分类号: 学校代码: 10165

密 级: 学 号: 201911010555

遼寧師範大學

硕 士 学 位 论 文

基于 Cox 比例风险模型的因果推断研究

Research on Causal Inference Based on Cox Proportional Hazard Model

作 者 姓 名: 谢 新 惠

学科、专业: 数 学 、概 率 论 与 数 理 统 计

研 究 方 向: 应 用 统 计

导 师 姓 名: 侯 文

2022 年 4 月

摘　要

探寻事物之间的因果关系是许多领域研究的最终目的.而因果推断问题对统计学的发展起到至关重要的作用.由于现实因素的影响,因果推断主要是利用观察数据研究事物之间的因果关系.潜在结果模型是研究因果推断的主要模型之一,通过对比分析潜在结果之间的差异,得到处理效应.而 Cox 比例风险模型是生存分析中最重要的模型之一,对很多研究领域都有着重要意义.本文基于 Cox 比例风险模型,利用逆概率加权法对观察数据中的协变量进行调整,并利用调整后的数据计算平均处理效应,研究处理变量与生存时间之间的因果关系.

观察数据中,处理组和对照组之间协变量的不平衡会使研究结果产生偏倚,本文在反事实框架下,利用 logistic 回归估计稳定的权重,在存在删失的情况下,对原始数据进行加权,以此来平衡协变量对处理效应估计的影响.利用 Cox 比例风险模型计算基于风险差异以及受限平均时间的平均处理效应.并利用含倾向得分的 Cox 比例风险回归模型进行统计推断,计算得到处理组和对照组之间的风险比,根据所得结果分析判断处理变量与结果之间的因果关系.运用 R 软件模拟生成不同生存时间分布的数据集,并对数据集进行匹配调整,对比协变量在数据集调整前后的均值差异以及其在处理组和对照组之间的分布情况,比较平均处理效应的变化情况,以及不同删失比率对匹配结果是否存在显著影响.结果显示,调整后的变量在处理组和对照组中的分布更加均衡,平均处理效应也更能反应处理变量对生存结果的影响.最后进行实例分析,利用绝对标准化均值差异量化两组之间的平衡情况,研究匹配前后的数据集中,处理变量与生存结果之间的因果关系.

关键词：Cox 比例风险模型；处理效应；倾向得分

Research on Causal Inference Based on Cox Proportional Hazard Model

Abstract

To explore causation is the ultimate goal of many fields of research. Causal inference plays an important role in the development of statistics. Due to the influence of realistic factors, causal inference mainly studies causation based on observational data. Potential outcome model is one of the main models to study causal inference. The average treatment effect can be obtained by comparing and analyzing the difference between potential outcomes. Cox proportional hazard model is one of the most important models in survival analysis, which is of great significance to many research fields. In this thesis, the inverse probability weighted method is used to adjust the covariates in the observational data based on Cox proportional hazard model, and the average treatment effect is calculated by the adjusted data to study the causation between treatment and survival time.

In observational data, the imbalance of covariates between the treatment group and the control group would bias the research results. In this thesis, under the counterfactual framework, logistic regression is used to estimate the stabilized weights. The original data is weighted in the case of censoring, so as to balance the influence of covariates on the estimation of the average treatment effect. Cox proportional hazard model is used to calculate the average treatment effect based on risk difference and restricted mean time. The Cox model with propensity score is used for statistical inference, and the hazard ratio between the two groups is calculated to analyze the causation between the treatment and the survival outcomes. R is used to simulate and generate datasets with different distributions of survival time. Adjust the datasets, and contrast the mean difference and distribution of covariate between two groups before and after adjustment. Compare the change of the average treatment effect, and analyze whether different censoring proportions has influence on matching results. The results show that the distribution of the adjusted variables is more balanced than before between the treatment group and the control group, and the average treatment effect also better reflects the influence of treatment variables on survival outcomes. Finally, an example is analyzed. The balance between the two groups is quantified by absolute standardized mean difference, and study causation between treatment and survival outcomes before and after the dataset adjustment.

Key Words: Cox proportional hazard model; treatment effect; propensity score

目　　录

引　言

从古至今,人们从未放弃对因果解释的探究.事物之间的因果关系是许多领域不断追寻的,无论是自然科学还是社会科学.随着社会的不断发展,科学技术的不断进步,人们对因果关系的研究也在不断深入,这就需要有科学的语言来对因果关系进行描述.Wright[1]首先在统计学中研究因果推断,将路径分析引入统计学,之后发展为结构方程模型.自二十世纪以来,在统计学中从未停止过对因果推断的研究,但相比于相关关系研究取得的显著成就,因果关系的进展则相对缓慢.近些年,因果推断正在重新走进人们的视野.因果推断主要包括两个模型:因果网络和潜在结果模型,本文主要研究后者.

潜在结果模型是由 Rubin[2]提出的,在经济学、生物医学等很多领域都有应用.潜在结果模型是对于每一个可能的处理都定义一个潜在结果变量,接着定义因果关系为两个潜在结果变量的差异,称为处理效应.因果推断的主要困难是混杂因素歪曲了处理与结果之间的因果关系,随机分配机制能够均衡混杂因素在各组间的分布,因而基于随机分配的统计推断可以揭示因果关系,从而估计处理效应.但在计量经济学、社会科学和公共卫生等领域,受道德伦理的约束,处理效应的评估主要基于自然实验条件下的观察数据得以实施.科学家只观察处理,而不操纵处理,因此这类实验通常不满足随机分配的条件.这时,只对各组之间的结果进行简单比较,可能会导致总体的处理效应估计产生严重偏倚.因此利用传统的统计推断方法并不可行,我们需要一种针对观察数据的因果推断方法.

在因果推断的研究中,对观察数据的研究至关重要,但观察数据中总会存在除我们关注的处理变量之外的其他变量(协变量),其对结果存在很大影响.Schield [3]认为,我们生活在一个多元的世界,对观察数据的混淆可能是推断结论过程中的一个严重问题.因此应该把对结果造成影响的协变量进行调整,调整模型和未调整模型之间的估计系数差异的大小和符号取决于因果结构模型中的真实系数[4].在 Cummiskey [5]的例子中,协变量年龄和性别都给出了这种情况:两者都在吸烟和强迫呼气量之间的非因果路径上,不调整这些因素,吸烟对强迫呼气量的因果关系就会得出错误的结果.

对于根据观察数据估计平均处理效应这一重要问题,几代统计学家在不同的框架下对此进行了研究.当处理选择过程完全依赖于可观测的协变量时,有两大类策略来估计平均处理效应,即倾向得分估计和回归分析.倾向得分,也就是在给定协变量的情况下接受处理的概率.Rosenbaum 和 Rubin[6]表明,调整真实的倾向得分可以消除所有由于混杂造成的偏差,真实的倾向得分平衡了两组之间的协变量分布.倾向得分可以用于子分类(Rosenbaum[7])、匹配(Abadie [8])和加权(Hirano[9])等,其中子分类的思想是确定一个“最优”子类的形式,使处理和对照的个体在同一子类尽可能相似.另一类策略是回归分析,当给定

协变量的结果假设为线性模型时,且所有相关混杂因素都是可控的,处理指标的系数表示对平均处理效应的估计.一般来说,更复杂的回归模型可以用来预测不可观测的潜在结果,Blinder[10]研究中平均处理效应可以通过平均预测结果来估计.目前,通过观察数据研究处理效应估计的基本方法是:以随机化实验为基准,在反事实框架下,通过“研究设计”以模拟随机分配机制,并通过逐渐放松识别条件发展出相应的估计方法.

生存分析已经发展成为生物统计学的主要研究领域之一,并在其他领域有着非常广泛的应用,例如经济学、药理学、心理和行为科学等.生存分析是指根据实验或调查得到的数据,对生物或人的生存时间进行分析和推断,是研究生存时间与众多影响因素之间关系以及影响程度大小的一门学科.生存时间通常无法精确观测,仅能观察到存在删失的数据,其常用的分布有:指数分布、Weibull 分布、Gamma 分布、对数正态分布等.如果不知道生存时间确切的函数分布类型,但需要分析不同因素对生存时间的影响,这种情况下一般采用半参数模型.半参数模型假定协变量对生存时间的影响具有参数形式,最常用的半参数生存分析模型是加速失效时间模型和 Cox 比例风险模型.

Cox 比例风险模型是生存分析中重要的半参数回归模型之一.比例风险模型是 Cox[11]在 1972 年提出的,如今,Cox 模型是生存分析、可靠性和生活质量研究、流行病学、临床试验和生物医学等研究中最重要的模型.同时在人口学、计量经济学、金融学、生物学、老年学、保险等方面也有巨大的应用.这些都标志着 Cox 比例风险模型的巨大成功,进一步引发了生存回归模型的扩展研究以及半参数估计理论、似然原理、计数过程建模和应用的相应发展.

在加速失效时间框架内,Robins 和 Tsiatis(1991)[12]提出了一种方法来量化存在选择性非依从时实际接受处理的效果.具体来说,他们构建了一个模型,以因果参数、观察到的暴露(暴露与协变量相互作用)以及已经观察到的事件时间（在个体没有接受处理的情况下）来表示个体的事件时间.推导了一类渐近无偏的估计量,在没有删失的情况下,这类最优估计达到了模型的半参数效率界.White[13]等人根据加速失效时间模型构建了一个人工数据集,在可选择处理方案下会观察到该数据集,并使用这个反事实数据集来估计比例风险效应.然而,基于 Cox 比例风险模型框架的估计量性质尚不清楚.

在假设删失没有不可测混杂因素的条件下,Robins 和 Finkelstein[14]提出了一种在比例风险框架中估计处理效应的方法.他们在个体第一次切换或停止治疗,或失去随访时,人为地认为个体删失.然后,他们使用逆概率删失加权(IPCW)的 Cox 偏似然来获得实际接受的处理效应的估计量.Hernan 等人[15]在使用边际结构 Cox 比例风险模型的观察性研究中,为实际接受的处理效果找到了一个因果估计量.因此,在这两种情况下,已经估计了整个研究人群受到特定暴露会发生的结果.

本文中,提出了一种基于随机化的因果推理方法,在不直接假设观察处理的分配机制的情况下,研究关于在比例风险模型框架中处理对实验结果的影响.文中利用稳定的权重对观察到的数据集进行加权匹配,比较调整前后的风险差异以及受限平均时间,计算处理组与对照组之间的风险比,推断处理变量对结果的影响.利用 R 语言程序模拟,用模拟产生的数据集进行研究,并比较数据调整前后不同变量在处理组和对照组中的分布情况,考察匹配结果,推断因果关系.

本文第一章介绍基础理论,包括因果推断,Cox 比例风险模型以及 bootstrap 方法,用于后续对标准差的估计;第二章基于 Cox 比例风险模型估计权重,对数据进行加权调整,计算平均处理效应;第三章是数值模拟,通过 R 软件模拟产生数据集,其中生存时间分别利用 Weibull 分布和对数正态分布生成,通过估计权重对数据进行匹配,对比匹配前后协变量的分布情况,比较数据在调整前后的计算结果;第四章为实例分析,对 rotterdam 数据集进行加权匹配,对比调整前后协变量的平衡情况,研究激素处理与无复发生存时间的因果关系;最后是对本文内容的总结与展望.

1 基础理论

1.1 因果推断

在日常生活中,人们经常会使用因果语言进行表述,但是这种表述方式很少出现在正式情形下.在统计学中,许多教科书,特别是较老的教科书,除了在随机实验的背景下,一般会避免提到因果关系这个术语.有时候提到它主要是为了强调其相关性,但相关关系并不等同于因果关系,因此对因果语言的运用通常较少.然而对于许多使用统计方法的人来说,因果关系正是他们所探究的,而探究的基本问题是研究因果关系与应用于一个个体的行为(例如操作、治疗或干预等)的相关联.这里的个体可以是在特定时间点的一个公司、一个人或一组个体.通常认为,相同的个体在不同的时间是不同的个体.从这个角度来看,可以表述为一个个体在特定的时间点受到或暴露于特定的行为、治疗或制度,但同一个体在同一时间点也可能暴露于另一种行为、治疗或制度.首先考虑单一个体的情况,这里只考虑该个体是否进行了某一特定行为,而忽略可能出现的其他情况.本文中将暴露于特定行为简单地称为“处理”,而将另一情况称为“对照”.

给定一个个体和一组行为,将每个行为个体与一个潜在的结果联系起来,称这些结果为潜在结果[16].潜在结果与实际采取的行为相对应,最终只有一个结果会实现,因此只能观察到一个结果,其他潜在结果因为没有采取导致这一结果实现的相应行为,从而无法观察到.

1.1.1 处理效应

因果关系的量化称为处理效应.处理效应的定义包括两个重要方面.首先,处理效应的定义取决于潜在结果,但并不取决于实际观察到的结果.第二,处理效应是对同一个体、同一时间、处理后潜在结果的比较.因此,因果推断的根本难点[17]是缺失数据的存在,即对于每个个体,最多只能观察到一个潜在结果.

为了对处理效应进行估计,需要比较观察到的结果,由于每个个体只能观察到一个潜在结果,因而这就需要考虑多个个体,并且要求这些个体中的一部分暴露于某一行为,其余暴露于与之对照的行为.因此了解观察到的是某个潜在结果而不是其他结果的原因非常重要,也就是说如何决定哪些个体接受处理,如何决定实现哪些潜在结果,这时就需要考虑分配机制,通过操纵思维实验,从而定义潜在结果.

具体介绍分配机制之前,首先定义处理效应.设A为一个二元处理指标,$A_i = 1$表示个体i接受处理(处理组),$A_i = 0$表示未接受处理(对照组),用Y表示潜在结果,则个体i的潜在结果为:

$$潜在结果=\begin{cases}Y_i(0) & A_i=0\\ Y_i(1) & A_i=1\end{cases}$$

此时处理效应定义为$\rho_i=Y_i(1)-Y_i(0)$.

潜在结果$Y_i(1),Y_i(0)$的比较就是单一个体层面的处理效应,可以简单地表示为差$Y_i(1)-Y_i(0)$或比值$Y_i(1)/Y_i(0)$.但一般来说,比较可以采取不同的形式,通常所说的因果估计主要是两个潜在结果的差异在整个总体上的平均值.

ATE（Average Treatment Effects）一般称为平均处理效应,ATE 问题的出现,最早是为了研究某种治疗方法是否有效或者改善程度.总体平均处理效应定义为:

$$ATE=E[Y_i(1)-Y_i(0)]=E[Y_i(1)]-E[Y_i(0)]$$

即所有个体处理效应的期望值.由于$Y_i(1),Y_i(0)$为潜在结果,不能同时被观察到.每一个个体都只能选择接受处理或者不接受处理,因此只能观察到$Y_i(1),Y_i(0)$的其中一个结果,例如$Y_i(1)$只在$A_i=1$的情况下观察到.这时观察到的结果可以表示为:

$$Y_i=A_iY_i(1)+(1-A_i)Y_i(0)$$

设X_i和Z_i分别表示未接受处理时可观测和不可观测的特征,假设完整数据集$\{A_i,Y_i(0),Y_i(1),X_i,Z_i\},(i=1,2,\dots n)$是独立同分布的,而实际观察到的数据集为$\{A_i,Y_i,X_i\},(i=1,2,\dots n)$.一般把未观测到的潜在结果称为反事实结果,由于反事实结果不能观测到,那么估计反事实结果就是估计处理效应的关键.

1.1.2 分配机制

随机分配被认为是最直接的一类分配机制,随机试验设计通常是因果推断最可靠的基础.如果采用随机分配机制,即个体被随机安排到处理组或对照组,这样潜在结果不会对处理分配产生影响,也就是说个体的潜在结果和处理指标是独立的:

$$\{Y_i(1),Y_i(0)\}\perp A_i$$

随机分配确保了处理组和对照组的可观测特征、不可观测特征和处理效应完全独立于是否接受处理:

$$\{X_i,Z_i,\tau_i\}\perp A_i$$

这样除处理变量之外的其他混杂因素的分布在两组间是平衡的,因此结果变量之间的差异可完全归于处理变量.

虽然随机试验在研究处理效应中非常实用,但在现实研究中由于研究成本、伦理约束等问题,随机实验的使用相对较少,更为常见的是观察性试验.在研究过程中运用的数据一般为观察性数据.与随机试验相比,观察性试验中存在着大量的混杂因素,会对处理效应的估计产生影响,这时对比运算便不能直接得到处理效应,可以通过匹配方法对观察数据进行处理.Rosenbaum 和 Rubin [18]在 1983 年提出假设:要使用匹配方法,需要满足两个前提

条件:非混淆假设和共同支撑域条件.

非混淆假设:在给定观测特征后,潜在结果与处理指标独立:

$$\{Y_i(1), Y_i(0)\} \perp A_i | X_i$$

这一假设表示不会因为潜在结果的不同对是否接受处理产生影响,这个假设也称为条件独立假设.根据上式,可以得到:

$$E(Y_i(0)|A_i = 1, X) = E(Y_i(0)|A_i = 0, X) = E(Y_i(0)|X) \quad (1.1)$$

$$E(Y_i(1)|A_i = 1, X) = E(Y_i(1)|A_i = 0, X) = E(Y_i(1)|X) \quad (1.2)$$

式(1.1)、式(1.2)保证了在给定处理组和对照组中可观测特征分布相同的情况下,不可观测特征在两组间的分布也相同.

共同支撑域条件:给定观测特征$X_i = x$,个体接受处理的概率大于 0 且小于 1:

$$0 < P(A_i|X_i = x) < 1$$

共同支撑域条件保证了在给定观察特征$X_i = x$条件下,处理组和对照组同时存在,即存在一些个体接受了处理,同时也存在个体没有接受处理.

同时,为了使得到的处理效应估计无偏,还需要两个关键假设:

稳定性假设:是指不同个体的潜在结果相互独立.这一假设包括两个方面,第一,任何个体的潜在结果都不会因为分配给其他个体的处理而变化,第二,对于每个个体,每个处理水平上没有不同的形式,从而导致不同的潜在结果.

假设倾向得分模型或权重估计模型的设定是正确的:要求倾向得分模型的形式是正确的,如果模型设定错误,得到的权重就会不准确,容易产生极端权重[19].

1.1.3 倾向得分匹配

接下来介绍匹配方法,匹配方法是一种非常直观的估计处理效应的方法.1.1.2 节提到随机分配中协变量在处理组和对照组中的分布是均衡的,在得到的观察数据中,对处理组中的每个个体,找到协变量相同的对照组个体与之对应,比较两个体之间的差异从而估计处理效应,这就是匹配方法的基本原理.

倾向得分匹配法是应用最普遍的匹配法,倾向得分可以将多维观测变量变换为一维的倾向得分,然后根据倾向得分进行匹配.倾向得分是协变量为$X_i = x$的个体接受处理的概率:

$$ps(X_i = x) = P(A_i = 1|X_i = x)$$

Rosenbaum 和 Rubin[18]证明了一条重要结论:

条件独立假设$\{Y_i(1), Y_i(0)\} \perp A_i|X_i$与$\{Y_i(1), Y_i(0)\} \perp A_i|ps(X_i)$是等价的.

倾向得分能够起到降维、匹配的效果,因此只需要对一维的倾向得分$ps(X_i)$进行匹配即可.

倾向得分匹配法包括估计倾向得分、匹配前均衡检验、评估共同支撑域条件、选择

匹配方法、匹配后均衡、最后计算处理效应等六个步骤.首先简要介绍第一步:需要选择模型来估计倾向得分,一般情况下使用 Probit 模型或 Logit 模型估计个体接受处理的概率,即倾向得分.

Probit 模型估计方程:

$$P(A_i = 1|X) = \Phi(\beta X)$$

$\Phi(\cdot)$是正态分布的累积概率函数;

Logit 模型估计方程

$$P(A_i = 1|X) = F(\beta X)$$

其中$F(\beta X) = \frac{e^{\beta X}}{1+e^{\beta X}}$是 logistic 分布的累积概率函数.同时还要对变量进行选择,但并没有一个公认的标准答案来展示如何选择变量,不过通常情况下,需遵循几个原则:第一,倾向得分中应包括能够同时影响处理选择和处理结果的变量;二,倾向得分不应包括受处理选择影响的变量;三,倾向得分的估计虽然一般使用 Probit 或 Logit 模型,但它的根本目的并不是要准确的估计接受处理的概率,而是使处理组和对照组之间的协变量达到均衡.

对于评估共同支撑域,就是需要在计算倾向得分后,评估处理组和对照组的倾向得分分布.因为如果两组样本没有重合的倾向得分或重合样本量太小,就会导致无法匹配或匹配偏差过大,因此评估共同支撑域是必要的.

倾向得分匹配法中的几种匹配方法,包括近邻匹配法、卡尺匹配法、核匹配法等都各有其利弊.对于模型评价标准,本文利用绝对标准化均值差异$(ASMD)$来量化处理组和对照组之间的协变量平衡情况,一般情况下,$ASMD < 0.2$即可认为协变量达到平衡.设有p个协变量,$X_i(i = 1, \dots p)$为协变量观察数据,$\bar{X}_{1i}, \bar{X}_{2i}$分别表示第$i$个协变量的处理组和对照组均值,$SD_i$为第$i$个变量的标准差,则第$i$个协变量的绝对标准化均值差异为:

$$ASMD_i = \frac{|\bar{X}_{1i} - \bar{X}_{2i}|}{SD_i} \tag{1.3}$$

最后,计算平均处理效应,一般公式为:

$$ATE = \frac{1}{N}\sum_{i \in S_p}\{Y_i - \omega_i Y_i\}$$

其中N为样本数,S_p是共同支撑域,Y_i是样本i的观测值,ω_i为匹配的权重.

1.1.4 逆概率加权

这里介绍另外一种匹配方法,称为逆概率加权法.逆概率加权法和倾向得分匹配法虽然是两种相互独立的方法,但又有紧密的联系.逆概率加权应用于因果研究,是在边际结构模型[20]框架下进行的.

边际结构模型:

$$logit(P[Y_A = 1]) = \beta_0 + \beta_1 A$$

可用来估计处理变量与结果变量间的因果关联强度,并可获得两个反事实变量估计值.基于上述倾向得分定义,用来估计ATE的逆概率权重为:

$$w = \frac{A}{ps} + \frac{(1-A)}{1-ps}$$

此时平均处理效应可表示为

$$\rho = E\left[\frac{AY}{ps} - \frac{(1-A)Y}{1-ps}\right]$$

为解决容易出现极端权重问题,引入稳定权重

$$sw = \frac{A \times P(A=1)}{ps} - \frac{(1-A) \times P(A=0)}{1-ps}$$

在非混淆假设下,广义倾向得分模型(GPS)可以表示在给定协变量条件下处理指标a的条件密度:$s(a,X) = f_{A|X}(a|X)$.由非混淆假设可知$Y(A) \perp A|s(a,X)$,就是说$s(a,X)$中包含了所有协变量信息,因此只需要调整广义倾向得分便可以控制观测到的协变量.

在广义倾向得分模型以及边际结构框架下,对于个体i,定义稳定权重:

$$sw_i = \frac{f_A(A_i)}{f_{A|X}(A_i|X_i)} - \frac{s(A_i)}{s(A_i,X_i)} \tag{1.4}$$

(1.4)式中$s(A_i)$是处理指标A的边际概率函数,$s(A_i,X_i)$是广义倾向得分.

1.2 Cox 比例风险模型

Cox 比例风险模型是生存分析中重要的模型之一.生存分析是一种流行的数据分析方法,用于特定种类的流行病学和其他数据分析.通常情况下,生存分析是一组数据分析的统计程序,其结果变量是事件发生之前的时间.这里的时间,是指从开始随访到某一事件发生的时间,可以以年、月、日等作为单位;而事件指的是死亡、疾病发生、缓解后的复发、康复或其他任何可能发生在个人身上感兴趣的指定经历,本文只考虑有一个感兴趣的经历.在生存分析中,通常将时间变量称为生存时间,因为它给出了个体在后续一段时间内“存活”的时间,也通常将事件称为失效,因为感兴趣的事件一般是死亡、疾病发生或其他一些负面的个人经历.当然失效有时也会是一个积极的事件,例如生存时间是“术后的恢复时间”.而在大多数的生存分析中,都必须要考虑一个关键问题:删失.如果对一个给定的个体,研究结束时个体没有达到我们感兴趣的事件,或者个体在中途失去随访,那么该患者的生存时间被认为删失.这样便不能知道这一个体完整的生存时间.

1.2.1 Cox 比例风险模型

令T表示临床试验中参与者的生存时间.$T=0$表示个体在临床试验中开始随访的时间.主要的问题通常是于对T的分布进行估计和检验.这种分布可以使用生存函数$S(t)=P(T>t), t>0$来描述.因为存在删失,这样更方便处理风险函数.

如果生存时间T和密度函数f是绝对连续的,那么风险函数定义为:

$$h(t)=\frac{f(t)}{S(t)} \tag{1.5}$$

即给定个体存活到t时刻的瞬时死亡(或失效)率.T的累积风险函数$H(t)=\int_0^t h(v)dv$,生存函数表示为$S(t)=exp\{-H(t)\}$.

设C为删失时间,即不能观察到超过此时间的个体.观察数据为$(\tilde{T},\delta)$,其中观测时间$\tilde{T}=min(T,C)$,删失指标$\delta=I(T\le C)$.虽然在数据未删失的情况下,通过经验分布函数可以一致地估计出分布函数$S(t)$.但 Tsiatis[21]提出,如果只观察$(\tilde{T},\delta)$,则$S(t)$和$h(t)$在不对C进行假设的情况下都不是一致估计的.对于只观察到$(\tilde{T},\delta)$,而不是所有个体的T,这时对所有的t,只能一致估计:

$$S^*(t)=exp\left\{-\int_0^t h^*(v)dv\right\}$$

使得$P(\tilde{T}>t)>0$,其中$h^*(t)=lim_{\triangle t\to 0}P(t\le T<t+\triangle t|T\ge t, C\ge t)/\triangle t$,这里的$h^*(t)$一般称为 crude hazard,而$h(t)$是 net hazard.在大多数生存分析应用中,都假设二者是相等的,本文同样假设$h^*(t)=h(t)$.这个假设成立的充分条件是T和C的独立性.

在生存分析中经常要评估处理效果,这时就需要调整其他可能与结果相关的协变量.本文中用A表示指定的处理(处理组$A=1$,对照组$A=0$),而X表示协变量,即研究开始时观察到的协变量.下面进一步介绍 Cox 比例风险模型,因为存在删失,通过风险函数对生存数据进行建模往往更便利.

Cox 比例风险模型[22]:

$$h(t)=h_0(t)\,exp(\beta^T x) \tag{1.6}$$

其中$h_0(t)$是一个未指定的基底风险函数,$\beta=(\beta_1,\beta_2,\dots\beta_p)^T$是回归参数向量.这里将(1.6)式描述为

$$h(t|A,X)=h_0(t)\,exp(\beta_1 A+\beta_2 X) \tag{1.7}$$

其中$h(t|A,X)$是在A和X条件下,时刻t的风险,$h_0(t)$为$A=0$和$X=0$时的风险函数.在该模型下,协变量对风险具有乘数效应,回归参数可解释为风险比的对数.如果指定生存时间的分布有有限数量的未知参数,生存模型(1.7)称为参数模型,否则称为半参数模型.

删失在比例风险模型中有着重要作用,因此需要其满足条件:

$$\lim_{\triangle t\to 0} P(t \le T < t+\triangle t | T \ge t, C \ge t)/\triangle t = \lim_{\triangle t\to 0} P(t \le T < t+\triangle t | T \ge t)/\triangle t$$

同样,对于生存分布的无偏估计,上述比例风险模型在协变量X以及处理指标A存在的情况下,需要相同的条件:

$$\lim_{\triangle t\to 0} P(t \le T < t+\triangle t | T \ge t, C \ge t, A, X)/\triangle t = \lim_{\triangle t\to 0} P(t \le T < t+\triangle t | T \ge t, A, X)/\triangle t$$

1.2.2 统计推断

下面对参数进行估计,设第i个个体的协变量为X_i, X_i为p维数据,事件与第i个个体的删失时间是独立的.t_i表示第i个个体的观察时间,δ_i为删失指标,当不存在删失时$\delta_i = 1$,存在右删失时$\delta_i = 0$,此时观察数据为

$$(t_i, \delta_i, X_i)(i = 1, \dots n)$$

不失一般性,假设没有相等的生存时间,并对生存时间进行排序:$t_{(1)} < \cdots < t_{(\tau)}$.令$D_i = \{i, \delta_i = 1 \text{ 且} t_i = t_{(i)}\}, d_j$表示$D_j$中所含元素个数,即在$t_{(j)}$时刻未删失的个体数.$H_j = \{i, t_i \ge t_{(j)}\}, (j = 1, \dots \tau), H_j$为$t_{(j)}$时刻的风险集,也就是在$t_{(j)}$时刻之前未删失且未失效的个体集.

为了估计参数β,Cox[23]提出的偏似然函数:

$$L(\beta) = \prod_{j=1}^{\tau} \frac{exp(\sum_{i\in D_j} \beta^T X_i)}{\left[\sum_{i\in H_j} exp(\beta^T X_i)\right]^{d_j}}$$

计算$\hat{\beta}$使偏似然函数达到最大,这时$\hat{\beta}$是β的最大偏似然估计,也称为 Cox 估计.接下来对$\hat{\beta}$进行求解,首先对$L(\beta)$取对数:

$$l(\beta) = ln\big(L(\beta)\big) = \sum_{j=1}^{\tau} \beta^T S_j - \sum_{j=1}^{\tau} d_j ln\left(\sum_{i\in H_j} exp(\beta^T X_i)\right) \tag{1.8}$$

其中$S_j = \sum_{i\in D_j} X_i, S_j = (S_{j1}, \dots S_{jp})^T, X_i = (x_{i1}, \dots x_{ip})^T$,然后对$\beta$求偏导:

$$\frac{\partial l}{\partial \beta_r} = \sum_{j=1}^{\tau} \left[S_{jr} - d_j \frac{\sum_{i\in H_j} x_{ir} exp(\beta^T X_i)}{\sum_{i\in H_j} exp(\beta^T X_i)} \right]$$

$$\frac{\partial^2 l}{\partial \beta_r \partial \beta_s} = \sum_{j=1}^{\tau} d_j \left[\frac{\sum_{i\in H_j} x_{ir} x_{is} exp(\beta^T X_i)}{\sum_{i\in H_j} exp(\beta^T X_i)} - \frac{\left(\sum_{i\in H_j} x_{ir} exp(\beta^T X_i)\right)\left(\sum_{i\in H_j} x_{is} exp(\beta^T X_i)\right)}{\left(\sum_{i\in H_j} exp(\beta^T X_i)\right)^2} \right]$$

其中$r, s = 1,2 \dots p$,令$\frac{\partial l}{\partial \beta_r} = 0$,得到估计$\hat{\beta}$.当样本$n$很大时,Cox 估计$\hat{\beta}$近似服从$N(\beta, I(\beta)^{-1})$,这里$I(\beta) = \left(-E\frac{\partial^2 l}{\partial \beta_i \partial \beta_j}\right)$是$p$阶矩阵.

下面利用上述结论对β进行检验,检验假设:

$$H_0: \beta = 0$$

引入协变量$x, S_1(t), S_2(t)$是两个生存函数,当$x = 0$时,生存时间T的生存函数是$S_1(t)$,当$x = 1$时,T的生存函数为$S_2(t)$.则考虑 Cox 模型:

$$S(t|x) = S_1(t)e^{\beta x}$$

则上述假设H_0等价于假设:

$$\widetilde{H}_0: S_1(t) = S_2(t)$$

即可以检验两个生存分布是否是相同的.

为了对假设H_0进行检验,取第一个总体的n_1个个体的观测值、第二个总体的n_2个个体的观测值,其中可能存在右删失.将两组个体合并作为上面 Cox 模型的一个样本.互异的生存时间$t_{(1)} < \cdots < t_{(\tau)}$,第$k$个总体在时刻$t_{(i)}$失效的个体有$d_{ki}$个,其中$k = 1,2$,则在时刻$t_{(i)}$失效的个体共有$d_i$个,$d_i = d_{1i} + d_{2i}$.$H_i$为$t_{(i)}$时刻的风险集,$n_i$为$H_i$中所含个体的数量,同理$n_i = n_{1i} + n_{2i}$.在这些假设下,(1.8)式变为:

$$l(\beta) = \beta h_2 - \sum_{i=1}^{\tau} d_i ln\left(n_{1i} + n_{2i}e^{\beta}\right)$$

其中$h_2 = \sum_{i=1}^{\tau} d_{2i}$,表示第二个总体中失效的个体总数.故

$$U(\beta) \stackrel{\text{def}}{=} \frac{\partial l}{\partial \beta} = h_2 - \sum_{i=1}^{\tau} \frac{d_i n_{2i} e^{\beta}}{n_{1i} + n_{2i}e^{\beta}}$$

$$I(\beta) \approx -\frac{\partial^2 l}{\partial \beta^2} = \sum_{i=1}^{\tau} \frac{d_i n_{1i} n_{2i} e^{\beta}}{(n_{1i} + n_{2i}e^{\beta})^2}$$

$U(\beta)$近似服从$N(\beta, I(\beta))$,因此在H_0下,$Z = \frac{U(0)}{\sqrt{I(0)}} \sim N(0,1)$,当$|Z|$值太大就拒绝假设$H_0: \beta = 0$.

$$U(0) = h_2 - \sum_{i=1}^{\tau} \frac{d_i n_{2i}}{n_{1i} + n_{2i}}, I(0) = \sum_{i=1}^{\tau} \frac{d_i n_{1i} n_{2i}}{{n_i}^2}$$

当有很多$d_i > 1$时,可使用统计量$\tilde{Z} = \frac{U(0)}{\sqrt{\tilde{I}(0)}} \sim N(0,1)$,其中$\tilde{I}(0) = \sum_{i=1}^{\tau} \frac{d_i(n_i - d_i)n_{1i}n_{2i}}{{n_i}^2(n_i - 1)}$

1.3 Bootstrap 方法

Bootstrap[24]方法是一种模拟抽样统计推断方法,就是利用重抽样方法得到新的样本.设$X = (x_1, x_2, \ldots x_n)$是来自某一未知总体$F$的样本,$\theta$是总体的参数,样本的经验分布函数为$F_n$,$\hat{\theta}$是$\theta$的估计,此时参数$\theta$的估计误差为

$$e(X, F) = \hat{\theta}(F_n) - \theta(F) = T_n$$

通过对观测样本进行重抽样,其样本含量同样为 n,此时得到一组新的观测值$X' = (x'_1, x'_2, \dots x'_n)$,利用新观测值构造统计量

$$e'(X, F) = \hat{\theta}(F'_n) - \hat{\theta}(F_n) = T'_n$$

其中F'_n是重抽样后样本的经验分布函数.重复上述抽样步骤,利用得到的统计量T'_n的分布模拟T_n的分布,得到未知参数$\theta(F)$的可能值,并对其进行统计推断.

2 基于 Cox 比例风险模型的因果推断方法

2.1 符号及模型

考虑将 n 个独立的个体分配到处理组和对照组,设第i个个体分配到处理组表示为$A_i = 1$、分配到对照组为$A_i = 0, X_i$表示协变量,设T_i和C_i分别表示第i个个体的生存时间和删失时间,则相应观测时间为$\tilde{T}_i = min(T_i, C_i), \delta_i$表示是否删失:未删失$(\delta_i = 1)$、删失$(\delta_i = 0)$.研究的目的是在观察到处理以及协变量的条件下判断处理变量的平均处理效应,对比处理组观察到的风险$h(t|A_i = 1, X_i = x)$和对照组观察到的风险$h(t|A_i = 0, X_i = x)$,有下面的比例风险模型:

$$h(t|A_i = 1, X_i = x) = h(t|A_i = 0, X_i = x)exp(\beta x) \tag{2.1}$$

β是未知参数,模型(2.1)中的一个隐藏假设就是 1.1.2 节中的稳定性假设,对于协变量X_i,在模型中纳入协变量有助于解决混杂(对于不平衡的随机组)和衰减性(如果不包括协变量,则处理效果偏倚).

利用观察到的处理,通过比例风险模型修改处理组和对照组中特定的生存分布,直至两随机组之间的生存分布达到相同.从生存分布角度改写模型(2.1):

$$S(t|A_i = 1, X_i = x) = S(t|A_i = 0, X_i = x)^{exp(\beta x)}$$

对照组的生存函数:

$$S(t|A_i = 0, X_i = x) = S(t|A_i = 1, X_i = x)^{exp(-\beta x)}$$

比例风险模型(2.1)本身以半参数方式将处理组中观察到的风险与对照组中的风险联系起来,如果假设处理和协变量之间相互作用,模型(2.1)可以扩展为:

$$h(t|A_i = 1, X_i = x) = h(t|A_i = 0, X_i = x)exp(\beta_1 a + \beta_2 x)$$

2.2 权重估计

设$L(t)$是时间相关的协变量,满足既是失效的风险影响因素,也影响随后的处理,同时过去的处理对风险因素存在影响,这时$L(t)$称为混杂因素.$\bar{A}(t) = \{A(u); 0 \leq u < t\}$是个体直到时刻$t$的处理,给定过去的处理$\bar{A}(t)$和基底协变量$X$,此时条件风险率为:

$$h_T(t|\bar{A}(t), X) = h_0(t)exp(\beta_1 A(t) + \beta_2 X)$$

首先假设无删失,在满足条件的时间相关协变量$L(t)$存在的情况下,通过最大化 Cox 偏似然函数得到的估计$\hat{\beta}_1$是参数β_1的渐近无偏估计.然而,即使模型中包括了时间相关的协变量$L(t)$作为回归,这仍然是对处理效应的一个有偏估计.

正如论文[21]中所讨论的,可以通过拟合上述随时间变化的 Cox 模型来消除或减少这种偏差,方法是让个体i参与在时刻t进行的风险集计算,该计算利用权重进行加权:

$$w_i(t)=\prod_{k=0}^{int(t)}\frac{1}{p(A(k)=a_i(k)|\bar{A}(k-1)=\bar{a}_i(k-1),\bar{L}(k)=l_i(k))}$$

通过加权以获得逆概率加权的偏似然估计.上式中$A(-1)$定义为 0,$int(t)$是小于等于t的最大整数.但通常情况下使用一种“稳定的”权重进行加权:

$$sw_i(t)=\prod_{k=0}^{int(t)}\frac{p(A(k)=a_i(k)|\bar{A}(k-1)=\bar{a}_i(k-1),X_i=x)}{p(A(k)=a_i(k)|\bar{A}(k-1)=\bar{a}_i(k-1),\bar{L}(k)=l_i(k))}$$

$sw_i(t)$的分子表示在给定过去的处理和协变量条件下,个体在时刻k时接受处理的概率.稳定的权重产生的置信区间通常更窄,且实际覆盖率更高.

定义处理$\bar{a}=\{a(t);0\le t<\infty\}$,对于每一个$\bar{a}$,其边际结构 Cox 比例风险模型:

$$h_{T_{\bar{a}}}(t|X)=h_0(t)exp(\beta_1 a(t)+\beta_2 X)$$

$h_{T_{\bar{a}}}(t|X)$表示在原始总体中,个体在协变量条件下,时刻t的失效风险,与事实相反即反事实结果:直到时刻t,所有的个体是否接受处理都与过去相同.β_1,β_2为未知参数,将这个模型称为边际结构模型 MSM,因为在X的水平内,它是反事实变量$T_{\bar{a}}$边际分布的一个结构模型.参数β_1是因果对数比率,故$exp(\beta_1)$具有因果解释,表示所有个体在时刻t持续暴露的风险率与所有个体在时刻t一直未暴露的风险率之比.在时间相关的协变量$L(t)$满足非混淆假设条件下,逆概率加权估计$\hat{\beta}_1$是β_1的一致估计.

由于每一个个体的权重$sw_i(t)$随时间变化,因此很难得到标准 Cox 模型软件来计算逆概率加权估计值$\hat{\beta}_1$,为了解决这一问题,利用权重$sw_i(t)$来拟合一个加权混合 logistic 回归:

$$\text{logit}\,P(D(t)=1|D(t-1)=0,\bar{A}(t-1),X)=\beta_0(t)+\beta_1 A(t-1)+\beta_2 X$$

其中如果个体在时刻t存活$D(t)=0,D(t)=1$表示个体在时刻t失效.

下面考虑存在右删失的情况,定义删失指标$\delta(t)$,当个体在时刻t右删失$\delta(t)=0$,未删失为 1.为了估计存在删失时的β_1,拟合一个加权的 Cox 模型,对在时刻t有风险的个体,利用权重$sw_i(t)\times sw_i'(t)$进行加权,其中

$$sw_i'(t)=\prod_{k=0}^{t}\frac{p\big(\delta(k)=1|\bar{\delta}(k-1)=1,\bar{A}(k-1)=\bar{a}_i(k-1),X_i=x\big)}{p\big(\delta(k)=1|\bar{\delta}(k-1)=1,\bar{A}(k-1)=\bar{a}_i(k-1),\bar{L}(k-1)=l_i(k-1)\big)}$$

$\delta(-1)$以及$A(-1)$定义为 0,$sw_i'(t)$是一个个体直到时刻t一直未删失的比率,除以一直未删失个体的条件概率.除了处理变量A之外没有其他的时间依赖决定因素,$sw_i(t)\times sw_i'(t)$的分母是已有个体直到时刻t的处理以及删失的概率.因为β_1是未知的,要利用观察数据进行估计.假设观察到的协变量足以调整由于失去随访造成的混杂和选择偏差,则通过$sw_i(t)\times sw_i'(t)$加权产生了因果参数β_1的一致估计.

首先考虑$sw_i(t)$的估计,要估计每个个体和时刻的$sw_i(t)$的分母和分子,假设任何个体一旦开始接受处理则一直接受处理,可以把开始接受处理的时间看作一个失效时间变量,通过混合 logistic 模型模拟开始接受处理的概率,拟合模型

$$\text{logit } P\left(A(k)=0\middle|\bar{A}(k-1)=0,\bar{L}(k)\right)=\alpha_0(k)+\alpha_1 L(k)+\alpha_2 X$$

得到未知参数的估计值$\hat{\alpha}=(\hat{\alpha}_0(k),\hat{\alpha}_1,\hat{\alpha}_2)$.此时只需要对时刻$k$是存活并且未删失的个体进行拟合.

个体i在时刻$k-1$没有开始接受处理,且在k时刻也没有接受处理的概率

$$\hat{p}_i(k)=expit(\hat{\alpha}_0(k)+\hat{\alpha}_1 L_i(k)+\hat{\alpha}_2 X_i) \tag{2.2}$$

其中$expit(x)=\frac{e^x}{1+e^x}$.当个体$i$直到时刻$k$也没有开始处理,个体$i$在时刻$k$时,$sw_i(t)$的分母等于$\bar{\hat{p}}_i(k)=\prod_{u=0}^{k}\hat{p}_i(u)$,当个体$i$在时刻$t(t\le k)$开始接受处理时,$sw_i(t)$的分母为$\bar{\hat{p}}_i(k)=(1-\hat{p}_i(k))\prod_{u=0}^{t-1}\hat{p}_i(u)$.计算分子时,利用上述 logistic 模型,只需除去时间相关协变量$L(k)$,则$sw_i(t)$的分子为$\bar{\hat{p}}_i(k)=(1-\hat{p}_i(k))\prod_{u=0}^{t-1}\hat{p}_i(u)$,其中$\hat{p}_i(k)=expit(\hat{\alpha}_0(k)+\hat{\alpha}_2 X_i)$.存在删失的情况下同理可得.

2.3 平均处理效应

通过加权匹配均衡协变量,下面计算平均处理效应.根据风险函数定义(1.5),对于处理a,定义风险:

$$h_a(t)=\lim_{\triangle t\to 0}P(t\le T^a<t+\triangle t|T^a>t)/\triangle t$$

其中$a=0,1$,表示是否接受处理,T^a是如果个体接受处理a的潜在结果;

累积风险率定义为:

$$CIF_a(t)=E\left[I_{(T^a\le t)}\right]=P(T^a\le t)$$

对于两个事件的风险比定义为:

$$HR(t)=\frac{h_1(t)}{h_0(t)}$$

这里的$HR(t)$能够依赖于时间,因为没有对风险函数$h_a(t)$假设任何模型.风险差异函数定义为:

$$RD(t)=E\left[I_{(T^1\le t)}\right]-E\left[I_{(T^0\le t)}\right]=CIF_1(t)-CIF_0(t) \tag{2.3}$$

如果想量化处理的效果,即从处理中得到获得(或损失)的天数或年数等,可以使用以下受限平均时间:

$$RMT_a(\tau)=\int_0^{\tau}CIF_a(t)dt$$

τ是某个指定感兴趣的时间点,$RMT_a(\tau)$可以解释为事件后受限平均时间.使用RMT来量化处理效果已经在单变量生存分析中变得越来越常用.如果事件是理想的,通常希望看到更长的平均事件后时间.基于RMT的平均处理效应可以定义为差异:

$$RMT(t) = RMT_1(\tau) - RMT_0(\tau) \tag{2.4}$$

2.4 基于倾向得分的 Cox 比例风险模型的回归分析

计算得到平均处理效应后,估计含有倾向得分的 Cox 比例风险模型系数:

$$h(t|A,X) = h_0(t)\,exp\{\beta A + \alpha ps(X,\gamma)\} \tag{2.5}$$

其中α,β,γ为未知的回归系数,$h_0(t)$为未指定的基底风险函数,$ps(X,\gamma)$为倾向得分$P(A=1|X)$.模型(2.5)表示X只通过倾向得分来对失效时间产生影响.根据倾向得分,处理变量A是条件独立于X的,因此这个假设是合理的.这意味着处理变量A和倾向得分一起影响失效时间.令$\theta = (\beta,\alpha,\gamma)'$,$Y_i(t) = I(Y_i \geq t)$,则$\theta$的偏似然函数为:

$$L_{ps}(\theta) = \prod_{i=1}^{n}\left[\frac{exp\{\beta A_i + \alpha ps(X_i,\gamma)\}}{\sum_{j=1}^{n} Y_j(T_i)exp\{\beta A_j + \alpha ps(X_j,\gamma)\}}\right]^{\delta_i}$$

偏似然函数的对数表示为:

$$l_{ps}(\theta) = \sum_{i=1}^{n}\delta_i\left[\beta A_i + \alpha ps(X_i,\gamma) - ln\sum_{j=1}^{n} Y_j(T_i)exp\{\beta A_j + \alpha ps(X_j,\gamma)\}\right] \tag{2.6}$$

因此,使$L_{ps}(\theta)$,$l_{ps}(\theta)$达到最大的值$\hat{\theta}_{ps}$就是最大偏似然估计量,但由于ps是未知的,因此不能直接最大化(2.6)式,得到估计量$\hat{\theta}_{ps}$,首先对倾向得分进行估计,假设倾向得分满足 logistic 回归模型:

$$ln\left(\frac{ps(X,\gamma)}{1-ps(X,\gamma)}\right) = -X'\gamma$$

对于观察数据,由于有完整的数据(A_i,X_i),可以使用卡方检验通过 logistic 回归模型来检验协变量是否与处理相关,根据(2.5)和(2.6)两式,得到模型:

$$h(t|A,X) = h_0(t)ex\,p\left\{\beta A + \frac{\alpha exp(-X'\gamma)}{1+exp(-X'\gamma)}\right\} = h_0(t)ex\,p\left\{\beta A + \frac{\alpha}{1+exp(X'\gamma)}\right\}$$

偏似然函数:

$$L_n(\theta) = \prod_{i=1}^{n}\left[\frac{exp\left\{\beta A_i + \frac{\alpha}{1+exp(X'\gamma)}\right\}}{\sum_{j=1}^{n} Y_j(T_i)exp\left\{\beta A_j + \frac{\alpha}{1+exp(X'\gamma)}\right\}}\right]^{\delta_i}$$

其对数:

$$l_n(\theta)=\sum_{i=1}^{n}\delta_i\left[\beta A_i+\frac{\alpha}{1+exp(X'\gamma)}-ln\sum_{j=1}^{n}Y_j(T_i)exp\left\{\beta A_j+\frac{\alpha}{1+exp(X'\gamma)}\right\}\right] \tag{2.7}$$

则参数$\theta=(\beta,\alpha,\gamma)'$的估计$\hat{\theta}_n=(\beta_n,\alpha_n,\gamma_n)'$可以利用最大化(2.7)式得到,且 Lu B.[25]认为这一估计是渐近正态的.

3 数值模拟

3.1 模拟实验设定

这一部分将基于 Cox 比例风险模型,利用加权匹配的数据,研究处理变量与生存时间之间的因果关系,使用 R 软件进行数值模拟.首先生成一个数据集,对数据集中的数据进行加权匹配, 对比调整和未调整的数据，计算调整前后数据集的平均处理效应,并进行比较.利用 Cox 比例风险回归模型探讨其中的处理变量对结果的影响.根据经验,常见的生存时间分布包括指数分布、Weibull 分布、Gamma 分布、对数正态分布等,本文主要研究了生存时间分布是 Weibull 分布和对数正态分布的情况.

设样本含量$n = 500, x_1, x_2$表示两个协变量,其中x_1服从均匀分布$U(0,1), x_2$服从两点分布$B(1,0.8)$.接着利用不同分布模拟生存时间T和删失时间C,其参数由协变量x_1, x_2确定.处理组中的生存时间服从 Weibull 分布时,其概率密度函数为:

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

设置形状参数$k_1 = 4$,比例参数$\lambda_1 = exp(-1 + 2x_1)$.对照组中的生存时间服从形状参数为$k_2 = 1$,比例参数$\lambda_2 = exp(-1.5 - 4x_1)$的 Weibull 分布.处理组和对照组的删失时间服从均匀分布$U(a, b)$,通过对均匀分布参数a, b进行调整,可以控制删失比率.这里设置删失时间C分别服从$U(x_2, 3)$、$U(x_2, 1.5)$、$U(x_2, 1)$,控制删失比率分别为 30%、50%、70%.对于生存时间为对数正态分布的数据集,对数正态分布的概率密度函数为

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} exp\left(-\frac{1}{2}\left(\frac{lnx - \mu}{\sigma}\right)^2\right), & x > 0 \\ 0, & x \leq 0 \end{cases}$$

设置参数:处理组中$\mu_1 = 2x_1, \sigma_1 = 1$,对照组中$\mu_2 = 1 - 3x_1, \sigma_2 = 3$.同样调节删失时间分布的参数,以保证不同的删失比率,则删失时间C分别服从$U(x_2, 14)$、$U(x_2, 7)$、$U(x_2, 1)$.根据模拟生成的生存时间以及删失时间的大小,可以确定是否删失,这里用δ来表示,删失为 0,未删失为 1,取对应的生存时间和删失时间中的较小值作为观测到的生存时间$\tilde{T}$.

下面利用一个两点分布$B(1, p)$生成处理变量A,其概率由 logistic 模型(2.2)得到,参数设为$\alpha = (0.5, 1, -1)$

$$p = expit(0.5 + x_1 - x_2)$$

p表示假设每个个体在其随访过程中一直未接受处理情况下反事实概率,通过此概率判断变量的值($A_i = 1$表示个体i接受处理,$A_i = 0$表示未接受处理).此时便产生了一个包含

协变量、观察时间、处理指标以及删失指标的一组数据集$\{\tilde{T}, A, \delta, x_1, x_2\}$,数据集中含有$n = 500$个个体的观察数据.

3.2 模拟实验结果

对数据集进行调整前,首先计算倾向得分,下图是处理组和对照组数据的倾向得分分布情况:

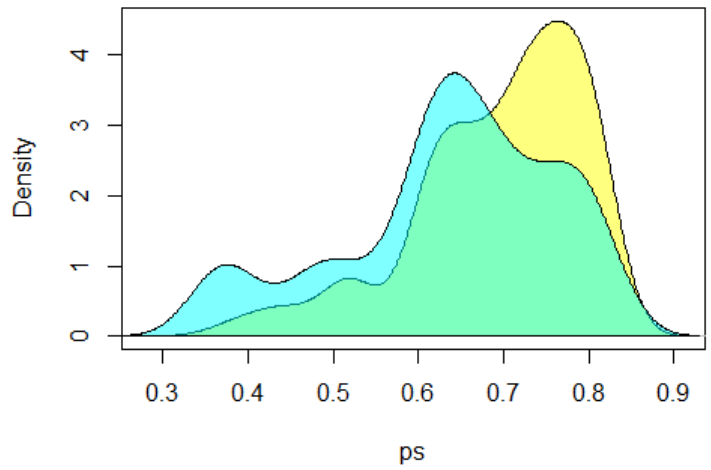

图 3.1　倾向得分分布

Fig. 3.1　Distribution of propensity scores

其中黄色区域表示个体接受处理的倾向得分分布,蓝色区域表示未接受处理的倾向得分.由图 3.1 可以看出,在这一模拟数据集中,个体接受处理的倾向较高,且重合区域较多,满足共同支撑域条件,便于后续研究.

接下来在 Cox 比例风险模型下,利用 2.2 节的估计权重对数据进行加权处理.表 3.1 中列出了数据集匹配前后协变量x_1、x_2在处理组和对照组的均值以及T统计量. 根据表中数据,数据集匹配之前,处理组和对照组中的协变量均值存在较大差异. 而对数据进行匹配调整之后,两组之间协变量的均值差异显著减小. 利用T检验对两组之间的差异进行检验,结果显示,协变量x_1、x_2在匹配之后检验的p值大于 0.05,认为两组间均值不存在显著差异,此时协变量在处理组和对照组中得到了较好的均衡. 图 3.2、图 3.3 展示了数据集调整前后协变量x_1、x_2在处理组和对照组中的分布情况,其中蓝色区域表示处理组,红色区域表示对照组.图中左侧图像为原始数据集,也就是未进行调整时的协变量在处理组和

对照组中的分布情况,根据图 3.2、图 3.3 中左右两图的对比,能够形象的看出数据集在经过调整后,协变量在处理组和对照组之间的分布更加均衡.

表 3.1　匹配前后两组间变量均值及差异检验

Tab. 3.1　The mean and difference tests of covariable between the two groups before and after adjustment

协变量	未匹配				匹配			
	均值		T 检验		均值		T 检验	
	处理组	对照组	T	P	处理组	对照组	T	P
x_1	0.52794	0.44206	3.1884	0.00152	0.50219	0.51007	-0.2488	0.8036
x_2	0.15385	0.31481	-4.2308	0.00003	0.20456	0.20677	-0.0528	0.9579

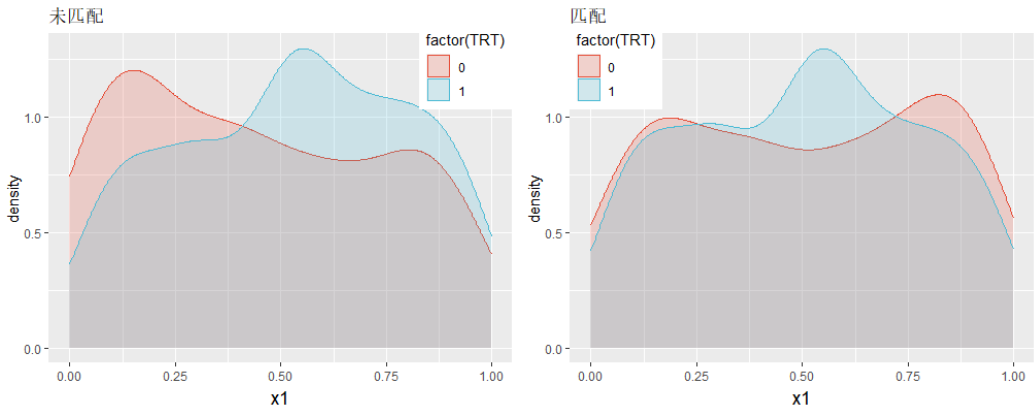

图 3.2　匹配前后协变量x_1在处理组和对照组中的分布情况

Fig. 3.2　Distribution of covariable x_1 in the treatment group and the control group before and after adjustment

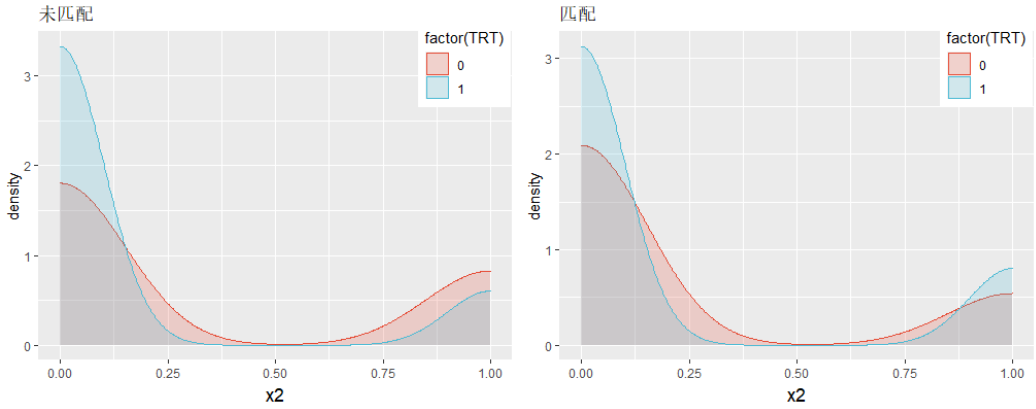

图 3.3　匹配前后协变量x_2在处理组和对照组中的分布情况

Fig. 3.3　Distribution of covariable x_2 in the treatment group and the control group before and after adjustment

3.2.1 基于风险差异的平均处理效应

在调整数据集使协变量达到均衡后,首先讨论风险差异$RD(t)$的值,考虑时刻$t=0.5$时的值,风险差异是处理组和对照组之间事件发生概率的绝对差值,属于绝对效应量.由于使用单一数据集得到的结果具有偶然性,因此生成$m=1000$组数据集,并分别对每一个数据集进行调整,利用公式(2.3)计算基于风险差异的平均处理效应及其标准差,则调整前后不同删失比率的数据计算结果对比如表 3.2 所示.

表 3.2 中列出了不同删失比率下,样本含量分别为 200,500,1000 的模拟结果,当$n=200$时,计算得出的标准差较大,因此增加了样本含量.根据表中数据可以看到,$n=500$时,已经可以很好地反应其结果,当$n>1000$时,其均值以及标准差基本保持稳定. 以 Weibull 分布为例进行具体说明,根据表中第三行数据结果可以看到,在$n=1000$时,未进行调整的数据集中的处理组和对照组的风险差异$RD_{前}(0.5)=-0.0869$. 这表示处理组的累积风险率相比对照组降低了 0.0869,而在进行加权匹配调整之后,两组之间风险差异变为$RD_{后}(0.5)=-0.1307$. 调整后数据的结果相比于调整前,风险差异的估计有所变化,数据在进行加权匹配后,处理变量所导致的累积风险率的降幅由 0.0869 变化为 0.1307,处理组和对照组之间的差异更为明显.也就是在对数据集中的协变量进行均衡后,处理变量实际对结果的积极影响更为显著.

表 3.2　匹配前后风险差异

Tab. 3.2　Risk difference before and after adjustment

删失比率	生存时间分布	样本含量	未匹配		匹配	
			均值(mean)	标准差(sd)	均值(mean)	标准差(sd)
30%	Weibull 分布	200	-0.0945	0.0627	-0.1369	0.0632
		500	-0.0889	0.0405	-0.1331	0.0407
		1000	-0.0869	0.0282	-0.1307	0.0277
	对数正态分布	200	-0.0868	0.0461	-0.0973	0.0501
		500	-0.0801	0.0293	-0.0905	0.0316
		1000	-0.0810	0.0207	-0.0911	0.0227
50%	Weibull 分布	200	-0.1891	0.0711	-0.2341	0.0704
		500	-0.1875	0.0461	-0.2345	0.0461
		1000	-0.1868	0.0325	-0.2338	0.0318
	对数正态分布	200	-0.1240	0.0496	-0.1351	0.0534
		500	-0.1283	0.0338	-0.1398	0.0361
		1000	-0.1192	0.0226	-0.1303	0.0248
70%	Weibull 分布	200	-0.2793	0.0796	-0.3249	0.0787
		500	-0.2813	0.0518	-0.3298	0.0510
		1000	-0.2808	0.0357	-0.3294	0.0348
	对数正态分布	200	-0.2660	0.0629	-0.2793	0.0669
		500	-0.2620	0.0408	-0.2765	0.0433
		1000	-0.2654	0.0297	-0.2795	0.0319

3.2.2 基于受限平均时间的平均处理效应

表 3.3 中列出了根据公式(2.4)计算得到的基于受限平均时间的平均处理效应,对比其在不同删失比率的数据集匹配前后的变化情况:

表 3.3　匹配前后受限平均时间差异

Tab. 3.3　Restricted mean time before and after adjustment

删失比率	生存时间分布	样本含量	未匹配		匹配	
			均值(mean)	标准差(sd)	均值(mean)	标准差(sd)
30%	Weibull 分布	200	-0.0237	0.0164	-0.0354	0.0175
		500	-0.0225	0.0108	-0.0347	0.0114
		1000	-0.0220	0.0075	-0.0342	0.0078
	对数正态分布	200	-0.0277	0.0154	-0.0313	0.0169
		500	-0.0265	0.0100	-0.0301	0.0109
		1000	-0.0270	0.0072	-0.0306	0.0080
50%	Weibull 分布	200	-0.0474	0.0197	-0.0606	0.0207
		500	-0.0476	0.0130	-0.0616	0.0137
		1000	-0.0476	0.0092	-0.0617	0.0095
	对数正态分布	200	-0.0394	0.0169	-0.0433	0.0184
		500	-0.0424	0.0117	-0.0465	0.0127
		1000	-0.0398	0.0080	-0.0437	0.0088
70%	Weibull 分布	200	-0.0701	0.0231	-0.0842	0.0242
		500	-0.0719	0.0154	-0.0871	0.0160
		1000	-0.0722	0.0107	-0.0875	0.0110
	对数正态分布	200	-0.0834	0.0226	-0.0883	0.0244
		500	-0.0863	0.0150	-0.0916	0.0161
		1000	-0.0884	0.0110	-0.0937	0.0120

根据表 3.3 中第三行数据,同样说明在数据集进行调整后,处理组和对照组之间平均受限时间差异增大,处理变量对结果影响更加显著.图 3.4 是数据匹配前后风险差异以及受限平均时间的差异随时间的变化情况,根据两图对比可以看到,随着时间的变化,调整前后的平均处理效应差异越来越大.

3.2.3 回归分析及风险比

利用 2.4 节中含倾向得分的 Cox 比例风险回归模型(2.5)估计数据处理前后变量的回归系数,用倾向得分ps统一表示匹配调整后数据集的协变量. 处理变量回归系数β的估计为$\hat{\beta}$,而$exp(\hat{\beta})$具有因果解释,为处理组与对照组的风险比. 通过对协变量以及处理变量A进行计算,不同删失比率下的结果如表 3.4 所示:

表 3.4　数据匹配前后的回归结果

Tab. 3.4　Regression results before and after adjustment

删失比率	生存时间分布		变量	回归系数	标准误	风险比	p值
30%	Weibull 分布	未匹配	x_1	-0.9232	0.2353	—	$<10^{-4}$ ***
			x_2	-0.0350	0.1242	—	0.778
			A	-0.2029	0.1483	0.8164	0.171
		匹配	ps	-0.6530	0.4680	—	0.1629
			A	-0.4895	0.1280	0.6129	0.0001 ***
	对数正态分布	未匹配	x_1	-1.0505	0.2089	—	$<10^{-4}$ ***
			x_2	0.0676	0.1412	—	0.632
			A	-0.1240	0.1285	0.8834	0.334
		匹配	ps	-1.1850	0.4458	—	0.0079 **
			A	-0.2503	0.1261	0.7786	0.0472 *
50%	Weibull 分布	未匹配	x_1	-1.0519	0.2755	—	0.0001***
			x_2	0.1980	0.1425	—	0.1647
			A	-0.1351	0.1602	0.8737	0.3991
		匹配	ps	-2.2086	0.7439	—	0.0030 **
			A	-0.3241	0.1458	0.7232	0.0263 *
	对数正态分布	未匹配	x_1	-1.3058	0.2291	—	$<10^{-4}$ ***
			x_2	0.0688	0.1463	—	0.638

删失比率	生存时间分布		变量	回归系数	标准误	风险比	p值
70%	Weibull 分布	匹配	A	-0.1938	0.1321	0.8238	0.142
			ps	-2.1482	0.5138	—	$<10^{-4}$ ***
		未匹配	A	-0.2734	0.1311	0.7608	0.037 *
			x_1	-1.3621	0.3022	—	$<10^{-4}$ ***
			x_2	-0.2104	0.1773	—	0.2354
	对数正态分布	匹配	A	-0.3040	0.1668	0.7378	0.0684
			ps	-0.8173	1.1853	—	0.4905
		未匹配	A	-0.5225	0.1621	0.5930	0.0013 **
			x_1	-1.1524	0.3025	—	0.0001 ***
			x_2	-0.4258	0.2115	—	0.0441 *
		匹配	A	-0.9670	0.1716	0.3802	$<10^{-4}$ ***
			ps	-0.4820	0.6131	—	0.432
		未匹配	A	-1.0374	0.1728	0.3544	$<10^{-4}$ ***

根据表 3.4 中第三行及第五行数据对比显示,在数据未进行加权匹配之前,p值为0.171,处理变量对结果不存在显著影响.而数据经过匹配调整后处理变量在 0.001 水平下具有显著统计学意义,表明处理对结果影响显著. 此时处理变量A的回归系数为$\hat{\beta}=-0.4895$,风险比$HR=exp(\hat{\beta})=0.6129$,其 95%置信区间为(0.477,0.7876).风险比HR表示接受处理个体的风险是未接受处理个体风险的 0.6129 倍,处理变量对于生存结果具有显著影响.处理组相比于对照组,风险减少了 0.3871,这一数据集中的处理变量对生存结果实际上起到了显著的积极影响.

表 3.2、表 3.3 和表 3.4 中列出了不同删失比率的数据结果,通过调整删失时间的分布参数控制删失比率.对比表中不同删失比率匹配前后的均值变化,表现的差异可能不够清晰.图 3.4、图 3.5 和图 3.6 分别是删失比率为 30%、50%、70%的风险差异和受限平均时间的差异变化,对比不同删失比率对于加权匹配结果的影响.根据对比,二者差异都随着时间的变化而增大,但删失率较低时,差异变化相对更加明显.

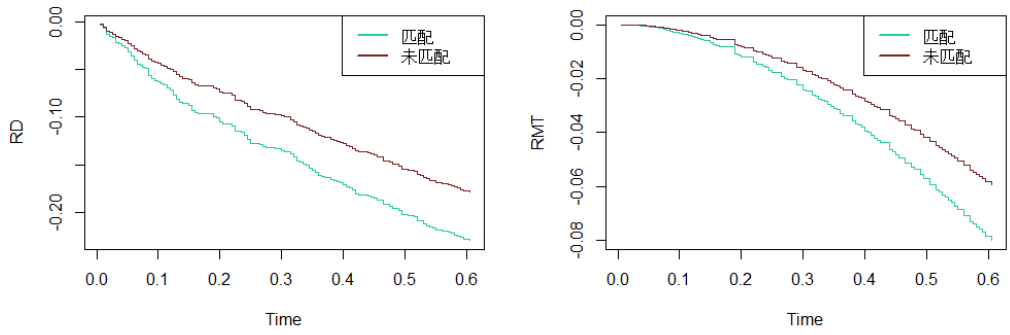

图 3.4　删失比率为 30%时风险差异和受限平均时间的变化情况

Fig. 3.4　The variation in risk difference and restricted mean time at a 30% Censoring Proportion

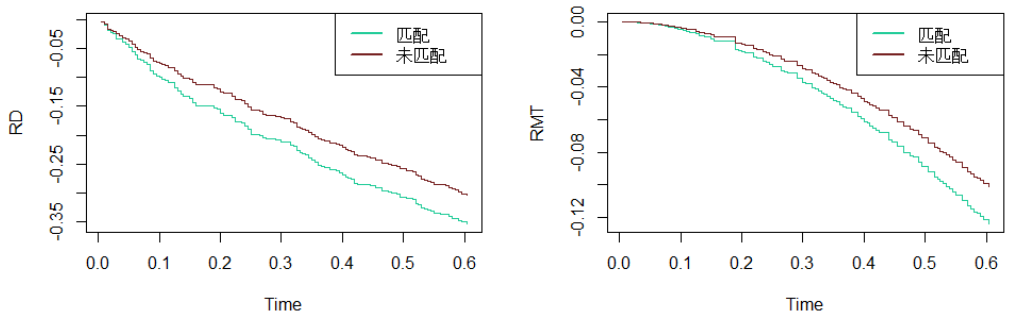

图 3.5　删失比率为 50%时风险差异和受限平均时间的变化情况

Fig. 3.5　The variation in risk difference and restricted mean time at a 50% Censoring Proportion

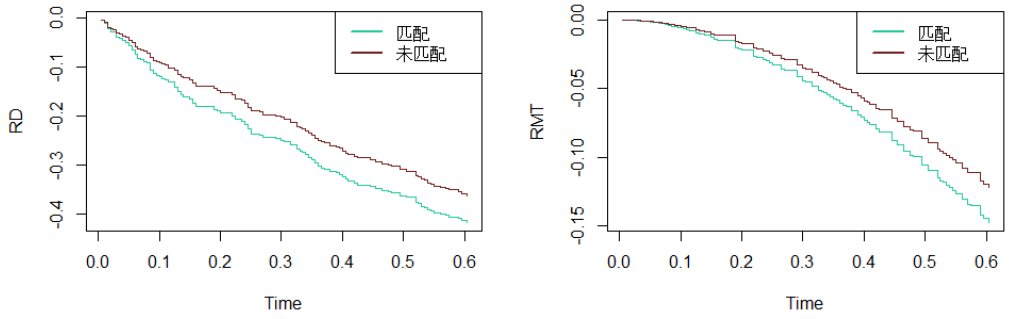

图 3.6　删失比率为 70%时风险差异和受限平均时间的变化情况

Fig. 3.6　The variation in risk difference and restricted mean time at a 70% Censoring Proportion

4 实例分析

4.1 rotterdam 数据集

考察 R 软件 survival 程序包中的 rotterdam 数据集[26],rotterdam 原始数据集包括 2982 名原发性乳腺癌患者,他们的记录列入了鹿特丹肿瘤库.随访时间 1~231 个月(中位数 107 个月),无复发生存时间定义为从初次手术到疾病早期复发或因任何原因死亡的时间.首先对数据集进行简单处理,例如数据集中肿瘤大小不能作为连续变量,因此分为三类:$< 20\,mm$、$20\sim 50\,mm$ 和 $> 50\,mm$,分别用数字 1、2、3 来表示.考察的所有变量如表 4.1 所示,数据为患者初次手术时的观测值.研究的处理变量为激素处理(hormone),目的是探究其对无复发生存时间的影响.

表 4.1　变量说明

Tab. 4.1　Variable specification

变量	变量定义	变量类型	变量取值	均值	最小值	最大值	中位数
hormone	激素处理	二分类	0(未接受) 1(接受)	0.1137	0	1	—
rtime	生存时间	连续	> 0	2097.9	36.0	7043	1940
death	状态	二分类	0(生存) 1(死亡)	0.4266	0	1	—
age	年龄	连续	> 0	55.1	24	90	54
meno	绝经状态	二分类	0(绝经前) 1(绝经后)	0.56	0	1	—
size	肿瘤大小	有序多分类	1($< 20\,mm$) 2($20\sim 50\,mm$) 3($> 50\,mm$)	1.64	1	3	—
qrade	肿瘤分级	二分类	2,3	2.73	2	3	—
pgr(fmol/l)	孕激素受体	连续	> 0	161.8	0	5004	41
nodes	阳性淋巴结数	连续	> 0	2.71	0	34	1

4.2 匹配结果

在对数据集进行了基本处理之后,首先比较接受激素处理(hormone)和未接受处理的两组倾向得分分布情况,如图 4.1.图中显示患者接受激素处理的倾向较低,存在重合部分,满足共同支撑域条件.

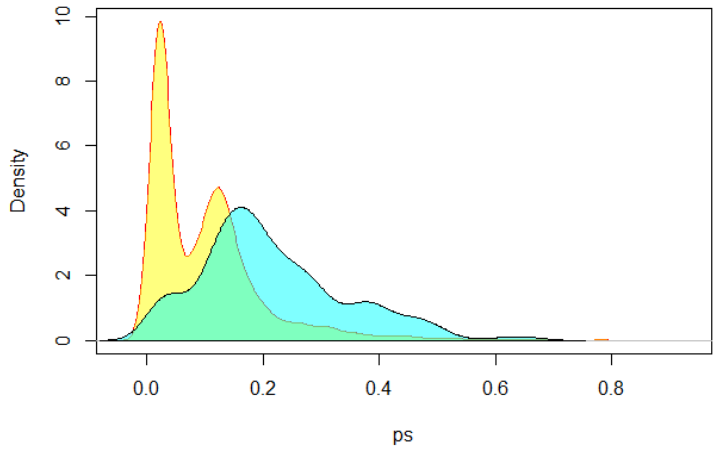

图 4.1　rotterdam 数据集倾向得分分布

Fig. 4.1　Distribution of propensity scores in the rotterdam dataset

为了评估匹配后获得的协变量平衡,考察绝对标准化均值差异($ASMD$),比较调整前后的数据,根据公式(1.3)计算得到两组数据绝对标准化均值差异对比如图 4.2:

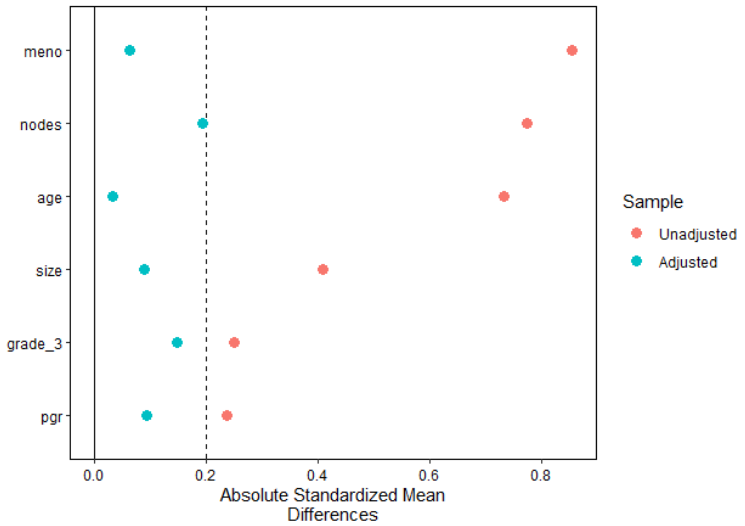

图 4.2　数据调整前后的绝对标准化均值差异

Fig. 4.2　Absolute standardized mean difference before and after data adjustment

由图 4.2,遵循既定惯例“小于 0.2 的标准化差异可以认为是平衡的”,图中绝对标准化均值差异都在 0.2 以下,因此可以认为稳定权重达到了良好的平衡.

为了对协变量达到的平衡情况做进一步的说明,表 4.2 列出了各个协变量在匹配前后处理组和对照组的均值情况,根据表中数据,除协变量肿瘤分级外,其他协变量在处理组和对照组中的均值差异都有不同程度的减小.对两组均值差异做 T 检验,结果也显示匹配调整后的数据集中,认为两组间均值不存在差异.这说明加权匹配对两组中的协变量起到了较好的平衡作用.图 4.3、图 4.4 分别是数据调整前后协变量年龄(age)和绝经状态(meno)在处理组和对照组中的分布情况.其中蓝色部分表示处理组,红色部分为对照组,根据图 4.3 和图 4.4 左右两图对比可以看出调整后两变量在两组中的分布达到了更好地均衡.

表 4.2　rotterdam 数据集匹配前后两组间变量均值及差异检验

Tab. 4.2　The mean value and difference test of variables between the two groups before and after adjustment in the rotterdam dataset

	未匹配				匹配			
	均值		T 检验		均值		T 检验	
协变量	处理组	对照组	T	P	处理组	对照组	T	P
age	62.549	54.098	11.558	$< 10^{-4}$	52.973	55.571	-1.551	0.121
meno	0.879	0.520	12.912	$< 10^{-4}$	0.512	0.567	-1.696	0.090
size	1.879	1.606	7.242	$< 10^{-4}$	1.624	1.655	-0.489	0.625
qrade	2.826	2.722	4.091	$< 10^{-4}$	2.549	2.755	-2.801	0.005
nodes	5.720	2.327	13.839	$< 10^{-4}$	3.640	2.987	1.267	0.205
pgr	108.233	168.706	-3.606	0.00032	176.892	163.019	0.656	0.512

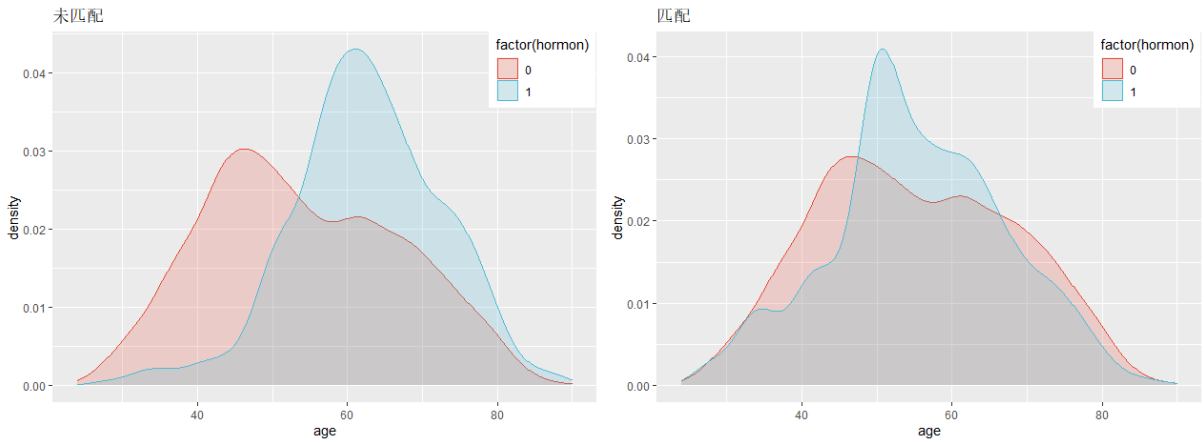

图 4.3　年龄变量在处理组与对照组之间的分布情况对比

Fig. 4.3　Comparison of distribution of age variables between the treatment group and the control group

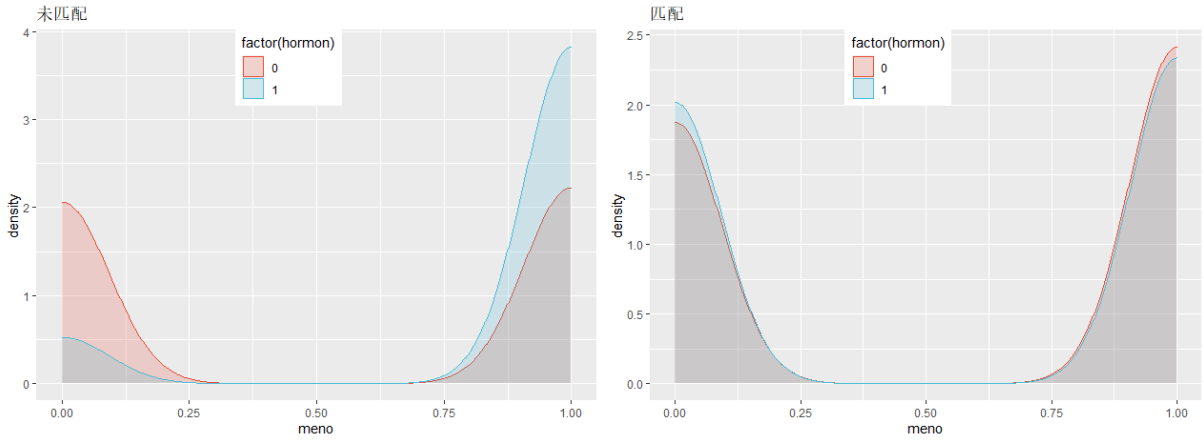

图 4.4　绝经状态变量在处理组与对照组之间的分布情况对比

Fig. 4.4　Comparison of distribution of meno variables between the treatment group and the control group

协变量在数据集中达到良好的均衡后，下面运用 R 软件计算平均处理效应,选择 Bootstrap 方法对其标准差进行估计,这样便不需要对原始数据的分布做任何假设.Austin [27]对比了三种方差估计方法,发现使用 Bootstrap 估计可以对标准误差和置信区间进行近似正确的估计.由于重复次数较少时,得到的结果波动性较大,因此计算重复抽样 1000 次时的平均处理效应的标准差.表 4.3 中是不同时间匹配前后计算得到的风险差异和受限平均时间结果及其标准误,根据表中第三行数据,第五年时,未调整的原始数据风险差异 $RD_{前}=0.0775$,相比对照组激素处理能够降低 0.0775 的风险.而调整后$RD_{后}=-0.0828$,接受激素处理比未接受激素处理的风险有所增加.这表明激素处理(hormone)对于生存结果并没有起到积极影响.

对基于RMT的平均处理效应,根据计算所得结果,$RMT_{前}$=88.8996、$RMT_{后}$=−93.2041.结果表明数据未进行调整时激素处理(hormone)可延缓复发约 89 天,而经过对协变量的调整匹配后,实际结果是激素处理(hormone)可能加速复发约 93 天.图 4.5 是数据匹配前后风险差异和受限平均时间差异随时间的变化情况:从随访开始,数据调整前后的差异随着时间的不断推移而不断增加.

表 4.3　rotterdam 数据集匹配前后的风险差异和受限平均时间

Tab. 4.3　The risk difference and restricted mean time before and after adjustment in the rotterdam dataset

时间	未匹配				匹配			
	RD	se_{RD}	RMT	se_{RMT}	RD	se_{RD}	RMT	se_{RMT}
1 年	0.0217	0.0079	2.8994	1.0783	-0.0220	0.0095	-2.8692	1.2410
3 年	0.0635	0.0222	36.6960	12.9886	-0.0664	0.0287	-37.7480	16.2190
5 年	0.0775	0.0267	88.8997	31.0607	-0.0828	0.0362	-93.2041	40.2500
7 年	0.0831	0.0283	147.5091	50.9054	-0.0898	0.0395	-156.2666	67.5728

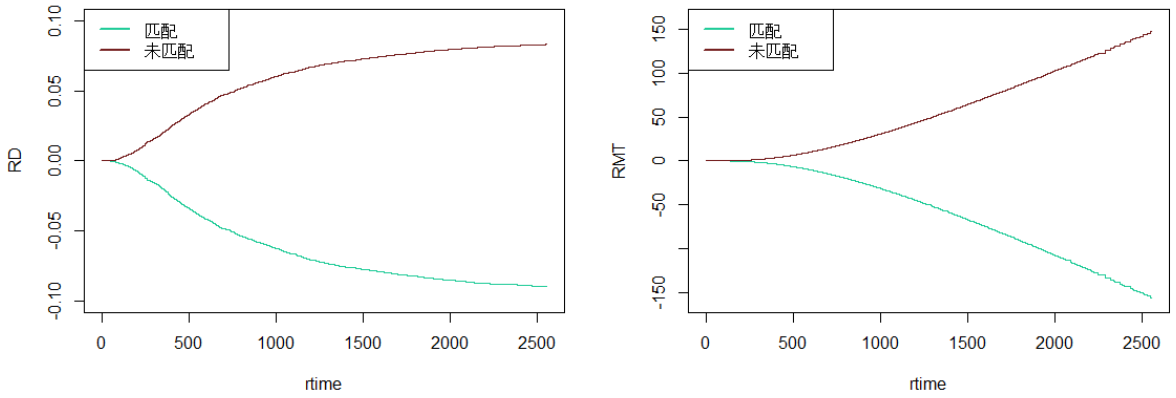

图 4.5　匹配前后风险差异和受限平均时间的变化情况

Fig. 4.5　Changes of risk difference and restricted mean time before and after adjustment

下面根据含倾向得分的 Cox 比例风险回归模型估计回归系数,结果如表 4.4 所示.表中第七行数据激素处理(hormone)在 0.01 水平下具有统计学价值,风险比为 0.7870 (95%置信区间 $(0.662, 0.935)$),表明激素处理对于生存结果具有积极影响.而对数据进行加权匹配后,表中最后一行数据中,激素处理(hormone)在 0.05 水平下具有统计学意义,风险比为

1.2458 (95%置信区间 $(1.053, 1.474)$).也就是在对协变量进行均衡后,结果显示激素处理对结果起到了消极的抑制作用,处理组的风险是对照组的 1.2458 倍,表明未接受激素处理的一组风险更小.

表 4.4　rotterdam 数据集匹配前后的回归结果

Tab. 4.4　Regression results before and after adjustment in the rotterdam dataset

	变量	回归系数	标准误	风险比	p值
未匹配	age	0.0073	0.0037	—	0.0472 *
	meno	0.1379	0.0972	—	0.1563
	size	0.4357	0.0442	—	$< 10^{-4}$ ***
	qrade	0.2941	0.0709	—	$< 10^{-4}$ ***
	nodes	0.0769	0.0047	—	$< 10^{-4}$ ***
	pgr	-0.0003	0.0001	—	0.0069 **
	hormone	-0.2395	0.0881	0.7870	0.0065 **
匹配	ps	2.5402	0.8652	—	0.0033 **
	hormone	0.2198	0.0859	1.2458	0.0105 *

结　论

因果推断的研究在各个领域都存在重要意义,而对观察性数据因果推断的研究更是研究的重点.因果推断在统计学中的不断发展,使得其统计方法在流行病学、生物医药、社会学和经济学等许多领域起到了不容忽视的重要作用.因果推断的方法有很多,本文研究的潜在结果模型是其中较为成熟,应用较广泛的方法.潜在结果模型的重要问题是对平均处理效应进行估计,通常采用的方法是线性回归或倾向得分.本文利用生存分析中的半参数模型 Cox 比例风险回归模型进行研究,估计权重,对观察到的数据进行加权,使其中的协变量在处理组和对照组之间达到均衡,接着利用均衡后的数据进行统计推断,计算平均处理效应,对比分析所得结果,并利用数值模拟以及实例分析进行研究.

在对数据集进行加权的计算中,本文只采用了一种稳定的权重进行加权,实际加权的方法有很多,之后可以考虑采用其他的加权方式进行分析研究.在估计平均处理效应的过程中,本文只考虑了混杂因素对处理变量产生影响的情况,在实际处理中还可能出现其他相关问题,比如在某研究的观察中,个体可能对接受的处理存在不依从现象,这时就需要考虑依从性对个体接受暴露以及试验结果的影响.类似此类问题在本文中没有讨论,因此在平均处理效应的估计中仍存在许多问题有待深入探讨.实例分析中的 rotterdam 数据集,本文只针对提供的数据变量进行了分析,由于对医学了解存在局限,可能忽略了其他相关问题,激素处理对结果实际起到的影响还需要进一步研究.

参 考 文 献

[1] Wright S. The theory of path coefficients a reply to Niles's criticism[J]. Genetics, 1923, 8(3): 239.

[2] Rubin D. B. Estimating causal effects of treatments in randomized and nonrandomized studies[J]. Journal of Educational Psychology, 1974, 66(5): 688-701.

[3] Schield M. Confounding and Cornfield: Back to the Future[C]//Looking Back, Looking Forward. Proceedings of the Tenth International Conference on Teaching Statistics. 2018.

[4] Pearl J. Linear Models: A Useful "Microscope" for Causal Analysis[J]. Journal of Causal Inference, 2013, 1(1): 155-170.

[5] Cummiskey K., Adams B., Pleuss J., et al. Causal inference in introductory statistics courses[J]. Journal of Statistics Education, 2020, 28(1): 2-8.

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

[7] Rosenbaum P. R. A characterization of optimal designs for observational studies[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1991, 53(3): 597-610.

[8] Abadie A., Imbens G. W. Large sample properties of matching estimators for average treatment effects[J]. econometrica, 2006, 74(1): 235-267.

[9] Hirano K., Imbens G. W., Ridder G. Efficient estimation of average treatment effects using the estimated propensity score[J]. Econometrica, 2003, 71(4): 1161-1189.

[10] Blinder A. S. Wage discrimination: reduced form and structural estimates[J]. Journal of Human resources, 1973, 8(4): 436-455.

[11] Cox D. R. Regression Models and Life-Tables[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1972, 34(2): 187-220.

[12] Robins J. M., Tsiatis A. A. Correcting for non-compliance in randomized trials using rank preserving structural failure time models[J]. Communications in Statistics - Theory and Methods, 1991, 20(8): 2609-2631.

[13] White I. R., Babiker A. G., Walker S., et al. Randomization-based methods for correcting for treatment changes: examples from the Concorde trial[J]. Statistics in Medicine, 1999, 18(19): 2617-2634.

[14] Robins J. M., Finkelstein D. M. Correcting for Noncompliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-Rank Tests[J]. Biometrics, 2000, 56(3): 779-788.

[15] Hernán M. A., Brumback B., Robins J. M. Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments[J]. Journal of the American Statistical Association, 2001, 96(454): 440-448.

[16] Imbens G. W., Rubin D. B. Causal Inference for Statistics, Social, and Biomedical Sciences[M]. Cambridge University Press, 2015

[17] Holland P. W., Rubin D. B. Causal Inference in Retrospective Studies[J]. Evaluation Review, 1988, 12(3): 203-231.

[18] Rosenbaum P. R., Rubin D. B. Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1983, 45(2): 212-218.

[19] Harder V. S., Stuart E. A., Anthony J. C. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research[J]. Psychological methods, 2010, 15(3): 234.

[20] Robins J. M., Hernan M. A., Brumback B. Marginal Structural Models and Causal Inference in Epidemiology[J]. Epidemiology, 2000, 11(5): 550-560.

[21] Tsiatis A. A nonidentifiability aspect of the problem of competing risks[J]. Proceedings of the National Academy of Sciences, 1975, 72(1): 20-22.

[22] Nikulin M., Wu H. D. I. The Cox Model and Its Applications[M]. Springer, Berlin, Heidelberg, 2016

[23] Cox D. R. Partial likelihood[J]. Biometrika, 1975, 62(2): 269-276.

[24] Efron B., Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy[J]. Statistical science, 1986, 1(1): 54-75.

[25] Lu B., Cai D., Wang L., et al. Inference for proportional hazard model with propensity score[J]. Communications in Statistics - Theory and Methods, 2018, 47(12): 2908-2918.

[26] Royston P., Altman D. G. External validation of a Cox prognostic model: principles and methods[J]. 2013, 13(1): 33.

[27] Austin P. C. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis[J]. Statistics in medicine, 2016, 35(30): 5642-5655.

攻读硕士学位期间发表学术论文情况

1. 谢新惠,蔡梦瑶,侯文.基于 R 软件的数理统计可视化教学研究——以最大似然估计为例[J]. 应用数学进展,2021,10(01):268-273.

2. 侯文,蔡梦瑶,谢新惠.零膨胀 Poisson 回归模型极大似然估计的相合性与渐近正态性[J].辽宁师范大学学报(自然科学版),2020,43(02):162-168.

致　　谢

三年的研究生生活马上就要画上句点,三年里,认识了新的朋友,也重新认识了自己.在校园生活即将结束的时刻,内心充满无限感慨,感谢一路上遇到的人们.

首先我要感谢我的导师侯文老师对我的悉心教导,对我各种问题的耐心解答,包容我的不足.感谢老师在论文选题、内容研究、程序修改以及论文写作方面给予的细致严格的指导,老师态度严谨、讲解生动,总能利用形象的事例解答我存在的疑惑.即使在假期,炎炎夏日,老师也会因为我遇到问题而赶来学校.感谢老师三年来的教诲,愿老师身体健康,万事顺遂.

我还要感谢我的父母和外婆给予我无私的爱与支持,在我成长的路上不管发生什么,你们始终都愿意相信我,给我无限的关心与鼓励,愿你们平安、健康、幸福.感谢我的好友新月一直以来的陪伴,给我各方面的帮助与建议;感谢我可爱的同学们,还有我的同门蔡梦瑶,感谢你们的温暖与能量;感谢美丽温馨的辽师大;感谢疫情期间辛苦的工作人员,感谢身边的一切.最后祝愿大家都能成为自己心中最满意的自己.