应用概率统计 第 40 卷
第 5 期 2024 年 10 月

Chinese Journal of Applied Probability and Statistics
Oct., 2024, Vol. 40, No. 5, pp. 800–818
doi: 10.12460/j.issn.1001-4268.aps.2024.2022083

协变量随机右删失时变系数模型的估计 *

柴旺[1] 尹俊平[2, 3, 4] 孙志华†[1]

[1] 中国科学院大学数学科学学院, 北京, 100049
[2] 北京应用物理与计算数学研究所, 北京, 100094
[3] 计算物理全国重点实验室, 北京, 100088
[4] 上海张江数学研究院, 上海, 201203

摘 要: 数据经常因为个体失访, 退出实验或者研究结束而出现右删失的现象. 右删失数据的研究吸引了很多研究者的兴趣. 文献中大部分研究集中在响应变量出现右删失的情况. 回归模型中的协变量也可能出现右删失, 但相关的研究并不多. 本文研究协变量随机右删失时变系数模型的估计问题. 我们利用逆概率加权方法直接对目标函数进行调整, 而不是调整被删失的协变量, 来处理数据的删失. 所得估计的渐近性质得到严格证明. 通过数值模拟和实例分析, 可以看到本文所提方法具有很好的有限样本性质.

关键词: 变系数模型; 局部线性估计; 随机右删失协变量; 渐近性质
中图分类号: O212.7

英文引用格式: CHAI W, YIN J P, SUN Z H. Estimation of varying coefficient model with randomly right-censored covariate [J]. *Chinese J Appl Probab Statist*, 2024, **40**(**5**): 800–818. (in Chinese)

1 引言

在临床医学、生物、经济等领域, 数据经常因为多种因素出现右删失, 比如实验对象死于非感兴趣的其他事件, 实验对象中途退出, 失访等. 右删失数据作为生存分析中一类重要数据类型, 吸引了很多研究者对其进行研究, 相关研究成果颇丰. 然而, 大多数研究集中在响应变量删失的情况, 具体可参考文献 [1–3] 及其相关参考文献.

在很多实际问题中, 协变量通常被随机右删失, 比如在研究子代与亲代之间遗传性疾病发病年龄的关系时, 亲代的发病年龄与患病类型往往是评估关系的重要因素, 但在实验结束时, 亲代并不一定总会患病, 或者由于其他因素使得研究者无法观察到亲代的发病年龄, 这时亲代的发病年龄 (协变量) 就被右删失. 文献 [4–5] 表示, 不加处理地直接使用删失数据会得到有偏的估计并使检验问题的第一类错误增大. 在实际应用中, 处理删失数据的一个简单方法是完全记录分析方法 (complete-case analysis), 即去掉含有被删失数据的

* 北京市自然科学基金重点项目 (批准号: Z200001), 国家自然科学基金项目 (12071457, 12271504), 国家自然科学基金重点项目 (批准号: 12031016), 国家重点研发计划 (批准号: 2023YFA1009000, 2023YFA1009004, 2020YFA0712203, 2020YFA0712201) 和国家自然科学基金重大项目 (批准号: 12292980, 12292984) 资助.

† 通讯作者, E-mail: sunzh@ucas.ac.cn.

本文 2022 年 6 月 9 日收到, 2023 年 2 月 12 日收到修改稿, 2023 年 4 月 20 日录用.

观测. Atem 等[6] 指出, 当删失变量和响应变量独立时, 该方法可以得到无偏估计, 但当删失比例比较大时, 这种方法会严重影响估计的效率. 文献 [7–8] 将删失数据替换为固定的数或函数, Lynn[9] 假定出现删失的变量的分布, 并将删失数据替换为删失数据的条件期望, May 等[10] 采用极大似然估计处理删失数据.

变系数模型是由经典的线性模型发展而来, 它将线性模型中的参数用协变量的函数代替, 相比参数模型, 能够有效地避免维数祸根的问题, 同时也具有线性模型的很好的可解释性. 自 Hastie 和 Tibshirani[11] 首次提出变系数模型以来, 关于变系数模型的研究已有一系列丰富的成果. Fan 和 Zhang[12] 提出了两阶段 (two-step) 估计方法用于处理系数函数具有不同光滑度时变系数模型的估计问题, Cai 等[13] 提出了局部极大似然估计, Fan 和 Zhang[14] 详细总结了变系数模型的估计方法及其应用. 据我们所知, 目前还没有关于协变量右删失时变系数模型估计问题的研究.

本文对协变量随机删失时的变系数模型的估计问题展开研究. 对于右删失协变量的处理, 本文所用方法不同于文献 [7–10] 中将被删失的协变量进行替代或假定分布的方法, 而是直接采用逆概率加权方法对目标函数进行调整. 采用上面方法基于下面的考虑, 用数值或函数对删失协变量进行替代需要基于模型的假定, 这样可能会因模型误定而导致估计是有偏的. 假定出现删失的变量的分布同样存在误定的风险. 文献中未见有利用逆概率加权方法对协变量的删失进行处理的相关工作, 但是基于逆概率加权对响应变量的删失进行处理是常见的. 基于逆概率加权方法处理响应变量删失的方法有两种, 一种称为综合数据 (SD, synthetic data) 方法, 参见文献 [1,15–16], 另一种称为加权最小二乘 (WLS, weighted least squares) 方法, 参见文献 [5,17]. 文献 [17] 显示基于 WLS 方法对目标函数进行矫正, 相比直接对变量进行矫正的 SD 方法, 具有更好的有限样本性质. 这里, 我们利用 WLS 方法构建目标函数, 进而构建回归系数函数的估计, 估计的渐近正态性得到严格的证明. 通过数值模拟研究, 验证了本文方法相比不处理删失数据的 Naive 方法具有更高的估计精度. 同时, 也比较了本文所提方法与完整数据下的估计方法. 这里完整数据下的方法对应数据是准确观测的, 不存在删失. 这个方法作为一个比较的标准. 数值模拟研究结果显示本文所提方法的效果和完整数据下估计效果相差不大. 本文所提方法也被用来分析两个实际数据: 弗雷明翰心脏研究 (FHS) 数据和原发性胆汁性肝硬化 (PBC) 数据.

本文剩余部分安排如下: 第 2 节对协变量随机右删失变系数模型进行介绍; 第 3 节构建了回归系数的估计; 第 4 节给出估计的渐近性质; 第 5 和第 6 节分别给出数值模拟和实例分析结果. 定理的证明和所需的条件在第 7 节中给出.

2　模型介绍

本文考虑具有下面形式的变系数模型:

$$Y = \boldsymbol{X}^{\top}\beta(U) + \varepsilon, \tag{1}$$

其中 Y 为 1 维响应变量, U 为 1 维协变量, $\boldsymbol{X}=(V,\boldsymbol{Z}^{\top})^{\top}$ 为 p 维协变量, 协变量 $\boldsymbol{X}$ 的分量 V 和 $\boldsymbol{Z}$ 分别为 1 维和 $p-1$ 维随机变量. 函数系数 $\beta(U)=\big(\beta_1(U),\beta_2(U),\cdots,\beta_p(U)\big)^{\top}$ 为 p 维向量, 每个分量均为未知的光滑函数. 模型误差 ε 满足: $\mathsf{E}(\varepsilon|\boldsymbol{X},U)=0,\quad \mathsf{Var}\,(\varepsilon|\boldsymbol{X},U)<\infty$.

在模型 (1) 中, 变量 Y, $\boldsymbol{Z}$ 和 U 都是完全准确地观察到的, 协变量 V 被随机右删失. 我们观察到的不是变量 V, 而是 $T=\min(V,C)$, 这里 C 是删失时间. 同时变量 V 是否被删失我们也是知道的, 也即, 我们还观察到删失指示变量 $\delta=\mathrm{I}(V\leqslant C)$, 这里 $\mathrm{I}(\cdot)$ 是示性函数. 假设删失变量的生存函数为 $G(\cdot)$.

本文我们假设随机删失机制成立, 即协变量 V 和删失变量 C 满足如下两个条件: 1) 独立性条件: 协变量 V 和删失变量 C 独立; 2) 条件独立性条件: $\mathsf{P}(V\leqslant C|V,\boldsymbol{Z},U,Y)=\mathsf{P}(V\leqslant C|V)$. 随机删失假设是一个很常见的假设, 更多细节可以参见文献 [15,17]. 上述独立性条件和文献 [17] 中响应变量出现删失时独立性条件稍微不同, 这是因为本文所考虑的不是响应变量出现删失, 而是协变量出现删失的情况. 这里给出的条件虽然和文献 [17] 给出的条件形式不同, 但都是为了保证估计的渐近正态性等渐近性质成立.

3 估计方法

记 $\widetilde{\boldsymbol{X}}=(T,\boldsymbol{Z}^{\top})^{\top}$. 假设我们有独立同分布样本数据 $\{Y_i,\widetilde{\boldsymbol{X}}_i,U_i,\delta_i\}_{i=1}^n$. 当数据准确观测时, 基于样本数据 $\{Y_i,\boldsymbol{X}_i,U_i\}_{i=1}^n$, 构建局部线性目标函数如下:

$$\sum_{i=1}^{n}\left[Y_i-\sum_{j=1}^{p}\Big\{a_j+b_j(U_i-u)\Big\}X_{ij}\right]^2 k\left(\frac{U_i-u}{h_n}\right),$$

这里 $k(\cdot)$ 为核函数, h_n 为带宽. 使上面目标函数关于 $\{(a_j,b_j),j=1,2,\cdots,p\}$ 达到最小就可求得系数函数 $\beta(u)$ 以及 $\beta'(u)$ 的估计. 这里 $\beta'(u)$ 是 $\beta(u)$ 的导数.

当协变量 V 出现右删失时, 上面的目标函数因为有些个体的变量 V 没有观测到而无法求解. 如果用观察到的生存时间 T 取代 V, 也即文献中的 Naive 方法, 得到的估计经常是有偏的, 参见文献 [4,15]. 我们利用逆概率加权方法来处理数据的删失从而消除数据的删失导致的估计是有偏的这一不利后果.

基于前面的讨论, 我们不考虑对删失变量直接进行逆概率加权的调整方法, 也即不用所谓的 SD 方法, 而是直接对目标函数进行调整, 得到调整的目标函数如下:

$$\sum_{i=1}^{n}\left[Y_i-\sum_{j=1}^{p}\Big\{a_j+b_j(U_i-u)\Big\}\widetilde{X}_{ij}\right]^2 \frac{\delta_i}{\widehat{G}_n(T_i)}k\left(\frac{U_i-u}{h_n}\right), \tag{2}$$

这里 $\widehat{G}_n(t)$ 为删失时间 C 的生存函数 $G(t)$ 的 Kaplan-Meier 估计:

$$\widehat{G}_n(t)=\prod_{j:T_j\leqslant t}\left\{1-\frac{1}{\sum\limits_{k=1}^{n}\mathrm{I}(T_k\geqslant T_j)}\right\}^{1-\delta_j}.$$

关于 Kaplan-Meier 估计的更多细节可参考文献 [1,18].

记 $\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)^\top$, $\widetilde{\boldsymbol{X}}_i=(T_i,\boldsymbol{Z}_i^\top)^\top$, $i=1,2,\cdots,n$. 定义矩阵 $\widetilde{\mathcal{X}}_u$ 和 W_u 分别具有如下形式:

$$\widetilde{\mathcal{X}}_u^\top=\begin{pmatrix}\widetilde{\boldsymbol{X}}_1^\top & \dfrac{U_1-u}{h_n}\widetilde{\boldsymbol{X}}_1^\top\\ \widetilde{\boldsymbol{X}}_2^\top & \dfrac{U_2-u}{h_n}\widetilde{\boldsymbol{X}}_2^\top\\ \vdots & \vdots\\ \widetilde{\boldsymbol{X}}_n^\top & \dfrac{U_n-u}{h_n}\widetilde{\boldsymbol{X}}_n^\top\end{pmatrix},$$

$$W_u=\mathtt{diag}\left(\frac{\delta_1}{\widehat{G}_n(T_1)}k\big(\frac{U_1-u}{h_n}\big),\frac{\delta_2}{\widehat{G}_n(T_2)}k\big(\frac{U_2-u}{h_n}\big),\cdots,\frac{\delta_n}{\widehat{G}_n(T_n)}k\big(\frac{U_n-u}{h_n}\big)\right)$$

记 $\theta(u)=\big(\beta(u)^\top,h_n\beta'(u)^\top\big)^\top$, 通过最小化目标函数 (2), 可得到 $\theta(u)$ 的估计:

$$\hat{\theta}_n(u)=(\widetilde{\mathcal{X}}_uW_u\widetilde{\mathcal{X}}_u^\top)^{-1}\widetilde{\mathcal{X}}_uW_u\boldsymbol{Y}. \tag{3}$$

取 $\hat{\theta}_n(u)$ 的前 p 个分量即为 $\beta(u)=\big(\beta_1(u),\beta_2(u),\cdots,\beta_p(u)\big)^\top$ 的估计, 记为 $\hat{\beta}_n(u)$.

下面考虑带宽的选择. 我们利用删一交叉验证 (leave-one-out cross-validation) 方法选择带宽, 即选择使得下面的交叉验证目标函数达到最小的带宽:

$$\hat{h}_n=\arg\min_{h_n}\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\widehat{G}_n(T_i)}\left\{Y_i-\widetilde{\boldsymbol{X}}_i^\top\hat{\beta}_n^{(-i)}(U_i)\right\}^2.$$

上面交叉验证目标函数中, 对 $i=1,2,\cdots,n$, $\hat{\beta}_n^{(-i)}(U_i)$ 为 $\beta(U_i)$ 的删一估计, 其计算细节在下面给出.

首先关于参数 $(a_1,a_2,\cdots,\ a_p,h_nb_1,h_nb_2,\cdots,h_nb_p)^\top$ 最小化下面的删一目标函数:

$$\sum_{k=1,k\neq i}^{n}\left[Y_k-\sum_{j=1}^{p}\left\{a_j+b_j(U_k-U_i)\right\}\widetilde{X}_{kj}\right]^2\frac{\delta_k}{\widehat{G}_n(T_k)}k\left(\frac{U_k-U_i}{h_n}\right).$$

假设上面问题的解为 $\hat{\theta}_n^{(-i)}(U_i)$. 取 $\hat{\theta}_n^{(-i)}(U_i)$ 的前 p 个分量, 即为 $\beta(U_i)$ 的删一估计 $\hat{\beta}_n^{(-i)}(U_i)$, $i=1,2,\cdots,n$. 上面构建删一估计的思想就是在估计 $\hat{\beta}_n^{(-i)}(U_i)$ 时, 构建的目标函数只用除了第 i 个个体的观测数据之外的 $n-1$ 个观测数据.

4 渐近性质

首先给出一些记号:

$$k_{ij}=\int u^i k^j(u)\mathrm{d}u, i,j=0,1,2,$$

$$H(t)=\mathsf{P}(T\leqslant t),\overline{H}(t)=1-H(t),$$

$$H_0(t)=\mathsf{P}(T\leqslant t,\delta=0),\ \overline{H}_0(t)=\mathsf{P}(T>t,\delta=0),$$

$$\eta\left(T,\delta;t\right)=-G(t)\left[\int_0^{T\wedge t}[\overline{H}(s)]^{-2}\mathrm{d}\overline{H}_0(s)+\frac{1}{\overline{H}(T)}I\left(T\leqslant t,\delta=0\right)\right].$$

记 $f_u(u)$ 是变量 U 的密度函数, $Q(u)=\mathsf{E}(\boldsymbol{X}\boldsymbol{X}^\top|U=u)$; $O=(T,\delta,U,\boldsymbol{X},\varepsilon)$, $O_i=(T_i,\delta_i,U_i,\boldsymbol{X}_i,\varepsilon_i)$,

$$\begin{aligned}W(O_i;O_j;u)=&\frac{1}{2\sqrt{h_n}}\frac{\delta_i}{G(T_i)}\left\{k\left(\frac{U_i-u}{h_n}\right)+\frac{1}{G(T_i)}\eta\left(T_j,\delta_j;T_i\right)k\left(\frac{U_i-u}{h_n}\right)\right\}\boldsymbol{X}_i\varepsilon_i\\&+\frac{1}{2\sqrt{h_n}}\frac{\delta_j}{G(T_j)}\left\{k\left(\frac{U_j-u}{h_n}\right)+\frac{1}{G(T_j)}\eta\left(T_i,\delta_i;T_j\right)k\left(\frac{U_j-u}{h_n}\right)\right\}\boldsymbol{X}_j\varepsilon_j,\end{aligned}$$

$i,j=1,2,\cdots,n$.

进一步记 $\Psi^{[1]}(O,u)=\mathsf{E}(W(O_i;O;u)\mid O_i)$, $\Sigma(u)=\mathsf{E}\left(\Psi^{[1]}(O,u)\right)^2$.

对上面所提函数型系数 $\beta(u)$ 的估计 $\hat{\beta}_n(u)$, 有下面的渐近正态性.

定理 1 当第 7 节中给出的条件 (C1)–(C7)成立时, 有

$$\sqrt{nh_n}\left(\hat{\beta}_n(u)-\beta(u)-\frac{1}{2}h_n^2k_{21}\beta^{(2)}(u)\right)\xrightarrow{L}\mathcal{N}(0,Q^{-1}(u)\Sigma(u)Q^{-1}(u)),$$

其中 $\beta^{(2)}(u)$ 表示 $\beta(u)$ 的 2 阶导数, $\xrightarrow{L}$ 表示依分布收敛.

在上面定理中, 估计的渐近方差 $Q^{-1}(u)\Sigma(u)Q^{-1}(u)$ 可用下式估计:

$$\left(\frac{1}{nh_n}\widetilde{\mathcal{X}}_uW_u\widetilde{\mathcal{X}}_u^\top\right)^{-1}\left\{\frac{1}{\sqrt{nh_n}}\widetilde{\mathcal{X}}_uW_u(Y-\widetilde{\mathcal{X}}_u^\top\theta)\right\}^{\otimes 2}\left(\frac{1}{nh_n}\widetilde{\mathcal{X}}_uW_u\widetilde{\mathcal{X}}_u^\top\right)^{-1}.$$

从定理 1 的证明容易得到上面渐近方差的估计的相合性.

5 数值模拟

接下来, 我们通过数值模拟研究来验证文章所提方法的有限样本性质.

我们考虑三个模型, 其中模型 3 的协变量 V 和删失变量 C 不满足独立性删失条件, 我们通过此模型来考察文章所提方法的稳健性, 具体细节如下:

模型 1 : $Y=(2U^2-4U+10)V+\varepsilon$, 这里 $X_1\sim\mathbf{U}(0,5)$, $U\sim\mathbf{U}(0,2)$, $\varepsilon\sim\mathbf{N}(0,1)$.

模型 2 : $Y=(\exp(U)+5)V+(U^2+5)Z+\varepsilon$, 这里 $U\sim\mathbf{U}(0,2)$, $V,Z\sim\mathbf{N}(1,1)$, $\varepsilon\sim\mathbf{N}(0,1)$.

模型 3 : $Y = (2U^2 - 4U + 10)V + \varepsilon$, 这里 $V \sim \mathbf{U}(0,5)$, $U \sim \mathbf{U}(0,2)$, $\varepsilon \sim \mathbf{N}(0,1)$. 当 $V \leqslant 2.5, C \sim \mathbf{E}(3)$, 如果 $V > 2.5, C \sim \mathbf{E}(\mu)$.

上面三个模型中, V 被随机右删失, 模型 1 和模型 2 的删失变量服从均值为 μ 的指数分布. 通过选取不同的 μ, 可以得到 V 的不同删失比例. 本文考虑删失比例为 20%, 40% 和 65% 的情况, 对应每一种删失比例, 样本量分别取 200 和 400. 采用文中所述的删一交叉验证 (leave-one-out cross-validation) 方法确定带宽. 为了进行对比研究, 我们还计算了样本数据未删失时的模拟结果和未对删失数据进行处理时的模拟结果, 我们记这两种方法为 Full 方法和 Naive 方法.

在估计参数时, 选取 Gauss 核函数. 我们基于 500 次模拟计算最后结果. 参考文献 [19] 和 [20] 中的方法, 我们采用积分均方误差 (integrated mean squared error, IMSE) 来评价估计的精度, IMSE 的定义如下:

$$\mathrm{IMSE}\big(\hat{\beta}_i(u)\big) = \frac{1}{500}\sum_{k=1}^{500}\frac{1}{100}\sum_{j=1}^{100}\Big\{\hat{\beta}_{ik}(u_j) - \beta_{ik}(u_j)\Big\}^2, \ i = 1 \text{ 或 } 2,$$

其中 $\hat{\beta}_{ik}(\cdot)$ 表示 $\beta_i(\cdot)$ 的第 k 次模拟的估计, u_j 为区间 (0, 2) 内均匀 (间隔 0.02) 取的 100 个点.

我们将模型 1 和模型 3 的模拟结果列于表 2. 模型 2 的模拟结果列于表 3, 其中 Proposed 表示本文所提方法, $\hat{\beta}_n^{\mathrm{ind}}(u)$ 为模型 1 中模型系数的估计, $\hat{\beta}_{1n}(u)$ 和 $\hat{\beta}_{2n}(u)$ 为模型 2 中模型系数的估计, $\hat{\beta}_n^{\mathrm{dep}}(u)$ 为模型 3 模型系数的估计. 进一步在区间 (0, 2) 内取 4 个点: $u_1 = 0.4, u_2 = 0.8, u_3 = 1.2$ 和 $u_4 = 1.6$. 我们计算了系数函数在 $u_i, i = 1, \cdots, 4$ 的值的估计的样本偏差 (Bias), 样本标准差 (SD) 以及根据定理 1 渐近方差的估计公式计算的渐近根方差的估计的均值 (MSD), 结果列于表 4–5. 模拟时所计算的带宽均值和删失变量分布所对应的 μ 值列于表 1 和表 6. 注意每一次模拟都需要计算出一个带宽, 然后基于 500 次模拟进一步算出一个结果. 因此无法一一展示每次模拟的带宽, 表 1 展示的是带宽的均值. 为了更直观地观察估计结果, 我们给出了样本量为 200 和 400 时函数系数的估计曲线, 如图 1–4.

从表 3 的结果可以看出, 当样本量一定, 删失比例减少, $\hat{\beta}_{1n}(u)$ 和 $\hat{\beta}_{2n}(u)$ 的积分均方误差 (IMSE) 也随之减小, 估计的精度提高. 当删失比例相同时, 随着样本量增加, $\hat{\beta}_{1n}(u)$ 和 $\hat{\beta}_{2n}(u)$ 的积分均方误差也同样减小. 对应于不同样本量和不同删失比例, 本文所提方法的估计效果都比直接使用删失数据的估计效果要好, 且估计精度与样本数据未删失时的模拟结果相差不大, 并且当删失比例增加时, 本文方法的估计精度变化不大, 直接使用删失数据的估计精度会显著变差. 从表 2 的结果可以看出, $\hat{\beta}_n(u)$ 的估计效果很好, 且 IMSE 值接近于 0, 并且当协变量 X_1 和删失变量 C 不满足独立性删失条件时也能够很好地估计函数系数, 可见本文所提方法具有很好的有限样本性质和稳健性. 从表 4–5 的结果可以看出, 当数据删失比例不大时, 文章给出的渐近标准差的估计结果和样本标准差的结果较为接近.

观察图 1–4, 可以观察到本文所提方法的非参数估计曲线接近于真实曲线, 并且和样本数据未删失时的估计曲线基本重合. 当样本量增加时, 估计的精度显著提高, 且删失比例对估计效果的影响不大, 当样本量不大时仍可以较好的估计函数系数. 由此可以看出, 本文方法在处理协变量随机右删失变系数模型的估计问题时具有很好的效果.

表 1 不同模型和删失率时、删失变量分布时的 μ 值

删失率/ 模型	20% Cens	40% Cens	65% Cens
1	10.864	4.375	1.836
2	4.453	1.758	0.585
3	46.875	5.656	0.986

表 2 不同样本量和删失率时, 非参数函数 $\hat{\beta}_n^{\mathrm{ind}}(u)$ 和 $\hat{\beta}_n^{\mathrm{dep}}(u)$ 的 IMSE 值

删失率	方法	200		400	
		$\hat{\beta}_n^{\mathrm{ind}}(u)$	$\hat{\beta}_n^{\mathrm{dep}}(u)$	$\hat{\beta}_n^{\mathrm{ind}}(u)$	$\hat{\beta}_n^{\mathrm{dep}}(u)$
20%	Full	0.006	0.006	0.003	0.003
	Proposed	0.008	0.006	0.004	0.003
	Naive	0.507	0.083	0.399	0.058
40%	Full	0.005	0.005	0.003	0.003
	Proposed	0.013	0.010	0.006	0.006
	Naive	2.782	1.767	2.488	1.573
65%	Full	0.005	0.006	0.003	0.003
	Proposed	0.057	0.067	0.022	0.032
	Naive	18.794	38.579	16.730	34.080

表 3 不同样本量和删失率时, 非参数函数 $\hat{\beta}_{1n}(u)$ 和 $\hat{\beta}_{2n}(u)$ 的 IMSE 值

删失率	方法	200		400	
		$\hat{\beta}_{1n}(u)$	$\hat{\beta}_{2n}(u)$	$\hat{\beta}_{1n}(u)$	$\hat{\beta}_{2n}(u)$
20%	Full	0.024	0.021	0.011	0.010
	Proposed	0.036	0.026	0.017	0.012
	Naive	0.376	0.783	0.248	0.643
40%	Full	0.022	0.021	0.011	0.010
	Proposed	0.068	0.040	0.033	0.018
	Naive	1.719	2.567	1.388	2.160
65%	Full	0.022	0.020	0.011	0.010
	Proposed	0.428	0.097	0.191	0.043
	Naive	17.401	6.710	14.992	6.607

6 实例分析

为说明本文所提方法的效果, 下面我们将上面所提方法应用于弗雷明翰心脏研究 (FHS) 数据集和原发性胆汁性肝硬化 (PBC) 数据集的分析.

表 4　不同样本量和删失率时, 模型 1 和模型 3 的参数估计在不同点 u 的样本偏差 (Bias), 样本标准差 (SD), 和渐近根方差的估计的均值 (MSD)

	删失率	u	200			400		
			Bias	SD	MSD	Bias	SD	MSD
$\hat{\beta}_n^{ind}(u)$	20%	0.4	0.037	0.071	0.066	0.028	0.052	0.046
		0.8	0.041	0.073	0.067	0.027	0.049	0.046
		1.2	0.036	0.072	0.064	0.029	0.049	0.048
		1.6	0.032	0.072	0.063	0.025	0.050	0.047
	40%	0.4	0.041	0.094	0.080	0.030	0.064	0.058
		0.8	0.047	0.090	0.085	0.029	0.065	0.060
		1.2	0.044	0.089	0.083	0.030	0.065	0.058
		1.6	0.040	0.096	0.082	0.026	0.065	0.056
	65%	0.4	0.061	0.165	0.139	0.038	0.115	0.102
		0.8	0.070	0.158	0.137	0.045	0.119	0.105
		1.2	0.085	0.164	0.143	0.052	0.117	0.103
		1.6	0.059	0.165	0.142	0.040	0.115	0.102
$\hat{\beta}_n^{dep}(u)$	20%	0.4	0.032	0.069	0.061	0.023	0.046	0.044
		0.8	0.027	0.061	0.056	0.022	0.046	0.041
		1.2	0.027	0.066	0.059	0.022	0.045	0.043
		1.6	0.030	0.061	0.054	0.024	0.048	0.044
	40%	0.4	0.043	0.082	0.073	0.031	0.060	0.054
		0.8	0.037	0.081	0.071	0.031	0.060	0.058
		1.2	0.031	0.087	0.075	0.037	0.058	0.055
		1.6	0.032	0.082	0.073	0.033	0.061	0.055
	65%	0.4	0.056	0.194	0.171	0.042	0.139	0.127
		0.8	0.080	0.190	0.183	0.057	0.142	0.136
		1.2	0.084	0.198	0.172	0.062	0.148	0.133
		1.6	0.037	0.199	0.165	0.055	0.135	0.120

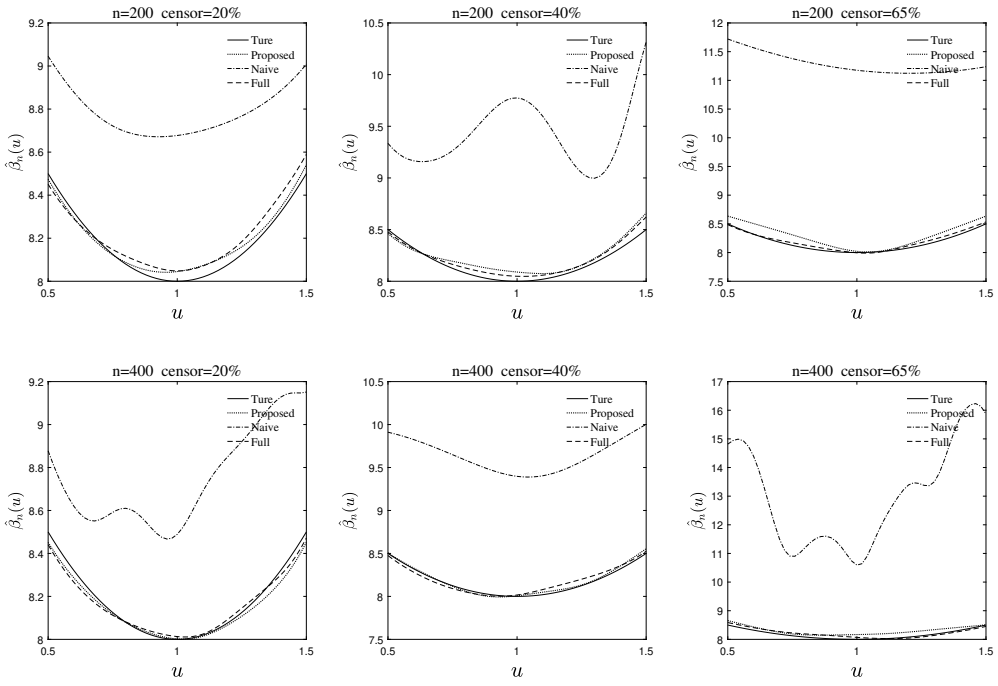

图 1　不同方法下模型 1 中函数系数 $\beta(u)$ 的估计曲线

表 5 不同样本量和删失率时, 模型 2 的参数估计在不同点 u 的样本偏差 (Bias), 样本标准差 (SD), 和渐近根方差的估计的均值 (MSD)

	删失率	u	200			400		
			Bias	SD	MSD	Bias	SD	MSD
$\hat{\beta}_{1n}(u)$	20%	0.4	0.022	0.138	0.117	0.017	0.101	0.099
		0.8	0.021	0.126	0.104	0.027	0.086	0.077
		1.2	0.043	0.142	0.120	0.032	0.096	0.087
		1.6	0.051	0.125	0.106	0.032	0.088	0.079
	40%	0.4	0.034	0.205	0.177	0.031	0.139	0.116
		0.8	0.033	0.150	0.126	0.029	0.106	0.088
		1.2	0.049	0.197	0.173	0.051	0.142	0.127
		1.6	0.054	0.167	0.135	0.036	0.103	0.090
	65%	0.4	0.049	0.474	0.346	0.036	0.338	0.285
		0.8	0.040	0.236	0.186	0.040	0.174	0.139
		1.2	0.053	0.482	0.350	0.074	0.315	0.266
		1.6	0.063	0.237	0.183	0.046	0.176	0.143
$\hat{\beta}_{2n}(u)$	20%	0.4	0.058	0.144	0.135	0.053	0.100	0.092
		0.8	0.045	0.128	0.112	0.036	0.091	0.081
		1.2	0.058	0.153	0.138	0.042	0.101	0.098
		1.6	0.037	0.127	0.108	0.028	0.087	0.078
	40%	0.4	0.066	0.195	0.168	0.063	0.140	0.129
		0.8	0.049	0.143	0.121	0.042	0.104	0.087
		1.2	0.060	0.202	0.177	0.053	0.147	0.136
		1.6	0.041	0.152	0.124	0.039	0.109	0.092
	65%	0.4	0.086	0.502	0.388	0.075	0.318	0.289
		0.8	0.050	0.243	0.184	0.045	0.186	0.147
		1.2	0.072	0.489	0.382	0.065	0.325	0.280
		1.6	0.055	0.247	0.182	0.053	0.187	0.142

表 6 不同模型、样本量和删失率下所取的带宽 h_n

删失率	方法	200			400		
		模型 1	模型 2	模型 3	模型 1	模型 2	模型 3
20%	Full	0.130	0.185	0.120	0.105	0.158	0.104
	Proposed	0.122	0.195	0.123	0.113	0.168	0.107
	Naive	0.248	0.414	0.181	0.199	0.299	0.149
40%	Full	0.145	0.188	0.122	0.104	0.155	0.102
	Proposed	0.123	0.221	0.139	0.123	0.184	0.118
	Naive	0.623	5.169	0.325	0.272	0.879	0.254
65%	Full	0.167	0.188	0.122	0.104	0.159	0.106
	Proposed	0.120	1.528	0.312	0.148	0.768	0.163
	Naive	7.372	14.415	8.377	2.163	5.229	3.783

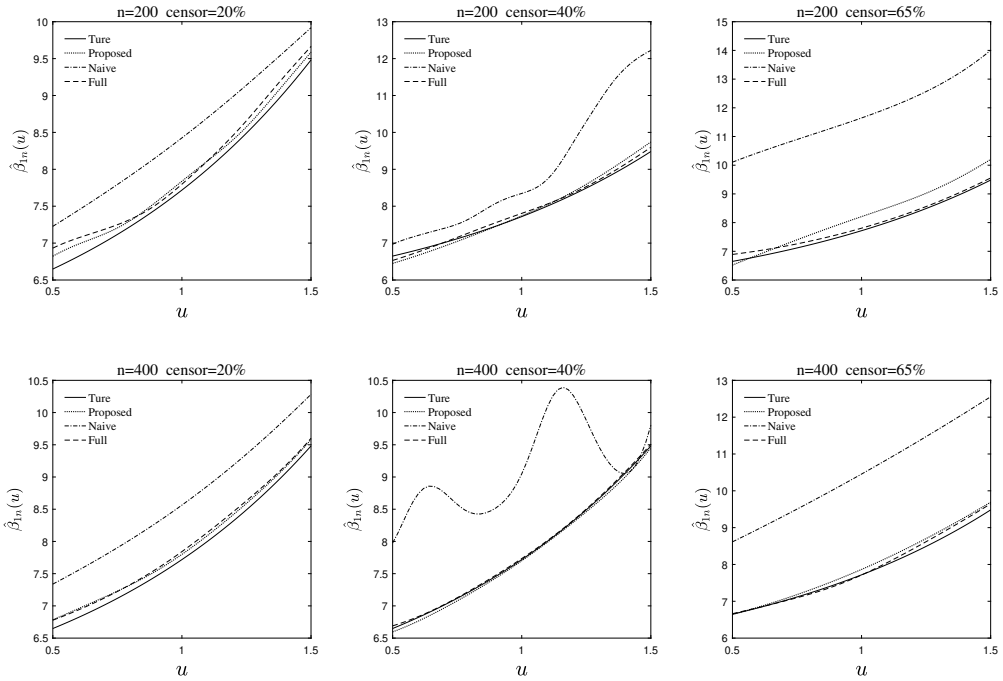

图 2　不同方法下模型 2 中函数系数 $\beta_1(u)$ 的估计曲线

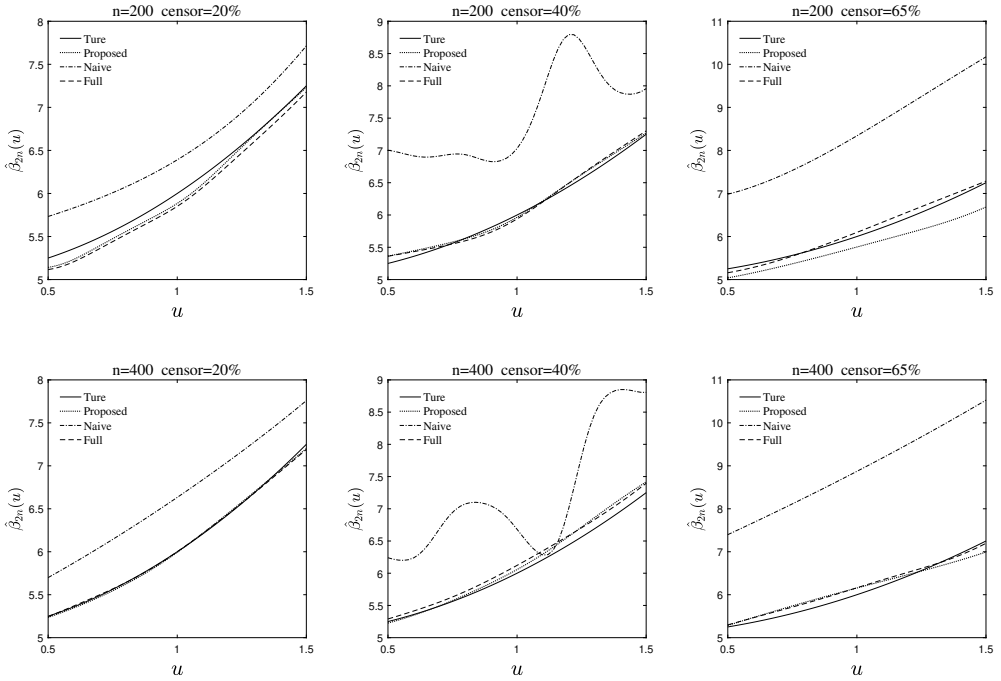

图 3　不同方法下模型 2 中函数系数 $\beta_2(u)$ 的估计曲线

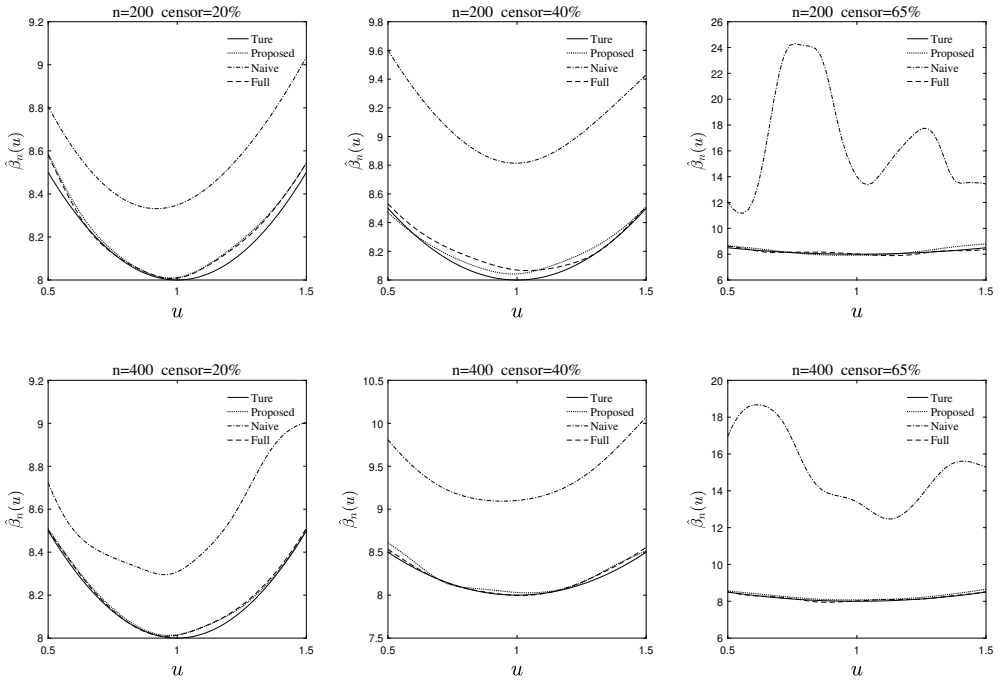

图 4 不同方法下模型 3 中函数系数 $\beta(u)$ 的估计曲线

6.1 弗雷明翰心脏研究 (FHS) 数据集

我们首先考虑弗雷明翰心脏研究 (FHS) 数据集的分析. 这个数据集来自于美国马萨诸塞州弗雷明翰社区的心血管疾病 (CVD) 病因学研究. 这是一项关于心血管疾病病因的前瞻性研究. 我们使用的是这个数据集的一个子集. 该数据集包含 4434 名受试者 1956 年至 1968 年共 3 个阶段的数据, 其中只有第 3 阶段包含有低密度脂蛋白数据. 故我们仅对受试者第 3 阶段的数据进行分析. 删除存在缺失值的数据之后, 数据集包含 1021 条记录. 目前已有很多学者分析了该数据. 例如, Atem 和 Matsouaka[21] 研究了低密度脂蛋白和患者心血管疾病发病年龄的关系; 张堉 等[22] 研究了吸烟对低密度脂蛋白和高密度脂蛋白的影响.

我们考虑低密度脂蛋白 (Y) 和发病年龄 (V)、高密度脂蛋白 (Z) 以及每日吸烟数量 (U) 之间的关系. 考虑具有截距项的变系数模型, 即采用模型:

$$Y = \beta_1(U)V + \beta_2(U)Z + \beta_3(U) + \varepsilon$$

来分析数据, 其中发病年龄 V 被随机右删失, 删失率为 76.7%. 注意在分析前我们对低密度脂蛋白 Y 取了对数再进行分析.

核函数取为 Gauss 核函数. 带宽通过文中所述的 CV 方法选取, 其值为 3.88. 对 Naive 方法, 经计算得到 CV 带宽为 3.50. 类似于文献 [23] 中的方法, 我们采用均方误差 (MSE)

来评价模型的拟合精度, MSE 的定义如下:

$$\mathrm{MSE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{Y_i} - \widehat{\mathrm{Y}}_\mathrm{i})^2 \delta_\mathrm{i},$$

这里 N 为未删失记录的个数. 对 $i = 1, 2, \cdots, n$, 当 $\delta_i = 1$ 时,$\widehat{Y}_i = \hat{\beta}_1(U_i)V_i + \hat{\beta}_2(U_i)Z_i + \hat{\beta}_3(U_i)$, $i = 1, \cdots, n$. 当 $\delta_i = 0$ 时, $\widehat{Y}_i$ 不用定义, 因为上式中对应的项 $(Y_i - \widehat{Y}_i)^2\delta_i = 0$. 对本文所提方法, 经计算得到 MSE 值为 0.068, 对 Naive 方法, 其 MSE 值为 0.071. 图 5 给出了系数函数的估计曲线, 图 6 是数据拟合的残差图和 QQ 图.

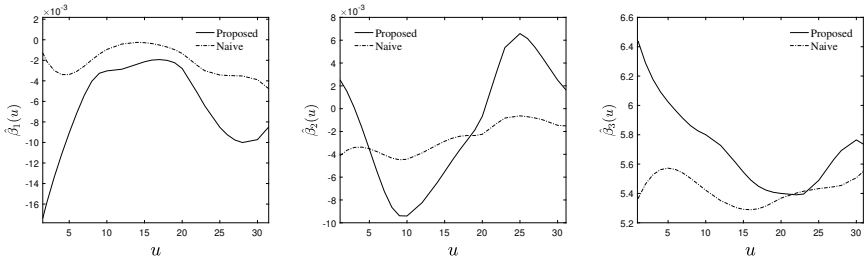

图 5　弗雷明翰心脏研究数据 $\beta(u)$ 曲线的估计

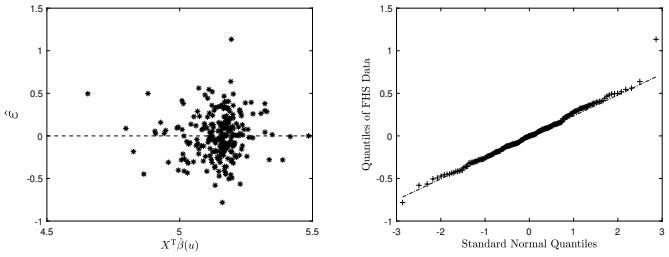

图 6　弗雷明翰心脏研究数据残差图和 QQ 图

从图 5 可以看到, 发病年龄对低密度脂蛋白的影响具有负效应, 且当每日吸烟数量较小时, 发病年龄对低密度脂蛋白的影响较大. 随着每日吸烟数量增加, 发病年龄对低密度脂蛋白的影响逐渐减小. 每日吸烟数量不同时, 高密度脂蛋白对低密度脂蛋白的影响变化较大. 另外从截距项的估计结果可以看到, 每日吸烟数量较多的受试者往往具有较低的低密度脂蛋白水平.

6.2　原发性胆汁性肝硬化 (PBC) 数据集

接下来, 我们利用所提方法分析原发性胆汁性肝硬化 (PBC) 数据集. 这个数据集来自于 1974 年至 1984 年进行的梅奥原发性胆汁性肝硬化 (PBC) 临床试验, 数据集及其详细描述见 http://lib.stat.cmu.edu/datasets. 该数据集存在竞争风险: 肝移植和死亡. 参考文献 [3] 的方法, 我们考虑生存时间为注册登记到死亡或者肝移植之间的时间, 其中生存时间因为实验结束被删失. 删除存在缺失值的数据后, 共有 318 个观察值, 删失率为 62.9%.

类似于文献 [24] 的设置, 我们研究血清胆红素 (Y), 生存时间 (V) 和年龄 (U) 之间的关系. 考虑用下面模型:

$$Y = \beta_1(U)V + \beta_2(U) + \varepsilon$$

来分析数据. 在分析前对 Y 取对数, 其余设置与 6.1 节相同. 经计算, 对本文所提方法, CV 带宽取为 7.02, MSE 值为 0.780. Naive 方法的 MSE 值为 1.042, CV 带宽为 18.27. 我们同样给出系数函数的估计曲线, 残差图和 QQ 图, 如图 7–8 所示.

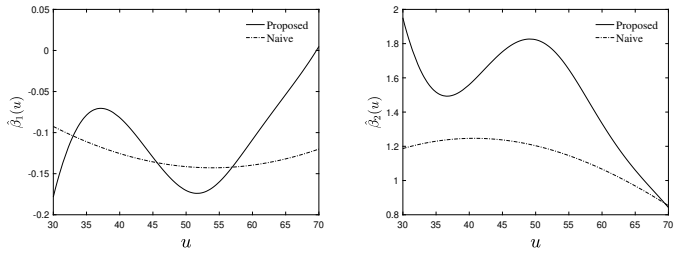

图 7 原发性胆汁性肝硬化数据 $\beta(u)$ 曲线的估计

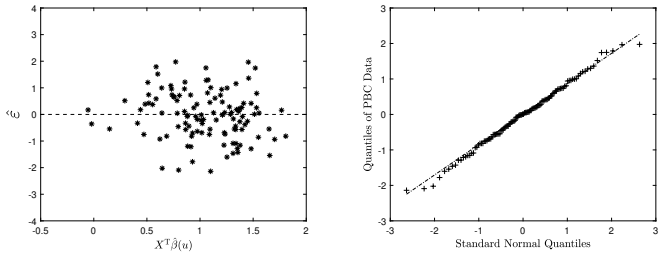

图 8 原发性胆汁性肝硬化数据残差图和 QQ 图

从图 7 可以看出, 生存时间对血清胆红素具有负效应影响, 随着年龄的增加, 生存时间对血清胆红素的影响逐渐变小, 而 Naive 的估计结果表明在不同的年龄水平下, 生存时间对血清胆红素的影响基本相同. 从所提方法的截距项估计结果可以看到, 血清胆红素水平随着年龄的增大而减小. Naive 方法的截距项估计结果则表示随着年龄的变化, 血清胆红素水平基本不变.

7 定理证明

下面列出定理成立所需的条件:

(C1) 函数 $\beta(u)$ 和 $\sigma^2(x,u)$, 以及条件密度 $f_{\boldsymbol{X}|U}(x|u)$, 关于 u 二阶连续可导;

(C2) $\sup_{x,u} \mathsf{E}(Y^2|\boldsymbol{X} = x, U = u) < \infty$;

(C3) $\sup_{x,u} \mathsf{E}\left(\frac{\varepsilon^2}{G(V)} \int_0^V \overline{H}^{-2}(s)\mathrm{d}H_0(s)\big|\boldsymbol{X} = x, U = u\right) < \infty$, 且关于 u 二阶连续可导;

(C4) 变量 U 的密度函数 $f_u(u)$ 满足: $0 < \inf_{u\in\mathbb{R}} f_u(u) \leqslant \sup_{u\in\mathbb{R}} f_u(u) < \infty$, 且关于 u 二阶连续可导;

(C5) 对任意分布函数 $L(t)$, 令 $\tau_L = \inf\{t : L(t) = 1\}$. 对 V 和 C 的分布函数 $\overline{F}(\cdot)$ 和 $\overline{G}(\cdot)$, 有 $0 < \tau_{\overline{F}} < \tau_{\overline{G}} < \infty$, 且 $\overline{F}(\cdot)$ 和 $\overline{G}(\cdot)$ 连续;

(C6) 核函数 $k(\cdot)$ 为有紧支撑的二阶核函数;

(C7) 带宽 h_n 满足下面的条件: 当 $n \to \infty$, $h_n \to 0$, $nh_n \to \infty$, $h_n^5 \ln n \to 0$, $nh_n^9 \to 0$.

引理 2　当条件 (C5) 成立时, 有 $\sup_{0<t<\tau_F} |\widehat{G}_n(t) - G(t)| = O\left(n^{-1/2}(\ln n)^{1/2}\right)$ a.s.

证明:　参见 Lo 和 Singh[18] 引理 3 的结论 (a). □

引理 3　在定理 1 的条件下, 对 $t < \tau_F < \tau_G$, 我们有

$$\sqrt{n}\left(\widehat{G}_n(t) - G(t)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \eta\left(T_i, \delta_i; t\right) + o_p(1).$$

证明:　参见 Lo 和 Singh[18] 定理 1 的证明. □

定理 1 的证明: 令 $\theta(u) = \left(\beta(u)^\top, h_n\beta'(u)^\top\right)^\top$. 根据式 (3), 有

$$\sqrt{nh_n}\big(\hat{\theta}_n(u) - \theta(u)\big) = \left(\frac{1}{nh_n}\widetilde{\mathcal{X}_u} W_u \widetilde{\mathcal{X}_u}^\top\right)^{-1} \frac{1}{\sqrt{nh_n}}\widetilde{\mathcal{X}_u} W_u (Y - \widetilde{\mathcal{X}_u}^\top \theta)$$

$$=: A_n(u)^{-1} B_n(u). \tag{4}$$

可以验证得到

$$A_n(u) = \frac{1}{nh_n}\sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_i \widetilde{\boldsymbol{X}}_i^\top \otimes \begin{pmatrix} 1 & (U_i - u)/h_n \\ (U_i - u)/h_n & \{(U_i - u)/h_n\}^2 \end{pmatrix} \frac{\delta_i}{\widehat{G}_n(T_i)} k\left(\frac{U_i - u}{h_n}\right).$$

注意到 $\delta_i T_i = \delta_i V_i$, $i = 1, 2, \cdots, n$. 因此有

$$A_n(u) = \frac{1}{nh_n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top \otimes \begin{pmatrix} 1 & (U_i - u)/h_n \\ (U_i - u)/h_n & \{(U_i - u)/h_n\}^2 \end{pmatrix} \frac{\delta_i}{\widehat{G}_n(T_i)} k\left(\frac{U_i - u}{h_n}\right).$$

由大数定律和引理 2, 可以验证

$$A_n(u) = f_u(u) \begin{pmatrix} Q(u) & 0 \\ 0 & Q(u)k_{21} \end{pmatrix} + o_p(1),$$

这里 $Q(u)$ 的定义在第 4 节给出. 进一步可以得到

$$A_n^{-1}(u) = \frac{1}{f_u(u)} \begin{pmatrix} Q^{-1}(u) & 0 \\ 0 & Q^{-1}(u)k_{21}^{-1} \end{pmatrix} + o_p(1).$$

对 $B_n(u)$, 根据 $\delta_i T_i = \delta_i V_i, i = 1, 2, \cdots, n$, 有下面的展开

$$B_n(u) = \begin{pmatrix} \dfrac{1}{\sqrt{nh_n}} \sum\limits_{i=1}^{n} \dfrac{\delta_i}{\widehat{G}_n(T_i)} k\left(\dfrac{U_i - u}{h_n}\right) \boldsymbol{X}_i \left(Y_i - \boldsymbol{X}_i^\top \beta(u) - \boldsymbol{X}_i^\top \beta^{(1)}(u)(U_i - u)\right) \\ \dfrac{1}{\sqrt{nh_n}} \sum\limits_{i=1}^{n} \dfrac{\delta_i}{\widehat{G}_n(T_i)} k\left(\dfrac{U_i - u}{h_n}\right) \dfrac{U_i - u}{h_n} \boldsymbol{X}_i \left(Y_i - \boldsymbol{X}_i^\top \beta(u) - \boldsymbol{X}_i^\top \beta^{(1)}(u)(U_i - u)\right) \end{pmatrix}$$
$$=: \begin{pmatrix} B_{n1}(u) \\ B_{n2}(u) \end{pmatrix},$$

这样我们可以得到

$$\sqrt{nh_n}\left(\hat{\beta}_n(u) - \beta(u)\right) = \frac{1}{f_u(u)} Q^{-1}(u) B_{n1}(u). \tag{5}$$

对 $B_{n1}(u)$, 由 $Y_i = X_i^\top \beta(U_i) + \varepsilon_i, i = 1, 2, \cdots, n$, 有下面的展开

$$\begin{aligned} B_{n1}(u) =& \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} \frac{\delta_i}{G(T_i)} k\left(\frac{U_i - u}{h_n}\right) \boldsymbol{X}_i \left\{\varepsilon_i + \frac{1}{2}\boldsymbol{X}_i^\top \beta^{(2)}(u)(U_i - u)^2\right\} \\ &+ \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} \left\{\frac{\delta_i}{\widehat{G}_n(T_i)} - \frac{\delta_i}{G(T_i)}\right\} k\left(\frac{U_i - u}{h_n}\right) \boldsymbol{X}_i \left\{\varepsilon_i + \frac{1}{2}\boldsymbol{X}_i^\top \beta^{(2)}(u)(U_i - u)^2\right\} \\ &+ o_p(1) \\ =& \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} \frac{\delta_i}{G(T_i)} k\left(\frac{U_i - u}{h_n}\right) \boldsymbol{X}_i \varepsilon_i \\ &+ \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} \frac{\delta_i}{G(T_i)} k\left(\frac{U_i - u}{h_n}\right) \frac{1}{2}\boldsymbol{X}_i \boldsymbol{X}_i^\top \beta^{(2)}(u)(U_i - u)^2 \\ &+ \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} \left\{\frac{\delta_i}{\widehat{G}_n(T_i)} - \frac{\delta_i}{G(T_i)}\right\} k\left(\frac{U_i - u}{h_n}\right) \boldsymbol{X}_i \varepsilon_i \\ &+ \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} \left\{\frac{\delta_i}{\widehat{G}_n(T_i)} - \frac{\delta_i}{G(T_i)}\right\} k\left(\frac{U_i - u}{h_n}\right) \frac{1}{2}\boldsymbol{X}_i \boldsymbol{X}_i^\top \beta^{(2)}(u)(U_i - u)^2 + o_p(1) \\ =:& \sum_{j=1}^{4} B_{n1,j}(u). \end{aligned} \tag{6}$$

对 $B_{n1,2}(u)$, 由大数定律和随机删失条件, 可得

$$\begin{aligned} B_{n1,2}(u) &= \sqrt{nh_n} \left\{\frac{1}{nh_n} \sum_{i=1}^{n} \frac{\delta_i}{G(T_i)} k\left(\frac{U_i - u}{h_n}\right) \frac{1}{2}\boldsymbol{X}_i \boldsymbol{X}_i^\top \beta^{(2)}(u) h_n^2 \left(\frac{U_i - u}{h_n}\right)^2\right\} \\ &= \sqrt{nh_n} \mathsf{E}\left\{\frac{1}{h_n} k\left(\frac{U_i - u}{h_n}\right) \frac{1}{2}\boldsymbol{X}\boldsymbol{X}^\top \beta^{(2)}(u) h_n^2 \left(\frac{U_i - u}{h_n}\right)^2\right\} + O_p(\sqrt{nh_n} h_n^4) \\ &= \sqrt{nh_n} \frac{1}{2} Q(u) k_{21} f_u(u) \beta^{(2)}(u) h_n^2 + o_p(1). \end{aligned} \tag{7}$$

接下来考虑 $B_{n1,3}(u)$, 利用引理 2 和引理 3 的结果, 可以得到

$$\begin{aligned}
B_{n1,3}(u) =&\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\frac{G(T_i)-\widehat{G}_n(T_i)}{G(T_i)\widehat{G}_n(T_i)}\delta_i k\left(\frac{U_i-u}{h_n}\right)\boldsymbol{X}_i\varepsilon_i\\
=&\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\frac{G(T_i)-\widehat{G}_n(T_i)}{G^2(T_i)}\delta_i k\left(\frac{U_i-u}{h_n}\right)\boldsymbol{X}_i\varepsilon_i+o_p(1)\\
=&\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\frac{1}{G^2(T_i)}\frac{1}{n}\sum_{j=1}^{n}\eta\left(T_j,\delta_j;T_i\right)\delta_i k\left(\frac{U_i-u}{h_n}\right)\boldsymbol{X}_i\varepsilon_i+o_p(1)\\
=&\frac{1}{n\sqrt{nh_n}}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{1}{G^2(T_i)}\eta\left(T_j,\delta_j;T_i\right)\delta_i k\left(\frac{U_i-u}{h_n}\right)\boldsymbol{X}_i\varepsilon_i\\
&+\frac{1}{n\sqrt{nh_n}}\sum_{i=1}^{n}\frac{1}{G^2(T_i)}\eta\left(T_i,\delta_i;T_i\right)\delta_i k\left(\frac{U_i-u}{h_n}\right)\boldsymbol{X}_i\varepsilon_i+o_p(1).
\end{aligned}$$

由大数定律可以证得 $\dfrac{1}{n\sqrt{nh_n}}\sum\limits_{i=1}^{n}\dfrac{1}{G^2(T_i)}\eta\left(T_i,\delta_i;T_i\right)\delta_i k\left(\dfrac{U_i-u}{h_n}\right)\boldsymbol{X}_i\varepsilon_i=O_p\left(\dfrac{1}{\sqrt{nh_n}}\right)=o_p(1)$. 因此我们有

$$B_{n1,3}(u)=\frac{1}{n\sqrt{nh_n}}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{1}{G^2(T_i)}\eta\left(T_j,\delta_j;T_i\right)\delta_i k\left(\frac{U_i-u}{h_n}\right)\boldsymbol{X}_i\varepsilon_i+o_p(1). \tag{8}$$

对 $B_{n1,4}(u)$, 由引理 2 的结果可以得出

$$B_{n1,4}(u)=\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\frac{G(T_i)-\widehat{G}_n(T_i)}{G^2(T_i)}\delta_i k\left(\frac{U_i-u}{h_n}\right)\frac{1}{2}\boldsymbol{X}_i\boldsymbol{X}_i^{\top}\beta^{(2)}(u)\left(\frac{U_i-u}{h_n}\right)^2h_n^2+o_p(1).$$

对上式的主项, 考虑其第 l 个分量, 记为 $B_{n1,4,l}(u)$, $l=1,2,\cdots,p$, 可以得到

$$\begin{aligned}
|B_{n1,4,l}(u)| =&\left|\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\frac{G(T_i)-\widehat{G}_n(T_i)}{G^2(T_i)}\delta_i k\left(\frac{U_i-u}{h_n}\right)\frac{1}{2}X_{il}\boldsymbol{X}_i^{\top}\beta^{(2)}(u)\left(\frac{U_i-u}{h_n}\right)^2h_n^2\right|\\
\leqslant&\sup_{0<t<\tau_F}\left|\widehat{G}_n(t)-G(t)\right|\frac{1}{2\sqrt{nh_n}}\sum_{i=1}^{n}\frac{\delta_i h_n^2}{G^2(T_i)}k\left(\frac{U_i-u}{h_n}\right)\left|X_{il}\boldsymbol{X}_i^{\top}\beta^{(2)}(u)\right|\left(\frac{U_i-u}{h_n}\right)^2\\
=&O\left(n^{-1/2}\ln(n)^{1/2}\right)O_p(\sqrt{nh_n}h_n^2)
\end{aligned}$$

由条件 (C7) 可以得到 $B_{n1,4,l}(u)=o_p(1)$, 进一步得到 $B_{n1,4}(u)=o_p(1)$. 这个结论, 和前面的式 (6)–(8)一起, 可以得到

$$\begin{aligned}
B_{n1}(u) =&\sqrt{nh_n}\frac{1}{2}Q(u)k_{21}f_u(u)\beta^{(2)}(u)h_n^2+\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\frac{\delta_i}{G(T_i)}k\left(\frac{U_i-u}{h_n}\right)\boldsymbol{X}_i\varepsilon_i\\
&+\frac{1}{n\sqrt{nh_n}}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{1}{G^2(T_i)}\eta\left(T_j,\delta_j;T_i\right)\delta_i k\left(\frac{U_i-u}{h_n}\right)\boldsymbol{X}_i\varepsilon_i+o_p(1)\\
=&\sqrt{nh_n}\frac{1}{2}Q(u)k_{21}f_u(u)\beta^{(2)}(u)h_n^2+\frac{2}{n\sqrt{n}}\sum_{i=1}^{n}\sum_{j<i}W(\Lambda_{iju})+o_p(1),
\end{aligned}\tag{9}$$

这里 $W(\Lambda_{iju})$ 可参考第四节给出的定义. 由文献 [25; 第 76 页定理 1] 容易证得定理 1 的结果. □

8 结论

在本文中, 我们研究了协变量随机右删失时变系数模型的估计问题. 根据文献 [17] 的方法, 我们利用逆概率加权方法对目标函数进行调整来处理数据的删失, 进而构建了回归系数的估计, 严格证明了所得估计的渐近正态性. 不同于以往文献 [7–10] 中处理删失协变量的方法, 本文方法不需要对模型或者出现删失的变量的分布进行任何假设, 从而不存在模型或分布误定的风险. 数值模拟和实例分析表明本文所提方法具有很好的有限样本性质. 本文发展的利用逆概率加权调整目标函数来处理协变量删失的方法可以推广至其他模型, 比如广义线性模型、部分线性模型等.

参考文献

[1] KOUL H, SUSARLA V, VAN RYZIN J. Regression analysis with randomly right-censored data [J]. *Ann Statist*, 1981, **9**(**6**): 1276–1288.

[2] MILLER R G. Least squares regression with censored data [J]. *Biometrika*, 1976, **63**(**3**): 449–464.

[3] SUN Z H, YE X, SUN L Q. Consistent test for parametric models with right-censored data using projections [J]. *Comput Stat Data An*, 2018, **118**: 112–125.

[4] RIGOBON R, STOKER T M. Bias from censored regressors [J]. *J Bus Econ Stat*, 2009, **27**(**3**): 340–353.

[5] AUSTIN P C, BRUNNER L J. Type I error inflation in the presence of a ceiling effect [J]. *Amer Statist*, 2003, **57**(**2**): 97–104.

[6] ATEM F D, QIAN J, MAYE J E, et al. Linear regression with a randomly censored covariate: Application to an Alzheimer's study [J]. *J R Stat Soc Ser C*, 2017, **66**(**2**): 313–328.

[7] MOULTON L H, HALSEY N A. A mixture model with detection limits for regression analyses of antibody response to vaccine [J]. *Biometrics*, 1995, **51**(**4**): 1570–1578.

[8] ATEM F D, MATSOUAKA R A, ZIMMERN V E. Cox regression model with randomly censored covariates [J]. *Biometrical J*, 2019, **61**(**4**): 1020–1032.

[9] LYNN H S. Maximum likelihood inference for left-censored HIV RNA data [J]. *Stat Med*, 2001, **20**(**1**): 33–45.

[10] MAY R C, IBRAHIM J G, CHU H T. Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits [J]. *Stat Med*, 2011, **30**(**20**): 2551–2561.

[11] HASTIE T, TIBSHIRANI R. Varying-coefficient models [J]. *J R Stat Soc Ser B*, 1993, **55**(**4**): 757–779.

[12] FAN J Q, ZHANG W Y. Statistical estimation in varying coefficient models [J]. *Ann Statist*, 1999, **27**(**5**): 1491–1518.

[13] CAI Z W, FAN J Q, LI R Z. Efficient estimation and inferences for varying-coefficient models [J]. *J Amer Statist Assoc*, 2000, **95**(**451**): 888–902.

[14] FAN J Q, ZHANG W Y. Statistical methods with varying coefficient models [J]. *Stat Interface*, 2008, **1**(**1**): 179–195.

[15] YANG S J, EI GHOUCH A, VAN KEILEGOM I. Varying coefficient models having different smooth-

ing variables with randomly censored data [J]. *Electron J Stat*, 2014, **8**(**1**): 226–252.

[16] CHEN M, YUEN K C, ZHU L X. Asymptotics of the goodness-of-fit test for a partial linear model with randomly censored data [J]. *Sci China Ser A*, 2003, **46**(**2**): 145–158.

[17] LOPEZ O, PATILEA V. Nonparametric lack-of-fit tests for parametric mean-regression models with censored data [J]. *J Multivariate Anal*, 2009, **100**(**1**): 210–230.

[18] LO S H, SINGH K. The product-limit estimator and the bootstrap: Some asymptotic representations [J]. *Probab Theory Rel*, 1986, **71**(**3**): 455–465.

[19] TANG Y L, WANG H X J, ZHU Z Y, et al. A unified variable selection approach for varying coefficient models [J]. *Stat Sinica*, 2012, **22**(**2**): 601–628.

[20] CAI Z W, FAN J Q, YAO Q W. Functional-coefficient regression models for nonlinear time series [J]. *J Amer Statist Assoc*, 2000, **95**(**451**): 941–956.

[21] ATEM F D, MATSOUAKA R A. Linear regression model with a randomly censored predictor: Estimation procedures [J]. *Biostat Biometric*, 2017, **1**(**2**): 21–32.

[22] 张堉, 郑诚东, 张代民, 等. 吸烟对健康人血清脂蛋白组分胆固醇水平的影响 [J]. 中国公共卫生学报, 1990, **9**(**5**): 261–263.

[23] MA X J, DU Y, WANG J L. Model detection and variable selection for mode varying coefficient model [J]. *Stat Methods Appl*, 2022, **31**(**2**): 321–341.

[24] TSIMIKAS J V, BANTIS L E, GEORGIOU S D. Inference in generalized linear regression models with a censored covariate [J]. *Comput Stat Data An*, 2012, **56**(**6**): 1854–1868.

[25] LEE A J. *U-Statistics: Theory and Practice* [M]. New York: Routledge, 1990.

Estimation of Varying Coefficient Model with Randomly Right-Censored Covariate

CHAI Wang[1] YIN Junping[2,3,4] SUN Zhihua[1]

[1] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, 100049, China

[2] Institute of Applied Physics and Computational Mathematics, Beijing, 100094, China

[3] National Key Laboratory of Computational Physics, Beijing, 100088, China

[4] Shanghai Zhangjiang Institute of Mathematics, Shanghai, 201203, China

Abstract: Data is often right-censored due to loss of follow-up, drop-out from experiments or end of clinical trials. The treatment of right-censored data has attracted the interest of many researchers. Most existing studies focus on the cases where the response variable is censored. Predictors in the regression model may also suffer from right-censoring. However, there are only sporadic works on the treatment of censored covariates. In this paper, the estimation of a varying coefficient model with randomly right-censored covariates is investigated. To deal with the right-censoring, we adjust the objective function directly through an inverse probability weighting, instead of adjusting the right-censored predictor. Estimation of the regression coefficient is proposed. The asymptotic properties of the proposed estimator are rigorously investigated. Numerical simulations and real-data analyses demonstrate that the proposed method has good finite sample properties.

Keywords: varying coefficient model; local linear method; randomly right-censored covariate; asymptotic property

2020 Mathematics Subject Classification: 62N02; 62G05