前　　言

具有缺失等复杂结构的数据很普遍, 这类数据的分析方法很快地被应用到生物、医学及相关的研究领域. 绝大多数微观数据都不可避免地存在数据缺失问题. 有部分研究者可能会选择直接删掉有数据缺失的样本, 利用剩余完全观测样本完成后续研究. 这样的处理方法也称为完全事例 (complete cases,CC) 方法. 虽然 CC 方法操作简单, 易于理解, 但在绝大多数实际数据缺失问题中, 由于直接忽略缺失样本, CC 方法建模往往会导致估计结果出现偏差, 影响后续实证分析. 可见, 如何更好地规避缺失数据带来的估计问题是亟待解决的难题, 在理论研究与实践应用中具有重要的研究意义. 数据缺失的外在表现形式可以分为单变量缺失、多变量缺失、单调缺失、非单调缺失等. 数据缺失的内在生成机制 [也称为缺失机制 (missing mechanism)] 同样包含很多类型, 如完全随机缺失 (missing complete at random, MCAR)、随机缺失 (missing at random,MAR)、非随机缺失 (missing not at random, MNAR)、基于随机效应的缺失以及在纵向数据中出现的 ‘‘掉队’’ 与 ‘‘间歇式缺失’’ 等 (Rubin,1976). 针对不同缺失机制下的缺失数据, 需要使用不同的缺失数据处理方法, 如 CC 方法仅适用于完全随机缺失机制. 本书主要关注随机缺失机制与非随机缺失机制. 关于缺失数据 (无回答) 的处理涉及很多方面, 如从抽样设计角度出发, 可以通过设计适合的抽样框降低无回答发生的可能性. 本书将在第 2 章与第 3 章从事后建模推断角度出发, 基于分位回归模型在不同的缺失机制下提出新

的估计方法, 规避缺失数据带来的不良影响, 提高模型参数的估计效果.

本书全面、系统、严格地阐明缺失分位回归建模理论与方法, 并尽力反映复杂缺失数据分析国际前沿研究. 本书所选取的缺失分位回归的主要内容包括缺失分层线性分位回归模型、缺失广义线性分位回归模型、缺失非线性分位回归模型、缺失分层半参数分位回归模型等内容.

一方面, 具有复杂缺失结构特征的数据普遍存在于现实世界中. 离开数据的这一缺失异质性, 常常会使传统的统计分析方法效果不佳, 甚至失效. 缺失结构数据分析是近年来迅速发展起来的一个分支, 在各学科领域越来越受到重视, 相关的研究显得异常活跃. 另一方面, 分位回归是一种稳健的统计建模方法, 它旨在对条件分位函数进行统计推断. 正如基于误差平方和最小化的经典线性回归方法能估计条件均值函数一样, 分位回归方法提供了一种估计条件分位函数的机制. 目前非常缺少全面、系统地介绍缺失分位回归建模理论与方法方面的书, 更谈不上找到尽力反映复杂删失缺失数据分析国际前沿研究的专著了. 本书正朝着这个目标努力, 用严格的数学语言对缺失数据分位模拟的现代面貌进行较为详细的介绍.

本书针对复杂缺失分位回归的统计建模问题, 从理论、模拟和实证方面进行了分析论证, 主要创新点、学术价值以及应用价值如下.

(1) 在分位回归的响应变量存在不可忽略缺失的情形下, 引入半参数指数倾斜模型刻画应答概率, 在此基础上提出三种卷积平滑分位数回归估计方程: 逆概率加权 (inverse probability weighting, IPW)、估计方程插补 (estimation equation imputation,EEI) 和增强逆概率加权 (augmented IPW, AIPW), 并在经验似然框架下得到倾斜参数和分位回归系数的估计量. 在理论上证明了三种分位回归估计量等价的渐近正态性和对应调整对数经验似然比函数的渐近 χ^2 性质. 数值模拟比较了上述估计量的有限样本表现, 验证了估计量的稳健性. 所提出方法应用于 HIV–CD4 数据分析, 考察不同治疗组中缺失机制的差异以及基线和前期的 CD4 和 CD8 细胞水平对当期 CD4 细胞水平的影响.

(2) 在随机缺失机制假定下, 从模型推断角度出发, 针对线性缺失分位回归模型, 提出逆概率多重加权 (inverse probability multiple weighted,

IPMW) 估计, 是在逆概率加权估计的基础上, 结合倾向得分匹配及模型平均思想, 经过多次估计加权确定最终参数估计结果, 适用于绝大多数缺失场景. 经过模拟研究发现, IPMW 估计量在继承 IPW 估计量的优势上具有更稳健的性质. 最后, 将该方法应用于含有缺失数据的微观调查数据中, 研究了经济较发达的准一线城市中等收入群体消费水平的影响因素, 对比两种估计方法的估计结果及置信带, 发现逆概率多重加权估计量的标准偏差更小, 估计结果更稳健.

(3) 考虑到有关随机缺失机制的研究已经很丰富, 以及随机缺失机制的假定过于强, 为了更加接近实际数据缺失的情形, 提出一种非随机缺失机制下的缺失数据处理方法. 首先总结梳理了分析非随机缺失问题的一般框架, 并在其中定义了更加一般化的倾向得分函数 (回答概率). 基于定义的倾向得分函数, 借助非参数工具变量函数, 提出了一种非参数逆概率加权缺失数据处理方法并证明估计量的渐近性质. 进一步, 考虑到分位回归下有关非随机缺失数据的问题研究较少, 继续将该方法拓展到分位回归模型中, 提出非随机缺失机制下的逆概率加权与增广逆概率加权分位回归估计量并证明回归系数估计量的渐近性质. 再进一步, 考虑到非随机缺失机制下缺失变量条件分布与总体分布不一致, 为了更加充分地利用已知信息, 尝试对前面提出的逆概率加权估计方法进行校准再加权. 在模拟研究中, 通过对比所提出的方法与所有已有方法, 发现所提出的方法在有限样本下估计效果更好. 在实证分析部分, 针对含有缺失数据的公共部门男性工资数据研究了是否有编制对工资水平的影响.

(4) 在响应变量非随机缺失假设下, 提出基于半参数指数倾斜应答模型的三种渐近无偏的分位回归估计方程和平滑经验似然估计方法, 证明了倾斜参数轮廓经验似然估计量和三种分位回归系数估计量等价的渐近正态性和相应对数似然比函数的渐近 χ^2 加权和性质.

(5) 基于半参数指数倾斜应答模型, 进一步考虑了工具变量未指定的情况, 利用充分降维方法构造工具变量识别模型参数, 避免多元核回归造成的维数灾难, 提出降维的 IPW、核辅助估计方程插补和 AIPW 渐近无偏分位回归估计方程, 并证明了与分位回归系数平滑经验似然估计量等价的渐近

正态性和对数经验似然比函数的渐近 χ^2 加权和性质.

(6) 基于半参数指数倾斜应答模型和观测部分的工作模型, 利用重要性重抽样方法提出插补分位回归估计方程和稳健的 AIPW 分位回归估计方程, 证明了模型正确指定情况下, 回归系数估计量的渐近服从正态分布, 对数经验似然比函数的渐近服从 χ^2 加权和分布, 以及在工作模型错误指定情况下, AIPW 估计方程的稳健性.

(7) 利用惩罚样条估计响应变量对缺失机制的非参数影响, 提出基于半参数 logistics(逻辑) 二项似然的 Pólya–Gamma(波利亚-伽马) 表示和非对称拉普拉斯分布正态混合表示的贝叶斯分位回归方法, 给出缺失响应变量的 Metropolis-adjusted Langevin 贝叶斯高效插补方法和未知参数后验分布的 Gibbs(吉布斯) 抽样算法.

(8) 在半参数非随机缺失假设下, 探讨 AIDS 研究队列中不同治疗组的缺失机制差异, 研究了 CD4 和 CD8 两种生物标志物之间的关系, 提出基于分位回归的 CD4 细胞数水平的预测模型; 考察了不同分位水平下的治疗臂效应和个体治疗效果的其他影响因素.

(9) 在纵向响应变量对缺失机制存在非参数影响的假设下, 比较了不同分位点上精神分裂症利培酮疗法的治疗效果随时间变化的规律, 探讨了感兴趣响应变量对缺失机制的影响形式.

本书内容全新、理论性强, 介绍了近年来国际上关于缺失数据分位回归建模理论与方法的许多新成果. 其中不乏作者本人在该领域研究中取得的一些成果. 本书取材全部来自国际一流学术期刊杂志上的代表性文章, 信息量大、内容权威.

自 2004 年中国人民大学统计学院在全国首开 ‘‘分位回归’’ 课程以来, 本书作者一直担任本课程的主讲老师. 本书的大部分材料在课堂上讨论过. 本书在写作过程中, 参加过翻译辅助工作的有 (不分先后): 杨澜、张若璇、苏鹏、刘一鉴、余颖、钱曼玲、张陶陶、陶丽、刘烨、王春雨、周雨菡、贾燕飞、郃凌楠、闫懋博、周照亚、芮荣祥、侯健、周旭、孟丽君、张丽平、刘林军; 参与过校对辅助工作的有 (不分先后): 张永霞、白永昕、梁晋雯、马少沛、王芝皓、郭婧璇、芮荣祥、虞祯、彭坤、杨鸣、姚雨薇、张馨月、

周照亚、周梦雨、曹苏周、梁永玉、侯健、周旭、张丽平、孟丽君、刘林军、刘硕、赵欣怡、刘淼柔、孟坦、徐宗文、胡晓雪等. 在此对他们表示衷心的感谢!

由于本人水平有限, 疏漏和不足在所难免, 甚望读者批评指正!

田茂再

2025 年 6 月 17 日

目　录

第一部分　缺失均值回归模型理论方法

第二部分　缺失分位回归模型理论方法

第一部分

缺失均值回归模型理论方法

第 1 章　缺失协变量回归

本节研究了在假定的均值函数中当某些个体协变量对应的值存在缺失时回归系数的估计问题. 通过利用不同的选择概率的估计值, 检验了 Horvitz 和 Thompson (1952) 类型的加权估计, 这些概率的估计值可以被视为冗余参数. 特别地, 研究了当选择概率是通过核光滑估计时回归系数估计结果的性质. 给出了相应的大样本性质, 并在模拟中对提出的估计量和极大似然估计以及多重表征在多种不同的模型假定以及缺失机制下进行了比较. 此外, 还展示了对本研究具有启发性的两个实际例子.

1.1　概　　述

在回归分析中, 缺失协变量数据的分析是一个实用的而且有趣的问题. 其中包含一元的响应变量 Y 以及协变量向量 $(\boldsymbol{X}, \boldsymbol{Z})$, 协变量 $\boldsymbol{X}$ 在某些记录的子集上可能存在缺失. 例如, 在夏威夷进行的关于饮食与甲状腺癌的病例对照研究中, 39 个个体 $(n = 468)$ 的平均每日总碘摄入量是缺失的. 本章后续会展示在华盛顿州西部进行的食管癌和贲门癌的病例对照研究, 其中大约 1/3 个体的酒精摄入量是缺失的. 本章考虑的回归模型是 $Y = g(\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\Theta}) + \varepsilon$, 其中 g 是确定的函数, 而 $E(\varepsilon|\boldsymbol{Z}, \boldsymbol{X}) = 0$. 该回归分析的目的是估计参数 $\boldsymbol{\Theta}$. 例如, 通常选取线性函数 $g(\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\Theta}) = \theta_0 + \boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{Z} + \boldsymbol{\theta}_2^{\mathrm{T}}\boldsymbol{X}$ 对连续协变量关于协变量向量的线性相关性进行建模, $\boldsymbol{\Theta} = (\theta_0, \boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}})^{\mathrm{T}}$,

或者选择 logistic(逻辑) 回归函数 $g(\boldsymbol{Z},\boldsymbol{X},\boldsymbol{\Theta}) = \{1+\exp(-\theta_0-\boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{Z}-\boldsymbol{\theta}_2^{\mathrm{T}}\boldsymbol{X})\}^{-1}$ 对二值响应变量对协变量向量的相关性进行建模.

在过去的十年里, 关于缺失协变量数据的回归分析模型的构建已经成为活跃的研究领域 (Little, 1992; Little and Rubin, 1987). 一个常见的方法是通过对观测到的数据 $\boldsymbol{Z}$ 进行线性回归来表示缺失的 $\boldsymbol{X}$(Dagenais, 1973) 或者利用完整观测通过利用对 $\boldsymbol{Z}$ 和 Y 的线性回归表示 (Buck, 1960). Rubin (1978, 1987) 提出了多重插补方法. 当该方法因为利用极大似然估计而需要迭代时, EM (expectation maximization, 最大期望) 算法成为一个非常有用的算法 (Dempster et al., 1977).

当 $\boldsymbol{X}$ 的测量存在误差且有一个 $\boldsymbol{X}$ 的代理变量可用时, 会出现一个与此密切相关的问题, 相应地提出了多个处理方法. 例如, 针对两阶段病例对照研究, Breslow 和 Cain(1988) 提出了拟条件似然方法. 当缺失不依赖 Y 和 $\boldsymbol{X}$ 时, Carroll 和 Wand(1991) 以及 Pepe 和 Fleming (1991) 对于连续和离散的替代变量提出了似然的非参数估计方法.

当 $\boldsymbol{X}$ 是随机缺失的 (missing at random, MAR) 时 (Rubin, 1976), Flanders 和 Greenland(1991) 以及 Zhao 和 Lipsitz(1992) 建议使用加权的估计量, 受 Horvitz 和 Thompson(1952) 的启发, 加权估计方法为

$$\boldsymbol{U}_n(\boldsymbol{\Theta},\pi) = n^{-1/2}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\mathcal{X}_i\{Y_i - g(\boldsymbol{Z}_i,\boldsymbol{X}_i,\boldsymbol{\Theta})\} = 0, \tag{1.1}$$

其中, $\mathcal{X}_i = (1,\boldsymbol{Z}_i^{\mathrm{T}},\boldsymbol{X}_i^{\mathrm{T}})^{\mathrm{T}}$ 是简便的选择, n 是独立观测的总数, δ_i 是一个二值指示函数且当协变量 $\boldsymbol{X}_i$ 观测到时为 1, 反之为 0. $\pi_i = P(\delta_i = 1|Y_i,\boldsymbol{Z}_i)$ 表示观测到 $\boldsymbol{X}_i$ 的概率 (严格大于 0). 显然, 当观测数据完整时, 式(1.1)的解与正态误差假定下的线性回归或 logistic 回归对应的极大似然估计是等价的. 在假定概率函数 π_i 是正确的前提下, 与 Zhao 和 Lipsitz(1992) 的研究一样, 式(1.1)的系数估计具有相合以及渐近正态性. 注意, 由于随机缺失, 选择概率 π_i 与未观测到的变量 $\boldsymbol{X}_i$ 是独立的. 在本章中 π_i 是一个需要估计冗余参数, 即使在某些应用中, π_i 可以事先给定. 为了估计 π, 需要对 δ_i 关于 $(Y_i,\boldsymbol{Z}_i)$ 的相关性通过 logistic 回归进行建模, 该模型同样包含需要利用数据估计的一些冗余参数. 在假定 π 事先正确给定的前提下, 该方法可

以得到模型中参数的相合估计. 然而, 错误给定可能会导致估计的回归系数有偏差. 为了克服这个困难, 考虑引入选择概率的非参数核平滑器.

Robins 等 (1994) 利用计算最优半参数得分函数提出了一个有效的估计方法. 在半参模型中, 选择概率包含在一个子模型中, 并假定其非零以及对应的参数空间为无穷维. 然而, 有效的得分通常没有显示表达, 而需要求解一个泛函积分方程. 本章考虑估计得分 $\boldsymbol{U}_n(\boldsymbol{\Theta}, \hat{\pi})$, 其中 $\boldsymbol{U}_n$ 见式(1.1), $\hat{\pi}$ 是 π 的核估计量. 虽然当 $(Y, \boldsymbol{Z})$ 不是分类时使用非参数平滑估计量 $\hat{\pi}$ 估计 π 的思想十分直观, 但包含最优带宽速度的渐近分布理论并不简单, 因而需要对 $\boldsymbol{\Theta}$ 构建相应的估计量. 特别地, 众所周知, 非参数光滑器的收敛速度低于 $\sqrt{n}$, 而在 $\boldsymbol{U}_n$ 的估计中使用了这个估计量. 基于合适的带宽比例, 得到的线性估计保证了参数 $\boldsymbol{\Theta}$ 的估计是 $\sqrt{n}$ 相合的. 进一步, 应用加权估计得分扩展了 Prentice 和 Pyke (1979) 关于回溯性抽样数据的预期分析结果.

通过加权调整处理缺失数据出现在调查抽样文献中, 其中无应答权重是乘以抽样权重的一个因子 (Little, 1986, 1991). 在这样的情形下, 使用估计的选择概率进行分析称为使用“估计的倾向分数”(Rosenbaum and Rubin, 1983). 而且, 在因果效应的观察性研究中, 匹配抽样和基于它的子分类是控制偏差的标准技术 (Rosenbaum and Rubin, 1984; Rubin and Thomas, 1992).

1.2节描述该方法. 1.3节研究 $\boldsymbol{\Theta}$ 的参数和半参数估计的渐近理论. 1.4节介绍加权估计方程在两阶段病例对照研究中的应用. 1.5节介绍模拟研究的一些结果, 包括适当带宽的选择、协方差估计方法以及与其他估计方法的比较. 1.6节将这些方法应用于甲状腺癌的病例对照研究以及食管癌和贲门癌的病例对照研究.

1.2　方　　法

CC 分析方法是基于数据 $\{Y_i, \boldsymbol{Z}_i, \boldsymbol{X}_i, \delta_i = 1, i = 1, 2, \cdots, n\}$ 且忽略所有缺失的变量的方法. 特别地, 令 $\hat{\boldsymbol{\Theta}}_{\mathrm{CC}}$ 表示基于完整数据的如下估计方程

的解

$$\sum_{i=1}^{n}\delta_i\mathcal{X}_i\{Y_i-g(\boldsymbol{Z}_i,\boldsymbol{X}_i,\boldsymbol{\Theta})\}=0. \tag{1.2}$$

CC 分析是简单的模型, 但其具有两个潜在的弊端: ①由于没有考虑数据的完整性导致有效性缺失; ②当验证数据并非原始数据的随机子样本时, 估计的结果可能是非相合的. 然而, 正如 Little (1992) 所言, 当缺失仅与协变量有关时, CC 分析的推断是可取的. 此外, 对于基于二值病例对照数据的 logistic 模型, CC 分析得到的斜率估计是有效的 (Prentice and Pyke, 1979).

1.2.1 选择概率的参数建模

在式(1.1)中, 缺失概率 $\pi_i(i=1,2,\cdots,n)$ 通常是未知的. 一种方法是如 Robins 等 (1994) 假定 π 的一个参数函数. 例如

$$\pi_i=\pi(Y_i,\boldsymbol{Z}_i)=P(\delta_i,Y_i,\boldsymbol{Z}_i)=\{1+\exp(-\alpha_0-\alpha_1Y_i-\alpha_2^{\mathrm{T}}\boldsymbol{Z}_i)\}^{-1}, \tag{1.3}$$

其中, α 是未知参数, 关于该模型的全局检验统计量见 Le Cessie 和 van Houwelingen (1991) 的相关研究. 通过替换 α 的极大似然估计得到 π 的估计. 当 $\pi(y,\boldsymbol{Z})$ 的模型如式(1.3)所示时, $\boldsymbol{\Theta}$ 的加权参数估计 $\hat{\boldsymbol{\Theta}}_{\mathrm{WP}}$ 的渐近分布见引理1.1.

1.2.2 选择概率的非参数光滑

当对均值函数已知晓的信息有限且样本量较大时, 非参数光滑是一个十分有效的工具, 例如, 若仅知道 π 是光滑的也就是连续可微. 两种常用的非参数光滑方法是核光滑和样条光滑. Copas (1983) 提出了用来绘制二值变量关于协变量图的核估计量. 本质上来说, 核估计是加权线性估计, 其权重与带宽的选择有关. 与核估计量的使用密切相关的一个重要问题是选择一个好的带宽值 (Eubank, 1987).

为简化符号, 令 $\boldsymbol{W}=(Y,\boldsymbol{Z}^{\mathrm{T}})^{\mathrm{T}}$, d 表示 $\boldsymbol{W}$ 中连续成分的个数, 若 Y 是 logistic 而 $\boldsymbol{Z}$ 是一元正态, 则 $d=1$. 令 K 表示第 r 阶核函数 $(r=2d)$, 令 h 表示带宽参数, 定义 $K_h(\cdot)=K(\cdot/h)$. 则 $\pi(\boldsymbol{w})$ 的核估计量 (Nadaraya,

1964; Watson, 1964)

$$\hat{\pi}(\boldsymbol{w}) = \frac{\sum_{i=1}^{n} \delta_i K_h(\boldsymbol{w} - \boldsymbol{W}_i)}{\sum_{i=1}^{n} K_h(\boldsymbol{w} - \boldsymbol{W}_i)}. \tag{1.4}$$

式中, 为了避免边界问题, 对边界点使用边界核 (Eubank, 1987) 从而使偏差比例保持为 h^r. 给定阶数 r, 核函数 K 的选择通常对非参数估计几乎没有影响, 因而对参数 $\boldsymbol{\Theta}$ 的影响更小. 然而, 带宽选择是至关重要的, $\boldsymbol{\Theta}$ 的加权半参数估计 $\hat{\boldsymbol{\Theta}}_{\mathrm{WS}}$ 的渐近分布见定理1.1.

1.3　渐 近 理 论

1.3.1　主要结论

注意到给定 $(\boldsymbol{X}_i, \boldsymbol{Z}_i)$, Y_i 的均值函数是通过均值函数 $g(\boldsymbol{Z}_i, \boldsymbol{X}_i, \boldsymbol{\Theta})$ 表示的, 且其分析最本质上是估计 $\boldsymbol{\Theta}$. 假定 $\boldsymbol{X}$ 是 MAR, 即 $P(\delta_i = 1|Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) = \pi(Y_i, \boldsymbol{Z}_i) = \pi(\boldsymbol{W}_i)$. 本章的主要结论见定理1.1, 为了进行比较, 首先在引理1.1中给出加权参数估计量的相关理论性质.

引理 1.1　*假定 $\pi(\boldsymbol{w})$ 满足模型(1.3)且含冗余参数 α. 令 $\hat{\boldsymbol{\Theta}}_{\mathrm{WP}}$ 表示解估计方程 $\boldsymbol{U}_n(\boldsymbol{\Theta}, \hat{\pi}) = 0$ 得到的 $\boldsymbol{\Theta}$ 的一个估计, 其中 $\hat{\pi}$ 是拟合模型(1.3)得到的 π_i 的极大似然估计. 则 $\hat{\boldsymbol{\Theta}}_{\mathrm{WP}}$ 是 $\boldsymbol{\Theta}$ 的一个相合估计且 $\sqrt{n}(\hat{\boldsymbol{\Theta}}_{\mathrm{WP}} - \boldsymbol{\Theta})$ 服从均值为 0, 方程如式(1.7)的正态分布.*

与引理1.1相似的结果见 Robins 等 (1994) 的相关研究. 对于任何非奇异矩阵 G, 定义 $G^{-\mathrm{T}} = (G^{-1})^{\mathrm{T}}$, 从而给出 $\boldsymbol{\Theta}$ 的加权半参数估计的重要结论.

定理 1.1　*假定存在常数 $c > 0$, $\pi(\boldsymbol{w}) \geqslant c$ 是 $\boldsymbol{w}$ 的光滑函数满足其第 r 阶导数存在且连续, $\boldsymbol{S}(\boldsymbol{\Theta}) = \mathcal{X}\{Y - g(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Theta})\}$ 有有限二阶矩, 且 $E(-\partial/\partial\boldsymbol{\Theta} S(\boldsymbol{\Theta}))$ 是正定的. 令 $\hat{\boldsymbol{\Theta}}_{\mathrm{WS}}$ 表示估计方程 $\boldsymbol{U}_n(\boldsymbol{\Theta}, \hat{\pi}) = 0$ 的解, 其中 $\hat{\pi}$ 是 π 的核估计, 且带宽参数 h 满足 $nh^{2d} \to \infty, nh^{2r} \to 0$. 则 $\hat{\boldsymbol{\Theta}}_{\mathrm{WS}}$ 是 $\boldsymbol{\Theta}$ 的相合估计, 且 $\sqrt{n}(\hat{\boldsymbol{\Theta}}_{\mathrm{WS}} - \boldsymbol{\Theta})$ 服从正态分布, 其中均值为 0, 方差为*

$$\boldsymbol{G}_n^{-1}(\hat{\boldsymbol{\Theta}}_{\mathrm{WS}}, \hat{\pi}) \left(\frac{1}{n} \sum_{i=1}^{n} \left[\frac{\delta_i}{\hat{\pi}_i^2} \{\boldsymbol{S}_i(\hat{\boldsymbol{\Theta}}_{\mathrm{WS}}) - \hat{\boldsymbol{S}}_i^*(\hat{\boldsymbol{\Theta}}_{\mathrm{WS}})\} \times \{\boldsymbol{S}_i(\hat{\boldsymbol{\Theta}}_{\mathrm{WS}}) \right. \right.$$

$$- \hat{\boldsymbol{S}}_i^*(\hat{\boldsymbol{\Theta}}_{\mathrm{WS}})\}^{\mathrm{T}} + \frac{\delta_i}{\pi_i}\hat{\boldsymbol{S}}_i^*(\hat{\boldsymbol{\Theta}}_{\mathrm{WS}})\{\hat{\boldsymbol{S}}_i^*(\hat{\boldsymbol{\Theta}}_{\mathrm{WS}})\}^{\mathrm{T}}\Big]\Big)\boldsymbol{G}_n^{-\mathrm{T}}(\hat{\boldsymbol{\Theta}}_{\mathrm{WS}}, \hat{\pi}) \tag{1.5}$$

其中

$$\boldsymbol{G}_n(\boldsymbol{\Theta}, \pi) = n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\Big\{-\frac{\partial}{\partial \boldsymbol{\Theta}^{\mathrm{T}}}\boldsymbol{S}_i(\boldsymbol{\Theta})\Big\}, \tag{1.6}$$

$$\hat{\boldsymbol{S}}_i^*(\boldsymbol{\Theta}) = \frac{\sum_k \frac{\delta_k \boldsymbol{S}_k(\boldsymbol{\Theta})}{\hat{\pi}_k}K_h(W_i - W_k)}{\sum_k K_h(W_i - W_k)}. \tag{1.7}$$

1.3.2 渐近无偏性和带宽选择

已知 π, 加权估计 $\hat{\boldsymbol{\Theta}}_{\mathrm{W}}$ 的估计得分为 $\boldsymbol{U}_n(\boldsymbol{\Theta}, \pi)$. 易得 $E(\boldsymbol{U}_n(\boldsymbol{\Theta}, \pi)) = 0$, 确保了 $\hat{\boldsymbol{\Theta}}_{\mathrm{W}}$ 的相合性. 当 π 通过核光滑估计时, 需要保证在合适的带宽下 $\boldsymbol{U}_n(\boldsymbol{\Theta}, \hat{\pi})$ 是渐近无偏的 (即 $E(\boldsymbol{U}_n(\boldsymbol{\Theta}, \hat{\pi}) \to 0)$. 注意, 定义 $\boldsymbol{W} = (Y, \boldsymbol{Z}^{\mathrm{T}})^{\mathrm{T}}$ 以及 $K_h(\cdot) = K(\cdot/h)$. 当 $nh^{2d} \to \infty, nh^{2r} \to 0$ 时, 有

$$E(\boldsymbol{U}_n(\boldsymbol{\Theta}, \hat{\pi})) = n^{1/2}E(\boldsymbol{S}_1(\boldsymbol{\Theta})) + o(1) = o(1).$$

从而保证了 $\boldsymbol{U}_n(\boldsymbol{\Theta}, \hat{\pi})$ 的渐近无偏性, 进而保证了 $\hat{\boldsymbol{\Theta}}_{\mathrm{WS}}$ 的渐近相合性. 这里的最优带宽比例与 Carroll 和 Wand(1991) 是一样的. 一种重要的情形是当 Y 是二值的, $\pi(\boldsymbol{w})$ 产生 Z 的两个函数: 一个是对于 $Y = 0$ 的, 另一个是关于 $Y = 1$ 的. 因此, 核光滑的维度是 $\boldsymbol{Z}$ 的维度. 核光滑的偏差率为阶 h^r, $\boldsymbol{U}_n(\boldsymbol{\Theta}, \hat{\pi}) - \boldsymbol{U}_n(\boldsymbol{\Theta}, \pi)$ 的偏差率为阶 $n^{1/2}h^r$. 线性部分 $\boldsymbol{U}_n(\boldsymbol{\Theta}, \hat{\pi})$ 的残差项其阶至多为 $O_p\{n^{1/2}(\hat{\pi} - \pi)^2\}$, 即为 $O_p\{n^{1/2}h^{2r} + n^{-1/2}h^{-d}\}$. 因此, $nh^{2d} \to \infty, nh^{2r} \to 0$ 是需要的.

1.3.3 估计选择概率的结果

$\hat{\boldsymbol{\Theta}}_{\mathrm{WP}}$ 和 $\hat{\boldsymbol{\Theta}}_{\mathrm{WS}}$ 的统计推断基于引理1.1和定理1.1. 例如, 为了基于加权半参数估计构建 $\boldsymbol{\Theta}$ 的置信区间, 首先得到 $\hat{\boldsymbol{\Theta}}_{\mathrm{WS}}$ 的点估计, 再采用标准正态分布理论 $\hat{\boldsymbol{\Theta}}_{\mathrm{WS}} \pm \boldsymbol{Z}_{\alpha/2}\hat{\sigma}_{\mathrm{WS}}$, 其中 $\hat{\sigma}^2_{\mathrm{WS}}$ 等于式(1.5)对角元素除以 n. 式(1.5)的计算过于庞大且在编程上容易遇到困难, 相应地, 可以选择

bootstrap (自举) 方法进行估计协方差矩阵. 在后续的模拟中, 将对二者进行详细比较并给出详细分析. 从 Wang 等 (1997) 的附录中可以注意到 $n^{1/2}(\hat{\boldsymbol{\Theta}}_{\mathrm{WP}}-\boldsymbol{\Theta})$ 和 $n^{1/2}(\hat{\boldsymbol{\Theta}}_{\mathrm{WS}}-\boldsymbol{\Theta})$ 有如下相似的协方差结构:

$$\boldsymbol{G}^{-1}(\boldsymbol{\Theta})[\boldsymbol{V}(\boldsymbol{\Theta},\pi)-\mathrm{cov}\{\boldsymbol{U}_n(\boldsymbol{\Theta},\hat{\pi})-\boldsymbol{U}_n(\boldsymbol{\Theta},\pi)\}]\boldsymbol{G}^{-\mathrm{T}}(\boldsymbol{\Theta}),$$

其中, $\boldsymbol{G}(\boldsymbol{\Theta})=-E[(\partial/\partial\boldsymbol{\Theta}^{\mathrm{T}})S]$ 和 $\boldsymbol{V}(\boldsymbol{\Theta},\pi)=\mathrm{cov}\boldsymbol{U}_n(\boldsymbol{\Theta},\pi)$.

从 Wang 等 (1997) 研究中可知 $\mathrm{cov}\{\boldsymbol{U}_n(\boldsymbol{\Theta},\pi),\boldsymbol{U}_n(\boldsymbol{\Theta},\hat{\pi})-\boldsymbol{U}_n(\boldsymbol{\Theta},\pi)\}$ $=\mathrm{cov}\{\boldsymbol{U}_n(\boldsymbol{\Theta},\hat{\pi})-\boldsymbol{U}_n(\boldsymbol{\Theta},\pi)\}+o(1)$. 因此, $\mathrm{cov}\{n^{1/2}(\hat{\boldsymbol{\Theta}}_{\mathrm{W}}-\boldsymbol{\Theta})\}-\mathrm{cov}\{n^{1/2}$ $\cdot(\hat{\boldsymbol{\Theta}}_{\mathrm{WP}}-\boldsymbol{\Theta})\}$ 和 $\mathrm{cov}\{n^{1/2}(\hat{\boldsymbol{\Theta}}_{\mathrm{W}}-\boldsymbol{\Theta})\}-\mathrm{cov}\{n^{1/2}(\hat{\boldsymbol{\Theta}}_{\mathrm{WS}}-\boldsymbol{\Theta})\}$ 是正定矩阵. 这意味着 $\boldsymbol{\Theta}$ 的估计精度可以通过 $\pi(\boldsymbol{w})$ 的数据调整得到提高. 这种现象是很常见的 (Robins et al., 1994; Rosenbaum, 1987).

1.4　两阶段病例对照研究中的应用

1.4.1　概述

1.3节所讨论的性质是针对含缺失 $\boldsymbol{X}$ 的 $(Y,\boldsymbol{Z},\boldsymbol{X})$ 样本. 本节将估计得分 $\boldsymbol{U}_n(\boldsymbol{\Theta},\hat{\pi})$ 应用到两阶段病例对照研究中, 其中样本基于回溯性研究获取. 在一项回溯性研究中, 获得了患有和不患有该疾病的人群, 然后确定了协变量的值. 分析基于来源人群具有如下性质

$$P(Y=1|\boldsymbol{Z},\boldsymbol{X})=H(\theta_0+\boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{Z}+\boldsymbol{\theta}_2^{\mathrm{T}}\boldsymbol{X}),\ H(u)=\{1+\exp(-u)\}^{-1}.$$

注意到在由两个子样本构成的病例对照研究中, 一部分来自 $Y=1$ 的总体而另一部分则来自 $Y=0$ 的总体. 在两阶段病例对照研究中, 第一阶段的协变量 $\boldsymbol{Z}$ 的 n_0 个控制以及 n_1 个病例是可观测到的, 而在第二阶段中, 协变量 $\boldsymbol{X}$ 则是基于设计选择计划的研究个体的一个子集上进行测量的. Prentice 和 Pyke(1979) 的研究表明当所有的 $\boldsymbol{X}$ 可观测到时, 给定 $Y=y$ 时 $(\boldsymbol{Z},\boldsymbol{X})$ 的回溯性密度函数满足

$$\mathrm{d}F_{\boldsymbol{Z},\boldsymbol{X}|Y=y}(\boldsymbol{z},\boldsymbol{x})$$

$$= \frac{n}{n_y} H^y(\theta_0^* + \boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{z} + \boldsymbol{\theta}_2^{\mathrm{T}}\boldsymbol{x}) \times (1 - H^y(\theta_0^* + \boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{z} + \boldsymbol{\theta}_2^{\mathrm{T}}\boldsymbol{x}))^{1-y} \mathrm{d}Q(\boldsymbol{z}, \boldsymbol{x}),$$

其中, θ_0^* 是一个新的截矩项, $Q(\boldsymbol{Z}, \boldsymbol{X})$ 是在病例对照研究机制中 $(\boldsymbol{Z}, \boldsymbol{X})$ 的边际分布函数. 注意到 θ_0 是不可识别的. Prentice 和 Pyke(1979) 还指出当 Y 是二值时, 可以利用普通 logistic 回归得到斜率 $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ 的估计. 因此, 病例对照数据的前瞻性分析给出了斜率的一致估计, 并且斜率的前瞻性协方差公式是渐近正确的.

1.4.2 前瞻性估计方程

需要指出的是, 引理1.1和定理1.1并未对关于两阶段病例对照研究中的加权估计给出任何必要的结论. 主要的不同点在于第一阶段的抽样是回溯性的而且其样本 n_0 和 n_1 都是固定的, 引理1.1和定理1.1则是基于前瞻性抽样的结果, 虽然总样本量是固定的, 但其对应的病例和对照的样本量是随机的. 令 $\boldsymbol{S}_i(\boldsymbol{\Theta}^*) = \mathcal{X}_i\{Y_i - g(\boldsymbol{Z}_i, \boldsymbol{X}_i, \boldsymbol{\Theta}^*)\}$, $\boldsymbol{\Theta}^* = (\theta_0^*, \boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}})^{\mathrm{T}}$. 下面的内容表明基于病例对照抽样设计的前瞻性估计方程的无偏性:

$$\begin{aligned}
&E(\boldsymbol{U}_n(\boldsymbol{\Theta}^*, \pi)|\tilde{Y}) \\
=&E(E(\boldsymbol{U}_n(\boldsymbol{\Theta}^*, \pi)|\tilde{Y}, \tilde{\boldsymbol{Z}})|\tilde{Y}) = E(n^{-1/2}\sum_{i=1}^{n} \boldsymbol{S}_i|\tilde{Y}) \\
=&n^{-1/2}\sum_{y=0}^{1} n_y \int (1, \boldsymbol{z}^{\mathrm{T}}, \boldsymbol{x}^{\mathrm{T}})^{\mathrm{T}}\{y - H(\theta_0^* + \boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{z} + \boldsymbol{\theta}_2^{\mathrm{T}}\boldsymbol{x})\} \\
&\times \mathrm{d}F_{\boldsymbol{Z},\boldsymbol{X}|Y=y}(\boldsymbol{z}, \boldsymbol{x}) \\
=&n^{-1/2}\Bigg[n_1 \int (1, \boldsymbol{z}^{\mathrm{T}}, \boldsymbol{x}^{\mathrm{T}})^{\mathrm{T}} \mathrm{d}F_{\boldsymbol{z},\boldsymbol{x}|Y=1}(\boldsymbol{z}, \boldsymbol{x}) \\
&- (1, \boldsymbol{z}^{\mathrm{T}}, \boldsymbol{x}^{\mathrm{T}})^{\mathrm{T}} H(\theta_0^* + \boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{z} + \boldsymbol{\theta}_2^{\mathrm{T}}\boldsymbol{x}) \times \sum_{y=0}^{1} n_y \mathrm{d}F_{\boldsymbol{z},\boldsymbol{x}|Y=y}(\boldsymbol{z}, \boldsymbol{x})\Bigg] \\
=&n^{-1/2}\Bigg[\int (1, \boldsymbol{Z}^{\mathrm{T}}, \boldsymbol{x}^{\mathrm{T}})^{\mathrm{T}} H(\theta_0^* + \boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{z} + \boldsymbol{\theta}_2^{\mathrm{T}}\boldsymbol{x}) \mathrm{d}Q(\boldsymbol{z}, \boldsymbol{x}) \\
&- \int (1, \boldsymbol{z}^{\mathrm{T}}, \boldsymbol{x}^{\mathrm{T}})^{\mathrm{T}} H(\theta_0^* + \boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{z} + \boldsymbol{\theta}_2^{\mathrm{T}}\boldsymbol{x}) \mathrm{d}Q(\boldsymbol{z}, \boldsymbol{x})\Bigg] = 0.
\end{aligned}$$