(EValue)
# IPW 的风险差和标准误
ate_ipw <- 0.032; se_ipw <- 0.022
r0 <- 0.4647  # 对照组基线死亡率

# 转换为 RR 尺度
rr_ipw <- (r0 + ate_ipw) / r0
rr_lo  <- (r0 + ate_ipw - 1.96 * se_ipw) / r0

ev_ipw <- evalues.RR(rr_ipw, lo = rr_lo)
print(ev_ipw)
```

IPW 的点估计 RR $\approx$ 1.069，对应 E-value 约为 1.31，比 AIPW 的 1.42 更小。置信区间下界的 E-value 接近 1.0，极弱的未测量混杂就能让显著性消失。IPW 的效应估计更小、标准误更大，对未测量混杂的抵抗力更弱。但两种方法的效应方向一致，跨方法的方向一致性本身也是稳健性的证据。

方法卡片：敏感性分析

分析目标：评估因果效应估计对未测量混杂的稳健程度，不产出新的 ATE。
E-value：$E = \mathrm{RR}_{\mathrm{obs}} + \sqrt{\mathrm{RR}_{\mathrm{obs}} \times (\mathrm{RR}_{\mathrm{obs}} - 1)}$。未测量混杂与处理和结局各需达到此 RR 才能翻盘。需要领域知识锚定。
sensemakr：用偏 R^2 量化遗漏变量偏差，稳健值 RV 表示使效应消失所需的最小混杂强度。可以用已观测协变量做基准校准。
R 实现：EValue 包的 `evalues.RR()`；sensemakr 包的 `sensemakr()`，直接接受 `lm()` 对象。
适用场景：所有依赖可交换性假设的观察性研究。审稿人越来越多地要求作者报告 E-value 或类似的敏感性指标。
局限：无法证明”没有未测量混杂”，只能量化”翻盘需要多强”。E-value 缺乏内部校准，sensemakr 依赖线性模型假设。

笔记 敏感性分析的思路可以追溯到 Cornfield 在 1959 年为吸烟致癌辩论提供的论证。烟草公司辩称吸烟与肺癌的关联可能由遗传因素造成，Cornfield 反驳：吸烟者肺癌发病率是非吸烟者的 9 倍以上，要用遗传混杂解释，该因素需要在吸烟者中流行率高 9 倍以上，这在生物学上不合理。这个论证和 E-value 一脉相承。Rosenbaum [22] 系统化了这套方法论，VanderWeele and Ding [27] 的 E-value 和 Cinelli and Hazlett [8] 的 sensemakr 则分别提供了更易用的工具。

下一章将转向异质性分析。前面所有方法估计的都是平均处理效应，但 ATE 是全体患者效应的均值，它可能掩盖了不同亚群之间的重要差异。也许对某些患者 RHC 确实有害，而对另一些患者反而有益，平均下来才是一个温和的正向风险差。因果森林可以估计个体化处理效应，帮助我们拆开 ATE 的”黑箱”。

本章知识地图

表 8.1: 第 8 章核心概念与常见误解

核心概念	核心内容	常见误解	为什么错
条件可交换性	在协变量 L 条件下潜在结果与处理独立，是因果识别的核心假设	可以通过某种统计检验来验证可交换性	验证需要观测反事实结局，而反事实在数据中永远缺失
E-value	未测量混杂与处理和结局各需达到的最小 RR 才能翻盘	E-value 大就说明没有未测量混杂	E-value 只量化翻盘所需的混杂强度，不能排除这种混杂存在
领域锚定	同一个 E-value 在不同领域含义不同，必须参照已知混杂的关联强度	E-value = 2 在任何领域都算稳健	药物试验中 RR > 2 的混杂罕见，社会科学中 RR > 2 的混杂常见
稳健值 RV	使效应估计降至零所需的最小偏 R^2	RV 小就说明结论一定错	RV 小只说明结论容易被推翻，但小效应的 RV 天然就小
基准校准	用已观测协变量的解释力作为参照系评估未测量混杂的合理强度	只要超过 1 倍基准就翻盘说明结论不可信	要看”k 倍基准”在领域内是否合理，而不只看倍数大小

核心概念	核心内容	常见误解	为什么错
敏感性分析的定位	压力测试，量化翻盘所需条件，搭配领域知识做判断	敏感性分析是因果推断的最后一步，过了就万事大吉	它回答的是”假设违反时结论多脆弱”，而不是”假设是否成立”

第 9 章 谁获益谁受害——因果森林与处理效应异质性

内容提要

- 理解从 ATE 到 CATE 的转变：为什么平均效应掩盖了个体差异
- 掌握因果森林的核心机制：诚实分裂与双重稳健评分
- 在 RHC 数据上用 grf 估计 5735 名患者的个体化处理效应
- 通过变量重要性和亚组分析识别谁因 RHC 受害最重、谁可能获益

上一章用敏感性分析检验了 RHC 平均效应的稳健性。E-value 和 sensemakr 都表明，要让 RHC 的不利效应消失，未测量混杂需要达到相当强的程度。这给了我们信心：RHC 对 ICU 患者整体而言确实增加了 180 天死亡风险。但” 整体” 这个词本身就是一种简化。ICU 里收治的患者差异极大：有 20 岁的创伤患者，也有 80 岁的多器官衰竭患者；有肝功能正常的心衰病人，也有胆红素飙到 10 以上的肝硬化患者。一个 ATE = 0.044 的结论，对这些截然不同的个体意味着什么？是每个人都被 RHC 害了 4.4 个百分点，还是有些人被害了 15 个百分点而另一些人其实获益了？本章要回答的就是这个问题。

9.1 从平均到个体：CATE 的提出

前八章估计的都是平均处理效应 ATE，即 $E[Y(1) - Y(0)]$。这个量把全部 5735 名患者的个体效应压缩成了一个数字。压缩的代价是信息损失：如果处理效应在不同人群中方向相反，ATE 可能接近零，让研究者误以为处理没有效果，实际上是正负效应相互抵消了。

临床实践中，” 同药不同效” 的现象随处可见。华法林在某些患者身上降低卒中风险 65%，在另一些患者身上引发严重颅内出血。抗肿瘤免疫治疗对 PD-L1 高表达的患者可能带来完全缓解，对低表达的患者可能无效甚至引发免疫性肺炎。精准医疗的核心挑战就是：在开始治疗之前，能否预测这个具体的病人会获益还是受害？

回答这个问题需要一个比 ATE 更精细的估计目标。

ATE 是所有人的平均效应。但不是每个人对 RHC 的反应都一样。CATE 问的是：对于这一类特定的病人，效应有多大？比如，对于”65 岁以上、APACHE 评分高于 50、胆红素正常” 的这群患者，RHC 增加死亡率多少？CATE 把 ATE 的单一数字拆成了一张效应的地图，每一类病人在地图上有自己的位置。

下面是 CATE 的正式定义。

比如 APACHE 评分低于 50 的年轻患者，RHC 可能只增加 2 个百分点的死亡风险；而 APACHE 评分超过 80 的高龄患者，增加的可能是 8 个百分点。CATE 就是捕捉这种差异的工具。

定义 9.1 (条件平均处理效应 CATE)

给定协变量向量 $X = x$，条件平均处理效应定义为

$$\tau(x) = E[Y(1) - Y(0) \mid X = x].$$

其中 $Y(1)$ 和 $Y(0)$ 是潜在结局，X 是基线特征变量。 [2]

这个公式的含义是：在协变量取值为 x 的那群人当中，接受处理与不接受处理的平均结局差异是多少。CATE 和 ATE 的关系很直接：ATE 就是 CATE 对 X 的边际分布取期望，ATE $= E[\tau(X)]$。如果 $\tau(x)$ 对所有 x 都相同，说明处理效应是同质的，CATE 退化为 ATE。如果 $\tau(x)$ 随 x 变化，说明存在处理效应异质性，英文称 Heterogeneous Treatment Effects，简称 HTE。

需要区分 CATE 和个体处理效应 ITE。ITE 是 $\tau_i = Y_i(1) - Y_i(0)$，对单个个体 i 而言是一个确定的数，但由于反事实缺失问题，我们永远观察不到 $Y_i(1)$ 和 $Y_i(0)$ 同时发生。CATE 是 ITE 在给定 $X = x$ 的条件下取期望，

是一个可以从数据中估计的统计量。当因果森林对张三输出 $\hat{\tau}(x_{\text{张三}}) = 0.06$ 时，这个值代表的是” 和张三基线特征相同的那群人” 的平均效应，而非张三本人的真实 ITE。

到这里的关键结论是：ATE 虽然方向一致，但它可能掩盖了重要的异质性。接下来我们介绍专门捕捉这种异质性的工具。

9.2 传统方法的局限：亚组分析与交互项

在因果森林出现之前，研究者探索异质性最常用的手段是亚组分析和回归交互项。亚组分析的做法是选定一个协变量，比如年龄，按中位数或临床阈值把样本分成两组，分别估计 ATE，看两组之间有没有差异。回归交互项的做法是在结果模型中加入处理变量与协变量的乘积项 $A \times X$，检验交互系数 β_3 是否显著。

这两种方法都有根本性的缺陷。亚组分析要求研究者预先指定按哪个变量分层，如果有 28 个协变量，按每个变量的中位数各分一次，就做了 28 次比较，多重检验问题严重，假阳性率膨胀。更关键的是，亚组分析每次只看一个变量，无法捕捉多变量联合驱动的异质性。比如 RHC 的效应可能在” 高龄 + 高胆红素 + 低白蛋白” 这个特定组合下才显著恶化，单看年龄或单看胆红素都发现不了这种模式。

回归交互项的问题是函数形式限制。β_3 假设处理效应与协变量之间是线性关系，而真实的异质性模式可能是高度非线性的。APACHE 评分从 30 到 50 的区间内处理效应可能平缓变化，到 60 以上突然跳升，线性交互项无法捕捉这种阈值效应。如果同时放入多个交互项，模型的自由度迅速消耗，估计不稳定，解释也变得困难。

因果森林的设计就是为了解决这些问题：它自动从数据中发现驱动异质性的协变量组合，不需要研究者预先指定分层变量，也不假设效应与协变量之间是线性关系。

9.3 因果森林：诚实分裂与双重稳健评分

因果森林由 Wager and Athey [28] 提出，是随机森林在因果推断领域的自然延伸。普通随机森林的目标是预测 $E[Y \mid X]$，因果森林的目标是预测 $\tau(x) = E[Y(1) - Y(0) \mid X = x]$。两者的算法骨架相同：构建大量决策树，每棵树在随机子样本和随机特征子集上生长，最终对所有树的预测取平均。但因果森林在两个关键环节做了专门的改造。

第一个改造是分裂准则。普通决策树在每个节点选择一个变量和阈值，使得分裂后子节点内的预测误差最小。因果森林的分裂准则不同：它选择的是让左右子节点的**处理效应差异**最大的切分方式。用数学语言说，节点选择使得 $(\hat{\tau}_{\text{left}} - \hat{\tau}_{\text{right}})^2$ 最大的变量和阈值。

用一个具体例子来说明。假设当前节点有 200 名患者，森林在考虑两种切分方式。首先按年龄 70 岁切分，左边组的平均效应为 0.02，右边组为 0.08，差异为 0.06。其次按 APACHE 评分 50 分切分，左边组效应 0.03，右边组 0.06，差异为 0.03。因果森林会选年龄这个切分点，因为它把” 对 RHC 反应不同的人” 分得更开。

这意味着因果森林不关心预测结局本身有多准，只关心哪些变量能把” 对处理反应不同的人” 分开。

想象考试出题和判卷是同一个人。他可以先看自己擅长什么，再围绕擅长的内容出题，考出来的分数自然好看，但完全不可信。普通决策树就有这个毛病：用同一批数据既决定树的结构，又估计叶节点的预测值，结果是过度拟合训练数据的噪声。

诚实分裂的做法是把出题和答题交给不同的人。出题的那一半样本只负责找到最佳的分裂变量和阈值，答题的那一半样本只负责在已经确定的叶节点内计算处理效应的均值。这样得到的效应估计不受分裂过程的选择偏差污染，具有渐近正态性和可构造置信区间的统计性质。

下面是诚实分裂的正式定义。

定义 9.2 (诚实分裂)

为什么需要诚实分裂？普通决策树用同一批数据既决定在哪里切分，又估计叶节点的效应值。这就像用同一批学生既出题又答题：树会迎合训练数据的噪声，导致效应估计过度乐观，置信区间也不可靠。

诚实分裂的做法是把每棵树的训练样本随机等分为两半。第一半称为结构样本，只负责决定树的分裂变量和阈值；第二半称为估计样本，只负责在已经确定的叶节点内计算处理效应 $\hat{\tau}$。构建结构和估计效应使用的是完全不同的样本，互不干扰。 [2]

第二个改造涉及效应的估计方式。Athey, Tibshirani, and Wager [3] 在广义随机森林框架中引入了双重稳健评分。每棵树在叶节点内估计效应时，跳过了处理组与对照组的简单均值差，直接使用类似 AIPW 的评分函数。这个评分同时利用了结果模型和倾向得分模型的信息，使得因果森林在两个辅助模型中至少有一个大致正确时，CATE 的估计就是一致的。这和第 6 章 AIPW 的双重稳健逻辑一脉相承。

定理 9.1 (雷区)

诚实分裂的代价是每棵树只用了一半样本来估计效应，另一半"浪费"在了决定树结构上。在样本量较小的数据集中，这会导致 CATE 估计的方差增大、置信区间变宽。grf 默认设置 `honesty = TRUE`，在 $n < 500$ 的小样本中，研究者需要权衡诚实性带来的方差增加与偏差减少之间的取舍。对于 RHC 数据的 $n = 5735$，样本量足够支撑诚实分裂而不会过度损失效率。

9.4 在 RHC 数据上拟合因果森林

本章继续使用第 1 章介绍的 RHC 数据集，$n = 5735$。下面的代码用 grf 包拟合因果森林，为每名患者估计个体化的 CATE，然后考察效应的分布、驱动异质性的关键变量、以及是否存在统计上显著的异质性。

```
set.seed(2026)
library(tidyverse)
library(grf)

d <- read_csv(here::here("data", "rhc.csv"), show_col_types = FALSE) |>
  mutate(death180_bin = if_else(death180 == "Yes", 1L, 0L),
         sex_bin      = if_else(sex == "Male", 1L, 0L),
         cancer_bin   = if_else(cancer == "No", 0L, 1L))

covs <- c("age", "sex_bin", "cancer_bin", "cardiovascular",
          "congestive_hf", "dementia", "psychiatric", "pulmonary",
          "renal", "hepatic", "gi_bleed", "tumor",
          "immunosupperssion", "transfer_hx", "mi",
          "apache_score", "glasgow_coma_score", "blood_pressure",
          "heart_rate", "respiratory_rate", "temperature",
          "albumin", "creatinine", "bilirubin", "wbc",
          "hematocrit", "das_index", "weight")

X <- as.matrix(d[, covs])
W <- d$rhc
Y <- d$death180_bin

# 2000 棵树，honesty = TRUE 保证统计推断合法
cf <- causal_forest(X, Y, W,
```

```
                    num.trees = 2000,
                    honesty = TRUE,
                    seed = 2026)

# 每名患者的 CATE 预测
cate <- predict(cf)$predictions
cat("CATE 均值:", round(mean(cate), 4),
    " SD:", round(sd(cate), 4), "\n")
cat("CATE > 0 (受害):", round(mean(cate > 0)*100, 1), "%\n")
cat("CATE < 0 (获益):", round(mean(cate < 0)*100, 1), "%\n")
```

结果解读

因果森林为 5735 名患者各自估计了一个 CATE 值。CATE 的均值为 0.043，标准差为 0.025，范围从 −0.042 到 0.137。96.9% 的患者 CATE 大于零，意味着绝大多数人因 RHC 而增加了死亡风险。只有 3.1% 的患者 CATE 为负，即可能从 RHC 中获益。整体分布明显偏向正值，这与前面各章 ATE 一致为正的结论吻合。但分布并非集中在一个点上：5% 分位数为 0.004，95% 分位数为 0.084，说明受害程度在不同患者之间存在差异。CATE 最高的患者死亡风险增加了约 14 个百分点，是均值的三倍多。

图 9.1 展示了 CATE 的分布。

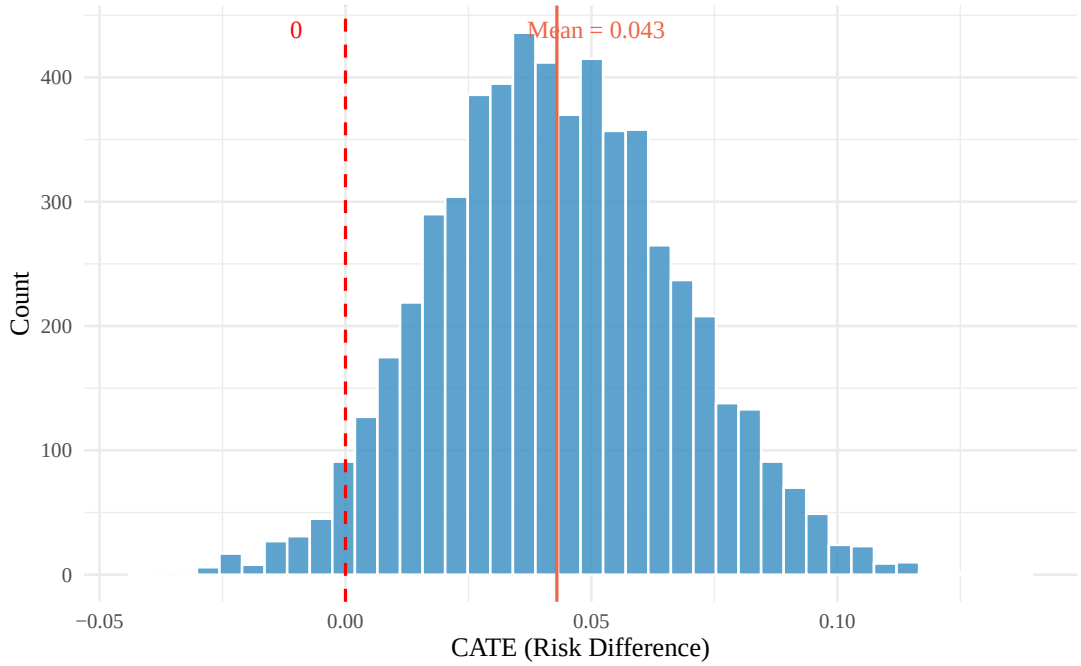

图 9.1: 5735 名患者的 CATE 分布。红色虚线为零参考线，橙红色实线为 CATE 均值 0.043。分布整体偏右，表明绝大多数患者因 RHC 增加了死亡风险，但受害程度存在差异。左尾有少数患者的 CATE 落入负值区域。

9.5 变量重要性：谁在驱动异质性

变量重要性衡量的是：如果把某个变量的信息抹掉，森林估计的效应异质性会减少多少。重要性高的变量意味着它能有效地把” 处理效应高的人” 和” 处理效应低的人” 区分开来。

这和普通随机森林的变量重要性在概念上类似，但含义不同。普通随机森林的变量重要性反映的是对结局预测的贡献。因果森林的变量重要性反映的是对效应异质性的贡献。一个变量可能对预测死亡率很重要，但对

区分”谁受害多、谁受害少”完全没用。因果森林只关心后者。

```
# 变量重要性：哪些协变量驱动了 CATE 的异质性
vimp <- variable_importance(cf)
vimp_df <- data.frame(Variable = covs,
                      Importance = as.numeric(vimp)) |>
  arrange(desc(Importance))
cat("Top 5 变量:\n")
print(head(vimp_df, 5))
```

结果解读

排名前五的变量依次是 *bilirubin*，重要性 0.114；*wbc*，重要性 0.092；*hematocrit*，重要性 0.087；*weight*，重要性 0.086；*blood_pressure*，重要性 0.072。*age* 排在第六位，重要性 0.067。

胆红素排在首位有临床意义。胆红素是肝功能的敏感指标，高胆红素往往意味着肝衰竭或胆道梗阻。RHC 作为一种有创的血流动力学监测手段，对肝功能严重受损的患者可能风险更大，因为这类患者的凝血功能本身就差，插管过程的并发症风险更高。白细胞计数排第二，反映了感染和炎症状态对 RHC 效应的调节作用。

图 9.2 展示了前 12 个变量的重要性排名。

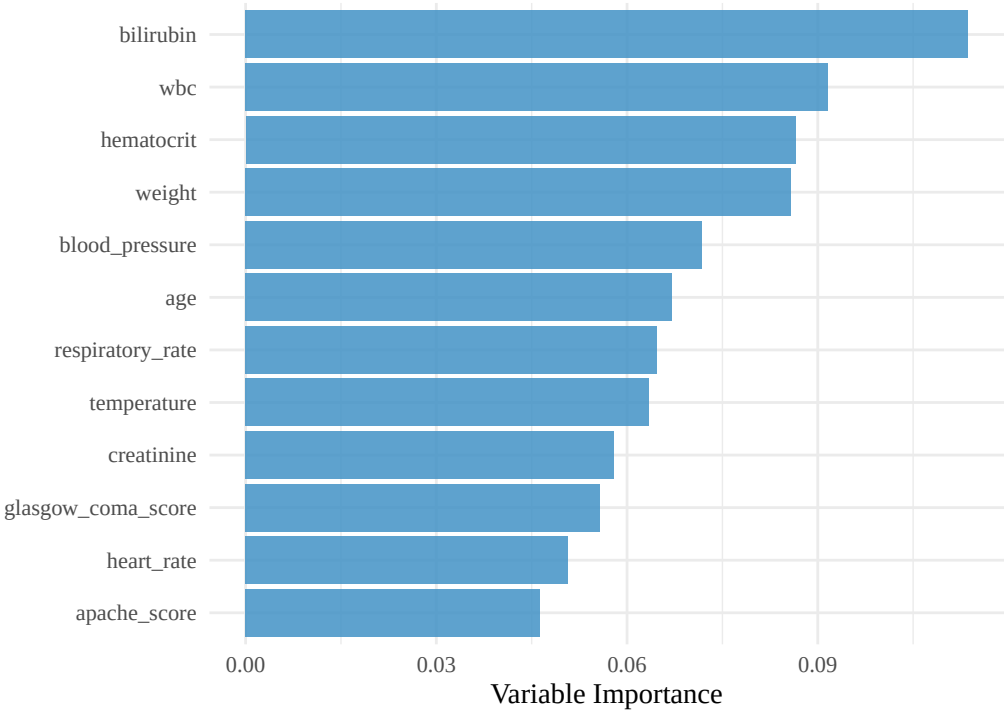

图 9.2: 因果森林变量重要性排名，展示驱动 CATE 异质性的前 12 个协变量。*bilirubin* 排名第一，其次是 *wbc* 和 *hematocrit*，三者均为实验室检验指标。

9.6 异质性的统计检验：BLP 方法

CATE 分布图显示了效应的变异，但我们还需要正式检验这种变异是否具有统计学意义，还是仅仅反映了估计的噪声。grf 提供的 `test_calibration()` 函数实现了 Best Linear Projection，简称 BLP 检验。BLP 把因果森林的 CATE 预测值作为自变量，用留出样本上的双重稳健得分作为因变量，拟合一个线性回归。回归中有两个系

数：*mean.forest.prediction* 检验因果森林的平均预测是否与真实 ATE 一致，*differential.forest.prediction* 检验 CATE 预测的变异是否反映了真实的异质性。

```r
# BLP 检验：CATE 的异质性是否具有统计学意义
blp <- test_calibration(cf)
print(blp)
```

结果解读

BLP 检验给出两个关键系数。*mean.forest.prediction* 的估计值为 1.006，标准误 0.318，$t = 3.16$，$p < 0.001$，高度显著。这说明因果森林的平均 CATE 预测与真实 ATE 校准良好，系数接近 1 意味着预测没有系统性的缩放偏差。

differential.forest.prediction 的估计值为 0.831，标准误 0.542，$t = 1.53$，单侧 $p = 0.063$。这个系数在 10% 水平上边缘显著，在 5% 水平上不显著。它的含义是：因果森林预测的 CATE 变异中，大约 83% 反映了真实的异质性，但由于标准误较大，我们不能在传统的 5% 水平上排除" 异质性为零" 的假设。

这个结果的实际解读是：RHC 数据中存在一定程度的处理效应异质性，但信号相对温和。这和 CATE 分布图传递的信息一致：大部分患者的 CATE 集中在 0.02–0.08 之间，变异存在但不算剧烈。

9.7 因果森林的平均 CATE 与前章 ATE 的校验

因果森林输出的是每个人的 CATE，把所有人的 CATE 平均起来应该接近前面各章估计的 ATE。grf 提供了 `average_treatment_effect()` 函数，它使用 AIPW 方法在因果森林的基础上计算总体 ATE 及其标准误。

```r
# 因果森林的 AIPW 平均效应
ate_cf <- average_treatment_effect(cf, target.sample = "all")
cat("ATE:", round(ate_cf[1], 4),
    " SE:", round(ate_cf[2], 4),
    " 95% CI: [", round(ate_cf[1] - 1.96*ate_cf[2], 4),
    ",", round(ate_cf[1] + 1.96*ate_cf[2], 4), "]\n")
```

结果解读

因果森林给出的 ATE 为 0.044，标准误 0.012，95% CI 为 $[0.020, 0.068]$，$p < 0.001$。这个数字和第 6 章手动实现 AIPW 得到的 0.044 几乎完全一致，与 G 计算的 0.052 和 IPW 的 0.032 也在同一区间内。跨方法的一致性进一步确认了 RHC 增加死亡风险这个因果结论。因果森林的标准误与 AIPW 相当，这是因为因果森林内部本质上也在做双重稳健估计。

9.8 亚组分析：谁受害最重

变量重要性告诉我们哪些协变量驱动了异质性，但还没有回答" 什么样的患者受害最重" 这个临床问题。一种直接的做法是把 5735 名患者按 CATE 预测值分成五等分，考察每组的 ATE 和临床特征。

```r
# 按 CATE 五等分做亚组分析
d$cate <- cate
d$cate_q <- cut(d$cate,
    breaks = quantile(d$cate, probs = seq(0, 1, 0.2)),
    labels = c("Q1", "Q2", "Q3", "Q4", "Q5"),
```

```
    include.lowest = TRUE)

for (q in c("Q1", "Q5")) {
  idx <- which(d$cate_q == q)
  ate_q <- average_treatment_effect(cf, subset = idx,
                                    target.sample = "all")
  cat(q, ": ATE =", round(ate_q[1], 4),
      ", 95% CI = [", round(ate_q[1] - 1.96*ate_q[2], 4),
      ",", round(ate_q[1] + 1.96*ate_q[2], 4), "]\n")
}
```

结果解读

Q1 是 CATE 预测值最低的 20% 患者，即因果森林认为受害最轻甚至可能获益的人群，其 AIPW 估计的 ATE 为 0.054，95% CI 为 $[0.001, 0.108]$，刚好显著。Q5 是 CATE 预测值最高的 20% 患者，即受害最重的人群，ATE 为 0.082，95% CI 为 $[0.028, 0.136]$，$p = 0.003$。Q5 的效应几乎是 Q1 的 1.5 倍。

两组的临床特征有明显差异。Q1 组的平均年龄为 57.8 岁，平均 APACHE 评分 53.2，平均胆红素 4.25。Q5 组的平均年龄为 64.9 岁，平均 APACHE 评分 56.7，平均胆红素 1.02。Q1 组的胆红素水平远高于 Q5 组，这和变量重要性的结果呼应。但注意方向：高胆红素的患者 CATE 反而更低，而低胆红素、高龄的患者 CATE 更高。这提示 RHC 的有害效应在”非肝脏病因、高龄、中高 APACHE”的患者身上表现得更集中。高胆红素的患者 CATE 较低，可能是因为这类患者即使不做 RHC，死亡率也已经很高，RHC 的边际伤害相对较小。从临床角度看，这些结果与 Connors, Speroff, Dawson, et al. [9] 的原始发现方向一致：RHC 并未改善任何亚组的生存。因果森林进一步量化了受害程度在不同人群中的差异，为未来的临床决策提供了更精细的证据。

定理 9.2 (雷区)

按 CATE 预测值分组再在组内估计 ATE，存在一个微妙的偏差来源。分组变量本身是从数据中估计出来的，而组内 ATE 也是用同一份数据估计的。如果不加处理，分组和估计使用了重叠的信息，会导致”自我实现”的偏差：Q5 组的 ATE 会被人为拉高，Q1 组的会被人为拉低。grf 的 `average_treatment_effect()` 通过样本分裂和诚实估计部分缓解了这个问题，但在解释亚组结果时仍需谨慎。稳健的做法是预先在独立的验证集上确认亚组差异。

笔记 因果森林的提出有清晰的学术背景。Athey and Imbens [2] 最先把决策树用于异质性效应估计，提出了诚实估计和自适应正则化的框架。Wager and Athey [28] 把这个想法扩展到随机森林，证明了因果森林在高维协变量下的一致性和渐近正态性。Athey, Tibshirani, and Wager [3] 进一步推广为广义随机森林框架，把双重稳健评分嵌入了分裂和估计过程。同期，Künzel et al. [15] 从 Metalearner 的角度提出了 S-Learner、T-Learner、X-Learner 等替代策略。因果森林和 Metalearner 解决的是同一个问题，区别在于因果森林直接对 $\tau(x)$ 建模，而 Metalearner 通过组合多个标准预测模型间接估计 $\tau(x)$。在实践中两者各有优势，因果森林的统计推断性质更好，Metalearner 的灵活性更高。

方法卡片：因果森林

估计目标：条件平均处理效应 $\tau(x) = E[Y(1) - Y(0) \mid X = x]$。

核心机制：随机森林 + 诚实分裂 + 双重稳健评分。分裂准则最大化子节点间的效应差异，诚实分裂把建树和估计分离，双重稳健评分同时利用结果模型和倾向得分。

核心假设：可交换性 + 正值性 + 一致性，与前几章的假设相同。额外要求样本量足以支撑诚实分裂的样本拆分。

R 实现：grf 包的 `causal_forest()`。关键参数包括 `num.trees`、`honesty`、`sample.fraction`。

适用场景：协变量较多、研究者不确定哪些变量驱动异质性、需要对 CATE 做统计推断。
失效场景：样本量太小导致诚实分裂后叶节点样本不足；处理效应真正同质时，模型会估出噪声驱动的虚假异质性；未测量混杂同样会传递到 CATE 估计中。

9.9 累积对比表

表 9.1: 方法演进对比表，截至第 9 章

方法	ATE 估计	95% CI	核心假设
回归调整	OR = 1.34	[1.18, 1.52]	模型设定正确 + 可交换性 + 正值性
G 计算	RD = 0.052	[0.027, 0.082]	结果模型设定正确 + 可交换性 + 正值性
IPW	RD = 0.032	[0.005, 0.064]	倾向得分模型正确 + 可交换性 + 正值性
AIPW	RD = 0.044	[0.017, 0.072]	两个模型至少一个正确 + 可交换性 + 正值性
因果森林	RD = 0.044	[0.020, 0.068]	可交换性 + 正值性 + 诚实分裂

五种方法的结论方向完全一致：RHC 增加了 ICU 患者的 180 天死亡风险。因果森林给出的 ATE 与 AIPW 几乎相同，置信区间宽度也相当。因果森林的额外价值在于它提供了 ATE 之外的信息：5735 个个体化的 CATE 预测值、驱动异质性的变量排名、以及亚组间效应差异的量化。

下一章将汇总全书所有方法的估计结果，对 RHC 的因果效应做最终结论。在累积了回归调整、G 计算、倾向得分、双重稳健、机器学习增强、敏感性分析和异质性分析之后，我们将回到最初的问题：ICU 里给危重病人插右心导管，到底是救人还是害人？

本章知识地图

表 9.2: 第 9 章核心概念与常见误解

核心概念	核心内容	常见误解	为什么错
CATE	在协变量 $X = x$ 条件下的平均处理效应，$\tau(x) = E[Y(1) - Y(0) \mid X = x]$	CATE 就是个体处理效应 ITE	CATE 是给定特征组合的群体平均，ITE 是不可观测的个体确定量
诚实分裂	建树和估计效应使用不同的样本，避免过拟合导致的推断失效	诚实分裂浪费了一半数据	代价是方差增大，但换来的是合法的统计推断和置信区间，在因果推断场景中推断比精度更重要
变量重要性	衡量协变量在分裂决策中被使用的频率，反映对异质性的贡献	变量重要性高等于这个变量对结局影响大	因果森林的变量重要性衡量的是对效应异质性的贡献，不是对结局预测的贡献

核心概念	核心内容	常见误解	为什么错
BLP 检验	用留出样本检验 CATE 预测的变异是否反映真实异质性	CATE 分布有变异就说明存在异质性	估计噪声也会产生变异，需要 BLP 或其他统计检验来区分真实信号和噪声
双重稳健评分	因果森林内部使用类似 AIPW 的评分函数估计叶节点效应	因果森林只用处理组和对照组的均值差	简单均值差不具备双重稳健性，grf 的评分函数同时利用结果模型和倾向得分
亚组分析偏差	按估计出的 CATE 分组再组内估计 ATE 会有自我实现偏差	分成五组后每组的 ATE 完全可信	分组变量和组内估计使用了重叠信息，需要诚实估计或独立验证集来校正

第 10 章 全书汇总——十种方法的终极对比

内容提要

- 汇总第 3–9 章所有方法的 ATE 估计，形成终极对比表
- 用森林图直观展示八种方法的点估计与置信区间
- 理解跨方法收敛与分歧背后的统计学原因
- 结合敏感性分析给出 RHC 因果效应的最终结论
- 梳理全书方法论脉络，明确每种方法的适用场景

第 1 章提出了一个问题：ICU 里给危重病人插右心导管，到底是救人还是害人？第 2 章用 DAG 画出了 RHC 与死亡率之间的混杂结构，确定了调整集。随后的七章用九种不同的因果推断方法回答了这个问题。回归调整从系数漂移中剥离混杂，G 计算构造了反事实人群，倾向得分方法从匹配、加权到重叠权重提供了三条路径，AIPW 拴上了两根保险绳，DML 和 TMLE 用机器学习替代了参数模型，因果森林把平均效应拆解为个体化的 CATE。第 8 章的敏感性分析用 E-value 和 sensemakr 量化了结论对未测量混杂的抵抗力。

现在是把所有结果放到同一张桌子上的时候了。本章不引入新方法，它的任务是整合、比较、反思。一张终极对比表汇总所有估计，一张森林图给出视觉全貌，然后我们讨论收敛意味着什么、分歧意味着什么、研究者该选哪种方法、RHC 到底是什么结论。

10.1 终极对比表

表 10.1 汇总了第 3–9 章所有方法的 ATE 估计、置信区间、核心假设和主要局限。回归调整报告的是条件 OR，其余方法报告的都是边际风险差 RD，两种尺度不直接可比，但 OR > 1 与 RD > 0 传递的因果方向相同。

表 10.1: 全书方法终极对比表

方法	估计	95% CI	核心假设	主要局限
回归调整	OR=1.34	[1.18, 1.52]	模型设定正确 + 可交换性 + 正值性	对函数形式敏感，无显式反事实，条件 OR 非边际效应
G 计算	RD=0.052	[0.027, 0.082]	结果模型设定正确 + 可交换性 + 正值性	单一模型依赖，模型错则估计错，无纠错机制
PSM	RD=0.076	[0.041, 0.109]	处理模型正确 + 可交换性 + 正值性	丢弃 36% 样本，目标人群改变，统计效力下降
IPW	RD=0.055	[0.025, 0.085]	处理模型正确 + 可交换性 + 正值性	极端权重导致方差膨胀，有效样本量下降
OW	RD=0.061	[0.033, 0.089]	处理模型正确 + 可交换性 + 正值性	估计的是 ATO 而非 ATE，目标人群为重叠人群
AIPW	RD=0.044	[0.017, 0.072]	两个模型至少一个正确 + 可交换性 + 正值性	两个模型同时错时失效，有限样本可能出界
DML	RD=0.040	[0.014, 0.065]	Neyman 正交 + 交叉拟合 + 可交换性 + 正值性	依赖 ML 学习器质量，随机折划分引入波动
TMLE	RD=0.088	[0.074, 0.103]	目标化更新 + SL + 可交换性 + 正值性	窄 CI 可能覆盖率偏低，对 SL 配置敏感
因果森林	RD=0.044	[0.020, 0.068]	可交换性 + 正值性 + 诚实分裂	侧重 CATE，ATE 为副产品，小样本方差大

九种方法的估计方向完全一致：RHC 增加了 ICU 患者的 180 天死亡风险。没有任何一种方法给出 RHC 有保护效应的结论。八个风险差估计落在 0.040 到 0.088 之间，中位数约为 0.053。回归调整的 OR = 1.34 对应的方向也是有害。所有置信区间的下界均大于零，统计显著性在每种方法下都成立。

这张表里有几个细节值得注意。第 5 章的倾向得分方法使用了 38 个协变量，第 6 章之后的方法使用了 28 个协变量，变量集的差异会影响点估计。PSM 和 OW 使用了更大的协变量集，IPW 也是如此。第 6 章的 AIPW 在缩减后的 28 个协变量上工作，这解释了同一份数据上不同章节的 G 计算和 IPW 估计为什么会有数值差异。变量集的选择本身就是建模决策的一部分，不同的变量集对应不同的可交换性条件。表中的数字差异既反映了方法差异，也反映了变量集差异，两者的贡献混在一起。

10.2 森林图：一张图看全貌

森林图是因果推断和 meta 分析中最常用的可视化工具之一。读法很简单：每一行代表一种方法，红色圆点是该方法的点估计，蓝色横线是 95% 置信区间。图中间有一条竖直虚线标记零点。如果某种方法的横线完全落在零线右侧，说明该方法在 95% 置信水平下认为 RHC 增加了死亡风险。如果横线跨过了零线，说明效应在统计上不显著。点估计越靠右，估计的有害效应越大；横线越短，估计的精度越高。

图 10.1 把八种风险差方法的点估计和 95% 置信区间放在同一张图上。回归调整因为报告的是 OR 而非 RD，没有纳入森林图，但它的方向与所有 RD 方法一致。

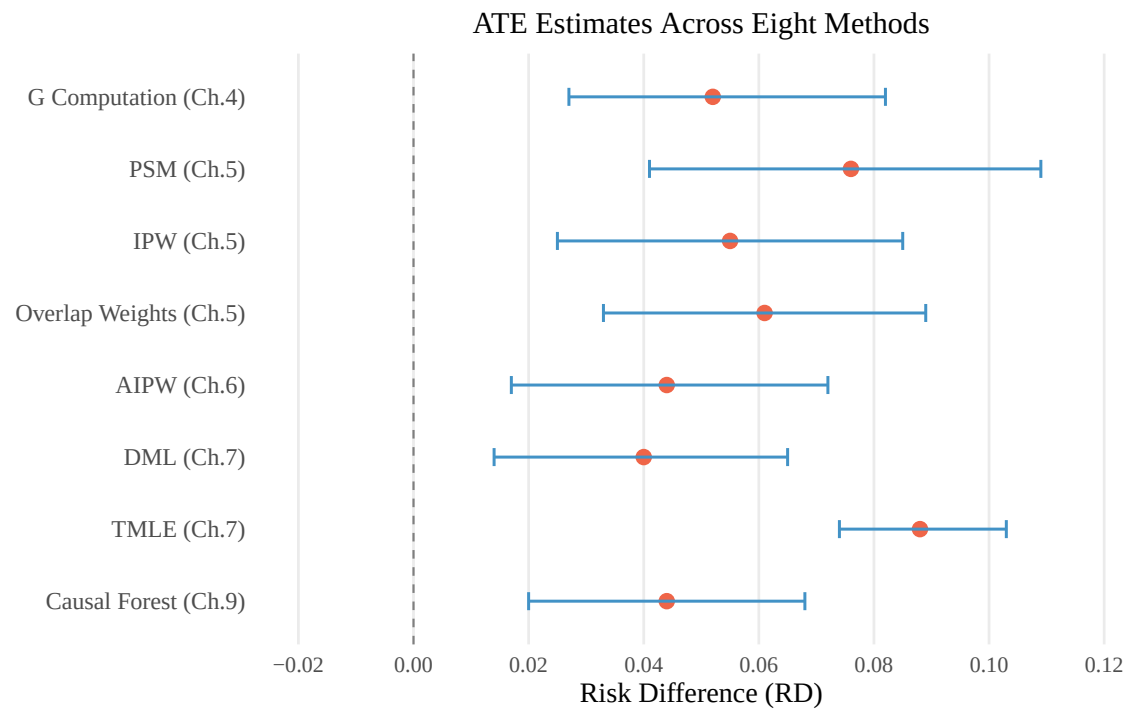

图 10.1: 八种方法的 ATE 估计及 95% CI 森林图。所有点估计均落在零的右侧，方向一致指向 RHC 增加死亡风险。虚线为零参考线。TMLE 的点估计最高但置信区间最窄，IPW 和 DML 的点估计最低但区间较宽。

```r
set.seed(2026)
library(ggplot2)

# 从第 3--9 章收集的真实估计值
methods <- c(
  "G Computation (Ch.4)", "PSM (Ch.5)", "IPW (Ch.5)",
  "Overlap Weights (Ch.5)", "AIPW (Ch.6)", "DML (Ch.7)",
  "TMLE (Ch.7)", "Causal Forest (Ch.9)")
est   <- c(0.052, 0.076, 0.055, 0.061, 0.044, 0.040, 0.088, 0.044)
ci_lo <- c(0.027, 0.041, 0.025, 0.033, 0.017, 0.014, 0.074, 0.020)
ci_hi <- c(0.082, 0.109, 0.085, 0.089, 0.072, 0.065, 0.103, 0.068)

df <- data.frame(method = factor(methods, levels = rev(methods)),
                 est = est, lo = ci_lo, hi = ci_hi)

ggplot(df, aes(x = est, y = method)) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "grey50") +
  geom_point(size = 3, color = "#EF6548") +
  geom_errorbar(aes(xmin = lo, xmax = hi), width = 0.25,
                color = "#4292C6", linewidth = 0.7, orientation = "y") +
  labs(x = "Risk Difference (RD)", y = NULL,
       title = "ATE Estimates Across Eight Methods") +
  scale_x_continuous(breaks = seq(-0.02, 0.12, 0.02)) +
  theme_minimal(base_size = 14, base_family = "serif") +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major.y = element_blank(),
```

```
27        plot.title = element_text(hjust = 0.5))
```

结果解读

森林图的视觉信息很清楚。八个点估计全部落在零线右侧，没有任何一个置信区间的下界穿过零。这意味着无论研究者选择哪种方法，都会得出”RHC 增加死亡风险” 的结论。点估计的分布呈现两个层级：TMLE 和 PSM 偏高，分别为 0.088 和 0.076；其余六种方法集中在 0.040 到 0.061 之间。TMLE 的置信区间明显窄于其他方法，这既反映了它利用似然结构的效率优势，也可能意味着有限样本覆盖率偏低。PSM 的点估计偏高可以部分归因于样本筛选效应：匹配丢弃了 36% 的样本，剩余人群的效应可能与全人群不同。

10.3 收敛与分歧：跨方法一致性说明了什么

九种方法在方向上完全一致：所有点估计都大于零，所有置信区间的下界都大于零。八个风险差估计落在 0.040 到 0.088 之间，中位数约为 0.053。这本身就是强有力的证据。

为什么方向一致这么重要？因为每种方法依赖不同的假设和建模路径。回归和 G 计算依赖结果模型，PSM、IPW 和 OW 依赖处理模型，AIPW、DML 和 TMLE 同时使用两个模型，因果森林用非参数方法估计异质性效应。如果 RHC 的有害效应只是某种特定建模选择的产物，比如逻辑回归的函数形式恰好制造了偏差，那么换用随机森林或 Super Learner 之后这个效应应该消失。但它没有消失。DML 用随机森林替代了逻辑回归，给出 RD = 0.040；TMLE 用 Super Learner 集成了多种算法，给出 RD = 0.088。参数方法和非参数方法指向同一个方向。

这种跨方法的收敛在因果推断中有一个术语叫做” 三角测量”，英文称 triangulation。当多种方法从不同角度逼近同一个因果量，且结果指向同一方向时，我们对因果结论的信心会增强。单一方法的估计永远可以被质疑：” 你的模型设定可能错了”” 你的倾向得分可能遗漏了关键变量”。但当九种方法同时犯同一方向的错误的概率远低于任一方法单独犯错的概率时，方向性的结论就比较可靠了。

置信区间的重叠模式也值得关注。森林图中八个方法的 95% CI 存在大面积的交叉重叠，中心区域大约在 0.04 到 0.07 之间。没有任何一个方法的置信区间与其他方法完全不重叠。这种重叠意味着各方法估计的差异可以被抽样变异性解释，它们在统计学意义上是相互兼容的。

如果某个方法的置信区间与其余所有方法完全分离，就应该认真排查该方法的建模假设是否被违反。

点估计的数值差异同样包含有用的信息。参数方法的 G 计算、IPW 和 AIPW 给出的 RD 在 0.032 到 0.052 之间，使用机器学习的 DML 和因果森林也落在这个区间内。

两个偏离主群的方法各有可追溯的原因。TMLE 的 0.088 偏高，这和第 7 章讨论过的原因一致：TMLE 在 logit 尺度上做目标化更新，与 DML 在线性尺度上做残差校正的路径不同，有限样本中两者可以产生分歧。PSM 的 0.076 偏高则有另一个原因：匹配丢弃了倾向得分极端的样本，剩余人群的基线特征与全人群不同，PSM 估计的严格来说是匹配样本的 ATT 而非全人群的 ATE。

如果去掉 TMLE 和 PSM 这两个有明确偏离原因的估计，剩余六种方法的 RD 集中在 0.040 到 0.061 之间，均值约为 0.049。这个区间可以看作 RHC 边际风险差的合理范围：接受 RHC 使 180 天死亡概率升高约 4 到 6 个百分点。

但跨方法收敛有一个重要的盲区。本书用到的九种方法全部依赖同一个不可检验的假设：条件可交换性，即在控制了观测协变量之后，处理分配与潜在结局独立。如果存在一个未被测量的混杂变量同时影响 RHC 使用和 180 天死亡，那么无论用多少种方法、无论模型多灵活，所有估计都会朝同一个方向偏。

九种方法一致为正，有两种解释：要么 RHC 确实增加了死亡率，要么所有方法共享的可交换性假设被违反了，而违反的方向恰好让效应偏正。这两种解释无法用数据本身区分开来，只能靠领域知识和敏感性分析来判断哪一种更合理。第 8 章的 E-value 分析就是在做这个判断。

10.4 方法选择指南

面对同一个因果问题，九种方法给出了方向一致但数值不同的估计。研究者在实际应用中该用哪一种？答案取决于研究场景的具体特征。

当协变量维度低于十个、研究者对变量之间的关系有充分的领域知识时，回归调整和 G 计算是合理的起点。它们的优势是透明和可解释：回归系数可以直接写进论文的表格，G 计算的反事实预测可以逐个检查是否合理。代价是研究者必须为函数形式负全责，交互项该不该加、非线性该怎么处理，都需要事先判断。

一个典型的应用场景：流行病学队列研究中暴露因素和三五个混杂变量之间关系明确，比如吸烟与肺癌的关联控制年龄、性别和社会经济地位。RHC 数据有 28 到 38 个协变量，线性 logistic 回归能否正确捕捉所有关系是一个合理的疑问，所以回归调整在本书中更适合用作基准线，放在附表里和主分析对照。

当研究者更有信心预测” 谁接受了处理” 而非预测结局时，倾向得分方法是自然的选择。一个典型的场景：ICU 医生根据什么指标决定是否插 RHC，有比较成熟的临床指南和经验总结；但 180 天死亡率受多少因素影响、每个因素的函数形式是什么，远不如处理决策透明。在这种情况下，把建模负担放在处理模型上比放在结果模型上更安全。

三种倾向得分方法各有取舍。IPW 保留全样本但对极端权重敏感，适合正值性表现良好的数据。OW 牺牲了全人群推断换取方差稳定性，适合倾向得分分布尾部很长的数据。PSM 最直观但损失样本，适合需要” 匹配后比较” 这种直觉叙事的研究，也适合审稿人对倾向得分匹配已经有成熟理解的期刊。

倾向得分方法的一个独特优势是平衡诊断的可视化。Love plot 可以一目了然地展示调整前后所有协变量的平衡状况，这种透明性是回归调整和 G 计算不具备的。在审稿过程中，一张 Love plot 比任何数字都更有说服力地展示了混杂控制的效果。

在大多数现代因果推断分析中，双重稳健方法应该是主分析的默认选项。AIPW、DML 和 TMLE 都具备双重稳健性，同时建两个模型让研究者不用在结果模型和处理模型之间做二选一的赌注。

三者的选择更多取决于学科传统和软件生态。流行病学领域倾向使用 TMLE，因为它的表述接近统计学家熟悉的影响函数和似然语言，tmle 包提供了开箱即用的实现。计量经济学领域倾向使用 DML，因为 Neyman 正交化和矩方法的表述与计量训练更契合，DoubleML 包基于 mlr3 生态支持灵活的学习器配置。手动实现的 AIPW 则适合教学和研究场景，比如本书第 6 章的做法：它让研究者看清公式里每一项的贡献，对理解方法的内部机制有不可替代的价值。

当研究目标从” 平均效应是多少” 转向” 谁获益谁受害” 时，因果森林是目前最成熟的工具。一个典型的场景：第 9 章发现高龄、低胆红素的患者 CATE 可达 8 个百分点以上，而高胆红素的肝病患者 CATE 相对较低。这种亚群差异只有因果森林能系统性地挖掘。但因果森林的 ATE 估计只是副产品，如果研究者只关心平均效应，直接用 AIPW 或 DML 更高效。

无论选择哪种方法，敏感性分析都是不可缺少的最后一步。因果推断的所有方法都建立在不可检验的假设之上，敏感性分析的作用是量化” 假设需要被违反到什么程度，结论才会翻转”。E-value 提供了一个快速的汇总数字，sensemakr 用已观测协变量的解释力做内部校准，两者配合使用可以给审稿人和读者一个关于结论稳健程度的定量判断。第 8 章用这两种工具检验了 AIPW 的估计，结论是 RHC 的有害效应在中等强度的未测量混杂下就可能被解释掉。这个信息和九种方法的方向一致性一样重要，它界定了因果结论的可信边界。

10.5 在论文中报告多方法比较

本书的森林图展示了九种方法的比较，但在正式的研究论文中，如何组织和报告这些结果是一个实操问题。常见的做法是在主文中报告一种事先确定的主分析方法，然后在附表中报告其余方法的结果作为敏感性分析。

主分析方法的选择应该在分析计划中预先指定，选择依据是方法的假设与研究场景的匹配度。如果研究者对结果模型和处理模型都没有强信心，双重稳健方法是安全的默认选项。如果领域内有成熟的倾向得分方法使

用传统，PSM 或 IPW 作为主分析也是合理的。关键是选择的理由要写清楚，让读者和审稿人能够评估这个选择是否恰当。

附表中的多方法对比用来回答两个问题：结论对方法选择是否敏感？如果敏感，差异的来源是什么？在 RHC 数据上，九种方法给出了方向一致的结论，这本身就是一个有价值的发现，值得在讨论部分提及。如果某种方法给出了明显不同的结果，比如 TMLE 的 0.088 高于其他方法的 0.040–0.061，研究者应该在讨论中解释可能的原因，比如学习器配置差异或目标化路径差异，而非简单地选择性忽略。

图 10.1 这样的森林图适合放在论文的正文或附录中，作为多方法敏感性分析的可视化总结。它让读者一眼看到所有方法的一致性程度，比单独列举数字更直观。

报告敏感性分析结果时，不需要对每种方法做同样详细的讨论。主分析的方法选择、模型设定、诊断步骤应该在方法部分完整报告。敏感性分析的结果可以简要概括方向一致性，然后引用附表或森林图，让感兴趣的读者自行查看细节。如果敏感性分析的某个方法给出了与主分析不同的结论，则需要在讨论部分认真解释差异的可能原因。

定理 10.1 (雷区)

方法选择应该在看到结果之前确定，写在分析计划或预注册方案中。一个常见的问题是”方法购物”：研究者跑完所有方法之后，挑选给出最显著或最符合预期的那个作为主分析报告。这和多重检验的问题本质相同，会膨胀假阳性率。正确的做法是事先确定一种方法作为主分析，其余方法作为敏感性分析附在附表中。本书的森林图之所以有意义，是因为它展示的是同一个问题的九种独立回答，而非一个研究者从中挑选最好看的那个。 ♡

10.6 RHC 的最终结论

经过九章的分析，RHC 的因果效应有了一个比较完整的画面。

从方向上看，九种方法一致指向 RHC 增加 180 天死亡风险。从效应大小看，排除 TMLE 和 PSM 的特殊偏离后，RD 的合理范围在 0.040 到 0.061 之间，即每 100 名接受 RHC 的 ICU 患者中约多 4 到 6 人在 180 天内死亡。回归调整的 OR = 1.34 和这个风险差范围也是匹配的。

从稳健性看，第 8 章的敏感性分析给出了 E-value = 1.42，置信区间下界的 E-value = 1.23。这意味着一个未测量的混杂因素与 RHC 使用和 180 天死亡各自的关联强度都达到 RR = 1.42 以上，就能把观测到的效应完全解释为混杂偏差。RR = 1.23 的残余混杂就能让统计显著性消失。在 ICU 医学领域，已知的强混杂如 APACHE 评分与 RHC 使用的关联可达 RR = 2.0，与死亡率的关联更强。这些强混杂已经被控制了。残余的未测量因素要同时在两端达到 1.42 的关联强度，可能性中等偏低但无法排除。sensemakr 的基准校准显示，3 倍 APACHE 评分强度的混杂才能让效应消失，2 倍就能让统计显著性消失。

从异质性看，第 9 章的因果森林显示 96.9% 的患者 CATE 大于零，只有 3.1% 的患者可能从 RHC 中获益。受害程度在不同人群中有差异：高龄、低胆红素、中高 APACHE 评分的患者受害最重，CATE 可达 8 个百分点以上。高胆红素的肝病患者 CATE 相对较低，可能是因为基线死亡率已经很高，RHC 的边际伤害有限。BLP 检验的 *differential.forest.prediction* 系数为 0.831，单侧 $p = 0.063$，异质性信号存在但强度温和。变量重要性排名显示胆红素、白细胞计数和血细胞比容是驱动 CATE 变异的前三个变量，这些实验室指标反映了肝功能、感染状态和血液系统的基线状况。

综合这些证据，本书的结论与 Connors, Speroff, Dawson, et al. [9] 原始论文的发现方向一致：RHC 与更高的 180 天死亡率相关，这个关联在多种因果推断方法下保持稳定。但我们不能宣称这是一个板上钉钉的因果效应，因为 E-value = 1.42 的保护强度有限。一个关联强度为 RR = 1.42 的未测量混杂就足以解释掉全部效应。ICU 临床决策中可能存在这样的未测量因素，比如主治医师的风险偏好、家属的治疗意愿、床旁对患者”能不能救”的主观判断。这些因素很难用结构化数据捕捉，但它们同时影响 RHC 的使用和患者的预后。

把 E-value 放到 ICU 医学的具体语境中看。APACHE 评分是 RHC 数据中已知最强的混杂变量，第 3 章的系数

漂移分析显示控制它之后 OR 从 1.38 降到 1.18，相当于吸收了约 14% 的混杂。sensemakr 的基准校准表明，一个与 APACHE 评分同等强度的残余混杂能让统计显著性消失，3 倍强度才能让效应消失。我们已经控制了 APACHE 评分本身，但 ICU 中是否存在一个我们没有测量、但与 APACHE 评分同等重要的混杂因素？考虑到 Connors 1996 的数据收集了当时临床上几乎所有可量化的患者特征，漏掉一个如此强大的混杂因素的可能性不高，但也不能排除。这种”可能性不高但无法排除”的判断，正是观察性研究中因果推断的典型结论形态。

命题 10.1 (RHC 因果效应的最终判断)

RHC 可能有害，但残余混杂不能被排除。九种方法的方向一致性和 E-value 的中等保护强度加在一起，给出的是一个”令人不安但不确定”的结论。这恰好是 Connors 1996 年论文发表后引发持续争论的原因：证据足够强到让人不敢随意使用 RHC，但不够强到让人确信 RHC 必然有害。后续的随机对照试验也没有推翻这个结论，因为伦理和操作上的困难使得 RHC 的 RCT 始终没有完成。 ♠

从方法论的角度看，RHC 数据恰好是因果推断教学的理想案例。它足够复杂：49 个变量、5735 个观测、处理分配高度非随机、正值性在尾部受到挑战。它足够真实：这不是模拟数据，每一行代表一个真实的 ICU 患者。它的结论足够微妙：效应方向一致但大小不确定，敏感性分析给出的保护有限。一个”效果很大、E-value 很高、所有方法完美收敛”的数据集反而不适合教学，因为它让学生误以为因果推断总能给出确定的答案。RHC 数据告诉我们的是：即使用了最先进的方法，观察性研究的因果结论仍然带有不确定性，而量化和报告这种不确定性是研究者的责任。

10.7 未覆盖的主题

本书聚焦于点处理的因果推断，即处理在单一时间点发生、结局在固定时间窗口内观测。这是因果推断最基本的设定，也是所有复杂方法的起点。但真实的研究场景远比这复杂，以下几个方向是本书没有覆盖但值得进一步学习的。

中介分析回答的是”处理通过什么机制影响结局”。RHC 增加了死亡率，但增加的机制是什么？可能的路径包括：RHC 插管操作本身导致并发症增加，比如气胸或血管损伤；RHC 提供的血流动力学数据引导了更激进的治疗策略，而这些策略本身有风险；或者 RHC 延长了 ICU 住院时间，增加了院内感染的机会。分解总效应为直接效应和间接效应需要额外的无混杂假设和专门的估计方法，VanderWeele [26] 的专著是这个领域的标准参考。

缺失数据在临床研究中无处不在。RHC 数据的完整性相对较好，但很多真实数据集存在大量缺失值。缺失的机制是完全随机的 MCAR、条件随机的 MAR、还是非随机的 MNAR，决定了什么样的处理方法是合理的。多重插补和逆概率加权是两种主流的处理策略，它们可以与本书介绍的因果推断方法结合使用。如果结局变量本身存在缺失，比如 180 天随访时部分患者失访，那么失访机制就成了另一种形式的混杂，需要专门的方法来处理。

纵向因果推断处理的是时变处理和时变混杂。如果 ICU 患者在第一天、第三天、第七天分别接受了不同的干预，而每次干预决策受到之前健康状态的影响，同时健康状态又受到之前干预的影响，传统的点处理方法就不再适用。控制中间时间点的健康状态会阻断因果路径，不控制又会留下混杂，回归调整在这种场景下左右为难。Robins 提出的边际结构模型，简称 MSM，和 G 方法的纵向推广是处理这类问题的标准框架。MSM 使用逆概率加权的思路，但权重反映的是整个处理历史的概率，而非单一时间点的倾向得分。G 方法的纵向版本则通过迭代的条件期望计算来处理时变混杂。Hernán and Robins [13] 的教科书对两种方法都有系统的讲解。

准实验方法在处理变量不受研究者控制时提供识别策略。工具变量利用一个外生的”推手”来估计因果效应，断点回归利用政策阈值附近的局部随机化，双重差分利用处理前后和处理组对照组的交叉比较。这些方法在经济学和政策评估中应用广泛，Angrist and Pischke [1] 是经典的入门读物。在 RHC 数据的语境下，如果某些医院的 RHC 使用率因为外部政策原因突然变化，就可以用双重差分或工具变量来利用这种外生变异估计因果效应，从而绕过可交换性假设的限制。

这些方向中的每一个都可以写一本书。本书的定位是把点处理场景下的方法讲透，让读者在自己的研究中能够选择合适的方法、正确实现、合理解读。掌握了本书介绍的九种方法之后，学习上述高级主题会顺畅得多，因为它们的核心逻辑都是本书方法的推广或变体。G 计算推广到纵向就是序贯 G 计算，IPW 推广到纵向就是 MSM 的时变权重，AIPW 的双重稳健思想在纵向场景中同样适用。因果森林的 CATE 估计可以扩展到生存分析的框架中。理解了点处理场景下每种方法的设计动机和失效机制，就能在更复杂的场景中做出有根据的方法选择。

练习 10.1 选择本书中任意两种方法在 RHC 数据上的估计结果，计算它们的点估计差异和置信区间重叠程度。讨论这种差异可能来自哪些来源：方法本身的差异、变量集的差异、目标估计量的差异，还是抽样变异。

解 以 G 计算的 RD = 0.052，95% CI [0.027, 0.082] 和 AIPW 的 RD = 0.044，95% CI [0.017, 0.072] 为例。点估计差 0.008，两个置信区间重叠区间为 [0.027, 0.072]，重叠很大。差异来源包括：G 计算只依赖结果模型，AIPW 同时使用了处理模型的 IPW 校正项，该校正项把 G 计算的估计往下拉了约 0.008。两者的变量集相同，目标估计量都是全人群的 ATE。置信区间的大面积重叠说明差异可以被抽样变异解释。

全书方法速查

探索性分析：回归调整，快速定位关键混杂变量，用作基准线。
需要边际效应：G 计算，直接在概率尺度上输出风险差，绕过非压缩性。
对结果模型没信心：IPW 或 OW，只需建处理模型。极端权重用 OW。
对任何单一模型没信心：AIPW / DML / TMLE，双重稳健，两根保险绳。
需要机器学习灵活性：DML 或 TMLE，配合 Super Learner 降低函数形式风险。
探索谁获益谁受害：因果森林，输出个体化 CATE 和变量重要性排名。
检验结论稳健性：E-value + sensemakr，量化翻盘所需的混杂强度。

笔记 回顾全书的方法演进，可以看到一条清晰的脉络。回归调整把全部赌注压在一个结果模型上；G 计算用标准化改进了效应提取方式但仍然依赖同一个模型；倾向得分方法把赌注从结果模型转移到处理模型；AIPW 同时建两个模型实现了双重稳健；DML 和 TMLE 用机器学习替代了参数模型，降低了函数形式错误的风险；因果森林在此基础上进一步拆解了平均效应，揭示了个体层面的异质性。每一步都是对前一步某个弱点的改进，后来的方法并不”替代”前面的方法，而是在更宽的适用范围内工作。理解了回归的局限才能理解 G 计算为什么存在，理解了单一模型的风险才能理解双重稳健的价值，理解了参数模型的天花板才能理解机器学习嵌入因果推断的必要性。

10.8 结语

本书从”相关不等于因果”这句每个研究者都听过的警告出发，在同一份 RHC 数据上依次展开了九种因果推断方法。这些方法从逻辑回归的一个系数开始，到因果森林的五千多个个体化效应预测结束，覆盖了参数与非参数、单一模型与双重稳健、平均效应与异质性效应的完整光谱。全书只用了一个问题和一份数据，但走过了从 Rosenbaum-Rubin 1983 到 Chernozhukov 2018 的三十五年方法论演进。方法在变，但因果推断的核心逻辑没有变：识别假设决定了你能从数据中读出什么，模型只是执行识别的计算工具，而敏感性分析告诉你结论离”被推翻”还有多远。掌握了这三层结构，读者在面对自己的数据时就有了一个可靠的分析框架：画 DAG 确定调整集，选择与研究场景匹配的估计方法，用敏感性分析量化结论的可信边界。这三步构成了观察性研究中因果推断的完整工作流。

本章知识地图

表 10.2: 全书核心概念与常见误解总览

核心概念	核心内容	常见误解	为什么错
跨方法收敛	九种方法方向一致增强因果结论的可信度	选数字最小或最大的那个方法就行	方法购物膨胀假阳性率，主分析应事先确定
三角测量	用不同假设的方法从不同角度逼近同一因果量	方法越多结论越可靠	如果所有方法共享同一个错误假设，比如遗漏了同一个混杂，收敛也可以是假的
E-value 与稳健性	未测量混杂需要 RR = 1.42 才能翻盘，保护强度中等	E-value 大就证明没有未测量混杂	E-value 只量化翻盘所需的混杂强度，不排除这种混杂存在
方法选择	根据研究场景匹配方法，双重稳健方法是现代因果分析的默认选项	存在一种普遍最优的方法	每种方法有自己的假设和局限，选择看的是方法和数据的匹配度
RHC 结论	RHC 可能有害但残余混杂不能排除，结论与 Connors 1996 一致	九种方法一致就等于因果关系确立	所有方法都依赖可交换性假设，未测量混杂是共同的盲区
识别与估计分离	识别假设决定能从数据读出什么，模型只是执行计算的工具	用了高级模型就不需要担心假设	Super Learner 和因果森林不能替代可交换性，假设错了模型再好也没用
模型设定正确	参数方法要求函数形式与真实数据生成过程一致	变量放对了就够了	变量正确但函数形式错误仍会产生设定偏误，高维场景下几乎不可能手动设定全对
双重稳健性	两个模型至少一个对就能给出一致估计	双重稳健 = 万无一失	两个模型同时错时保护失效，用参数模型时两个模型的误差往往高度相关
正值性	每个协变量组合下接受处理和不接受处理的概率都大于零	没有精确的 0 或 1 就没问题	接近 0 或 1 就足以让 IPW 方差爆炸，重叠权重通过压低极端个体权重来缓解
效应异质性	ATE 可能掩盖个体差异，CATE 揭示谁获益谁受害	ATE 适用于每一个个体	ATE 是人群平均，如果效应正负相消，ATE 可能接近零但个体效应很大
条件 OR vs 边际 RD	回归给条件 OR，G 计算和后续方法给边际 RD	控制混杂后 OR 下降就是混杂被消除	非压缩性让条件 OR 和边际 OR 在非线性模型中天然不等，下降可能是数学性质而非混杂

核心概念	核心内容	常见误解	为什么错
敏感性分析定位	量化翻盘所需的混杂强度，是压力测试而非验证	做了敏感性分析就证明了因果关系	敏感性分析不能排除未测量混杂存在，只能量化推翻结论的门槛
论文报告规范	主分析事先确定，多方法比较作为敏感性分析附在附表	跑完所有方法挑最好看的报告	方法购物等价于多重检验，膨胀假阳性率

附录 A 附录

A.1 R 包版本清单

表 A.1 列出了本书各章使用的全部 R 包及其用途。建议使用 R ≥ 4.5.1，并通过 `install.packages()` 安装最新稳定版。

表 A.1: 全书 R 包清单

R 包	版本	用途	章节
tidyverse	≥ 2.0.0	数据读写、清洗、管道操作、ggplot2 可视化	1–10
here	≥ 1.0.1	项目根目录路径管理	1–10
tableone	≥ 0.13	Table 1 基线比较与 SMD 计算	1
dagitty	≥ 0.3	DAG 定义与最小调整集推导	2
broom	≥ 1.0.5	模型输出整理（tidy / glance）	3
MatchIt	≥ 4.5	倾向得分匹配	5
WeightIt	≥ 1.0	倾向得分加权（IPW / OW）	5
cobalt	≥ 4.5	协变量平衡诊断与 Love plot	5
SuperLearner	≥ 2.0	集成学习预测框架	7
DoubleML	≥ 0.7	Double Machine Learning 框架	7
mlr3	≥ 0.18	机器学习后端（DoubleML 依赖）	7
mlr3learners	≥ 0.7	mlr3 学习器扩展	7
data.table	≥ 1.15	高效数据框（DoubleML 输入格式）	7
ranger	≥ 0.16	随机森林实现（SL / DML 后端）	7
glmnet	≥ 4.1	Lasso / 弹性网络（SL 候选学习器）	7
tmle	≥ 2.0	Targeted Maximum Likelihood Estimation	7
EValue	≥ 4.1	E-value 计算	8
sensemakr	≥ 0.1.4	遗漏变量偏差分析与等高线图	8
grf	≥ 2.3	因果森林与 CATE 估计	9
ggplot2	≥ 3.5	全书图形绑定（含于 tidyverse）	1–10

A.2 各章完整 R 代码

以下代码按章节汇编，每章的全部 `rcode` 片段合并为一个可直接运行的脚本。运行前请确保工作目录为项目根目录（含 `data/rhc.csv`），并已安装上表所列的全部 R 包。

A.2.1 第 1 章：问题与数据

```r
# ===== 第 1 章：问题与数据 =====
library(tidyverse)
set.seed(2026)

# 读入数据——here::here() 保证路径相对于项目根目录
```

```r
d <- read_csv(here::here("data", "rhc.csv"), show_col_types = FALSE)
dim(d)  # 5735 行 x 49 列

# 创建二分类结局变量
d <- d |> mutate(death180_bin = ifelse(death180 == "Yes", 1, 0))

# 预览 10 个关键变量的前 6 行
key_vars <- c("rhc", "death180", "age", "sex", "apache_score",
              "blood_pressure", "creatinine", "albumin",
              "heart_rate", "respiratory_rate")
head(d[, key_vars], 6)

# --- 基线比较 Table 1 ---
library(tableone)

# 选取 12 个关键协变量
vars <- c("age", "sex", "apache_score", "blood_pressure",
          "heart_rate", "respiratory_rate", "creatinine",
          "albumin", "hematocrit", "wbc", "temperature",
          "das_index")

# 按 RHC 分组计算 Table 1，同时输出 SMD
d <- d |> mutate(rhc_label = ifelse(rhc == 1, "RHC", "No RHC"))
tab1 <- CreateTableOne(vars = vars, strata = "rhc_label",
                       data = d, test = FALSE, smd = TRUE)
print(tab1, smd = TRUE)

# --- 粗死亡率 ---
# 按 RHC 分组计算 180 天死亡率
d |>
  group_by(rhc) |>
  summarise(
    n        = n(),
    deaths   = sum(death180_bin),
    mortality = mean(death180_bin),
    .groups = "drop"
  )
```

A.2.2 第 2 章：因果结构与识别条件

```r
# ===== 第 2 章：因果结构与识别条件 =====
set.seed(2026)
library(dagitty)

# 定义 RHC 的因果结构——三组协变量都是混杂
# 每组既影响医生是否决定上 RHC，也影响病人结局
rhc_dag <- dagitty("dag {
  severity   [pos=\"1,0\"]
```

```
  comorbidity [pos=\"2,0\"]
  demographics [pos=\"0,0\"]
  A           [pos=\"0.5,1.5\"]
  Y           [pos=\"2,1.5\"]
  severity -> A
  severity -> Y
  comorbidity -> A
  comorbidity -> Y
  demographics -> A
  demographics -> Y
  A -> Y
}")
exposures(rhc_dag) <- "A"
outcomes(rhc_dag)  <- "Y"

# dagitty 自动推导最小调整集
adjustmentSets(rhc_dag, type = "minimal")
```

A.2.3 第 3 章：回归调整

```
# ===== 第 3 章：回归调整 =====
set.seed(2026)
library(tidyverse)
library(broom)

d <- read_csv(here::here("data", "rhc.csv"), show_col_types = FALSE) |>
  mutate(death180_bin = if_else(death180 == "Yes", 1L, 0L),
         sex_bin      = if_else(sex == "Male", 1L, 0L))

# 模型 1：粗模型，只看 RHC 与死亡的边际关联
m1 <- glm(death180_bin ~ rhc, data = d, family = binomial)

# 模型 2：加入人口学——年龄和性别本身是混杂还是精度变量？
m2 <- glm(death180_bin ~ rhc + age + sex_bin,
          data = d, family = binomial)

# 模型 3：加入疾病严重度——APACHE 和 GCS 是最强的混杂源
m3 <- glm(death180_bin ~ rhc + age + sex_bin +
            apache_score + glasgow_coma_score,
          data = d, family = binomial)

# 模型 4：加入合并症和全部生理指标
m4 <- glm(death180_bin ~ rhc + age + sex_bin +
            apache_score + glasgow_coma_score +
            cancer + cardiovascular + congestive_hf + dementia +
            pulmonary + renal + hepatic + blood_pressure +
            heart_rate + respiratory_rate + temperature +
            albumin + creatinine + bilirubin + wbc + hematocrit +
```

```
            das_index + dnr_status + medical_insurance + race +
            income + edu + transfer_hx + mi + gi_bleed +
            tumor + immunosupperssion + psychiatric,
          data = d, family = binomial)

# 提取四个模型中 RHC 的 OR 和 95% CI
bind_rows(
  tidy(m1, conf.int = TRUE, exponentiate = TRUE) |>
    filter(term == "rhc") |> mutate(model = "Model 1"),
  tidy(m2, conf.int = TRUE, exponentiate = TRUE) |>
    filter(term == "rhc") |> mutate(model = "Model 2"),
  tidy(m3, conf.int = TRUE, exponentiate = TRUE) |>
    filter(term == "rhc") |> mutate(model = "Model 3"),
  tidy(m4, conf.int = TRUE, exponentiate = TRUE) |>
    filter(term == "rhc") |> mutate(model = "Model 4")
) |> select(model, estimate, conf.low, conf.high)

# --- 练习: APACHE 二次项 ---
# 在模型 3 基础上加入 APACHE 的二次项
m3b <- glm(death180_bin ~ rhc + age + sex_bin +
             apache_score + I(apache_score^2) +
             glasgow_coma_score,
           data = d, family = binomial)

# 比较 RHC 的 OR
cat("Model 3  OR:", exp(coef(m3)["rhc"]), "\n")
cat("Model 3b OR:", exp(coef(m3b)["rhc"]), "\n")
cat("AIC  M3:", AIC(m3), " M3b:", AIC(m3b), "\n")
```

A.2.4 第 4 章: G 计算

```
# ===== 第 4 章: G 计算 =====
set.seed(2026)
library(tidyverse)

d <- read_csv(here::here("data", "rhc.csv"), show_col_types = FALSE) |>
  mutate(death180_bin = if_else(death180 == "Yes", 1L, 0L),
         sex_bin      = if_else(sex == "Male", 1L, 0L))

# 建模: 与第 3 章模型 4 相同的结局模型
# G 计算的全部"因果推断负担"都压在这个模型上
outcome_mod <- glm(death180_bin ~ rhc + age + sex_bin +
  apache_score + glasgow_coma_score +
  cancer + cardiovascular + congestive_hf + dementia +
  pulmonary + renal + hepatic + blood_pressure +
  heart_rate + respiratory_rate + temperature +
  albumin + creatinine + bilirubin + wbc + hematocrit +
  das_index + dnr_status + medical_insurance + race +
```

```r
  income + edu + transfer_hx + mi + gi_bleed +
  tumor + immunosupperssion + psychiatric,
  data = d, family = binomial)

# 预测反事实：构造两个"平行世界"的数据集
# 关键操作——只改处理变量，协变量保持每个人的真实值
d1 <- d |> mutate(rhc = 1L)   # 所有人接受 RHC
d0 <- d |> mutate(rhc = 0L)   # 所有人不接受 RHC

Y1 <- predict(outcome_mod, newdata = d1, type = "response")
Y0 <- predict(outcome_mod, newdata = d0, type = "response")

# 边际化：对全人群取算术平均
EY1 <- mean(Y1)
EY0 <- mean(Y0)
RD  <- EY1 - EY0

cat("E[Y(1)] =", round(EY1, 4), "\n")
cat("E[Y(0)] =", round(EY0, 4), "\n")
cat("Risk Difference =", round(RD, 4), "\n")

# --- Bootstrap 置信区间 ---
# Bootstrap 置信区间：重复整个 G 计算流程 1000 次
# 每次有放回抽样 -> 重新拟合模型 -> 重新预测 -> 重新取均值
n_boot <- 1000
boot_rd <- numeric(n_boot)

for (i in seq_len(n_boot)) {
  idx <- sample(nrow(d), replace = TRUE)
  bd  <- d[idx, ]

  mod <- glm(death180_bin ~ rhc + age + sex_bin +
    apache_score + glasgow_coma_score +
    cancer + cardiovascular + congestive_hf + dementia +
    pulmonary + renal + hepatic + blood_pressure +
    heart_rate + respiratory_rate + temperature +
    albumin + creatinine + bilirubin + wbc + hematocrit +
    das_index + dnr_status + medical_insurance + race +
    income + edu + transfer_hx + mi + gi_bleed +
    tumor + immunosupperssion + psychiatric,
    data = bd, family = binomial)

  bd1 <- bd |> mutate(rhc = 1L)
  bd0 <- bd |> mutate(rhc = 0L)
  boot_rd[i] <- mean(predict(mod, bd1, "response")) -
                mean(predict(mod, bd0, "response"))
}

ci <- quantile(boot_rd, c(0.025, 0.975))
```

```r
cat("Bootstrap 95% CI:", round(ci, 4), "\n")

# --- 练习：交互项 ---
# 在结局模型中加入 RHC 与 APACHE 的交互项
# 如果交互项显著，说明 RHC 的效应因病情严重程度而异
outcome_mod2 <- glm(death180_bin ~ rhc * apache_score +
  age + sex_bin + glasgow_coma_score +
  cancer + cardiovascular + congestive_hf + dementia +
  pulmonary + renal + hepatic + blood_pressure +
  heart_rate + respiratory_rate + temperature +
  albumin + creatinine + bilirubin + wbc + hematocrit +
  das_index + dnr_status + medical_insurance + race +
  income + edu + transfer_hx + mi + gi_bleed +
  tumor + immunosupperssion + psychiatric,
  data = d, family = binomial)

d1 <- d |> mutate(rhc = 1L)
d0 <- d |> mutate(rhc = 0L)
RD2 <- mean(predict(outcome_mod2, d1, "response")) -
       mean(predict(outcome_mod2, d0, "response"))

cat("RD (no interaction):", round(0.0595, 4), "\n")
cat("RD (with interaction):", round(RD2, 4), "\n")
cat("Change:", round((RD2 - 0.0595) / 0.0595 * 100, 1), "%\n")
```

A.2.5 第 5 章：倾向得分

```r
# ===== 第 5 章：倾向得分——匹配、加权与平衡诊断 =====
library(tidyverse)
set.seed(2026)

d <- read_csv(here::here("data", "rhc.csv"), show_col_types = FALSE) |>
  mutate(death180_bin = ifelse(death180 == "Yes", 1, 0))

# 38 个协变量——第 2 章 DAG 确定的调整集
covs <- c("age", "sex", "edu", "das_index", "apache_score",
  "glasgow_coma_score", "blood_pressure", "wbc", "heart_rate",
  "respiratory_rate", "temperature", "pa_o2vs_fio2",
  "albumin", "hematocrit", "bilirubin", "creatinine",
  "sodium", "potassium", "pa_co2", "ph", "weight",
  "dnr_status", "medical_insurance", "race", "income",
  "cancer", "cardiovascular", "congestive_hf", "dementia",
  "psychiatric", "pulmonary", "renal", "hepatic",
  "gi_bleed", "tumor", "immunosupperssion", "transfer_hx", "mi")

fml <- as.formula(paste("rhc ~", paste(covs, collapse = " + ")))

# 倾向得分模型——对处理分配机制建模，不是对结局建模
```

```r
ps_model <- glm(fml, data = d, family = binomial)
d$ps <- predict(ps_model, type = "response")
summary(d$ps)

# --- PSM ---
library(MatchIt)

# 1:1 最近邻匹配，卡钳值 0.2 倍 logit PS 标准差
m_out <- matchit(fml, data = d, method = "nearest",
                 distance = "glm", caliper = 0.2, ratio = 1)

# 匹配后样本
m_data <- match.data(m_out)
cat("Matched sample:", nrow(m_data), "\n")
cat("RHC:", sum(m_data$rhc), "  No RHC:", sum(m_data$rhc == 0), "\n")

# 匹配样本上的风险差
rd_psm <- mean(m_data$death180_bin[m_data$rhc == 1]) -
          mean(m_data$death180_bin[m_data$rhc == 0])
cat("PSM Risk Difference:", round(rd_psm, 4), "\n")

# --- IPW ---
library(WeightIt)

# IPW: 估计 ATE
w_ipw <- weightit(fml, data = d, method = "glm", estimand = "ATE")
summary(w_ipw)

d$w_ipw <- w_ipw$weights

# 加权后的风险差
ate_ipw <- weighted.mean(d$death180_bin[d$rhc == 1], d$w_ipw[d$rhc == 1]) -
           weighted.mean(d$death180_bin[d$rhc == 0], d$w_ipw[d$rhc == 0])
cat("IPW Risk Difference:", round(ate_ipw, 4), "\n")

# --- OW ---
# OW: 估计 ATO，即重叠人群的平均效应
w_ow <- weightit(fml, data = d, method = "glm", estimand = "ATO")
summary(w_ow)

d$w_ow <- w_ow$weights

ate_ow <- weighted.mean(d$death180_bin[d$rhc == 1], d$w_ow[d$rhc == 1]) -
          weighted.mean(d$death180_bin[d$rhc == 0], d$w_ow[d$rhc == 0])
cat("OW Risk Difference:", round(ate_ow, 4), "\n")

# --- Love plot ---
library(cobalt)

```

```r
# Love plot: 同时展示 PSM / IPW / OW 的平衡
love.plot(fml, data = d,
          stats = "m", abs = TRUE,
          thresholds = c(m = 0.1),
          weights = list(PSM = m_out, IPW = w_ipw, OW = w_ow),
          colors = c("#999999", "#EF6548", "#4292C6", "#66C2A5"),
          shapes = c(17, 16, 15, 18),
          sample.names = c("Unadjusted", "PSM", "IPW", "OW"))
```

A.2.6 第 6 章：双重稳健 AIPW

```r
# ===== 第 6 章: 双重稳健 AIPW =====
set.seed(2026)
library(tidyverse)

d <- read_csv(here::here("data", "rhc.csv"), show_col_types = FALSE) |>
  mutate(death180_bin = if_else(death180 == "Yes", 1L, 0L),
         sex_bin      = if_else(sex == "Male", 1L, 0L),
         cancer_bin   = if_else(cancer == "No", 0L, 1L))

covs <- c("age", "sex_bin", "cancer_bin", "cardiovascular",
          "congestive_hf", "dementia", "psychiatric", "pulmonary",
          "renal", "hepatic", "gi_bleed", "tumor",
          "immunosupperssion", "transfer_hx", "mi",
          "apache_score", "glasgow_coma_score", "blood_pressure",
          "heart_rate", "respiratory_rate", "temperature",
          "albumin", "creatinine", "bilirubin", "wbc",
          "hematocrit", "das_index", "weight")

# 第一根绳: 结果模型——预测 Y 在 A 和 L 条件下的期望
out_mod <- glm(death180_bin ~ rhc + .,
               data = d |> select(death180_bin, rhc, all_of(covs)),
               family = binomial)

# 第二根绳: 处理模型——预测谁更可能接受 RHC
ps_mod <- glm(rhc ~ .,
              data = d |> select(rhc, all_of(covs)),
              family = binomial)
d$ps <- predict(ps_mod, type = "response")

# 反事实预测: 让每个人分别"接受"和"不接受" RHC
d1 <- d0 <- d
d1$rhc <- 1; d0$rhc <- 0
d$m1 <- predict(out_mod, newdata = d1, type = "response")
d$m0 <- predict(out_mod, newdata = d0, type = "response")

# 组装 AIPW: 三个部分逐项计算
d$aipw_score <- with(d, {
```

```r
  (m1 - m0) +                                  # A: G 计算部分
    rhc / ps * (death180_bin - m1) -           # B: 处理组校正
    (1 - rhc) / (1 - ps) * (death180_bin - m0)  # C: 对照组校正
})

ate_aipw <- mean(d$aipw_score)
se_aipw  <- sd(d$aipw_score) / sqrt(nrow(d))
cat("ATE:", round(ate_aipw, 4),
    " SE:", round(se_aipw, 4),
    " 95% CI: [", round(ate_aipw - 1.96*se_aipw, 4),
    ",", round(ate_aipw + 1.96*se_aipw, 4), "]\n")

# --- 练习: 倾向得分截断 ---
# 倾向得分截断——牺牲少量偏差换取方差稳定
d$ps_trim <- pmax(0.025, pmin(0.975, d$ps))

d$aipw_trim <- with(d, {
  (m1 - m0) +
    rhc / ps_trim * (death180_bin - m1) -
    (1 - rhc) / (1 - ps_trim) * (death180_bin - m0)
})

ate_trim <- mean(d$aipw_trim)
se_trim  <- sd(d$aipw_trim) / sqrt(nrow(d))
cat("Trimmed ATE:", round(ate_trim, 4),
    " SE:", round(se_trim, 4), "\n")
```

A.2.7 第 7 章：ML 增强——Super Learner、DML 与 TMLE

```r
# ===== 第 7 章: ML 增强 =====

# --- Super Learner ---
set.seed(2026)
library(SuperLearner)

# 结果模型: 预测 death180 在 rhc 和 28 个协变量条件下的概率
sl_out <- SuperLearner(
  Y = d$death180_bin,
  X = d |> select(rhc, all_of(covs)),
  family = binomial(),
  # 四个候选学习器: 均值基准、逻辑回归、Lasso、随机森林
  SL.library = c("SL.mean", "SL.glm", "SL.glmnet", "SL.ranger"),
  cvControl = list(V = 5)   # 5 折交叉验证评估各学习器
)

# 查看各学习器的交叉验证风险和集成权重
sl_out

```

```r
# --- DML ---
set.seed(2026)
library(DoubleML); library(mlr3); library(mlr3learners)
library(data.table)

# DoubleML 要求 data.table 格式
dt <- as.data.table(d |> select(death180_bin, rhc, all_of(covs)))

dml_data <- DoubleMLData$new(
  data = dt,
  y_col = "death180_bin",   # 结局
  d_cols = "rhc",           # 处理
  x_cols = covs             # 28 个协变量
)

# IRM: Interactive Regression Model, 适用于二分类处理
# ml_g 估计 E[Y|X], ml_m 估计 P(A=1|X)
dml_irm <- DoubleMLIRM$new(
  data = dml_data,
  ml_g = lrn("classif.ranger", predict_type = "prob", num.trees = 500),
  ml_m = lrn("classif.ranger", predict_type = "prob", num.trees = 500),
  score = "ATE",
  n_folds = 5,    # 5 折交叉拟合
  n_rep = 3       # 重复 3 次取平均, 减小随机分裂的波动
)

dml_irm$fit()
dml_irm$summary()
print(dml_irm$confint())

# --- TMLE ---
set.seed(2026)
library(tmle); library(SuperLearner)

SL_lib <- c("SL.glm", "SL.glmnet", "SL.ranger", "SL.mean")

tmle_fit <- tmle(
  Y = d$death180_bin,
  A = d$rhc,
  W = d |> select(all_of(covs)),
  Q.SL.library = SL_lib,   # 结果模型用 Super Learner
  g.SL.library = SL_lib,   # 倾向得分模型也用 Super Learner
  family = "binomial"       # 二分类结局
)

tmle_fit
```

A.2.8 第 8 章：敏感性分析

```r
# ===== 第 8 章：敏感性分析 =====
set.seed(2026)
library(tidyverse)
library(EValue)

d <- read_csv(here::here("data", "rhc.csv"), show_col_types = FALSE) |>
  mutate(death180_bin = if_else(death180 == "Yes", 1L, 0L))

# 对照组基线死亡率
r0 <- mean(d$death180_bin[d$rhc == 0])

# AIPW 估计的风险差和 95% CI（来自第 6 章）
ate  <- 0.0442
se   <- 0.0139
ci_lo <- ate - 1.96 * se
ci_hi <- ate + 1.96 * se

# 转换为相对风险尺度——E-value 需要 RR 作为输入
rr_point <- (r0 + ate) / r0
rr_lo    <- (r0 + ci_lo) / r0
rr_hi    <- (r0 + ci_hi) / r0

# 计算 E-value：点估计和置信区间下界各一个
ev <- evalues.RR(rr_point, lo = rr_lo, hi = rr_hi)
print(ev)

# --- sensemakr ---
library(sensemakr)

d <- d |>
  mutate(sex_bin    = if_else(sex == "Male", 1L, 0L),
         cancer_bin = if_else(cancer == "No", 0L, 1L))

covs <- c("age", "sex_bin", "cancer_bin", "cardiovascular",
          "congestive_hf", "dementia", "psychiatric", "pulmonary",
          "renal", "hepatic", "gi_bleed", "tumor",
          "immunosupperssion", "transfer_hx", "mi",
          "apache_score", "glasgow_coma_score", "blood_pressure",
          "heart_rate", "respiratory_rate", "temperature",
          "albumin", "creatinine", "bilirubin", "wbc",
          "hematocrit", "das_index", "weight")

# 线性概率模型——sensemakr 需要 lm 对象
lin_mod <- lm(death180_bin ~ rhc + .,
              data = d |> select(death180_bin, rhc, all_of(covs)))

# 以 APACHE 评分为基准，分别看 1 倍、2 倍、3 倍强度的混杂
```

```r
sens <- sensemakr(model = lin_mod,
                  treatment = "rhc",
                  benchmark_covariates = "apache_score",
                  kd = c(1, 2, 3))
summary(sens)

# --- 练习：IPW 的 E-value ---
library(EValue)
# IPW 的风险差和标准误
ate_ipw <- 0.032; se_ipw <- 0.022
r0 <- 0.4647  # 对照组基线死亡率

# 转换为 RR 尺度
rr_ipw <- (r0 + ate_ipw) / r0
rr_lo  <- (r0 + ate_ipw - 1.96 * se_ipw) / r0

ev_ipw <- evalues.RR(rr_ipw, lo = rr_lo)
print(ev_ipw)
```

A.2.9 第 9 章：因果森林与异质性

```r
# ===== 第 9 章：因果森林与异质性 =====
set.seed(2026)
library(tidyverse)
library(grf)

d <- read_csv(here::here("data", "rhc.csv"), show_col_types = FALSE) |>
  mutate(death180_bin = if_else(death180 == "Yes", 1L, 0L),
         sex_bin      = if_else(sex == "Male", 1L, 0L),
         cancer_bin   = if_else(cancer == "No", 0L, 1L))

covs <- c("age", "sex_bin", "cancer_bin", "cardiovascular",
          "congestive_hf", "dementia", "psychiatric", "pulmonary",
          "renal", "hepatic", "gi_bleed", "tumor",
          "immunosupperssion", "transfer_hx", "mi",
          "apache_score", "glasgow_coma_score", "blood_pressure",
          "heart_rate", "respiratory_rate", "temperature",
          "albumin", "creatinine", "bilirubin", "wbc",
          "hematocrit", "das_index", "weight")

X <- as.matrix(d[, covs])
W <- d$rhc
Y <- d$death180_bin

# 2000 棵树，honesty = TRUE 保证统计推断合法
cf <- causal_forest(X, Y, W,
                    num.trees = 2000,
                    honesty = TRUE,
```

```r
                      seed = 2026)

# 每名患者的 CATE 预测
cate <- predict(cf)$predictions
cat("CATE 均值:", round(mean(cate), 4),
    " SD:", round(sd(cate), 4), "\n")
cat("CATE > 0 (受害):", round(mean(cate > 0)*100, 1), "%\n")
cat("CATE < 0 (获益):", round(mean(cate < 0)*100, 1), "%\n")

# --- 变量重要性 ---
# 变量重要性: 哪些协变量驱动了 CATE 的异质性
vimp <- variable_importance(cf)
vimp_df <- data.frame(Variable = covs,
                      Importance = as.numeric(vimp)) |>
  arrange(desc(Importance))
cat("Top 5 变量:\n")
print(head(vimp_df, 5))

# --- BLP 检验 ---
# BLP 检验: CATE 的异质性是否具有统计学意义
blp <- test_calibration(cf)
print(blp)

# --- 平均效应 ---
# 因果森林的 AIPW 平均效应
ate_cf <- average_treatment_effect(cf, target.sample = "all")
cat("ATE:", round(ate_cf[1], 4),
    " SE:", round(ate_cf[2], 4),
    " 95% CI: [", round(ate_cf[1] - 1.96*ate_cf[2], 4),
    ",", round(ate_cf[1] + 1.96*ate_cf[2], 4), "]\n")

# --- 亚组分析 ---
# 按 CATE 五等分做亚组分析
d$cate <- cate
d$cate_q <- cut(d$cate,
    breaks = quantile(d$cate, probs = seq(0, 1, 0.2)),
    labels = c("Q1", "Q2", "Q3", "Q4", "Q5"),
    include.lowest = TRUE)

for (q in c("Q1", "Q5")) {
  idx <- which(d$cate_q == q)
  ate_q <- average_treatment_effect(cf, subset = idx,
                                    target.sample = "all")
  cat(q, ": ATE =", round(ate_q[1], 4),
      ", 95% CI = [", round(ate_q[1] - 1.96*ate_q[2], 4),
      ",", round(ate_q[1] + 1.96*ate_q[2], 4), "]\n")
}
```

A.2.10 第 10 章：全书汇总

```r
# ===== 第 10 章：全书汇总——森林图 =====
set.seed(2026)
library(ggplot2)

# 从第 3--9 章收集的真实估计值
methods <- c(
  "G Computation (Ch.4)", "PSM (Ch.5)", "IPW (Ch.5)",
  "Overlap Weights (Ch.5)", "AIPW (Ch.6)", "DML (Ch.7)",
  "TMLE (Ch.7)", "Causal Forest (Ch.9)")
est   <- c(0.052, 0.076, 0.055, 0.061, 0.044, 0.040, 0.088, 0.044)
ci_lo <- c(0.027, 0.041, 0.025, 0.033, 0.017, 0.014, 0.074, 0.020)
ci_hi <- c(0.082, 0.109, 0.085, 0.089, 0.072, 0.065, 0.103, 0.068)

df <- data.frame(method = factor(methods, levels = rev(methods)),
                 est = est, lo = ci_lo, hi = ci_hi)

ggplot(df, aes(x = est, y = method)) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "grey50") +
  geom_point(size = 3, color = "#EF6548") +
  geom_errorbar(aes(xmin = lo, xmax = hi), width = 0.25,
                color = "#4292C6", linewidth = 0.7, orientation = "y") +
  labs(x = "Risk Difference (RD)", y = NULL,
       title = "ATE Estimates Across Eight Methods") +
  scale_x_continuous(breaks = seq(-0.02, 0.12, 0.02)) +
  theme_minimal(base_size = 14, base_family = "serif") +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major.y = element_blank(),
        plot.title = element_text(hjust = 0.5))
```

Bibliography

[1] Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

[2] Susan Athey and Guido Imbens. "Recursive partitioning for heterogeneous causal effects". In: *PNAS* 113.27 (2016), pp. 7353–7360.

[3] Susan Athey, Julie Tibshirani, and Stefan Wager. "Generalized random forests". In: *Annals of Statistics* 47.2 (2019), pp. 1148–1178.

[4] Peter C. Austin. "An introduction to propensity score methods for reducing the effects of confounding in observational studies". In: *Multivariate Behavioral Research* 46.3 (2011), pp. 399–424.

[5] Philipp Bach et al. "DoubleML: An object-oriented implementation of double machine learning in R". In: *Journal of Statistical Software* 108.3 (2024), pp. 1–56.

[6] Heejung Bang and James M. Robins. "Doubly robust estimation in missing data and causal inference models". In: *Biometrics* 61.4 (2005), pp. 962–973.

[7] Victor Chernozhukov et al. "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21.1 (2018), pp. C1–C68.

[8] Carlos Cinelli and Chad Hazlett. "Making sense of sensitivity: Extending omitted variable bias". In: *Journal of the Royal Statistical Society: Series B* 82.1 (2020), pp. 39–67.

[9] Alfred F. Connors, Theodore Speroff, Neal V. Dawson, et al. "The effectiveness of right heart catheterization in the initial care of critically ill patients". In: *JAMA* 276.11 (1996), pp. 889–897.

[10] Michele Jonsson Funk et al. "Doubly robust estimation of causal effects". In: *American Journal of Epidemiology* 173.7 (2011), pp. 761–767.

[11] Sander Greenland, Judea Pearl, and James M. Robins. "Causal diagrams for epidemiologic research". In: *Epidemiology* 10.1 (1999), pp. 37–48.

[12] Susan Gruber and Mark J. van der Laan. "tmle: An R package for targeted maximum likelihood estimation". In: *Journal of Statistical Software* 51.13 (2012), pp. 1–35.

[13] Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

[14] Paul W. Holland. "Statistics and causal inference". In: *Journal of the American Statistical Association* 81.396 (1986), pp. 945–960.

[15] Sören R. Künzel et al. "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *PNAS* 116.10 (2019), pp. 4156–4165.

[16] Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. "Super Learner". In: *Statistical Applications in Genetics and Molecular Biology* 6.1 (2007).

[17] Mark J. van der Laan and Daniel Rubin. "Targeted maximum likelihood learning". In: *The International Journal of Biostatistics* 2.1 (2006).

[18] Judea Pearl. "Causal diagrams for empirical research". In: *Biometrika* 82.4 (1995), pp. 669–688.

[19] Judea Pearl. *Causality: Models, Reasoning, and Inference*. 2nd. Cambridge University Press, 2009.

[20] James M. Robins. "A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect". In: *Mathematical Modelling* 7 (1986), pp. 1393–1512.

[21] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. “Estimation of regression coefficients when some regressors are not always observed”. In: *Journal of the American Statistical Association* 89.427 (1994), pp. 846–866.

[22] Paul R. Rosenbaum. *Observational Studies*. 2nd. Springer, 2002.

[23] Paul R. Rosenbaum and Donald B. Rubin. “The central role of the propensity score in observational studies for causal effects”. In: *Biometrika* 70.1 (1983), pp. 41–55.

[24] Donald B. Rubin. “Estimating causal effects of treatments in randomized and nonrandomized studies”. In: *Journal of Educational Psychology* 66.5 (1974), pp. 688–701.

[25] Johannes Textor et al. “Robust causal inference using directed acyclic graphs: the R package ‘dagitty’”. In: *International Journal of Epidemiology* 45.6 (2016), pp. 1887–1894.

[26] Tyler J. VanderWeele. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, 2015.

[27] Tyler J. VanderWeele and Peng Ding. “Sensitivity analysis in observational research: Introducing the E-value”. In: *Annals of Internal Medicine* 167.4 (2017), pp. 268–274.

[28] Stefan Wager and Susan Athey. “Estimation and inference of heterogeneous treatment effects using random forests”. In: *JASA* 113.523 (2018), pp. 1228–1242.